Data Analytics

Systems Engineering • Cybersecurity • Project Management

C. GRECO

Data Analytics

LICENSE, DISCLAIMER OF LIABILITY, AND LIMITED WARRANTY

By purchasing or using this book (the "Work"), you agree that this license grants permission to use the contents contained herein, but does not give you the right of ownership to any of the textual content in the book or ownership to any of the information, files, or products contained in it. *This license does not permit uploading of the Work onto the Internet or on a network (of any kind) without the written consent of the Publisher.* Duplication or dissemination of any text, code, simulations, images, etc. contained herein is limited to and subject to licensing terms for the respective products, and permission must be obtained from the Publisher or the owner of the content, etc., in order to reproduce or network any portion of the textual material (in any media) that is contained in the Work.

Mercury Learning and Information ("MLI" or "the Publisher") and anyone involved in the creation, writing, production, accompanying algorithms, code, or computer programs ("the software"), and any accompanying Web site or software of the Work, cannot and do not warrant the performance or results that might be obtained by using the contents of the Work. The author, developers, and the Publisher have used their best efforts to ensure the accuracy and functionality of the textual material and/or programs contained in this package; we, however, make no warranty of any kind, express or implied, regarding the performance of these contents or programs. The Work is sold "as is" without warranty (except for defective materials used in manufacturing the book or due to faulty workmanship).

The author, developers, and the publisher of any accompanying content, and anyone involved in the composition, production, and manufacturing of this work will not be liable for damages of any kind arising out of the use of (or the inability to use) the algorithms, source code, computer programs, or textual material contained in this publication. This includes, but is not limited to, loss of revenue or profit, or other incidental, physical, or consequential damages arising out of the use of this Work.

The sole remedy in the event of a claim of any kind is expressly limited to replacement of the book and only at the discretion of the Publisher. The use of "implied warranty" and certain "exclusions" vary from state to state and might not apply to the purchaser of this product.

Data Analytics

Systems Engineering • Cybersecurity • Project Management

Christopher Greco

Mercury Learning and Information

Dulles, Virginia
Boston, Massachusetts
New Delhi

Publisher: David Pallai
Mercury Learning And Information
22841 Quicksilver Drive
Dulles, VA 20166
info@merclearning.com
www.merclearning.com
(800) 232-0223

C. Greco. *Data Analytics: Systems Engineering • Cybersecurity • Project Management*
ISBN: 978-1-6839264-81

The publisher recognizes and respects all marks used by companies, manufacturers, and developers as a means to distinguish their products. All brand names and product names mentioned in this book are trademarks or service marks of their respective companies. Any omission or misuse (of any kind) of service marks or trademarks, etc. is not an attempt to infringe on the property of others.

The use of US government Websites and data are permitted under the rules governed at this site: https://www.usa.gov/government-works. Using the information on these sites does not mean this book is endorsed in any way by any agency of the US Federal Government. The opinions in this text are only those of the author.

Library of Congress Control Number: 2020952869

212223 321 This book is printed on acid-free paper in the United States of America.

Our titles are available for adoption, license, or bulk purchase by institutions, corporations, etc. For additional information, please contact the Customer Service Dept. at (800) 232-0223(toll free). Digital versions of our titles are available at: www.academiccourseware.com and other electronic vendors.

The sole obligation of Mercury Learning And Information to the purchaser is to replace the book and/or disc, based on defective materials or faulty workmanship, but not based on the operation or functionality of the product.

CONTENTS

PREFACE

The idea of data analytics has been applied to data since the beginning of human existence. People may say that we are just maturing in this endeavor, but time has shown that humans are more than ready to accept data and analyze that data to achieve greatness. When Wilbur and Orville Wright flew in 1903, they did not do so on a whim, but gathered data to see how and when their airplane would fly. There was a book written several decades ago about the owner of a TV dinner brand that made a million dollars by understanding the demographics of the time. The author saw the data on homemakers who were tired of making dinner every night. The TV dinner was a way to provide a solution to this issue. As I recall, the work of selling the product was hard, but the outcome was successful. Another book that showed data is the focus of human existence is the book *Outliers: The Story of Success* by Malcolm Gladwell [Gladwell-2008]. Gladwell highlights the idea that creativity, linked with data analysis, could lead the user and analyst to new perspectives, pointing out the successes of this activity.

If you have not heard Gladwell's talk on *Pepsi*, watch it online at (*https://www.youtube.com/watch?v=VkhFh5Ms1vc*). You will forever be changed about the use of data in any form.

Tool-centric learning has to be mitigated since "buttonology" is not the way to really teach analytics (or any other subject for that matter – other than those classes that teach the tool details). Therefore, this book is helpful to an instructor of the craft. Please buy this book and use it as a classroom study guide. Please enjoy and share the content in this book with your students and other teachers, and feel free to provide feedback. I would love to hear from you about the book and how to make it better. You can contact me through my website, *www.grectech.com*, and I would be happy to answer your questions. Thank you in advance for your help and support.

One last item, there are screenshots of US government websites that are permitted under the rules governed at this site: *https://www.usa.gov/government-works*. Using these sites and the information on them does not mean this book is endorsed by any agency of the US Federal Government. The opinions in this text are only those of the author.

ACKNOWLEDGMENTS

In this, and every other endeavor, there are people who are the unsung heroes who help and provide support in the background, but in the end, are never listed on the cover of the book. Of course, without my family supporting me, this would have been a failure. My wife, daughters, son-in-law, and grandson are all part of the family unit that helped me get through this. It is not a sprint but a marathon, and thanks to you all, I ended up finishing the race. My siblings are part of this support unit and I thank them for the consistent positive support they give me. In addition, I want to thank my friends, like Greg, Barbara, Rich, and others, whose consistent kindness and generosity make me want to share my work. My company motto is "Learn, Offer, Value, and Educate." You likely figured out that the first letters of each word spell "LOVE," and that is what makes us all a success. I would also like to thank Jennifer Blaney and the copy editors for their diligence and work on this product.

CHAPTER 1

INTRODUCTION TO STATISTICS FOR DATA ANALYSTS

Data analytics is creeping into the lexicon of our daily language. If the remarks heard throughout the day include “big data” or “AI” or “algorithm,” then data analytics is probably involved. This book gives the reader a perspective as to the overall data analytics skill set, starting with a primer on statistics, and works toward the application of those methods. However, and most importantly, this book includes the use of data that is available to anyone to analyze, and it can be used as a reference for more complex data analytics tasks. As the reader will experience in this book, there are a variety of formulas and algorithms used in the data analytics process. These formulas are there so that no matter what technology the reader uses, they will be able to plug the formula into the software application to get the answer they need. There are several demonstrations of this process in the book, so please do not worry about knowing the formulas in detail as much as being able to recognize and use them.

Every statistical formula contains some symbols that may not be recognized by the reader, and this is OK. An explanation of those symbols is included in the formula description so the reader can understand what the formula is trying to calculate. Think of the formulas as a universal key to a problem. Once the formula is translated into “buttonology” or the procedure for the software, the reader will be able to achieve the result for that specific analysis. Although data science tools are not covered in detail in this book, some are introduced. Many who are reading this text have been inundated with suggestions for using one or two common statistical tools, but this book is more "tool agnostic" than "tool endorsing." By using formulas, the tool chosen makes no difference. The main point is that no matter what tool the reader uses, the formula can be used with that tool.

1.1 Objectives

This section describes the objectives of this book, starting with the most fundamental of definitions. There are analysts who do not understand what data entails, and thereby jump to conclusions without really understanding what the data implies or contains. In law enforcement, there is a modus operandi (or method of operation, otherwise known as the "MO") that criminals use in order to steal, scam, or otherwise fool the law-abiding citizen. It is the same way with data, since data has trending or anomalous events that reveal what the analyst is trying to answer.

After understanding the definition of data definition, we review statistical fundamentals. This is a perfunctory approach to the fundamentals, but introduces certain concepts that were probably not raised in your "Introduction to Statistics" courses.

The "Occam's Razor" approach to data analysis is a very quick overview of keeping it simple. Named after William of Occam, this adage is the quintessential reason to follow to make data analytics simple but effective. There are colleagues who have addressed the yearning of students to go straight to regression analysis instead of trying a simpler approach to data analytics. A story making the rounds through the statistics realm is one where a junior analyst comes to the senior analyst and expresses excitement over using regression analysis to find an answer. The student is met with the following from the senior analyst:

"Why would you use such a complex method to find the answer? There has got to be a simpler way to figure this out!"

The student was hurt.

"That is just mean, sir."

The senior analyst replied,

"You are correct! The mean or average is a much better indication in this case!"

Simple should be the watchword of the analyst. This section will get into a little more detail on this concept.

Data origination and data validity raise the specter of what is called *data dilution*. In the old days, there was a drink made with a packet of artificial color, sugar, and water. The more water added, the more the drink tasted more like red (or green or orange) water and not the flavor intended. That is the same with data dilution. If the data is *transformed* three or four times, it becomes a shadow of what it once was, especially if the last analyst did not perform a data dictionary and left it up to the next analyst to figure out what the data meant in the first place!

1.2 The Three Professions

This book presents a variety of methods and techniques, as well as case studies, to enrich the knowledge of data analytics for project managers, systems engineers, and cybersecurity professionals. There are many times when a project manager needs to analyze data obtained from a product owner or show a stakeholder the consequences of actions implemented during the project. The development of any product process can be richly accompanied by data that the systems engineers need to analyze. The cybersecurity professional is inundated with data regarding the confidentiality, integrity, and availability of the software or hardware application. This can range from the port use to the access frequency. In all these cases, data analytics skills can, and should, be applied to each of the professions. This text separates the case studies so that each profession can practice some straightforward data analytics. The main purpose of this text is not bore the professional with elementary statistics but refresh the knowledge necessary to build models for data analytics. Along with that, this book encompasses the analytics thinking that is essential to all the professions. Without this basic critical thinking skill, the project manager, systems engineer, or cybersecurity professional will spin their wheels in useless analytics that waste time and money. The one section that covers Occam's Razor helps the analyst go straight to the simple rather than the complex. Not many people go to the Grand Canyon to see the composition of the rock formations. They go for the view. Think of the stakeholder like that tourist. Give them an overview of the data. If they want more, then massage the requirements to do so.

CHAPTER 2

WHAT IS DATA?

The beginning of a data analytics book should begin with exactly what "data" refers to and what to do with it. The working definition for this book is that data refers to *information gathered for analysis and subsequent discussion.* The reader is not reading this book to determine the factual verification of the data, but to use the data with conventional techniques or methods. Performing the process of the verification and validation of datasets is prudent and diligent, but the preliminary stages of analysis take place after the verification and validation process. For instance, if the reader obtains data from a federal website (like *www.cdc.gov* for the Centers for Disease Control), then the validation and verification do not need to be completed, since that data is already prepared and distributed via the Cross Industry Standard Process for Data Mining (CRISP-DM) method. Current data analytics training for government analysts includes the CRISP-DM methods.

2.1 Data Types

There are two types of data: *quantitative* and *qualitative*. These two types are broken down further into *discrete* or *continuous* (for quantitative) and *categorical* or *nominal* (for qualitative). Categorical data may refer to a range of numbers that could be considered quantitative, but for the most part, categorical data is qualitative in nature for the purposes of this book (and from experience). For instance, in a dataset that includes all tornados that occurred in 1950 (from the website *www.ncdc.noaa.gov*), the damage in dollars was categorized with K or M to denote thousands or millions, respectively. By assigning that designator, the data originators changed a quantitative measure to a qualitative measure, even though the field focuses on dollars. In this instance, by changing the type of data, the originators limited that data description because the mean and standard deviation cannot be calculated on that field without changing it to a quantitative measure. That is why it is important to consider the analysis value at the outset of a data analysis task.

2.1.1 Quantitative values

The quantitative values in our lives present us with meaning. Getting an 80 on a test may not mean anything, but 80 out of 100 gives us a better understanding. In most situations, more information may lead to a better understanding. Understanding also applies to perspective. If the analyst can understand the perspective of the data, then that analyst can find meaning as the data applies to other parts of the narrative. For example, during the COVID-19 pandemic, the number of cases (or deaths) was often discussed, but this number was not related to entire populations. At one point, there were over 100,000 cases of the COVID-19 reported in the US. Taken by itself, this number is frightening, but taken as a percentage of the total US population, which is 320,000,000, the number is put into context. (That percentage, by the way, is 100000/320000000 or .0003125 or .031%, which is a very small percentage when we take into consideration the entire population of the US.) If that is the situation, then that would mean quantitative measures are better than qualitative values. Of course, there is a way to transform qualitative to quantitative, which is discussed later in this book.

There are times when quantitative refers to different types of numbers. For instance, *discrete* quantitative measures denote those numbers that are integers. If the data describes how many TVs there are in a home, then that would be a discrete measure. There is no way to have 1.5 TVs in a home. By having discrete measures, there are certain statistical methods that may apply while others do not apply. Continuous quantitative data takes into consideration numbers that apply to other types of quantities. For instance, it is essential that the analyst determine the type of quantitative measure prior to applying certain methods to this data. For example, in the CRISP-DM process, one of the first steps is to understand the data. As obvious as this sounds, some analysts do not understand the data that they are analyzing, but conduct tests and build models on the data regardless of the analytical requirements. Although this is relatively common, the results from such models and tests may not be the results that align with verified and valid requirements.

Please remember that the quantitative data is important to determine measurement. For instance, during the COVID-19 outbreak, it was common to report the numbers of individuals that tested positive for the virus. This was an important measurement, but alone it did not give the relative impact on either the United States or the world. Another example of the quantitative nature of the statistics related to the virus is that, at one point, there were 3,000 individuals who tested positive for the virus in the US. This number means 3,000 people out of the entire US population of 320,000,000 tested positive for COVID-19. From a percentage viewpoint, the amount of people testing positive is less than .0009%, which is .000009! This number is so small that it would be considered miniscule in other situations. At that point in time, less than .0009% of the US population tested positive for this virus. Notice the phrase "tested positive" rather than "had the virus." Analysts need to be clear about this phrasing, since there have been indications that there are false negatives in the virus testing, as well as false positives[1]. What a false negative means is that the test may not be able to detect the existence of COVID-19 in the individuals who have it. Understanding the data and determining the type of quantitative measure used is the analyst's responsibility. However, this decision may have implications far beyond the narrow application of methods on a single source of data.

2.1.2 Qualitative values

Many readers have been asked to fill in surveys on services received from restaurants, hotels, and other businesses. These surveys usually have a range of numbers (i.e., from 1-10) that associate numbers to feelings. These measures are qualitative values transformed into quantitative values to measure customer service. After all, the ability to quantify "like" or "dislike" is very challenging. The first part of the challenge is that one person's "like" might be another person's "dislike." Unfortunately, because we are accustomed to these types of somewhat arbitrary definitions of rating personal preferences, people accept the range of numbers as something that can be transformed from feelings. The types of qualitative data can be placed into two different types – nominal and ordinal. For instance, nominal values can be "male" or "female" or "yes" and "no." Ordinal values describe a sequence, such as "no high school" to "graduate school." In these instances, as in all instances of qualitative measures, these measures can be changed into quantitative measures. For example, substituting a "0" for male and "1" for female is acceptable as long as the methods used to analyze do not include descriptive statistics (like mean and median) because it would look

[1] *https://www.bbc.com/news/health-51491763*

like females have more value than males (because they are 1 more than males on the numeric scale).

When two choices are used for qualitative data, it is considered a "binary" choice. If the choices include discrete numbers, then we use the term *binomial*. This type of data considers a problem like a multiple-choice test that has a right or wrong answer. This book does not discuss these types of data analytics.

Qualitative data includes the binary data mentioned above as well as categorical data. *Categories* are data that usually have a range of values and are used to denote certain types of behavior or characteristics. The two types of categorical data are nominal and ordinal. *Nominal data* (nominal refers to the "name") is a label of data that does not consider the ordering of the data, such as labeling male and female characteristics[2]. In this way, labeling can be accomplished without any worry about which category goes first. *Ordinal data* is not just labeled, but placed in a sequence, such as "high school," "some college," or "college graduate." In this case, there is a sequence and a range for the data. The real issue here is arbitrarily trying to make nominal data numerical data because that could lead to certain types of conclusions. For instance, say that an analyst assigns "1" to male and "0" to female. A less experienced (or inexperienced) analyst tries to do a descriptive analysis with this data and finds (surprise!) that the males have a larger mean than the females. Well, that was a mistake waiting to happen. If anything, the proportion would be a much better way of analyzing this data, in the same way that having nominal data for Democrats and Republicans would be something of an issue in this instance. What if the analyst assigns a "1" to a high school graduate, and a "5" to a post-grad student? This could lead to a very skewed analysis and show that high school graduates are not as important as post-grad students.

2.1.3 Application of Each Type of Data

Now that each type of data has been reviewed, when would an analyst use them and for what purpose? Quantitative data is used to determine data-set configurations (such as skew or kurtosis) as well as trends and other models, while qualitative is used to satisfy a specific need, such as a survey or other requirement. Quantitative data also satisfies a specific need, but

[2] *https://corporatefinanceinstitute.com/resources/knowledge/other/nominal-data/*

is less judgmental in the approach, relying on actual numbers rather than feelings transformed into numbers. There are many times when qualitative data is translated into quantitative data, and this can lead to problems. For instance, if the dataset has a gender category with "male" and "female," this is fine for a qualitative measure. Unfortunately, the natural instinct is to use a number instead of the gender, making male = 1 and female = 0. Although this is acceptable for proportional or probability measures (i.e., What is the percentage of males in the company? What is the percentage of the females?), this would not be acceptable when making a quantitative measurement, such as the mean, median, mode, standard deviation, kurtosis, or other measure. What happens is that the mean of the result of this measure might be 1.2. What does that mean? What does it mean when there are more males than females? What would a kurtosis of this dataset show? If one is truly an analyst, then the assignment of these values is something that makes this type of analysis of a qualitative value somewhat useless. If a project manager were to rate every success of their projects a 0 and failures a 1, what happens when the mean comes out 1.2? Without any qualifying aspects of this measurement, the numbers could be read as if the project manager were a failure, when in fact that may not be the case. What if the probability of failed projects is 60%? Does this make more sense? The same examples could be done with systems engineering and cybersecurity. The way that the data is translated and the form that it takes can make what seems to be a very similar transition raise issues with the resulting calculations and analysis.

In project management, systems engineering, and cybersecurity, there is a risk analysis (which is discussed in more detail later). Part of this risk analysis is translating subjective ratings, like "low," "moderate," and "high," into actual numerical ratings, or worse, a series of ratings. These are great for the short term, but can cause contradictions when used in isolation without any accompanying narrative. The idea behind any subjective number is that it is just that – the number *is* subjective, and in many cases a guess, educated or not.

This qualitative vs. quantitative concept battle has been going on in data science and will continue well into the next century. There are methods to analyze qualitative values, and most of them are in the probability section of this text. Remember that, if it is qualitative, then proportion or probability is going to be the way to analyze it.

The next section starts down a road that many readers have traveled previously – statistics. This is a must for data analysts because there are some methods within the statistical realm that are not just very applicable to data analytics, but very simple to apply. Remember that the whole focus of this book is to begin with the simple and move toward the complex.

CHAPTER 3

STATISTICS REVIEW – MEASURES OF THE CENTRAL TENDENCY

It is interesting when someone draws a circle. The circle takes a shape, but if the person wants a perfect circle, they should use a protractor and start with a center point to ensure that the area of the circle is the same in all directions. This is the same as the concept of the "central tendency in statistics" (or *central tendency*). The idea of finding the center of the data, much like having the center of the circle, is the perfect beginning to the analysis of that data. These measures are not only vital to determining the center of the data, but essential for the fundamentals of data science. The first measure of the central tendency that most people have some familiarity with using is the mean, or average. Before delving into this concept, remember back in high school when the teacher was handing out the test results? She may have described the average of the test. This is not only a measure of center, but also a measure of how well the student did in relation to other students. An average or mean is the focal point of many analytic beginnings and should *always* be considered at the beginning of the analysis because it helps describe the data. Some analysts gloss over the mean as just a forced calculation to get to the more detailed analysis, but in actuality, the mean is probably the best way to determine some basics about the data. For example, in many convenience stores, there is a height chart so that the police can make a quick description of an individual should that individual rob the store. Think of the mean as this chart, a place to quickly observe and describe data now and in the future.

3.1 Mean

Most people use the mean and do not even realize they are doing so. They prefer using the mean because it is relatively straightforward and gives a quick result. However, and this is important, the mean is impacted by high and low numbers. These very high and low numbers in a dataset are called *outliers* and are described in a handbook by a federal agency (*www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm*). Outliers exist in almost every dataset and can impact the mean in a very profound way. An example is an ad for a certification for which the announcer stated that the average salary of someone with this certification was approximately $65,000 per year. Although this sounds inviting, the announcer failed to adequately describe the data, since salaries in some states are much more than other states, but ultimately buy the same amount. The high salaries allowed the average to creep up, showing a much more inviting result than if the outliers were removed. Some analysts remove outliers to make the data more realistic, but in actuality, removing any data from the dataset does affect the future models. If the analyst decides to remove the outliers, great care must be taken to mention this is the methodology notes to tell other analysts that this was done to make the data have a better central tendency. Otherwise, those values are lost forever and may be necessary to understand other parts of the data.

The following formula shows the calculation of the mean. Please do not let the symbology make it seem as if the calculation is difficult. What the symbols mean is explained below:

$$\frac{\Sigma x_i}{N}$$

The funny looking "E" is the capital "S" in Greek, called *sigma*. Many analysts are already familiar with this symbol because it represents the "sum" in statistical software, including Open Office and Excel.

The "x" with the subscript "i" means that the analyst takes the sum of the values in the column or field of the dataset. The "x" represents the value and the "i" represents the number of that value in the dataset. For example, if one of the values of the dataset lies in the 5th position, and that is the only value calculated, then the analyst would see the "x" with a "5" as the subscript. Basically, the symbols state that all the values are to be summed. The last symbol is the "N" as the divisor (that is the number on the bottom of the fraction). This letter means the population of values. In statistics classes, it is sometimes called the number of numbers or the number of values. If the dataset contains 10 numbers or values, then the mean would be the sum of those values over 10. This formula can be used with any statistical software.

This book does not focus on software as much as formulas because formulas are the algorithms of statistics. Once the analyst understands the formulas, then the analyst can apply those formulas to any software available on the market. The idea of this book is to give the analyst the flexibility to try numerous software products to help the statistics methodologies.

One area that the analyst should understand is estimation. In this case, estimation is relatively straightforward because the analyst just needs to round the numbers, add them, and then divide by the number of values. For instance, assume the analyst has a dataset. The analyst could quickly take numbers like 80, 90, 50, 70, and 90 and add them together, making 360, and divide that number by 5, which would have a mean of around 72 (the actual mean is 77). For a quick estimation, at least the analyst knows that the dataset should have a range in the 70s. If the answer is 120, then the analyst knows that something is wrong. Another method for estimation is to take the lowest number and the highest number, add them together, and then divide by 2. In this case, it would be 90+50, which would be 140, and then divide it by 2, which would be 70. Again, this estimate is only a method to arrive at a quick result in order to make a comparison. However, in time-sensitive situations, estimation can be a very valuable tool that helps the project manager, systems engineer, or cybersecurity analyst make a quick decision or quickly see the range of the result to check the calculation process.

3.1.0 Averaging with the PERT Method

A quick note about the project manager's or systems engineer's use of the mean. In project management, as in systems engineering, there is a method called the PERT method (derived from the Program Evaluation and Review Technique), from which whole sections of classes are derived (*sites.psu.edu/se505/lesson-8-project-planning-and-scheduling/*). The PERT form of the cost and scheduling estimation takes the *Pessimistic* amount of the cost or time, the *Optimistic* amount of cost or time, and the *Most Likely* amount of cost or time. The Pessimistic, Optimistic, and Most Likely (which is multiplied by 4) values are added together and divided by 6. This method basically makes 6 values out of 3 by multiplying the Most Likely by 4. The formula is (O + 4 * M + P)/6. In this way, the whole idea is to do simple averaging with a twist. The twist is what is called *weighting* a value. The value of Most Likely is weighted 4 times more than Optimistic or Pessimistic. In this way, there is a faux normal distribution that is obtained from this method (*www.projectcubicle.com/pert-method/*).

3.1.1 Geometric Mean

A type of mean that is rarely raised or discussed in a statistics class is the geometric mean. The previous mean is sometimes called the *arithmetic mean* because it is the sum of values. The *geometric mean* is the multiplication of values. Instead of taking the division route, the geometric mean takes the root of the values based on the *N* that was previously used as a divisor. The formula for this calculation is shown below, and again, please do not be distracted by the various symbols.

$$\sqrt[N]{\Pi\, x_i}$$

The joined pillar symbol that looks like a large pi symbol is actually the capital Greek letter *pi*. What this symbol denotes is the multiplication of all events shown here as x with a subscript of I representing the next value to be multiplied. For instance, if there are 4 values, then the i would be substituted for 4, with the first value at $x_1 * x_2 * x_3 * x_4$ covering all four values. Once all values or events are multiplied, then the analyst takes the root of that number based on the number of events. As an example, if there are 3 numbers – 3, 6, and 2 – and the analyst wants to take a geometric mean, they would multiply the three numbers, 3 * 6 * 2 = 36, and then take the cube (3) root of that number to get the answer, which is approximately 3.30. Now, the analyst could also use logarithms if a calculator is not available, but in this day and age, both calculators or computers are available to virtually anyone. The questions that arise with this type of mean are "When is it used?" and "Why would any analyst use this formula?" In essence, this formula is used with averaging numbers that represent the growth or decline of things, like the interest rate. When a geometric mean not used? When there is a "0" in the data, because multiplying anything by "0" equals "0." This would negate any value that the geometric mean may bring to the analysis.

For example, we can calculate the geometric mean by multiplying 1.1 * .8 * 1.3 = 1.144 and then taking the cube root (since the total number of the values is equal to 3) of 1.144, which is approximately 1.05. If this were an interest rate on a retirement account, then even though the rate falls the second year by 20%, overall, the investment grew by 5% over the three-year period (thanks to the 30% growth the third year – whew!). Otherwise, the geometric mean could be an issue if *any* of the numbers are 0, because anything multiplied by 0 equals 0. In other words, under normal situations, if there are any zeroes in the data, geometric mean will not work as stated previously. The repetition of these caveats is important so the analyst can understand that there are limitations to using a geometric mean.

