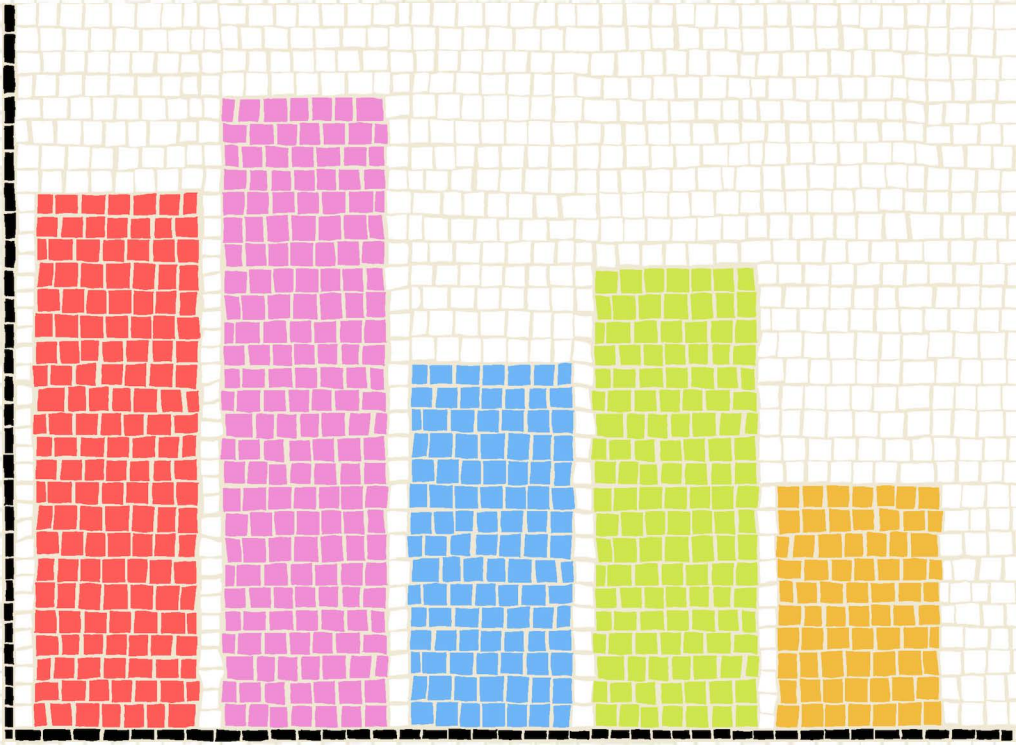
The Art of Data Science

A Guide for Anyone Who Works with Data



Roger D. Peng & Elizabeth Matsui

The Art of Data Science

A Guide for Anyone Who Works with Data

Roger D. Peng and Elizabeth Matsui

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

Also By Roger D. Peng

R Programming for Data Science

Exploratory Data Analysis with R

Report Writing for Data Science in R

Special thanks to Maggie Matsui, who created all of the artwork for this book.

Contents

1. Data Analysis as Art

Data analysis is hard, and part of the problem is that few people can explain how to do it. It's not that there aren't any people doing data analysis on a regular basis. It's that the people who are really good at it have yet to enlighten us about the thought process that goes on in their heads.

Imagine you were to ask a songwriter how she writes her songs. There are many tools upon which she can draw. We have a general understanding of how a good song should be structured: how long it should be, how many verses, maybe there's a verse followed by a chorus, etc. In other words, there's an abstract framework for songs in general. Similarly, we have music theory that tells us that certain combinations of notes and chords work well together and other combinations don't sound good. As good as these tools might be, ultimately, knowledge of song structure and music theory alone doesn't make for a good song. Something else is needed.

In Donald Knuth's legendary 1974 essay *Computer Programming as an Art*[1], Knuth talks about the difference between art and science. In that essay, he was trying to get across the idea that although computer programming involved complex machines and very technical knowledge, the act of writing a computer program had an artistic component. In this essay, he says that

> Science is knowledge which we understand so well that we can teach it to a computer.

[1] http://www.paulgraham.com/knuth.html

Everything else is art.

At some point, the songwriter must inject a creative spark into the process to bring all the songwriting tools together to make something that people want to listen to. This is a key part of the *art* of songwriting. That creative spark is difficult to describe, much less write down, but it's clearly essential to writing good songs. If it weren't, then we'd have computer programs regularly writing hit songs. For better or for worse, that hasn't happened yet.

Much like songwriting (and computer programming, for that matter), it's important to realize that *data analysis is an art*. It is not something yet that we can teach to a computer. Data analysts have many *tools* at their disposal, from linear regression to classification trees and even deep learning, and these tools have all been carefully taught to computers. But ultimately, a data analyst must find a way to assemble all of the tools and apply them to data to answer a relevant question—a question of interest to people.

Unfortunately, the process of data analysis is not one that we have been able to write down effectively. It's true that there are many statistics textbooks out there, many lining our own shelves. But in our opinion, none of these really addresses the core problems involved in conducting real-world data analyses. In 1991, Daryl Pregibon, a prominent statistician previously of AT&T Research and now of Google, said in reference to the process of data analysis[2] that "statisticians have a process that they espouse but do not fully understand".

Describing data analysis presents a difficult conundrum. On the one hand, developing a useful framework involves characterizing the elements of a data analysis using abstract

[2] http://www.nap.edu/catalog/1910/the-future-of-statistical-software-proceedings-of-a-forum

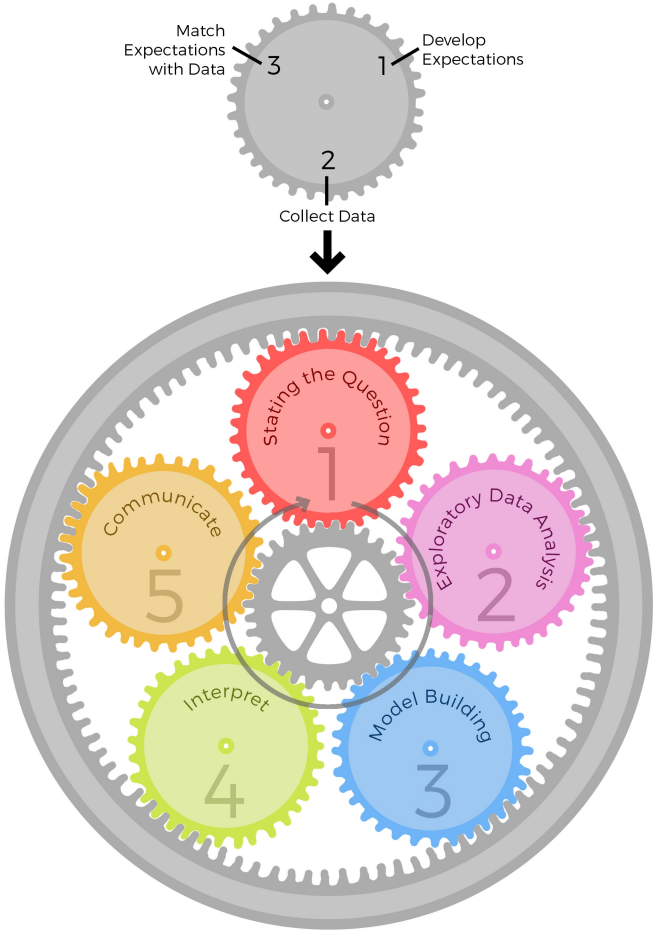
language in order to find the commonalities across different kinds of analyses. Sometimes, this language is the language of mathematics. On the other hand, it is often the very details of an analysis that makes each one so difficult and yet interesting. How can one effectively generalize across many different data analyses, each of which has important unique aspects?

What we have set out to do in this book is to write down the process of data analysis. What we describe is not a specific "formula" for data analysis—something like "apply this method and then run that test"— but rather is a general process that can be applied in a variety of situations. Through our extensive experience both managing data analysts and conducting our own data analyses, we have carefully observed what produces coherent results and what fails to produce useful insights into data. Our goal is to write down what we have learned in the hopes that others may find it useful.

2. Epicycles of Analysis

To the uninitiated, a data analysis may appear to follow a linear, one-step-after-the-other process which at the end, arrives at a nicely packaged and coherent result. In reality, data analysis is a highly iterative and non-linear process, better reflected by a series of epicycles (see Figure), in which information is learned at each step, which then informs whether (and how) to refine, and redo, the step that was just performed, or whether (and how) to proceed to the next step.

An epicycle is a small circle whose center moves around the circumference of a larger circle. In data analysis, the iterative process that is applied to all steps of the data analysis can be conceived of as an epicycle that is repeated for each step along the circumference of the entire data analysis process. Some data analyses appear to be fixed and linear, such as algorithms embedded into various software platforms, including apps. However, these algorithms are final data analysis products that have emerged from the very non-linear work of developing and refining a data analysis so that it can be "algorithmized."



Epicycles of Analysis

2.1 Setting the Scene

Before diving into the "epicycle of analysis," it's helpful to pause and consider what we mean by a "data analysis." Although many of the concepts we will discuss in this

book are applicable to conducting a *study*, the framework and concepts in this, and subsequent, chapters are tailored specifically to conducting a *data analysis*. While a study includes developing and executing a plan for collecting data, a data analysis presumes the data have already been collected. More specifically, a study includes the development of a hypothesis or question, the designing of the data collection process (or study protocol), the collection of the data, and the analysis and interpretation of the data. Because a data analysis presumes that the data have already been collected, it includes development and refinement of a question and the process of analyzing and interpreting the data. It is important to note that although a data analysis is often performed without conducting a study, it may also be performed as a component of a study.

2.2 Epicycle of Analysis

There are 5 core activities of data analysis:

1. Stating and refining the question
2. Exploring the data
3. Building formal statistical models
4. Interpreting the results
5. Communicating the results

These 5 activities can occur at different time scales: for example, you might go through all 5 in the course of a day, but also deal with each, for a large project, over the course of many months. Before discussing these core activities, which will occur in later chapters, it will be important to first understand the overall framework used to approach each of these activities.

Although there are many different types of activities that you might engage in while doing data analysis, every aspect of the entire process can be approached through an interative process that we call the "epicycle of data analysis". More specifically, for each of the five core activities, it is critical that you engage in the following steps:

1. Setting Expectations,
2. Collecting information (data), comparing the data to your expectations, and if the expectations don't match,
3. Revising your expectations or fixing the data so your data and your expectations match.

Iterating through this 3-step process is what we call the "epicycle of data analysis." As you go through every stage of an analysis, you will need to go through the epicycle to continuously refine your question, your exploratory data analysis, your formal models, your interpretation, and your communication.

The repeated cycling through each of these five core activities that is done to complete a data analysis forms the larger circle of data analysis (See Figure). In this chapter we go into detail about what this 3-step epicyclic process is and give examples of how you can apply it to your data analysis.

	Set Expectations	Collect Information	Revise Expectations
Question	Question is of interest to audience	Literature Search/Experts	Sharpen question
EDA	Data are appropriate for question	Make exploratory plots of data	Refine question or collect more data
Formal Modeling	Primary model answers question	Fit secondary models, sensitivity analysis	Revise formal model to include more predictors
Interpretation	Interpretation of analyses provides a specific & meaningful answer to the question	Interpret totality of analyses with focus on effect sizes & uncertainty	Revise EDA and/or models to provide specific & interpretable answer
Communication	Process & results of analysis are understood, complete & meaningful to audience	Seek feedback	Revise analyses or approach to presentation

Epicycles of Analysis

2.3 Setting Expectations

Developing expectations is the process of deliberately thinking about what you expect before you do anything, such as inspect your data, perform a procedure, or enter a command. For experienced data analysts, in some circumstances, developing expectations may be an automatic, almost subconscious process, but it's an important activity to cultivate and be deliberate about.

For example, you may be going out to dinner with friends at a cash-only establishment and need to stop by the ATM to withdraw money before meeting up. To make a decision about the amount of money you're going to withdraw, you have to have developed some expectation of the cost of dinner. This may be an automatic expectation because you dine at this establishment regularly so you know what the

typical cost of a meal is there, which would be an example of *a priori* knowledge. Another example of *a priori* knowledge would be knowing what a typical meal costs at a restaurant in your city, or knowing what a meal at the most expensive restaurants in your city costs. Using that information, you could perhaps place an upper and lower bound on how much the meal will cost.

You may have also sought out external information to develop your expectations, which could include asking your friends who will be joining you or who have eaten at the restaurant before and/or Googling the restaurant to find general cost information online or a menu with prices. This same process, in which you use any a priori information you have and/or external sources to determine what you expect when you inspect your data or execute an analysis procedure, applies to each core activity of the data analysis process.

2.4 Collecting Information

This step entails collecting information about your question or your data. For your question, you collect information by performing a literature search or asking experts in order to ensure that your question is a good one. In the next chapter, we will discuss characteristics of a good question. For your data, after you have some expectations about what the result will be when you inspect your data or perform the analysis procedure, you then perform the operation. The results of that operation are the data you need to collect, and then you determine if the data you collected matches your expectations. To extend the restaurant metaphor, when you go to the restaurant, getting the check is collecting the data.

2.5 Comparing Expectations to Data

Now that you have data in hand (the check at the restaurant), the next step is to compare your expectations to the data. There are two possible outcomes: either your expectations of the cost matches the amount on the check, or they do not. If your expectations and the data match, terrific, you can move onto the next activity. If, on the other hand, your expectations were a cost of 30 dollars, but the check was 40 dollars, your expectations and the data do not match. There are two possible explanations for the discordance: first, your expectations were wrong and need to be revised, or second, the check was wrong and contains an error. You review the check and find that you were charged for two desserts instead of the one that you had, and conclude that there is an error in the data, so ask for the check to be corrected.

One key indicator of how well your data analysis is going is how easy or difficult it is to match the data you collected to your original expectations. You want to setup your expectations and your data so that matching the two up is easy. In the restaurant example, your expectation was $30 and the data said the meal cost $40, so it's easy to see that (a) your expectation was off by $10 and that (b) the meal was more expensive than you thought. When you come back to this place, you might bring an extra $10. If our original expectation was that the meal would be between $0 and $1,000, then it's true that our data fall into that range, but it's not clear how much more we've learned. For example, would you change your behavior the next time you came back? The expectation of a $30 meal is sometimes referred to as a sharp hypothesis because it states something very specific that can be verified with the data.

2.6 Applying the Epicyle of Analysis Process

Before we discuss a couple of examples, let's review the three steps to use for each core data analysis activity. These are :

1. Setting expectations,
2. Collecting information (data), comparing the data to your expectations, and if the expectations don't match,
3. Revising your expectations or fixing the data so that your expectations and the data match.

Example: Asthma prevalence in the U.S.

Let's apply the "data analysis epicycle" to a very basic example. Let's say your initial question is to determine the prevalence of asthma among adults, because your company wants to understand how big the market might be for a new asthma drug. You have a general question that has been identified by your boss, but need to: (1) sharpen the question, (2) explore the data, (3) build a statistical model, (4) interpret the results, and (5) communicate the results. We'll apply the "epicycle" to each of these five core activities.

For the first activity, refining the question, you would first develop your expectations of the question, then collect information about the question and determine if the information you collect matches your expectations, and if not, you would revise the question. Your expectations are that the answer to this question is unknown and that the question is answerable. A literature and internet search, however, reveal that this question has been answered (and is continually answered by the Centers for Disease Control (CDC)), so you

reconsider the question since you can simply go to the CDC website to get recent asthma prevalence data.

You inform your boss and initiate a conversation that reveals that any new drug that was developed would target those whose asthma was not controlled with currently available medication, so you identify a better question, which is "how many people in the United States have asthma that is not currently controlled, and what are the demographic predictors of uncontrolled asthma?" You repeat the process of collecting information to determine if your question is answerable and is a good one, and continue this process until you are satisfied that you have refined your question so that you have a good question that can be answered with available data.

Let's assume that you have identified a data source that can be downloaded from a website and is a sample that represents the United States adult population, 18 years and older. The next activity is exploratory data analysis, and you start with the expectation that when you inspect your data that there will be 10,123 rows (or records), each representing an individual in the US as this is the information provided in the documentation, or codebook, that comes with the dataset. The codebook also tells you that there will be a variable indicating the age of each individual in the dataset.

When you inspect the data, though, you notice that there are only 4,803 rows, so return to the codebook to confirm that your expectations are correct about the number of rows, and when you confirm that your expectations are correct, you return to the website where you downloaded the files and discover that there were two files that contained the data you needed, with one file containing 4,803 records and the second file containing the remaining 5,320 records. You download the second file and read it into your statistical

software package and append the second file to the first.

Now you have the correct number of rows, so you move on to determine if your expectations about the age of the population matches your expectations, which is that everyone is 18 years or older. You summarize the age variable, so you can view the minimum and maximum values and find that all individuals are 18 years or older, which matches your expectations. Although there is more that you would do to inspect and explore your data, these two tasks are examples of the approach to take. Ultimately, you will use this data set to estimate the prevalence of uncontrolled asthma among adults in the US.

The third activity is building a statistical model, which is needed in order to determine the demographic characteristics that best predict that someone has uncontrolled asthma. Statistical models serve to produce a precise formulation of your question so that you can see exactly how you want to use your data, whether it is to estimate a specific parameter or to make a prediction. Statistical models also provide a formal framework in which you can challenge your findings and test your assumptions.

Now that you have estimated the prevalence of uncontrolled asthma among US adults and determined that age, gender, race, body mass index, smoking status, and income are the best predictors of uncontrolled asthma available, you move to the fourth core activity, which is interpreting the results. In reality, interpreting results happens along with model building as well as after you've finished building your model, but conceptually they are distinct activities.

Let's assume you've built your final model and so you are moving on to interpreting the findings of your model. When you examine your final predictive model, initially your expectations are matched as age, African American/black race,

body mass index, smoking status, and low income are all positively associated with uncontrolled asthma.

However, you notice that female gender is *inversely* associated with uncontrolled asthma, when your research and discussions with experts indicate that among adults, female gender should be positively associated with uncontrolled asthma. This mismatch between expectations and results leads you to pause and do some exploring to determine if your results are indeed correct and you need to adjust your expectations or if there is a problem with your results rather than your expectations. After some digging, you discover that you had thought that the gender variable was coded 1 for female and 0 for male, but instead the codebook indicates that the gender variable was coded 1 for male and 0 for female. So the interpretation of your results was incorrect, not your expectations. Now that you understand what the coding is for the gender variable, your interpretation of the model results matches your expectations, so you can move on to communicating your findings.

Lastly, you communicate your findings, and yes, the epicycle applies to communication as well. For the purposes of this example, let's assume you've put together an informal report that includes a brief summary of your findings. Your expectation is that your report will communicate the information your boss is interested in knowing. You meet with your boss to review the findings and she asks two questions: (1) how recently the data in the dataset were collected and (2) how changing demographic patterns projected to occur in the next 5-10 years would be expected to affect the prevalence of uncontrolled asthma. Although it may be disappointing that your report does not fully meet your boss's needs, getting feedback is a critical part of doing a data analysis, and in fact, we would argue that a good data analysis requires communication, feedback, and then

actions in response to the feedback.

Although you know the answer about the years when the data were collected, you realize you did not include this information in your report, so you revise the report to include it. You also realize that your boss's question about the effect of changing demographics on the prevalence of uncontrolled asthma is a good one since your company wants to predict the size of the market in the future, so you now have a new data analysis to tackle. You should also feel good that your data analysis brought additional questions to the forefront, as this is one characteristic of a successful data analysis.

In the next chapters, we will make extensive use of this framework to discuss how each activity in the data analysis process needs to be continuously iterated. While executing the three steps may seem tedious at first, eventually, you will get the hang of it and the cycling of the process will occur naturally and subconsciously. Indeed, we would argue that most of the best data analysts don't even realize they are doing this!

3. Stating and Refining the Question

Doing data analysis requires quite a bit of thinking and we believe that when you've completed a good data analysis, you've spent more time thinking than doing. The thinking begins before you even look at a dataset, and it's well worth devoting careful thought to your question. This point cannot be over-emphasized as many of the "fatal" pitfalls of a data analysis can be avoided by expending the mental energy to get your question right. In this chapter, we will discuss the characteristics of a good question, the types of questions that can be asked, and how to apply the iterative epicyclic process to stating and refining your question so that when you start looking at data, you have a sharp, answerable question.

3.1 Types of Questions

Before we delve into stating the question, it's helpful to consider what the different types of questions are. There are six basic types of questions and much of the discussion that follows comes from a paper[1] published in *Science* by Roger and Jeff Leek[2]. Understanding the type of question you are asking may be the most fundamental step you can take to ensure that, in the end, your interpretation of the results is correct. The six types of questions are:

[1] http://www.sciencemag.org/content/347/6228/1314.short
[2] http://jtleek.com

1. Descriptive
2. Exploratory
3. Inferential
4. Predictive
5. Causal
6. Mechanistic

And the type of question you are asking directly informs how you interpret your results.

A *descriptive* question is one that seeks to summarize a characteristic of a set of data. Examples include determining the proportion of males, the mean number of servings of fresh fruits and vegetables per day, or the frequency of viral illnesses in a set of data collected from a group of individuals. There is no interpretation of the result itself as the result is a fact, an attribute of the set of data that you are working with.

An *exploratory* question is one in which you analyze the data to see if there are patterns, trends, or relationships between variables. These types of analyses are also called "hypothesis-generating" analyses because rather than testing a hypothesis as would be done with an inferential, causal, or mechanistic question, you are looking for patterns that would support proposing a hypothesis. If you had a general thought that diet was linked somehow to viral illnesses, you might explore this idea by examining relationships between a range of dietary factors and viral illnesses. You find in your exploratory analysis that individuals who ate a diet high in certain foods had fewer viral illnesses than those whose diet was not enriched for these foods, so you propose the hypothesis that among adults, eating at least 5 servings a day of fresh fruit and vegetables is associated with fewer viral illnesses per year.

An *inferential* question would be a restatement of this proposed hypothesis as a question and would be answered by analyzing a different set of data, which in this example, is a representative sample of adults in the US. By analyzing this different set of data you are both determining if the association you observed in your exploratory analysis holds in a different sample and whether it holds in a sample that is representative of the adult US population, which would suggest that the association is applicable to all adults in the US. In other words, you will be able to infer what is true, on average, for the adult population in the US from the analysis you perform on the representative sample.

A *predictive* question would be one where you ask what types of people will eat a diet high in fresh fruits and vegetables during the next year. In this type of question you are less interested in what causes someone to eat a certain diet, just what predicts whether someone will eat this certain diet. For example, higher income may be one of the final set of predictors, and you may not know (or even care) why people with higher incomes are more likely to eat a diet high in fresh fruits and vegetables, but what is most important is that income is a factor that predicts this behavior.

Although an inferential question might tell us that people who eat a certain type of foods tend to have fewer viral illnesses, the answer to this question does not tell us if eating these foods causes a reduction in the number of viral illnesses, which would be the case for a *causal* question. A causal question asks about whether changing one factor will change another factor, on average, in a population. Sometimes the underlying design of the data collection, by default, allows for the question that you ask to be causal. An example of this would be data collected in the context of a randomized trial, in which people were randomly assigned to eat a diet high in fresh fruits and vegetables or one that

was low in fresh fruits and vegetables. In other instances, even if your data are not from a randomized trial, you can take an analytic approach designed to answer a causal question.

Finally, none of the questions described so far will lead to an answer that will tell us, if the diet does, indeed, cause a reduction in the number of viral illnesses, *how* the diet leads to a reduction in the number of viral illnesses. A question that asks how a diet high in fresh fruits and vegetables leads to a reduction in the number of viral illnesses would be a *mechanistic* question.

There are a couple of additional points about the types of questions that are important. First, by necessity, many data analyses answer multiple types of questions. For example, if a data analysis aims to answer an inferential question, descriptive and exploratory questions must also be answered during the process of answering the inferential question. To continue our example of diet and viral illnesses, you would not jump straight to a statistical model of the relationship between a diet high in fresh fruits and vegetables and the number of viral illnesses without having determined the frequency of this type of diet and viral illnesses and their relationship to one another in this sample. A second point is that the type of question you ask is determined in part by the data available to you (unless you plan to conduct a study and collect the data needed to do the analysis). For example, you may want to ask a causal question about diet and viral illnesses to know whether eating a diet high in fresh fruits and vegetables causes a decrease in the number of viral illnesses, and the best type of data to answer this causal question is one in which people's diets change from one that is high in fresh fruits and vegetables to one that is not, or vice versa. If this type of data set does not exist, then the best you may be able to do is either apply causal

analysis methods to observational data or instead answer an inferential question about diet and viral illnesses.

