Statistical Foundations for Engineers and Scientists

Eric M Reyes

Last Updated: 2017-08-31

Contents

I Unit I: Language and Logic of Inference **5**

1 The Statistical Process **7**
1.1 Overview of Drawing Inference . . . 8
1.2 Anatomy of a Dataset . . . 9
1.3 A Note on Codebooks . . . 10

2 Case Study: Health Effects of the Deepwater Horizon Oil Spill **11**

3 Asking the Right Questions **13**
3.1 Characterizing a Variable . . . 13
3.2 Framing the Question . . . 15

4 Gathering the Evidence (Data Collection) **17**
4.1 What Makes a Sample Reliable . . . 17
4.2 Poor Methods of Data Collection . . . 19
4.3 Preferred Methods of Sampling . . . 21
4.4 Two Types of Studies . . . 21

5 Presenting the Evidence (Summarizing Data) **27**
5.1 Characteristics of a Distribution (Summarizing a Single Variable) . . . 27
5.2 Summarizing Relationships . . . 32

6 Assessing the Evidence (Quantifying the Variability in Estimates) **35**
6.1 Conceptualizing the Sampling Distribution . . . 36
6.2 Example of a Sampling Distribution . . . 38
6.3 Modeling the Sampling Distribution . . . 41
6.4 Using a Model for the Sampling Distributions (Confidence Intervals) . . . 43
6.5 Bringing it All Together . . . 45

7 Quantifying the Evidence (Rejecting Bad Models) **47**
7.1 Some Subtleties . . . 47
7.2 Assuming the Null Hypothesis . . . 48
7.3 Using the Null Distribution . . . 50
7.4 Sampling Distributions vs. Null Distributions . . . 52

8 Using the Tools Together **53**
8.1 Framing the Question (Fundamental Idea I) . . . 53
8.2 Getting Good Data (Fundamental Idea II) . . . 54
8.3 Presenting the Data (Fundamental Idea III) . . . 54
8.4 Quantifying the Variability in the Estimate (Fundamental Idea IV) . . . 55
8.5 Quantifying the Evidence (Fundamental Idea V) . . . 56
8.6 Summary . . . 57

II Unit II: Comparing the Average Response Across Groups 59

9 Case Study: Organic Foods and Superior Morals 61

10 Framing the Question 63
10.1 General Setting 64

11 Study Design 67
11.1 Aspects of a Well Designed Experiment 67
11.2 Collecting Observational Data 70

12 Presenting the Data 71

13 Quantifying the Evidence 75
13.1 Partitioning Variability 75
13.2 Forming a Standardized Test Statistic 77
13.3 Obtaining a P-value 78

14 Building the Statistical Model 81
14.1 General Formulation 81
14.2 Statistical Model for A Quantitative Response and a Categorical Predictor 82
14.3 Conditions on the Error Distribution 84
14.4 Simulating the Null Distribution 85
14.5 Recap 86

15 Assessing Modeling Assumptions 87
15.1 Assessing Independence 87
15.2 Assessing Homoskedasticity 89
15.3 Assessing Normality 89
15.4 General Tips for Assessing Assumptions 90

Part I

Unit I: Language and Logic of Inference

Chapter 1

The Statistical Process

Is driving while texting as dangerous as driving while intoxicated? Is there a relationship between a student's college grade point average and their starting salary following graduation? If so, does that relationship differ across academic departments? Regardless of your future career path, you will eventually need to answer a question. The discipline of statistics is about using data to address questions by converting that data into valuable information.

> **Key Idea:** Statistics is the discipline of converting data into information.

It might be natural at this point to ask "do I really need an entire class about answering questions with data? Isn't this simple?" Sometimes, it is simple; other times, it can be far from it. Let's illustrate with the following example from Tintle et al. (2015).

Example 1.1 (Organ Donation)**.** Even though organ donations save lives, recruiting organ donors is difficult. Interestingly, surveys show that about 85% of Americans approve of organ donation in principle and many states offer a simple organ donor registration process when people apply for a driver's license. However, only about 38% of licensed drivers in the United States are registered to be organ donors. Some people prefer not to make an active decision about organ donation because the topic can be unpleasant to think about. But perhaps phrasing the question differently could affect people's willingness to become a donor.

Johnson and Goldstein (2003) recruited 161 participants for a study, published in the journal *Science*, to address the question of organ donor recruitment. The participants were asked to imagine they had moved to a new state and were applying for a driver's license. As part of this application, the participants were to decide whether or not to become an organ donor. Participants were presented with one of three different default choices:

- Some of the participants were forced to make a choice of becoming a donor or not, without being given a default option (the "neutral" group).
- Other participants were told that the default option was not to be a donor but that they could choose to become a donor if they wished (the "opt-in" group).
- The remaining participants were told that the default option was to be a donor but that they could choose not to become a donor if they wished (the "opt-out" group).

The results of this study were 79% agreeing to become donors in the neutral group, 42% for the opt-in group, and 82.0% for the opt-out group.

The results of the study are presented in Figure 1.1. It seems obvious that using the "opt-in" strategy results in fewer people agreeing to organ donation. However, does the "opt-out" strategy, in which people are by default declared organ donors, result in more people agreeing to organ donation compared to the "neutral" strategy? On the one hand, a higher percentage did agree to organ donation under the "opt-out" (82% compared to 79%). However, since this study involved only a subset of Americans, is this enough evidence to

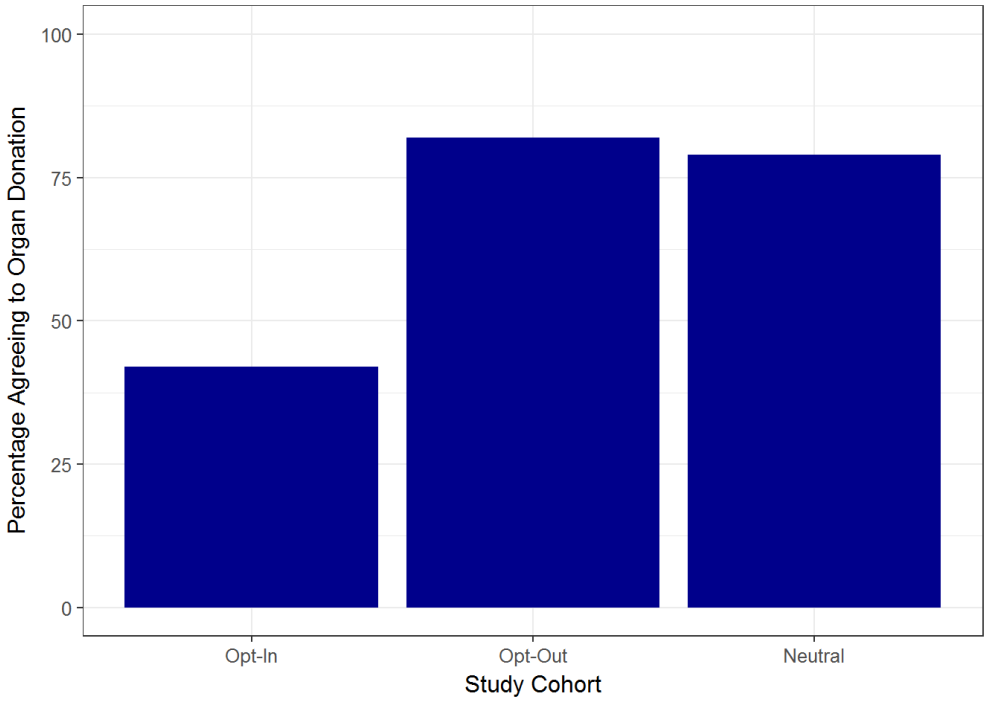


Figure 1.1: Summary of the responses for the Organ Donation Study described in Example 1.1.

claim the "opt-out" strategy is really superior compared to the "neutral" strategy? The discipline of statistics provides a framework for addressing such ambiguity.

1.1 Overview of Drawing Inference

Let's begin by taking a step back and considering the big picture of how data is turned into information. Every research question we pose, at its heart, is trying to characterize a **population**, the group of subjects of ultimate interest.

Definition 1.1 (Population)**.** The collection of subjects we would like to say something about.

In the Organ Donation study, the researchers would like to say something about Americans who are of the age to consent to organ donation; in particular, they would like to quantify how likely it is that someone from this group agrees to organ donation. Therefore, the population is the all Americans who are of the age to consent to organ donation. The subjects in a population need not be people; the population could just as easily be a collection of screws, sheet metal... whatever characterizes the objects from which we would *like to* obtain measurements. We use the phrase "like to" because in reality, it is often impossible (or impractical) to observe the entire population. Instead, we make observations on a subset of the population; this smaller group is known as the **sample**.

Definition 1.2 (Sample)**.** The collection of subjects for which we actually obtain measurements (data).

For each subject within the sample, we obtain a collection of measurements forming our set of data. The goal of statistical modeling is to use the sample (the group we actually observe) to say something about the population of interest (the group we wish we had observed); this process is known as **statistical inference**. This process is illustrated in Figure 1.2.

Definition 1.3 (Statistical Inference)**.** Sometimes referred to as "inference," the process of using a sample to characterize some aspect of the population.

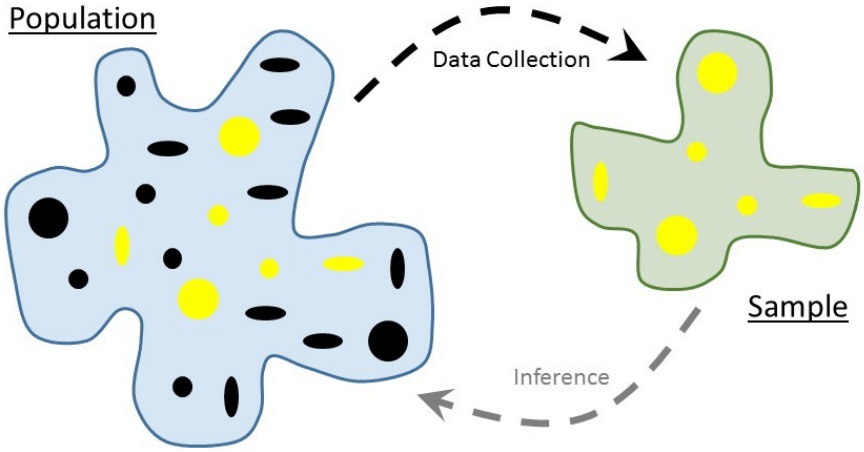


Figure 1.2: Illustration of the statistical process.

1.2 Anatomy of a Dataset

Once we have our sample, we take measurements on each of the subjects. These measurements form the data. When we hear the word "data," most of us envision a large spreadsheet. In reality, data can take on many forms — spreadsheets, images, text files, unstructured text from a Twitter feed, etc. Regardless of the form, all datasets contain information for each subject in the sample; this information, the various measurements, are called **variables**.

Definition 1.4 (Variable)**.** A measurement, or category, describing some aspect of the subject.

Variables come in one of two flavors. **Categorical** variables are those which denote a grouping to which the subject belongs. Examples include marital status, brand, and experimental treatment group. **Numeric** variables are those which take on values for which ordinary arithmetic (addition, multiplication) makes sense. Examples include height, age of a product, and diameter. Note that sometimes numeric values are used to represent the levels of a categorical variable in a dataset; for example, 0 may indicate "No" and 1 may indicate "Yes" for a variable capturing whether a person is a registered organ donor. Therefore, just because a variable has a numeric value does not make it a numeric variable; the key here is that numeric variables are those for which arithmetic makes sense.

Definition 1.5 (Categorical Variable)**.** Also called a "qualitative variable," a measurement on a subject which denotes a grouping or categorization.