3.2 Median

The second measure of the central tendency is commonly called the *median*, but is also considered to be the physical center of the data. Unlike the mean, the median is not impacted by very large or very small numbers, because the physical center is always the physical center based on the number of values, not the amount of the values. This is an important distinction that deserves some extra page space.

The whole reason for the median is to find the value of the physical center, but not by adding or dividing the values of the different events. It is to find where those events lie in the dataset once the values are place in numerical order. This is important because not doing the first step, that is, placing the values in numerical order, will lead to erroneous median results. Current computer and calculator programs do not require that numerical order be maintained, but it is important for the analyst to understand that this does take place in case some manual calculation is required.

The median is crucial to finding not just the center of the data, but also the percentiles within the data. The median is considered the 50th percentile, which means that the "median of the median" can be the 25th or 75th percentile based on whether that value is above or below the median. There will be more on this later in the text, but the median is probably one of the most important values for the analyst (besides the mean).

The formula for revealing the median is the actual process of finding the median, irrespective of the tools employed. Once the data is sorted from least to greatest, the center of that sorting is the median. Does it sound too easy? That is because not all datasets, once sorted, end up with just one number in the center of the set. For instance, if the number of values is 9, then the 5th number is the median. Simple enough for this example, but what happens if the number of values is 10? That means that the 5th and 6th values are the center of the set. In that case, the median would be the mean of those two numbers. This brings us back to the mean, which is the total of the values divided by the number of values. For instance, if the center of a 10-value dataset is the numbers 9 and 11, then the median would be 9 + 11 divided by 2. Why 2, when it is a 10-value set? Because the analyst only wants to find the arithmetic mean between the two physically centered values. This sounds complicated, but in actuality, this works every time. The median in this case is 20/2 or 10. It makes sense when the analyst analyzes the different case studies using the median.

In this exercise, the analyst is faced with 5 house prices, not sorted in order. The first thing to do in order to find the median is to sort these values going from least to greatest. This would give the analyst the following:

50,000; 75,000; 100,000; 150,000, 300,000

From this list, the analyst would pick the middle number, which is 100,000. This number represents the median of this dataset. To show you the difference between the median and the mean, the mean of the dataset is 135,000. This is quite a difference from 100,000 and shows how much just a few numbers on either side of the dataset can "draw" the average to one side or the other. In this case, the 300,000 drew the average to the right of the median. Later on, we focus on how the dataset is "skewed" to the right if the mean is to the right of the median. This means that the dataset is not symmetrical and may not relate to a "normal" or standard distribution. More on this later.

3.3 Mode

The mode is probably one of the most misunderstood forms of the measures of the central tendency. In my many years of statistics instruction, the mode always seems to take a back seat to the more complicated methods that exist and there is no reason to ever dismiss the mode, specifically when trying to find a pattern of events. For instance, if a government agency wanted to know when most of the customers came into a field office, it could track that number over a period of weeks, months, years, or decades and see when the curve showed *humps*. Those humps represent when the most customers visited the field office. In other words, the mode is a value that represents the most prevalent value in that dataset. There is no magical formula, and in most instances, this is much easier to do with a computer than by hand, but it is just an important (if not more important) than many other measures. The best way to illustrate this is not through formulas and calculation, but looking at visual representations of the data. Figure 3.1 shows a graph that tracks customer data, with the x-axis being time and the y-axis being the number of customers. What does this graph show the analyst?

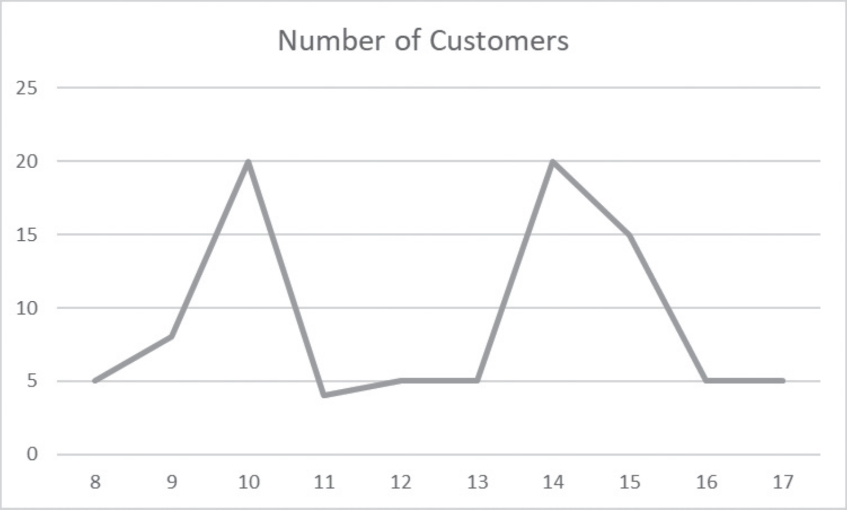


FIGURE 3.1 Graph of contrived data showing a bi-modal curve

When the analyst looks at Figure 3.1, it is unmistakable that the two humps (commonly called a bi-modal graph) represent when the most customers came into the office. What does this mean? This would mean that very little office visitation is done during lunch, indicating that management can give the employees a good lunch hour that would allow them to take a break from work. Statistics is not just a method; it is way to see what decisions could benefit others using data (at least in this case).

In a more recent use of modes, most of the COVID-19 data was related to how many cases there were in any particular part of the country or in a specific part of the United States per state. Analysts used the mode to follow any patterns that might occur in any county, city, or other community within a particular state. The mode is something that should never be ignored or shoved aside in favor of a more complicated statistical method. Sometimes, it is the simplest solution that really is the best.

3.4 Data Skew

It is important at this juncture to address data skewing since it is something that happens when a great amount of data values pools toward one end or the other of the dataset.

The formula for calculating skew in a dataset is relatively straightforward and is illustrated in the following section, courtesy of the National Institute of Standards and Technology (NIST) site on statistics.[3] This formula was developed by the same person who gave us the correlation formula, according to this site.

$$skew = \frac{3 * (mean - median)}{standard\ deviation}$$

Although the formula uses the mean and median, as it should, the skew also relates to the mode since it is the mode that determines the skewing of the data. Skewing can be seen in Figures 3.2 and 3.3.

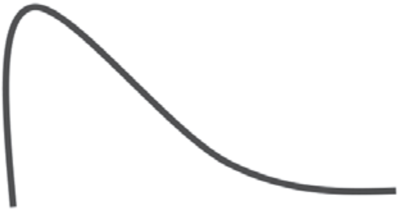


FIGURE 3.2 Right skew

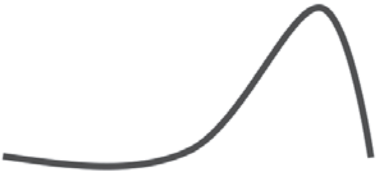


FIGURE 3.3 Left skew

In these cases, there is either a right or left skew that exists in the data. The intuitive answer to which of these is the right skew would be Figure 3.3, but this would be wrong! Right or left skewing is based on the amount of data that exists on each side; the more data there is, the greater the skew. The

[3] *https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm*

reality is that Figure 3.3 shows *left*-skewed data. It seems counter intuitive until the data is tracked, and it is shown that most of the data appears to the left of the mode (the hump). Collect a left-skewed dataset and prove it through making a graph. It always turns out the same.

3.4.1 Kurtosis

Kurtosis is not covered in detail in this book. *Kurtosis* shows whether the data has more values in the "tails" than in the center, but the formula and definition is better explained by NIST at their statistics site that was cited earlier.[4]

The NIST has an entire handbook that explains a number of concepts, and gives some formulae that matches the ones presented here. Kurtosis is something that should be considered when performing a descriptive statistical study of the values in the dataset because it allows the analyst a "view" of the dataset without having to create a graph or other illustration. However, the calculation for the Kurtosis may take more time than is really necessary to see that the data has a certain shape. If the analyst's software tool is one that will automatically perform Kurtosis, then this is a worthwhile endeavor. Otherwise, it may be best to leave the Kurtosis for another day.

3.5 Measures of Variation

Where there is a center, there are two edges. In the Air Force, there were a series of questions that were part of a "nonsense" test that went something like this:

Question: "How long is a piece of string?"

Answer: "Twice the distance from the middle to each end."

Variation is something like this piece of string. In order to understand the variation in a dataset, the analyst must understand the center, which was explained with different methods in an earlier section. What variation in a dataset means is how much of a difference is between the center and each end of the data. If the difference is large, the events within the dataset have a lot of variation. If the difference is small, then the dataset does not have a lot of variation. Is variation within a dataset important? The answer to that, unfortunately, is that it depends on what the dataset is trying to measure or depict. The more varied the dataset, the more the dataset could

[4] *https://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm*

accurately depict a situation or circumstance. For instance, if the set of weights among toddlers is not varied, that may mean that the toddlers are being raised in similar environments (at least in the food category), or that there is not enough sampling to make a varied dataset. In another instance, varied datasets could mean that there is no real pattern, or a way of measuring or predicting a future event. This could put a damper on the data analysis from a long-term perspective and, by virtue of that, make the data useless to predict future events.

In any event, the measure of variation within a (or between) dataset(s) is something that is necessary to gain an accurate perspective of that dataset. This section addresses those methods and explains how the variation relates to the measures of the central tendency. There are plenty of introductory statistics books that cover this topic, and this one describes some additional variation methods that are not covered in a conventional statistics text.

3.5.1 Variance

The first measure of variation that is necessary is variance. The variance of a dataset is the overall distance between the object and the mean. This distance is squared and then added together and divided by the number of objects. This determines the width of the data in a very general sense. Besides determining the width, it may reveal the result of the outliers in a dataset. The variance is necessary to determine the standard deviation, which is the square root of the variance and is covered in a future section.

$$\frac{\Sigma\,(x_i - \boldsymbol{\mu})^2}{N}$$

The variance formula looks complex, but an analyst can understand the symbols and use this formula in any software program. Simply put, the formula means that we sum the squares of the differences between each object's value in the dataset and the mean; then, we take that sum and divide it by the number of events or objects in the dataset. Sometimes, this variance formula is known by its shortened name, the sum of the squares. In this case, the name does fit the formula (at least a good bit of it anyways). If the dataset contains a series of values on either side of the mean, then the method for finding the variance consists of subtracting the mean from each of those values, squaring those values, and then adding up those values.

The last step is dividing that added value from the number of values in the dataset. For this illustration, we are going to use the following information:

Mean: 5 (The real mean is 4.9, so 5 is close enough)

Data values: 3, 4, 6, 7, 8, 1, 3, 4, 6, 7

Step 1: Take each of the values minus the mean.

3 – 5 = –2, square 4

4 – 5 = –1, square 1

6 – 5 = 1, square 1

7 – 5 = 2, square 4

8 – 5 = 3, square 9

1 – 5 = –4, square 16

3 – 5 = –2, square 4

4 – 5 = –1, square 1

6 – 5 = 1, square 1

7 – 5 = 2, square 4

The next step is to square those results and add those squares, which is 45. That result is divided by the numbers of values, which is 10, so 45/10 = 4.5 for the variance. Hence, the sum of the squares has great value in describing this method. This is again a valuable way of remembering this method.

What is the reason for squaring the differences between each event and the mean? Squaring the result of a number, whether the number is negative or positive, results in a positive answer. A variance is of little use if the result is negative or 0, which could happen if the difference between the object and the mean is not squared.

Why take the variance at all? One reason is that the variance is particularly useful when comparing two or more datasets. By comparing the variance, the analyst can determine if the datasets are comparable. Some of the tests to do this include the Levene's F-Test,[5] which is not covered in detail in this text. Besides that, there is another reason for taking the variance and that is to get to the standard deviation, which is the conventional method for measuring the variance in a dataset. This standard deviation is

[5] *https://www.itl.nist.gov/div898/handbook/eda/section3/eda35a.htm)*

something that is accepted and applied to just about every analytical effort, and that is because it is much smaller than the variance. The standard deviation is covered in the next section.

3.5.2 Standard Deviation

Standard deviation is one of those statistical terms that everyone seems to recognize, but cannot able to explain its importance. It is a formula that relies on the mean as the main source of values, which, as pointed out before, can be skewed by outliers. The standard deviation is a direct derivation of the variance, specifically the square root of the variance. The word "standard" in standard deviation has been known to create some confusion for statistics students who try to find the origination of that name. As an instructor, one way of looking at this is to state that the standard deviation is a major component of the standard normal curve, which also employs the standard deviation to convert the dataset's standard deviation to what is called a "Z-score." By connecting the standard normal curve to the standard deviation, we see that one cannot exist without the other.

$$\sqrt{\frac{\Sigma\,(x_i - \boldsymbol{\mu})^2}{N}}$$

If the standard deviation formula looks familiar, just copy the formula to another sheet of paper and cover the square root sign; the resulting formula is the variance formula. The standard deviation is the square root of the variance. All the analyst has to do is to determine the variance and take the square root of that value. There is something that must be mentioned at this time, although it has been discussed before: the difference between the population and sample standard deviation. The population and sample variances differ by placing n-1 instead of N in the denominator. What this does is prevent sample bias in the resulting standard deviation. In statistical terms, this is called the "degrees of freedom," and it is used in other tests, like the students t-test, to analyze a dataset that may not be a normal distribution. An illustration of this can be found on the Internet.[6] The concept of the degrees of freedom is discussed in detail in a later section.

3.5.2.1 Real-World Use of the Standard Deviation

Project managers have a method for assessing the standard deviation by taking the pessimistic time minus the optimistic time for the project task

[6] *http://vortex.ihrc.fiu.edu/MET4570/members/Lectures/Lect05/m10divideby_nminus1.pdf*

and dividing it by 6. The pessimistic time is the most time that the task can take to be completed and the optimistic time is the least amount of time necessary for the task to be completed [PMI-2017]. This is a very rough estimate, but it can help with some quick calculations, specifically if the project manager needs them for the team meeting. The systems engineer and cybersecurity analyst can also use this formula for a quick estimation and further normal curve developments (also, for the systems engineer to check the process efficiency and for the cybersecurity analyst to check the availability statistics).

3.6 Standard Normal Curve vs. Normal Curve

The standard normal curve is illustrated at the web site shown as a footnote on this page. This curve always has a mean of 0 and a standard deviation of 1. The different segmentation of the curve shows how much of the percentage of area under that curve is in each segment. This is important because the area of the standard normal curve is 1. This results in the area of the curve being converted to a probability, which can only have a 0 to 1 value. The curve therefore represents the probability of something happening, which also varies between 0 (impossible) and 1 (not just possible, but will definitely occur). Instead of saying that 34.13% of the segment exists between the center of the curve and first line to the right of that point, the real use for this is if a value exists within that segment, it has a .3413 chance of being between the center of the curve and the line to the right of that curve. This is an essential function of the standard normal curve when analyzing datasets related to a proposed mean or comparing two or more datasets. The formula for the standard normal curve is shown, but there is no need to memorize this one (most computers do this for you). If you would like to see how this is graphed (to confirm that it is, in fact, the standard normal curve), then plug this formula into any graphing calculator or computer software and the standard normal curve will appear. However, with data analysts having more than their share of computer applications, why bother with this formula?

$$f(x) = \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}}$$

If this formula gives the standard normal curve, what does a normal curve look like? The normal curve is similar in most features of the standard normal curve, with one major exception. The normal curve's mean is the mean generated by the dataset, and the standard deviation is the standard deviation of the dataset. So, instead of having a mean of 0 and a standard deviation of 1, the mean and standard deviation are those of the dataset. For example, if the mean of a dataset is 250, and the standard deviation is 10, then the curve will have a specific configuration, except that instead of 0 in the center, it will have 250 and each line to the right will increase by 10 (so the first line will be 260, the second line will be 270), and the lines to the left will decrease by 10 (so the first line to the left will be 240, the second line 230).

What does all this mean? If you have a dataset with the mean and standard deviation as above and want to know what the probability is of a value of 275, you can place this value in relation to the different segments to see the general probability. In this case, the value 275 lies to the right of the 270 value, which is the end of the second segment and the beginning of the third. What this means is that approximately 98% of the values are below the value 275. This would result in realizing that 275 is a "good" result or "bad" result based on whether high scores mean something good or bad.

How did the 98% appear? You can add all the different segments from the normal curve and add all the segment values from the second line to the right of the center line to the left side of the curve to obtain the following:

$13.59 + 34.13 + 34.13 + 13.59 + 2.14 + .13 = 97.71$

If you have trouble following this logic, you can refer to a math site[7] for more information.

How does an analyst compare two different datasets that have different means and standard deviations, and therefore different normal curves? As stated above, the standard normal curve has a set mean and standard deviation, but there is a way to compare two different datasets using the Z-score formula, which takes any point (called a *random variable*) and subtracts the mean from this value and divides that result by the standard deviation. In other words, if you have two different test results and want to compare one student getting a 70 with another getting an 80, but both in different classes with different means and standard deviations, you would simply take each score and subtract the respective means from that score and then divide it by the respective standard deviations. The resulting scores can be compared on a standard normal curve to determine if one score is really

[7] *https://mathbitsnotebook.com/Algebra2/Statistics/STstandardNormalDistribution.html*

lower than the other when taking into consideration the separate tests and separate measures.

3.7 Other Measures of Variation

The next calculations are not normally taught at the undergraduate level in statistics. However, they are an important part of the analyst's tool box since they provide an alternative way of seeing the very essence of data description. Here, the formula is presented first, then the technique is described, and finally, the technique is applied to an example.

When working through these different methods, you should begin to consider how these methods might help your overall analysis and how the application of these methods could help verify results. In this way, the following material is much more than just introducing new techniques: it is providing the foundation for future use and interpretation.

3.7.1 Mean Absolute Deviation

The formula for the mean absolute deviation may appear somewhat confusing, but there is a very simple explanation for it. Instead of the "sum of squares" that is part of the variance and standard deviation, this formula uses absolute values to prevent negative numbers from appearing in the overall summation. In other words, in the variance formula, there might be a value 4 and the mean 5, which indicates that the difference of those two numbers is -1 (which is squared to give the final value of 1, and later summed with the other mean differences). In this case, the squaring is not done, and so the final number is given in terms of the absolute value. That means that the -1 in the previous solution is changed to 1, giving a positive result. There is no need to sum the squares, but a need to sum the absolute values. The example that follows shows how to apply the formula to actual numbers.

$$\frac{\sum_{i=1}^{n} |x_i - \bar{x}_i|}{n}$$

All the values and mean are included in the problem, so you simply take each value and subtract the mean from each of those values. Doing so results in the following values:

25000

50000

60000

–20000

Changing all the values to positive numbers using the absolute value results in the following list:

25000

50000

60000

20000

You then sum the numbers and divide that result by 4, which is the number of values in the dataset. This is the final calculation:

155000/4 = 38750

This is the mean absolute deviation. You could use this number is place of the standard deviation when applying the normal curve or finding the Z-score (discussed later). There is one caution. The number of the mean absolute deviation is based on the mean, which is affected by outliers. No matter the derivation of the value, the mean is always affected by those outliers. To mitigate these outlier affects, use the median and the next measure of the central tendency takes the median to measure the variance. This is done through the median absolute deviation.

3.7.2 Median Absolute Deviation

In the previous section on the median, it was revealed that the median is never impacted by outliers as the mean is impacted. This is an asset to obtaining an unbiased data center and is also useful in obtaining an unbiased data variation. The *median absolute deviation* is rarely used by data analysts, but it is available in such data science tools as KNIME (*www.knime.com*) and R (*www.cran.org*). It has also been used in the Geospatial Information Systems (GIS) in some agencies of the federal government. This measure is not affected by outliers as is the mean absolute deviation.

The formula for calculating the median absolute deviation is as follows:

$$Median\ (|x_1 - m|, |x_2 - m|, \dots |x_n - m|)$$

In this example, the first step is to take the data and place it in numerical order. That order is as follows:

1, 3, 5, 8, 10, 15, 16, 18, 20

That would mean that the "physical center" is 10, which would mean that the median (or 50th percentile) of the data is 10. This is a simple example for demonstration, but remember that if the data had an additional number (i.e., 25), then the order is as follows:

1, 3, 5, 8, 10, 15, 16, 18, 20, 25

This would place the physical center at both 10 and 15 (the 10 has 4 values to the left; the 15 has 4 values to the right). This requires the analyst to take the average of the two center values, which is 12.5 (10+15/2). So, now the 50th percentile or median is 12.5.

The median is established, so what is the median absolute deviation? This encompassed the following steps:

1. Take the absolute value of the difference between each value and the median. An example of this is as follows:

   1-10, 3-10, 5-10, 8-10, 10-10, 15-10, 16-10, 18-10, 20-10

   9, 7, 5, 2, 0, 5, 6, 8, 10

2. Now that the calculation is complete, the list of values attained by Step 1 should be placed in numerical order as shown below:

   0, 2, 5, 5, 6, 7, 8, 9, 10

3. Once the list is ordered, you obtain the median of that ordered list. In this case, that is 6.

4. The way to use this result in the normal curve is the same way as in the standard deviation. The result is placed with the mean at one point of the normal curve, then the mean + one median absolute deviation, then the mean + two median absolute deviations, and so on.

5. Now, you find the mean of the original list of values (the ones in Step 1), which is 96/9 or 10.67. Then the first median absolute deviation is 16.67 (that is, the mean of 10.67 plus the median absolute deviation of 6), the second median absolute deviation is 22.67, and so on.

What about the median absolute deviation? How would the analyst find that? In this case, because this is a learning text, you would go back to the section on the median absolute deviation and follow the steps in that calculation (hint: take the sum of the absolute values of each data object or event minus the mean instead of the sum of the squares).

3.7.3 Still More Tests for Variation

There are other tests for variation besides the ones that have been discussed. These are given to afford the analyst more tools for the data tool chest, but these are not unique. However, they are sometimes placed at the rear of analytical priorities when they should be at the forefront of analysis.

3.7.3.1 Range

Range is one of those measures of variation. The *range* of the data is the maximum value minus the minimum value. Although simple in execution, range is one of the most misunderstood measures because it is that simple. Basically, the range determines which values are included in the dataset. If there is another value that enters that dataset that lies outside that range, then that value is considered an outlier. The range is there to measure what is in the dataset as well as what can be placed there. It is a measure of pure simplicity that says volumes about a dataset without using a complicated formula. For instance, if the range of the data is 50, then that means that the maximum minus the minimum is 50 and that there are 50 "steps" in values between the minimum and the maximum. If the range is 1000, then that means that the variation in that dataset is probably more than the variation in the dataset with 50 as the range. The larger the range, the more probable that dataset contains values that span more space than the one with the smaller range. As an analyst, any measure, no matter how simple, is worth considering when it comes to interpretation in that dataset.

Besides its importance, the range is also a stepping stone to further analysis using this type of variation. A specific type of analytical tool using range is the Inter-Quartile Range (IQR) which is on the docket next.

3.7.3.2 Inter-Quartile Range (IQR)

Before addressing the IQR, it is important to just discuss the range of the data. Although calculating the range is probably one of the simplest operations, the real value of the range comes into play with other tools like the IQR. Calculating the range is as simple as taking the highest value minus the lowest value. Comparing the range of two datasets brings more clarity to the differences (or the similarities) between the datasets.

The IQR is something that is (again) normally glossed over in the study of undergraduate statistics, and is usually covered as part of the percentile portion of the course. We do much the same, but go into a little more detail of how this method can be useful in determining outliers from a formula standpoint.

However, in order to understand quartiles, we must first understand percentiles. The best way to describe percentiles is to comprehend what a percentile incorporates. A quartile is ¼ of a percentile, which means that the first quartile (Q1) is the 25th percentile, the second quartile (also called the median) is the 50th percentile, and the third quartile (Q3) is the 75th percentile. Students have argued that this is just three quartiles, not four, but the quartiles split the data into four equal parts. The first part is from the minimum to Q1; the second part is from Q1 to the median; the third part is from median to Q3; and finally, the fourth part is from Q3 to the maximum. Although this explanation seems elementary, it will help students understand the term quartile and the application of the term. The next section explains the formula behind the percentile, but for this section on IQR, the next logical step is the formula for determining that range.

The IQR formula is relatively straightforward, with the answer being Q3 minus Q1. What this does is give you a measure of the variation that can be used to determine outliers in the dataset. The formula below the IQR formula is split into two parts. The first part determines the outliers on the right side of the dataset and the second part determines the outliers on the left side of the dataset.

Although this seems relatively straightforward, it needs to pass the “reasonable” test: “Do the results make sense given what the dataset contains?” If the result passes the reasonable test, the analyst can continue with the process. However, if the results do not pass the reasonable test, then the analyst needs to re-evaluate the results.

For instance, if the data has a Q3 of 75 and a Q1 of 25, then the IQR would be 75-25 or 50. Once the IQR is obtained, the outliers are determined by taking Q3 plus (1.5 times IQR). In this case, it would be 75 + (1.5 * 50) or 75 + 75 or 150. If there is a value that is greater than 150 in the dataset, then that value is considered an outlier. To find the left side of the outlier group, the we perform the following:

$$Q1 - (1.5 * IQR) = 25 - (1.5 * 50) = -25$$

If the values go no lower than 0, then we know there are no outliers to the left side of the data. This is called the “IQR Rule” and is illustrated nicely at Kahn Academy site (*www.khanacademy.org/math/statistics-probability/summarizing-quantitative-data/box-whisker-plots/a/identifying-outliers-iqr-rule.*)

As an example, imagine a dataset with a Q3 of 43 and a Q1 of 22. What are those values that the analyst should be alert to notice that could indicate outliers? In this case, the IQR is Q3 – Q1, which is 43-22 or 21. Taking 1.5 times 21 would give the result of 31.5. To get the upper outlier, take 43 + 31.5 to get 74.5 and to get the lower outlier take 21 – 31.5 which would get -10.5. The next thing the analyst needs to consider is the type of data in the dataset and the reason for the analysis. If the dataset is the temperature of the glaciers at certain times of the year, then the lower outlier at -10.5 would be pertinent since temperature can be negative. However, if the dataset is the age of individuals when they start to drive, then a negative outlier would be impossible, which means that there are no lower outliers in this dataset. The dataset range is 0 to 74.5, which means any number lower than 0 or higher than 74.5 would be considered outliers.

What this does is give the analyst the ability to spot outliers without any graphical representation of the data. This is very important, since the more than analyst can determine without extensive analytical time, the more time is saved for other analytical tests.