3.2 Applying the Epicycle to Stating and Refining Your Question

You can now use the information about the types of questions and characteristics of good questions as a guide to refining your question. To accomplish this, you can iterate through the 3 steps of:

1. Establishing your expectations about the question
2. Gathering information about your question
3. Determining if your expectations match the information you gathered, and then refining your question (or expectations) if your expecations did not match the information you gathered

3.3 Characteristics of a Good Question

There are five key characteristics of a good question for a data analysis, which range from the very basic characteristic that the question should not have already been answered to the more abstract characteristic that each of the possible answers to the question should have a single interpretation and be meaningful. We will discuss how to assess this in greater detail below.

As a start, the question should be of **interest** to your audience, the identity of which will depend on the context and environment in which you are working with data. If you are in academia, the audience may be your collaborators, the scientific community, government regulators, your

funders, and/or the public. If you are working at a start-up, your audience is your boss, the company leadership, and the investors. As an example, answering the question of whether outdoor particulate matter pollution is associated with developmental problems in children may be of interest to people involved in regulating air pollution, but may not be of interest to a grocery store chain. On the other hand, answering the question of whether sales of pepperoni are higher when it is displayed next to the pizza sauce and pizza crust or when it is displayed with the other packaged meats would be of interest to a grocery store chain, but not to people in other industries.

You should also check that the question has **not already been answered.** With the recent explosion of data, the growing amount of publicly available data, and the seemingly endless scientific literature and other resources, it is not uncommon to discover that your question of interest has been answered already. Some research and discussion with experts can help sort this out, and can also be helpful because even if the specific question you have in mind has not been answered, related questions may have been answered and the answers to these related questions are informative for deciding if or how you proceed with your specific question.

The question should also stem from a **plausible** framework. In other words, the question above about the relationship between sales of pepperoni and its placement in the store is a plausible one because shoppers buying pizza ingredients are more likely than other shoppers to be interested in pepperoni and may be more likely to buy it if they see it at the same time that they are selecting the other pizza ingredients. A less plausible question would be whether pepperoni sales correlate with yogurt sales, unless you had some prior knowledge suggesting that these should be correlated.

If you ask a question whose framework is not plausible, you are likely to end up with an answer that's difficult to interpret or have confidence in. In the pepperoni-yogurt question, if you do find they are correlated, many questions are raised about the result itself: is it really correct?, why are these things correlated- is there another explanation?, and others. You can ensure that your question is grounded in a plausible framework by using your own knowledge of the subject area and doing a little research, which together can go a long way in terms of helping you sort out whether your question is grounded in a plausible framework.

The question, should also, of course, be **answerable**. Although perhaps this doesn't need stating, it's worth pointing out that some of the best questions aren't answerable - either because the data don't exist or there is no means of collecting the data because of lack of resources, feasibility, or ethical problems. For example, it is quite plausible that there are defects in the functioning of certain cells in the brain that cause autism, but it not possible to perform brain biopsies to collect live cells to study, which would be needed to answer this question.

Specificity is also an important characteristic of a good question. An example of a general question is: Is eating a healthier diet better for you? Working towards specificity will refine your question and directly inform what steps to take when you start looking at data. A more specific question emerges after asking yourself what you mean by a "healthier" diet and when you say something is "better for you"? The process of increasing the specificity should lead to a final, refined question such as: "Does eating at least 5 servings per day of fresh fruits and vegetables lead to fewer upper respiratory tract infections (colds)?" With this degree of specificity, your plan of attack is much clearer and the answer you will get at the end of the data analysis will be

more interpretable as you will either recommend or not recommend the specific action of eating at least 5 servings of fresh fruit and vegetables per day as a means of protecting against upper respiratory tract infections.

3.4 Translating a Question into a Data Problem

Another aspect to consider when you're developing your question is what will happen when you translate it into a data problem. Every question must be operationalized as a data analysis that leads to a result. Pausing to think through what the results of the data analysis would look like and how they might be interpreted is important as it can prevent you from wasting a lot of time embarking on an analysis whose result is not interpretable. Although we will discuss many examples of questions that lead to interpretable and meaningful results throughout the book, it may be easiest to start first by thinking about what sorts of questions *don't* lead to interpretable answers.

The typical type of question that does not meet this criterion is a question that uses inappropriate data. For example, your question may be whether taking a vitamin D supplement is associated with fewer headaches, and you plan on answering that question by using the number of times a person took a pain reliever as a marker of the number of headaches they had. You may find an association between taking vitamin D supplements and taking less pain reliever medication, but it won't be clear what the interpretation of this result is. In fact, it is possible that people who take vitamin D supplements also tend to be less likely to take other over-the-counter medications just because they are "medication avoidant," and not because they are actually

getting fewer headaches. It may also be that they are using less pain reliever medication because they have less joint pain, or other types of pain, but not fewer headaches. Another interpretation, of course, is that they are indeed having fewer headaches, but the problem is that you can't determine whether this is the correct interpretation or one of the other interpretations is correct. In essence, the problem with this question is that for a single possible answer, there are multiple interpretations. This scenario of multiple interpretations arises when at least one of the variables you use (in this case, pain reliever use) is not a good measure of the concept you are truly after (in this case, headaches). To head off this problem, you will want to make sure that the data available to answer your question provide reasonably specific measures of the factors required to answer your question.

A related problem that interferes with interpretation of results is confounding. Confounding is a potential problem when your question asks about the relationship between factors, such as taking vitamin D and frequency of headaches. A brief description of the concept of confounding is that it is present when a factor that you were not necessarily considering in your question is related to both your exposure of interest (in the example, taking vitamin D supplements) and your outcome of interest (taking pain reliever medication). For example, income could be a confounder, because it may be related to both taking vitamin D supplements and frequency of headaches, since people with higher income may tend to be more likely to take a supplement and less likely to have chronic health problems, such as headaches. Generally, as long as you have income data available to you, you will be able to adjust for this confounder and reduce the number of possible interpretations of the answer to your question. As you refine your question,

spend some time identifying the potential confounders and thinking about whether your dataset includes information about these potential confounders.

Another type of problem that can occur when inappropriate data are used is that the result is not interpretable because the underlying way in which the data were collected lead to a biased result. For example, imagine that you are using a dataset created from a survey of women who had had children. The survey includes information about whether their children had autism and whether they reported eating sushi while pregnant, and you see an association between report of eating sushi during pregnancy and having a child with autism. However, because women who have had a child with a health condition recall the exposures, such as raw fish, that occurred during pregnancy differently than those who have had healthy children, the observed association between sushi exposure and autism may just be the manifestation of a mother's tendency to focus more events during pregnancy when she has a child with a health condition. This is an example of recall bias, but there are many types of bias that can occur.

The other major bias to understand and consider when refining your question is selection bias, which occurs when the data your are analyzing were collected in such a way to inflate the proportion of people who have both characteristics above what exists in the general population. If a study advertised that it was a study about autism and diet during pregnancy, then it is qute possible that women who both ate raw fish and had a child with autism would be more likely to respond to the survey than those who had one of these conditions or neither of these conditions. This scenario would lead to a biased answer to your question about mothers' sushi intakes during pregnancy and risk of autism in their children. A good rule of thumb is that if

you are examining relationships between two factors, bias may be a problem if you are more (or less) likely to observe individuals with both factors because of how the population was selected, or how a person might recall the past when responding to a survey. There will be more discussion about bias in subsequent chapters on (Inference: A Primer and Interpreting Your Results), but the best time to consider its effects on your data analysis is when you are identifying the question you will answer and thinking about how you are going to answer the question with the data available to you.

3.5 Case Study

Joe works for a company that makes a variety of fitness tracking devices and apps and the name of the company is Fit on Fleek. Fit on Fleek's goal is, like many tech start-ups, to use the data they collect from users of their devices to do targeted marketing of various products. The product that they would like to market is a new one that they have just developed and not yet started selling, which is a sleep tracker and app that tracks various phases of sleep, such as REM sleep, and also provides advice for improving sleep. The sleep tracker is called Sleep on Fleek.

Joe's boss asks him to analyze the data that the company has on its users of their health tracking devices and apps to identify users for targeted Sleep on Fleek ads. Fit on Fleek has the following data from each of their customers: basic demographic information, number of steps walked per day, number of flights of stairs climbed per day, sedentary awake hours per day, hours of alertness per day, hours of drowsiness per day, and hours slept per day (but not more detailed information about sleep that the sleep tracker would track).

Although Joe has an objective in mind, gleaned from a

discussion with his boss, and he also knows what types of data are available in the Fit on Fleek database, he does not yet have a question. This scenario, in which Joe is given an objective, but not a question, is common, so Joe's first task is to translate the objective into a question, and this will take some back-and-forth communication with his boss. The approach to informal communications that take place during the process of the data analysis project, is covered in detail in the Communication chapter. After a few discussions, Joe settles on the following question: "Which Fit on Fleek users don't get enough sleep?" He and his boss agree that the customers who would be most likely to be interested in purchasing the Sleep on Fleek device and app are those who appear to have problems with sleep, and the easiest problem to track and probably the most common problem is not getting enough sleep.

You might think that since Joe now has a question, that he should move to download the data and start doing exploratory analyses, but there is a bit of work Joe still has to do to refine the question. The two main tasks Joe needs to tackle are: (1) to think through how his question does, or does not, meet the characteristics of a good question and (2) to determine what type of question he is asking so that he has a good understanding of what kinds of conclusions can (and cannot) be drawn when he has finished the data analysis.

Joe reviews the characteristics of a good question and his expecations are that his question has all of these characteristics: -of interest -not already answered -grounded in a plausible framework -answerable -specific

The answer that he will get at the end of his analysis (when he translates his question into a data problem) should also be interpretable.

He then thinks through what he knows about the question and in his judgment, the question is of interest as his boss expressed interest.

He also knows that the question could not have been answered already since his boss indicated that it had not and a review of the company's previous data analyses reveals no previous analysis designed to answer the question.

Next he assesses whether the question is grounded in a plausible framework. The question, Which Fit on Fleek users don't get enough sleep?, seems to be grounded in plausibility as it makes sense that people who get too little sleep would be interested in trying to improve their sleep by tracking it. However, Joe wonders whether the duration of sleep is the best marker for whether a person feels that they are getting inadequate sleep. He knows some people who regularly get little more than 5 hours of sleep a night and they seem satisfied with their sleep. Joe reaches out to a sleep medicine specialist and learns that a better measure of whether someone is affected by lack of sleep or poor quality sleep is daytime drowsiness. It turns out that his initial expectation that the question was grounded in a plausible framework did not match the information he received when he spoke with a content expert. So he revises his question so that it matches his expectations of plausibility and the revised question is: Which Fit on Fleek users have drowsiness during the day?

Joe pauses to make sure that this question is, indeed, answerable with the data he has available to him, and confirms that it is. He also pauses to think about the specificity of the question. He believes that it is specific, but goes through the exercise of discussing the question with colleagues to gather information about the specificity of the question. When he raises the idea of answering this question, his col-

leagues ask him many questions about what various parts of the question mean: what is meant by "which users"? Does this mean: What are the demographic characteristics of the users who have drowsiness? Or something else? What about "drowsiness during the day"? Should this phrase mean any drowsiness on any day? Or drowsiness lasting at least a certain amount of time on at least a certain number of days? The conversation with colleagues was very informative and indicated that the question was not very specific. Joe revises his question so that it is now specific: "Which demographic and health characteristics identify users who are most likely to have chronic drowsiness, defined as at least one episode of drowsiness at least every other day?"

Joe now moves on to thinking about what the possible answers to his questions are, and whether they will be interpretable. Joe identifies two possible outcomes of his analysis: (1) there are no characteristics that identify people who have chronic daytime drowsiness or (2) there are one or more characteristics that identify people with chronic daytime drowsiness. These two possibilities are interpretable and meaningful. For the first, Joe would conclude that targeting ads for the Sleep on Fleek tracker to people who are predicted to have chronic daytime drowsiness would not be possible, and for the second, he'd conclude that targeting the ad is possible, and he'd know which characteristic(s) to use to select people for the targeted ads.

Now that Joe has a good question in hand, after iterating through the 3 steps of the epicycle as he considered whether his question met each of the characteristics of a good question, the next step is for him to figure out what type of question he has. He goes through a thought process similar to the process he used for each of the characteristics above. He starts thinking that his question is an exploratory one, but as he reviews the description and examples of

an exploratory question, he realizes that although some parts of the analysis he will do to answer the question will be exploratory, ultimately his question is more than exploratory because its answer will predict which users are likely to have chronic daytime drowsiness, so his question is a prediction question. Identifing the type of question is very helpful because, along with a good question, he now knows that he needs to use a prediction approach in his analyses, in particular in the model building phase (see Formal Modeling chapter).

3.6 Concluding Thoughts

By now, you should be poised to apply the 3 steps of the epicycle to stating and refining a question. If you are a seasoned data analyst, much of this process may be automatic, so that you may not be entirely conscious of some parts of the process that lead you to a good question. Until you arrive at this point, this chapter can serve as a useful resource to you when you're faced with the task of developing a good question. In the next chapters, we will discuss what to do with the data now that you have good question in hand.

4. Exploratory Data Analysis

Exploratory data analysis is the process of exploring your data, and it typically includes examining the structure and components of your dataset, the distributions of individual variables, and the relationships between two or more variables. The most heavily relied upon tool for exploratory data analysis is visualizing data using a graphical representation of the data. Data visualization is arguably the most important tool for exploratory data analysis because the information conveyed by graphical display can be very quickly absorbed and because it is generally easy to recognize patterns in a graphical display.

There are several goals of exploratory data analysis, which are:

1. To determine if there are any problems with your dataset.
2. To determine whether the question you are asking can be answered by the data that you have.
3. To develop a sketch of the answer to your question.

Your application of exploratory data analysis will be guided by your question. The example question used in this chapter is: "Do counties in the eastern United States have higher ozone levels than counties in the western United States?" In this instance, you will explore the data to determine if there are problems with the dataset, and to determine if you can answer your question with this dataset.

To answer the question of course, you need ozone, county, and US region data. The next step is to use exploratory data analysis to begin to answer your question, which could include displaying boxplots of ozone by region of the US. At the end of exploratory data analysis, you should have a good sense of what the answer to your question is and be armed with sufficient information to move onto the next steps of data analysis.

It's important to note that here, again, the concept of the epicycle of analysis applies. You should have an expectation of what your dataset will look like and whether your question can be answered by the data you have. If the content and structure of the dataset doesn't match your expectation, then you will need to go back and figure out if your expectation was correct (but there was a problem with the data) or alternatively, your expectation was incorrect, so you cannot use the dataset to answer the question and will need to find another dataset.

You should also have some expectation of what the ozone levels will be as well as whether one region's ozone should be higher (or lower) than another's. As you move to step 3 of beginning to answer your question, you will again apply the epicycle of analysis so that if, for example, the ozone levels in the dataset are lower than what you expected from looking at previously published data, you will need to pause and figure out if there is an issue with your data or if your expectation was incorrect. Your expectation could be incorrect, for example, if your source of information for setting your expectation about ozone levels was data collected from 20 years ago (when levels were likely higher) or from only a single city in the U.S. We will go into more detail with the case study below, but this should give you an overview about the approach and goals of exploratory data analysis.

4.1 Exploratory Data Analysis Checklist: A Case Study

In this section we will run through an informal "checklist" of things to do when embarking on an exploratory data analysis. As a running example I will use a dataset on hourly ozone levels in the United States for the year 2014. The elements of the checklist are

1. Formulate your question
2. Read in your data
3. Check the packaging
4. Look at the top and the bottom of your data
5. Check your "n"s
6. Validate with at least one external data source
7. Make a plot
8. Try the easy solution first
9. Follow up

Throughout this example we will depict an ongoing analysis with R code and real data. Some of the examples and recommendations here will be specific to the R statistical analysis environment, but most should be applicable to any software system. Being fluent in R is not necessary for understanding the main ideas of the example. Feel free to skip over the code sections.

4.2 Formulate your question

Previously in this book, we have discussed the importance of properly formulating a question. Formulating a question can be a useful way to guide the exploratory data analysis

process and to limit the exponential number of paths that can be taken with any sizeable dataset. In particular, a *sharp* question or hypothesis can serve as a dimension reduction tool that can eliminate variables that are not immediately relevant to the question.

For example, in this chapter we will be looking at an air pollution dataset from the U.S. Environmental Protection Agency (EPA). A general question one could as is

> Are air pollution levels higher on the east coast than on the west coast?

But a more specific question might be

> Are hourly ozone levels on average higher in New York City than they are in Los Angeles?

Note that both questions may be of interest, and neither is right or wrong. But the first question requires looking at all pollutants across the entire east and west coasts, while the second question only requires looking at single pollutant in two cities.

It's usually a good idea to spend a few minutes to figure out what is the question you're *really* interested in, and narrow it down to be as specific as possible (without becoming uninteresting).

For this chapter, we will consider the following question:

> Do counties in the eastern United States have higher ozone levels than counties in the western United States?

As a side note, one of the most important questions you can answer with an exploratory data analysis is "Do I have the right data to answer this question?" Often this question is difficult ot answer at first, but can become more clear as we sort through and look at the data.

4.3 Read in your data

The next task in any exploratory data analysis is to read in some data. Sometimes the data will come in a very messy format and you'll need to do some cleaning. Other times, someone else will have cleaned up that data for you so you'll be spared the pain of having to do the cleaning.

We won't go through the pain of cleaning up a dataset here, not because it's not important, but rather because there's often not much generalizable knowledge to obtain from going through it. Every dataset has its unique quirks and so for now it's probably best to not get bogged down in the details.

Here we have a relatively clean dataset from the U.S. EPA on hourly ozone measurements in the entire U.S. for the year 2014. The data are available from the EPA's Air Quality System web page[1]. I've simply downloaded the zip file from the web site, unzipped the archive, and put the resulting file in a directory called "data". If you want to run this code you'll have to use the same directory structure.

The dataset is a comma-separated value (CSV) file, where each row of the file contains one hourly measurement of ozone at some location in the country.

NOTE: Running the code below may take a few minutes. There are 7,147,884 rows in the CSV file. If it takes too long,

[1] http://aqsdr1.epa.gov/aqsweb/aqstmp/airdata/download_files.html

you can read in a subset by specifying a value for the `n_max` argument to `read_csv()` that is greater than 0.

```
> library(readr)
> ozone <- read_csv("data/hourly_44201_2014.csv",
+                   col_types = "ccccinnccccccncnnccccccc")
```

The `readr` package by Hadley Wickham is a nice package for reading in flat files (like CSV files) *very* fast, or at least much faster than R's built-in functions. It makes some tradeoffs to obtain that speed, so these functions are not always appropriate, but they serve our purposes here.

The character string provided to the `col_types` argument specifies the class of each column in the dataset. Each letter represents the class of a column: "c" for character, "n" for numeric", and "i" for integer. No, I didn't magically know the classes of each column—I just looked quickly at the file to see what the column classes were. If there are too many columns, you can not specify `col_types` and `read_csv()` will try to figure it out for you.

Just as a convenience for later, we can rewrite the names of the columns to remove any spaces.

```
> names(ozone) <- make.names(names(ozone))
```

4.4 Check the Packaging

Have you ever gotten a present *before* the time when you were allowed to open it? Sure, we all have. The problem is that the present is wrapped, but you desperately want to know what's inside. What's a person to do in those circumstances? Well, you can shake the box a bit, maybe knock it with your knuckle to see if it makes a hollow sound, or even

weigh it to see how heavy it is. This is how you should think about your dataset before you start analyzing it for real.

Assuming you don't get any warnings or errors when reading in the dataset, you should now have an object in your workspace named `ozone`. It's usually a good idea to poke at that object a little bit before we break open the wrapping paper.

For example, you should check the number of rows

```
> nrow(ozone)
[1] 7147884
```

and columns.

```
> ncol(ozone)
[1] 23
```

Remember when we said there were 7,147,884 rows in the file? How does that match up with what we've read in? This dataset also has relatively few columns, so you might be able to check the original text file to see if the number of columns printed out (23) here matches the number of columns you see in the original file.

Another thing you can do in R is run `str()` on the dataset. This is usually a safe operation in the sense that even with a very large dataset, running `str()` shouldn't take too long.

```
> str(ozone)
Classes 'tbl_df', 'tbl' and 'data.frame':       7147884 obs. of  23 variab\
les:
 $ State.Code         : chr  "01" "01" "01" "01" ...
 $ County.Code        : chr  "003" "003" "003" "003" ...
 $ Site.Num           : chr  "0010" "0010" "0010" "0010" ...
 $ Parameter.Code     : chr  "44201" "44201" "44201" "44201" ...
 $ POC                : int  1 1 1 1 1 1 1 1 1 ...
 $ Latitude           : num  30.5 30.5 30.5 30.5 30.5 ...
 $ Longitude          : num  -87.9 -87.9 -87.9 -87.9 -87.9 ...
 $ Datum              : chr  "NAD83" "NAD83" "NAD83" "NAD83" ...
 $ Parameter.Name     : chr  "Ozone" "Ozone" "Ozone" "Ozone" ...
 $ Date.Local         : chr  "2014-03-01" "2014-03-01" "2014-03-01" \
"2014-03-01" ...
 $ Time.Local         : chr  "01:00" "02:00" "03:00" "04:00" ...
 $ Date.GMT           : chr  "2014-03-01" "2014-03-01" "2014-03-01" \
"2014-03-01" ...
 $ Time.GMT           : chr  "07:00" "08:00" "09:00" "10:00" ...
 $ Sample.Measurement : num  0.047 0.047 0.043 0.038 0.035 0.035 0.0\
34 0.037 0.044 0.046 ...
 $ Units.of.Measure   : chr  "Parts per million" "Parts per million"\
 "Parts per million" "Parts per million" ...
 $ MDL                : num  0.005 0.005 0.005 0.005 0.005 0.005 0.0\
05 0.005 0.005 0.005 ...
 $ Uncertainty        : num  NA NA NA NA NA NA NA NA NA NA ...
 $ Qualifier          : chr  "" "" "" "" ...
 $ Method.Type        : chr  "FEM" "FEM" "FEM" "FEM" ...
 $ Method.Name        : chr  "INSTRUMENTAL - ULTRA VIOLET" "INSTRUME\
NTAL - ULTRA VIOLET" "INSTRUMENTAL - ULTRA VIOLET" "INSTRUMENTAL - U\
LTRA VIOLET" ...
 $ State.Name         : chr  "Alabama" "Alabama" "Alabama" "Alabama"\
 ...
 $ County.Name        : chr  "Baldwin" "Baldwin" "Baldwin" "Baldwin"\
 ...
 $ Date.of.Last.Change: chr  "2014-06-30" "2014-06-30" "2014-06-30" \
"2014-06-30" ...
```

The output for `str()` duplicates some information that we already have, like the number of rows and columns. More importantly, you can examine the *classes* of each of the columns to make sure they are correctly specified (i.e. num-

bers are `numeric` and strings are `character`, etc.). Because we pre-specified all of the column classes in `read_csv()`, they all should match up with what we specified.

Often, with just these simple maneuvers, you can identify potential problems with the data before plunging in head first into a complicated data analysis.

4.5 Look at the Top and the Bottom of your Data

It's often useful to look at the "beginning" and "end" of a dataset right after you check the packaging. This lets you know if the data were read in properly, things are properly formatted, and that everthing is there. If your data are time series data, then make sure the dates at the beginning and end of the dataset match what you expect the beginning and ending time period to be.

In R, you can peek at the top and bottom of the data with the `head()` and `tail()` functions.