Definition 1.6 (Numeric Variable)**.** Also called a "quantitative variable," a measurement on a subject which takes on a numeric value *and* for which ordinary arithmetic makes sense.

While it may be natural to think of a dataset as a spreadsheet, not all spreadsheets are created equal. Here, we consider datasets which have the following characteristics:

- Each column contains a unique variable.
- Each record (row in the dataset) corresponds to a different observation of the variables.
- If you have multiple datasets, they should include a column in the table that allows them to be linked (subject identifier).

These are characteristics of "tidy data." Even unstructured data such as images or Twitter feeds must be processed, often converted to tidy data, prior to performing a statistical analysis. The above description eliminates a common method of storing data in engineering ans scientific disciplines — storing each sample in a different column. To illustrate, suppose we conduct a study comparing the lifetime (in hours) of two brands of batteries. We measure the lifetime of five batteries of Brand A and six of Brand B. It is common to see a

Table 1.1: Example of a common data structure which does not represent tidy data. Data is from a hypothetical study comparing battery lifetimes (hours).

Brand A	Brand B
8.3	8.4
5.1	8.6
3.3	3.8
5.3	4.1
5.7	4.5
	4.0

Table 1.2: Example of a tidy dataset, a good way of storing data. Data is from a hypothetical study comparing battery lifetimes (hours).

Battery	Brand	Lifetime
1	A	8.3
2	A	5.1
3	A	3.3
4	A	5.3
5	A	5.7
6	B	8.4
7	B	8.6
8	B	3.8
9	B	4.1
10	B	4.5
11	B	4.0

dataset like that in Table 1.1; the problem here is that the first record of the dataset contains information on two different observations. We have the lifetime from a battery of Brand A in the same row as the lifetime from a battery of Brand B.

In order to adhere to the tidy structure, we can reformat this dataset as illustrated in Table 1.2. Here, each record represents a unique observation and each column is a different variable. We have also added a unique identifier.

It may take some time to get used to storing data in this format, but it makes analysis easier and avoids time spent managing the data later.

1.3 A Note on Codebooks

A dataset on its own is meaningless if you cannot understand what the values represent. *Before* you access a dataset, you should always review any available **codebooks**.

Definition 1.7 (Codebook)**.** Also called a "data dictionary," these provide complete information regarding the variables contained within a dataset.

Some codebooks are excellent, with detailed descriptions of how the variables were collected and appropriate units. Other codebooks are not as good, giving only an indication of what the variable represents. Whenever you are working with previously collected data, reviewing a codebook is the first step; and, you should be prepared to revisit the codebook often throughout an analysis. When you are collecting your own dataset, constructing a codebook is essential for others to make use of your data.

Chapter 2

Case Study: Health Effects of the Deepwater Horizon Oil Spill

On the evening of April 20, 2010, the *Deepwater Horizon*, an oil drilling platform positioned off the coast of Louisiana, was engulfed in flames as the result of an explosion. The drilling rig, leased and operated by BP, had been tasked with drilling an oil well in water nearly 5000 feet deep. Eleven personnel were killed in the explosion. The following clip is from the initial coverage by the *New York Times*[1]:

The incident is considered the worst oil spill in US history, creating an environmental disaster along the Gulf Coast. In addition to studying the effects on the local environment, researchers have undertaken studies to examine the short and long-term health effects caused by the incident. As an example, it is reasonable to ask whether volunteers who were directly exposed to oil, such as when cleaning wildlife, are at higher risk of respiratory irritation compared to those volunteers who were helping with administrative tasks and therefore were not directly exposed to the oil. An article appearing in *The New England Journal of Medicine* (B. D. Goldstein, Osofsky, and Lichtveld 2011) reported the results from a health symptom survey performed in the Spring and Summer of 2010 by the National Institute for Occupational Safety and Health. Of 54 volunteers assigned to wildlife cleaning and rehabilitation, 15 reported experiencing "nose irritation, sinus problems, or sore throat." Of 103 volunteers who had no exposure to oil, dispersants, cleaners, or other chemicals, 16 reported experiencing "nose irritation, sinus problems, or sore throat."

While a larger fraction of volunteers cleaning wildlife *in the study* reported respiratory symptoms compared to those who were not directly exposed to irritants, would we expect similar results if we were able to interview all volunteers? What about during a future oil spill? Is there evidence that more than 1 in 5 volunteers who clean wildlife will develop respiratory symptoms? What is a reasonable value for the increased risk of respiratory symptoms for those volunteers with direct exposure compared to those without?

In the first part of this text, we use this case study as the context for discussing how research questions should be framed, methods for data collection, summarizing and presenting data clearly, quantifying the variability in an estimate, and quantifying the degree to which the data disagrees with a proposed model. We capture these ideas in what we call the *Five Fundamental Ideas of Inference*. We see that any statistical analysis iterates between the components of what we call the *Distributional Quartet*. These two frameworks allow us to describe the language and logic of inference, serving as a foundation for the statistical thinking and reasoning needed to address more complex questions encountered later in the text.

[1] http://www.nytimes.com/2010/04/22/us/22rig.html?rref=collection%2Ftimestopic%2FOil%20Spills&action=click&contentCollection=timestopics&region=stream&module=stream_unit&version=search&contentPlacement=1&pgtype=collection

Search Continues After Oil Rig Blast

By CAMPBELL ROBERTSON APRIL 21, 2010

The rig burned Wednesday about 50 miles southeast of Venice, La. Firefighting efforts were causing it to take on water and list. Gerald Herbert/Associated Press

NEW ORLEANS — An explosion on an oil drilling rig off the coast of southeast Louisiana left at least 3 people critically injured and 11 others missing as of Wednesday night.

Figure 2.1: *New York Times* coverage of the *Deepwater Horizon* oil spill.

Chapter 3

Asking the Right Questions

The discipline of statistics is about turning data into information in order to address some question. While there may be no such thing as a stupid question, there are ill-posed questions — those which cannot be answered as stated. Consider the Deepwater Horizon Case Study. It might seem natural to ask "if a volunteer cleans wildlife, will she develop adverse respiratory symptoms?" However, we quickly see that this is an ill-posed question. Of the Of 54 volunteers assigned to wildlife cleaning and rehabilitation, 15 reported experiencing adverse respiratory symptoms ("nose irritation, sinus problems, or sore throat"). So, while some volunteers developed symptoms, others did not. What makes the question ill-posed is *variability*, the fact that not every volunteer had the same reaction when directly exposed to oil.

It is variability that creates a need for statistics; in fact, you could think of statistics as the study and characterization of variability. We must therefore learn to ask the *right* questions — those which can be answered in the presence of variability.

Definition 3.1 (Variability)**.** The notion that measuremenets differ from one observation to another.

> **Key Idea:** The presence of variability makes some questions ill-posed; statistics concerns itself with how to address questions in the presence of variability.

3.1 Characterizing a Variable

Recall that the goal of statistical inference is to say something about the population; as a result, any question we ask should then be centered on this larger group. The first step to constructing a well-posed question is then to identify the population of interest for the study. For the Deepwater Horizon Case Study, it is unlikely that we are only interested in these 54 observed volunteers assigned to wildlife cleaning. In reality, we might want to say something about volunteers for any oil spill. In this case, the 54 volunteers in our dataset form the sample, a subset from all volunteers who clean wildlife following an oil spill. That is, our population of interest is comprised of all volunteers who clean wildlife following an oil spill.

> **Tip:** When identifying the population of interest, be specific! Are you really interested in *all* trees, for example? Or, are you interested in Maple trees within the city limits of Terre Haute, Indiana?

Since we expect that the reaction to oil exposure — the primary variable of interest for this study, sometimes called the **response** — to vary from one individual to another, we cannot ask a question about the *value* of

the reaction (whether they experienced symptoms or not). Instead, we want to characterize the **distribution** of the response.
Definition 3.2 (Response)**.** The primary variable of interest within a study.
Definition 3.3 (Distribution)**.** The pattern of variability corresponding to a set of values.

Notice that in this case, the response is a categorical variable; describing such a variable is equivalent to describing how individuals in are divided among the possible groups. With a finite number of observations, we could present the number of observations (**frequeny**) within each group. For example, of the 54 volunteers, 15 experienced adverse symptoms and 39 did not. This works well within the sample; however, as our population is infinitely large (all volunteers cleaning wildlife on any oil spill), reporting the frequencies is not appropriate. In this case, we report the fraction of observations (**relative frequency**) falling within each group.
Definition 3.4 (Frequency)**.** The number of observations falling into a particular level of a categorical variable.
Definition 3.5 (Relative Frequency)**.** Also called the "proportion," the fraction of observations falling into a particular level of a categorical variable.

Numeric quantities, like the proportion, which summarize the distribution of a variable within the population are known as **parameters**.
Definition 3.6 (Parameter)**.** Numeric quantity which summarizes the distribution of a variable within the *population* of interest.

While the *value* of a variable may vary across the population, the *parameter* is a single fixed constant which summarizes the variable for that population. Therefore, well-posed questions can be constructed if we limit ourselves to questions about the parameter. The second step in constructing well-posed questions is then to identify the parameter of interest.

The questions we ask then generally fall into one of two categories:

- Estimation: what *proportion* of volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms?
- Model Consistency: is it reasonable that no more than 1 in 5 volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms?

Now, since we do not get to observe the population (we only see the sample), we cannot observe the value of the parameter. That is, we will never know the true proportion of volunteers who will experience symptoms. However, we can determine what the data suggests about the population (that's inference).
Definition 3.7 (Estimation)**.** Using the sample to approximate the value of a parameter from the underlying population.
Definition 3.8 (Hypothesis Testing)**.** Using a sample to determine if the data is consistent with a working theory or if there is evidence to suggest the data is not consistent with the theory.

> **Key Idea:** Parameters are unknown values and can, in general, never be known. Parameters are generally denoted by Greek letters in statistical formulas.

It turns out, the vast majority of research questions can be framed in terms of a parameter. In fact, this is the first of what we consider the *Five Fundamental Ideas of Inference*.

> **Fundamental Idea:** **Fundamental Idea I**: A research question can often be framed in terms of a parameter which characterizes the population. Framing the question should then guide our analysis.

3.2 Framing the Question

In engineering and scientific applications, many questions fall under the second category of model consistency. Examining such questions is known as **hypothesis testing**. Specifically, data is collected to help the researcher choose between two competing theories for the parameter of interest. In this section, we consider the terminology surrounding specifying such questions.

For the Deepwater Horizon Case Study suppose we are interested in addressing the following question:

> Is there evidence that more than 1 in 5 volunteers who clean wildlife following an oil spill will develop adverse respiratory symptoms?

The question itself is about the population (all volunteers assigned to clean wildlife following an oil spill) and is centered on a parameter (the proportion who develop adverse respiratory symptoms). That is, this is a well-posed question that can be answered with appropriate data. The overall process for addressing these types of questions is similar to conducting a trial in a court of law. In the United States, a trial has the following essential steps:

1. Assume the defendant is innocent.
2. Evidence to establish guilt (to the contrary of innocence) is presented by the prosecution.
3. The jury considers the weight of the evidence.
4. If the evidence is "beyond a reasonable doubt," the jury declares the defendant guilty; otherwise, the jury declares the subject not guilty.

The process of conducting a hypothesis test has similar essential steps:

1. Assume the innocent of what we want the data to show (develop a working theory).
2. Gather data and compare it to the proposed model.
3. Quantify the likelihood of our data under the proposed model.
4. If the likelihood is small, conclude the data is not consistent with the working model (there is evidence for what we want to show); otherwise, conclude the data is consistent with the working model (there is no evidence for what we want to show).