3.7.3.3 Percentile

Now that the quartile and percentile concepts have already been broached, it is time to get into more detail on the percentile aspect of statistics. The percentile is something that we see in commonly, but sometimes do not understand the real interpretation of the result. For instance, when students take the Scholastic Aptitude Test (SAT), their scores are compared with a population of scores from a number of years and then the percentile is calculated. If the student receives a percentile of 90, this means that 90% of the people that took the test scored BELOW the student in question. This is an important difference between percentile and percentage. Percentage is the probability that a particular event might happen, such as saying that there is an 80% chance of rain. The percentile is a measure of success for a particular event, and the ranking of that score (or object) against all other objects. It is interesting to note that if someone ranks 50 as a percentile, that means that they are average. In other words, the 50th percentile is the center of the data, which in most cases is considered the mean, or average. If the analyst wants to picture the 50th percentile, picture the "hump" of the normal curve. The center of this hump (or mode) is the 50th percentile. If the analyst wanted to know if their child was of an average height or weight, they would need to take the weight and height of their child (along with the child's age and gender), and then look up the resulting population of children's height and weight (which can be downloaded at *www.cdc.gov/growthcharts/clinical_charts.htm*). As you can see, anything that is needed from a wide point of view can often be obtained through government websites or data ports.

The formula for obtaining the percentile of a value in a dataset is much like obtaining the median in that you are obtaining the position of the value, not the value itself. The formula for percentile is

Percentile = Percentile (in decimal) * N

N = population number of events

If you want to know what the 20th percentile in a dataset containing 50 values, then the position of the 20th percentile would be .20 * 50, which is the value between the 10th and 11th values since the percentile is a round number. If the percentile location ends up to not be a round number, you can just round up. This works for any percentile that you want to find, even if the percentile is strange, like the 76th percentile or something similar.

For example, say you are given the following values:

2, 4, 6, 8, 10, 12, 14, 15, 16, 25

You want to know what the 25th percentile is for this value set. The first step is to take .25 (the decimal equivalent of 25th percentile) and multiply it by the number of values, which in this case is 10. This result is 2.5, which you round up to 3. You then count the third value of the dataset (after sorting the data in ascending order). The third value is 6, so that is the 25th percentile of the dataset. What if you want to know the 80th percentile of the dataset? The first step is to multiply .80 by the number of values (10) to get 8. With a whole number, you must average the result and the number above it to get the LOCATION of the 80th percentile. The average of the 8th and 9th values is (15 + 16)/2, which is 31/2, resulting in 15.5. This means that the value of 15.5 is the 80th percentile of the dataset.

3.7.4 Five Number Summary

With all the preceding information on the median and quartiles, it is easy to overlook the method of taking some descriptive statistics to create a visual perspective of the data. This method, using the median, quartiles, and the minimum and maximum values, is called the *Five Number Summary*. Basically, the five numbers in this summary are the minimum value of the dataset, the maximum value of the dataset, the 25th, the 50th, and the 75th percentiles. Together, and properly plotted, these five numbers encompass a data visualization called a *box plot*.

The first thing to do is to determine the Five Number Summary, and then visualize these numbers. For example, taking the previous example, we can determine that the Five Number Summary is as follows:

2, 4, 6, 8, 10, 12, 14, 15, 16, 25

The first thing is to determine the minimum and maximum values, which is easy enough if we sorted the dataset in ascending order (as it is here). The minimum value is 2 and the maximum value is 25. We next determine the 25th, 50th, and 75th percentiles, which was explained in the previous sections. The 25th percentile is 6 (from the last example). The 50th percentile is also the median, which is .50 * 10 = 5, so the median (and the 50th percentile) is the average of the 5th and 6th values, which are 10 and 12. The average of 10 and 12 is (10 + 12)/2 = 22/2 = 11, so that is the median. The 75th percentile is .75 * 10, which is 7.5, and we round up to 8, which makes the value of 15 the 75th percentile. Voila! The Five Number Summary is

2, 6, 11, 15, 25

These figures can be made into a box plot using a series of steps that can be found on the Internet.[8] Of course, some tools, like Excel and KNIME, give the analyst a ready-made box plot function, but it is always nice to go back to the manual method just to review some basic concepts.

[8] *https://www.wikihow.com/Make-a-Box-and-Whisker-Plot*

CHAPTER 4

PROBABILITY PRIMER

Probability can be daunting, but understanding it is essential to understanding data analytics and statistics. In fact, not understanding how probability affects something can be a detriment for a data scientist or analyst, especially since people seem to read the end of a study and then share those findings, which are almost certainly based on probability. Therefore, understanding probability is something that is critical, since without that, data analytics is not informative.

Many students of statistics have had little exposure to probability because it can become a complex area of study. In some cases, the negative influence of probability can be attributed to the complex examples that accompany such a curriculum. Whether it is the classic *false positive of the flu* example or the *sock drawer* example, students have become accustomed to probability being both hard to learn and hard to do.

That is why this section is as straightforward as possible, beginning with the very basic probability formula. This formula is a very simple and fundamental to the understanding of probability. What the formula portrays is the frequency of an occurrence divided by the total number of outcomes. For instance, the probability of throwing a 3 from a balanced die is 1 (the frequency of outcomes, since only one side is read) divided by 6 (the total outcomes, since there are 6 sides of the die). Because there is only one 3 on a die, and therefore one side of the die, the probability of throwing a 3 is 1/6. The formula takes on a different perspective when addressed in this fashion.

$$\frac{f}{N}$$

4.1 Addition Method in Probability

Although one would think that probability just stops with the basic formula, there is more. In fact, probability plays a major role in the analysis of several variables that affect our lives. For instance, does getting a flu shot provide some protection against COVID-19? What about a flu shot and wearing a mask? These are the types of analytical questions that apply probability.

To properly introduce the different types of problems that have to do with probability, the first example should be relatively straightforward. If we wanted to know the probability of a person in the US of dying from heart disease (as defined by the Centers for Disease Control or CDC), then we can go to a site that lists the types and numbers of deaths (*www.cdc.gov/nchs/fastats/deaths.htm*) and see that the number of deaths from heart disease is approximately 647,000. With the total population of the United States being approximately 320 million, the probability of dying from heart disease is 647,000/320000000 or .002. If we want to know what the probability is that a person died from cancer OR heart disease, then we apply a formula, with P being the probability of something happening (in this case either cancer or heart disease). We then take these probabilities and add them together.

$$P(cancer\ OR\ heart\ disease) = P(cancer) + P(heart\ disease)$$

The equation looks like this:

$$P(cancer\ OR\ heart\ disease) = \frac{599000}{320000000} + \frac{599000}{320000000} = .0018 + .0020 = .0038$$

What this means is that a person has a .0038 probability of dying from either cancer or heart disease in that specific year. If we want to convert this to a percentage, we could say that a person would have a little less than a .4% chance of dying from either of the aforementioned maladies in that year. This is not even one-half of one percent, a very small chance indeed. Of course, this is a fundamental probability and does not take into consideration any other factors, such as the location within the United States, age, gender, or occupation. Hence, this is a very general guess of this probability, but this example is solely for academic purposes.

4.2 Multiplication Property of Probability

In discussions on probability, "or" means addition, so probability problems with "and" in the problem involve multiplication. The confusion that this causes with students is understandable, since in their math experience they have learned that "and" means addition, not multiplication. Now, they have to learn that the "and" means multiplication. This will take some doing, but with some examples, specifically using language as the learning key, the students will come around.

The formula for this probability property is

$$P(A)\ AND\ P(B) = P(A) * P(B)$$

In plain language, this means that the probability of something (A) happening AND the probability of something else (B) happening is equal to the multiplication of the probability of A and B.

Moving to another state can be harrowing. If the chance that a tornado will occur in that state is about 20% and the chance of hail is 30%, what is the probability that both will occur at the same time, understanding that hail does occur during a tornado?

The expression would be .20 * .30, or the probability of a tornado AND the probability of hail appearing at the same time. One note on this: hail is a common occurrence with a tornado, so these percentages were contrived for illustration only. If you want to know the real probabilities, please see the National Weather Service (*www.noaa.nws.gov*) data on tornados.

The probability of a tornado AND hail is approximately .06, or given another way, there is approximately a 6% chance of seeing a tornado and hail in this state in that specific year. Another note at this juncture is that this (again) does not take into consideration the time of year, the number of days of rain, and other factors that may change this percentage. This is just an educated guess.

Probability is to statistics as air is to breathing. They are inextricably linked through inferential statistics. To give a real example of this type of probability application, and using the tornado database at the NWS, what are the chances of having a tornado in 1950 that caused more than $250,000 in damage? Using the correct dataset for the year 1950 (*www.ncdc.noaa.gov/stormevents/ftp.jsp*), the number is approximately 67 out of 224 total tornadoes. This would mean that there was approximately a 30% chance of a tornado causing more than $250,000 in damage in 1950. Remember that the year is very specific, which focuses the data "question" so that the analysis is centered on the answer to the question rather than seemingly endless evaluation. But now you can see the application of probability to statistics and therefore to data analytics.

4.3 Bayesian Probability

The content of this text would be incomplete without a section on Bayesian probability. The sections on probability that preceded this section focused on what is called *frequentist probability*, where there are certain assumptions already built-into the probability, such as the results of flipping a coin resulting in a 50/50 chance of getting a head or tail. However, Thomas Bayes, the originator of the probability named after him, felt that there were other factors that were contained within the basic probability and these needed to be included in probability discussions and formulations. As a result, the Bayesian probability was born. There are plenty of books that address Bayesian probability, but one that is both entertaining and educational is Will Kurt's *Bayesian Statistics the Fun Way* [Kurt-2019]. Because of the complexity of the concept, and the fact that there are books written on the subject, this text focuses on only one concept called the *Bayes factor* (BF), which looks complicated, but can be explained in a straightforward manner. First, let's discuss the conditional probability.

Conditional probability is the type of probability that exists in real life. Rarely do we just throw a die once or pick one card. The conditional probability takes into consideration the chance of picking one card should another card already be picked. The language for conditional probability is the probability of event A *given* event B already happened. For example, what is the probability that someone picks a red card given that a jack has already been picked? The formula for such a question looks like the formula below.

$$P(A|B)$$

Trying to make sense of this is somewhat confusing, but there is a derivation of this formula that encompasses some of the concepts that have already been covered. The formula above can be distilled to the following formula.

$$\frac{P(A\&B)}{P(B)}$$

Taking the example about the red card and the jack, the way to address this problem with the formula is to state the probability of drawing a red card AND a jack (probability of A & B), which is 2/52 (jack of hearts and jack of diamonds), which is about .04 (.038 to be exact). The denominator of the formula is the probability of picking a Jack, which is 4/52 or .076. Dividing .038 by .076 gives a result of .50, which means that the person has a 50% of picking a red card given that a jack has already been chosen.

However, if the analyst used the formula above, then they used a Bayesian probability. This is just the first part of the Bayesian probability, but also conforms to the conditional probability concept. The next formula helps determine whether the data can help change a belief or change someone's mind with data.

The Bayesian probability can be a very complex concept, but for the purposes of this text, we discuss only one area of Bayesian probability called the Bayes factor. This is a concept discussed at length in many texts, but it is accompanied by a formula that may look complicated but is straightforward to explain. The formula is as follows:

$$\frac{P(D_n|H)}{P(D_n|\overline{H})}$$

What the formula compares is the data given one hypothesis (that the entity can answer all questions correctly) over the data given another hypothesis (that the entity cannot answer all the questions correctly). For example, what if there was a machine that can supposedly perform feats of prognostication that seems to show it is clairvoyant? However, given the formula above, there can actually be numbers to support this belief. For instance, if the belief is that the machine does get the future guesses correct, then the numerator will be 1 (and will always be 1 if the belief is that the machine, person, or other entity gets it right every time). If the person believes that at worst case basis (denominator), the machine has an equal chance of picking a right or wrong answer, then the denominator is .5n. The "n" denotes how many questions the machine gets right at any particular time. What the formula will look like, based on the above assumptions, is as follows [Kurt-2019, p. 169]:

$$\frac{1}{.5^n}$$

This means that if the machine gets three questions correct, we substitute 3 for n and solve. This would be something that a project manager could use for task completion or other parts of a project. For instance, if the team that is doing module A has a past probability of 50% completion on tasks, while you as project manager seem to think that they can complete all their tasks on time, the formula would be something like this one.

Using this formula, the project manager fills in the "n" with the number of tasks that are part of the team's schedule and determines the probability of getting those tasks completed on time. For instance, if there are 10 tasks, and the team has completed all 10 tasks on time, then it is $1/.5^{10}$, which is over 1,024. What does this mean?

Anything over 150 means that there is a great deal of confidence that the team can get their tasks completed on time (or that there is great confidence in the hypothesis *stating* that the team can get all the tasks completed) even at a 50% probability [Kurt-2019, p. 169]. This could mean that the team has done something to get their tasks completed more efficiently or there is some other area they have improved in.

The last part of this concept is adding in the previous belief or perception about the problem. For example, the analyst asks someone who knows the team what the odds are that they will get the tasks completed on time, and the person says about 10% or 1 in 10. At this point, those odds are multiplied by the BF that we determined earlier to be 1,024. This results in a new BF of 102, which would lower the confidence of the BF, but still make it somewhat confident [Kurt-2019, p. 169]. In essence, this is a very useful formula that can be used to determine the perceptions before and after data is collected on a specific project phase or gate (to use the systems engineering phrase).

CHAPTER 5

OCCAM'S RAZOR AND DATA ANALYTICS

To properly introduce the next section, we first introduce the concept of Occam's Razor, attributed to William of Occam (1287-1347), a British philosopher who is best known for his attributed (albeit according to one source, not found in his writings) adage that translates to be that the simplest solution is the best [Spade-2019]. This is the foundation upon which data analytics is constructed. (Ironically, William of Occam also felt that situations should not be overly repetitive). Many data analysts want to immediately do a simple linear regression or a correlation on the data, as if these methods replace the ability to look at data summaries and arrive at a description of that data. The reason for starting with some basic statistics is not just as a refresher, but serves as a reminder that when provided data, the true analyst looks to the simplest form of analysis first and then moves into more complex methods. There are many times that analysts perform a number of evaluations on data, only to discover that they are far beyond the requirements that were first proffered by the product owner. In project management (and systems engineering), this is called *scope creep*, and it is alive and well within the data analytics community. For example, according to the *INCOSE Systems Engineering Handbook*, there are a number of attributes of good requirements, including whether the evaluation is necessary, unambiguous, complete, singular, achievable, and verifiable (just to mention a few) [Walden-2015, p. 61]. In project management, the concept of requirements is part of the scope planning [PMI-2017]. There are many references to SMART requirements, with SMART being an acronym that means Specific, Measurable, Attainable, Realistic, and Time-bound (*pmhut.com/smart-requirements-introduction*). Cybersecurity has very similar requirements, but these are based on the initialism CIA: confidentiality, integrity, and availability (*re-magazine.ireb.org/articles/cyber-security-requirements-engineering*). Because of the relationship between each of these, it is vital that the data analyst, which could be the project manager, systems engineer, or cybersecurity analyst, needs to understand how to simplify the requirements without diluting them to the point of uselessness. This is where Occam's Razor seems to work well with all of these different professions.

For example, if you have a teenager who discusses wanting a Ferrari, the parent will naturally revert to the simple requirement. That is, *wanting* a Ferrari translates into *needing* a car. In the same way, the data analyst must adjust the requirements for a data analytical project to conform to the needs, not the wants.

What happens should the data analyst succumb to scope creep and include useless, redundant, and maybe even erroneous data analytic evaluations? One of the results of the analyst not properly considering the requirements is *overfitting*. One source describes overfitting as the inability of the core or training data to fit other generalized and independent data.[9] It is the willingness of the analyst to try all the variables at once to get a correlation with one or more of those variables. In essence, it would be to take age, gender, location, ethnic background, occupation, and smoking to try to find one or more of those variables' relationship with lung cancer. Of course, there would be at least one of those variables that would correlate with lung cancer, but what was the requirement in this instance – find *any* variable that relates to lung cancer? That would be too general and would be allow one instance of a correlation to be translated into a generalization. That is why hypothesis generation and testing are essential.

5.1 Data Origination

One of the main concerns when performing data analytics, and a major part of Occam's Razor, is the origin of the data that the analyst uses. Some analysts disregard the origin of the data as long as they can get the data for models and such, but origins play a big part in the overall analysis results.

Take, for instance, data concerning COVID-19. The data from Johns Hopkins University (*www.jhu.edu*) was compared with the data from the Centers for Disease Control (*www.cdc.gov*) for a single day. The origin, in this case, resulted in a wide disparity on that one day for one location to the

[9] *https://www.futurelearn.com/courses/data-mining-with-weka/3/steps/290129*

tune of over 20,000 cases! Now, this could have been the way that the data was collected, and a quick look at the data sources showed that both entities did not count the cases in the same way. Just a quick look at the sources can prevent an analyst from stating that one of the datasets is wrong when it may just be different. It was just good fortune that both datasets had the sources plainly described at the respective sites.

Another example is the dataset from the National Weather Service (*www.noaa.nws.gov*), which describes the number of storms (including tornadoes) from the years 1950 until 2018. This data was combined into one large dataset to show the differences in tornado activity during those years. The origin of the data was the same, seeing as it was from the NWS, but the analyst must realize that data collected in 1950 was not collected the same way as it was in 2018. First, there were no weather satellites in 1950. In fact, there were no satellites in space above the Earth in 1950 at all, since Sputnik was the first artificial satellite and that was not launched by the Soviet Union until 1957. Sputnik was not a weather satellite. Today, there are more satellites in space than one can seemingly count or track, with weather satellite data being fed directly to our cell phones. Second, the type of recording was different in 1950 than it was in 2018. The 1950 dataset contains only tornado tracking; the 2018 dataset contains a variety of storms from hail, thunderstorms, and, of course, tornadoes. Lastly, the amount of damages in 1950 cannot be compared with the damages in 2018 because of a few factors, including the population density, cost of money, and the reporting of those damages.

Therefore, to compare the number of tornadoes in 1950 (or even 1970) and today would demand some research to see which factors affect the numbers. In other words, the origin of the data (which includes the time and location) affects the results of any analysis. This is something that analysts must consider before conducting any explicit analysis, and they should qualify such findings in their analysis reports.

How would an analyst avoid this pitfall? There are several steps that can prevent data origin bias. First, do the research necessary to make the data collection as inclusive as possible. This is important when it comes to analyzing data that changes radically because of location and time. This is critical if performing a longitudinal study (based on time) because the season can affect the results. Second, look at similar studies to get a baseline of the types of challenges that might accompany such a project. In project management and systems engineering, this is called an *analogous estimate*, where similar projects are considered when tackling a new project of similar structure. Finally, ask questions. This sounds simple, but critical question asking is really an art, honed through experience and knowledge. The best way to ask a critical question, without insulting the person wanting the information, is to repeat what the requester wants. In most cases, hearing the actual requirement, the requester will see that they have to phrase their request a little differently. When she was very young, my daughter was traveling with her mother and her mother's friend on a train from Baltimore to New York City. She found out that there were railways throughout the United States and wondered out loud how great it would be to travel to Hawaii by train. It was not until she heard that statement aloud that she realized her faux pas. The requirement was far more complicated than it would have seemed in the beginning.

CHAPTER 6

DATA ANALYSIS TOOLS

This is a brief compilation of some data analysis tools (data science tools by another name) that are used by data analysts in the performance of their jobs. For those who would like a more detailed application of these tools for analytical methods, one source is *Data Science Tools,* also from Mercury Learning (*www.merclearning.com*). There are many books that cover the use of data analysis tools; this is just one that will assist in the implementation of specific analytical methods including t-tests, confidence intervals, multiple regression, correlation, and even Levene's Test and other analytical techniques.

6.1 Microsoft Excel

Excel seems to be the go-to data analytics tool. It is often a part of any computer system and comes with myriad learning possibilities, from full credit courses to a number of learning websites specifically geared towards the software application.

The advantage of Excel is the Analysis ToolPak, which comes as an add-on to Excel and contains a number of one-stop approaches for often-used statistical techniques. For instance, an analyst can determine an entire descriptive statistics array from just one command within the Analysis ToolPak. In fact, in order to do this manually, an analyst would need to do an average, median, mode, variance, standard deviation, count, and confidence interval all individually within the Excel spreadsheet. The time savings is exceptional.

In addition to the descriptive statistics, the Analysis ToolPak also has a number of functions for statistical models including regression, ANOVA, and t-tests. All of these are essential in some analysis circles, so it is nice to know that they are included in this add-on.

6.2 R Stats

It seems that R is a statistical tool that is becoming more prevalent in analysis. One advantage to this tool is that it is free and open source, and can be downloaded from the CRAN site (*www.r-project.org*). This tool is also covered in the *Data Science Tools* book mentioned earlier, so it is not covered in detail here. However, the functionality within this tool expands to even text mining, which is used when there is a need to track text data, such as when surveys contain textual answers. In addition to text mining, the R Stats tool is also attached to python-related applications, and has a functionality for geospatial functions. The software is command-line driven, but there is a function called RATTLE (also covered in the *Data Science Tools* book) that takes some of the programming effort and turns it into a one-click function. In addition to RATTLE, there is RStudio (*www.rstudio.com*), which makes the functionality of R even more enticing. If the analyst has not explored R before, now is the time to download and use this tool.

6.3 Open Office

Open Office is a free download from *www.openoffice.org*; if you have experience with Excel, then Open Office is a cinch to learn. However, Open Office does not have the Analysis ToolPak, so all commands have to be manually inputted. There are some slight differences between Open Office and Excel, but these are more syntax than functionality. For instance, to separate parameters in Excel formulas, the user uses a comma, whereas in the Open Office spreadsheet program, the user uses a semi-colon. Other than those small differences, Open Office provides a relatively robust tool for data analytics.

6.4 Minitab

Minitab is a very dynamic statistical software that provides a click-and-drag functionality that far exceeds many of the other applications aforementioned. There is some expense to Minitab, but there is a price for functionality. The real issue here is whether the analyst can apply these functions to the analysis of everyday datasets; the answer is a resounding "yes." There are many books on this software application, but the best way to get a good introduction is through the Minitab website (*www.Minitab.com*). Use this software in a variety of datasets, specifically in implementing models such as regression and ANOVA.

The setup is similar to other spreadsheet software, but the choices in the tool bars are focused on statistical test and modeling, unlike some of the other software.

6.5 Tableau, SPSS, QLIK, and others

There are so many other types of Statistical software that it would be impossible to mention them all in this book. The choice of software is as personal a choice as what to wear during the day. Most statistical software has many of the basic functions that are common among all of these applications, including descriptive statistics, correlation, t-tests, and regression. The main differences in these applications range from their ability to analyze a great volume of data to their display of said data. Rather than expound on any one of these, it is better to download a trial of these software applications (they all come with a time limited trial for the user) and see if they fit your needs. (Then would be a good time to consider the economics of such an investment.)

One addition to many of these applications is the ability to use an online version of these tools, which in many cases is free. This might be a perfect opportunity to test the software and determine its feasibility for the analysis of a specific dataset. Tableau is one of those software applications that comes with this option. QLIK software has the QlikView free download that is relatively useful, but does have some programming requirements.

The software programs for statistical purposes are both plentiful and highly functional. The real trick is choosing the application that helps with specific data and can translate the results into a readable, and reliable, display.

6.6 Geospatial Statistical Systems

This section was originally relegated to a special section about geospatial statistics, but after some thought, it was placed near the section on statistical software because geospatial software can help give a geospatial perspective that uses location as one of the factors. For instance, seeing tornado data as a table or a variety of charts is fine, with the states listed along with the number of tornadoes experienced in them. However, presenting this same data on a map brings a much clearer picture of the data with respect to location. Geospatial analytics are discussed in more detail in a later section.

For example, if you want to see the damage caused by tornadoes for a specific year, you can see the data in a table format or use the 3D mapping function available within Excel (the button to access this feature is shown in Figure 6.1). Which would be a better way of presenting the data?

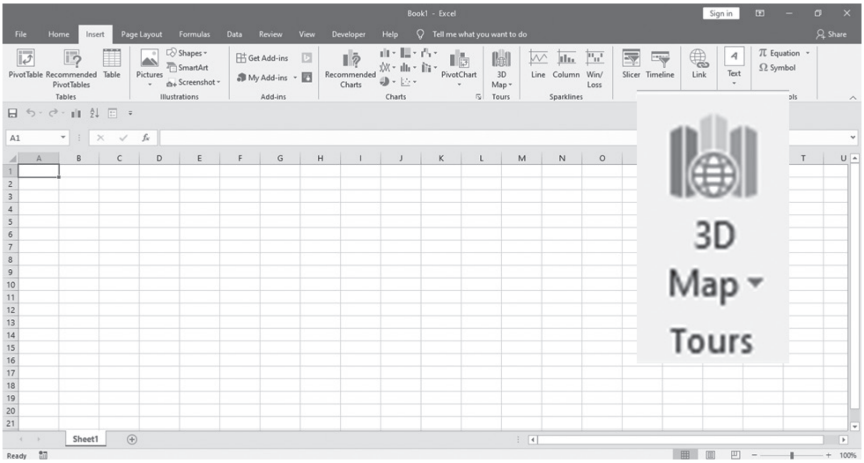


FIGURE 6.1 Zoom of toolbar option for 3D mapping in Excel

The presentation of a map sometimes gives a clearer picture of the data than any chart or graph would. Figure 6.1 leads you to 3D mapping, which in turn can display the data as a *choropleth* map, which shows the darker or lighter colors on states that have more (or less) tornadoes.

There are plenty of software packages available for geospatial analysis, some which require a subscription and others free and open source. We address each of these briefly.