Here's the top.

```
> head(ozone[, c(6:7, 10)])
  Latitude Longitude Date.Local
1   30.498 -87.88141 2014-03-01
2   30.498 -87.88141 2014-03-01
3   30.498 -87.88141 2014-03-01
4   30.498 -87.88141 2014-03-01
5   30.498 -87.88141 2014-03-01
6   30.498 -87.88141 2014-03-01
```

For brevity I've only taken a few columns. And here's the bottom.

```
> tail(ozone[, c(6:7, 10)])
        Latitude Longitude Date.Local
7147879 18.17794 -65.91548 2014-09-30
7147880 18.17794 -65.91548 2014-09-30
7147881 18.17794 -65.91548 2014-09-30
7147882 18.17794 -65.91548 2014-09-30
7147883 18.17794 -65.91548 2014-09-30
7147884 18.17794 -65.91548 2014-09-30
```

The `tail()` function can be particularly useful because often there will be some problem reading the end of a dataset and if you don't check that specifically you'd never know. Sometimes there's weird formatting at the end or some extra comment lines that someone decided to stick at the end. This is particularly common with data that are exported from Microsoft Excel spreadsheets.

Make sure to check all the columns and verify that all of the data in each column looks the way it's supposed to look. This isn't a foolproof approach, because we're only looking at a few rows, but it's a decent start.

4.6 ABC: Always be Checking Your "n"s

In general, counting things is usually a good way to figure out if anything is wrong or not. In the simplest case, if you're expecting there to be 1,000 observations and it turns out there's only 20, you know something must have gone wrong somewhere. But there are other areas that you can check depending on your application. To do this properly, you need to identify some *landmarks* that can be used to check against your data. For example, if you are collecting data on people, such as in a survey or clinical trial, then you should know how many people there are in your study. That's something you should check in your dataset, to make

sure that you have data on all the people you thought you would have data on.

In this example, we will use the fact that the dataset purportedly contains *hourly* data for the *entire country*. These will be our two landmarks for comparison.

Here, we have hourly ozone data that comes from monitors across the country. The monitors should be monitoring continuously during the day, so all hours should be represented. We can take a look at the `Time.Local` variable to see what time measurements are recorded as being taken.

```
> head(table(ozone$Time.Local))

 00:00  00:01  01:00  01:02  02:00  02:03
288698      2 290871      2 283709      2
```

One thing we notice here is that while almost all measurements in the dataset are recorded as being taken on the hour, some are taken at slightly different times. Such a small number of readings are taken at these off times that we might not want to care. But it does seem a bit odd, so it might be worth a quick check.

We can take a look at which observations were measured at time "00:01".

```
> library(dplyr)
> filter(ozone, Time.Local == "13:14") %>%
+         select(State.Name, County.Name, Date.Local,
+                Time.Local, Sample.Measurement)
Source: local data frame [2 x 5]

  State.Name County.Name Date.Local Time.Local
       (chr)       (chr)      (chr)      (chr)
1   New York    Franklin 2014-09-30      13:14
2   New York    Franklin 2014-09-30      13:14
Variables not shown: Sample.Measurement (dbl)
```

We can see that it's a monitor in Franklin County, New York and that the measurements were taken on September 30, 2014. What if we just pulled all of the measurements taken at this monitor on this date?

```
> filter(ozone, State.Code == "36"
+        & County.Code == "033"
+        & Date.Local == "2014-09-30") %>%
+         select(Date.Local, Time.Local,
+                Sample.Measurement) %>%
+         as.data.frame
   Date.Local Time.Local Sample.Measurement
1  2014-09-30      00:01              0.011
2  2014-09-30      01:02              0.012
3  2014-09-30      02:03              0.012
4  2014-09-30      03:04              0.011
5  2014-09-30      04:05              0.011
6  2014-09-30      05:06              0.011
7  2014-09-30      06:07              0.010
8  2014-09-30      07:08              0.010
9  2014-09-30      08:09              0.010
10 2014-09-30      09:10              0.010
11 2014-09-30      10:11              0.010
12 2014-09-30      11:12              0.012
13 2014-09-30      12:13              0.011
14 2014-09-30      13:14              0.013
15 2014-09-30      14:15              0.016
16 2014-09-30      15:16              0.017
17 2014-09-30      16:17              0.017
```

```
18 2014-09-30      17:18            0.015
19 2014-09-30      18:19            0.017
20 2014-09-30      19:20            0.014
21 2014-09-30      20:21            0.014
22 2014-09-30      21:22            0.011
23 2014-09-30      22:23            0.010
24 2014-09-30      23:24            0.010
25 2014-09-30      00:01            0.010
26 2014-09-30      01:02            0.011
27 2014-09-30      02:03            0.011
28 2014-09-30      03:04            0.010
29 2014-09-30      04:05            0.010
30 2014-09-30      05:06            0.010
31 2014-09-30      06:07            0.009
32 2014-09-30      07:08            0.008
33 2014-09-30      08:09            0.009
34 2014-09-30      09:10            0.009
35 2014-09-30      10:11            0.009
36 2014-09-30      11:12            0.011
37 2014-09-30      12:13            0.010
38 2014-09-30      13:14            0.012
39 2014-09-30      14:15            0.015
40 2014-09-30      15:16            0.016
41 2014-09-30      16:17            0.016
42 2014-09-30      17:18            0.014
43 2014-09-30      18:19            0.016
44 2014-09-30      19:20            0.013
45 2014-09-30      20:21            0.013
46 2014-09-30      21:22            0.010
47 2014-09-30      22:23            0.009
48 2014-09-30      23:24            0.009
```

Now we can see that this monitor just records its values at odd times, rather than on the hour. It seems, from looking at the previous output, that this is the only monitor in the country that does this, so it's probably not something we should worry about.

Because the EPA monitors pollution across the country, there should be a good representation of states. Perhaps we should see exactly how many states are represented in this

dataset.

```
> select(ozone, State.Name) %>% unique %>% nrow
[1] 52
```

So it seems the representation is a bit too good—there are 52 states in the dataset, but only 50 states in the U.S.!

We can take a look at the unique elements of the `State.Name` variable to see what's going on.

```
> unique(ozone$State.Name)
 [1] "Alabama"              "Alaska"
 [3] "Arizona"              "Arkansas"
 [5] "California"           "Colorado"
 [7] "Connecticut"          "Delaware"
 [9] "District Of Columbia" "Florida"
[11] "Georgia"              "Hawaii"
[13] "Idaho"                "Illinois"
[15] "Indiana"              "Iowa"
[17] "Kansas"               "Kentucky"
[19] "Louisiana"            "Maine"
[21] "Maryland"             "Massachusetts"
[23] "Michigan"             "Minnesota"
[25] "Mississippi"          "Missouri"
[27] "Montana"              "Nebraska"
[29] "Nevada"               "New Hampshire"
[31] "New Jersey"           "New Mexico"
[33] "New York"             "North Carolina"
[35] "North Dakota"         "Ohio"
[37] "Oklahoma"             "Oregon"
[39] "Pennsylvania"         "Rhode Island"
[41] "South Carolina"       "South Dakota"
[43] "Tennessee"            "Texas"
[45] "Utah"                 "Vermont"
[47] "Virginia"             "Washington"
[49] "West Virginia"        "Wisconsin"
[51] "Wyoming"              "Puerto Rico"
```

Now we can see that Washington, D.C. (District of Columbia) and Puerto Rico are the "extra" states included in the dataset.

Since they are clearly part of the U.S. (but not official states of the union) that all seems okay.

This last bit of analysis made use of something we will discuss in the next section: external data. We knew that there are only 50 states in the U.S., so seeing 52 state names was an immediate trigger that something might be off. In this case, all was well, but validating your data with an external data source can be very useful. Which brings us to….

4.7 Validate With at Least One External Data Source

Making sure your data matches something outside of the dataset is very important. It allows you to ensure that the measurements are roughly in line with what they should be and it serves as a check on what *other* things might be wrong in your dataset. External validation can often be as simple as checking your data against a single number, as we will do here.

In the U.S. we have national ambient air quality standards, and for ozone, the current standard[2] set in 2008 is that the "annual fourth-highest daily maximum 8-hr concentration, averaged over 3 years" should not exceed 0.075 parts per million (ppm). The exact details of how to calculate this are not important for this analysis, but roughly speaking, the 8-hour average concentration should not be too much higher than 0.075 ppm (it can be higher because of the way the standard is worded).

Let's take a look at the hourly measurements of ozone.

[2] http://www.epa.gov/ttn/naaqs/standards/ozone/s_o3_history.html

```
> summary(ozone$Sample.Measurement)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.02000 0.03200 0.03123 0.04200 0.34900
```

From the summary we can see that the maximum hourly concentration is quite high (0.349 ppm) but that in general, the bulk of the distribution is far below 0.075.

We can get a bit more detail on the distribution by looking at deciles of the data.

```
> quantile(ozone$Sample.Measurement, seq(0, 1, 0.1))
   0%   10%   20%   30%   40%   50%   60%   70%
0.000 0.010 0.018 0.023 0.028 0.032 0.036 0.040
  80%   90%  100%
0.044 0.051 0.349
```

Knowing that the national standard for ozone is something like 0.075, we can see from the data that

- The data are at least of the right order of magnitude (i.e. the units are correct)
- The range of the distribution is roughly what we'd expect, given the regulation around ambient pollution levels
- Some hourly levels (less than 10%) are above 0.075 but this may be reasonable given the wording of the standard and the averaging involved.

4.8 Make a Plot

Making a plot to visualize your data is a good way to further your understanding of your question and your data. Plotting can occur at different stages of a data analysis. For

example, plotting may occur at the exploratory phase or later on in the presentation/communication phase.

There are two key reasons for making a plot of your data. They are *creating expectations* and *checking deviations from expectations*.

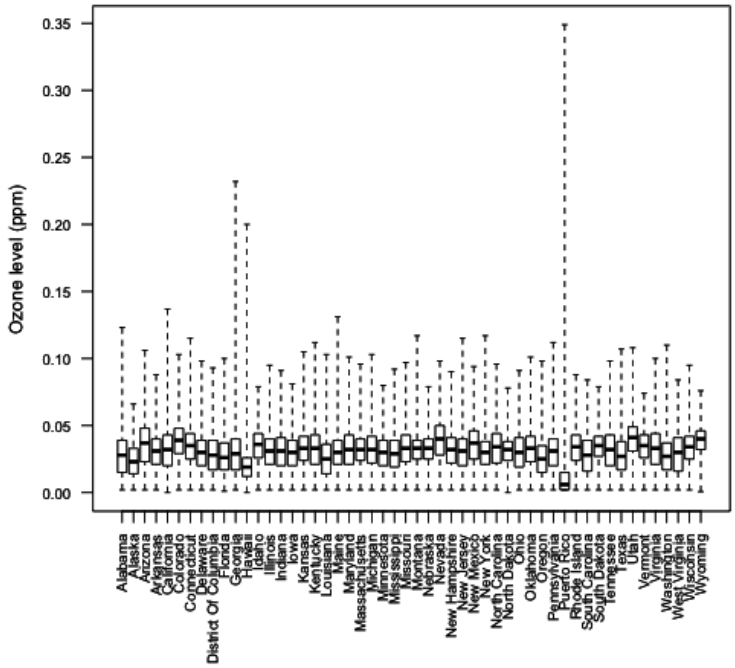
At the early stages of analysis, you may be equipped with a question/hypothesis, but you may have little sense of what is going on in the data. You may have peeked at some of it for sake of doing some sanity checks, but if your dataset is big enough, it will be difficult to simply look at all the data. So making some sort of plot, which serves as a summary, will be a useful tool for *setting expectations for what the data should look like*.

Once you have a good understanding of the data, a good question/hypothesis, and a set of expectations for what the data should say vis a vis your question, making a plot can be a useful tool to see how well the data match your expectations. Plots are particularly good at letting you see *deviations* from what you might expect. Tables typically are good at *summarizing* data by presenting things like means, medians, or other statistics. Plots, however, can show you those things, as well as show you things that are far from the mean or median, so you can check to see if something is *supposed* to be that far away. Often, what is obvious in a plot can be hidden away in a table.

Here's a simple boxplot[3] of the ozone data, with one boxplot for each state.

[3] https://en.wikipedia.org/wiki/Box_plot

```
> par(las = 2, mar = c(10, 4, 2, 2), cex.axis = 0.8)
> boxplot(Sample.Measurement ~ State.Name, ozone, range = 0, ylab = \
"Ozone level (ppm)")
```



Boxplot of ozone values by state

From the plot, we can see that for most states the data are within a pretty narrow range below 0.05 ppm. However, for Puerto Rico, we see that the typical values are very low, except for some extremely high values. Similarly, Georgia and Hawaii appear to experience an occasional very high value. These might be worth exploring further, depending on your question.

4.9 Try the Easy Solution First

Recall that our original question was

> Do counties in the eastern United States have higher ozone levels than counties in the western United States?

What's the simplest answer we could provide to this question? For the moment, don't worry about whether the answer is correct, but the point is how could you provide *prima facie* evidence for your hypothesis or question. You may refute that evidence later with deeper analysis, but this is the first pass. Importantly, if you do not find evidence of a signal in the data using just a simple plot or analysis, then often it is unlikely that you will find something using a more sophisticated analysis.

First, we need to define what we mean by "eastern" and "western". The simplest thing to do here is to simply divide the country into east and west using a specific longitude value. For now, we will use -100 as our cutoff. Any monitor with longitude less than -100 will be "west" and any monitor with longitude greater than or equal to -100 will be "east".

```
> library(maps)
> map("state")
> abline(v = -100, lwd = 3)
> text(-120, 30, "West")
> text(-75, 30, "East")
```

Map of East and West Regions

Here we create a new variable called `region` that we use to indicate whether a given measurement in the dataset was recorded in the "east" or the "west".

```
> ozone$region <- factor(ifelse(ozone$Longitude < -100, "west", "eas\
t"))
```

Now, we can make a simple summary of ozone levels in the east and west of the U.S. to see where levels tend to be higher.

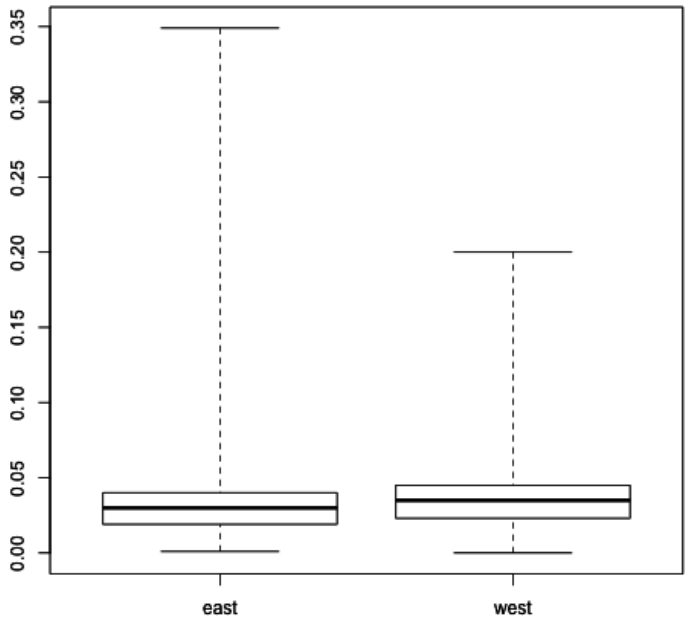
```
> group_by(ozone, region) %>%
+         summarize(mean = mean(Sample.Measurement, na.rm = TRUE),
+                   median = median(Sample.Measurement, na.rm = TRUE\
))
Source: local data frame [2 x 3]

  region       mean median
  (fctr)      (dbl)  (dbl)
1   east 0.02995250  0.030
2   west 0.03400735  0.035
```

Both the mean and the median ozone level are higher in the western U.S. than in the eastern U.S., by about 0.004 ppm.

We can also make a boxplot of the ozone in the two regions to see how they compare.

```
> boxplot(Sample.Measurement ~ region, ozone, range = 0)
```



Boxplot of Ozone for East and West Regions

We can see from the boxplots that the variability of ozone in the east tends to be a lot higher than the variability in the west.

Challenge Your Solution

The easy solution is nice because it is, well, easy, but you should never allow those results to hold the day. You should always be thinking of ways to challenge the results, especially if those results comport with your prior expectation.

Recall that previously we noticed that three states had some

unusually high values of ozone. We don't know if these values are real or not (for now, let's assume they are real), but it might be interesting to see if the same pattern of east/west holds up if we remove these states that have unusual activity.

```
> filter(ozone, State.Name != "Puerto Rico"
+        & State.Name != "Georgia"
+        & State.Name != "Hawaii") %>%
+         group_by(region) %>%
+         summarize(mean = mean(Sample.Measurement, na.rm = TRUE),
+                   median = median(Sample.Measurement, na.rm = TRUE\
))
Source: local data frame [2 x 3]

  region       mean median
  (fctr)      (dbl)  (dbl)
1   east 0.03003692  0.030
2   west 0.03406880  0.035
```

Indeed, it seems the pattern is the same even with those 3 states removed.

4.10 Follow-up Questions

In this chapter we've presented some simple steps to take when starting off on an exploratory analysis. The example analysis conducted in this chapter was far from perfect, but it got us thinking about the data and the question of interest. It also gave us a number of things to follow up on in case we continue to be interested in this question.

At this point it's useful to consider a few followup questions.

1. **Do you have the right data?** Sometimes at the conclusion of an exploratory data analysis, the conclusion is that the dataset is not really appropriate for this

question. In this case, the dataset seemed perfectly fine for answering the question of whether counties in the eastern U.S. have higher levels in the western U.S.

2. **Do you need other data?** While the data seemed adequate for answering the question posed, it's worth noting that the dataset only covered one year (2014). It may be worth examining whether the east/west pattern holds for other years, in which case we'd have to go out and obtain other data.
3. **Do you have the right question?** In this case, it's not clear that the question we tried to answer has immediate relevance, and the data didn't really indicate anything to increase the question's relevance. For example, it might have been more interesting to assess which counties were in violation of the national ambient air quality standard, because determining this could have regulatory implications. However, this is a much more complicated calculation to do, requiring data from at least 3 previous years.

The goal of exploratory data analysis is to get you thinking about your data and reasoning about your question. At this point, we can refine our question or collect new data, all in an iterative process to get at the truth.

5. Using Models to Explore Your Data

The objectives of this chapter are to describe what the concept of a model is more generally, to explain what the purpose of a model is with respect to a set of data, and last, to describe the process by which a data analyst creates, assesses, and refines a model. In a very general sense, a model is something we construct to help us understand the real world. A common example is the use of an animal which mimics a human disease to help us understand, and hopefully, prevent and/or treat the disease. The same concept applies to a set of data–presumably you are using the data to understand the real world.

In the world of politics a pollster has a dataset on a sample of likely voters and the pollster's job is to use this sample to predict the election outcome. The data analyst uses the polling data to construct a model to predict what will happen on election day. The process of building a model involves imposing a specific structure on the data and creating a summary of the data. In the polling data example, you may have thousands of observations, so the model is a mathematical equation that reflects the shape or pattern of the data, and the equation allows you to summarize the thousands of observations with, for example, one number, which might be the percentage of voters who will vote for your candidate. Right now, these last concepts may be a bit fuzzy, but they will become much clearer as you read on.

A statistical model serves two key purposes in a data analysis, which are to provide a *quantitative summary* of your

data and to impose a specific *structure* on the population from which the data were sampled. It's sometimes helpful to understand what a model is and why it can be useful through the illustration of extreme examples. The trivial "model" is simply **no model at all**.

Imagine you wanted to conduct a survey of 20 people to ask them how much they'd be willing to spend on a product you're developing. What is the goal of this survey? Presumably, if you're spending time and money developing a new product, you believe that there is a large *population* of people out there who are willing to buy this product. However, it's far too costly and complicated to ask everyone in that population what they'd be willing to pay. So you take a *sample* from that population to get a sense of what the population would pay.

One of us (Roger) recently published a book titled *R Programming for Data Science*[1]. Before the book was published, interested readers could submit their name and email address to the book's web site to be notified about the books publication. In addition, there was an option to specify how much they'd be willing to pay for the book. Below is a random sample of 20 response from people who volunteered this information.

```
25 20 15 5 30 7 5 10 12 40 30 30 10 25 10 20 10 10 25 5
```

Now suppose that someone asked you, "What do the data say?" One thing you could do is simply hand over the data—all 20 numbers. Since the dataset is not that big, it's not like this would be a huge burden. Ultimately, the answer to their question is in that dataset, but having all the data isn't a summary of any sort. Having all the data is important, but

[1] https://leanpub.com/rprogramming

is often not very useful. This is because the trivial model provides no reduction of the data.

The first key element of a statistical model is *data reduction*. The basic idea is you want to take the original set of numbers consisting of your dataset and transform them into a smaller set of numbers. If you originally started with 20 numbers, your model should produce a summary that is fewer than 20 numbers. The process of data reduction typically ends up with a *statistic*. Generally speaking, a statistic is any summary of the data. The sample mean, or average, is a statistic. So is the median, the standard deviation, the maximum, the minimum, and the range. Some statistics are more or less useful than others but they are all summaries of the data.

Perhaps the simplest data reduction you can produce is the mean, or the simple arithmetic average, of the data, which in this case is $17.2. Going from 20 numbers to 1 number is about as much reduction as you can do in this case, so it definitely satisfies the summary element of a model.

5.1 Models as Expectations

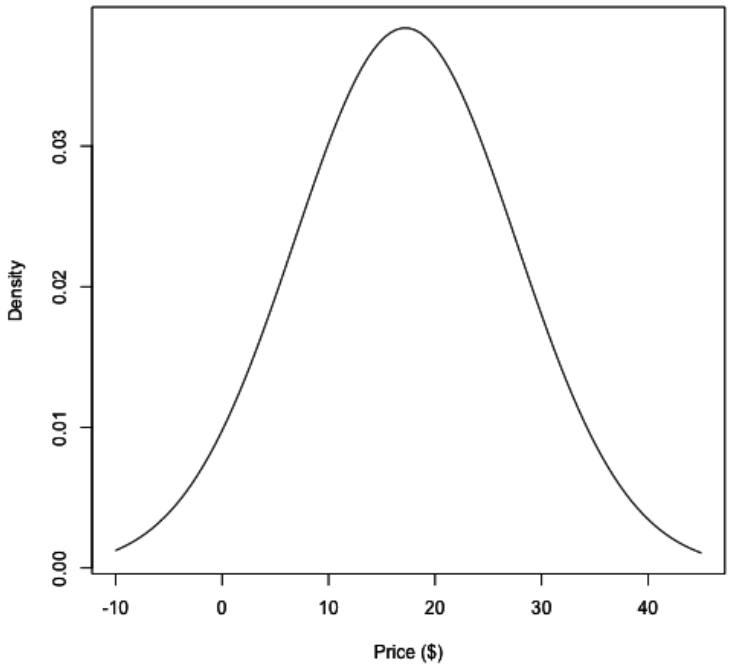
But a simple summary statistic, such as the mean of a set of numbers, is not enough to formulate a model. A statistical model must also impose some structure on the data. At its core, **a statistical model provides a description of how the world works and how the data were generated**. The model is essentially an *expectation* of the relationships between various factors in the real world and in your dataset. What makes a model a *statistical model* is that it allows for some randomness in generating the data.

Applying the normal model

Perhaps the most popular statistical model in the world is the Normal model. This model says that the randomness in a set of data can be explained by the Normal distribution, or a bell-shaped curve. The Normal distribution is fully specified by two parameters—the mean and the standard deviation.

Take the data that we described in the previous section—the amount of money 20 people were willing to pay for a hypothetical new product. The hope is that these 20 people are a representative sample of the entire population of people who might purchase this new product. If that's the case, then the information contained in the dataset can tell you something about everyone in the population.

To apply the Normal model to this dataset, we just need to calculate the mean and standard deviation. In this case, the mean is $17.2 and the standard deviation is $10.39. Given those parameters, our expectation under the Normal model is that the distribution of prices that people are willing to pay looks something like this.



Normal Model for Prices

According to the model, about 68% of the population would be willing to pay somewhere between $6.81 and $27.59 for this new product. Whether that is useful information or not depends on the specifics of the situation, which we will gloss over for the moment.