Notice that a trial focuses not on proving guilt but on disproving innocence; similarly, in statistics, we are able to establish evidence *against* a specified theory. This process has several subtle points including this one. We will discuss these subtleties at various points throughout the text and revisit the overall concepts often. Here, we focus solely on that first step — developing a working theory that we want to *disprove*.

Consider the above question for the Deepwater Horizon Case Study. We want to find evidence that the proportion experiencing adverse symptoms exceeds 0.20 (1 in 5). Therefore, we would like to *disprove* (or provide evidence *against*) the statement that the proportion experiencing adverse symptoms is no more than 0.20. This is known as the **null hypothesis**; the opposite of this statement, called the **alternative hypothesis**, captures what we would like to establish.

Definition 3.9 (Null Hypothesis)**.** The statement (or theory) that we would like to *disprove*. This is denoted H_0, read "H-naught" or "H-zero".

Definition 3.10 (Alternative Hypothesis)**.** The statement (or theory) capturing what we would like to provide evidence *for*; this is the opposite of the null hypothesis. This is denoted H_1 or H_a, read "H-one" and "H-A" respectively.

For the Deepwater Horizon Case Study, we write:

> H_0 : The proportion of volunteers assigned to wildlife following an oil spill who experience adverse respiratory symptoms is no more than 0.20.
> H_1 : The proportion of volunteers assigned to wildlife following an oil spill who experience adverse respiratory symptoms exceeds 0.20.

Each hypothesis is a well-posed statement (about a parameter characterizing the entire population), and the two statements are exactly opposite of one another meaning only one can be a true statement. We can now collect data to determine if it is consistent with the null hypothesis (a statement similar to "not guilty") or if

the data provides evidence against the null hypothesis and in favor of the alternative (a statement similar to "guilty").

Often these statements are written in a bit more of a mathematical structure in which a Greek letter is used to represent the parameter of interest. For example, we might write

> Let θ be the proportion of volunteers (assigned to wildlife following an oil spill) who experience adverse respiratory symptoms.
> $H_0 : \theta \leq 0.20$
> $H_1 : \theta > 0.20$

In the above statements, θ represents the parameter of interest; the value 0.20 is known as the **null value**.
Definition 3.11 (Null Value)**.** The value associated with the equality component of the null hypothesis; it forms the threshold or boundary between the two hypothesis. Note: not all questions of interest require a null value be specified.

Key Idea: Hypothesis testing is a form of statistical inference in which we quantify the evidence *against* a working theory (captured by the null hypothesis). We essentially argue that the data supports the alternative if it is not consistent with the working theory.

Tip: Process for Framing a Question In order to frame a research question, consider the following steps:

1. Identify the population of interest.
2. Identify the parameter(s) of interest.
3. Determine if you are interested in estimating the parameter(s) or quantifying the evidence against some working theory.
4. If you are interested in testing a working theory, make the null hypothesis the working theory and the alternative the exact opposite statement (what you want to provide evidence for).

Chapter 4

Gathering the Evidence (Data Collection)

Consider again the goal of statistical inference — to use a sample as a snapshot to say something about the underlying population (Figure 4.1). This generally provokes unease in people, leading to a distrust of statistical results. In this section we attack that distrust head on.

4.1 What Makes a Sample Reliable

If we are going to have some amount of faith in the statistical results we produce, we must have data in which we can place our trust. *The Treachery of Images* (Figure @**??**fig:data-pipe-img)) is a canvas painting depicting a pipe, below which the artist wrote the French phrase for "This is not a pipe." Regarding the painting, the artist said

> The famous pipe. How people reproached me for it! And yet, could you stuff my pipe? No, it's just a representation, is it not? So if I had written on my picture "This is a pipe," I'd have been lying!

Just as a painting is a representation of the object it depicts, so a sample should be a representation of the underlying population from which it was taken. This is the primary requirement if we are to rely on the resulting data.

> **Key Idea:** In order for a statistical analysis to be reliable, the sample must be *representative* of the underlying population.

We need to be careful to not get carried away in our expectations. What constitutes "representative" really depends on the question, just as an artist chooses his depiction based on how he wants to represent the object. Let's consider the following example.

Example 4.1 (School Debt)**.** In addition to a degree, college graduates also tend to leave with a large amount of debt due to college loans. In 2012, a graduate with a student loan had an average debt of $29,400; for graduates from private non-profit institutions, the average debt was $32,300[1].

Suppose we are interested in determining the average amount of debt in student loans carried by a graduating senior from Rose-Hulman Institute of Technology, a small private non-profit engineering school. There are many faculty at Rose-Hulman who choose to send their children to the institute. Since I am also on the

[1] http://ticas.org/sites/default/files/pub_files/Debt_Facts_and_Sources.pdf

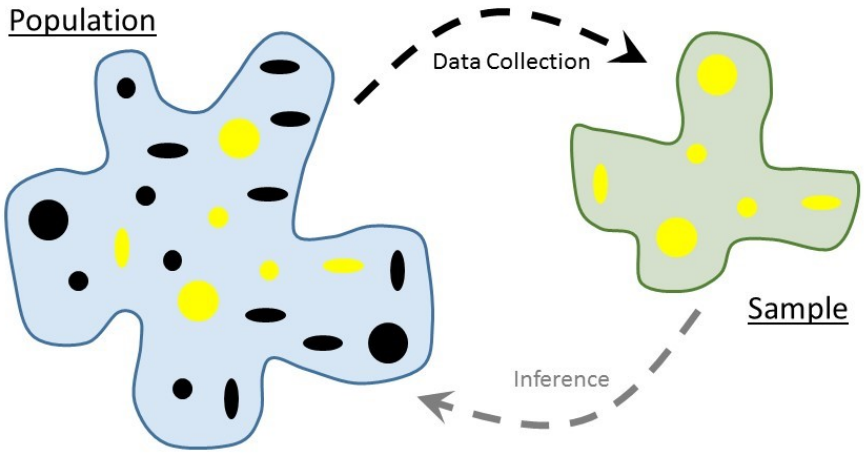


Figure 4.1: Illustration of the statistical process (reprinted from Chapter 1).

Figure 4.2: *The Treachery of Images* by René Magritte.

faculty, I know many of these individuals. Suppose I were to ask each to report the amount of student loans their children carried upon graduation from Rose-Hulman. I compile the 25 responses and compute the average amount of debt. Further, I report that based on this study, there is significant evidence that the average debt carried by a graduate of Rose-Hulman is far below the $32,300 reported above (great news for this year's graduating class)! Why might we be hesitant to trust these results?

When our distrust of a statistical result stems from a distrust of the data on which it is based, it is generally a result of our doubting the sample is representative of the population. Rose-Hulman, like many other universities, has a policy that the children of faculty may attend their university (assuming admittance) tuition-free. We would therefore expect their children to carry much less debt than the typical graduating senior.

This provides a nice backdrop for discussing what it means to be representative. First, let's define our population; in this case, we are interested in graduating seniors. The variable of interest is the amount of debt carried in student loans; the parameter of interest is then the average amount of debt in student loans carried by graduating seniors. With regard to the grade point average of the students in our sample, it is probably similar to all graduating seniors. Their starting salary is probably similar; the fraction of mechanical engineering majors versus math majors is probably similar. So, in many regards the sample is representative of the population; however, it fails to be representative with regard to the variable of interest. This is our concern. The amount of debt carried by students in our sample is not representative of that debt carried by all graduating seniors.

> **Tip:** When thinking about whether a sample is representative, focus your attention to the characteristics specific to your research question.

Does that mean the sample is useless? Yes and no. The sample collected cannot be used to answer our initial question of interest. No statistical method can fix bad data; statistics adheres to the "garbage-in, garbage-out" phenomena. If the data is bad, no analysis will undo that. While the sample cannot be used to answer our initial question, it could be used to address a different question however:

> What is the average amount of debt in student loans carried by graduating seniors from Rose-Hulman whose parent is a faculty member at the university?

For this revised question, the sample may indeed be representative. If we are working with previously collected data, we must consider the population to which our results will generalize. That is, for what population is the given sample representative. If we are collecting our data, we need to be sure we collect data in such a way that the data is representative. Let's first look at what *not* to do.

4.2 Poor Methods of Data Collection

Example 4.1 is an example of a "convenience sample," when the subjects in the sample are chosen simply due to ease of collection. Examples include surveying students only in your sorority when you are interested in all females who are part of a sorority on campus; taking soil samples from only your city when you are interested in the soil for the entire state; and, obtaining measurements from only one brand of phone, because it was the only one you could afford on your budget, when you are interested in studying all cell phones on the market. A convenience sample is unlikely to be representative if there is a relationship between the ease of collection and the variable under study. This was true in the School Debt example; the relationship of a student to a faculty member was directly related to the amount of debt they carried. As a result, the resulting sample was not representative of the population.

When conducting a survey with human subjects, it is common to only illicit responses from volunteers. Such "volunteer samples" tend to draw in those with extreme opinions. Consider product ratings on Amazon. Individual ratings tend to cluster around 5's and 1's. This is because those customers who take time to submit a review (which is voluntary) tend to be those who are really thrilled with their product (and want to

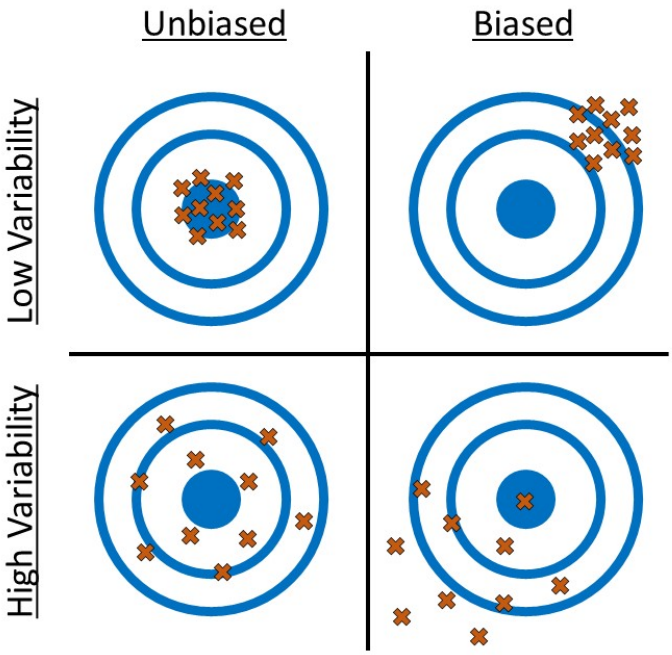


Figure 4.3: Illustration of bias and variability.

encourage others to purchase it) and those who are really disappointed with their purchase (and want to encourage others to avoid it). Such surveys often fail to capture those individuals in the population who have intermediate opinions.

We could not possibly name all the poor methods for collecting a sample; but, these methods all share something in common — it is much more likely the resulting sample is not representative. Failing to be representative results in **biased** estimates of the parameter.

Definition 4.1 (Bias)**.** A set of measurements, or an estimate of a parameter, is said to be biased if they are *consistently* too high (or too low).

To illustrate the concept of bias, consider shooting at a target as in Figure 4.3. We can consider the center of our target to be the parameter we would like to estimate within the population. The values in our sample (the strikes on the target) will vary around the parameter; while we do not expect any one value to hit the target precisely, a "representative" sample is one in which the values tend to be clustered about the parameter (unbiased). When the sample is not representative, the values in the sample tend to cluster off the mark (biased). Notice that to be unbiased, it may be that not a single value in the sample is perfect, but aggregated together, they point in the right direction. So, bias is not about an individual measurement being an "outlier," (more on those in a later chapter) but about repeatedly shooting in the wrong direction.