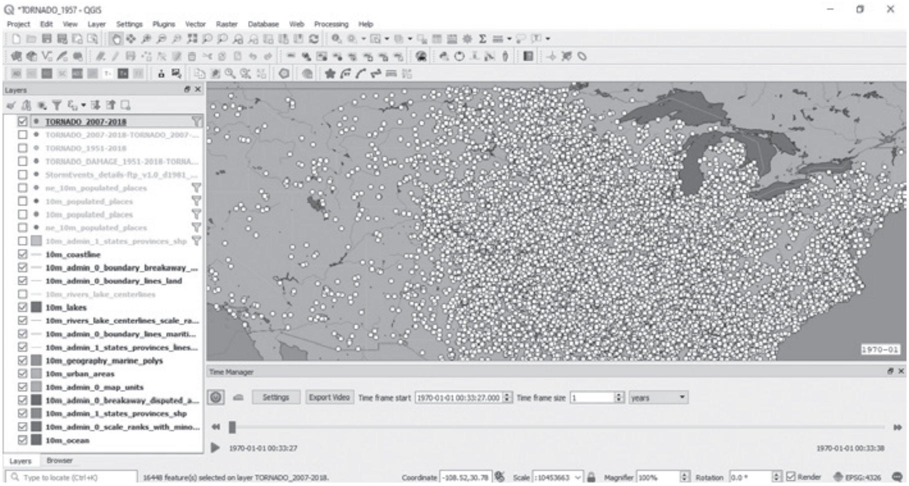
6.6.1 ARCGIS

ARCGIS is the preferred GIS software that is used by the US government. An individual can obtain a yearly subscription for the ARCGIS system for about $100 (as of the writing of this book), and this includes a download of the desktop version, which is very flexible and adaptable to a variety of databases including comma separated value (CSV) files (as the examples above use). A sample screenshot can be seen, along with other information about downloads of the GIS software at *www.esri.com/en-us/home*. With some practice and some of the references included at the back of this book, there is no reason why any analyst should not be able to use this application successfully. ARCGIS is considered by many geospatial analysts to be the pinnacle of GIS applications, and it is used throughout the world by both public and private organizations. It is used in intelligence, crop surveillance, disease surveillance, and even something as innocuous as house hunting, but this application is very useful for any analyst who is interested in geospatial analysis.

A discussion of how these different geospatial statistics are compiled and calculated is included in a later chapter, but this section explains how data can be visualized in a manner that has an immediate impact. In addition to a static display, the data can also be displayed as a time lapse using the years that the data was collected (in this case between 2007 and 2018). The analyst will have to discover for themselves the power and flexibility of this tool.

6.6.2 QGIS

Think of QGIS as the free version of ARCGIS. Actually, no thinking is necessary: QGIS is a free version of that geospatial analysis application. Unfortunately, and this is true of almost all free and open source tools, the dynamics and usability of QGIS are not quite the same as ARCGIS. The main reason is that QGIS is designed for the analyst who does not mind learning the tool from the ground up. However, there is a lot to be said for the functionality and usability of QGIS. Once downloaded from *www.qgis.org*, the application screen will look something like this (of course, with data that was included in the previous ARCGIS example). Notice the similarities of the display and the map. ARCGIS, however, comes with the ability to load Open Street Map, while with QGIS, you will have to look around for a base map. Nonetheless, QGIS does come in handy for those who do not have the money to spend for a subscription or who need to use the application for commercial purposes.



☑ TORNADO_2007-2018
☐ TORNADO_2007-2018-TORNADO_2007-...
☐ TORNADO_1951-2018
☐ TORNADO_DAMAGE_1951-2018-TORNA...
☐ StormEvents_details-ftp_v1.0_d1981_...
☐ ne_10m_populated_places
☐ 10m_populated_places
☐ 10m_populated_places
☐ ne_10m_populated_places
☐ 10m_admin_1_states_provinces_shp
☑ 10m_coastline
☑ 10m_admin_0_boundary_breakaway_...
☑ 10m_admin_0_boundary_lines_land
☐ 10m_rivers_lake_centerlines
☑ 10m_lakes
☑ 10m_rivers_lake_centerlines_scale_ra...
☑ 10m_admin_0_boundary_lines_mariti...
☑ 10m_admin_1_states_provinces_lines...
☑ 10m_geography_marine_polys
☑ 10m_urban_areas
☑ 10m_admin_0_map_units
☑ 10m_admin_0_breakaway_disputed_a...
☑ 10m_admin_1_states_provinces_shp
☑ 10m_admin_0_scale_ranks_with_mino...
☑ 10m_ocean

Time frame start 1970-01-01 00:33:27.000 Time frame size 1 years

1970-01-01 00:33:27 1970-01-01 00:33:38

16448 feature(s) selected on layer TORNADO_2007-2018. Coordinate -108.52,30.78 Scale :10453663 Magnifier 100% Rotation 0.0 ° ☑ Render EPSG:4326

FIGURE 6.2 Quantum GIS (QGIS) map showing tornadoes between 2007 and 2018. *https://creativecommons.org/licenses/by-sa/3.0/*

Figure 6.2 shows the visual of tornadoes that have occurred in the United States between 2007 and 2018. The Time Manager node below the map shows a time lapse of the tornadoes as they occur, giving the viewer the ability to see when the tornadoes occur, not just as a lump sum approach that is depicted above.

CHAPTER 7

EFFECT SIZE

There are entire books on the effect size, and this topic can be found in a variety of blogs and statistics articles; some are listed in the reference section of this book. The only thing that you need to know about the effect size is the amount of association that one variable has with another. A few measures of this include Pearson's R, which is the conventional method of correlation, and Cohen's D, which is another method of attaining that association with the t-test. We address Pearson's R and Cohen's D here, which is a very common approach to arriving at the effect size. The importance of the effect size relates to the amount of sampling necessary to attain the numerical relationship. For example, if you want to know if one variable has a .80 relationship to another variable, you need to understand the effect size; in this case, there is a strong positive relationship between those two variables. You use that .80 in a sampling formula (to be addressed later) to determine the number of samples necessary to achieve this effect size. But how would you know where the data association existed at the moment? That would require the use of the formula addressed in the next section on correlation.

7.1 Correlation

Correlation is a method for finding a relationship between two (or more) variables. In this book, the focus is on using a formula for evaluating whether two variables are associated. A good book on how correlations can be abused (known as *spurious correlations*), please read *Spurious Correlations* [Vigen-2015]. The author of this book, Tyler Vigen, has his own website, *www.tylervigen.com*, that will amuse even the most entry level data analyst for hours. Basically, the book and the website poke fun at the seemingly endless associations between things such as movies by Nicholas Cage and the number of people who drowned by falling in a swimming pool. The very nature of correlations is to show an association, NOT a causation! This is crucial before even starting this discussion and now that is out of the way, on to the formula, which is as follows:

$$r = \frac{n\sum(xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

The formula looks extremely difficult, but the concepts in that formula are those that have already been covered in preceding text. In simple terms, we are taking the number of samples times the sum of the values of x times y and subtracting that from the sum of the x values times the sum of the y values. We then divide this result by the square root of the sample times the sum of x-squared minus the sum of x-collectively squared. We then repeat that for the y factor. At this point, it is very important to reiterate that the "sum of the x values times the y values" is NOT the same as the "sum of the x values times the sum of the y values." For example, Table 7.1 shows what happens if you have three x values and three y values (the values are in two separate columns).

Table 7.1 The multiplication of the X and Y variables

X	Y	X times Y
10	5	50
5	8	40
3	9	27
Sum X = 18	Sum Y = 22	

The "sum of X values * Y values" is the sum of 50, 40, and 27, which is 117. The "sum of the X values * the sum of the Y values" is 18 times 22, which is 396. You can see that these are very different results.

Once that is completed for the numerator, the denominator is also a little confusing. The brackets show the left side of the denominator as being "n times the sum of X squared" minus the "sum of X collectively squared." The same is on the right side, but using Y instead of X. What is the difference between "X squared" and "X collectively squared?" Please see Table 7.2.

Table 7.2 The steps for correlation formula in table format

X	X-squared	Sum of X Collectively Squared	Y	Y-squared	Sum of Y Collectively Squared
10	100	10	5	25	5
5	25	5	8	64	8
3	9	3	9	81	9
Sum X = 18	Sum of X-squared (or $\sum X^2$) = 134	Sum of X-Collectively Squared (or $(\sum X)^2 = 18^2 = 396$	Sum Y = 22	Sum of Y- Squared = 170	Sum of Y Collectively Squared = $22^2 = 484$

Understanding the differences in these symbols can make the correlation equation easier to make into an algorithm so that you can use it in any software. Once the formula is completed, the result (for the above table of values) is approximately -.98. This would be a very strong association between X and Y, which could be translated as "when X increases, so does Y."

7.1.1 Correlation does not mean causation, but ...

If you take anything away from the preceding formula and subsequent discussion, it should be that correlation is not causation. Just because something is correlated does not mean that one factor causes the other. To reverse that, however, *causation means correlation*. In other words, if one thing causes another, then they are correlated.

The hardest part of being an analyst is to not jump to conclusions. Once data is analyzed, the results are sometimes extrapolated into either disaster or dream. Those types of hyperbole help neither the analyst nor the decision maker. If there is a correlation (or effect size), then that will lead to another analytical evaluation. In the case above, there are only three samples, which should never be the case in a valid dataset. The example used here was designed to help you understand the formula without arduous calculations. This confirms that it is imperative that the analyst not jump to conclusions, since the validity of the data may not be confirmed at the time of the first analysis. In fact, there is a *type* error that is reserved for the misinterpretation of data. This is a Type IV error, and it is detailed in the article by Stephanie Glenn [Glenn-2015]. Basically, a Type IV error is a misinterpretation of the data for a variety of reasons, but one of them is using the wrong data for the analysis. This error is rampant within data analysis, and it is made by analysts in both the public and private sector. This leads to the analyst leaping to a conclusion that in all probability is wrong, or at best, only partially right.

The next section of this book addresses various methods to avoid a Type IV error and use the methods that have been addressed to make the best possible analysis, thereby providing a way to give decision makers results that they can use to make intelligent decisions.

CHAPTER 8

ANALYSIS PROCESS METHODS

This book addresses two methods for data analysis that apply to project managers, systems engineers, and cybersecurity specialists. The first one is the more conventional and standard method called the *Cross Industry Standards of Performance for Data Mining* or CRISP-DM. The second method is one that was described in a GIS analysis text by Andy Mitchell [Mitchell-2009]. In this text, Mitchell describes a method that best exemplifies the requirements process commonly implemented by all three professions. Although the method is introduced in a GIS specific context, the method can be applied to any project, whether that project is engineering or cybersecurity focused.

The most essential aspect of analytics is to understand the question. Please remember that every statistical error in statistics has to do with questions. The Type I error is considered a *false positive*, where the question that would normally be false is pronounced true. The Type II error is the *false negative*, where the question that would be really true is pronounced false. From the many years of teaching Statistics, the Type III error is the right answer to the wrong question, and the Type IV error is the wrong answer to the right question. Critical questioning is vital to a successful analysis project. Without further ado, here are the two analytical methods for subsequent analytical application.

8.1 CRISP-DM Method

The CRISP-DM method is a process that standardizes the different aspects of data analytics [Shearer-2000]. Because it is standardized, it should be used by all organizations whenever there is a need for data analytics. For project managers, think of the project process (Initiate, Plan, Execute, Monitor/Control, and Close) [PMI-2017]; for systems engineers, think of the systems engineering process (Concept, Development, Production, Utilization/Support, and Retirement) [Walden-2015]; and for the cybersecurity expert, the main areas of consideration are the Confidentiality, Integrity, and Availability (CIA) of the system in question. In the course of describing the CRISP-DM process, try to associate this process with the respective profession.

8.1.1 Understand the Organization

The first step in the CRISP-DM process is to know and comprehend the organization. This is true whether the work is in the public or private sector. For instance, the understanding of the organization starts with the different policies and procedures that exist within that organization. If the data that is analyzed contains sensitive private data, then you have to understand the limits placed on the sharing of that data. The same is true for data that is sensitive, such as that related to national security or classified data. You need to know how the organization handles that data in order to present those results. In each of the professions of project management, systems engineering, and cybersecurity, understanding the organization is something that must be understood prior to initiating any project. For example, if the project manager does not understand the organization, how would she know if the normal project organization is a matrix or project-centric organization? How would the systems engineer know the type of process necessary to develop specific software? How would the cybersecurity professional know what the network infrastructure is and how that relates to the security of the email server? All these must be considered when understanding the organization.

One of the areas that many professions seem to disregard is that of the political arena. Who wields the most power within a certain department? Who is more reasonable when it comes to project priorities? Who is the manager who gives their user ID and password to the administrative staff so that they can gain access to the manager's email and print it off for the manager every day (real life event!)? All these considerations are part of understanding the organization, especially when the analyst must gather data for the project, systems, or cyber areas.

8.1.2 Understanding the Data

After years of teaching statistics and data analytics, students seem to be stuck on this concept as something that is so intuitive, there is no need to even address it. Understanding the data *prior* to the actual analysis is both preferred and advantageous. For example, having seen and used the data on tornadoes released by the National Weather Service, the one column that contained damages caused by each tornado was in a "categorical" form, such as "25 K." By placing a letter next to the numbers, the only way to analyze this variable is as a category, not as part of a descriptive statistics initiative. You would be hard pressed to come up with a "mean" of a column with entries like 25 K, .25 K, or .3 M. As a result of this, a data transformation has to take place that converts these figures into numbers so that they can be used in numerical models. The time saved in analyzing transformed data is a major benefit of understanding the data. Understanding the data also includes the tools available to analyze said data. If the only tool available is Excel, you should know the advantages and disadvantages of such a tool in analyzing the data available through the organization. In addition to knowing the tools, you must recognize that if the data contains variables external to the organization, then there should be a data dictionary in the dataset that explains and describes those variables. If a column (or field in data speak) is titled FST_QTR, does that mean it is the *First Quarter* or *Fast Quarter*? Maybe it is the *Fastest Quarter*! By having a data dictionary, the author of the dataset can explain what this title describes in the data. It takes somewhere between 10 minutes and an hour to do this type of work for a dataset. Analysts don't always remember why they named a column as they did, so a data dictionary is a great memory jogger.

8.1.3 Preparing the Data

We discussed preparing the data in the previous section when referring to transforming the damaged data into numerical data, but preparing the data means much more than just transformation. The preparation of the data goes hand-in-hand with the requirements levied at the start of the project. By attaining and scrubbing the requirements for the data analysis, the analyst can prepare the data by ridding the dataset of unnecessary variables (such as columns and fields) that have no value in the subsequent analysis. This is not just an optional step, but something that saves time, bandwidth, and capacity. There is a finite capacity for any infrastructure and the least amount used will help with faster processing and, consequently, faster results. By not overwhelming the system with unnecessary data parts, the models are able to streamline the work. There are real-world situations that show that overwhelming the infrastructure brings it to a screeching halt. This happened to a federal agency that wanted to bring a new application online and did not sufficiently estimate the client response; the new app brought the system down until more memory was purchased. The time of unavailability was minimal, but the possible reputation cost was huge. Therefore, understanding the data also means understanding the audience for that data, which is covered in greater detail in a later section.

8.1.4 Analyze and Interpret the Data

As stated in the previous section, it is imperative that the analyst do the optimum work on the data. Too much analysis leads to data bias, such as overfitting, while too little analysis misses those patterns that may be a key to discovering something new about the data. The analysis begins with a view of the data in its most primal state. That means you must first look at the raw data and make a rough chart to see the data in a perspective that helps to describe it in limited terms. In order to analyze, the analyst must first be able to describe. This is the same with everything in life. Instructions on how to open a door require someone to know what a door looks like and where the door knob is located (for that matter, a description of the door knob). The same applies to analyzing data, where a description of that data (the mean, median, mode, standard deviation, and maybe even the IQR) would help the analyst to better analyze the data. For example, if the average number of minutes a computer is open is more than other computers in the same department, this could be a red flag to the cybersecurity personnel to take a closer look at that computer. If the average number of weeks required to complete a sprint on a specific project is more than the average of a similar project sprint, the project manager might want to do further research into the reasons for this disparity. If a systems engineer, after establishing a development plan, finds that the average programming time on that system is going to be more than a similar project with less functionality, then that will lead to some further discussions with the stakeholder. The analysis of data is much simpler than providing a regression or time series analysis on that data. It may be something that helps kick off the analysis process, but does not take an inordinate amount of time. What happens when there is bad analysis? A bad analysis leads to a bad interpretation. This then leads to a Type IV error, or the wrong answer to the right question. It is not enough to analyze data and interpret that data. We must ensure that the preliminary analysis is something that places the analyst on the right road.

8.1.5 Evaluate the Analysis

This step most often raises some questions from students. What is the difference between evaluation and analysis? Is analysis evaluation?

Evaluation is ensuring the analysis is valid. For instance, if a teacher wants to see how the students did on a test, the teacher would average the scores and then see how the students' scores individually compared with the average (much like the variance – hint, hint). However, to evaluate that result, the teacher would have to include some factors, such as the number of students in the class where the test occurred, as well as the time of day and the questions that were missed (was there that pesky question concerning evaluation and analysis on the test?). This, some say, is further analysis, but it is actually the *evaluation* of the analysis. Evaluation leads to recommendations; analysis does not always lead to recommendations. Analysis is part of a process whereas evaluation is the validation of that process.

8.1.6 Communicate and Deploy the Results

Once the data is analyzed and evaluated, the analyst must present the results to the stakeholder. This is probably the most important step of the process because it is this communication that tells the stakeholder what the data meant, not just what it represented. For example, let's assume that a project manager for the federal government was responsible for presenting the status of various projects, including on-time completion rates and assorted other requirements that the stakeholders needed to know. The presentation showed a pie chart (or other similar visual) that described the completion rates. The audience consisted of high-ranking government executives, all with their own questions. The most optimum approach was to present the data with two or three charts to ensure that most of the obvious questions were answered, such as "How many projects were undertaken?" or "How many projects were completed?" or "How many projects were completed on time?" But not just those questions were part of the deployment part of the presentation. There were handouts of the presentation with some detailed data that explained the charts should the stakeholder need that for later or for reference during the presentation. In addition, the source of the data was included, along with a short methodology on how the data was collected and organized. All of this preparation set the stage for the stakeholders.

In one instance, there was a slide presentation consisting of a line chart showing the time frame as the "x" axis and cost as the "y" axis. The visual showed two lines, one for the planned cost and the other for the actual cost of the project by month. The normal presentation would have included the cumulative cost, but this visual showed the monthly costs individually. When there were dissimilar peaks between the planned and actual cost, there was a cartoon balloon at that point with a word or two of the explanation of the difference. This presentation proved to be a big hit with the executives, who normally did not have the time for long presentations. In addition, the visual was a great way for the executives receiving the presentation to review with their supervisors. Some analysts may find the perfect combination for communicating the data results, while most will have to adjust to each audience, which is par for the course.

This is the last step in the CRISP-DM process, but it is extremely important to note that just because an analyst thinks they understand the data does not mean that fact will continue to be true throughout the CRISP-DM process. There will probably be a point along the process where the analyst has to study the data again. This is not abnormal in analytics, whether the analyst is studying programming timeframes or security access lists. The bottom line is that the CRISP-DM is not necessarily sequential until the end, and can move forward or backward along the process. Being aware of this will prevent the analyst from driving ahead in a snowstorm instead of going back and staying one more night in the hotel.

8.2 Alternative Method

This section follows Andy Mitchell's text describing how to perform statistics on geographic data [Mitchell-2009]. More important than the actual method is the process for the geographic analysis, which is described starting on page 6 of Mitchell's text. The steps that the book uses take a little from the CRISP-DM, but the process is so descriptive that it would be neglectful not to include them as an alternative to the CRISP-DM method. Each step will be described using real-world situations whenever possible.

8.2.1 Framing the Question

Those with children likely already understand *framing the question*. Framing the question, or in project management, determining the scope, is widely accepted but rarely applied. In both the public and private sector, the question is seldom framed well enough in the beginning of the project to be able to finish the project with that question intact. In data science, the question is sometimes a statement, otherwise known as a hypothesis. Questions like "Is climate change related to carbon emissions?" sometimes take on the declarative form, "Climate change is related to carbon emissions," which demands the necessary research and data analysis to either prove this statement true or false. Analysts understand these statements are hypotheses, whereas those that are not data analysts take the statements as fact. Framing the question not only helps to focus the data collection and analysis, but also provides a goal for the analysis. The scope of the data analysis is set with one statement, and the analysis is done with this statement in mind. Project managers try to attain this, but in many situations the frame of the question takes the form of a list of requirements. These requirements are then transformed into the scope and then the cost and schedule are determined. In the case of systems engineering, the Concept and Development Stages are those that form the basis for the engineering process, but the important aspect is that, according to the *Systems Engineering Handbook*, there are a number of processes that mirror the CRISP-DM and this alternative in the framing the question area. The *Technical Processes* in the handbook reflect a focus on the business as well as the requirements for the system [Walden-2015, p. 2]. It is interesting that the project management handbook, the PMBOK, does not have requirements as a major focus, but relies on the Scope Management for including the requirements [PMI-2017]. Now, this is not a major concern, because in both instances, the requirements do form the basis for a project or system. The main reason for raising the requirements issue is that requirements are essential to help frame the question. The better the requirements, the better the probability that the project will be a success.

8.2.2 Understanding Data

Here we go again! In this case, the understanding falls along the lines of the types of data. Remember the discussion about discrete and continuous data? Well, this is what the understanding the data is all about. If the analyst wanted to take the mean, median, and mode of the damage data from the tornado spreadsheets offered by the National Weather Service, that would be a problem because the damage data is in categorical form (nominal form) and not numerical form. This would mean some transformation is necessary to satisfy the requirements of the descriptive statistics that the analyst needs. The realization of the "framing of the question" part comes to light here while trying to understand the data.

One other aspect of understanding the data is the idea of origination. Where did the data originate? If the data came from the federal government, or a reputable data repository like Johns Hopkins University (*www.jhu.edu*), then that might make the analyst a little more comfortable than an organization that has a bias towards certain data that only offers one point of view. It is up to the analyst to be careful with the data that they have collected and check for verification.

8.2.3 Choose a Method

When flying a Piper Cub, or any tail-wheel aircraft, the pilot knows that there are two ways of landing the airplane. The first method is a full stall landing, where the pilot must establish a landing speed on the final approach to the runway. The second method is a wheel landing, where that landing speed is faster than the full stall approach. It is wrong to use a method of last-minute decision making when deciding on the type of landing, just the same as it is wrong to figure out which data analytics method to use as other methods are being attempted.

With project management, systems engineering, and cybersecurity, the available guidance (available in the reference section) discuss the different planning aspects, and some of the planning is determining the different methods to conduct the project or system. For instance, if the team is a matrix management team, then the documentation and stakeholders are different than a projectized team. In the same way, determining the analytical method takes some research to optimize the time and effort for analyzing the data. For instance, there is a visual presentation called Gapminder (*www.gapminder.org*) that demands that the data be in a particular format to display the time lapse function (this app is particularly great for a free and open source application). If the analyst does not make this adjustment, the visualization will not appear. Knowing the data and choosing the method of presentation are two aspects that are necessary in this instance. A free download of the tool to get a very good depiction of data can be found at the Gapminder site mentioned above. There is a

YouTube video that shows how to get the COVID-19 data from the Johns Hopkins University Corona Virus Data Set [Dong-2020] and how to visualize this data in Gapminder (*youtu.be/asMT1jYl3F0*). The citation requirements for the data taken from the Johns Hopkins data port are available online.[10]

8.2.4 Calculate the Statistics

This step sounds like a natural step to attain the analysis necessary, but there is more to this than just calculating blindly. In this text, as in all statistics texts, the formula is presented not just as a way to take up space, but to tell the analyst how the result was obtained. The idea of formula memorization is both ridiculous and unnecessary. Understanding the formula is part of the process of understanding the theory of statistics and how to apply that theory. For instance, if you want to calculate the standard deviation, it is important to note that the standard deviation directly involves the mean, which is affected by the extreme values in the data. Understanding this now gives you the ability to also calculate the median absolute deviation, which is not highly affected by the values, and with which you can compare results. When you know the very basics about the formula, you can make any adjustments in the parameters to allow for any peculiarities in the formula's structure.

8.2.5 Interpret the Statistics

Calculating is one thing, but interpreting is the most important part of both data analysis and analysis in general. One of the main reasons for presenting the statistics refresher as a formula-centric format is to ensure that you understand the interpretation as part of the function of analysis. When you calculate a mean, that produces a number. What you do with that number is the interpretation of the statistic. For instance, during the COVID-19 pandemic, there were many times when there was the reporting of the total number of cases, which was sometimes the highlight of the report. In one instance, the report showed that there were over 3.5 million cases of COVID-19 in the United States.[11] Although this figure looks daunting, the total population of the United States is 320 million, which means that the total number of cases is just slightly more than 1% of the total population of the United States. When interpreted in different ways, the result of a data

[10] *https://github.com/CSSEGISandData/COVID-19*

[11] *https://www.globaltimes.cn/content/1194837.shtml*

analysis can sway opinions and even drive certain decisions. The interpretation is also part of a Type error, specifically Type IV which, as stated previously, is the wrong answer to the right question.

8.2.6 Test the Significance of the Statistics

If you have had any experience with statistics, you have heard the phrase *statistically significant.* What does that mean? Statistical significance is a phrase that means that the results did not occur by chance. What would happen if an analyst produced results, reported those results, and then found out that the results were a product of just good (or bad) luck? The analyst would never be able to reproduce those results and, probably more importantly, the decision made because of that data could be an erroneous decision. The normal language used in statistics is that *the result is statistically significant at the 95% confidence level* or something similar. But what does this really mean? The translation is that there is only a 5% probability that the result of the statistics test is the result of chance. Statistical significance is important, but certainly not the only way to test the statistical significance of the data. In fact, one source [Reinhart-2015] details how the *p-value* through which statistical significance is based has its flaws and should be only one of many evaluations of the data. When something like statistical significance is raised, it should include a number of different evaluations that encompass a hand-in-hand approach to data analysis. Relying on just one factor when it comes to statistical significance is much like relying on just a hammer to build a room addition.

8.2.7 Question the Results

The hardest part of finishing a job is to self-assess that job. By questioning the results of the data analysis, the analyst is raising possible concerns that will undoubtedly be raised by others in the field. Questioning the results does not mean to second guess the process, but to honestly evaluate the analytic effort so that any possible issues and concerns can be addressed or qualified in the final report.

CHAPTER 9

DATA ANALYTICS THINKING

There was a guest on National Public Radio who talked about how young people are not learning good analytical techniques, blaming it partially on the digital age where computers do the thinking for us as individuals. Project managers, systems engineers, and cybersecurity professionals must have the ability to do analytical thinking to address the many risks that are associated with their respective professions. They do this through a method that systems engineers call *decomposition*. Decomposition means to take something large and break it down into smaller components. By doing this, the task seems smaller to complete and can be completed one task at a time. This works whether it is building a house or programming software. Decomposition can be used for analytical thinking, specifically for data analytics. There are several ways to decompose tasks into subtasks. One of these is historical decomposition, which is used in project management and systems engineering. Cybersecurity uses historical references as a means of possible penetration tests and social engineering, specifically against the employees of an organization.