You can use the statistical model to answer more complex questions if you want. For example, suppose you wanted to know "What proportion of the population would be willing to pay more than $30 for this book?" Using the properties of the Normal distribution (and a little computational help from R), we can easily do this calculation.

```
pnorm(30, mean = mean(x), sd = sd(x), lower.tail = FALSE)
```

```
[1] 0.1089893
```

So about 11% of the population would be willing to pay more than $30 for the product. Again, whether this is useful to you depends on your specific goals.

Note that in the picture above there is one crucial thing that is missing—the data! That's not exactly true, because we used the data to draw the picture (to calculate the mean and standard deviation of the Normal distribution), but ultimately the data do not appear directly in the plot. In this case **we are using the Normal distribution to tell us what the population looks like**, not what the data look like.

The key point here is that we used the Normal distribution to setup the shape of the distribution that we *expect* the data to follow. The Normal distribution is our expectation for what the data should look like.

5.2 Comparing Model Expectations to Reality

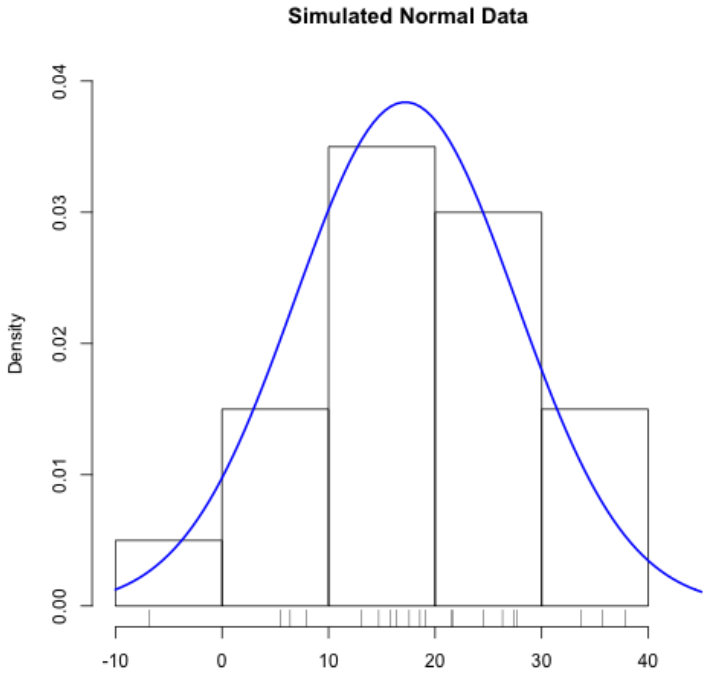
We may be very proud of developing our statistical model, but ultimately its usefulness will depend on how closely it mirrors the data we collect in the real world. How do we know if our expectations match with reality?

Drawing a fake picture

To begin with we can make some pictures, like a histogram of the data. But before we get to the data, let's figure out

what we *expect* to see from the data. If the population followed roughly a Normal distribution, and the data were a random sample from that population, then the distribution estimated by the histogram should look like the theoretical model provided by the Normal distribution.

In the picture below, I've simulated 20 data points from a Normal distribution and overlaid the theoretical Normal curve on top of the histogram.



Histogram of Simulated Normal Data

Notice how closely the histogram bars and the blue curve match. This is what we want to see with the data. If we see

this, then we might conclude that the Normal distribution is a **good statistical model for the data**.

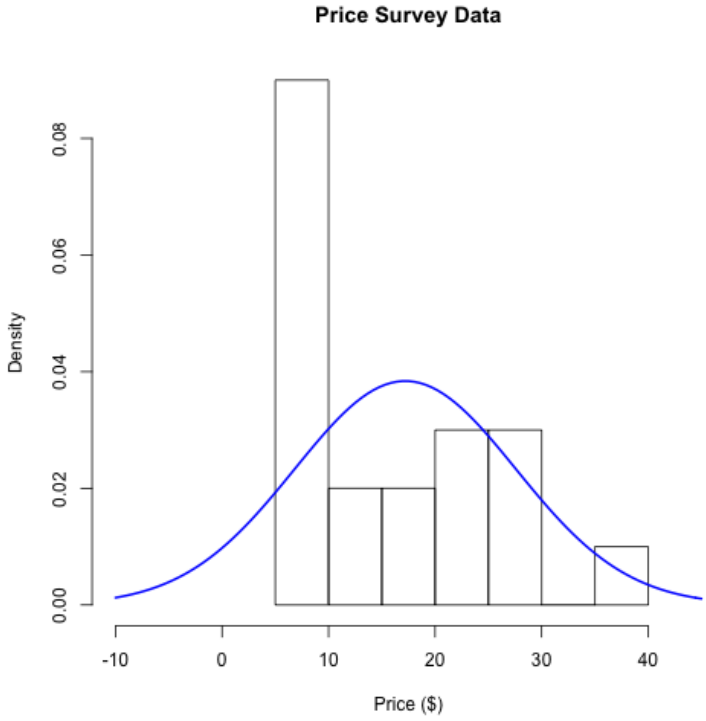
Simulating data from a hypothesized model, if possible, is a good way to setup expectations *before* you look at the data. Drawing a fake picture (even by hand, if you have to) can be a very useful tool for initiating discussions about the model and what we expect from reality.

For example, before we even look at the data, we might suspect the Normal model may not provide a perfect representation of the population. In particular, the Normal distribution allows for *negative* values, but we don't really expect that people will say that they'd be willing to pay negative dollars for a book.

So we have some evidence already that the Normal model may not be a *perfect* model, but no model is perfect. The question is does the statistical model provide a reasonable approximation that can be useful in some way?

The real picture

Here is a histogram of the data from the sample of 20 respondents. On top of the histogram, I've overlaid the Normal curve on top of the histogram of the 20 data points of the amount people say they are willing to pay for the book.



Histogram of Price Survey Data

What we would *expect* is that the histogram and the blue line should roughly follow each other. How do the model and reality compare?

At first glance, it looks like the histogram and the Normal distribution don't match very well. The histogram has a large spike around $10, a feature that is not present with the blue curve. Also, the Normal distribution allows for negative values on the left-hand side of the plot, but there are no data points in that region of the plot.

So far the data suggest that the Normal model isn't really a very good representation of the population, given the data

that we sampled from the population. It seems that the 20 people surveyed have strong preference for paying a price in the neighborhood of $10, while there are a few people willing to pay more than that. These features of the data are not well characterized by a Normal distribution.

5.3 Reacting to Data: Refining Our Expectations

Okay, so the model and the data don't match very well, as was indicated by the histogram above. So what do do? Well, we can either

1. Get a different model; or
2. Get different data

Or we could do both. What we do in response depends a little on our beliefs about the model and our understanding of the data collection process. If we felt strongly that the population of prices people would be willing to pay should follow a Normal distribution, then we might be less likely to make major modifications to the model. We might examine the data collection process to see if it perhaps led to some bias in the data. However, if the data collection process is sound, then we might be forced to re-examine our model for the population and see what could be changed. In this case, it's likely that our model is inappropriate, especially given that it's difficult to imagine a valid data collection process that might lead to negative values in the data (as the Normal distribution allows).

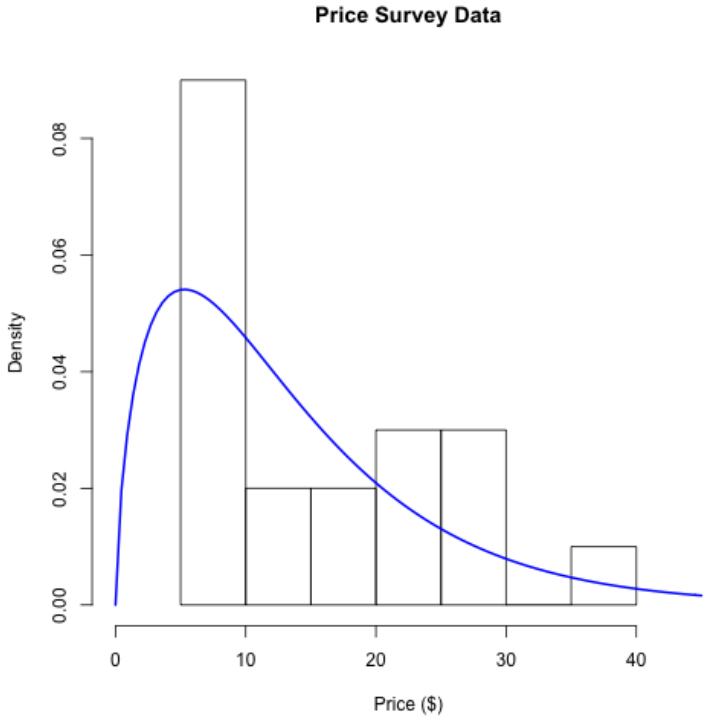
To close the loop here, we will choose a different statistical model to represent the population, the *Gamma distribution*. This distribution has the feature that it only allows positive

values, so it eliminates the problem we had with negative values with the Normal distribution.

Now, we should go back to the top of our iteration and do the following:

1. Develop expectations: Draw a fake picture—what do we expect to see before looking at the data?
2. Compare our expectations to the data
3. Refine our expectations, given what the data show

For your reference, here is a histgram of the same data with the Gamma distribution (estimated using the data) overlaid.



Price Survey Data with Gamma Distribution

How do the data match your expectations now?

You might ask what difference does it make which model I use to represent the population from which the data were generated? Well, for starters it might affect the kinds of predictions that you might make using the model. For example, recall before that were interested in what proportion of the population might be willing to pay at least $30 dollars for the book. Our new model says that only about 7% of the population would be willing to pay at least this amount (the Normal model claimed 11% would pay $30 or more). So different models can yield different predictions based on

the same data, which may impact decisions made down the road.

5.4 Examining Linear Relationships

It's common to look at data and try to understand linear relationships between variables of interest. The most common statistical technique to help with this task is *linear regression*. We can apply the principles discussed above—developing expectations, comparing our expectations to data, refining our expectations—to the application of linear regression as well.

For this example we'll look at a simple air quality dataset containing information about tropospheric ozone levels in New York City in the year 1999 for months of May through 1999. Here are the first few rows of the dataset.

```
     ozone     temp month
1 25.37262 55.33333     5
2 32.83333 57.66667     5
3 28.88667 56.66667     5
4 12.06854 56.66667     5
5 11.21920 63.66667     5
6 13.19110 60.00000     5
```

The data contain daily average levels of ozone (in parts per billion [pbb]) and temperature (in degrees Fahrenheit). One question of interest that might motivate the collection of this dataset is "How is ambient temperature related to ambient ozone levels in New York?"

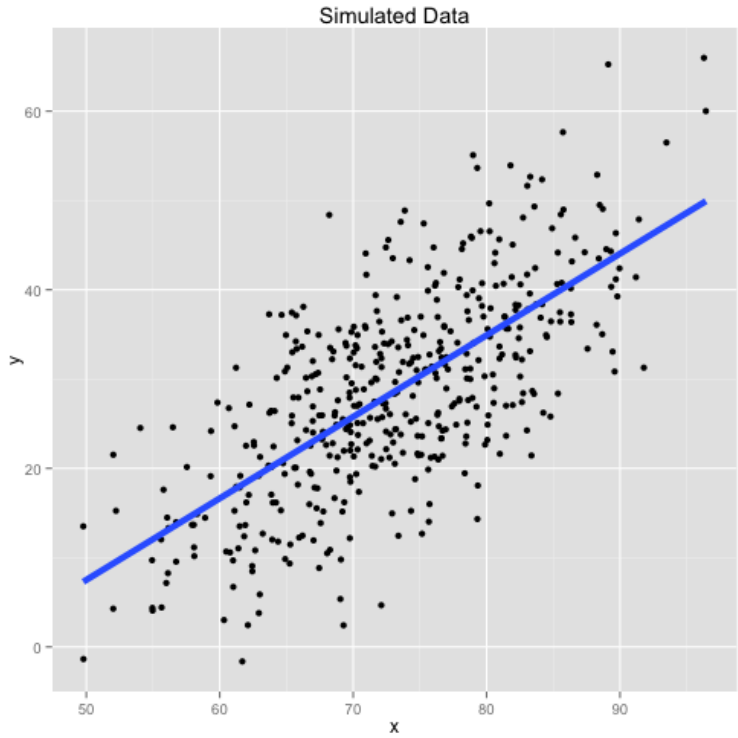
Expectations

After reading a little about ozone formation in the atmosphere[2], we know that the formation of ozone depends critically on the presence of sunlight. Sunlight is also related to temperature in the sense that on days where there is a lot of sunlight, we would expect the average temperature for that day to be higher. Cloudy days have both lower temperatures on average and less ozone. So there's reason to believe that on days with higher temperatures we would expect there to be higher ozone levels. This is an indirect relationship—we are using temperature here as essentially a proxy for the amount of sunlight.

The simplest model that we might formulate for characterizing the relationship between temperature and ozone is a *linear model*. This model says that as temperature increases, the amount of ozone in the atmosphere increases linearly with it. What do we expect this to look like?

We can simulate some data to make a *fake picture* of what the relationship between ozone and temperature should look like under a linear model. Here's a simple linear relationship along with the simulated data in a scatterplot.

[2] https://en.wikipedia.org/wiki/Tropospheric_ozone



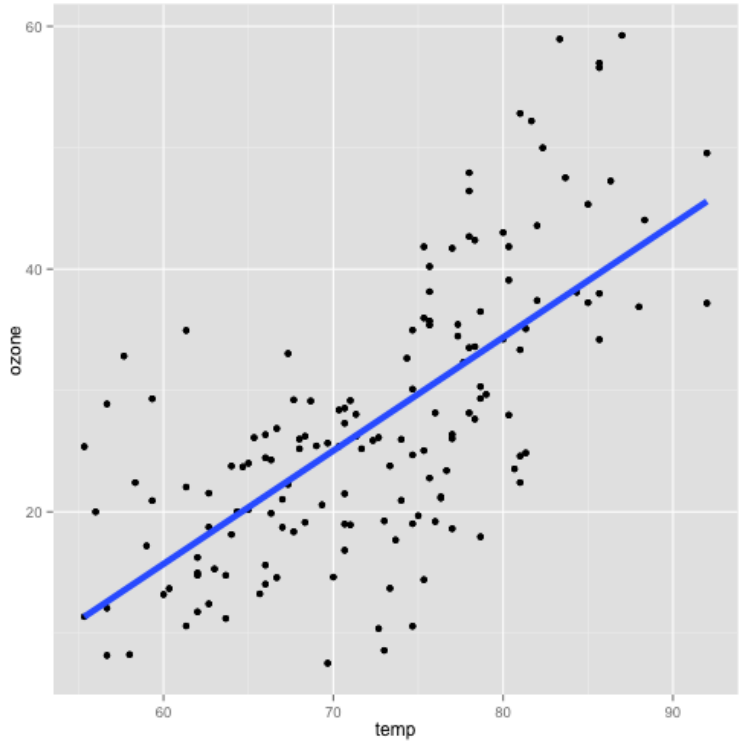
Simulated Data with a Linear Model

Note that if you choose any point on the blue line, there are roughly the same number of points above the line as there are below the line (this is also referred to as unbiased errors). Also, the points on the scatterplot appear to increase linearly as you move towards the right on the x-axis, even if there is a quite a bit of noise/scatter along the line.

If we are right about our linear model, and that is the model that characterizes the data and the relationship between ozone and temperature, then roughly speaking, this is the picture we should see when we plot the data.

Comparing expectations to data

Here is the picture of the actual ozone and temperature data in New York City for the year 1999. On top of the scatterplot of the data, we've plotted the fitted linear regression line estimated using the data.



Linear Model for Ozone and Temperature

How does this picture compare to the picture that you were expecting to see?

One thing is clear: There does appear to be an increasing trend in ozone as temperature increases, as we hypothe-

sized. However, there are a few deviations from the nice fake picture that we made above. The points don't appear to be evenly balanced around the blue regression line.

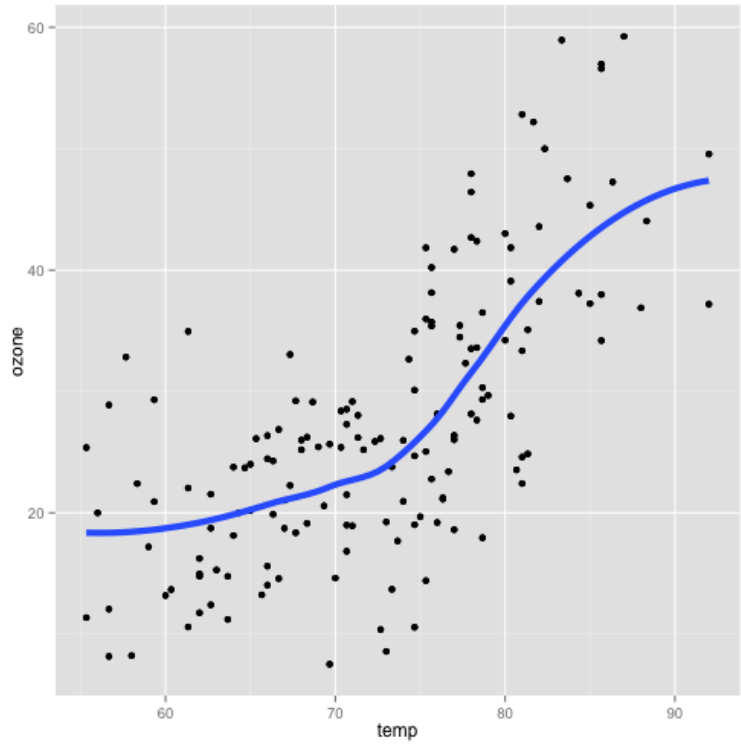
If you draw a vertical line around a temperature of 85 degrees, you notice that most of the points are above the line. Drawing a vertical line around 70 degrees shows that most of the points are below the line. This implies that at higher temperatures, our model is biased downward (it underestimates ozone) and at moderate temperatures our model is biased upwards. This isn't a great feature–in this situation we might prefer that our model is not biased anywhere.

Our simple linear regression model appears to capture the general increasing relationship between temperature and ozone, but it appears to be biased in certain ranges of temperature. It seems that there is room for improvement with this model if we want to better characterize the relationship between temperature and ozone in this dataset.

Refining expectations

From the picture above, it appears that the relationship between temperature and ozone may not be linear. Indeed, the data points suggest that maybe the relationship is flat up until about 70 degrees and then ozone levels increase rapidly with temperature after that. This suggest a *nonlinear* relationship between temperature and ozone.

The easiest way we can capture this revised expectation is with a smoother, in this case a loess smoother.



Loess Smoother for Ozone and Temperature

This plot shows a different picture–the relationship is slowly increasing up until about 75 degrees, and then sharply increases afterwards. Around 90 degrees, there's a suggestion that the relationship levels off again.

Smoothers (like loess) are useful tools because they quickly capture trends in a dataset without making any structural assumptions about the data. Essentially, they are an automated or computerized way to sketch a curve on to some data. However, smoothers rarely tell you anything about the mechanism of the relationship and so may be limited in that sense. In order to learn more about the relationship

between temperature and ozone, we may need to resort to a more detailed model than the simple linear model we had before.

5.5 When Do We Stop?

In the examples above, we completed one iteration of the data analysis process. In some cases, a single iteration may be sufficient, but in most real-life cases, you'll need to iterate at least a few times. From the examples above, there are still some things left to do:

- **Price Survey Data**: We ended the example by fitting a Gamma distribution model. But how does that fit the data? What would we expect from the data if they truly followed a Gamma distribution (we never made that plot)? Is there a better way to capture that spike in the distribution right around $10?
- **Ozone and Temperature**: The smoother suggested a nonlinear relationship between temperature and ozone, but what is the reason for this? Is the nonlinearity real or just a chance occurrence in the data? Is there a known physical process that explains the dramatic increase in ozone levels beyond a certain temperature and can we model that process?

Ultimately, you might be able to iterate over and over again. Every answer will usually raise more questions and require further digging into the data. When exactly do you stop the process then? Statistical theory suggests a number of different approaches to determining when a statistical model is "good enough" and fits the data well. This is not what we will discuss here, but rather we will discuss a few high-level criteria to determine when you might consider stopping the data analysis iteration.

Are you out of data?

Iterative data analysis will eventually begin to raise questions that simply cannot be answered with the data at hand. For example, in the ozone/temperature analysis, the modeling suggested that there isn't just a simple relationship between the two variables, that it may be nonlinear. But the data can't explain precisely why such a nonlinear relationship might exist (although they can suggest certain hypotheses). Also, you may need to collect additional data to determine whether what you observe is real or simply a fluke or statistical accident. Either way, you need to go back out into the world and collect new data. More data analysis is unlikely to bring these answers.

Another situation in which you may find yourself seeking out more data is when you've actually completed the data analysis and come to satisfactory results, usually some interesting finding. Then, it can be very important to try to *replicate* whatever you've found using a different, possibly independent, dataset. In the ozone/temperature example, if we concluded that there were a nonlinear relationship between temperature and ozone, our conclusion might be made more powerful if we could show that this relationship were present in other cities besides New York. Such independent confirmation can increase the strength of evidence and can play a powerful role in decision making.

Do you have enough evidence to make a decision?

Data analysis is often conducted in support of decision-making, whether in business, academia, government, or elsewhere, we often collect an analyze data to inform some sort of decision. It's important to realize that the analysis that you perform to get yourself to the point where you

can make a decision about something may be very different from the analysis you perform to achieve other goals, such as writing a report, publishing a paper, or putting out a finished product.

That's why it's important to always keep in mind the *purpose* of the data analysis as you go along because you may over- or under-invest resources in the analysis if the analysis is not attuned to the ultimate goal. The purpose of a data analysis may change over time and there may in fact be multiple parallel purposes. The question of whether you have enough evidence depends on factors specific to the application at hand and your personal situation with respect to costs and benefits. If you feel you do not have enough evidence to make a decision, it may be because you are out of data, or because you need to conduct more analysis.

Can you place your results in any larger context?

Another way to ask this question is "Do the results make some sort of sense?" Often, you can answer this question by searching available literature in your area or see if other people inside or outside your organization have come to a similar conclusion. If your analysis findings hew closely to what others have found, that may be a good thing, but it's not the only desirable outcome. Findings that are at odds with past results may lead down a path of new discovery. In either case, it's often difficult to come to the right answer without further investigation.

You have to be a bit careful with how you answer this question. Often, especially with very large and complex datasets, it's easy to come to a result that "makes sense" and conforms to our understanding of how a given process *should* work. In this situation, it's important to be hyper-critical of our findings and to challenge them as much as possible. In our ex-

perience, when the data very closely match our expectation, it can be a result of either mistakes or misunderstandings in the analysis or in the data collection process. It is critical to question every aspect of the analysis process to make sure everything was done appropriately.

If your results do *not* make sense, or the data do not match your expectation, then this is where things get interesting. You may simply have done something incorrectly in the analysis or the data collection. Chances are, that's exactly what happened. For every diamond in the rough, there are 99 pieces of coal. However, on the off-chance that you've discovered something unusual that others have not yet seen, you'll need to (a) make sure that the analysis was done properly and (b) replicate your findings in another dataset. Surprising results are usually met with much scrutiny and you'll need to be prepared to rigorously defend your work.

Ultimately, if your analysis leads you to a place where you can definitively answer the question "Do the results make sense?" then regardless of how you answer that question, you likely need to **stop your analysis and carefully check every part of it**.

Are you out of time?

This criterion seems arbitrary but nevertheless plays a big role in determining when to stop an analysis in practice. A related question might be "Are you out of money?" Ultimately, there will be both a time budget and a monetary budget that determines how many resources can be committed to a given analysis. Being aware of what these budgets are, even if you are not necessarily in control of them, can be important to managing a data analysis. In particular, you may need to argue for more resources and to persuade others to given them to you. In such a situation,

it's useful to know when to stop the data analysis iteration and prepare whatever results you may have obtained to date in order to present a coherent argument for continuation of the analysis.