Key Idea: Biased results are typically due to poor sampling methods that result in a sample which is not representative of the underlying population.

The catch (there is always a catch) is that we will never know if a sample is representative or not. But, we can employ methods that help to minimize the chance that the sample is biased.

4.3 Preferred Methods of Sampling

No method guarantees a perfectly representative sample; but, we can take measures to reduce or eliminate bias. A useful strategy is to employ *randomization*. This is summarized in our second Fundamental Idea:

> **Fundamental Idea:** **Fundamental Idea II**: If data is to be useful for making conclusions about the population, a process referred to as drawing inference, proper data collection is crucial. Randomization can play an important role ensuring a sample is representative and that inferential conclusions are appropriate.

Consider the School Debt example again. Suppose instead of the strategy described there, we had constructed a list of all graduating seniors from the university. We placed the name of each student on an index card; then, I thoroughly shuffle the cards and choose the top 25 cards. For these 25 individuals, I record the amount of debt in student loans each carries. Using a lottery to select the sample is known as a **simple random sample**. By conducting a lottery, we make it very unlikely that our sample consists of only students with a very small amount of student debt (as occurred when we used a convenience sample).

Definition 4.2 (Simple Random Sample)**.** Often abbreviated SRS, this is a sample of size n such that *every* collection of size n is equally likely to be the resulting sample. This is equivalent to a lottery.

There are situations in which a simple random sample does not suffice. Again, consider our School Debt example. The Rose-Hulman student body is predominantly domestic, with only about 3% of the student body being international students. But, suppose we are interested in comparing the average debt carried between international and domestic students. It is very likely that in a simple random sample of 25 students, none will be international by chance alone. Instead of a simple random sample, we might consider taking a sample of, say 13, domestic students and a sample of 12 international students; this is an example of a **stratified random sample**. This approach is useful when there is a natural grouping of interest within the population.

Definition 4.3 (Stratified Random Sample)**.** A sample in which the population is first divided into groups, or strata, based on a characteristic of interest; a simple random sample is then taken within each group.

There are countless sampling techniques used in practice. The two described above can be very useful starting point for developing a custom method suitable for a particular application. Their benefit stems from their use of randomization.

This section is entitled "Preferred Methods" because while these methods are ideal, they are not always practical. For example, consider the Deepwater Horizon Case Study; conceptually, we can take a simple random sample of the volunteers for our study. However, as with any study involving human subjects, researchers would be required to obtain consent from each subject in the study. That is, a volunteer has the right to refuse to participate in the study. Therefore, it is unlikely that a simple random sample as described above could be obtained. Again, the key is to obtain a *representative* sample; while random selection may be a nice tool for accomplishing this, we may need to appeal to the composition of the sample itself to justify its use. Based on the characteristics of those willing to participate in the study, do we feel the study participants form a representative group of all volunteers? That is the essential question. This is often why studies report a table summarizing subject demographics such as age, gender, etc. It is also why it is extremely important for researchers to describe how subjects were selected so that readers may make the judgement for themselves whether the sample is representative.

4.4 Two Types of Studies

Thinking about how the data was collected helps us determine how the results generalize beyond the sample itself (to what population the results apply). When our question of interest is about the relationship between two variables (as most questions are), we must also carefully consider the study design. Too often separated

from the statistical analysis that follows, keeping the study design in mind should guide the analysis as well as inform us about the conclusions we can draw.

In order to illustrate how study design can impact the results, consider the following example.
Example 4.2 (Kangaroo Care)**.** At birth, infants have low levels of Vitamin K, a vitamin needed in order to form blood clots. Though rare, without the ability for her blood to clot, an infant could develop a serious bleed. In order to prevent this, the American Academy of Pediatrics recommends that all infants be given a Vitamin K shot shortly after birth in order to raise Vitamin K levels. As with any shot, there is typically discomfort to the infant, which can be very discomforting to new parents.

Kangaroo Care is a method of holding a baby which emphasizes skin-to-skin contact. The child, who is dressed only in a diaper, is placed upright on the parent's bare chest; a light blanket is draped over the child. The method was initially recognized for its benefits in caring for pre-term infants. Suppose suppose we are interested in determining if utilizing the method while giving the child a Vitamin K shot reduces the discomfort in the infant, as measured by the total amount of time the child cries following the shot. Contrast the following two potential study designs:

(A) We allow the attending nurse to determine whether Kangaroo Care is initiated prior to giving the Vitamin K shot. Following the shot, we record the total time (in seconds) the child cries.
(B) We flip a coin. If it comes up heads, the nurse should have the parents implement Kangaroo Care prior to giving the Vitamin K shot; if it comes up tails, the nurse should give the Vitamin K shot without implementing Kagaroo Care. Following the shot, we record the total time (in seconds) the child cries.

Note, in both study designs (A) and (B), we only consider term births which have no complications to avoid potential complications that might alter the timing of the Vitamin K shot or the ability to implement Kangaroo Care.

Note that there are some similarities in the two study designs:

- The underlying population is the same for both designs — infants born at term with no complications.
- There are two treatment groups in both designs — the "Kangaroo Care" group and the "no Kangaroo Care" group.
- The response (variable of interest) is the same in both designs — the time (in seconds) the infant cries.
- There is action taken by the researcher in both designs — a Vitamin K shot is given to the child.

There is one prominent difference between the two study designs:

- For design (A), the choice of Kangaroo Care is left up to the nurse (self-selected); for design (B), the choice of Kangaroo is *assigned* to the nurse by the researcher, and this selection is made *at random*.

Design (A) is an example of an **observational study**; design (B) is a **controlled experiment**.
Definition 4.4 (Observational Study)**.** A study in which the subjects self-select into the treatment groups under study.
Definition 4.5 (Controlled Experiment)**.** A study in which the subjects are *randomly* assigned to the treatment groups under study.

It is common to think that anytime the environment is "controlled" by the researcher that an experiment is taking place, but the defining characteristic is the random assignment to treatment groups (sometimes referred to as the *factor* under study). In the example above, both study designs involved a controlled setting (the delivery room of a hospital) in which trained staff (the nurse) delivered the shot. However, only design (B) is a controlled experiment because the researchers randomly determined into which group the infant would be placed.

To understand the impact of random allocation, suppose that we had conducted a study as in design (A); further, the results suggest that those infants who were given a shot while using Kangaroo Care cried for a shorter time period, on average. Can we conclude that it was the Kangaroo Care that led to the shorter crying time? Maybe. Consider the following two potential explanations for the resulting data:

(1) Kangaroo Care is very effective; as a result, those children who were given Kangaroo Care had reduced crying time following the Vitamin K shot.

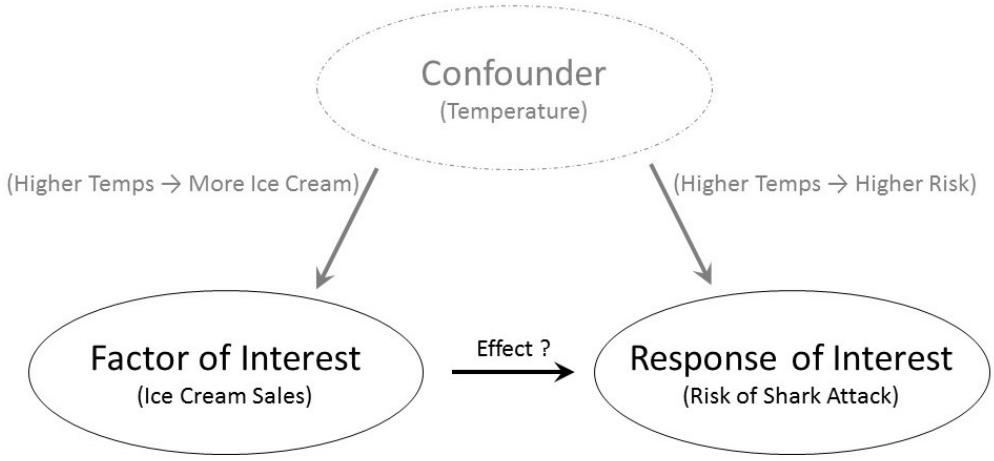


Figure 4.4: Illustration of a confounding variable. The confounder, related to both the factor and the treatment can make it appear as though there is a causal relationship when none exists.

(2) It turns out that those nurses who chose to implement Kangaroo Care (remember, they have a choice under design (A) whether they implement the method) were also the nurses with a gentler bedside manner. Therefore, these nurses tended to be very gentle when giving the Vitamin K shot whereas the nurses who chose not to implement Kangaroo Care tended to just jab the needle in when giving the shot. As a result, the reduced crying time is not a result of the Kangaroo Care but the manner in which the shot was given.

The problem is that we are unable to determine which of the explanations is true. Given the data we have collected, we are unable to tease out the effect of the Kangaroo Care from that of the nurse's bedside manner. As a result, we are able to say we observed a *relationship* between the use of Kangaroo Care and reduced crying time, but we are unable to conclude that Kangaroo Care *caused* a reduction in the crying time. In this hypothetical scenario, the nurse's bedside manner is called a **confounder**.

Definition 4.6 (Confounding)**.** When the effect of a variable on the response is mis-represented due to the presence of a third, potentially unobserved, variable known as a confounder.

Confounders can mask the relationship between the factor under study and the response. There is a documented relationship between ice cream sales and the risk of shark attacks. As ice cream sales increase, the risk of a shark attack also increases. This does not mean that if a small city in the Midwest increases its ice cream sales that the citizens are at higher risk of being attacked by a shark. As Figure 4.4 illustrates, there is a confounder — temperature. As the temperatures increase, people tend to buy more ice cream; as the temperature increases, people tend to go to the beach increasing the risk of a shark attack. Two variables can appear to be related as a result of a confounder.

Tip: Confounders are variables that influence *both* the factor of interest and the response.

Observational studies are subject to confounding; thus, controlled experiments are often considered the gold standard in research because they allow us to infer cause-and-effect relationships from the data. Why does the random allocation make such an impact? Because it removes the impact of confounders. Let's return to the hypothetical study. Suppose there are nurses with a gentle bedside manner and those who are a little less gentle. If the infants are randomly assigned to one of the two treatment groups, then for every gentle nurse who is told to implement Kangaroo Care while giving the shot, there is a gentle nurse who is told to not implement Kangaroo Care. Similarly, for every mean nurse who is told to implement Kangaroo Care

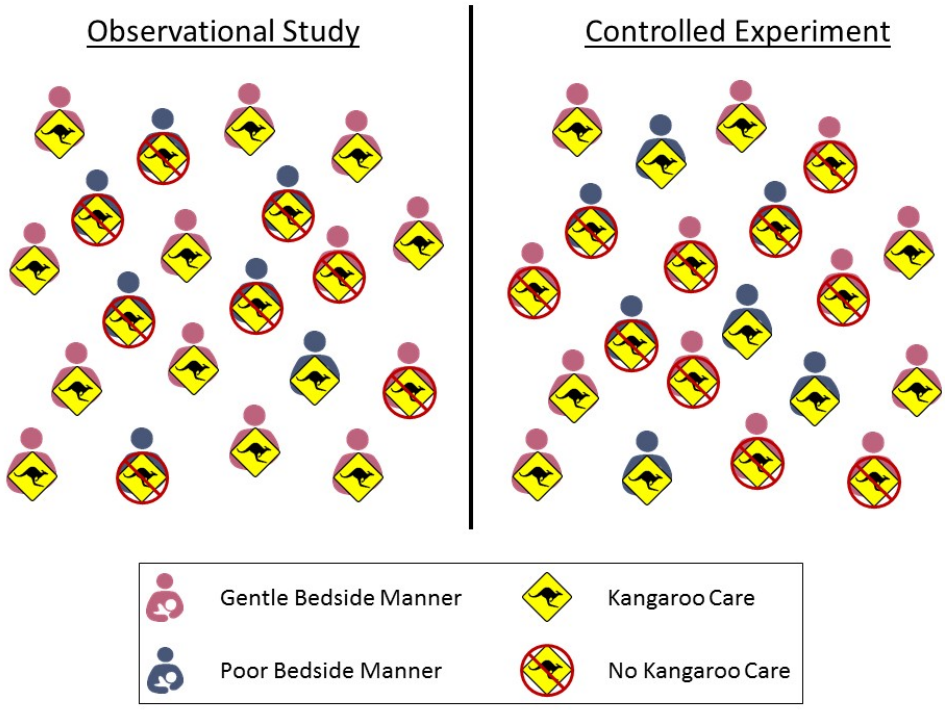


Figure 4.5: Illustration of the impact of random assignment in study design. For the observational study, the treatment groups are unbalanced. For the controlled experiment, the treatment groups are balanced.

while giving a shot, there is a mean nurse who is told to not implement Kangaroo Care. This is illustrated in Figure 4.5. For an observational study, the treatment groups are unbalanced; there is a higher fraction (11/12 compared to 1/4) of friendly nurses in the Kangaroo Care group compared to the No Kangaroo Care group. For the controlled experiment however, the treatment groups are balanced; there is approximately the same fraction of friendly nurses in both groups. Random assignment is the great equalizer; it tends to result in groups which are similar in all respects; therefore, any differences we observe between the groups *must* be due to the grouping and not an underlying confounding variable.