Decomposing by historical comparison is used in project management when estimating costs or work. For instance, when determining the cost of a project, historical references can give the project manager a foundation for estimating the cost of the current project. The same is true of systems engineering when estimating costs or work. For cybersecurity, the historical reference can change the course of the analysis. In one example, it was shown that a specific malware was spread by what is known as *sneakerware,* where it was transferred from machine to machine through the insertion of a USB drive [Walden-2015, p. 43]. The historical aspect of this reference is that no matter how secure one thinks their network is, the manual insertion of outside objects can (and will) increase the risk of malware distribution.

9.1 Elements of Data Analytic Thinking

There are several elements of analytical thinking that need to be delineated so that analysts can better understand the nature of this type of thinking. In the same way that CRISP-DM tries to drive the analyst to a better standard of analysis, the mindset of a data analyst should drive the analyst to a better way of thinking of data (not just as data, but part of a larger part of the overall business reasoning and development). The elements of data analytics thinking include determining the structure of the data, determining the elements of analysis that can be found in that data, and finally determining the elements of the analysis to be found outside of the data. Each of these areas will be discussed, with particular attention to the three professions mentioned throughout this text. Please pay special attention to the fact that many of these elements overlap in these professions.

9.1.1 Data Structure

The structure of the data is much like determining the structure of a building. How many floors does the building have? How many bathrooms, bedrooms, closets? What are the dimensions of each of the rooms? Does the building contain the necessary items to house the pertinent organizations? All these questions can be directed toward the data structure.

For instance, does the data have a data dictionary or data definition (as has been discussed previously)? Without a standard for the column (field) headings, there could be confusion as to what the data contains and what variables are important. Trying to decipher a field name is like decrypting a 20-character password without any hints. Many data field names are relatively standard, like FNAME for first name and DATE for date. Some, like PID, could mean "personal identification" or "person in detention." A data dictionary prevents this type of confusion.

All three professions need to always have a data dictionary to prevent the confusion of commonly used acronyms. POV could mean "point of view" or "privately owned vehicle." PORT could mean a real shipping port in the case of project management, a computer port in cybersecurity, or an input port in the case of systems engineering. Understanding the perspective of the analysis can help with determining the structure of the data.

9.1.2 Analysis Elements Inside Data

Computers both help and hurt when determining the elements for the analysis. Computer technology has allowed us to download data that is voluminous compared to that of time past. This technology helps find only those fields that are pertinent, but it can detract from the analysis because the analyst should have already surveyed the data fields to determine what they would need to do the analysis. What happens in this case is that the analyst has additional fields that may not pertain to the current analysis, but the analyst may deem those fields useful because they are available in the data download. This could lead to some bias on the part of the analyst and skew the data in favor of the hypothesis rather than objectively analyze it.

For example, if a project manager needs cost estimates and sees that there is a dataset with historical cost on projects similar to his, but then sees that there are additional costs that do not pertain to his project, but *could* pertain, he may use those costs in the cost estimates even those they are not pertinent to the current analysis. If a systems engineer is performing a process evaluation and sees from the data that there is an additional field that (again) could pertain to her analysis, she may use that field even though preliminary survey would have discounted this item. What about the cybersecurity professional who wants to check historical data on port activation and sees a field that concerns document access and decides to include that analysis? In all three cases, the data fields are not the main ones that should be considered and, consequently, will affect the overall data analysis results.

9.1.3 Analysis Elements Outside Data

There is a story about a young cybersecurity professional talking with a more mature (older) cybersecurity professional about possible malware on a firewall. The younger individual was flummoxed about the tracking data, which seemed to point to a very specific module, but the malware was not there. In frustration, the younger professional asked the older professional about the problem.

“I would look in this other module,” the older man said.

“Why would you do that? There is no data that points in that direction,” the younger one retorted.

“Because that is where we put the malware when we tested that code years ago.”

The fact that the analyst does not take data from other sources outside of the ones that they are using limits their overall knowledge of the data.

In the same way, project managers, systems engineers, and cybersecurity analysts should take into consideration data outside of the normal data that they see daily.

For instance, a project manager might talk with another project manager who had a similar project and see their data on cost and work estimates; a systems engineer might look at case studies on system successes and failures [Walden-2015, pp. 39-45], and the cybersecurity analyst might also read case studies as above, or seek substantive help provided by outside agencies like the National Institute of Standards and Technology (NIST) through their web site that contains helpful methods for analysis of cybersecurity for businesses.[12] There is something to be said for the amount of data and references that exist outside of downloaded datasets.

9.2 There is a "Why" in Analysis

The hardest part of analysis is asking the "why" questions. In all three professions that apply these ideas, the *why* is the most important. Without asking (and answering) the why questions, projects, systems, and cybersecurity would fail.

For example, when was the last time that a project manager experienced a failure and asked why that failure occurred? If a system does not function as it should, does the systems engineer ask why that function did not perform? If there is a computer security breach, does the cybersecurity analyst not ask why that happened?

Is the aspect of why questions different from the phrase "why question?" If the project manager fails to question correctly, the project has a great probability of failure. This happens in daily life when having a flat tire and asking why that flat tire occurred. There are those who do not question the situation, just fix the tire and move on. However, maybe parking in the lumber section of the home store parking lot will increase the probability of getting a nail in the tire. The one question asked based on the "why" is the one that will save the individual a possible future flat tire.

Data is no different in this regard. There are "why" questions that need to be asked concerning the data and these questions should be part of the

[12] *https://www.nist.gov/news-events/news/2020/02/nist-offers-strategies-help-businesses-secure-their-cyber-supply-chains*

pre-analytic checklist that go through the head of the analyst. Remember that the computer gives the analyst the ability, not the reason for the analysis. If the word requirements have been recirculated throughout this text, it is because it is the most important part of the projects, systems, or cybersecurity. Rather that reiterate what has been written above about requirements, let's take a look at the "V's" that are associated with data and use that as a starting point to talk the "why" question.

9.2.1 The "V's" in Data

There are whole articles written on the variety of "V's" in data science. Some say that there are 10[13], while others say that there are 42.[14] For the purposes of this book, there will be only 4 "V's" that will pertain to the application of the analytical methods discussed here. The four will be velocity, variety, volume, and vulnerability (the last one is added here). Let's discuss each one individually.

9.2.1.1 Data Velocity

For the purposes of this book, velocity with relation to data means the currency of the target data. Current data, for the most part, is much more valuable than older (outdated) data. For instance, if the project manager is looking at scheduling data on a project that is more than 10 years old, then there are certain qualifying aspects of that data that are necessary to comprehend. Remember that any cost data that is in that project data file must be "normalized" against the current monetary value. If the project manager just considers inflation in the 10-year calculation, the cost data could be skewed positively or negatively, so it is imperative that these types of standardization be achieved for any data that is not current. Another aspect to consider is demographics. Today, the United States has approximately 320,000,000 people that live either in the continental United States or its outlying areas. If a systems engineer takes into consideration a process that incorporates access for US citizens over the age of 65, this number will have changed over the course of several years so that must be considered in the overall calculation. Finally, what about computer capacity? If a cybersecurity analyst does not consider the Random Access Memory (RAM) capacity in a computer (PC), and the subsequent changes to this capacity, there is a possibility that certain PCs are not up to date and lack the needed capacity

[13] *https://tdwi.org/articles/2017/02/08/10-vs-of-big-data.aspx*

[14] *https://www.kdnuggets.com/2017/04/42-vs-big-data-data-science.html*

to manage the data. Therefore, the data from the past would have to be "weighted" in order to match the capacity of today.

Analysis of data that is not current presents some challenges to the analyst. None of these challenges are overwhelming, but they are necessary in order to sufficiently allow for the passage of time.

9.2.1.2 Data Variety

The variety of data, for the purposes of this book, is the data collected with *entropy* in mind. Entropy in data science denotes how much the data is dissimilar. The term "entropy" was used by Claude Shannon in his penultimate treatise on information in the digital age [Shannon-1948]. In his article, *A Mathematical Theory of Communication*, Shannon raises entropy as a standard to base the transference of data. The theory, now called *Shannon's Entropy* in many federal government cybersecurity password policy documents, is a method to determine the strength of a password. If the data is collected from one specific area, then the sampling should be different enough to make a comparison. For example, taking a sample from the students at one university does not give the analyst a perspective of the students at another university (if the result was supposed to show that). By establishing a variety, the analyst is mitigating a possible sampling bias that may exist.

An example of this is the sampling done for the 1948 presidential election between Thomas Dewey (then governor of New York) and Harry Truman (the acting president after Franklin Delano Roosevelt's death). The sampling was done by phone, which biased the sampling since most phones were owned by individuals in the upper income level, who supported Dewey. A newspaper took those results, which showed Dewey winning the election. Truman ended up winning and showed a copy of the faulty headline. An example of biased sampling and, consequently, not a variety of data.

9.2.1.3 Data Volume

There is an old adage that "if a little garlic is good, a lot must be great." If only life were that simple, but sometimes a lot of something does not make it good. In fact, there is a logic fallacy known as the *Fallacy of Composition*, which means that making something better can be a fallacy.[15]

[15] *https://www.merriam-webster.com/dictionary/fallacy%20of%20composition*

Unfortunately, there are exceptions to the rule and data is one of those exceptions. The more data that an analyst can use, the greater the ability to use that data for various models and evaluations. An analyst does not want to use all the data at once, but can split the data (using random sampling or similar sampling method) into a smaller dataset for testing and evaluation. Once that testing is completed, then the larger dataset can be evaluated. What this does is give the analyst an idea of the validity of the test and the usefulness of the methods in the testing. The smaller dataset is called a "training dataset." However, the analyst must have a large dataset with which to start the overall testing, and so is using volume as a method to ensure the validity of the analysis.

In project management, a greater volume does not mean historical data but data that can be applied directly to the project or the team. For instance, at one federal agency, there was a software application that allowed retired members of that agency to maintain a competency rating on certain skills and knowledge. In this specific case, it was their experience with certain countries and languages. The agency kept this data current and had the ability to recall the employees if they needed them for a special project that required someone of their skill level. The application was so successful that other agencies were using it or developing their own. In other words, the volume of data was huge, but the result of such a repository of data was that any project manager had the ability to seek and find individuals of certain skills for their project.

In the case of systems engineers, data volume is their friend. Simulation and modeling data are used extensively by systems engineers to test and validate their systems. The data from these tests are used to ensure that all requirements are satisfied, and therefore ensure the stakeholder is satisfied [Walden-2015, pp 89-91]. The idea that a volume of data might be a hindrance is not in the eye of the systems engineer (or the project manager, for that matter).

What about the cybersecurity analyst? He must use data to analyze computer security, and the more data available, the more the longitudinal information is useful. By having several years of data on accessing certain applications, it is nice to see the time the applications were accessed placed in a time series graph and compared year-to-year or month-to-month to see if there were any anomalies or problem patterns. The more data that is there, the more that can be analyzed. It is literally that simple.

9.2.1.4 Data Vulnerability

Data vulnerability is not included in any of the data science "V's". With so many technological advances, and the miniaturization of everything, it is important to consider the vulnerability of data. One might say immediately that vulnerability focuses on the cybersecurity analyst and that, in part, is true. Data vulnerability is something that the cybersecurity analyst must be aware of and manage. The Common Vulnerability Scoring System (or CVSS) is a way of rating data and the system where it resides. There are several web sites that discuss the CVSS, but the government's National Vulnerability Database[16] is good because it includes a calculator for rating the data and system. Although there are plenty of scoring systems out there, this one is used by the NIST. NIST has some of the best cybersecurity professionals, and we recommend the analyst dig deep into the NIST website for calculators, training articles, and formulas to help them do their work. The NIST site also has a section for data scientists to analyze cyber data located at the site.[17] This isn't just for cybersecurity professionals, but it is a good place to start.

Data vulnerability is not just for cybersecurity analysts. Project managers and systems engineers both have a need to understand data vulnerability. If project data is vulnerable, it could be accessed by those who are not on the team or, worse, by those that could place malware on the data to void the validity of the data. This is something that every project manager should be aware of and attempt to prevent. Ways of preventing this would be to limit access to certain folders and files that contain project data. At one federal agency, there are timelines that are kept in a folder out of sight of the project team because those timelines were crucial to the daily workload tracking and presentation. Later, when a task was questioned, the timeline for that day was recovered and shown to the government sponsor, alleviating any consternation about the project team involvement. Systems engineers have data that contain simulation and testing data, along with any phase development or validation and verification. These should be kept current and the access should be controlled so that testing data maintains a high degree of veracity. The cybersecurity analyst has data that contains a variety of accesses and, if the analyst is a systems administrator, is extremely sensitive. The confidentiality of data is paramount for highly vulnerable

[16] *https://nvd.nist.gov/vuln-metrics/cvss*

[17] *https://www.nist.gov/topics/mathematics-statistics*

datasets. By protecting those datasets, the vulnerability decreases and the veracity increases.

9.3 Risk

As with some of the other subjects addressed in this book, risk is something that whole volumes have been written on. All professionals should understand that risk is something that exists in all endeavors, whether it is taking a shower in the morning or driving to work at night.

Data has a distinct and inextricable link to risk, since it is data that establishes the probability the risk will happen and the consequences of that risk actually happening. The two areas of likelihood and consequences (or probability and impact in some circles) make up risk in its most elemental form. "Risk" stands for ***R***ecognition of ***I***n-***S***ufficient ***K***nowledge. No matter how much data is presented to mitigate (lessen) risk, project teams cannot eliminate risk as a factor in project completion.

The three professions go about risk in relatively similar ways. The project manager takes their cue from the *Project Management Body of Knowledge* (PMBOK), which describes the process of approaching the risk portion of the project by establishing a Risk Management Plan. This plan consists of several stages, from recognizing and listing the risks to mitigating the risks [PMI-2017, p. 395]. Although there are specific steps within the Risk Management philosophy, the real data exists around the risk identification, which is central to the overall risk management process. The rule of thumb for personal project management is that *for every requirement, there must be at least one risk*. This is true for every project undertaken, whether it is to build a room, remove a closet, or fix a personal computer. How many times has a tech support specialist told someone that it should only take 5 minutes for a task? How long did it really take?

For systems engineers, the process is outlined in the INCOSE *Systems Engineering Handbook*, which focuses on using the analysis of risks as an accompanying step to identifying the risks [Walden-2015, p. 116]. This makes sense in that identification and analysis go together like peas and carrots. The project manager is someone who has the natural tendency to use the identification of risks to start analyzing those risks, and the same is true for systems engineers. In both situations, the likelihood and consequences of qualifying (and later quantifying) risks is something that data collection and analysis solve.

What about cybersecurity analysts? The risk analysis starts with a combination of describing the system and then listing the risks.[18] This may seem intuitive, but the first step in this process is the brainstorming of risks against known threats, not known threats against a specific architecture or infrastructure. The description of the systems sets the stage for the requirements, which then translates into risks against those requirements. Additionally, cybersecurity has an entire set of government guides and regulations that govern risk assessment through the National Institute of Standards and Technology (NIST). The NIST governing document, 800-53, is commonly used by US government agencies to determine risk at any point of the project. The NIST Risk Management Framework, of which 800-53 is just one reference, is relatively straight forward and is composed of steps that range from categorizing systems to selecting and monitoring the security controls of that system. The framework is well defined (*csrc.nist.gov/Projects/Risk-Management/rmf-overview*). One note concerning the cybersecurity risk framework. The first step is categorizing systems, but the only way to do that is to understand the business (back to CRISP-DM), along with gathering requirements for any future systems. The linking of these concepts is so evident that one can see how important data is to making these types of logical decisions.

From what has been described thus far, it is evident that, in order to adequately address risk, data is necessary. Where do you get this data? How is the data combined with the risk assessment to present a cogent representation of the risk to decision makers?

It all depends on what the analysis needs to address. For instance, if the project manager is concerned with the project team getting ill from COVID-19, then that manager should get data that shows the percentage of positive COVID-19 tests in the area where the team resides. For virtual teams, this is even more complicated, because the area may be anywhere in the United States or even abroad. If the systems engineer wants to know the possibility of foul weather in the area where the servers will be located, then that data can be obtained from the National Weather Service and would consist of hazardous weather maps and what the probability of foul weather at the desired location. The cybersecurity analyst wants data on the administration of the servers, access control lists, employee trustworthiness

[18] *https://www.tylercybersecurity.com/blog/6-steps-to-a-cybersecurity-risk-assessment*

certificates (or something similar), and any other data that could affect the CIA areas.

What about the two areas mentioned earlier – likelihood and consequences? We will address them as probability and impact. Since the subject of probability has already been breached (sorry about the use of "breach," cybersecurity professionals!), the next two sections address the probability of risk and the impact of risk using data analysis as the foundation for the sections.

9.3.1 Probability of Risk

The probability of risk is relatively straight forward. If one of the risks is a weather-related disaster, then the probability of that risk occurring is the frequency of that type of disaster occurring divided by the total number of this type of disaster. For example, if the number of tornadoes in the county where the project or system is scheduled to be deployed is seven per year, this really does not say much about the probability of the tornado happening in the area. However, if the total number of tornadoes in the United States is 2000, then the probability of the tornado happening in that specific area of the United States is 7 out of 2000 or 7/2000 or .0035 or approximately .35%. This number is less than a half of a percent! The project manager or systems engineer could make it a little more specific by taking into consideration the 7 tornadoes divided by the number of tornadoes happening in the state. However, when a data analyst does this, that is, she limits the data comparison, then that could affect the overall probability percentage. The best way to avoid this type of data difference is to discuss the data plan with your project champion or other major stakeholder to reconcile any concerns that they might have with the data analysis.

In the systems engineering aspect of data analysis having to do with the probability of risk, the data used here might be more module specific, such as a validity or verification check and how many times the module functions within specifications. Again, a discussion with the stakeholder would help to alleviate any concerns. If the data is for the systems engineer, not all risks may be found in the *Systems Engineering Handbook* [Walden-2015, p. 117]. Sometimes, risks result in opportunities and, as such, these should also be listed on the risk table.

What about cybersecurity analysts? The probability of a risk occurring takes into consideration data about the computer infrastructure. For instance, the probability of breaking a 14-character password composed of letters, numbers, and special characters, is much less than a password that is composed of 8 characters of the same composition. There is a formula available to the cybersecurity analyst to address this issue on the Scientific American website [Delahaye-2019]. It is interesting that in this situation, the data is not found on a website or repository, but is found by generating data from a formula, which is used to predict possibilities of the future.[19] However, please understand that forecasting and prediction have their disadvantages, which are not addressed in this book. Data is the focus of calculating any probability of risk, no matter the profession.

One last thing before moving on: please refer to the first formula on probability, presented at the beginning of the probability section. Notice how simple it is and how it says all that needs to be said. This formula is all you need to determine the probability of risk for the future risk table. The next step is not so simple, and demands the use of data to make sense of the overall risk impact or consequence.

9.3.2 Risk Impact

The second part of risk qualification (or risk quantification) is the impact (or consequences) for each risk. The probability of the risk is attained from the previous calculations, but how is the impact calculated? One must look at the data to calculate the risk impact. For instance, if a project manager is trying to ascertain the impact of fraud on a project location, one of the ways to calculate the impact is to research how much fraud is reported in the proximity of that location. In this way, the project manager can obtain an earned value for that impact. An earned value is obtained by multiplying the probability by the amount; if the probability of fraud is 20% and the average amount of fraud is $100,000, then the earned value would be .20 * 100,000 or 20,000. This would more than suffice for the impact of fraud as part of the risk calculation. If the project manager wanted to state that the impact was $100,000, then that would be fine, but should include a qualifying remark that this was the average amount of fraud for the calculation.

For the systems engineer, the risk impact would probably be related to time, such as the impact would cost 3.5 hours more computer time per week. In this case, the systems engineer could also use the earned value formula by multiplying the probability of that risk occurring by the average loss in time. For example, if the average loss is 3.5 hours and the probability

[19] *https://www.mathworks.com/discovery/predictive-modeling.html*

is 10% or .10, then the formula is 3.5 * .10 or .35 hours. However, and for the purposes of this book, taking the average time loss is sufficient for this example.

What about the cybersecurity analyst? The cybersecurity community has data available to show a variety of vulnerabilities and can use that data to arrive at a quantitative risk analysis. For example, the probability that a 14-character password can be hacked is a formula that is available on the Internet. How can the cybersecurity analyst use this for the impact risk analysis? The impact is what the password is protecting. If the password (or access in general) is protecting financial data for the company, then the impact of that data breach would be the entire financial foundation of the company. For example, if a private citizen has a bank account with a password that takes less time to crack than 90 days, then that bank account could be hacked since all passwords should be changed every 90 days. In other words, the probability of the data breach is the number of days for cracking the password divided by 90 days, and the impact is the amount in that bank account. Again, this is based on data that is available concerning the amount in the bank account, as well as the data that describes the password.

9.3.3 The Risk Chart

The risk chart is common to all three professions. However, it is essential that the low, moderate, or high-risk areas are properly standardized prior to the chart being completed. For instance, if the $100,000 impact mentioned above is considered low impact, then that needs to be mentioned in the chart. Color coding is done to differentiate between low, moderate, and

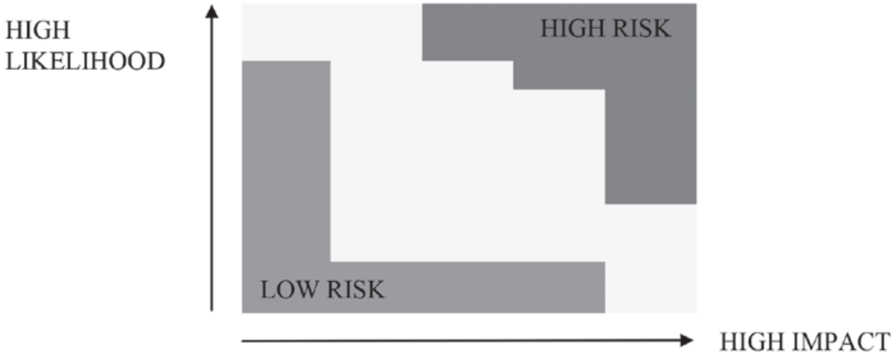


FIGURE 9.1 Example of a Risk Table

high, and uses intuitive colors that immediately show whether a risk needs no, some, or all of the attention of the project manager, systems engineer, or cybersecurity analyst. An example of a risk chart is shown in Figure 9.1, but realize that this is just one example and should not be taken as the only way to display this data. This is a form of displaying data that the analyst collected to make the chart. It is the same as gathering data for any other chart, including bar and line charts. Remember that collecting and analyzing data does not necessarily mean that the data is in some spreadsheet or other medium. It could be data that the analyst compiles at the moment for analysis. The two lines along the side and on the bottom of the chart denote probability and impact. As the probability (here, the y axis) increases, the rating changes from green to yellow. As the impact (here, the x axis) increases, the rating also changes from green to yellow.

CHAPTER 10

WHERE'S THE DATA?

One of the most difficult decisions that any professional dealing with data has to make is what data to get and where to get it. This is especially true of risk data, specifically, the probability of a risk occurring. For instance, if there is listed as a project risk the probability of weather-related disasters, what does the project manager, systems engineer, or cybersecurity analyst do in this situation?

The first way of getting valid data is to ensure the data is relatively current, but if it has an historical component that is fine also. For instance, the risk of storms ranges from hail to tornadoes and beyond, so where does the professional get this data? The first question, taken from the previous sections, should be what are the requirements for the data?

We now explore the very fundamentals of the locations of data, the methods for retrieving the data, and the reasons for retrieving the data. Each profession has their methods and techniques for collecting and using data.

10.1 Data Locations

Collecting data in this day and age is one of the easiest tasks on the analysts "to do" list. The proliferation of data, by both public and private sources, is so pervasive that data is aggregated by businesses daily. If one just looks around at the news broadcasts, social media, or any talk radio show, the data is often the focal point of the discussion. Because of this, it is essential that the data analyst get a variety of data (remember the "V" discussed earlier) and ensure that the data is not just from one source, if at all possible.

Where would a project manager collect data on projects that compare with the current one? Where would a systems engineer gather data on systems? Where would a cybersecurity professional get data on system access and protection?

The answer to all three of these questions, at least as part of the preliminary research, is *www.data.gov*, a massive repository of data available to anyone with a computer (permissions found here: *www.data.gov/privacy-policy#site_security*).

This government website has a search bar that points the analyst to any type of data that may meet the needs of the analyst. For instance, if the project manager wanted to know the cost of facility projects in 2018 to form an estimate for a possible facility project that they are undertaking, the search engine will output the datasets from the submitted search term.

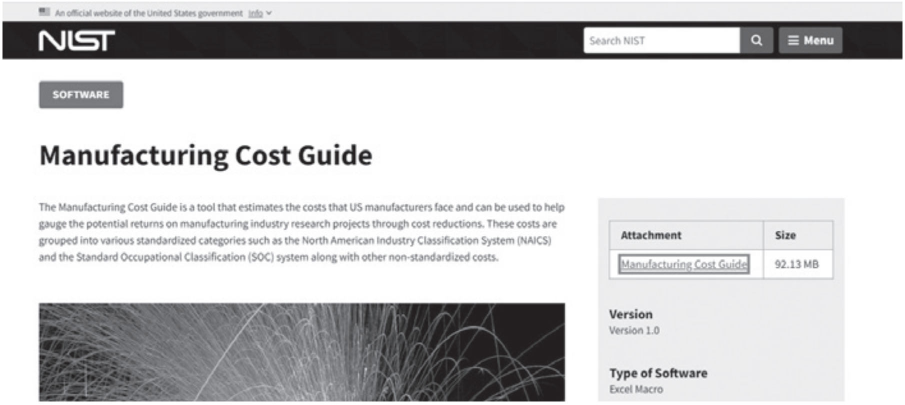
For the search term "Cost of Building Facility Projects," there are over 1100 datasets included in the search. If you focus on the first entry, you can see the "CSV" as one of the choices below the Washington State entry. This means that the data is in a comma separated value file, which is easy to import into most data analytics tools.

Let us assume that the analyst is not happy with this result, so he decides to change the search to "Cost of Constructing a Building" in the search box and get results that seem to fit the requirements a little better. The number of datasets is still above 1100, but the datasets are more refined. Remember that the better the requirements are delineated, the better the results from the search. It is all about the way the request is made to get the best result possible.