5.6 Summary

Model building, like the entire process of data analysis itself, is an iterative process. Models are used to provide data reduction and to give you some insight into the population about which you are trying to make inference. It's important to first set your expectations for a how a model should characterize a dataset before you actually apply a model to data. Then you can check to see how your model conforms to your expectation. Often, there will be features of the dataset that do not conform to your model and you will have to either refine your model or examine the data collection process.

6. Inference: A Primer

Inference is one of many possible goals in data analysis and so it's worth discussing what exactly is the act of making inference. Recall previously we described one of the six types of questions you can ask in a data analysis is an **inferential** question. So what is inference?

In general, the goal of inference is to be able to make a statement about something that is *not observed*, and ideally to be able to characterize any uncertainty you have about that statement. Inference is difficult because of the difference between what you are able to observe and what you ultimately want to know.

6.1 Identify the population

The language of inference can change depending on the application, but most commonly, we refer to the things we cannot observe (but want to know about) as the **population** or as features of the population and the data that we observe as the **sample**. The goal is to use the sample to somehow make a statement about the population. In order to do this, we need to specify a few things.

Identifying the population is the most important task. If you cannot coherently identify or describe the population, then you cannot make an inference. Just stop. Once you've figured out what the population is and what feature of the population you want to make a statement about (e.g. the mean), then you can later translate that into a more specific

statement using a formal statistical model (covered later in this book).

6.2 Describe the sampling process

How did the data make its way from the population to your computer? Being able to describe this process is important for determining whether the data are useful for making inferences about features of the population. As an extreme example, if you are interested in the average age of women in a population, but your sampling process somehow is designed so that it only produces data on men, then you cannot use the data to make an inference about the average age of women. Understanding the sampling process is key to determining whether your sample is *representative* of the population of interest. Note that if you have difficulty describing the population, you will have difficulty describing the process of sampling data from the population. So describing the sampling process hinges on your ability to coherently describe the population.

6.3 Describe a model for the population

We need to have an abstract representation of how elements of the population are related to each other. Usually, this comes in the form of a statistical model that we can represent using mathematical notation. However, in more complex situations, we may resort to algorithmic representations that cannot be written down neatly on paper (many machine learning approaches have to be described this way). The simplest model might be a *simple linear model*, such as

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Here, x and y are features of the population and β_0 and β_1 describe the relationship between those features (i.e. are they positively or negatively associated?). The final element ε is a catch-all that is intended to capture all of the factors that contribute to the difference between the y and what we *expect* y to be, which is $\beta_0 + \beta_1 x$. It is this last part that makes the model a statistical model because we typically allow ε to be random.

Another characteristic that we typically need to make an assumption about is how different units in the population interact with each other. Typically, without any additional information, we will assume that the units in the population are *independent*, meaning that the measurements of one unit do not provide any information about the measurements on another unit. At best, this assumption is approximately true, but it can be a useful approximation. In some situations, such as when studying things that are closely connected in space or time, the assumption is clearly false, and we must resort to special modeling approaches to account for the lack of independence.

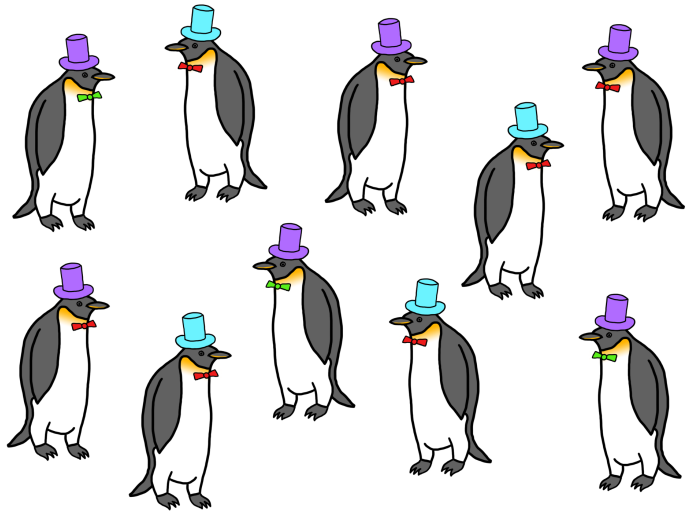
George Box, a statistician, once said that[1] "all models are wrong, but some are useful". It's likely that whatever model you devise for describing the features of a population, it is technically wrong. But you shouldn't be fixated on developing a *correct* model; rather you should identify a model that is useful to you and tells a story about the data and about the underlying processes that you are trying to study.

6.4 A Quick Example

Consider this group of penguins below (because penguins are awesome), each wearing either a purple or turquoise hat.

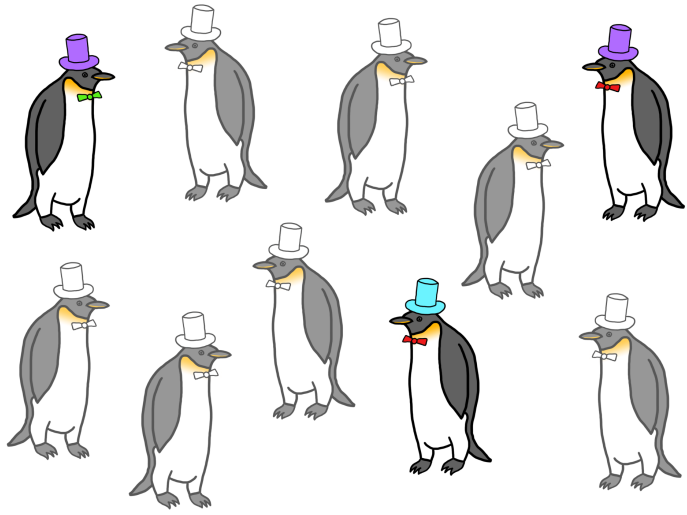
[1] https://en.wikipedia.org/wiki/All_models_are_wrong

There are a total of 10 penguins in this group. We'll call them the *population*.



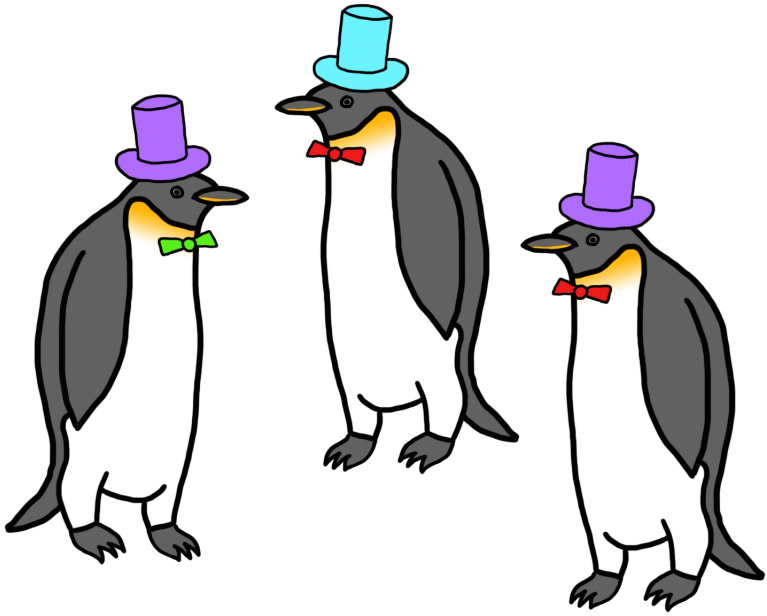
Population of Penguins with Turquoise and Purple Hats

Now suppose you wanted to know how what proportion of the *population* of penguins wears turqoise hats. But there's a catch–you don't have the time, money, or ability to take care of 10 penguins. Who does? You can only afford to take care of three penguins, so you randomly sample three of these 10 penguins.



Sample of 3 Penguins from Population

The key point is that you never observe the full population of penguins. Now what you end up with is your *dataset*, which contains only three penguins.



Dataset of Penguins

At this point an easy question to ask is "What proportion of the penguins *in my dataset* are wearing turquoise hats?". From the picture above, it's clear that 1/3 of the penguins are wearing turquoise hats. We have *no uncertainty* about that proportion because the data are sitting right in front of us.

The hard question to ask is "Based on the data I have, what proportion of the penguins in the *original population* are wearing turqoise hats?" At this point, we just have our sample of three penguins and do not observe the full population. What can we do? We need to make an *inference* about the population using the data we have on hand.

The three things that we need to do to make an inference are:

1. **Define the population**. Here, the population is the original 10 penguins from which we sampled our dataset of three penguins.
2. **Describe the sampling process**. We haven't explicitly mentioned this, but suppose for now that our "sampling process" consisted of taking the first three penguins that walked up to us.
3. **Describe a model for the population**. We will assume that the hats the penguins wear are *independent* of each other, so the fact that one penguin has a purple hat doesn't influence whether another penguin has a turquoise hat. Since we only want to estimate a simple proportion of penguins with turquoise hats, we don't need to make any more complex assumptions about how penguins relate to each other.

Given the three ingredients above, we might estimate the proportion of penguins with turquoise hats to be 1/3. How good of an estimate is this? Given that we know the truth here–2/5 of the penguins have turquoise hats in the population–we might ask whether 1/3 is a reasonable estimate or not. The answer to that question depends on a variety of factors that will be discussed in the next section.

6.5 Factors Affecting the Quality of Inference

The key factors affecting the quality of an inference you might make relate to violations in our thinking about the sampling process and the model for the population. Obviously, if we cannot coherently define the population, then any "inference" that we make to the population will be similarly vaguely defined.

A violation of our understanding of how the sampling process worked would result in our having collected data that did not represent the population in the way that we thought it would. This would affect our inference in that the inference we would make would apply not to the entire population, but to a specific selection of the population. This phenomenon is sometimes referred to as **selection bias** because the quantities that you estimate are biased toward the selection of the population that you *did* sample.

A violation of the model that we posit for the population could result in us estimating the wrong relationship between features of the population or underestimating the uncertainty of our estimates. For example, if it's true that penguins can influence what color hats other penguins wear, then that would violate the assumption of independence between penguins. This would result in an increase in the uncertainty of any estimates that we make from the data. In general, dependence between units in a population reduce the "effective sample size" of your dataset because the units you observe are not truly independent of each other and do not represent independent bits of information.

A final reason for a difference between our estimate from data and the truth in the population is **sampling variability**. Because we randomly sampled penguins from the population, it's likely that if we were to conduct the experiment again and sample another three penguins, we would get a different estimate of the number of penguins with turquoise hats, simply due to random variation in the sampling process. This would occur even if our description of the sampling process were accurate and our model for the population were perfect.

In most cases, differences between what we can estimate with data and what the truth is in the population can be

explained by a combination of all three factors. How big a role each plays in a given problem can be difficult to determine sometimes due to a lack of information, but it is usually worth putting some thought into each one of these factors and deciding which might be playing a dominant role. That way, one may be able to correct the problem, for example, in future studies or experiments.

6.6 Example: Apple Music Usage

On August 18, 2015, consumer market research firm MusicWatch released a study[2] about a new music service launched by Apple, Inc. called Apple Music. The service was a new streaming music service designed to give users streaming access to a large catalog of music for $9.99 per month. However, there was a free trial period that lasted for 3 months. At the time there was much speculation over how many users would ultimately continue to pay the $9.99 per month once the free trial ended.

MusicWatch's study claimed, among other things, that

> Among people who had tried Apple Music, 48 percent reported they are not currently using the service.

This would suggest that almost half of people who had signed up for the free trial period of Apple Music were not interested in using it further and would likely not pay for it once the trial ended. If it were true, it would be a blow to the newly launched service.

[2] http://www.businesswire.com/news/home/20150818005755/en#.VddbR7Scy6F

But how did MusicWatch arrive at its number? It claimed to have surveyed 5,000 people in its study. Shortly before the survey by MusicWatch was released, Apple claimed that about 11 million people had signed up for their new Apple Music service (because the service had just launched, everyone who had signed up was in the free trial period). Clearly, 5,000 people do not make up the entire population, so we have but a small sample of users.

What is the target that MusicWatch was trying to answer? It seems that they wanted to know the percentage of *all people who had signed up for Apple Music* that were still using the service. Because it would have been enormously expensive to survey all 11 million people, they had to resort to a much smaller sample of 5,000. Can they make inference about the entire population from the sample of 5,000?

Let's consider the three ingredients for inference:

1. **Population**: We are interested in the behavior of the entire Apple Music user base, which is approximately 11 million people, according to Apple.
2. **Sampling process**: It's not clear from the press release how the study was conducted and the data collected. It's likely this was a telephone survey and so people were randomly selected to be called and asked about their use of the service. Do you think this process led to a sample of respondents that is representative of the entire population of Apple Music users?
3. **Model for the population**: Given the relatively small size of the sample relative to the entire population, it's likely that the individuals in the survey could be thought of being independent of each other. In other words, it's unlikely that one respondent in the survey could have influenced another respondent.

If the sample is representative and the individuals are independent, we could use the number 48% as an estimate of the percentage in the population who no longer use the service. The press release from MusicWatch did not indicate any measure of uncertainty, so we don't know how reliable the number is.

Interestingly, soon after the MusicWatch survey was released, Apple released a statement to the publication *The Verge*, stating that 79% of users who had signed up were still using the service (i.e. only 21% had stopped using it, as opposed to 48% reported by MusicWatch). Now, the difference between Apple and MusicWatch is that Apple has easy access to the entire population of Apple Music users. If they want to know what percentage of the *population* of users is still using it, they simply need to count the number of active users of the service and divide by the total number of people who signed up. There is *no uncertainty* about that particular number, because no sampling was needed to estimate it (I assume Apple did not use sampling to estimate the percentage).

If we believe that Apple and MusicWatch were measuring the same thing in their analyses (and it's not clear that they were), then it would suggest that MusicWatch's estimate of the population percentage (48%) was quite far off from the true value (21%). What would explain this large difference?

1. **Random variation**. It's true that MusicWatch's survey was a small sample relative to the full population, but the sample was still big with 5,000 people. Furthermore, the analysis was fairly simple (just taking the proportion of users still using the service), so the uncertainty associated with that estimate is unlikely to be that large.

2. **Selection bias**. Recall that it's not clear how MusicWatch sampled its respondents, but it's possible that the way that they did it led them to capture a set of respondents who were less inclined to use Apple Music. Beyond this, we can't really say more without knowing the details of the survey process.
3. **Measurement differences**. One thing we don't know is how either MusicWatch or Apple defined "still using the service". You could imagine a variety of ways to determine whether a person was still using the service. You could ask "Have you used it in the last week?" or perhaps "Did you use it yesterday?" Responses to these questions would be quite different and would likely lead to different overall percentages of usage.
4. **Respondents are not independent**. It's possible that the survey respondents are not independent of each other. This would primiarily affect the uncertainty about the estimate, making it larger than we might expect if the respondents were all independent. However, since we do not know what MusicWatch's uncertainty about their estimate was in the first place, it's difficult to tell if dependence between respondents could play a role.

6.7 Populations Come in Many Forms

There are a variety of strategies that one can employ to setup a formal framework for making inferential statements. Often, there is literally a population of units (e.g. people, penguins, etc.) about which you want to make statements. In those cases it's clear where the uncertainty comes from (sampling from the population) and what exactly it is you're trying to estimate (some feature of the population). However, in other applications it might not be so clear what

exactly is the population and what exactly it is you're trying to estimate. In those cases, you'll have to be more explicity about defining the population because there may be more than one possibility.

Time series

Some processes are measured over time (every minute, every day, etc.). For example, we may be interested in analyzing data consisting of Apple's daily closing stock price for calendar year 2014. If we wanted to make an inference from this dataset, what would the population be? There are a few possibilities.

1. We might argue that the year 2014 was randomly sampled from the population of *all possible years* of data, so that inferences that we make apply to other years of the stock price.
2. We might say the Apple's stock represents a sample from the *entire stock market*, so that we can make inference about *other stocks* from this dataset.

Regardless of what you choose, it's important to make clear what population you are referring to before you attempt to make inference from the data.

Natural processes

Natural phenomena, such as earthquakes, fires, hurricanes, weather-related phenomena, and other events that occur in nature, often are recorded over time and space. For purely temporal measurements, we might define the population in the same way that we defined the population above with the time series example. However, we may have data that

is only measured in space. For example, we may have a map of the epicenters of all earthquakes that have occurred in an area. Then what is the population? One common approach is to say that there is an *unobserved stochastic process* that randomly drops earthquakes on to the area and that our data represent a random sample from this process. In that case, we are using the data to attempt to learn more about this unobserved process.

Data as population

One technique that is always possible, but not commonly used, is to treat the dataset as a population. In this case, there is no inference because there's no sampling. Because your dataset *is* the population, there's no uncertainty about any characteristic of the population. This may not sound like a useful strategy but there are circumstances where it can be used to answer important questions. In particular, there are times where we do not care about things outside the dataset.

For example, it is common in organizations to analyze salary data to make sure that women are not being paid less than men for comparable work or that there are not major imbalances between employees of different ethnic groups. In this setting, differences in salaries between different groups can be calculated in the dataset and one can see if the differences are large enough to be of concern. The point is that the data directly answer a question of interest, which is "Are there large salarly differences that need to be addressed?" In this case there's no need to make an inference about employees outside the organization (there are none, by definition) or to employees at other organizations over which you would not have any control. The dataset is the population and answers to any question regarding the population are in that dataset.

7. Formal Modeling

This chapter is typically the part of the statistics textbook or course where people tend to hit a wall. In particular, there's often a lot of math. Math is good, but gratuitous math is not good. We are not in favor of that.

It's important to realize that often it is useful to represent a model using mathematical notation because it is a compact notation and can be easy to interpret once you get used to it. Also, writing down a statistical model using mathematical notation, as opposed to just natural language, forces you to be precise in your description of the model and in your statement of what you are trying to accomplish, such as estimating a parameter.

7.1 What Are the Goals of Formal Modeling?

One key goal of formal modeling is to develop a precise specification of your question and how your data can be used to answer that question. Formal models allow you to identify clearly what you are trying to infer from data and what form the relationships between features of the population take. It can be difficult to achieve this kind of precision using words alone.

Parameters play an important role in many formal statistical models (in statistical language, these are known as *parametric statistical models*). These are numbers that we use to represent features or associations that exist in the population.

Because they represent population features, parameters are generally considered unknown, and our goal is to estimate them from the data we collect.

For example, suppose we want to assess the relationship between the number of ounces of soda consumed by a person per day and that person's BMI. The slope of a line that you might plot visualizing this relationship is the parameter you want to estimate to answer your question: "How much would BMI be expected to increase per each additional ounce of soda consumed?" More specifically, you are using a *linear regression model* to formulate this problem.

Another goal of formal modeling is to develop a rigorous framework with which you can challenge and test your primary results. At this point in your data analysis, you've stated and refined your question, you've explored the data visually and maybe conducted some exploratory modeling. The key thing is that you likely have a pretty good sense of what the answer to your question is, but maybe have some doubts about whether your findings will hold up under intense scrutiny. Assuming you are still interested in moving forward with your results, this is where formal modeling can play an important role.

7.2 General Framework

We can apply the basic epicycle of analysis to the formal modeling portion of data analysis. We still want to set expectations, collect information, and refine our expectations based on the data. In this setting, these three phases look as follows.

1. **Setting expectations**. Setting expectations comes in the form of developing a *primary model* that represents

your best sense of what provides the answer to your question. This model is chosen based on whatever information you have currently available.

2. **Collecting Information**. Once the primary model is set, we will want to create a set of secondary models that challenge the primary model in some way. We will discuss examples of what this means below.
3. **Revising expectations**. If our secondary models are successful in challenging our primary model and put the primary model's conclusions in some doubt, then we may need to adjust or modify the primary model to better reflect what we have learned from the secondary models.

Primary model

It's often useful to start with a *primary model*. This model will likely be derived from any exploratory analyses that you have already conducted and will serve as the lead candidate for something that succinctly summarizes your results and matches your expectations. It's important to realize that at any given moment in a data analysis, the primary model is *not necessarily the final model*. It is simply the model against which you will compare other secondary models. The process of comparing your model to other secondary models is often referred to as *sensitivity analyses*, because you are interested in seeing how sensitive your model is to changes, such as adding or deleting predictors or removing outliers in the data.

Through the iterative process of formal modeling, you may decide that a different model is better suited as the primary model. This is okay, and is all part of the process of setting expectations, collecting information, and refining expectations based on the data.

Secondary models

Once you have decided on a primary model, you will then typically develop a series of secondary models. The purpose of these models is to test the legitimacy and robustness of your primary model and potentially generate evidence against your primary model. If the secondary models are successful in generating evidence that refutes the conclusions of your primary model, then you may need to revisit the primary model and whether its conclusions are still reasonable.

7.3 Associational Analyses

Associational analyses are ones where we are looking at an association between two or more features in the presence of other potentially confounding factors. There are three classes of variables that are important to think about in an associational analysis.

1. **Outcome**. The outcome is the feature of your dataset that is thought to change along with your **key predictor**. Even if you are not asking a causal or mechanistic question, so you don't necessarily believe that the outcome *responds* to changes in the key predictor, an outcome still needs to be defined for most formal modeling approaches.
2. **Key predictor**. Often for associational analyses there is one key predictor of interest (there may be a few of them). We want to know how the outcome changes with this key predictor. However, our understanding of that relationship may be challenged by the presence of potential confounders.

3. **Potential confounders**. This is a large class of predictors that are both related to the key predictor and the outcome. It's important to have a good understanding what these are and whether they are available in your dataset. If a key confounder is not available in the dataset, sometimes there will be a proxy that is related to that key confounder that can be substituted instead.

Once you have identified these three classes of variables in your dataset, you can start to think about formal modeling in an associational setting.

The basic form of a model in an associational analysis will be

$$y = \alpha + \beta x + \gamma z + \varepsilon$$

where

- y is the outcome
- x is the key predictor
- z is a potential confounder
- ε is independent random error
- α is the intercept, i.e. the value y when $x = 0$ and $z = 0$
- β is the change in y associated with a 1-unit increase x, adjusting for z
- γ is the change in y associated with a 1-unit increase in z, adjusting for x

This is a linear model, and our primary interest is in estimating the coefficient β, which quantifies the relationship between the key predictor x and the outcome y.

Even though we will have to estimate α and γ as part of the process of estimating β, we do not really care about the

values of those α and γ. In the statistical literature, coefficients like α and γ are sometimes referred to as *nuisance parameters* because we have to use the data to estimate them to complete the model specification, but we do not actually care about their value.

The model shown above could be thought of as the primary model. There is a key predictor and one confounder in the model where it is perhaps well known that you should adjust for that confounder. This model may produce sensible results and follows what is generally known in the area.

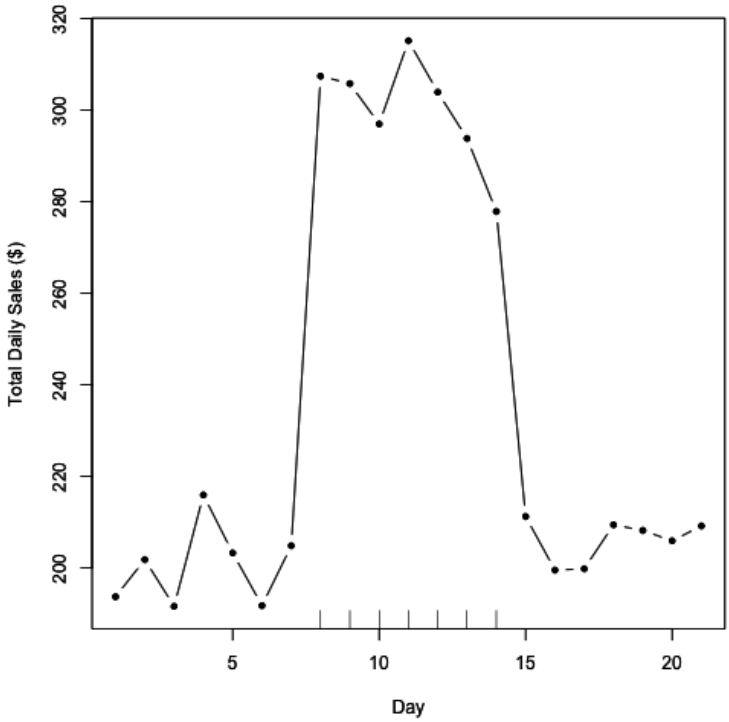
Example: Online advertising campaign

Suppose we are selling a new product on the web and we are interested in whether buying advertisements on Facebook helps to increase the sales of that product. To start, we might initiate a 1-week pilot advertising campaign on Facebook and gauge the success of that campaign. If it were successful, we might continue to buy ads for the product.