Key Idea: Randomly assigning treatment groups balances the groups with respect to the confounders; that is, the treatment groups are similar. Therefore, any differences between the two groups can be attributed to the grouping factor itself.

While controlled experiments are a fantastic study design, we should not discount the use of observational studies. Consider the Deepwater Horizon Case Study; suppose we are interested in the following question:

> Is there evidence that volunteers who are directly exposed to oil have an increased risk of developing adverse respiratory symptoms compared to those who are not directly exposed to oil?

The response is whether a volunteer develops adverse respiratory symptoms; the factor of interest is whether the volunteer has direct exposure to oil. We could conduct a controlled experiment by randomly determining which volunteers are assigned to wildlife clean up and which are assigned to administrative tasks, for example. However, it may be that volunteer tasks need to be determined by skillset or by greatest need at the time of the person volunteers. It may not be feasible to randomly assign volunteers to specific positions. Or, it could be that the data was obtained after the fact; that is, the data is not the result of a planned study in which case random assignment is not possible because volunteers self-selected into positions in the past. If random assignment is not possible, it does not mean the data is useless. But, it does mean we will need to be sure we

address the potential confounding when performing the analysis and discussing the results. The latter half of the text will discuss methods for addressing confounding.

The big idea is that in order to make causal conclusions, we must be able to state that the two treatment groups are balanced with respect to any potential confounders; random assignment is one technique for accomplishing this.

Chapter 5

Presenting the Evidence (Summarizing Data)

If you open any search engine and look up "data visualization," you will be quickly overwhelmed by a host of pages, texts, and software filled with tools for summarizing your data. Here is the bottom line: a good visualization is one that helps you answer your question of interest. It is both that simple and that complicated.

> **Fundamental Idea:** **Fundamental Idea III**: The use of data for decision making requires that the data be summarized and presented in ways that address the question of interest.

Whether simple or complex, all graphical and numerical summaries should help turn the data into usable information. Pretty pictures for the sake of pretty pictures are not helpful. In this section, we will consider various simple graphical and numerical summaries to help build a case for addressing the question of interest.

5.1 Characteristics of a Distribution (Summarizing a Single Variable)

Remember that because of *variability*, the key to asking good questions is to not ask questions about individual values but to characterize the underlying *distribution* (see Definition 3.3). Therefore, characterizing the underlying distribution is also the key to a good visualization or numeric summary. For the Deepwater Horizon Case Study, the response (whether a volunteer experienced adverse respiratory symptoms) is categorical. As we stated previously, summarizing the distribution of a categorical variable reduces to showing how individual subjects fall into the various groups. Figure 5.1 displays a *bar chart* summarizing the rate of respiratory symptoms for volunteers cleaning wildlife.

In general, it does not matter whether the frequency or the relative frequencies are reported; however, if the relative frequencies are plotted, some indication of the sample size should be provided with the figure either as an annotation or within the caption. Statisticians tend to agree that bar charts are preferable to pie charts (see this whitepaper and this blog for further explanation). More importantly, we all agree that the graphic should help address the question. From the above graphic, we see that nearly 28% of volunteers assigned to wildlife experienced adverse respiratory symptoms.

Summarizing the distribution of a numeric variable requires a bit more thought. Consider the following example.

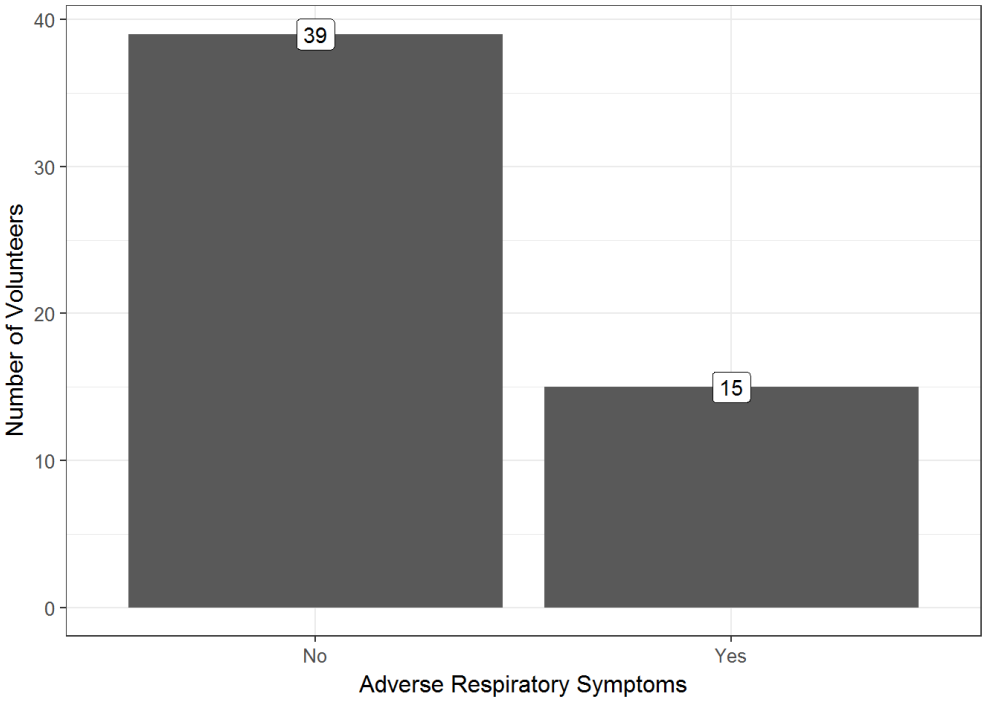


Figure 5.1: Frequency of adverse respiratory symptoms for volunteers cleaning wildlife following the Deepwater Horizon oil spill.

Table 5.1: Breaking length (km) for first 5 specimens in the Paper Strength study.

Specimen	Breaking Length
1	21.312
2	21.206
3	20.709
4	19.542
5	20.449

Example 5.1 (Paper Strength)**.** While electronic records have become the predominant means of storing information, we do not live in a paperless society. Paper products are still used in a variety of applications ranging from printing reports and photography to packaging and bathroom tissue. In manufacturing paper for a particular application, the strength of the resulting paper product is a key characteristic.

There are several metrics for the strength of paper. A conventional metric for assessing the inherent (not dependent upon the physical characteristics, such as the weight of the paper, which might have an effect) strength of paper is the *breaking length*. This is the length of a paper strip, if suspended vertically from one end, that would break under its own weight. Typically reported in kilometers, the breaking length is computed from other common measurements. For more information on paper strength measurements and standards, see the following website: http://www.paperonweb.com

A study was conducted at the University of Toronto to investigate the relationship between pulp fiber properties and the resulting paper properties (Lee 1992). The breaking length was obtained for each of the 62 paper specimens, the first 5 measurements of which are shown in Table 5.1. The complete data is available online at the following website: https://vincentarelbundock.github.io/Rdatasets/doc/robustbase/pulpfiber.html

While there are several questions one might ask with the available data, here we are primarily interested in characterizing the breaking length of these paper specimens.

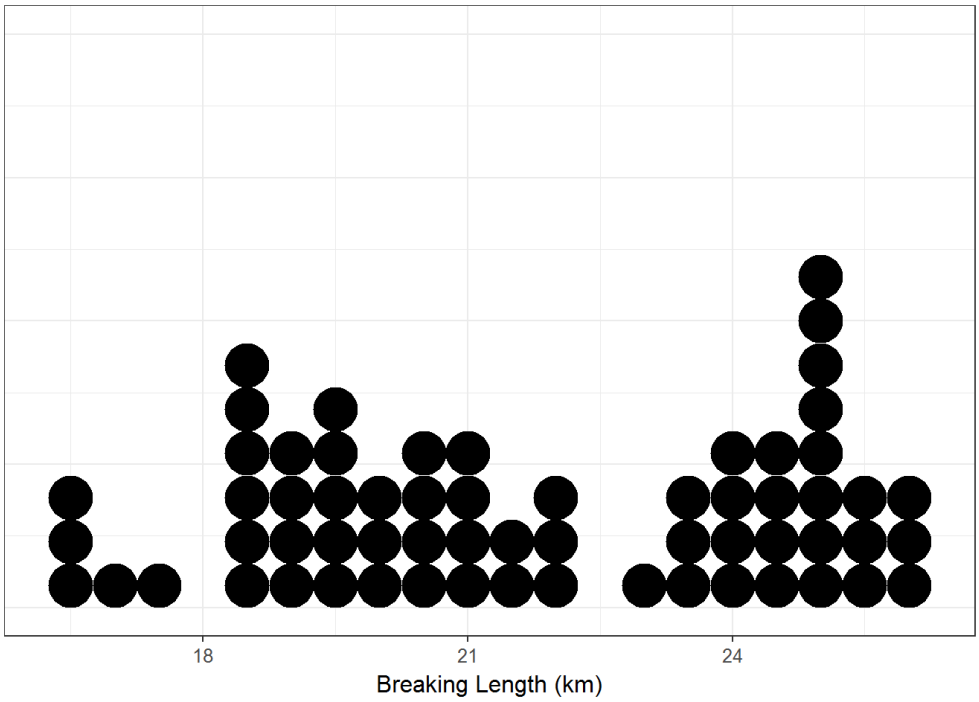


Figure 5.2: Breaking Length (km) for 62 paper specimens.

Figure 5.2 presents the breaking length for all 62 paper specimens in the sample through a *dot plot*.

With any graphic, we tend to be drawn to three components:

- *where* the values tend to be,
- *how tightly* the values tend to be clustered there, and
- *the way* the values tend to cluster.

Notice that about half of the paper specimens in the sample had a breaking length longer than 21.26 km. Only about 25% of paper specimens had a breaking length less than 19.33 km. These are measures of *location*. In particular, these are known as **percentiles**, of which the **median**, **first quartile** and **third quartile** are commonly used examples.

Definition 5.1 (Percentile)**.** The value q such that k% of the values in the distribution are less than or equal to q. For example,

- 25% of values in a distribution are less than or equal to the 25-th percentile (known as the first quartile).
- 50% of values in a distribution are less than or equal to the 50-th percentile (known as the median).
- 75% of values in a distribution are less than or equal to the 75-th percentile (known as the third quartile).