What about the systems engineer? There are datasets available for every professional at *www.data.gov*. For example, if the systems engineer wants to know the potential cost for a new advanced movement detection system, the search begins with "Cost of Advanced Movement Detection System," which produces over 1300 datasets.

The results here include a complete manufacturing guide that gives data on manufacturing certain categories of items (also found here *www.nist.gov/services-resources/software/manufacturing-cost-guide*).

When clicking on the "HTML" link on the output page, another page appears showing the NIST page (Figure 10.1). Note that there is a download of data in Excel that includes a macro, which probably helps to change certain parameters so that different scenarios are shown. The analyst now has some data to at least do a preliminary study of the system in question.



Manufacturing Cost Guide

The Manufacturing Cost Guide is a tool that estimates the costs that US manufacturers face and can be used to help gauge the potential returns on manufacturing industry research projects through cost reductions. These costs are grouped into various standardized categories such as the North American Industry Classification System (NAICS) and the Standard Occupational Classification (SOC) system along with other non-standardized costs.

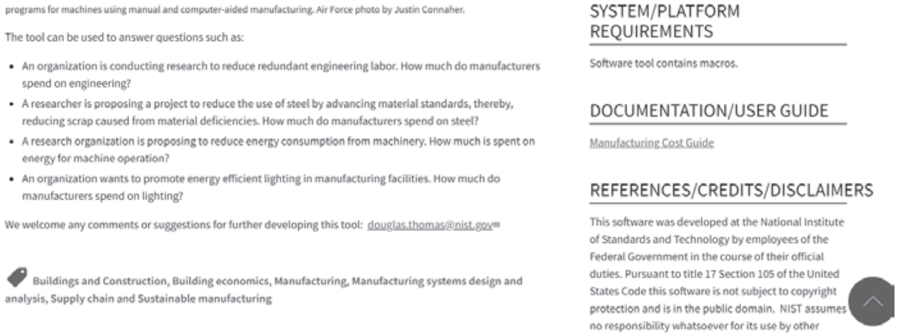
Attachment	Size
Manufacturing Cost Guide	92.13 MB

Version
Version 1.0

Type of Software
Excel Macro

FIGURE 10.1 NIST's Manufacturing Cost Guide

The most interesting thing about the NIST website is that further down the page, there is a list of questions that the data should answer. Remember the questions that needed to be asked? Well, here they are for the reading! This makes it easier for the systems engineer to see if this is, in fact, the data for which they have been searching. The questions are shown in Figure 10.2.



programs for machines using manual and computer-aided manufacturing. Air Force photo by Justin Connaher.

The tool can be used to answer questions such as:

- An organization is conducting research to reduce redundant engineering labor. How much do manufacturers spend on engineering?
- A researcher is proposing a project to reduce the use of steel by advancing material standards, thereby, reducing scrap caused from material deficiencies. How much do manufacturers spend on steel?
- A research organization is proposing to reduce energy consumption from machinery. How much is spent on energy for machine operation?
- An organization wants to promote energy efficient lighting in manufacturing facilities. How much do manufacturers spend on lighting?

We welcome any comments or suggestions for further developing this tool: douglas.thomas@nist.gov

Buildings and Construction, Building economics, Manufacturing, Manufacturing systems design and analysis, Supply chain and Sustainable manufacturing

SYSTEM/PLATFORM REQUIREMENTS

Software tool contains macros.

DOCUMENTATION/USER GUIDE

Manufacturing Cost Guide

REFERENCES/CREDITS/DISCLAIMERS

This software was developed at the National Institute of Standards and Technology by employees of the Federal Government in the course of their official duties. Pursuant to title 17 Section 105 of the United States Code this software is not subject to copyright protection and is in the public domain. NIST assumes no responsibility whatsoever for its use by other

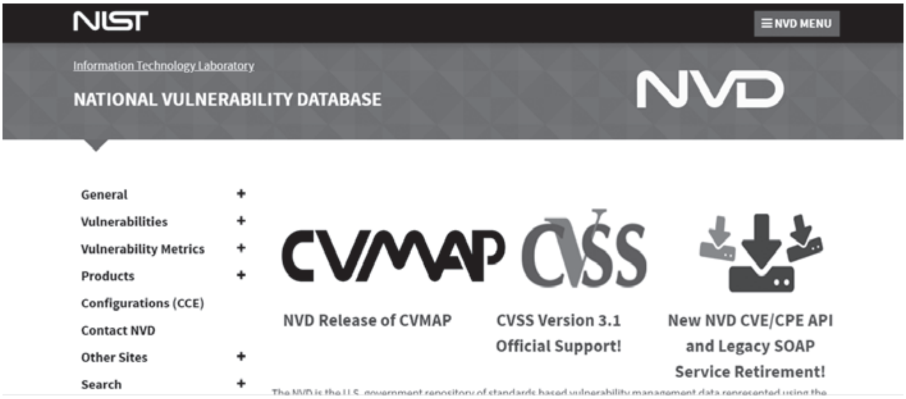
FIGURE 10.2 NIST website showing critical questions

The page with the list of critical questions has the potential to eliminate hours of searching and provides a tool to gauge the cost of manufacturing for cost estimations, as discussed earlier. By providing a macro-level Excel spreadsheet, the work hour cost of a simulation is eliminated.

The spreadsheet referred to in this section is very specific as to the types of costs necessary for producing certain commodities. The main concern would be having a macro-level Excel spreadsheet activated on your computer. If the analyst's computer has an active anti-virus program, and connection to the internet is off, then this would be a safer way of opening and activating the worksheet. Which brings us to the last profession that needs to gather data – the cybersecurity analyst.

Does *www.data.gov* have any data that could help the cybersecurity analyst? The site does have some datasets that pertain directly to the cybersecurity analyst (specifically, one that links to the National Vulnerability Database (NVD)). Simply put "cybersecurity vulnerability" into the search box and a number of datasets will appear.

The NVD website shows the results of vulnerability scoring on a variety of computer software. This site helps the analyst get a normalization on vulnerability to score their own software or infrastructure.



The NVD is the U.S. government repository of standards based vulnerability management data represented using the

FIGURE 10.3 National Vulnerability Database (NVD) site from NIST

The search for data does not have to be limited to that found at government sites, but can expand to other areas that possess data repositories, such as KAGGLE (*www.kaggle.com*). KAGGLE is a wiki-like community effort that takes datasets from all parts of society and uses them as a foundation for research (and even competition). For the project manager, there are datasets on construction or costs for certain industries. For systems engineers, there are such datasets on process tracking or mission tracking. For the cybersecurity analyst, there are datasets on the vulnerabilities of certain computer functions and software.

To download these datasets, you need to create an account. You should look through the many datasets available using the search bar to obtain the most pertinent dataset for use.

What other resources are there for analysis? The analyst can usually find resources on federal websites, whether the analysis concerns disease (*www.cdc.gov*) or weather (*www.nws.gov*) or anything in between. You might be asking what disease data would help a project manager or systems engineer or even a cybersecurity analyst, and the answer to that is easy. All project teams are vulnerable to disease, even a simple cold. It would benefit the project manager to know what the probability that people of a certain gender and age in a certain location can get sick. That would mean the risk of illness would have the likelihood of a certain percentage, and the impact would be the timing on the project schedule. In other words, data that was previously thought to be useless is now within reach and useful to the project manager, systems engineer, or the cybersecurity analyst.

10.2 How Much Data?

Once we understand where to get the data, we need to know how much data is necessary to satisfy the requirement. In some cases, the data may be compact, consisting of enough rows and columns (records and fields in "data" speak) that the analyst can use any tool to analyze the dataset. However, in other instances, the data may be too voluminous to use with some tools, so the analyst has to sample the data to get a smaller dataset from the larger one. Now that the word "sample" has been used, a refresher on that term is necessary to continue this concept.

10.3 Sampling

Sampling is a way of obtaining analytical results that will mirror the results from a larger dataset. Some people refer to sampling as "random sampling," even though the sampling is anything but random. *Random sampling* means that every event is considered in the selection of the sample. Sampling is both necessary to work with a large dataset and important because, without it, analysts would be faced with countless hours of trying to work with a dataset that is too large for their needs.

There is a story about a person who refuses to believe that sampling actually works. A data analyst tells the person that, if that is the case, the next time the person goes to the doctor to have their blood drawn, the doctor should take it all! Random sampling is done in many parts of our lives; it is just that we do not perceive it as random sampling. If one thinks about it, a blood sample is something that mirrors our entire blood supply. When a blood test says that a person's cholesterol is 148, that means that it reflects the person's blood content – period. That is why there are so many different ways of sampling, but for the purpose of this book, the main sampling topics are random sampling and systematic sampling.

10.3.1 Random Sampling

The simplest illustrations of random sampling are the "hat draw" and the "bingo draw." The hat draw involves putting slips of paper into a hat and then have someone draw from the hat. The true meaning of random is evident in this example because all the pieces of paper have an equal chance of being chosen; the one that is chosen is truly chosen by chance. In the bingo draw, different enumerated balls are placed in a cage, the cage is spun, and a numbered ball is drawn out. The chance of any one number being drawn is no different than any of the others. One might add the lottery draw, which uses forced air to choose the different numbers, leaves human choice out of the equation. So, whether you are drawing to see who goes first in a game, playing bingo, or playing the lottery, the random sampling method is visible in those specific situations.

What about sampling televisions to test them for endurance or suitability? What about sampling military personnel for drug urinalysis? In every situation, the analyst can input a random approach to sampling to ensure that random picks are being conducted. In the television example, each television has a serial number, and those serial numbers can be placed in a spreadsheet or database and then placed into a tool that does random sampling. Which tools do this? Excel, KNIME, R, and others have that ability to choose a random sample from a larger dataset. The specifics for performing the random sampling for R, KNIME, and Excel is in *Data Science Tools* (from Mercury Learning) or many other books that are available on the market for each of these tools.

Random sampling is random when every event is equally considered. If that is not the case, for whatever reason, then the sampling is not random. What happens if the sampling is not random? There is one event that illustrates the bias that comes with a non-random sampling; the 1948 Truman-Dewey election. The newspaper came up with a biased result that led to a very embarrassing situation. The interpretation of the question of whether Truman or Dewey won the election was the problem. Please refer back to the "Type" errors discussed earlier and see that this was a Type IV error, basically the wrong answer to the right question. In this situation, the data was correct, but biased based on the sampling. The result showed this bias.

10.3.2 Systematic Sampling

Systematic sampling can be done either in conjunction with random or non-random methods. *Systematic sampling* takes a number derived from a formula (that is not be discussed in this book) and derives the sampling from that number. For instance, if the analyst decides to take every fifth event, then 5 is the number for the systematic sampling.

The main problem with this method is that not every event is considered evenly. If the systematic sampling calculation comes out a 6, then every sixth element is considered, leaving the first through fifth elements as not considered. One can see that, if not carefully evaluated, the systematic sampling could be biased.

10.3.3 Sampling Bias

The Truman vs. Dewey race is the epitome of what bias sampling can do, but how does the analyst detect data bias, and what can they do to mitigate this risk? Data bias can come down to a very simple question: "Where did the data originate?"

In data collection and gathering, the discussion has been focused on where to get the data and how much data to get, but this section focuses on "who" is providing the data. In many datasets, there is a separate section on where the data originated. During the COVID-19 pandemic, the main source of data was the Johns Hopkins University dataset (*www.jhu.edu*). In the dashboard on the Johns Hopkins University site, there is an explanation of where the data originated. The Centers for Disease Control (*www.cdc.gov*) also has data on COVID-19 and explains where that data originated in their dataset proper. The main thing to consider whenever an analyst wants to verify the data is to check the data source.

One example in the COVID-19 is the culmination of fatalities that were caused by COVID-19. Where did that data originate? Does the data signify who died ***from*** COVID-19 or who died ***with*** COVID-19? If there is no distinction, should there be or does that make the data contrived and unverifiable? The analyst must evaluate the origination of the data in order to adequately qualify the findings. Even in clinical studies, the association of the authors to the drug company that is sponsoring the study is revealed in the document. This is incredibly important because it tells the data analyst that the results might be biased or at the very least there might be a perceived conflict of interest. It also allows the analyst to note that in their results as part of the transparency of the analysis.

Another example of bias is simply not taking into consideration the change in technology that might accompany a longitudinal study. If the analyst goes to the National Weather Service and evaluates the tornado dataset, they will find that there were approximately 200 tornados in 1951, yet in 1971, this number grew almost 5 times as many. Does that mean that the number of tornadoes grew because of climate change? Or could it be that in 1951, there was no man made satellites, and in 1971, there were actual weather satellites? These are the types of bias that can exist without really understanding the underlying reasons for that bias.

10.3.3.1 Mitigating Data Bias

What are some of the ways that an analyst can reduce this data bias? First, ensure that the sampling is truly random to allow for all events to be considered fairly. Second, compare "apples with apples" using the available data to "normalize" the data. The term "normalize" means that all variables are considered against standard measures. An example would be trying to compare the price of an item in dollars to the same item priced in yen. In this case, there is a wide disparity between the items being compared. The analyst would have to convert the yen to dollars or the dollars to yen to make the comparison standard. This is where "standard" deviation comes into play. No matter how different the datasets seem, the standard deviation, with the mean and placed in the normal curve, can standardize the results of the analysis.

There are several methods to normalize two different datasets or variables. The first one is weighting the different variables to normalize the values. For instance, with the tornado dataset, possibly taking the number of tornadoes and using the amount per 100,000 population might help normalize the data. There is no doubt that there were fewer than 320,000,000 people in the United States in 1951 (in fact, there were less than 170,000,000, according to one reference[20]). To normalize the datasets between 1951 and 2019, the analyst takes 170,000,000 and divides it

[20] *https://www.populationpyramid.net/united-states-of-america/1951/*

by 320,000,000 to get a result of .53 (or the population in 1951 was 53% of the population in 2019). Once this is done, then multiply the number of tornadoes in 2019 by .53 to help normalize the tornado per population ratio; or conversely, multiply the tornadoes in 1951 by 1.88 (taking 320 million divided by 170 million). Either way, the analyst needs to adjust the overall tornado occurrence to equalize the two years by weighting the population. The second way to perform this normalization is to do the same thing with the population, but use the 10 states where tornadoes occur the most and use that in the weighting.

10.3.4 Determinism

Coming off the heels of sampling is the idea of determinism. Data analysts want to associate one variable with another. This can sometimes be a trap, and this book addresses some of the more conventional techniques for association, such as correlation, but there are some others that are necessary to discuss, including lift, leverage, support, and strength.

10.3.4.1 Lift

Lift is the process of seeing how one action affects a result. For a more detailed study of this concept, see Provost and Fawcett's book *Data Science for Business* [Provost-2013]. The concepts presented in this book can help any professional in their daily duties. Lift is an aspect of probability where two events happen, not by chance. The example in the Provost and Fawcett book made reference to a person who went into a store to buy peanuts and beer. What is the probability that the person was buying beer because they bought peanuts (or vice versa)? Is there just a chance that they are buying them together, as a coincidence, or are they purposely buying one because of the other? Lift helps to calculate the probability of this happening. If lift results in a value greater than 1, then the event did not happen by chance. The formula is relatively straightforward:

$$\frac{P(A\&B)}{P(A) * P(B)}$$

What the formula depicts is the probability of both event A and event B happening, divided by the probability of event A multiplied by the probability of event B. As an illustration, the analyst must determine what the lift is for a tornado to occur with $250,000 in damage. If the chance of a tornado happening that causes $250,000 in damage is 20%, then the next step is determining what the probability is for a tornado and the probability of damage of $250,000. If the probability of a tornado happening is 30% and the probability of $250,000 is 5%, then the formula is as follows:

$$\frac{.20}{.30 * .05}$$

The result of this formula is 13.3. This is well above the value of 1 that would signify that this occurrence is well beyond chance. For a more graphic demonstration, creating a lift chart is a much better way of determining the overall difference between the probabilities of the data and chance. A lift chart is demonstrated in a number of software applications; one in particular is KNIME, which is illustrated in the Mercury Learning and Information book, *Data Science Tools*.

10.3.4.2 Leverage

Leverage is another way to determine associations in data that is unbiased. Again, it is important to understand that biased data will not reveal the same associations as would good, unbiased data. The formula for leverage is focused on the differences between probabilities rather than the ratio of the lift formula.

$$P(A, B) = P(B\&A) - P(A) * P(B)$$

Given the same probabilities as used in the lift example, substituting the numbers for the symbols produces the following equation:

$$.20 - (.30 * .05)$$

The result is .185, which means that there is an 18.5% increased probability that one variable occurs with the other, regardless of chance.

Leverage is another method to discover associations, without going through the correlation drill, but all this is based on the data being sampled without bias.

10.3.4.3 Support

The formula for the support is probably one of the most straightforward deterministic formulas presented in this book. *Support* is just the prevalence of two variables occurring together. In our tornado example, it is a tornado with $250,000 of damage, which yields a probability of 20%.

And that is it for support! As written, this is just the probability of two events happening together and is composed of a litany of techniques to determine the association of those two events.

10.3.4.4 Strength

Strength uses conditional probability, and is the probability of something happening *given* that something has already happened. In this case, it is that there was a tornado that did $250,000 of damage. Written as a formula, this is as follows:

$$P(A|B) = \frac{P(A\&B)}{P(B)}$$

What this means is that the probability of A given that B has occurred is equal to the probability of A occurring with B, divided by the probability of B happening alone. To translate this into numbers, the probability of A is .30 (the probability of a tornado), the probability of B is .05 (the probability of $250,000 of damage), and the probability of A and B is .20 (the probability of a tornado causing $250,000 of damage). Placing these numbers in the appropriate places in the formula results in the formula below:

$$P(A|B) = \frac{.20}{.05}$$

The result is 4.0, which means that it is well above 1.0 or 100%. What this means is the strength of the relationship between the tornado and the damage is extremely high. This actually makes sense, since a tornado always causes some damage, no matter how small the tornado.

With all these techniques to try to find the association, it should be evident that there are many ways to perform analysis other than the conventional ones taught in a statistics class. The whole reason for this book is to give the analyst more tools to perform a true analysis of the data. However, once analyzed, it is essential to display the data so that your audience can understand and possibly make decisions with the data. That is the topic of the next section.

CHAPTER 11

DATA PRESENTATION

The title of this chapter could be *Data Visualization*, but the true purpose for any visualization is the presentation of the data. Before delving into the many possibilities of data presentation, it is essential that one point be made clear. There is no more important stakeholder in a data presentation than the audience. Why would presenters give charts and graphs that were entertaining but meant nothing to the audience (oftentimes their boss)? The objective of presenting the data is to ensure the audience can understand the data. If the audience does not understand the data, then more questions will arise, usually ones that show some frustration about the content of the presentation.

In other words – KNOW THE AUDIENCE! If you do not understand the level of data science education (and concern) of the audience, then you are in for a very long presentation, trying to answer questions that would have never been raised had you considered the audience. This is whole purpose of this section, with plenty of illustrations.

11.1 The Good, The Bad, and The OMG

There are many examples of poor data presentation. Figure 11.1 provides an example of poor data presentation: it's impossible to understand what this map shows. Guessing will not help, because there is no reference point, no title, and no explanation of the colors.

Figure 11.2 shows a range for the data, but does not provide any more useful information. Figure 11.3 shows a heat map. The map uses shades of gray, which is better than using color because some people in the audience may have color blindness. While color can be used effectively, the solution for audience members with color blindness is to use a shades of gray or draw hatch marks to differentiate the values.

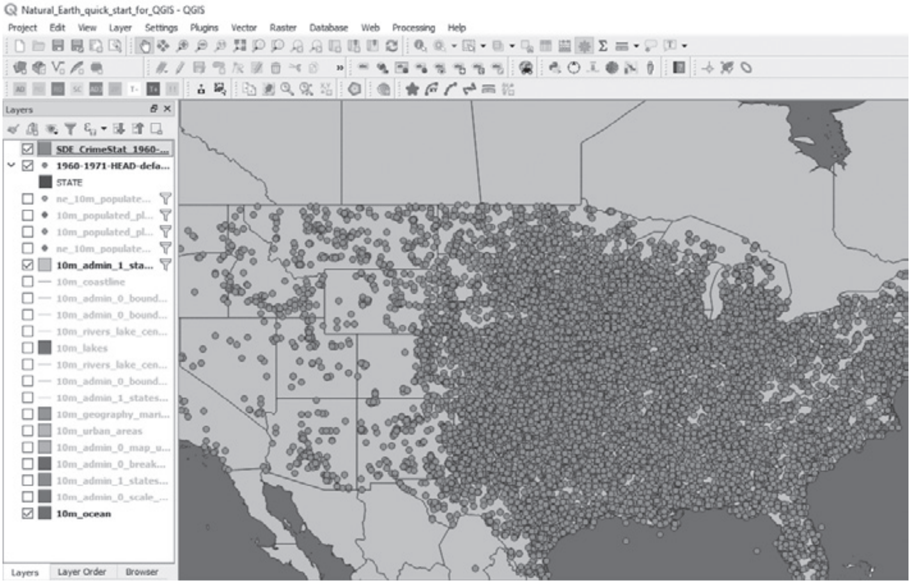


FIGURE 11.1 A poorly-designed QGIS map

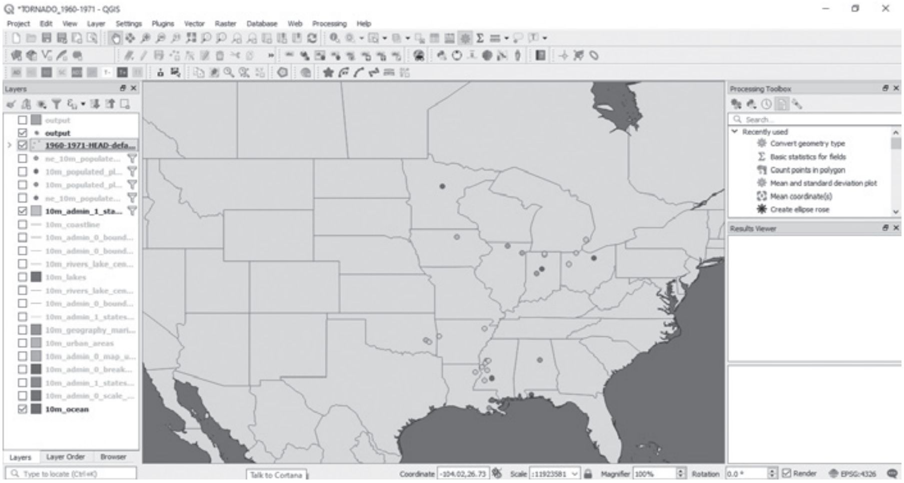
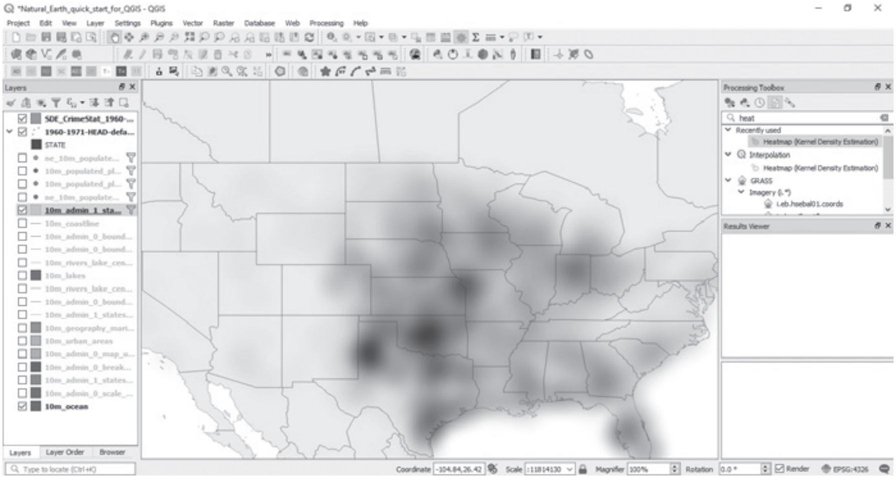


FIGURE 11.2 QGIS map with a specific range of data, but no legend



Layers:

- ☑ SDE_CrimeStat_1960-...
- ☑ 1960-1971-HEAD-defa...
  - STATE
- ☐ ne_10m_populate...
- ☐ 10m_populated_pl...
- ☐ 10m_populated_pl...
- ☐ ne_10m_populate...
- ☑ 10m_admin_1_sta...
- ☐ 10m_coastline
- ☐ 10m_admin_0_bound...
- ☐ 10m_admin_0_bound...
- ☐ 10m_rivers_lake_cen...
- ☐ 10m_lakes
- ☐ 10m_rivers_lake_cen...
- ☐ 10m_admin_0_bound...
- ☐ 10m_admin_1_states...
- ☐ 10m_geography_mari...
- ☐ 10m_urban_areas
- ☐ 10m_admin_0_map_u...
- ☐ 10m_admin_0_break...
- ☐ 10m_admin_1_states...
- ☐ 10m_admin_0_scale_...
- ☑ 10m_ocean

Processing Toolbox search: heat

- Recently used
  - Heatmap (Kernel Density Estimation)
- Interpolation
  - Heatmap (Kernel Density Estimation)
- GRASS
  - Imagery (i.*)
    - i.eb.hsebal01.coords

Coordinate -104.84,26.42 Scale 1:11814130 Magnifier 100% Rotation 0.0 ° ☑ Render EPSG:4326

FIGURE 11.3 Example of a heat map in QGIS using shades of gray

Figure 11.4 is a pie chart showing a portion of the same data used in the maps. Although a map would probably be better for this content, you could use other charts to show the information in different ways. The pie chart should be used in conjunction with the map for more granularity.