One simple approach might be to track daily sales before, during, and after the advertising campaign (note that there are more precise ways to do this with tracking URLs and Google Analytics, but let's leave that aside for now). Put simply, if the campaign were a week long, we could look at the week before, the week during, and the week after to see if there were any shift in the daily sales.

Expectations

In an ideal world, the data might look something like this.



Hypothetical Advertising Campaign

The tick marks on the x-axis indicate the period when the campaign was active. In this case, it's pretty obvious what effect the advertising campaign had on sales. Using just your eyes, it's possible to tell that the ad campaign added about $100 per day to total daily sales. Your primary model might look something like

$$y = \alpha + \beta x + \varepsilon$$

where y is total daily sales and x is and indicator of whether a given day fell during the ad campaign or not. The hypo-

thetical data for the plot above might look as follows.

```
      sales campaign day
1  193.7355        0   1
2  201.8364        0   2
3  191.6437        0   3
4  215.9528        0   4
5  203.2951        0   5
6  191.7953        0   6
7  204.8743        0   7
8  307.3832        1   8
9  305.7578        1   9
10 296.9461        1  10
11 315.1178        1  11
12 303.8984        1  12
13 293.7876        1  13
14 277.8530        1  14
15 211.2493        0  15
16 199.5507        0  16
17 199.8381        0  17
18 209.4384        0  18
19 208.2122        0  19
20 205.9390        0  20
21 209.1898        0  21
```

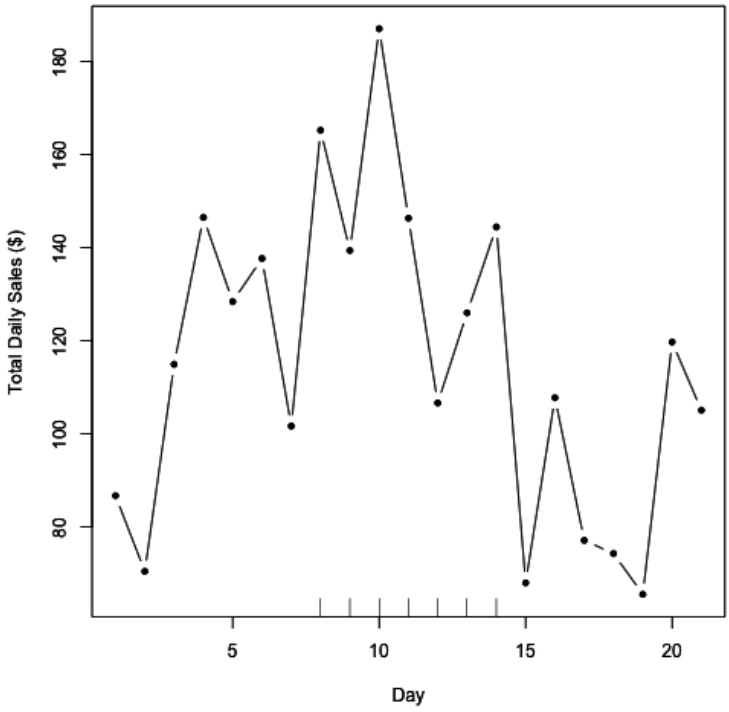
Given this data and the primary model above, we'd estimate β to be \$96.78, which is not far off from our original guess of \$100.

Setting Expectations. The discussion of this ideal scenario is important not because it's at all likely to occur, but rather because it instructs on what we would *expect* to see if the world operated according to a simpler framework and how we would analyze the data under those expectations.

More realistic data

Unfortunately, we rarely see data like the plot above. In reality, the effect sizes tend to be smaller, the noise tends

to be higher, and there tend to be other factors at play. Typically, the data will look something like this.



More Realistic Daily Sales Data

While it does appear that there is an increase in sales during the period of the ad campaign (indicated by the tick marks again), it's a bit difficult to argue that the increased sales are *caused* by the campaign. Indeed, in the days before the campaign starts, there appears to be a slight increase in sales. Is that by chance or are there other trends going on in the background? It's possible that there is a smooth background trend so that daily sales tend to go up and down throughout the month. Hence, even without the ad campaign in place,

it's possible we would have seen an increase in sales anyway. The question now is whether the ad campaign increased daily sales *on top of* this existing background trend.

Let's take our primary model, which just includes the outcome and the indicator of our ad campaign as a key predictor. Using that model we estimate β, the increase in daily sales due to the ad campaign, to be \$44.75.

However, suppose we incorporated a background trend into our model, so instead of our primary model, we fit the following.

$$y = \alpha + \beta x + \gamma_1 t + \gamma_2 t^2 + \varepsilon$$

where t now indicates the day number (i.e. $1, 2, \ldots, 21$). What we have done is add a quadratic function of t to the model to allow for some curvature in the trend (as opposed to a linear function that would only allow for a strictly increasing or decreasing pattern). Using this model we estimate β to be \$39.86, which is somewhat less than what the primary model estimated for β.

We can fit one final model, which allows for an even more flexible background trend–we use a 4th order polynomial to represent that trend. Although we might find our quadratic model to be sufficiently complex, the purpose of this last model is to just push the envelope a little bit to see how things change in more extreme circumstances. This model gives us an estimate of β of \$49.1, which is in fact larger than the estimate from our primary model.

At this point we have a primary model and two secondary models, which give somewhat different estimates of the association between our ad campaign and daily total sales.

Model	Features	Estimate for β
Model 1 (primary)	No confounders	$44.75
Model 2 (secondary)	Quadratic time trend	$39.86
Model 3 (secondary)	4th order time trend	$49.1

Evaluation

Determining where to go from here may depend on factors outside of the dataset. Some typical considerations are

1. **Effect size**. The three models present a range of estimates from $39.86 to $49.1. Is this a large range? It's possible that for your organization a range of this magnitude is not large enough to really make a difference and so all of the models might be considered equivalent. Or you might consider these 3 estimates to be significantly different from each other, in which case you might put more weight on one model over another. Another factor might be the cost of the advertising campaign, in which case you would be interested in the return on your investment in the ads. An increase in $39.86 per day might be worth it if the total ad cost were $10 per day, but maybe not if the cost were $20 per day. Then, you might need the increase in sales to be higher to make the campaign worthwhile. The point here is that there's some evidence from your formal model that the ad campaign might only increase your total daily sales by 39.86, however, other evidence says it might be higher. The question is whether you think it is worth the risk to buy more ads,

given the range of possibilities, or whether you think that even at the higher end, it's probably not worth it.

2. **Plausibility**. Although you may fit a series of models for the purposes of challenging your primary model, it may be the case that some models are more plausible than others, in terms of being close to whatever the "truth" about the population is. Here, the model with a quadratic trend seems plausible because it is capable of capturing a possible rise-and-fall pattern in the data, if one were present. The model with the 4th order polynomial is similarly capable of capturing this pattern, but seems overly complex for characterizing a simple pattern like that. Whether a model could be considered more or less plausible will depend on your knowledge of the subject matter and your ability to map real-world events to the mathematical formulation of the model. You may need to consult with other experts in this area to assess the plausibility of various models.
3. **Parsimony**. In the case where the different models all tell the same story (i.e. the estimates are β are close enough together to be considered "the same"), it's often preferable to choose the model that is simplest. There are two reasons for this. First, with a simpler model it can be easier to tell a story about what is going on in the data via the various parameters in the model. For example, it's easier to explain a linear trend than it is to explain an exponential trend. Second, simpler models, from a statistical perspective, are more "efficient", so that they make better use of the data per parameter that is being estimated. Complexity in a statistical model generally refers to the number of parameters in the model–in this example the primary model has 2 parameters, whereas the most complex model has 6 parameters. If no model

produces better results than another, we might prefer a model that only contains 2 parameters because it is simpler to describe and is more parsimonious. If the primary and secondary models produce significant differences, then might choose a parsimonious model over a more complex model, but not if the more complex model tells a more compelling story.

7.4 Prediction Analyses

In the previous section we described associational analyses, where the goal is to see if a key predictor x and an outcome y are associated. But sometimes the goal is to use all of the information available to you to predict y. Furthermore, it doesn't matter if the variables would be considered unrelated in a causal way to the outcome you want to predict because the objective is prediction, not developing an understanding about the relationships between features.

With prediction models, we have outcome variables–features about which we would like to make predictions–but we typically do not make a distinction between "key predictors" and other predictors. In most cases, any predictor that might be of use in predicting the outcome would be considered in an analysis and might, *a priori*, be given equal weight in terms of its importance in predicting the outcome. Prediction analyses will often leave it to the prediction algorithm to determine the importance of each predictor and to determine the functional form of the model.

For many prediction analyses it is not possible to literally write down the model that is being used to predict because it cannot be represented using standard mathematical notation. Many modern prediction routines are structured as algorithms or procedures that take inputs and trans-

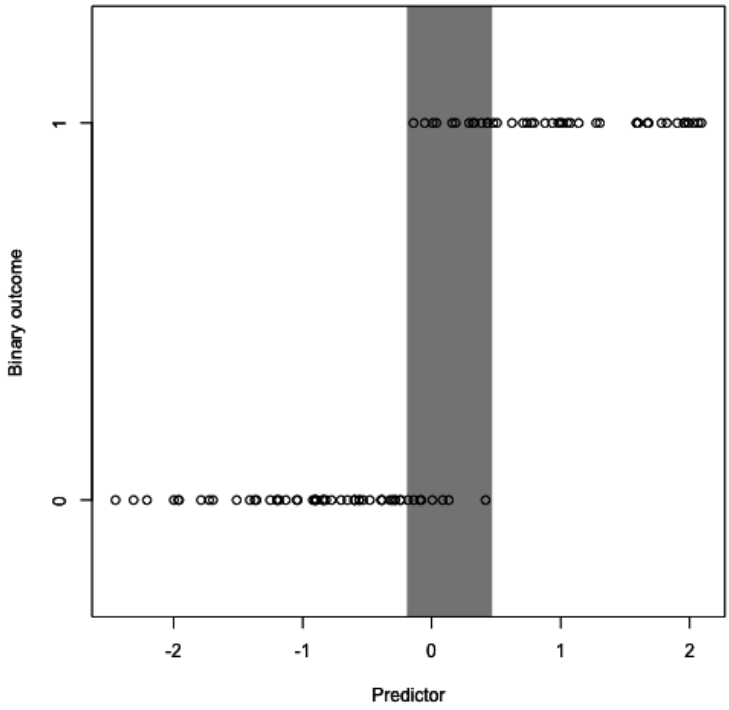
form them into outputs. The path that the inputs take to be transformed into outputs may be highly nonlinear and predictors may interact with other predictors on the way. Typically, there are no parameters of interest that we try to estimate–in fact many algorithmic procedures do not have any estimable parameters at all.

The key thing to remember with prediction analyses is that we usually do not care about the specific details of the model. In most cases, as long as the method "works", is reproducible, and produces good predictions with minimal error, then we have achieved our goals.

With prediction analyses, the precise type of analysis you do depends on the nature of the outcome (as it does with all analyses). Prediction problems typically come in the form of a **classification problem** where the outcome is binary. In some cases the outcome can take more than two levels, but the binary case is by far the most common. In this section, we will focus on the binary classification problem.

Expectations

What's the ideal scenario in a prediction problem? Generally, what we want is a predictor, or a set of predictors, to produce *good separation* in the outcome. Here's an example of a single predictor producing reasonable separation in a binary outcome.



Ideal Classification Scenario

The outcome takes values of 0 and 1, while the predictor is continuous and takes values between roughly -2 and 2. The gray zone indicated in the plot highlights the area where values of the predictor can take on values of 0 or 1. To the right of the gray area you'll notice that the value of the outcome is always 1 and to the left of the gray area the value of the outcome is always 0. In prediction problems, it's this gray area where we have the most uncertainty about the outcome, given the value of the predictor.

The goal of most prediction problems to identify a set of predictors that minimizes the size of that gray area in

the plot above. Counterintuitvely, it is common to identify predictors (particularly categorical ones) that *perfectly separate* the outcome, so that the gray area is reduced to zero. However, such situations typically indicate a degenerate problem that is not of much interest or even a mistake in the data. For example, a continuous variable that has been dichotomized will be perfectly separated by its continuous counterpart. It is a common mistake to include the continuous version as a predictor in the model and the dichotomous version as the outcome. In real-world data, you may see near perfect separation when measuring features or characteristics that are known to be linked to each other mechanistically or through some deterministic process. For example, if the outcome were an indicator of a person's potential to get ovarian cancer, then the person's sex might be a very good predictor, but it's not likely to be one of great interest to us.

Real world data

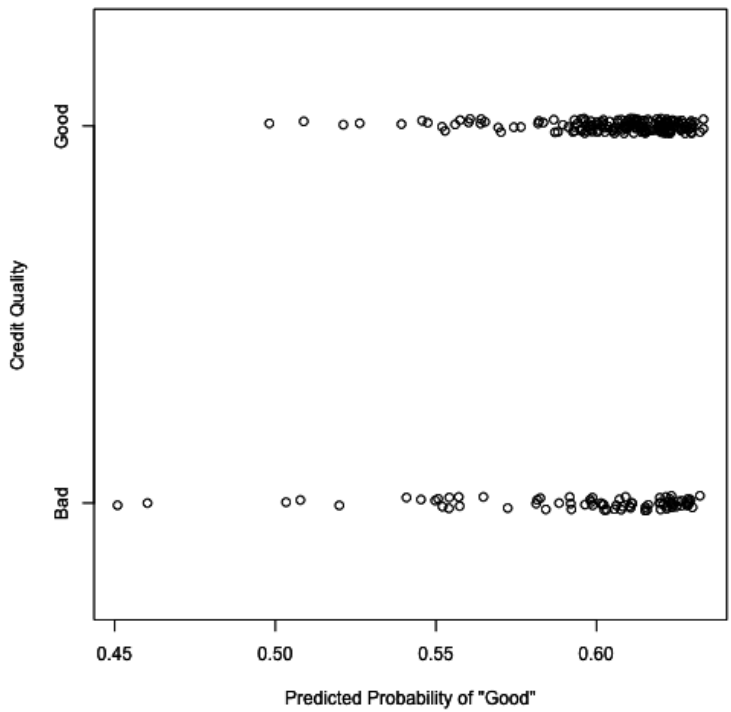
For this example we will use data on the credit worthiness of individuals. The dataset is taken from the UCI Machine Learning Repository[1]. The dataset classifies individuals into "Good" or "Bad" credit risks and includes a variety of predictors that may predict credit worthiness. There are a total of 1000 observations in the dataset and 62 features. For the purpose of this exposition, we omit the code for this example, but the code files can be obtained from the Book's web site.

The first thing we do for a prediction problem is to divide the data into a training dataset and a testing dataset. The training dataset is for developing and fitting the model and the testing dataset is for evaluating our fitted model and

[1] https://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data)

estimating its error rate. In this example we use a random 75% of the observations to serve as the training dataset. The remaining 25% will serve as the test dataset.

After fitting the model to the training dataset we can compute the predicted probabilities of being having "Good" credit from the test dataset. We plot those predicted probabilities on the x-axis along with each individuals true credit status on the y-axis below. (The y-axis coordinates have been randomly jittered to show some more detail.)



Prediction vs. Truth

Here we can see that there isn't quite the good separation

that we saw in the ideal scenario. Across the range of predicted probabilities, there are individuals with both "Good" and "Bad" credit. This suggests that the prediction algorithm that we have employed perhaps is having difficulty finding a good combination of features that can separate people with good and bad credit risk.

We can compute some summary statistics about the prediction algorithm below.

```
Confusion Matrix and Statistics

          Reference
Prediction Bad Good
      Bad    2    1
      Good  73  174

               Accuracy : 0.704
                 95% CI : (0.6432, 0.7599)
    No Information Rate : 0.7
    P-Value [Acc > NIR] : 0.4762

                  Kappa : 0.0289
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99429
            Specificity : 0.02667
         Pos Pred Value : 0.70445
         Neg Pred Value : 0.66667
             Prevalence : 0.70000
         Detection Rate : 0.69600
   Detection Prevalence : 0.98800
      Balanced Accuracy : 0.51048

       'Positive' Class : Good
```

We can see that the accuracy is about 70%, which is not great for most prediction algorithms. In particular, the algorithm's specificity is very poor, meaning that if you are a

"Bad" credit risk, the probability that you will be classified as such is only about 2.6%.

Evaluation

For prediction problems, deciding on the next step after initial model fitting can depend on a few factors.

1. **Prediction quality**. Is the model's accuracy good enough for your purposes? This depends on the ultimate goal and the risks associated with subsequent actions. For medical applications, where the outcome might be the presence of a disease, we may want to have a high sensitivity, so that if you genuinely have the disease, the algorithm will detect it. That way we can get you into treatment quickly. However, if the treatment is very painful, perhaps with many side effects, then we might actually prefer a high specificity, which would ensure that we don't mistakenly treat someone who *doesn't* have the disease. For financial applications, like the credit worthiness example used here, there may be asymmetric costs associated with mistaking good credit for bad versus mistaking bad credit for good.
2. **Model tuning**. A hallmark of prediction algorithms is their many tuning parameters. Sometimes these parameters can have large effects on prediction quality if they are changed and so it is important to be informed of the impact of tuning parameters for whatever algorithm you use. There is no prediction algorithm for which a single set of tuning parameters works well for all problems. Most likely, for the initial model fit, you will use "default" parameters, but these defaults may not be sufficient for your purposes. Fiddling with the tuning parameters may greatly change the quality of

your predictions. It's very important that you document the values of these tuning parameters so that the analysis can be reproduced in the future.

3. **Availability of Other Data**. Many prediction algorithms are quite good at exploring the structure of large and complex datasets and identifying a structure that can best predict your outcome. If you find that your model is not working well, even after some adjustment of tuning parameters, it is likely that you need additional data to improve your prediction.

7.5 Summary

Formal modeling is typically the most technical aspect of data analysis, and its purpose is to precisely lay out what is the goal of the analysis and to provide a rigorous framework for challenging your findings and for testing your assumptions. The approach that you take can vary depending primarily on whether your question is fundamentally about estimating an association develoing a good prediction.

8. Inference vs. Prediction: Implications for Modeling Strategy

Understanding whether you're answering an inferential question versus a prediction question is an important concept because the type of question you're answering can greatly influence the modeling strategy you pursue. If you do not clearly understand which type of question you are asking, you may end up using the wrong type of modeling approach and ultimately make the wrong conclusions from your data. The purpose of this chapter is to show you what can happen when you confuse one question for another.

The key things to remember are

1. For **inferential questions** the goal is typically to estimate an association between a predictor of interest and the outcome. There is usually only a handful of predictors of interest (or even just one), however there are typically many potential confounding variables to consider. They key goal of modeling is to estimate an association while making sure you appropriately adjust for any potential confounders. Often, sensitivity analyses are conducted to see if associations of interest are robust to different sets of confounders.
2. For **prediction questions** the goal is to identify a model that *best predicts* the outcome. Typically we do not place any *a priori* importance on the predictors, so long as they are good at predicting the outcome.

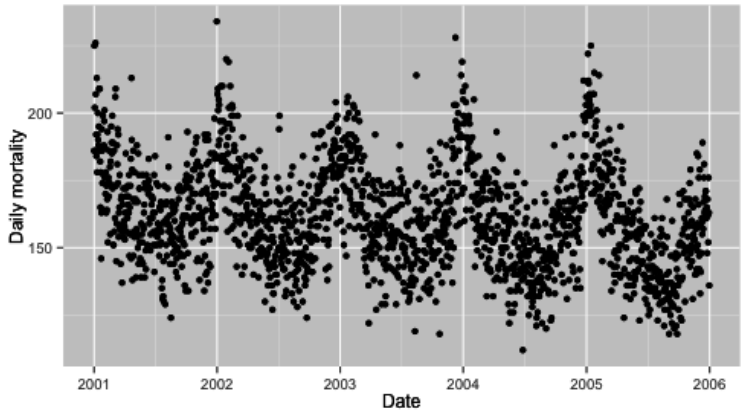
There is no notion of "confounder" or "predictors of interest" because all predictors are potentially useful for predicting the outcome. Also, we often do not care about "how the model works" or telling a detailed story about the predictors. The key goal is to develop a model with good prediction skill and to estimate a reasonable error rate from the data.

8.1 Air Pollution and Mortality in New York City

The following example shows how different types of questions and corresponding modeling approaches can lead to different conclusions. The example uses air pollution and mortality data for New York City. The data were originally used as part of the National Morbidity, Mortality, and Air Pollution Study[1] (NMMAPS).

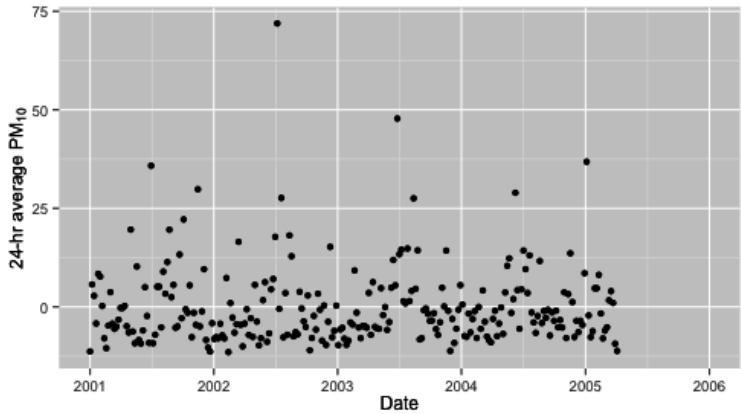
Below is a plot of the daily mortality from all causes for the years 2001–2005.

[1] http://www.ihapss.jhsph.edu



Daily Mortality in New York City, 2001–2005

And here is a plot of 24-hour average levels of particulate matter with aerodynamic diameter less than or equal to 10 microns (PM10).



Daily PM10 in New York City, 2001–2005

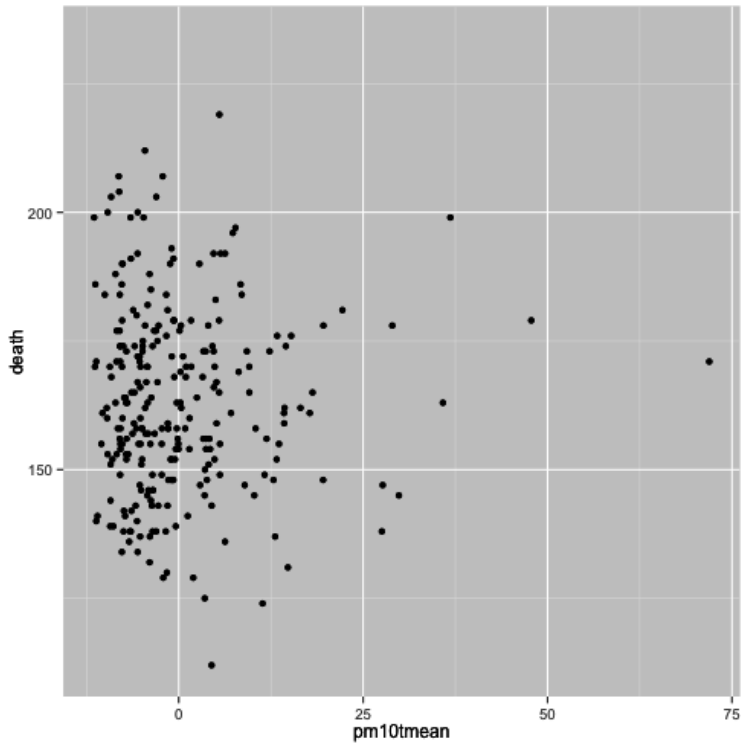
Note that there are many fewer points on the plot above than there were on the plot of the mortality data. This is because PM10 is not measured everyday. Also note that

there are negative values in the PM10 plot–this is because the PM10 data were mean-subtracted. In general, negative values of PM10 are not possible.