The **average** is also a common measure of location. The breaking length of a paper specimen is 21.72 km, on average. In this case, the average breaking length and median breaking length are very close; this need not be the case. The average is not describing the "center" of the data in the same way as the median; they capture different properties.

Definition 5.2 (Average)**.** Also known as the "mean," this measure of location represents the balance point for the distribution. It is denoted by $\bar{x}$.

For a sample of size n, it is computed by

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

where x_i rerpesents the i-th value in the sample.

When referencing the average for a population, it can also be called the "Expected Value," and is often denoted by μ.

Clearly, the breaking length is not equivalent for all paper specimens; that is, there is variability in the measurements. Measures of *spread* quantify the variability of values within a distribution. Common examples include the **standard deviation** (related to **variance**) and **interquartile range**. For the Paper Strength example, the breaking length varies with a standard deviation of 2.88 km; the interquartile range for the breaking length was 5.2 km. Neither of these values has a natural interpretation; instead, larger values of these measures simply indicate a higher degree of variability in the data. The standard deviation is often reported more often than the variance since it is on the same scale as the original data; however, as we will see later, the variance is useful from a mathematical perspective for derivations.
Definition 5.3 (Variance)**.** A measure of spread, this roughly captures the average distance values in the distribution are from the mean.

For a sample of size n, it is computed by

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

where $\bar{x}$ is the sample mean and x_i is the i-th value in the sample. The division by $n-1$ instead of n reduces the bias in the statistic.
Definition 5.4 (Standard Deviation)**.** A measure of spread, this is the square root of the variance.
Definition 5.5 (Interquartile Range)**.** The distance between the first and third quartiles. This measure of spread indicates the range over which the middle 50% of the data is spread.

The measures we have discussed so far are illustrated in Figure 5.3. While some authors suggest that which values are reported depend on the shape of the distribution, we argue that it is best to report the values that align with the question of interest. It is the question that should be shaped by the beliefs about the underlying distribution.

Finally, consider the *shape* of the distribution of breaking length we have observed. The breaking length tends to be clustered in two locations; we call this *bimodal* (each mode is a "hump" in the distribution). Other terms used to describe the shape of a distribution are *symmetric* and *skewed*. Symmetry refers to cutting a distribution in half (at the median) and the lower half being a mirror image of the upper half; skewed distributions are those which are not symmetric.

Observe then that the dot plot above gives us some idea of the location, spread, and shape of the distribution, in a way that the table of values could not. This makes it a useful graphic as it is characterizing the **distribution of the sample** we have observed. This is one of the distributions in what we call the *Distributional Quartet*.
Definition 5.6 (Distribution of the Sample)**.** The pattern of variability in the observed values of a variable.

When the sample is not large, a dot plot is reasonable. Other common visualizations for a single variable include a *jitter plot*, *box plot*, *histogram*, or *density plot* (smoothed histogram). To illustrate, the breaking length for the Paper Strength example is summarized using various methods in Figure 5.4. The latter three visualizations are more helpful when the dataset is very large and plotting the raw values actually hides the distribution. There is no right or wrong graphic; it is about choosing the graphic which addresses the question and adequately portrays the distribution.

The numeric summaries of a distribution are known as **statistics**. While parameters characterize a variable at the population level, statistics characterize a variable at the sample level.
Definition 5.7 (Statistic)**.** Numeric quantity which summarizes the distribution of a variable within a *sample*.

Why would we compute numerical summaries in the sample if we are interested in the population? Remember the goal of this discipline is to use the sample to say something about the underlying population. As

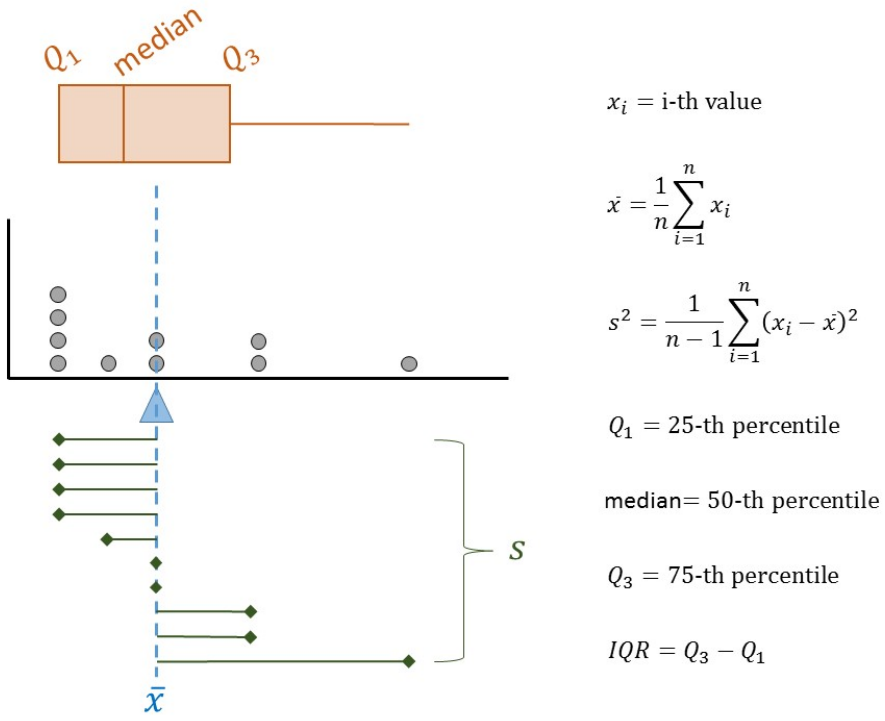


Figure 5.3: Illustration of measures of location and spread for a distribution of values.

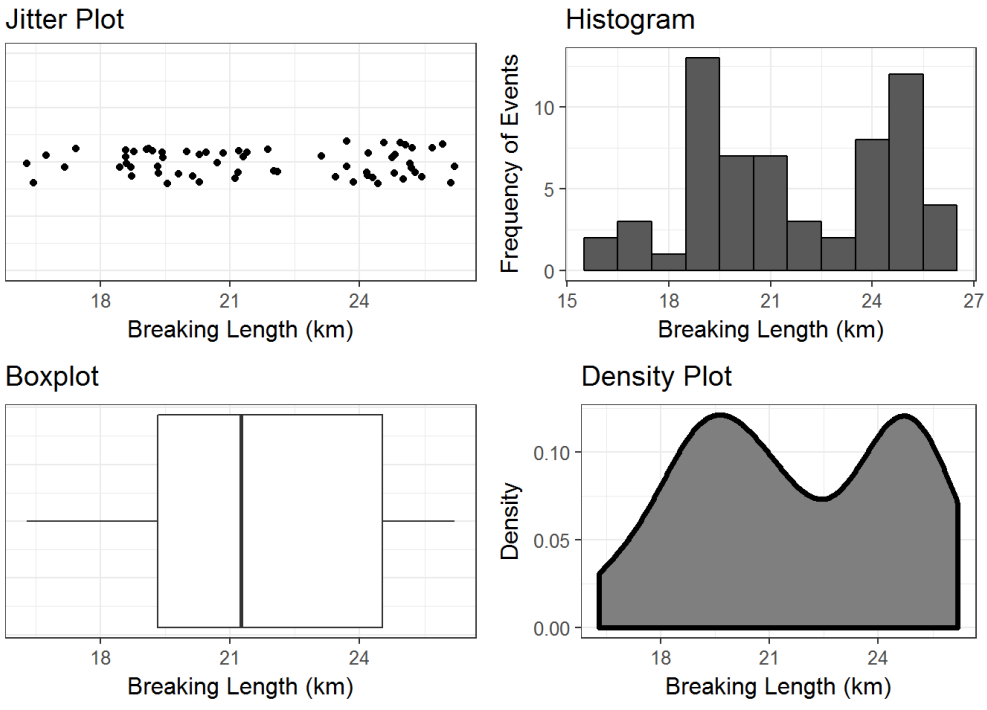


Figure 5.4: Four graphical summaries of the breaking length for the Paper Strength example.

long as the sample is representative, the distribution of the sample should reflect the **distribution of the population**; therefore, summaries of the sample should roughly equate to the analogous summaries of the population. Now we see the real importance of having a representative sample; it allows us to say that what we observe in the sample is a good proxy for what is happening in the population.
Definition 5.8 (Distribution of the Population)**.** The pattern of variability in values of a variable at the population level. Generally, this is impossible to know, but we might model it.

That is, the mean in the sample should approximate (estimate) the mean in the population; the standard deviation of the sample should estimate the standard deviation in the population; and, the shape of the sample should approximate the shape of the population, etc. The sample is acting as a representation in all possible ways of the population.

> **Key Idea:** A representative sample reflects the population; therefore, we can use statistics as estimates of the population parameters.

5.2 Summarizing Relationships

The summaries discussed above are nice for examining a single variable. In general, research questions of interest typically involve the relationship between two or more variables. Most graphics are two-dimensional (though 3-dimensional graphics and even virtual reality are being utilized now); therefore, summarizing a rich set of relationships may require the use of both axes, color, shape, size, and even multiple plots in order to tell the right story. We will explore these various features in upcoming units of the text. Here, we focus on the need to tell a story that answers the question of interest instead of getting lost in making a graphic. Consider the following question from the Deepwater Horizon Case Study:

> What is the increased risk of developing adverse respiratory symptoms for volunteers cleaning wildlife compared to those volunteers who do not have direct exposure to oil?

Consider the graphic in Figure 5.5; this is *not* a useful graphic. While it compares the number of volunteers with symptoms in each group, we cannot adequatly address the question because the research question involves comparing the rates for the two groups.

Instead, Figure 5.6 compares the rates within each group. Notice that since we are reporting relative frequencies for comparison, we report the sample size for each group.

From the graphic, it becomes clear that a higher fraction of volunteers cleaning wildlife experienced adverse symptoms compared with those without oil exposure. In fact, volunteers cleaning wildlife were 1.79 times more likely to experience adverse respiratory symptoms.

The key to a good summary is understanding the question of interest and addressing this question through a useful characterization of the variability.

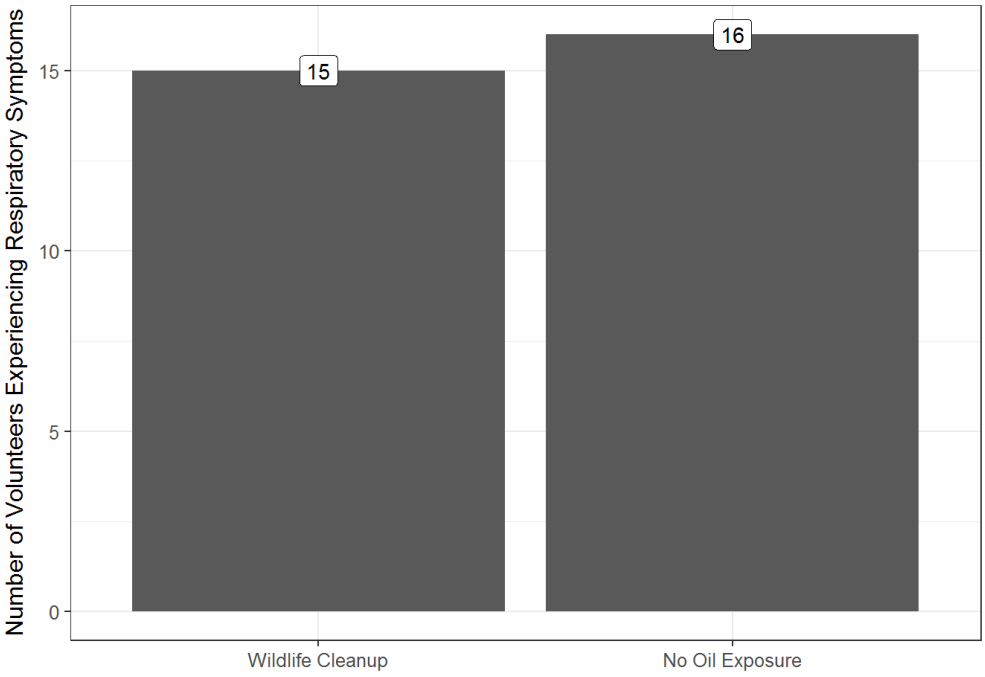


Figure 5.5: Illustration of a poor graphic; the graphic does not give us a sense of the rate within each group in order to make a comparison.