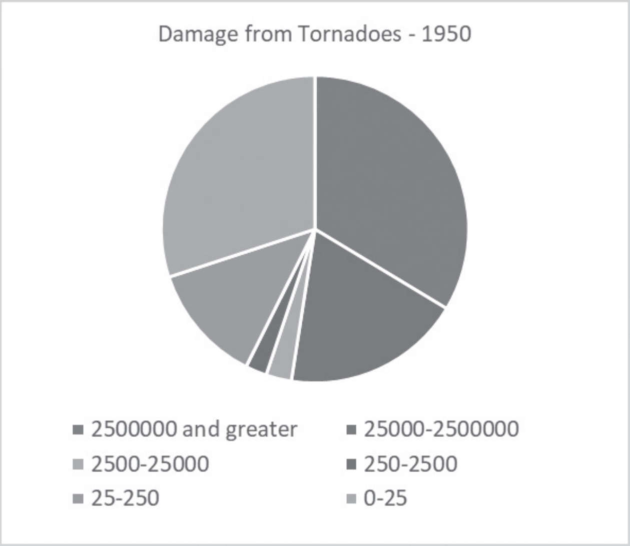


FIGURE 11.4 Pie chart with tornado damage data

The main issue with the pie chart in Figure 11.4 is that there are no units for the numbers, so the audience does not know if the numbers represent money, people, or buildings. At least one author has advised against the use of pie charts [Knaflic-2015]. However, by clearly labeling the main sections of the pie chart with the numbers, this chart can be more useful. The chart should also be enlarged to accommodate those numbers, and should only include the data that is relevant to the audience.

11.2 Real-World Example from a Project Management Perspective

Project managers, systems engineers, and cybersecurity analysts must deliver numerous brief presentations, both technical and non-technical. This means that the visuals for the presentations need to be succinct.

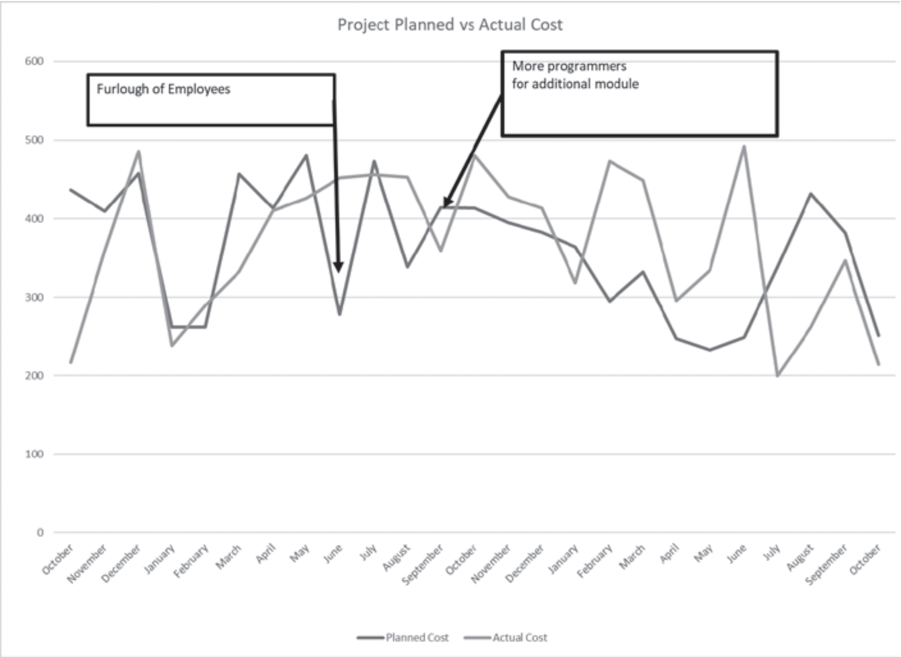


FIGURE 11.5 Example of a chart describing project costs

Let's consider the following example. A one-slide presentation was needed to discuss a project. The favorite tool of the project manager was *Earned Value Management* (EVM) software, which catalogued all the costs for the project in the form of the EVM template. Basically, this consisted of Actual Costs, Planned Costs, Actual Schedule, Planned Schedule, and then a cumulative look at these measures. It is important to realize that just because the data is there does *not* mean that the analyst has to use it all. The main elements used for this presentation were the Actual Costs and Planned Costs, by month. The finished graph did not use the cumulative perspective of the project, but showed the month-to-month progress. The graph showed the line chart of the costs in different colors (including different types of lines to allow for audience members who were color blind), with small square "balloons" that captured what happened at the junction of the planned and actual costs. The presentation of this data prompted the project manager to remember what happened in conjunction with the Actual Cost being less than the Planned Cost or vice versa. Figure 11.5 shows this type of chart, which was generated from random numbers.

The audience enjoyed the presentation because the project manager was able to narrate using the slide as a prompt and the stakeholder was able to understand the slide. The stakeholder was able to take that chart and explain it to other stakeholders without any further involvement from the presenter. This is called the *trifecta of good presentation*, where the presentation benefits the presenter, the audience, and the stakeholder.

Not all presentations allow for this type of template, but if you look at the data and really think carefully about what the stakeholder is interested in knowing, you will have many more successful presentations.

CHAPTER 12

GEOSPATIAL DATA ANALYTICS

The next few sections highlight the use of geospatial information systems (GIS) to promote geospatial data analytics. The formulas for calculating the geospatial statistics are discussed in the same way as the formulas were discussed for non-geospatial statistics. The topics are separated into sections on mean, median, and standard deviation, but focus on the geospatial formulas. Additional references for these formulas are included in the reference section, but you should read the books *Elementary Statistics for Geographers* and *The Esri Guide to GIS Analysis, Volume 2*. These books are excellent references for those who want to expand their overall knowledge of geospatial analysis and learn how to apply it. Each section here has real-life examples of how this type of has been used in project management, systems engineering, and cybersecurity.

12.1 Geospatial Mean Center

The main focus of the section on the mean was on summing the number of values and then dividing that by the count of the values. The same thing is true of the *mean center* in geospatial analysis. However, geographic coordinates are composed of two different values corresponding to the latitude (lat) and longitude (long). The longitude refers to the imaginary lines that go from north to south on the globe and measure how far east or west a place is in regards to the Prime Meridian. Since the globe is oblong (more of an egg shape than circular), these lines are longer than those of the latitude, which means that the line is “long” (“longitude”). Another way to remember longitude is to associate it with the term from statistics *longitudinal,* which means "time based." The Earth's longitude is what time is based on, specifically, the time zones. Therefore, if you remember "longitudinal" is based on the long lines going from north to south as in time zones, then remembering the longitude should not be a problem.

The calculation of the geospatial mean center is basically the same as the mean, but it is done in two steps. The first step is calculating the mean of the latitude, which is the "x" value, by summing the x values and dividing that value by the count of the x values. The second step is the same for the "y" value or longitude. In this way, the result of the calculation is the xy coordinate, which give a location [Mitchell-2009, p. 33]. Now, what this calculation produces is a location that could be skewed by the values that precede it. Being geospatial does not mean that it is impervious to skewing, because as a mean calculation, it is affected by the values that compose it. If there are more easterly values, then those values drive the geospatial mean center to the right; more westerly values drive it to the left. The very nature of the mean is such that it incorporates the same limitations as those for any derivation of that mean.

The geospatial mean center incorporates both the latitude and longitude, along with any height (called the "z" value), which means that the analyst can also add the height of the terrain and average that along with the xy value.

The formula for these calculations is the same as presented in the mean section at the beginning of this text, with the substitution of y and z for the x values in the formula. If there are many coordinates (whether they be x, y, or z) that are the same, then you can perform a "weighted" geospatial mean center. The only difference is that the weight of the coordinates is multiplied by the coordinate and then added to other values to be divided by the sum of the weights of the coordinates. For example, if there are three instances of the x value being 32.339, and 1 instance of 42.223, then the resulting formula for this x-value is as follows:

$$\frac{(3 * 32.339) + (1 * 42.223)}{4}$$

This formula reflects the original mean formula, but instead of counting each value separately, it counts the same values as a weighted sum. The calculation and result are the same if each of these values is listed separately.

What does this mean from a project management, systems engineering, or cybersecurity perspective? Project managers often tend to see things from a project perspective, but not from a project team perspective. The same applies to the systems engineer or the cybersecurity professional. The geospatial analysis gives the analyst the ability to "see" from the perspective of the project team. For example, if a stakeholder wants the server farm to be placed in the center of the department, then this method comes in very handy. If the department is split into several areas or rooms by grids, then the grids act as a longitude and latitude where this formula can show the mean center of the grid, and therefore the center of the different departments. For the systems engineer, location can be in the form of the different process points, especially if those process points are in different parts of the room. If the functionality and efficiency of the system are based on the proximity of sequential steps, then this method is relatively valuable. In the cybersecurity area, the geospatial mean can help find a possible hacker by distance from any wireless point where attack occurred. Again, a grid is established and, by taking the geospatial mean, the analyst can locate the center of the wireless intercepts based on the optimum distance from the wireless access point.

12.1.1 Real-World Application of Geospatial Mean

A project or system may be related to a problem that needs to be solved or something that needs to be improved. Tracking criminal activity is one of those hobbies that a data analyst can conduct with the data being available through open source. At times, you can start the analysis with just the news reports that give the address of the robbery. Years ago, there was a bank robber in Baltimore called the "Bandage Bandit" because he would wear a bandage while robbing a bank. After getting the addresses of the banks robbed and using Google to associate latitudes and longitudes with these addresses (*georeferencing*), the dots were plotted using QGIS and several types of analysis were completed on the geospatial points. Figure 12.1 shows one of the slides submitted to the FBI after the study was completed. The reason for finding the mean center was to try and locate where the robber lived, since many criminals commit crimes in their immediate vicinity for work, home, or entertainment.[21] Finding the mean center in this case might lead to either the workplace, home, or area where the criminal spent

[21] *https://web.archive.org/web/20150518074711/http://www.ceamos.cl/ceamos/images/stories/actividadesyeventos/pattern_theory1.pdf*)

his leisure time. This is just one use of the geospatial mean in analysis projects. Figure 12.2 shows the same information as Figure 12.1, but includes streets and other information. The name of the town where the robberies took place has been removed for anonymity.

The geospatial mean can be a part of data analysis whether the project manager, systems engineer, or cybersecurity professional is measuring the actual geographic location or a grid system that can be translated into x and y coordinates. Professionals should be creative in their use of geospatial analysis.

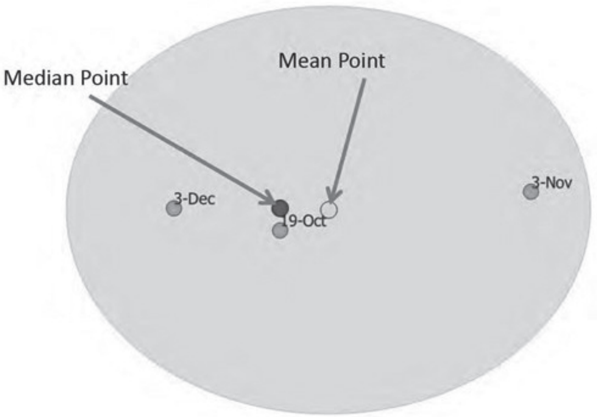


FIGURE 12.1 Map showing the mean and median center

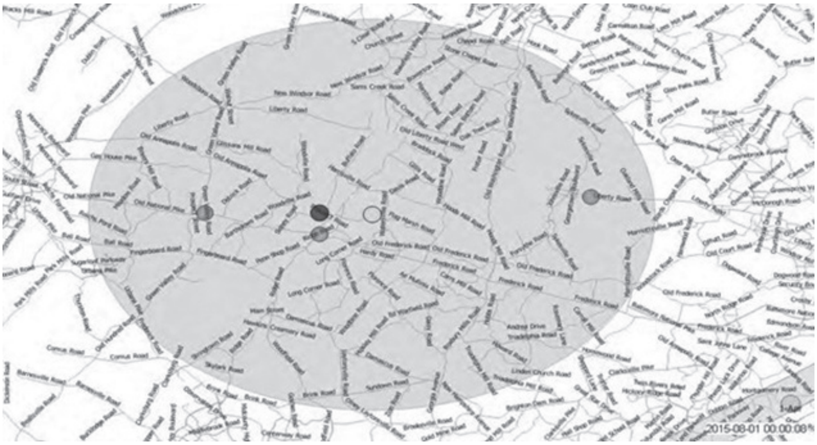


FIGURE 12.2 Map with street names

12.2 Standard Distance

The *standard distance* is a circle that is formed by the distance calculated between the x and y sum of squares calculation. Although this seems complex on the surface, it is basically taking the square root of the additions of the squared difference between the x coordinates and x-mean and then adding them to the squared difference between the y coordinates and the y-mean. The equation is as follows:

$$\sqrt{\frac{\sum (X-\overline{X})^2}{n}+\frac{\sum (Y-\overline{Y})^2}{n}}$$

The result is an area of extent for the coordinates that are either on a map or grid. You can obtain an area of interest for the project or system to help estimate possible costs that would be incurred on that project. For instance, if the project is a building and outlying infrastructure, then this formula would help get a quick extent of the grounds needed for the overall project as an estimate. The standard distance is in the form of a circle, but the standard deviational ellipse has the resulting form in the name – ellipse.

12.3 Standard Deviational Ellipse

The *standard deviational ellipse* is a formula to determine the length and breadth of the distance that the coordinates provide, whether those coordinates are on a map or part of a grid system. The following formula is given as two parts, the x part and the y part, so that the ellipse provided reflects the distance provided by the x and the y coordinates.

$$\sqrt{\frac{\sum (X-\overline{X})^2}{n}}$$

$$\sqrt{\frac{\sum (Y-\overline{Y})^2}{n}}$$

Figure 12.3 shows the ellipse created using part of the information from the Bandage Bandit study. The ellipse is extremely large based on the coordinates. Please remember that the standard deviational ellipse is based on the mean, which can be skewed based on outliers. Few points on the map or grid can lead to an extraordinarily large ellipse.

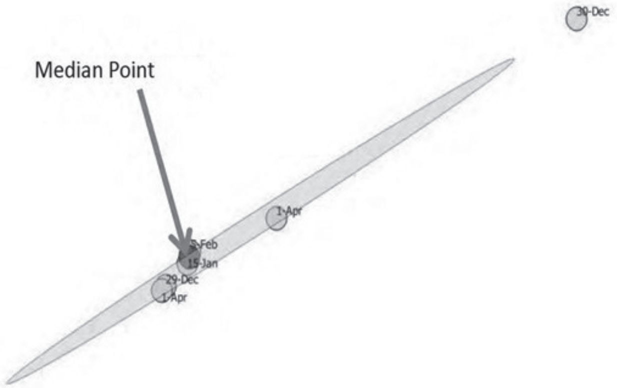


FIGURE 12.3 Example of a standard deviational ellipse

This formula not only provides a circle of the events that occurred, but provides a direction that those values appear to be favoring. This is important if the professional wants to determine the direction of something, whether it be weather, disease, or some other function. Think about the uses for this other than crime statistics. What about team member locations or process steps? What about IP addresses or wireless connection points? All these things can be considered when applying these formulas to xy coordinate planes. Many analysts can look at a scatter chart and make sense out of the data, but some have challenges when that same xy data is placed on a map or grid. With practice, the use of geospatial analysis will be as easy as using statistical tools from the conventional tool chest.

There is one more measure that we will cover, called the *Geary's C*, which can determine if values are clustered or random. The formula is included in this book because, like forecasting or prediction, the pattern finding of geospatial analysis is vital to help determine next steps or make decisions.

12.4 Geary's C

There are many pattern-finding formulas in geospatial analysis, but this is a relatively straightforward one to implement. The formula is as follows:

$$\frac{n \sum\sum w_{ij}(X_i - X_j)^2}{2 \sum\sum w \sum(X_i - \overline{X})^2}$$

The numerator takes the number of values (n) and multiplies that number times the sum of the weights of the values of the sector (area of the value) and the neighboring sector. This is then multiplied by the difference between the two values, specifically between the value of the one sector and the value of another sector. This is then divided by two times the sum of the weights of the values multiplied by the sum of the squares of the values minus the mean of the values.

This formula has nothing to do with the coordinates, but the values that are located within each sector that is being studied. This can be hard to understand, but think of sectors as things like counties or states or regions. For example, if you must find out if there is a cluster of COVID-19 patients in the surrounding counties to a county that has an outbreak, this formula would be useful.

For example, let's say we want to know if there is a cluster between three regions, one containing nine COVID cases, one containing three, and one containing five. The formula uses the target, which in this case is three, and the neighbors are nine and five.

Each neighbor is subtracted from the target:

3 – 9 = – 6

3 – 5 = – 2

Each of those results is squared:

– 6 squared = 36

– 2 squared = 4

These are multiplied by the weight of each of these values, which in this case is 1, but could be any weight assigned to the values (for instance, if there is more emphasis on certain regions, those weights could be increased). The sum of the previous is 40 multiplied by 1, which is 40. That result is then multiplied by the number of features, 3, which makes the numerator 120. The denominator is a little different with the sum of the squares for all the features:

9 – 5.7 = 3.3

3 – 5.7 = –2.7

5 – 5.7 = –.7

This result is squared and summed to give the following:

3.3 squared = 10.89

−2.7 squared = 7.29

−.7 squared = .49

10.89 + 7.29 + .49 = 18.67

The next step is to multiply this result by 2, which is 37.34.

The finished equation, with the substitutions of the values for the numerator and the denominator, is as follows:

$$\frac{120}{37.34}$$

The answer is 3.2. A value of greater than 1 denotes dispersion, which means that these are not clustered. This is important because the result shows that one region does not affect the other, at least with this measure [Mitchell-2009, p. 124].

What is the value of this measure to professionals? This formula helps to determine patterns so cybersecurity specialists can find vulnerabilities in data stores or even problems with capacity management. In the project management and systems engineering areas, the measures of cluster patterns could be an indicator of either issues or risks in both the project and process areas. In fact, clustering could be a factor in risk analysis. For example, if the project is in an area that is plagued by severe weather, the analyst could determine the clustering effect of that weather on neighboring regions or areas. Think about the decision made by a project manager to move a project team to a nearby area only to realize that it has a clustering effect with the area they just left!

CHAPTER 13

ADDITIONAL DATA ANALYTIC METHODS

Some data analytic texts utilize the regression and correlation methods to discuss ways to analyze data, but we are going to discuss some methods that are relatively rare in data analytic texts. There is also a section here on the effect size, but the term effect size is not used with the method. Finally, there is a very brief section on modeling, more as a refresher than introducing new material. Modeling is the foundation for much of the data analytics in public service.

13.1 Entropy

Entropy is a relatively new method in the data analytics community. In fact, one PhD in logistics who had a background in statistics said he had never heard of entropy being used in analytics.

Entropy, however, is a very old concept in the information world and was first identified by Claude Shannon in 1948 [Shannon-1948]. Shannon proved that as information becomes more diverse, the security of that information is strengthened. The paper later led to a concept known as *Shannon's Entropy*, which is still used today to determine the strength of digital security, including passwords [Shannon-1948]. This section addresses the concept of entropy as it relates to data and to data analytics. For those who want a deeper discussion on this issue, please see Provost and Fawcett's book in the reference section [Provost-2013].

The formula for entropy is as follows:

$$-[p_1 log_2(p_1) + p_2 log_2(p_2) + \cdots]$$

This formula involves taking the probability of one variable multiplied by the log of the probability of that same variable added to the probability of the second variable times the log of the probability of that same variable, and on until the variables are exhausted.

For example, let us say you want to know if the length of a tornado in 1951 was different *enough* from the width of the tornado in that same year. In this example, the probability of a length of a tornado being more than 5 miles is .20, while the width of a tornado of 5 miles is .10. In this case, .20 is placed where p_1 exists, and .10 is placed where p_2 resides. The following is the result of this substitution.

$$-[.20\ log_2(.20) + .10\ log_2(.10)]$$

The negative sign at the beginning of the formula is needed because decimals in logarithmic numbers are negative, so the result of the formula above is .7966. This means that the entropy is sufficiently high so they are not related. In other words, this is a way of checking for collinearity as a secondary method of evaluation. Entropy can be used for *Information Gain*, as explained by Provost and Fawcett [3, p. 53]. Information Gain (IG) entails taking the entropy and adding a "child" entropy that would show whether the variable in the child calculation affords any information gain to the overall data comparison between the *parent* variables. The formula for the IG is provided below [Provost-2013, p. 53]:

$$Entropy\ (parent) - [\,p(c_1) * entropy(c_1) + p(c_2)\ ...]$$

Let's continue with the example above. The entropy of the parent is .7966. We then add in the probability of a tornado doing over $250,000 of damage in 1951 (.30). We place these values in the formula [Provost-2013, p. 53] to obtain the following:

$$.7966 - [.30 * (.30 * log(.30))]$$

The result of this formula is .9050, which means that, given the probabilities in this example, the additional variable of the property damage above $250,000 would add to the information from the other two variables of the length and width of the tornado.

The Internet includes websites that have further explanations of these concepts.[22] However, Provost gives some detailed explanations of the different probabilities.

13.2 Effect Size, Part 2

The effect size is not commonly addressed in undergraduate statistics classes, but it is very useful. The most visible effect size is the *correlation coefficient*, which determines the strength of the relationship between two or more variables. Because correlation is addressed earlier in this text, a reiteration of this topic is not necessary. However, the effect size can be used to evaluate a test or model so the result can be as valid as possible. If the correlation results in .80, the analyst knows that this is a strong relationship. It also determines the effect size of the target dataset.

The other form of the effect size includes Cohen's D, and a detailed explanation can be found online (*www.statology.org/effect-size/*). *Cohen's D* is used to evaluate such tests as the t-test, which measures the differences between two datasets. The formula for Cohen's D is as follows:

$$D = \frac{Mean\ 1 - Mean\ 2}{Pooled\ Standard\ Deviations}$$

The formula for obtaining the pooled standard deviations, given the number of values in each of the datasets are equal, is below. The formula combines the standard deviations from both datasets to determine the differences between those datasets. We can use the "plug and chug" method to solve this formula, meaning you only need the number of values for each dataset and the standard deviation for each of those datasets. The rest is substituting the numbers in for the formula symbols.

$$Pooled\ SD\ (If\ Samples\ are\ Equal) = \frac{(SD_1^2 + SD_2^2)}{2}$$

If the numbers of values in each dataset are not equal, which in many situations would be the case, then you need to use another formula. (Please

[22] *https://towardsdatascience.com/entropy-and-information-gain-b738ca8abd2a*

see this site, which has more information on the formula and its rationale: *toptipbio.com/cohens-d/.*)

$$Pooled\ SD\ (If\ Samples\ are\ Unequal) = \sqrt{\frac{(n_1 - 1) * SD_1^2 + (n_2 - 1) * SD_2^2)}{n_1 + n_2 - 2}}$$

With this effect size formula, you can now evaluate the models of both the correlation and t-test, which encompass at least two of the more conventional tests that are completed on the datasets. This leads us to our final section of this text, which is on modeling.

13.3 Modeling and Simulation

Analysts can sometimes confuse modeling with simulation. *Modeling* is a representative of reality, while simulation places a wide variety of inputs into the model to see what the output is with these inputs. Modeling a dataset involves using an algorithm and having the simulation apply the algorithm.

13.3.1 Model Type

One of the most conventional models is the *Simple Linear Regression* (SLR). SLR is used more than is needed. However, there are conditions when SLR is beneficial. Those conditions are as follows:

1. When a prediction or trend is requested or required
2. When the data is collected from a reputable and verifiable source
3. When the need arises because other simpler models have been applied and are not giving the information required by the stakeholder

Do not confuse #1 and #3. Just because the trend does not show what the stakeholder expects does not mean that it is wrong. The result that the stakeholder expects may not match the model. Just hope that the stakeholder does not hold you responsible for faulty data or requirements.

This text does not address the procedure for conducting an SLR, since there are plenty of references and tools available to do that. The reason for mentioning the SLR was to ensure that you understand the limitations of this model for all data analysis. Sometimes it is necessary for other, simpler, models to prevail. For instance, sometimes a simple correlation model can be done that gives both a model and effect size to the data. In other instances, just using the Five Number summary will suffice.

When assuming that a model type is required, it is best to see which one is necessary based on the data composition. Remember our earlier discussion about understanding the data? This is where that concept is applied. In *process modeling*, you take a dependent variable (commonly called “y”) and one or many independent variables (called “x”) and relate those two (or three) through a series of visual depictions as well as probability simulations (a very detailed description is available online[23]). However, for this text, just one example is discussed to show the application of many of the concepts that were discussed earlier.

For instance, suppose that the dataset encompasses a comparison of a set of budget figures from different departments. These departments include accounting, human resources, operations, systems, and occupational health. Table 13.1 shows the budget for each. What type of visualization (model) would you use to depict this data? If a line graph came to mind, that might be a concern because these are separate departments that do not show the data in longitudinal (time based) form. Therefore, the line graph would not be a good representation of the data. In fact, the visual might be misinterpreted as a time changing value when it is really one value for each category.

Table 13.1 Data for visualization

Department	Accounting	HR	Systems	Operations
Budget ($)	130000	250000	500000	750000

The first visual to model the data in many instances is a bar graph or histogram. What is the difference? Every histogram is a bar graph, but not every bar graph is a histogram. A *histogram* usually shows ranges while bar graphs show singular values. Figure 13.1 shows an example of a bar graph based on the data from the example.

The process modeling is a combination of both the x and y component, but it also includes a probability component. For instance, if the budget numbers above are presented without any way to compare the budget per department, this could be misleading. What if the budget is given per person in each department? Or if the departments are compared with respect to each other in population? This could then present the visual with a whole new perspective. Table 13.2 depicts the population of the different departments.

[23] *https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd11.htm*

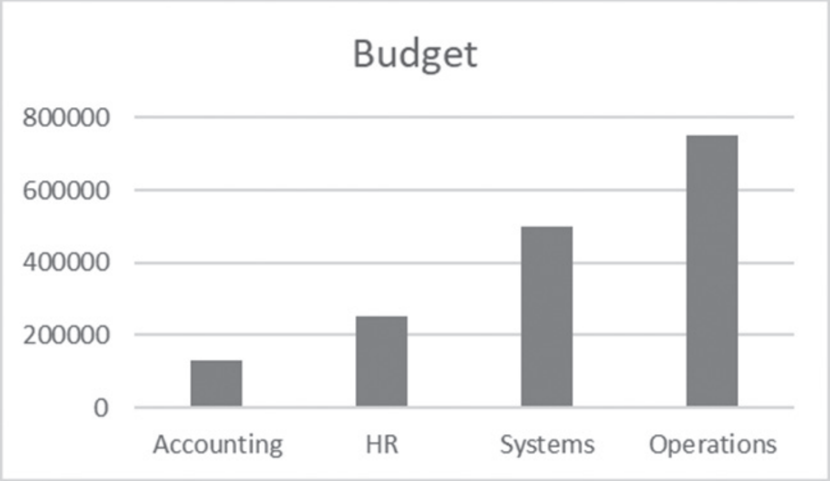


FIGURE 13.1 Bar graph of the data from Table 13.1

Table 13.2 Proportionally normalizing data based on population

Department	Accounting	HR	Systems	Operations
Budget ($)	130000	250000	500000	750000
Population (#)	100	50	200	500
Proportion of budget to population	1300	5000	2500	1500
Proportion of population to budget	.08%	.02%	.04%	.07%

What does this table show? The first part shows that Operations has the largest budget, but look at the proportion of the budget to the population of the department (the last row). Operations has more budget per population, but it has less population served per budget dollar than Accounting. The model should show the different perspectives given a different process model of the data. Given the information above, you could show what amount of the budget would serve each department population equally just by setting the goal .02% for the least or .08% for the most. This creates our simulation, since the scenario is being changed to suit a specific circumstance or situation. Although a deep discussion of model simulation is not part of this text, setting up this specific data for simulation is possible in Excel using the scenario function or in KNIME using a readily available example.[24]

[24] *https://www.nuwavesolutions.com/baking-an-approximate-pi-with-knime-using-a-monte-carlo-recipe/*

Process modeling encompasses an entire book full of procedures and testing, but it is important that you understand what the process modeling entails and where to look to encounter the information necessary for analysis. Since the representation of reality is what a prediction is meant to achieve, the simulation of the modeling should be conducted with verifiable and reputable data.