8.2 Inferring an Association

The first approach we will take will be to ask "Is there an association between daily 24-hour average PM10 levels and daily mortality?" This is an inferential question and we are attempting to estimate an association. In addition, for this question, we know there are a number of potential confounders that we will have to deal with.

Let's take a look at the bivariate association between PM10 and mortality. Here is a scatterplot of the two variables.



PM10 and Mortality in New York City

There doesn't appear to be much going on there, and a simple linear regression model of the log of daily mortality and PM10 seems to confirm that.

```
                Estimate   Std. Error      t value  Pr(>|t|)
(Intercept) 5.08884308354 0.0069353779 733.75138151 0.0000000
pm10tmean   0.00004033446 0.0006913941   0.05833786 0.9535247
```

In the table of coefficients above, the coefficient for `pm10tmean` is quite small and its standard error is relatively large. Effectively, this estimate of the association is zero.

However, we know quite a bit about both PM10 and daily mortality, and one things we do know is that *season* plays a large role in both variables. In particular, we know that mortality tends to be higher in the winter and lower in the summer. PM10 tends to show the reverse pattern, being higher in the summer and lower in the winter. Because season is related to *both* PM10 and mortality, it is a good candidate for a confounder and it would make sense to adjust for it in the model.

Here are the results for a second model, which includes both PM10 and season. Season is included as an indicator variable with 4 levels.

```
               Estimate   Std. Error    t value     Pr(>|t|)
(Intercept)  5.166484285 0.0112629532 458.714886 0.000000e+00
seasonQ2    -0.109271301 0.0166902948  -6.546996 3.209291e-10
seasonQ3    -0.155503242 0.0169729148  -9.161847 1.736346e-17
seasonQ4    -0.060317619 0.0167189714  -3.607735 3.716291e-04
pm10tmean    0.001499111 0.0006156902   2.434847 1.558453e-02
```

Notice now that the `pm10tmean` coefficient is quite a bit larger than before and its `t value` is large, suggesting a strong association. How is this possible?

It turns out that we have a classic example of Simpson's Paradox[2] here. The overall relationship between P10 and mortality is null, but when we account for the seasonal variation in both mortality and PM10, the association is positive. The surprising result comes from the opposite ways in which season is related to mortality and PM10.

So far we have accounted for season, but there are other potential confounders. In particular, weather variables, such as temperature and dew point temperature, are also both related to PM10 formation and mortality.

[2] https://en.wikipedia.org/wiki/Simpson%27s_paradox

In the following model we include temperature (`tmpd`) and dew point temperature (`dptp`). We also include the `date` variable in case there are any long-term trends that need to be accounted for.

```
                Estimate    Std. Error    t value     Pr(>|t|)
(Intercept)  5.62066568788 0.16471183741 34.1242365 1.851690e-96
date        -0.00002984198 0.00001315212 -2.2689856 2.411521e-02
seasonQ2    -0.05805970053 0.02299356287 -2.5250415 1.218288e-02
seasonQ3    -0.07655519887 0.02904104658 -2.6361033 8.906912e-03
seasonQ4    -0.03154694305 0.01832712585 -1.7213252 8.641910e-02
tmpd        -0.00295931276 0.00128835065 -2.2969777 2.244054e-02
dptp         0.00068342228 0.00103489541  0.6603781 5.096144e-01
pm10tmean    0.00237049992 0.00065856022  3.5995189 3.837886e-04
```

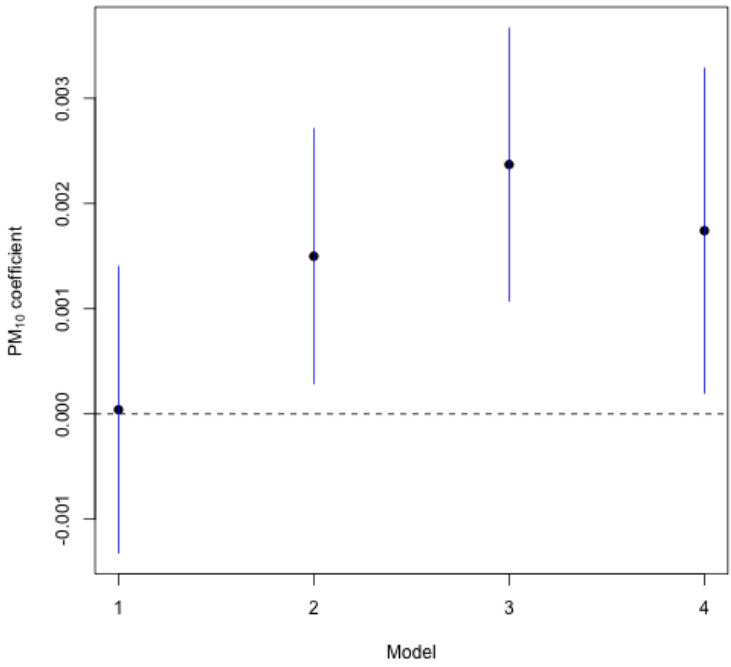
Notice that the `pm10tmean` coefficient is even bigger than it was in the previous model. There appears to still be an association between PM10 and mortality. The effect size is small, but we will discuss that later.

Finally, another class of potential confounders includes other pollutants. Before we place blame on PM10 as a harmful pollutant, it's important that we examine whether there might be another pollutant that can explain what we're observing. NO2 is a good candidate because it share some of the same sources as PM10 and is known to be related to mortality. Let's see what happens when we include that in the model.

```
                Estimate    Std. Error    t value     Pr(>|t|)
(Intercept)  5.61378604085 0.16440280471 34.1465345 2.548704e-96
date        -0.00002973484 0.00001312231 -2.2659756 2.430503e-02
seasonQ2    -0.05143935218 0.02338034983 -2.2001105 2.871069e-02
seasonQ3    -0.06569205605 0.02990520457 -2.1966764 2.895825e-02
seasonQ4    -0.02750381423 0.01849165119 -1.4873639 1.381739e-01
tmpd        -0.00296833498 0.00128542535 -2.3092239 2.174371e-02
dptp         0.00070306996 0.00103262057  0.6808599 4.965877e-01
no2tmean     0.00126556418 0.00086229169  1.4676753 1.434444e-01
pm10tmean    0.00174189857 0.00078432327  2.2208937 2.725117e-02
```

Notice in the table of coefficients that the `no2tmean` coefficient is similar in magnitude to the `pm10tmean` coefficient, although its `t value` is not as large. The `pm10tmean` coefficient appears to be statistically significant, but it is somewhat smaller in magnitude now.

Below is a plot of the PM10 coefficient from all four of the models that we tried.



Association Between PM10 and Mortality Under Different Models

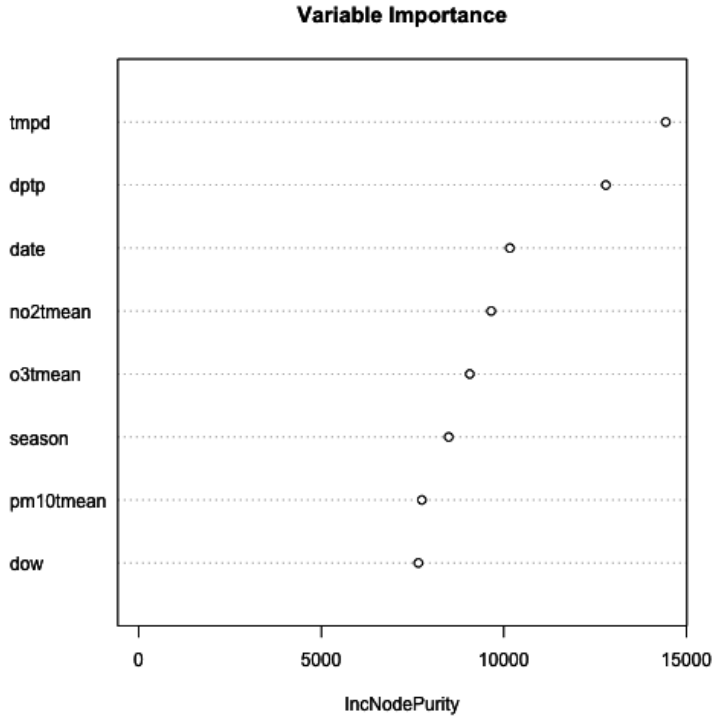
With the exception of Model 1, which did not account for any potential confounders, there appears to be a positive association between PM10 and mortality across Models 2–4. What this means and what we should do about it depends on what our ultimate goal is and we do not discuss that in detail here. It's notable that the effect size is generally small, expecially compared to some of the other predictors in the model. However, it's also worth noting that presumably, everyone in New York City breaths, and so a small effect could have a large impact.

8.3 Predicting the Outcome

Another strategy we could have taken is to ask "What best predicts mortality in New York City?" This is clearly a prediction question and we can use the data on hand to build a model. Here, we will use the random forests[3] modeling strategy, which is a machine learning approach that performs well when there are a large number of predictors. One type of output we can obtain from the random forest procedure is a measure of *variable importance*. Roughly speaking, this measure indicates how important a given variable is to improving the prediction skill of the model.

Below is a variable importance plot, which is obtained after fitting a random forest model. Larger values on the x-axis indicate greater importance.

[3] https://en.wikipedia.org/wiki/Random_forest



Random Forest Variable Importance Plot for Predicting Mortality

Notice that the variable `pm10tmean` comes near the bottom of the list in terms of importance. That is because it does not contribute much to predicting the outcome, mortality. Recall in the previous section that the effect size appeared to be small, meaning that it didn't really explain much variability in mortality. Predictors like temperature and dew point temperature are more useful as predictors of daily mortality. Even NO2 is a better preditor than PM10.

However, just because PM10 is not a strong predictor of mortality doesn't mean that it does not have a relevant association with mortality. Given the tradeoffs that have

to made when developing a prediction model, PM10 is not high on the list of predictors that we would include–we simply cannot include every predictor.

8.4 Summary

In any data analysis, you want to ask yourself "Am I asking an inferential question or a prediction question?" This should be cleared up *before* any data are analyzed, as the answer to the question can guide the entire modeling strategy. In the example here, if we had decided on a prediction approach, we might have erroneously thought that PM10 was not relevant to mortality. However, the inferential approach suggested a statistically significant association with mortality. Framing the question right, and applying the appropriate modeling strategy, can play a large role in the kinds of conclusions you draw from the data.

9. Interpreting Your Results

Although we have dedicated an entire chapter to interpreting the results of a data analysis, interpretation is actually happening continuously throughout an analysis. Experienced data analysts may not even be aware of how often they are interpreting their findings because it has become second nature to them.

By now the 3 step epicyclic process of: setting expecations, collecting information (data), and then matching expectations to the data, should be very familiar to you, so you will recognize that the third step, matching expectations to the data, is itself interpretation. In some ways, we have addressed the topic of interpreting results throughout the book. However, it deserves its own chapter because there is much more to interpretation than matching expectations to results and because it is, in and of itself, a major step of data analysis. Because interpretation happens most deliberately after completing your primary and supportive analyses, including formal modeling, but before communicating results, we have placed this chapter in between these respective chapters.

9.1 Principles of Interpretation

There are several principles of interpreting results that we will illustrate in this chapter. These principles are:

1. Revisit your original question

2. Start with the primary statistical model to get your bearings and focus on the nature of the result rather than on a binary assessment of the result (e.g. statistically significant or not). The nature of the result includes three characteristics: its directionality, magnitude, and uncertainty. Uncertainty is an assessment of how likely the result was obtained by chance.
3. Develop an overall interpretation based on (a) the totality of your analysis and (b) the context of what is already known about the subject matter.
4. Consider the implications, which will guide you in determining what action(s), if any, sbould be taken as a result of the answer to your question.

It is important to note that the epicycle of analysis also applies to interpretation. At each of the steps of interpretation, you should have expectations prior to performing the step, and then see if the result of the step matches your expectations. Your expectations are based on what you learned in the process of your exploratory data analysis and formal modeling, and when your interpretation doesn't match your expectations, then you will need to determine whether they don't match because your expectations are incorrect or your interpretation is incorrect. Even though you may be on one of the last steps of data analysis when you are formally interpreting your results, you may need to go back to exploratory data analysis or modeling to match expectations to data.

9.2 Case Study: Non-diet Soda Consumption and Body Mass Index

It is probably easiest to see the principles of interpretation in action in order to learn how to apply them to your own

data analysis, so we will use a case study to illustrate each of the principles.

Revisit the Question

The first principle is reminding yourself of your original question. This may seem like a flippant statement, but it is not uncommon for people to lose their way as they go through the process of exploratory analysis and formal modeling. This typically happens when a data analyst wanders too far off course pursuing an incidental finding that appears in the process of exploratory data analysis or formal modeling. Then the final model(s) provide an answer to another question that popped up during the analyses rather than the original question.

Reminding yourself of your question also serves to provide a framework for your interpretation. For example, your original question may have been "For every 12 ounce can of soda drunk per day, how much greater is the average BMI among adults in the United States?". The wording of the question tells you that your original intent was to determine how much greater the BMI is among adults in the US who drink, for example, two 12 ounce cans of sodas per day on average, than among adults who drink only one 12 ounce soda per day on average. The interpretation of your analyses should yield a statement such as: For every 1 additional 12 ounce can of soda that adults in the US drink, BMI increases, on average, by X kg/m^2. But it should not yield a statement such as: "For every additional *ounce* of soda that adults in the US drink, BMI increases, on average, by X kg/m^2."

Another way in which revisiting your question provides a framework for interpreting your results is that reminding yourself of the type of question that you asked provides

an explicit framework for interpretation (See Stating and Refining the Question for a review of types of questions). For example, if your question was: "Among adults in the US, do those who drink 1 more 12 ounce serving of non-diet soda per day have a higher BMI, on average?", this tells you that your question is an *inferential* question and that your goal is to understand the average effect of drinking an additional 12 ounce serving of non-diet soda per day on BMI among the US adult population. To answer this question, you may have performed an analysis using cross-sectional data collected on a sample that was representative of the US adult population, and in this case your interpretation of the result is framed in terms of what the association is between an additional 12 ounce serving of soda per day and BMI, on average in the US adult population.

Because your question was not a causal one, and therefore your analysis was not a causal analysis, the result cannot be framed in terms of what would happen if a population started consuming an additional can of soda per day. A *causal* question might be: "What effect does drinking an additional 12 ounce serving of non-diet soda per day have on BMI?", and to answer this question, you might analyze data from a clinical trial that randomly assigned one group to drink an additional can of soda and the other group to drink an additional can of a placebo drink. The results from this type of question and analysis could be interpreted as what the causal effect of drinking additional 12 ounce can of soda per day would be on BMI. Because the analysis is comparing the average effect on BMI between the two groups (soda and placebo), the result would be intepreted as the average causal effect in the population.

A third purpose of revisiting your original question is that it is important to pause and consider whether your approach to answering the question could have produced a **biased** re-

sult. Although we covered bias to some degree in the chapter on Stating and Refining the Question, sometimes new information is acquired during the process of exploratory data analysis and/or modeling that directly affects your assessment of whether your result might be biased. Recall that bias is a systematic problem with the collection or analysis of the data that results in an incorrect answer to your question.

We will use the soda-BMI example to illustrate a simpler example of bias. Let's assume that your overall question about the soda-BMI relationship had included an initial question which was: What is the mean daily non-diet soda consumption among adults in the US? Let's assume that your analysis indicates that in the sample you are analyzing, which is a sample of all adults in the US, the average number of 12 ounce servings of non-diet soda drunk per day is 0.5, so you infer that the average number of 12 ounce servings of soda drunk per day by adults in the US is also 0.5. Since you should always challenge your results, it is important to consider whether your analysis has an inherent bias.

So how do you do this? You start by imagining that your result is incorrect, and then think through the ways in which the data collection or analysis could have had a systematic problem that resulted in an incorrect estimate of the mean number of 12 ounces cans of non-diet soda drunk per day by adults in the US. Although this exercise of imagining that your result is wrong is discussed as an approach to assessing the potential for bias, this is a terrific way to **challenge your results at every step of the analysis**, whether you are assessing risk of bias, or confounding, or a technical problem with your analysis.

The thought experiment goes something like this: imagine that the *true* average number of 12 ounce servings of non-

diet soda drunk per day by adults in the US is 2. Now imagine how the result from your analysis of the sample, which was 0.5, might be so far off from the true result: for some reason, the sample of the population that comprises your dataset is not a random sample of the population and instead has a disproportionate number of people who do not drink any non-diet soda, which brings down the estimated mean number of 12 ounces of servings of non-diet soda consumed per day. You might also imagine that if your sample result had been 4, which is much higher than the true amount drunk per day by adults in the US, that your sample has a disproportionate number of people who have high consumption of non-diet soda so that the estimate generated from your analyses is higher than the true value. So how can you gauge whether your sample is non-random?

To figure out if your sample is a non-random sample of the target population, think about what could have happened to attract more people who don't consume non-diet soda (or more people who consume a lot of it) to be included in the sample. Perhaps the study advertised for participation in a fitness magazine, and fitness magazine readers are less likely to drink non-diet soda. Or perhaps the data were collected by an internet survey and internet survey respondents are less likely to drink non-diet soda. Or perhaps the survey captured information about non-diet soda consumption by providing a list of non-diet sodas and asking survey respondents to indicate which ones they had consumed, but the survey omitted Mountain Dew and Cherry Coke, so that those people who drink mostly these non-diet sodas were classified as not consuming non-diet soda (or consuming less of it than they actually do consume). And so on.

Although we illustrated the simplest scenario for bias, which

occurs when estimating a prevalence or a mean, you of course can get a biased result for an estimate of a relationship between two variables as well. For example, the survey methods could unintentionally oversample people who both don't consume non-diet soda and have a high BMI (such as people with type 2 diabetes), so that the result would indicate (incorrectly) that consuming non-diet soda is not associated with having a higher BMI. The point is that pausing to perform a deliberate thought experiment about sources of bias is critically important as it is really the only way to assess the potential for a biased result. This thought experiment should also be conducted when you are stating and refining your question and also as your are conducting exploratory analyses and modeling.

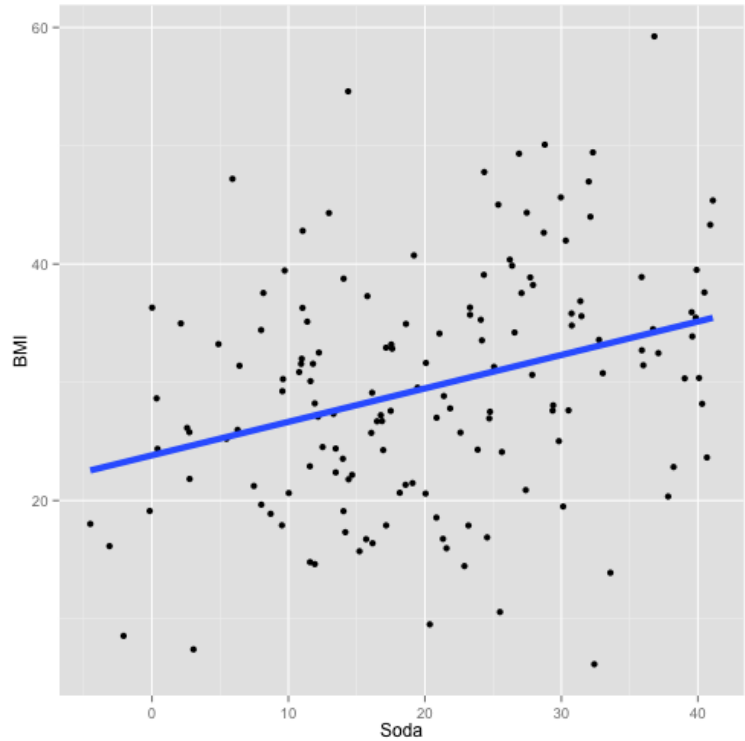
Start with the primary model and assess the directionality, magnitude, and uncertainty of the result

The second principle is to start with a single model and focus on the full continuum of the result, including its directionality and magnitude, and the degree of certainty (or uncertainty) there is about whether the result from the sample you analyzed reflects the true result for the overall population. A great deal of information that is required for interpreting your results will be missed if you zoom in on a single feature of your result (such as the p-value), so that you either ignore or gloss over other important information provided by the model. Although your interpretation isn't complete until you consider the results in totality, it is often most helpful to first focus on interpreting the results of the model that you believe best answers your question and reflects (or "fits") your data, which is your primary model (See Formal Modeling). Don't spend a lot of time worrying about which single model to start with, because in

the end you will consider all of your results and this initial interpretation exercise serves to orient you and provide a framework for your final interpretation.

Directionality

Building on the soda-BMI example, take a look at sample dataset below with a fitted model overlaid.



Sample Data for BMI-soda Example

We will focus on what the model tells us about the **directionality** of the relationship between soda consumption and BMI, the **magnitude** of the relationship, and the **uncer-**

tainty of the relationship, or how likely the model's depiction of the relationship between non-diet soda consumption and BMI is real vs. just a reflection of random variation you'd expect when sampling from a larger population.

The model indicates that the directionality of the relationship is positive, meaning that as non-diet soda consumption increases, BMI increases. The other potential results could have been a negative directionality, or no directionality (a value of approximately 0). Does the positive directionality of the result match your expectations that have been developed from the exploratory data analysis? If so, you're in good shape and can move onto the next interpretation activity. If not, there are a couple of possible explanations. First your expectations may not be correct because either the exploratory analysis was done incorrectly or your interpretation of the exploratory analyses were not correct. Second, the exploratory analysis and your interpretation of it may be correct, but the formal modeling may have been done incorrectly. Notice that with this process, you are once again applying the epicycle of data analysis.

Magnitude

Once you have identified and addressed any discrepancies between your expectations and interpretation of the directionality of the relationship, the next step is to consider the **magnitude** of the relationship. Because the model is a linear regression, you can see that the slope of the relationship, reflected by the beta coefficient, is 0.28. Interpreting the slope requires knowing the units of the "soda" variable. If the units are 12 ounce cans of soda per day, than the interpetation of this slope is that BMI increases by 0.28 kg/m^2 per additional 12 ounce can of non-diet soda that is consumed per day. However, the units are in ounces of soda,

so the interpretation of your model is that BMI increases by 0.28 kg/m^2 for each additional ounce of non-diet soda that is consumed per day.

Although you're comfortable that you understand the units of your soda variable correctly and have the correct interpretation of the model, you still don't quite have the answer to your question, which was framed in terms of the association of each additional 12 ounce can of soda and BMI, not each additional ounce of non-diet soda. So you'll need to convert the 0.28 slope so that it pertains to a 12 ounce, rather than 1 ounce, increase in soda consumption. Because the model is a linear model, you can simply multiply the slope, or beta coefficient, by 12 to get 3.36, which tells you that each additional 12 ounce can of soda consumed per day is associated with a BMI that is 3.36 kg/m^2 higher.

The other option of course is to create a new soda variable whose unit is 12 ounces rather than 1 ounce, but multiplying the slope is a simple mathematical operation and is much more efficient. Here again you should have had some expectations, based on the exploratory data analysis you did, about the magnitude of the relationship between non-diet soda consumption and BMI, so you should determine if your interpretation of the magnitude of the relationship matches your expectations. If not, you'll need to determine whether your expectations were incorrect or whether your interpretation was incorrect and act accordingly to match expectations and the result of your interpretation.

Another important consideration about the magnitude of the relationship is whether it is meaningful. For example, a 0.01 increase in BMI for every additional 20 ounces consumed per day is probably not particularly meaningful as a large amount of soda is associated with a very small increase in BMI. On the other hand, if there were a 0.28 kg/m^2

increase in BMI for every 1 ounce increase in soda consumption, this would in fact be quite meaningful. Because you know BMI generally ranges from the high teens to the 30's, a change of 0.01 kg/m^2 is small, but a change of 0.28 kg/m^2 could be meaningful.