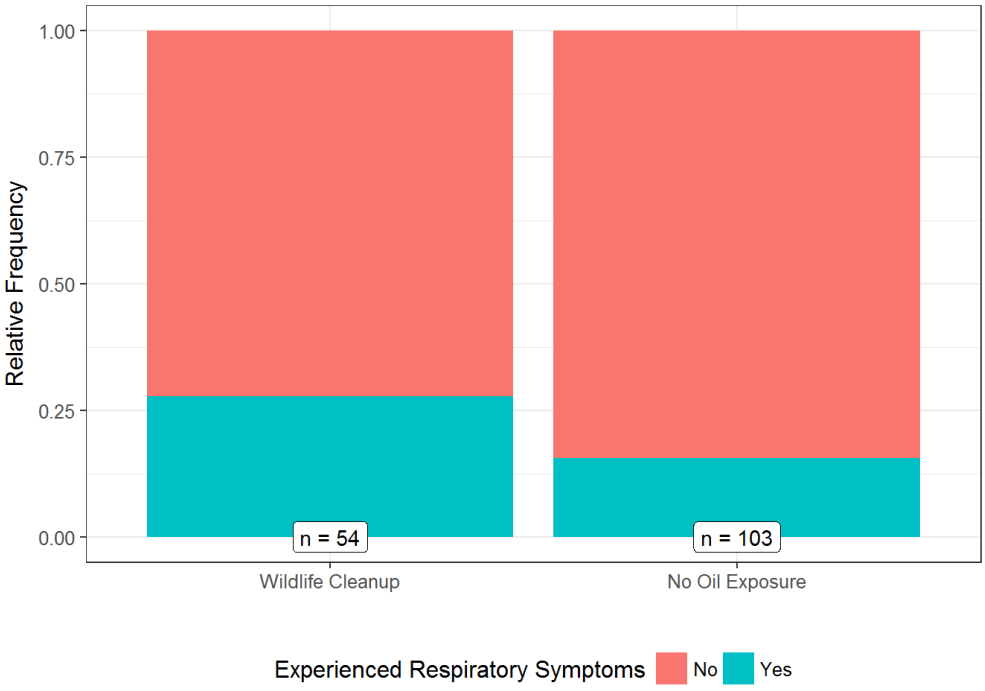


Figure 5.6: Comparison of the rate of adverse respiratory symptoms among volunteers assigned to different tasks.

Chapter 6

Assessing the Evidence (Quantifying the Variability in Estimates)

Again, the goal of statistical inference is to use the sample as a snapshot of the underlying population (Figure 6.1). There are generally three reasons people distrust this process:

1. Fear that the sample does not represent what is going on in the population.
2. Fear that we cannot make a conclusion with a sample of size n (wanting more data).
3. Fear that one study is not enough to make a conclusion.

We have already tackled the first reason in Chapter @ref(#Data); if we are to trust statistical results, we must collect data that is representative of the underlying population. The second and third fears above are tied together, though maybe not obviously. Before launching into a slightly more formal discussion, consider the following thought experiment.

Example 6.1 (Free Throws)**.** Your friends Dave lives for his Wednesday "pick-up" basketball games at the gym. One afternoon, while waiting for a few more players to arrive Dave shoots 10 free throws, of which he makes 3.

I imagine no one is ready to claim *definitively* that Dave has a 30% success rate from the free throw line. So, what can we say? Well, if this set of 10 free throws is representative of Dave's free throw performance, then we would say that 30% is an estimate for his success rate; that is, the statistic 30% is a good guess at the unknown parameter (overall success rate). There are two ways we might impove our confidence in this estimate. First, we might consider a larger sample size (make Dave shoot more free throws).

Example 6.2 (Free Throws (cont.))**.** Joe has also been waiting for a few more players to arrive; however, Joe shoots 100 free throws (clearly he has more time on his hands) of which he makes 30.

Again, we probably wouldn't claim *definitively* that Joe has a 30% success rate from the free throw line. But, assuming this set of 100 free throws is representative of his overall performance, then we would be more confident in our guess for Joe's overall performance compared with our guess for Dave's. The more shots we observe, the more confidence we have in our estimate. This idea is known as the **Law of Large Numbers**.

Definition 6.1 (Law of Large Numbers)**.** For our purposes, this essentially says that as a sample size gets really large, a statistic will become arbitrarily close (extremely good guess) of the parameter it estimates.

Unfortunately, we may not be able to take a really large sample. It is probably not feasible to have Dave or Joe shoot thousands of free throws, for example. Our goal then becomes to somehow quantify the confidence we have in our estimates *given the sample size we have available.* That is, given that we only saw Dave shoot 10 free throws, can we quantify our confidence in that 30% estimate of his free throw success? Our "confidence" in an estimate is tied to the estimate's repeatability — "if we were to repeat the study, how much would we expect our estimate to change?" This gets at the last fear; we know that if we repeat a study, the results will change. Our job is to quantify (keeping the sample size in mind) the degree to which

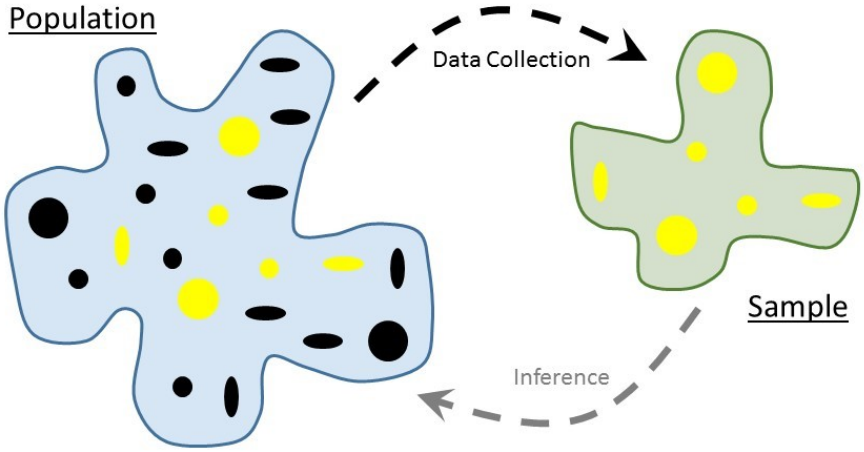


Figure 6.1: Illustration of the statistical process (reprinted from Chapter 1).

the results will change. That is, we need to quantify the *variability* in the estimate across repeated studies (known as sampling variability; we told you statistics was all about variability). This is known as a **sampling distribution**.

Definition 6.2 (Sampling Distribution)**.** The distribution of a *statistic* across repeated samples.

This is perhaps the most important of the *Distributional Quartet*; it is the holy grail of statistical inference. Once we have the sampling distribution, inference is straight-forward.

> **Fundamental Idea:** **Fundamental Idea IV**: Variability is inherent in any process, and as a result, our estimates are subject to sampling variability. However, these estimates often vary across samples in a predictable way; that is, they have a distribution that can be modeled.

6.1 Conceptualizing the Sampling Distribution

The sampling distribution of a statistic is one of the most fundamental, and yet one of the most abstract, concepts in statistics. It's name is even confusing; the "distribution of the sample" (Definition 5.6) and the "sampling distribution" (Definition 6.2) are two different things. In this section, we develop the idea of a sampling distribution; then, we turn toward actually constructing it.

For the Deepwater Horizon Case Study, consider the following question:

> What proportion of volunteers assigned to clean wildlife will develop adverse respiratory symptoms?

In the sample, we observed 15 out of 54 such volunteers (27.8% or a proportion of 0.278). This proportion is a good estimate of the rate of adverse symptoms in the population (assuming the sample is representative, of course). Now, imagine randomly selecting 54 new volunteers from the population (repeating the study). We could determine what fraction of volunteers in this new sample experienced adverse symptoms, expecting this value to be a bit different than what we obtained in the first sample. Since this second sample is also representative, it provides a good estimate of the parameter. Now, we could take a third random sample of 54 volunteers and compute the fraction in this third sample which experienced adverse symptoms. This third sample also provides a good (and potentially unique) estimate of the parameter. We could continue this process m times, for some large number m. This process is illustrated in Figure 6.2.

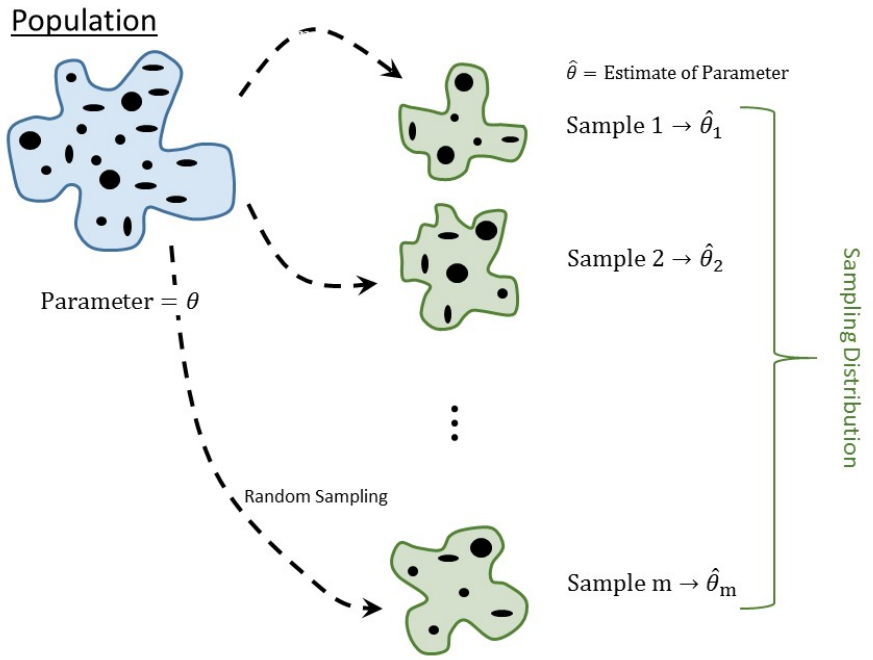


Figure 6.2: Illustration of repeatedly sampling from a population.

Consider what we are describing. With each representative sample, we have constructed an estimate of the parameter. What we have kept from each repetion is *not* the values of the variables themselves (whether the volunteers experienced adverse respiratory symptoms) but rather we have retained the *statistic* from each of m whole new studies. So, which of these m estimates do we trust? All of them. Since each sample is representative of the population, each estimate is a good (not perfect) estimate of the parameter. Since we have all these estimates, we could think about pooling the information from all of them; describing the way in which they change from one sample to another is the sampling distribution.

Notice that the sampling distribution is not describing a variable, it is describing a *statistic*. In order to construct a sampling distribution, we would go through the following steps:

1. Take a sample; record variables of interest.
2. Compute the statistic which estimates the parameter.
3. Repeat steps 1 and 2 a large number of times.
4. Examine the statistics collected.