13.3.2 Simulation

This section is just a brief look at a simulation, using MATLAB as an example of this method. MATLAB is an off-the-shelf (OTS) software that has several types of downloads, the most inexpensive one being for home use.

Before starting with MATLAB, it is important to note that people use simulations in their daily life, whether it be driving to work or using a variety of products. One such example of this is using a specific route to work, which has a longitudinal and historical element to it, making it relatively easy to predict whether the route will be the quickest way to work. Of course, now that data is collected immediately through cell phone apps, it is part of the input to this decision making. However, people also use simulations when plotting scenarios in their head of "what if" something occurs during the drive. The result of that conjecture helps the decision-making process, and consequently helps the individual form more scenarios.

The simulation examples presented here are based on the Monte Carlo simulations illustrated in Paul Nahin's book *Digital Dice* [Nahin-2008]. The program that Mr. Nahin presented in his text was adjusted so that what is placed in the simulation is .08 to see how the error decreases with more sampling. This example shows that the more sampling is conducted, the more accurate the simulation becomes. Simulations conducted with computers are the only way to conduct such a massive sampling with relative ease. Indeed, the idea of a formula for simulations is certainly not a separate conversation. Formulas are necessary for simulations and are adjusted according to what the analyst wants to test. Sometimes the only question that should be asked is "how good is my simulation?" This is, of course, based on the sampling that is placed into the simulation. For instance, is 1000 iterations enough to produce a simulation that would be as accurate as possible? What about 10,000 iterations? Is there a difference between the two?

Nahin conducted a study on these questions and found that the difference between 1000 and 10,000 iterations from the standpoint of accuracy of the simulations is negligible. He used pi as the standard and used random numbers to generate simulations to detect errors in calculating pi. By doing this, he was able to substantiate that 1000 or 10,000 simulations provide essentially the same results [Nahin-2008, p. 224]. Given this information, if the analyst wants to run a simulation, then more simulations are better, but if the resources are not there to run more simulations, then running approximately 1000 simulations should produce the "good enough" results necessary. Running simulations is comparable to sampling, which shows that more sampling can lead to more accurate results, but the degree of accuracy may not be practical with the cost of computer capacity.

CHAPTER 14

SUMMARY

The last section completes the body of this book, and there was so much discussed that it is challenging to summarize it all, so this is an executive summary – short and to the point.

First, data analysts must consider the audience when creating a presentation. Some stakeholders just want a short and concise view of what the data shows and what their requirements produced. In this case, an Exploratory Data Analysis (EDA), otherwise known as descriptive statistics, would suffice. Other stakeholders want details on the tests conducted, why they were conducted, and the results of those tests. In this case, the analyst needs to prepare a report that shows that the methods chosen were needed to satisfy the requirements of the stakeholder. Much like the excessive use of unnecessary medical diagnostic tests, extraneous testing can lead to extra costs. The methods introduced here included the formulas because the tool for the data analyst can change and, therefore, the formula provides the algorithm for the analyst to load it into any tool to get a result.

Second, the analyst must understand the data before trying to pursue the methods in this book. If you just want to conduct testing (i.e., simple regression) without understanding the reason for this test, then you will waste valuable analysis time pursuing a lead that goes nowhere. Some forethought always benefits the analyst.

Finally, there are many more methods than those introduced in this book for analysts to use in their analyses. Please explore books written by data analysts for data analysts. A good one to start this journey is *How to Lie with Statistic*s, written by Darren Huff [Huff-1954]. This book was written in the 1950s and still rings true today. This book shows that approximately

70 years ago, people had the same concerns about data being manipulated or exploited. The methods in this book can help to thoroughly understand how to describe data. Hopefully, this text can spark the fire within you to pursue something better.

CHAPTER 15

CASE STUDIES

The best way to understand data analytics methods is through a case study based on a real-world project. The amount of data presented here is limited to allow for the practice of the formulas in this book without specifying the data science tool needed. In this way, the analyst can apply the methods to any tool that are currently the standard for the organization, using the formula to make algorithms in that tool. The tables used in this section contain data that was drawn from the author's knowledge and experience and do not represent a real-world data set. The data presented is to help the reader perform the necessary exercises without the need of a computer.

15.1 Case Study Scenario

You are a project manager selected to plan and implement a software-based project for your organization. Your staff consists of a systems engineer and a cybersecurity analyst to properly implement the process in a secure manner. The company's software is used to identify the expertise of certain employees so they can participate in future work groups. The employees access the software through the graphic user interface (GUI) and score their competencies or skills through a screen that allows them to rate their skills from entry level to expert. There are approximately 30,000 employees who need to have access to this software, and the availability of the software should be uninterrupted to the greatest extent possible, with a service level agreement of 99% availability (as agreed upon by the stakeholder). The systems engineer has a staff to determine the processes necessary to successfully implement the infrastructure and the cybersecurity analyst has a staff to determine the security needed for the software. There is a staff of programmers, approximately 5, that program the interface but are "matrix-managed" so they do not belong to the project manager but to other departments. The project manager will only be able to access these programmers on a part-time basis, and the project will not interfere with their primary duties. The timeframe is one year from initiation of the project and all the staff members are employees of the organization (there are no contractors or "1099" employees involved). The total staff allocated to the task includes the three department heads (project management, systems engineering, and cybersecurity) with two staff members for each of those departments and five programmers for a total of 14 (including the project manager, systems engineer, and cybersecurity professional) people needed to complete the project. The product owner wants to know if the software should be produced or procured ("make or buy"), and has a few other questions that are addressed in the following sections.

15.2 Case Study: Description of Data

The following considers all of the concepts covered in the section on central tendency (mean, median, and mode). There are no specific tools used in these case studies, so you are free to use whatever tools are at your disposal.

Scenario - Project Management:

You are a project manager in charge of an Agile project consisting of software development for a client. This software is focused on a specific programming language and user stories, but the main reason for the project meeting is to estimate the possible schedule for the project. You decide to use the Fibonacci Method (where each member has a set of numbers 1, 2, 3, 5, 8, 13, 21, 34, 55, 89) to estimate the time frame for each task (or user story), and find the following values from the team consisting of 10 people. You conduct a two-part estimation, where the 10 team members indicate a number that they feel best represents the time necessary to get the task completed (in days). Each part's statistics are listed after that specific part.

Part 1:

Mean = 23, Median = 18, Standard Deviation: 5, Range: 53

Part 2:

Mean = 21, Median = 22, Standard Deviation: 3, Range: 48

You decide to use the mean, median, and standard deviation to determine the central tendency of these estimates. You want to see if the values show a central tendency or if there are outliers that might skew the data to the left or right (depending on where the mean is in relation to the median). You decide to use the formula for the skew to determine the extent of the skewing. After all, if the skewing is to the right, that would mean that the estimates could be overestimated.

From the given descriptive statistics, what do you see is the difference between these two datasets? How would you address this with the project team?

Scenario - Systems Engineering:

You are the engineer on a system that requires complicated tasks and you want to calculate the time of these tasks. You know that one task occurs in systems engineering as a part of the system engineering process. As a result of your research, you come across 15 times the task has been accomplished and record the hours needed for this task. The results are as follows:

4, 5, 10, 15, 20, 4, 6, 8, 8, 8, 6, 7, 13, 19, 16

The different measures are as follows:

Mean = 9.93

Median = 8

Standard Deviation = 5.17

Range = 16

From the different measures, what did you find? What was the skew of the data? Does the skew point out anything about the data? Perform the outlier formula and remove possible outliers.

Scenario - Cybersecurity:

You are the cybersecurity analyst developing the software explained in the Case Study. You want to find out if a machine within the company is taking too much time to access some routine software application. The information in Table 15.1 was retrieved from the active port (i.e., port 124) from each of the machines that have access to the document to see if the data outflow seems peculiar. You decide to take both the mean, median, and standard deviation of each user machine's outflow in megabytes. From your calculations, what conclusion can you give the company? In your estimation, does there seem to be excess use by any of the machines in question? If so, why? How did you come to these conclusions? What is the probability of Machine 2 accessing the data, given that it is a Monday?

Table 15.1 Cybersecurity data for the case study

DAY	Machine 1	Machine 2	Machine 3	Machine 4	MEAN	MEDIAN	STD DEV
SUNDAY	300	300	4000	200	1200	300	1617.096
MONDAY	575	1000	580	900	763.75	740	189.5842
TUESDAY	700	1000	1500	9000	3050	1250	3447.1
WEDNESDAY	9000	1000	400	900	2825	950	3572.377
THURSDAY	800	400	700	1000	725	750	216.5064
FRIDAY	1500	1000	700	400	900	850	406.2019
SATURDAY	400	300	5500	100	1575	350	2268.673
MEAN	1896.429	714.2857	1911.429	1785.714			
MEDIAN	700	1000	700	900			
STAND DEV	2922.463	331.3547	1867.064	2964.277			

15.3 Case Study: Normal Curve

Scenario - Project Management:

Now that you have the standard deviation and the mean, you decide to make a normal distribution to check some of the values individually and see how they line up with the distribution. You take the value of 25 and, using the mean and standard deviation of Part 1, transform the value into a standard normal curve (Z-Score) and check it against the same value of 25, using the mean and standard deviation of Part 2 and check it against the standard normal distribution after transforming it into a Z-Score. What kind of observation can you make about the two values in relation to their standing in the standard normal distribution curve, even though both values are the same? Why would two values that are the same be placed at different locations within the same standard normal distributions? What does this signify with relation to those values?

Scenario - Systems Engineering:

Use the raw data given in Section 15.2 and use the mean and standard deviation that resulted from the previous case study to see how the first two values in the dataset look when placed on the standard normal curve. Which percentile is the first value of the dataset? Does this seem to coincide with the standard normal distribution?

Scenario - Cybersecurity:

Your supervisor asked about the relationship between the data for all the users in your study. She wants to know if one of the users seem to access the document as much or more than User 2, who should be accessing the document the most. You decide to use the standard normal distribution to compare the different values to determine if one of the users seem to be accessing the document more than another. Use the data from Table 15.1.

What did you find when you compared the data using the normal distribution? Did you find any outliers? What made you think they were outliers?

15.4 Case Study: Variation Measures

Scenario - Project Management:

Now that the time estimate has been evaluated, the next area to consider is the team composition, specifically the number of team members necessary to complete the project. You, as the project manager, decide to take a look at similar projects to decide which type of qualifications and number of people are necessary to get the job done. The following five projects were provided.

Table 15.2 Data for project management case study

Project	Engineers	Programmers
1	10	5
2	9	3
3	5	10
4	7	8
5	5	8
Median	7	8
.25 Percentile	5	4
.75 Percentile	9.5	9
Minimum	5	3
Max	10	10

You decide to do a Five Number Summary (already provided) and box plot for engineers vs. programmers on the projects. Please remember that this is a very small number of events and would not normally be used for such a method. However, limiting the data helps you become familiar with the process to automate it. What do you see from the data? Is there what seems to be a "happy medium" for the project in the form of numbers of engineers and programmers?

Scenario - Systems Engineering:

You are appointed as a Systems Engineer/Project Manager on an IT project that concerns automating a function within a customer service area of a company. You take a data analysis approach using a Five Number Summary and box plot to compare two separate tools that are options to improve the overall functionality and efficiency of the customer service application. The data shows each tool with the time it took for the customer service function in minutes. Currently, the customer service function takes 30 minutes to complete, and the goal is to get at least 10% better, on the average. What is the decision that would seem best, given the results of the data analysis? Is one option preferable over the other?

Table 15.3 Data for the systems engineer case study

User Number	Option 1 (minutes)	Option 2 (minutes)
1	25	22
2	26	28
3	31	35
4	20	22
5	25	30
6	38	33
7	20	26
8	27	20
9	24	20
10	29	30
Median	25.5	27
.25 Percentile	23	21.5
.75 Percentile	29.5	30.75
Minimum	20	20
Maximum	38	35
Range	18	15

Scenario - Cybersecurity:

You are a cybersecurity consultant looking into a company's use of computers during the day. You decide to take a sample of 4 machines over a period of 10 days to see how much bandwidth is used each day by each machine. You want to compare the machines because one is from Management, one is from Production, one is from HR, and one is from Accounting. The numbers represent the number of gigabytes per day that each of the machines are using in memory. Upper management is concerned that one department is using too much of the available memory and that could be an indicator that unorthodox downloads are taking place in that department. You decide to do a Five Number Summary and box plots on each of the machines and visualize the data for upper management. The collected data is in Table 15.4.

Table 15.4 Data for cybersecurity case study

Event	Management	Accounting	Production	Human Resources
1	900	500	600	700
2	400	700	800	1000
3	900	400	800	200
4	500	700	700	800
5	300	600	800	1000
6	800	300	700	300
7	800	1000	400	200
8	700	1000	500	700
9	800	300	800	900
10	1000	1000	900	700
Median	800	650	750	700
.25 Percentile	475	375	575	275
.75 Percentile	900	1000	800	925
Maximum	1000	1000	900	1000
Minimum	300	300	400	200
Range	700	700	500	800

From the analysis, is one computer using a lot more memory than another? What do the box plots show?

15.5 Case Study: Probability

Scenario - Project Management:

You are a project manager for an IT project that encompasses the use of software that tracks employees needed for specific workgroups. The software incorporates a user interface to self-evaluate certain competencies or skills that the employee possesses and report those competencies on a 1-5 skill level, 1 being entry level and 5 being expert. Your job is to evaluate whether it is better to buy or build the software within a year. You evaluated a software vendor and rated them on the expense and project completion. The vendor has completed 5 projects that are about the same as yours, with 3 of them being on time and 4 of them completed on budget. Given the joint probability, what is the probability that the vendor will get the project completed on time AND on budget? If the standard for such a probability is 50% on time and on budget, does the vendor meet the prescribed standard? If you have government programmers who have completed 10 projects, with 7 of them on time and 6 of them on budget, what is the comparison between them and the vendors?

Scenario - Systems Engineering:

You are a systems engineer on the same project as the one in the first scenario. The project manager has asked you to determine if a make vs. buy decision is possible through the comparison of the processes for a vendor of a software product and the proposed architecture for the in-house solution.

You decide to use the value formula, value = function/cost [Walden-2015, 242], but you find that the cost of the vendor software changes with certain functions, and find that in approximately 3 cases out of 50, the vendor software has functions that are not suitable for the user. The in-house history shows that is 5 cases out of 100 where the functionality is not suited to the user. Using basic probability, which one is the better alternative, even though there will be functions that may not be suitable to the user? If the process used in the software takes a user 43 minutes to complete 5 steps, and it takes 54 minutes to complete 7 steps in the in-house solution, which one takes less time (using probability)?

What is the cost if there are 5 major functions in the vendor software and the cost of those functions is $100,000 and each of those functions take 3 weeks to complete? Use basic probability to complete this example.

Scenario - Cybersecurity:

You have audited the software vendor and found that their Common Vulnerability Scoring System (CVSS) score (a score evaluating the security of the software; an explanation can be found at *nvd.nist.gov/vuln-metrics/cvss/v3-calculator*) has been the following when evaluated in the past 5 years of the same software: 5.3, 6.2, 4.5, 3.5, and 3.3. Using probability, what are the chances that the software passes a standard score (which is less than 5.5)? What other factors should be considered when attempting this probability?

15.6 Case Study: Occam's Razor

Scenario - Project Management, Systems Engineering, and Cybersecurity:

You are working on your project for the organization as covered in the Case Study description when your boss calls you into her office to discuss some of the requirements. One of the requirements is that there should be a way to communicate with the person who has the competencies necessary for a specific workgroup. You want to ensure that the communication method meets the specifications necessary to make the contact with the employee as quickly as possible, again based on the requirements from the boss. You meet with the systems engineer and the cybersecurity analyst to gather the information necessary to go back to the boss with more specific requirements. What questions should each department ask to get as specific a requirement as possible? Why not just go with email as the solution and start making plans to incorporate that into the project? If email is an option, what questions should be addressed by the systems engineer and the cybersecurity analyst to adequately cover the boss's requirements?

APPENDIX

RECOMMENDED SOLUTIONS FOR CASE STUDIES

Introduction

These examples are simple because the purpose is the focus on the process, not the results. In keeping with this, the answers should contain the recommended formulas and process to get to the answer. This includes any calculations to attain the answer to the question.

A.1 Recommended Approach for Case Study 15.2

All Professions: The outlier formula would be a good start (found in Section 3.7.3.2), remembering that the data has to be sorted and the 25th and 75th percentiles have to be calculated. Then, using the IQR formula, the outliers are revealed. Remember that if the outlier is in the negative, and the data goes no further than 0, then that means that there are no negative outliers to the data, so 0 is the “floor” of the data.

For the skew formula, in order to find if the data is skewed right or left (positive or negative, respectively), the easiest way to do this is to use the formula located in Section 3.0.1 (basically using the mean, median, and standard deviation). Remember that if the number is positive, the data is probably right skewed; if the number is negative, the data is left skewed. A chart with the data is very helpful to further illustrate any peculiarities in the data. (Most chart tools are functions within the software package, like KNIME, R, and Excel.)

Since the rest of the case study examples contain the necessary data to do the tests above, the recommended process is the same for all of them. The goal is to develop a natural instinct as a data analyst to look at the mean, median, and mode and use those to get a perspective of the data. The more that an analyst can picture the data, the more that they can describe it in better detail.

A.2 Recommended Approach for Case Study 15.3

The first thing to do is use the normal curve to compare two different datasets. How is that done? Please refer to section 3.6 for the process of taking each mean and standard deviation and using the Z-score formula to compare the two datasets. Remember that once the Z-scores are attained, take each of them and compare them with one another; that comparison will them determine if the one score is greater or less than the other by using the standard normal curve. For instance, if one of the random variables (any one value in the dataset) has a Z-score of 2.0 and the other random variable has a Z-score of 1.5, then the second value is less than the first value. For clarification, "less" means that there are less values below the second Z-Score than then the first Z-Score. Since the answer to each of the exercises uses the same process, please repeat this process for each. Even though the values may seem to be more or less than one another, once the standard normal curve is determined, those respective values may be the reverse of their actual number. This may seem counterintuitive, but what you are doing is comparing two disparate datasets using a very standard form of data transformation.

A.3 Recommended Approach for Case Study 15.4

The Five Number Summary is provided for each of the examples, so the recommendation is to ignore those results and figure the results using the percentile formula provided in Section 3.7.3.3. Remember that all of the figures that are calculated from the summary are part of the box plot. The translation of the Five Number Summary to the box plot is provided in Section 3.7.4. Once that is constructed, there are a few areas to consider, including how one box plot might "overlap" another box plot. If that is the situation, then that means that there might be little difference between those different variables (or fields or columns). If this is the situation, it might be worthwhile to perform an ANOVA (analysis of variance) on the different variables. ANOVA is not addressed in this text, but it compares more than two variables with relatively good accuracy. The main limitation to doing an ANOVA is that the result will show if there are differences between variables, but not show which variable is different.

A.4 Recommended Approach for Case Study 15.5

Project Management:

Remember that joint probability, the “and” calculation, means that you have to multiply the two probabilities together to get the joint probability. Since the two probabilities in the project management section are provided, you can take the number of on-time projects (3) and divide that by the total number of projects (5) to get the first probability and then do the same for the on-budget projects (4) and the total projects (5). Then you take those two probabilities and multiply them together to see if they meet the 50% criteria. The next example of using the government programmers uses the same process, and then you can compare that answer with 50% and the vendor probability. You can also refer to Sections 4.0, 4.1, and 4.2 for any further help in this area.

Systems Engineering:

This is basic probability, so no further explanation is really necessary other than the comparison of both proportions. Using proportions is a form of analysis. There is a way to get both the mean and standard deviation from proportions, although they were not really covered in this text. The mean of the proportion is the probability of success (normally “p”) in the proportion equations. The standard deviation of the proportion (or probability) is the square root of the probability of the success times the probability of failure (1-p or q), divided by the number in the sample. For instance, if the proportion (probability of success) of producing an error-free function for the system is 70% or .70 then that would be the mean of the proportion. To find the standard deviation, you would have to know the sample; in this case, make it 100. That means that the standard deviation is the square root of (.70*.30)/100. Remember the square root! Once the standard deviation and the mean are determined, you can convert the proportion into a Z-score to determine the comparison to the standard normal distribution (use this site to help you: *www.tutorialspoint.com/statistics/one_proportion_z_test.htm).*

Cybersecurity:

This is a bit tricky, since one answer might be to take the mean of the scores and determine if the mean has ever been more than 5.5. However, the probability of the problem comes from how many times the score has been more than 5.5. You use that number to get a proportion or probability of whether it is feasible to get more than a 5.5. Remember to use Occam's Razor on this one.

A.5 Recommended Approach for Case Study 15.6

This is a real-world situation that happened on software made specifically for communicating between the user, the supervisor, and the workgroup leader. In the real-world case, the following questions were part of the overall reasoning behind the email inclusion in the project.

1. Will this email be solely for government use?
2. Will the email need to be secured at a level commensurate with current email security?
3. Will the email need to be approved by the Resource Management Office? Will the email be part of the eDiscovery initiative (the eDiscovery initiative established ANY email is discoverable by legal counsel should the employee end up in court)?
4. Will it be necessary to have the Privacy Office be involved in this email portion of the software?
5. Since this agency has an established union, should the office that interacts with the union be notified of this as part of the software development?

As you can see, just including what seemed to be a small addition to the software compounded the complexity of the software development. However, and most importantly, using email was the simplest and most effective method to include all parties in any possible information gathering that was necessary in the process of using employee background for selection to a workgroup. Since this section concerns the simplest solution, the email solution is the correct approach. However, the questions above point out that important information can be gleaned by simply realizing the basic components of the organization and how they affect any decision.

REFERENCES

[Kurt-2019] W. Kurt, Baysian Statistics the Fun Way, No Starch Press, 2019.

[Spade-2019] P. V. a. P. C. Spade, “William of Occam,” *The Stanford Encyclopedia of Philosophy,* 2019.

[Glen-2015] S. Glen, “Type III Error and Type IV Error in Statistical Tests,” 16 January 2015. [Online]. Available: Stephanie Glen. “Type III Error and Type IV Error in Statistical Teshttps://www.statisticshowto.com/type-iii-error-in-statistical-tests/.

[Mitchell-2009] A. Mitchell, The Esri Guide to GIS Analysis: Volume 2: Spatial Measurements & Statistics, Redlands: Esri Press, 2009.

[Shearer-2000] C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining,” *Journal of Data Warehousing,* vol. 5, no. 4, pp. 13-22, 2000.

[PMI-2017] Project Management Institute (PMI), A Guide to the Project Management Body of Knowledge: PMBOK GUIDE (6th Edition), Newtown Square: Project Management Institute, 2017.

[Walden-2015] DD Walden, GJ Roedler, KJ Forsberg, RD Hamelin, and TM Shortell (Eds.), Systems Engineering Handbook: A Guide for System Life Cycle Processes and Activities, 4th ed., San Diego, California: John Wiley & Sons, Inc., 2015.

[Dong-2020] H. D. L. G. Ensheng Dong, “An interactive web-based dashboard to track COVID-19 in real time,” *The Lancet,* pp. 30120-1, 19 February 2020.

[Reinhart-2015] A. Reinhart, Statistics Done Wrong: The Woefully Complete Guide, San Francisco: No Starch Press, 2015.

[Shannon-1948] C. Shannon, “A Mathematical Theory of Communication,” *The Bell System Technical Journal,* pp. 379-423, 623-656, 1948.

[Delahaye-2019] J.-P. Delahaye, “The Mathematics of (Hacking) Passwords,” 19 April 2019. [Online]. Available: https://www.scientificamerican.com/article/the-mathematics-of-hacking-passwords/.

[Provost-2013] F. a. T. F. Provost, Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking, O’Reilly Publishing, 2013.

[Knaflic-2015] C. N. Knaflic, Storytelling with Data: A Data Visualization Guide for Business Professionals, Hoboken: J. Wiley and Sons, Inc., 2015.

[Nahin-2008] P. J. Nahin, Digital Dice: Computational Solutions to Practical Probability Problems, Oxford: Princeton University Press, 2008.

[Carlberg-2013] C. Carlberg, Predictive Analytics: Microsoft Excel, Que Publishing, 2013.

[Cirillo-2017] A. Cirillo, R Data Mining, Packt Publishing, 2017.

[Teetor-2011] P. Teetor, R Cookbook, O'Reilly Publishing, 2011.

[Bruce-2017] P. a. A. B. Bruce, Practical Statistics for Data Scientists: 50 Essential Concepts, New York: O'Reilly Publishing, 2017.

[Carlberg-2014] C. Carlberg, Decision Analytics: Microsoft Excel, Indianapolis: Que, 2014.

[Poundstone-2019] W. Poundstone, The Doomsday Calculation, New York: Little, Brown Spark, 2019.

[Gladwell-2008] M. Gladwell, Outliers: The Story of Success, New York: Little, Brown, and Company, 2008.

[Vigen-2015] T. Vigen, Spurious Correlations: Correlation Does Not Equal Causation, New York: Hachette Books, 2015.

[Huff-1954] Huff, D., How to Lie with Statistics, WW Norton and Company, 1954.

INDEX