When taken in the context of the kinds of volumes of soda people might consume, a 0.01 kg/m^2 for each 20 ounce increase in soda consumption is small since people are (hopefully) not drinking 10 twenty ounce servings per day, which is how much someone would need to drink in order to observe even a 0.1 kg/m^2 increase in BMI. On the other hand a 0.28 kg/m^2 increase in BMI for every additional ounce of soda would add up quickly for people who consumed an extra 20 ounce non-diet soda per day - this would equate to an expected increase in BMI of 5.6 kg/m^2. A key part of interpreting the magnitude of the result, then, is understanding how the magnitude of the result compares to what you know about this type of information in the population you're interested in.

Uncertainty

Now that you have a handle on what the model says about the directionality and magnitude of the relationship between non-diet soda consumption and BMI, the next step it to consider what the degree of **uncertainty** is for your answer. Recall that your model has been constructed to fit data collected from a *sample* of the overall population and that you are using this model to understand how non-diet soda consumption is related to BMI in the *overall* population of adults in the US.

Let's get back to our soda-BMI example, which does involve using the results that are obtained on the sample to make inferences about what the true soda-BMI relationship is in

the overall population of adults in the US. Let's imagine that the result from your analysis of the sample data indicates that *within your sample*, people who drink an additional ounce of non-diet soda per day have a BMI that is 0.28 kg/m^2 greater than those who drink an ounce less per day. However, how do you know whether this result is simply the "noise" of random sampling or whether it is a close approximation of the true relationship among the overall population?

To assess whether the result from the sample is simply random "noise", we use measures of uncertainty. Although some might expect that all random samples serve as excellent surrogates for the overall population, this is not true. To illustrate this idea using a simple example, imagine that the prevalence of females in the overall US adult population is 51%, and you draw a random sample of 100 adults. This sample may have 45% females. Imagine that you draw a new sample of 100 adults and your sample has 53% females. You could draw many samples like this and even draw samples of 35% or 70% females. The probability of drawing a sample with a prevalence of females that is this different from the overall population prevalence of females is very small, while the probability of drawing a sample that has close to 51% females is much higher.

It is this concept—**the probability that your sample reflects the answer for the overall population varies depending on how close (or far) your sample result is to the true result for the overall population**—that is the bedrock of the concept of uncertainty. Because we don't know what the answer is for the overall population (that's why we're doing the analysis in the first place!), it's impossible to express uncertainty in terms of how likely or unlikely it is that your sample result reflects the overall population. So there are other approaches to measuring uncertainty

that rely on this general concept, and we will discuss two common approaches below.

One tool that provides a more continuous measure of uncertainty is the confidence interval. A confidence interval is a range of values that contains your sample result and you have some amount of confidence that it also contains the true result for the overall population. Most often statistical modeling software provides 95% confidence intervals, so that if the 95% CI for the sample estimate of 0.28 kg/m^2 from above is 0.15–0.42 kg/m^2, the approximate interpretation is that you can be 95% confident that the true result for the overall population is somewhere between 0.15 and 0.42 kg/m^2.

> A more precise definition of the 95% confidence interval would be that over repeated samples, if we were to conduct this experiment many times (each time collecting a dataset of the same size) then a confidence interval constructed in this manner would cover the truth 95% of the time. It's important to realize that because the confidence interval is constructed from the data, the *interval itself is random*. Therefore, if we were to collect new data, the interval we'd construct would be slightly different. However, the truth, meaning the population value of the parameter, would always remain the same.

Another tool for measuring uncertainty is, of course, the p-value, which simply is the probability of getting the sample result of 0.28 kg/m^2 (or more extreme) when the true relationship between non-diet soda consumption and BMI in the overall population is 0. Although the p-value is a continuous measure of uncertainty, many people consider

a p-value of <0.05, which indicates that there is a less than 5% probability of observing the sample result (or a more extreme result) when there is no relationship in the overall population, as "statistically significant". This cutpoint is arbitrary and tells us very little about the *degree* of uncertainty or about where the true answer for the overall population lies. Focusing primarily on the p-value is a risky approach to interpreting uncertainty because it can lead to ignoring more important information needed for thoughtful and accurate interpretation of your results.

The CI is more helpful than the p-value, because it gives a range, which provides some quantitative estimate about what the actual overall population result is likely to be, and it also provides a way to express how certain it is that the range contains the overall population result.

Let's walk through how the p-value vs. 95% CI would be used to interpret uncertainty about the result from the soda-BMI analysis. Let's say that our result was that BMI was 0.28 kg/m^2 higher on average among our sample who drank one ounce more of non-diet soda per day and that the p-value associated with this result was 0.03. Using the p-value as a tool for measuring uncertainty and setting a threshold of statistical significance at 0.05, we would interpet the uncertainty as follows: there is a less than 5% chance that we would get this result (0.28) or something more extreme if the true population value was 0 (or in other words, that there was really not an association between soda consumption and BMI in the overall population).

Now let's go through the same exercise with the 95% CI. The 95% CI for this analysis is 0.15–0.42. Using the CI as the tool for interpreting uncertainty, we could say that we are 95% confident that the true relationship between soda consumption and BMI in the adult US population lies

somewhere between a 0.15 and 0.42 kg/m^2 increase in BMI on average per additional ounce of non-diet soda that is consumed. Using this latter approach tells us something about the range of possible effects of soda on BMI and also tells us that it is very unlikely that soda has no association with BMI in the overall population of adults in the US. Using the p-value as the measure of uncertainty, on the other hand, implies that we have only two choices in terms of interpreting the result: either there is a good amount of uncertainty about it so we must conclude that there is no relationship between soda consumption and BMI, or there is very little uncertainty about the result so we must conclude that there is a relationship between soda consumption and BMI. Using the p-value constrains us in a way that does not reflect the process of weighing the strength of the evidence in favor (or against) a hypothesis.

Another point about uncertainty is that we have discussed assessing uncertainty through more classical statistical approaches, which are based on the Frequentist paradigm, which is the most common approach. The Bayesian framework is an alternate approach in which you update your prior beliefs based on the evidence provided by the analysis. In practice, the Frequentist approach we discussed above is more commonly used, and in real-world setting rarely leads to conclusions that would be different from those obtained by using a Bayesian approach.

One important caveat is that sometimes evaluating uncertainty is not necessary because some types of analyses are not intended to make inferences about a larger overall population. If, for example, you wanted to understand the relationship between age and dollars spent per month on your company's products, you may have all of the data on the entire, or "overall" population you are interested in-which is your company's customers. In this case you do not

have to rely on a sample, because your company collects data about the age and purchases of ALL of their customers. In this case, you would not need to consider the uncertainty that your result reflects the truth for the overall population because your analysis result **is the truth** for your overall population.

Develop an overall interpretation by considering the totality of your analyses and external information

Now that you have dedicated a good amount of effort interpreting the results of your primary model, the next step is to develop an overall interpretation of your results by considering both the totality of your analyses and information external to your analyses. The interpretation of the results from your primary model serves to set the expectation for your overall interpretation when you consider all of your analyses. Building on the soda-BMI example, let's assume that your interpretation of your primary model is that BMI is 0.28 kg/m^2 higher on average among adults in the US who consume an average one additional ounce of soda per day. Recall that this primary model was constructed after gathering information through exploratory analyses and that you may have refined this model when you were going through the process of interpreting its results by evaluating the directionality, magnitude and uncertainty of the model's results.

As discussed in the Formal Modeling chapter, there is not one single model that alone provides **the** answer to your question. Instead, there are additional models that serve to challenge the result obtained in the primary model. A common type of secondary model is the model which is constructed to determine how sensitive the results in your

primary model are to changes in the data. A classic example is removing outliers to assess the degree to which your primary model result changes. If the primary model results were largely driven by a handful of, for example, very high soda consumers, this finding would suggest that there may not be a linear relationship between soda consumption and BMI and that instead soda consumption may only influence BMI among those who have very high consumption of soda. This finding should lead to a revision of your primary model.

A second example is evaluating the effect of potential confounders on the results from the primary model. Although the primary model should already contain key confounders, there are typically additional potential confounders that should be assessed. In the soda-BMI example, you may construct a secondary model that includes income because you realize that it is possible that the relationship you observe in your primary model could be explained entirely by socioeconomic status: people of higher socioeconomic status might drink less non-diet soda and also have lower BMIs, but it is not because they drink less soda that this is the case. Instead, it is some other factor associated with socioeconomic status that has the effect on BMI. So you can run a secondary model in which income is added to the primary model to determine if this is the case. Although there are other examples of uses of secondary models, these are two common examples.

So how do you interpret how these secondary model results affect your primary result? You can fall back on the paradigm of: directionality, magnitude, and uncertainty. When you added income to the soda-BMI model, did income change the directionality of your estimated relationship between soda and BMI from the primary model - either to a negative association or no association? If it did, that

would be a dramatic change and suggest that either something is not right with your data (such as with the income variable) or that the association between soda consumption and BMI is entirely explained by income.

Let's assume that adding income did not change the directionality and suppose that it changed the magnitude so that the primary model's estimate of 0.28 kg/m^2 decreased to 0.12kg/m^2. The magnitude of the relationship between soda and BMI was reduced by 57%, so this would be interpreted as income explaining a little more than half, but not all, of the relationship between soda consumption and BMI.

Now you move on to uncertainty. The 95% CI for the estimate with the model that includes income is 0.01–0.23, so that we can be 95% confident that the true relationship between soda and BMI in the adult US population, independent of income, lies somewhere in this range. What if the 95% CI for the estimate were -0.02–0.26, but the estimate was still 0.12 kg/m^2? Even though the CI now includes 0, the result from the primary model, 0.12, did not change, indicating that income does not appear to explain any of the association between soda consumption and BMI, but that it did increase the uncertainty of the result. One reason that the addition of income to the model could have increased the uncertainty is that some people in the sample were missing income data so that the sample size was reduced. Checking your n's will help you determine if this is the case.

It's also important to consider your overall results in the context of external information. External information is both general knowledge that you or your team members have about the topic, results from similar analyses, and information about the target population. One example discussed above is that having a sense of what typical and plausible volumes of soda consumption are among adults in

the US is helpful for understanding if the magnitude of the effect of soda consumption on BMI is meaningful. It may also be helpful to know what percent of the adult population in the US drinks non-diet soda and the prevalence of obesity to understand the size of the population for whom your results might be pertinent.

One interesting example of how important it is to think about the size of the population that may be affected is air pollution. For associations between outdoor air pollution and critical health outcomes such as cardiovascular events (stroke, heart attack), the magnitude of the effect is small, but because air pollution affects hundreds of millions of people in the US, the numbers of cardiovascular events attributable to pollution is quite high.

In addition, you probably are not the first person to try and answer this question or related questions. Others may have done an analysis to answer the question in another population (adolescents, for example) or to answer a related, but different question, such as: "what is the relationship between non-diet soda consumption and blood sugar levels?" Understanding how your results fit into the context of the body of knowledge about the topic helps you and others assess whether there is an overall story or pattern emerging across all sources of knowledge that point to non-diet soda consumption being linked to high blood sugar, insulin resistance, BMI, and type 2 diabetes. On the other hand, if the results of your analysis differ from the external knowledge base, that is important too. Although most of the time when the results are so strikingly different from external knowledge, there is an explanation such as an error or differences in methods of data collection or population studied, sometimes a distinctly different finding is a truly novel insight.

Implications

Now that you've interpreted your results and have conclusions in hand, you'll want to think about the implications of your conclusions. After all, the point of doing an analysis is usually to inform a decision or to take an action. Sometimes the implications are straightforward, but other times the implications take some thought. An example of a straightforward implication is if you performed an analysis to determine if purchasing ads increased sales, and if so, did the investment in ads result in a net profit. You may learn that either there was a net profit or not, and if there were a net profit, this finding would support continuing the ads.

A more complicated example is the soda-BMI example we've used throughout this chapter. If soda consumption turned out to be associated with higher BMI, with a 20 ounce additional serving per day associated with a 0.28 kg/m^2 greater BMI, this finding would imply that if you could reduce soda consumption, you could reduce the average BMI of the overall population. Since your analysis wasn't a causal one, though, and you only demonstrated an association, you may want to perform a study in which you randomly assign people to either replacing one of the 20 ounce sodas they drink each day with diet soda or to not replacing their non-diet soda. In a public health setting, though, your team may decide that this association is sufficient evidence to launch a public health campaign to reduce soda consumption, and that you do not need additional data from a clinical trial. Instead, you may plan to track the population's BMI during and after the public health campaign as a means of estimating the public health effect of reducing non-diet soda consumption. The take-home point here is that the action that results from the implications often depends on the mission of the organization that requested the analysis.

10. Communication

Communication is fundamental to good data analysis. What we aim to address in this chapter is the role of routine communication in the process of doing your data analysis and in disseminating your final results in a more formal setting, often to an external, larger audience. There are lots of good books that address the "how-to" of giving formal presentations, either in the form of a talk or a written piece, such as a white paper or scientific paper. In this chapter, though, we will focus on:

1. How to use routine communication as one of the tools needed to perform a good data analysis; and
2. How to convey the key points of your data analysis when communicating informally and formally.

Communication is both one of the tools of data analysis, and also the final product of data analysis: there is no point in doing a data analysis if you're not going to communicate your process and results to an audience. A good data analyst communicates informally multiple times during the data analysis process and also gives careful thought to communicating the final results so that the analysis is as useful and informative as possible to the wider audience it was intended for.

10.1 Routine communication

The main purpose of routine communication is to gather data, which is part of the epicyclic process for each core

activity. You gather data by communicating your results and the responses you receive from your audience should inform the next steps in your data analysis. The types of responses you receive include not only answers to specific questions, but also commentary and questions your audience has in response to your report (either written or oral). The form that your routine communication takes depends on what the goal of the communication is. If your goal, for example, is to get clarity on how a variable is coded because when you explore the dataset it appears to be an ordinal variable, but you had understood that it was a continous variable, your communication is brief and to the point.

If, on the other hand, some results from your exploratory data analysis are not what you expected, your communication may take the form of a small, informal meeting that includes displaying tables and/or figures pertinent to your issue. A third type of informal communication is one in which you may not have specific questions to ask of your audience, but instead are seeking feedback on the data analysis process and/or results to help you refine the process and/or to inform your next steps.

In sum, there are three main types of informal communication and they are classified based on the objectives you have for the communication: (1) to answer a very focused question, which is often a technical question or a question aimed at gathering a fact, (2) to help you work through some results that are puzzling or not quite what you expected, and (3) to get general impressions and feedback as a means of identifying issues that had not occurred to you so that you can refine your data analysis.

Focusing on a few core concepts will help you achieve your objectives when planning routine communication. These concepts are:

1. **Audience**: Know your audience and when you have control over who the audience is, select the right audience for the kind of feedback you are looking for.
2. **Content**: Be focused and concise, but provide sufficient information for the audience to understand the information you are presenting and question(s) you are asking.
3. **Style**: Avoid jargon. Unless you are communicating about a focused highly technical issue to a highly technical audience, it is best to use language and figures and tables that can be understood by a more general audience.
4. **Attitude**: Have an open, collaborative attitude so that you are ready to fully engage in a dialogue and so that your audience gets the message that your goal is not to "defend" your question or work, but rather to get their input so that you can do your best work.

10.2 The Audience

For many types of routine communication, you will have the ability to select your audience, but in some cases, such as when you are delivering an interim report to your boss or your team, the audience may be pre-determined. Your audience may be composed of other data analysts, the individual(s) who initiated the question, your boss and/or other managers or executive team members, non-data analysts who are content experts, and/or someone representing the general public.

For the first type of routine communication, in which you are primarily seeking factual knowledge or clarification about the dataset or related information, selecting a person (or people) who have the factual knowledge to answer the

question and are responsive to queries is most appropriate. For a question about how the data for a variable in the dataset were collected, you might approach a person who collected the data or a person who has worked with the dataset before or was responsible for compiling the data. If the question is about the command to use in a statistical programming language in order to run a certain type of statistical test, this information is often easily found by an internet search. But if this fails, querying a person who uses the particular programming language would be appropriate.

For the second type of routine communication, in which you have some results and you are either unsure whether they are what you'd expect, or they are not what you expected, you'll likely be most helped if you engage more than one person and they represent a range of perspectives. The most productive and helpful meetings typically include people with data analysis and content area expertise. As a rule of thumb, the more types of stakeholders you communicate with while you are doing your data anlysis project, the better your final product will be. For example, if you only communicate with other data analysts, you may overlook some important aspects of your data analysis that would have been discovered had you communicated with your boss, content experts, or other people.

For the third type of routine communication, which typically occurs when you have come to a natural place for pausing your data analysis. Although when and where in your data analysis these pauses occur are dictated by the specific analysis you are doing, one very common place to pause and take stock is after completing at least some exploratory data analysis. It's important to pause and ask for feedback at this point as this exercise will often identify additional exploratory analyses that are important for in-

forming next steps, such as model building, and therefore prevent you from sinking time and effort into pursuing models that are not relevant, not appropriate, or both. This sort of communication is most effective when it takes the form of a face-to-face meeting, but video conferencing and phone conversations can also be effective. When selecting your audience, think about who among the people available to you give the most helpful feedback and which perspectives will be important for informing the next steps of your analysis. At a minimum, you should have both data analysis and content expertise represented, but in this type of meeting it may also be helpful to hear from people who share, or at least understand, the perspective of the larger target audience for the formal communication of the results of your data analysis.

10.3 Content

The most important guiding principle is to tailor the information you deliver to the objective of the communication. For a targeted question aimed at getting clarification about the coding of a variable, the recipient of your communication does not need to know the overall objective of your analysis, what you have done up to this point, or see any figures or tables. A specific, pointed question along the lines of "I'm analyzing the crime dataset that you sent me last week and am looking at the variable "education" and see that it is coded 0, 1, and 2, but I don't see any labels for those codes. Do you know what these codes for the "education" variable stand for?"

For the second type of communication, in which you are seeking feedback because of a puzzling or unexpected issue with your analysis, more background information will be

needed, but complete background information for the overall project may not be. To illustrate this concept, let's assume that you have been examining the relationship between height and lung function and you construct a scatterplot, which suggests that the relationship is non-linear as there appears to be curvature to the relationship. Although you have some ideas about approaches for handling non-linear relationships, you appropriately seek input from others. After giving some thought to your objectives for the communication, you settle on two primary objectives: (1) To understand if there is a best approach for handling the non-linearity of the relationship, and if so, how to determine which approach is best, and (2) To understand more about the non-linear relationship you observe, including whether this is expected and/or known and whether the non-linearity is important to capture in your analyses.

To achieve your objectives, you will need to provide your audience with some context and background, but providing a comprehensive background for the data analysis project and review of all of the steps you've taken so far is unnecessary and likely to absorb time and effort that would be better devoted to your specific objectives. In this example, appropriate context and background might include the following: (1) the overall objective of the data analysis, (2) how height and lung function fit into the overall objective of the data analysis, for example, height may be a potential confounder, or the major predictor of interest, and (3) what you have done so far with respect to height and lung function and what you've learned. This final step should include some visual display of data, such as the aforementioned scatterplot. The final content of your presentation, then, would include a statement of the objectives for the discussion, a brief overview of the data analysis project, how the specific issue you are facing fits into the overall data analysis project, and

then finally, pertinent findings from your analysis related to height and lung function.

If you were developing a slide presentation, fewer slides should be devoted to the background and context than the presentation of the data analysis findings for height and lung function. One slide should be sufficient for the data analysis overview, and 1-2 slides should be sufficient for explaining the context of the height-lung function issue within the larger data analysis project. The meat of the presentation shouldn't require more than 5-8 slides, so that the total presentation time should be no more than 10-15 minutes. Although slides are certainly not necessary, a visual tool for presenting this information is very helpful and should not imply that the presentation should be "formal." Instead, the idea is to provide the group sufficient information to generate discussion that is focused on your objectives, which is best achieved by an informal presentation.

These same principles apply to the third type of communication, except that you may not have focused objectives and instead you may be seeking general feedback on your data analysis project from your audience. If this is the case, this more general objective should be stated and the remainder of the content should include a statement of the question you are seeking to answer with the analysis, the objective(s) of the data analysis, a summary of the characteristics of the data set (source of the data, number of observations, etc.), a summary of your exploratory analyses, a summary of your model building, your interpretation of your results, and conclusions. By providing key points from your entire data analysis, your audience will be able to provide feedback about the overall project as well as each of the steps of data analysis. A well planned discussion yields helpful, thoughtful feedback and should be considered a success if

you are left armed with additional refinments to make to your data analysis and thoughtful perspective about what should be included in the more formal presentation of your final results to an external audience.

10.4 Style

Although the style of communication increases in formality from the first to the third type of routine communication, all of these communications should largely be informal and, except for perhaps the focused commmunication about a small technical issue, jargon should be avoided. Because the primary purpose of routine communication is to get feedback, your communication style should encourage discussion. Some approaches to encourage discussion include stating up front that you would like the bulk of the meeting to include active discussion and that you welcome questions during your presentation rather than asking the audience to hold them until the end of your presentation. If an audience member provides commentary, asking what others in the audience think will also promote discussion. In essence, to get the best feedback you want to hear what your audience members are thinking, and this is most likely accomplished by setting an informal tone and actively encouraging discussion.

10.5 Attitude

A defensive or off-putting attitude can sabotage all the work you've put into carefully selecting the audience, thoughtfully identifying your objectives and preparing your content, and stating that you are seeking discussion. Your audience will be reluctant to offer constructive feedback if they

sense that their feedback will not be well received and you will leave the meeting without achieving your objectives, and ill prepared to make any refinements or additions to your data analysis. And when it comes time to deliver a formal presentation to an external audience, you will not be well prepared and won't be able to present your best work. To avoid this pitfall, deliberately cultivate a receptive and positive attitude prior to communicating by putting your ego and insecurities aside. If you can do this successfully, it will serve you well. In fact, we both know people who have had highly successful careers based largely on their positive and welcoming attitude towards feedback, including constructive criticism.

11. Concluding Thoughts

You should now be armed with an approach that you can apply to your data analyses. Although each data set is its own unique organism and each analysis has its own specific issues to contend with, tackling each step with the epicycle framework is useful for any analysis. As you work through developing your question, exploring your data, modeling your data, interpreting your results, and communicating your results, remember to always set expectations and then compare the result of your action to your expectations. If they don't match, identify whether the problem is with the result of your action or your expectations and fix the problem so that they do match. If you can't identify the problem, seek input from others, and then when you've fixed the problem, move on to the next action. This epicycle framework will help to keep you on a course that will end at a useful answer to your question.

In addition to the epicycle framework, there are also activities of data analysis that we discussed throughout the book. Although all of the analysis activities are important, if we had to identify the ones that are most important for ensuring that your data analysis provides a valid, meaningful, and interpretable answer to your question, we would include the following:

1. Be thoughtful about developing your question and use the question to guide you throughout all of the analysis steps.
2. Follow the ABCs:
    1. Always be checking

2. Always be challenging
3. Always be communicating

The best way for the epicycle framework and these activities to become second nature is to do a lot of data analysis, so we encourage you to take advantage of the data analysis opportunities that come your way. Although with practice, many of these principles will become second nature to you, we have found that revisiting these principles has helped to resolve a range of issues we've faced in our own analyses. We hope, then, that the book continues to serve as a useful resource after you're done reading it when you hit the stumbling blocks that occur in every analysis.

12. About the Authors

Roger D. Peng is an Associate Professor of Biostatistics at the Johns Hopkins Bloomberg School of Public Health. He is also a Co-Founder of the Johns Hopkins Data Science Specialization[1], which has enrolled over 1.5 million students, and the Simply Statistics blog[2] where he writes about statistics and data science for the general public. Roger can be found on Twitter and GitHub @rdpeng[3].

Elizabeth Matsui is a Professor of Pediatrics, Epidemiology and Environmental Health Sciences at Johns Hopkins University and a practicing pediatric allergist/immunologist. She directs a data management and analysis center with Dr. Peng that supports epidemiologic studies and clinical trials and is co-founder of Skybrude Consulting, LLC[4], a data science consulting firm. Elizabeth can be found on Twitter @eliza68[5].

[1] http://www.coursera.org/specialization/jhudatascience/1
[2] http://simplystatistics.org/
[3] https://twitter.com/rdpeng
[4] http://skybrudeconsulting.com
[5] https://twitter.com/eliza68