So, the sampling distribution is not a plot of the raw values of a variable on individual subjects but a plot of statistics which summarize entire samples. That is, the unit of observation has changed. While a sample consists of individual subjects from the population, the sampling distribution consists of individual samples from the population.

Tip: Re-read the description of a sampling distribution several times, and return to it often as you read through the text. It takes a while for this to sink in, but if you truly grasp this one concept, the remainder of statistical inference becomes much more accessible.

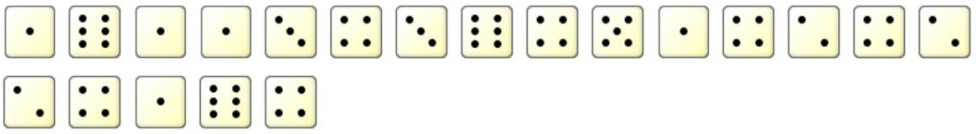


Figure 6.3: Potential sample of rolling a cie 20 times.

6.2 Example of a Sampling Distribution

Since this idea is so critical to grasping statistical inference, we are going to walk through the process of generating a sampling distribution for a known data generating process.

Example 6.3 (Dice Experiment)**.** Consider an ordinary six-sided die; we are interested in the proportion of times that rolling the die will result in a 1. Putting this in the language of the statistics, we have the following:

- The *population* of interest is all rolls of the die. Notice that this population is infinitely large as we could roll the die forever.
- The *variable* is the resulting value from the roll. Since this can take on only one of six values, this is a categorical variable.
- The *parameter* of interest is the proportion of rolls that result in a 1.

Our goal is to construct the sampling distribution of the proportion of rolls that result in a 1 when the die is rolled 20 times.

What makes this example unique is that we know the value of the parameter. Because of the physical properties of a die, we know that the probability a roll results in a 1 is $\theta = 1/6$. So, statistical inference is not needed here. This example simply provides a simple vehicle for studying sampling distributions. Going back to the steps for creating a sampling distribution described in the previous section, we have the following steps:

1. Roll a die 20 times, each time recording the resulting value.
2. Compute the proportion of times (out of the 20) the resulting value was a 1.
3. Repeat steps 1 and 2 a large number of times (let's say 500).
4. Plot the resulting values; there should be 500 proportions that we are keeping.

Notice that we are actually rolling a die 10000 times (20 rolls repeated 500 times); we only keep 500 values (one proportion for each set of 20 rolls). This is something you could physically do at home. For example, the first sample might look like that in Figure 6.3.

For this particular sample, the proportion in the sample (our statistic of interest) would be 0.25 (5/20). That is the value we would record. We then repeat this 499 more times. You could try a few out yourself using an online simulator. Figure 6.4 shows the resulting proportions for 500 samples of size 20 each.

With modern computing power, there is no need to restrain ourselves to repeating the study 500 times. A simple computer program could replicate rolling the dice thousands of times. Figure 6.5 is the sampling distribution for the proportion of rolls that result in a 1 based on a sample of size 20 repeating the study 50000 times.

Notice that the sampling distribution is centered around the true value of the parameter ($\theta = 1/6$). In general, the sampling distribution of statistics, when taken from a random sample, center on the true value of the parameter. This is the unbiased nature of the data coming out; random samples are representative of the population. Similarly, note that while no one sample (remember, each value in the distribution represents a statistic from a sample of 20 values) is perfect, no samples produce values which are far from the true parameter. That is, a representative sample may not be perfect, but it will give a *reasonable* estimate of the parameter. Notice that these properties hold even though we had a relatively small sample size ($n = 20$ coin flips).

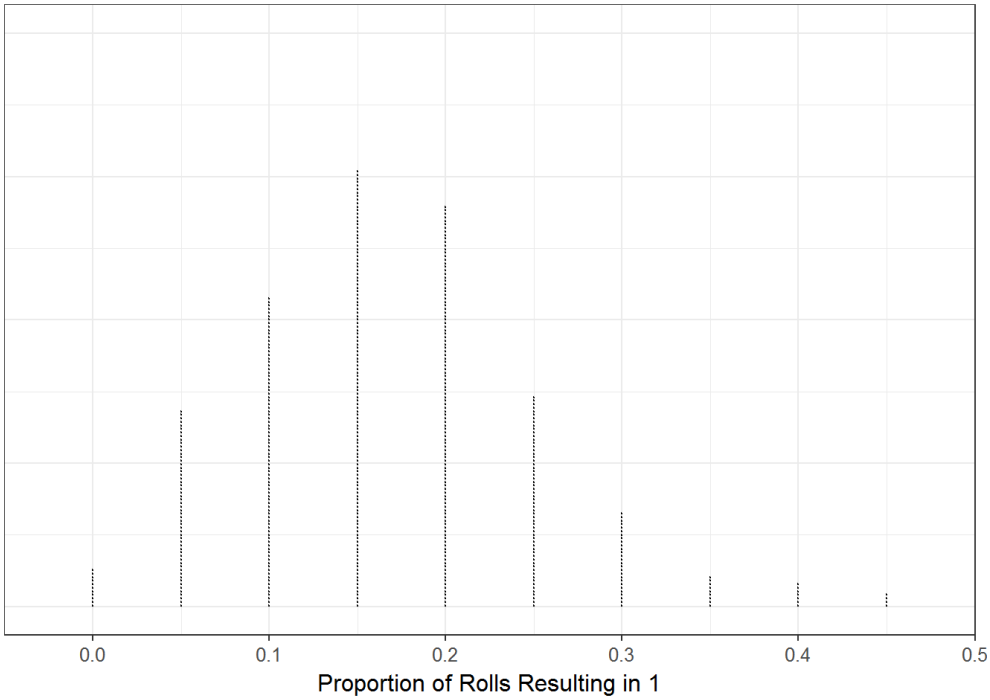


Figure 6.4: Sampling distribution for the proportion of 20 dice rolls which result in a 1. The distribution is based on repeating the sampling process 500 times.

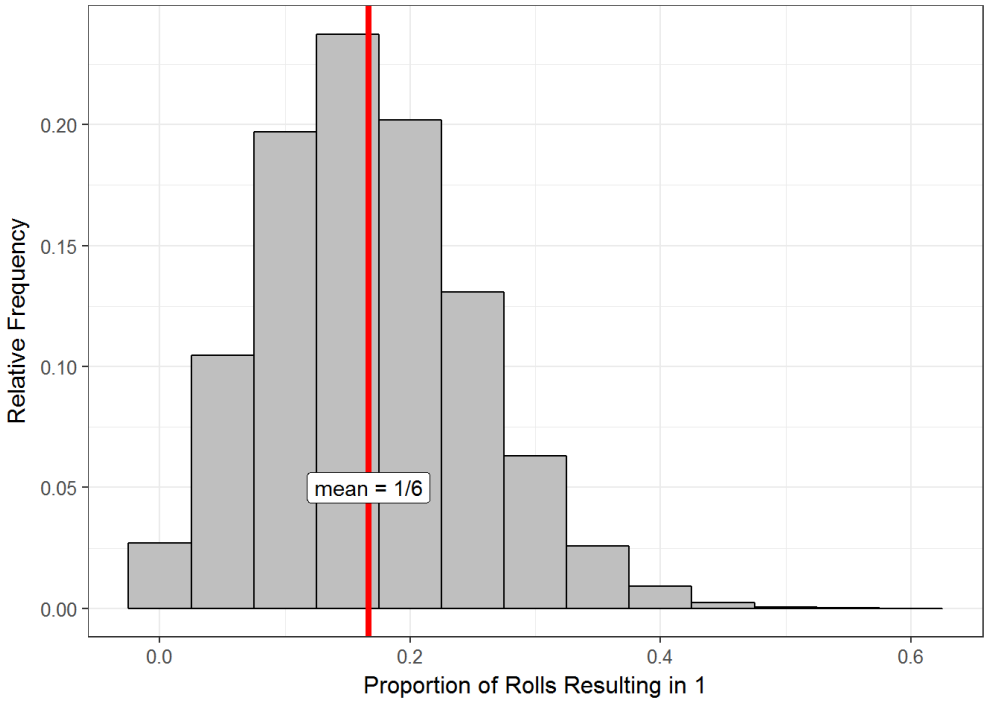


Figure 6.5: Sampling distribution for the proportion of 20 dice rolls which result in a 1. The distribution is based on repeating the sampling process 50000 times.

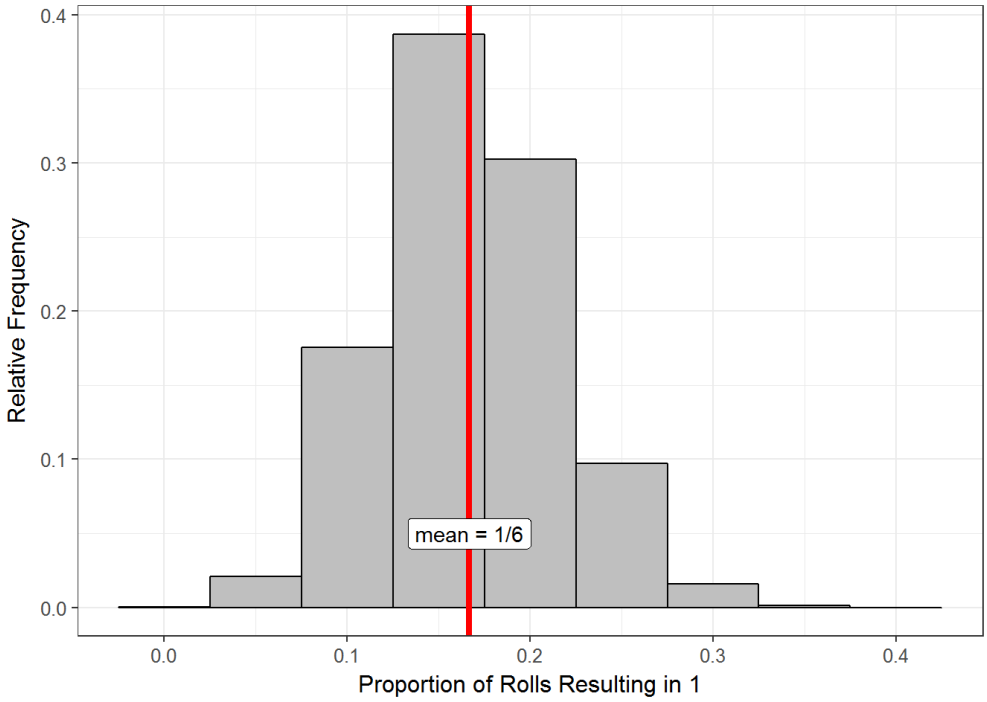


Figure 6.6: Sampling distribution for the proportion of 60 dice rolls which result in a 1. The distribution is based on repeating the sampling process 50000 times.

Key Idea: The size of the sample is not as important as whether it is representative. A small representative sample is better for making inference than a large sample which is biased.

One of the most useful things about the sampling distribution is that it gives us an idea of how much we might expect our statistic to change from one sample to another. Based on Figure 6.5, we could say that if we roll a die 20 times, the proportion of rolls which result in a 1 is most likely to be between 0.05 and 0.30 (so somewhere between 1 and 6 ones out of the 20 rolls). It would be *extremely* rare to have 12 of the 20 rolls result in a 1 (notice how small the bar is on the 0.6 proportion). The sampling distribution is therefore giving us an idea of the variability in our statistic.

Remember, our goal was to account for the variability in the statistic (how much it changes from one sample to another) *while accounting for the sample size*. How is this done? When forming the sampling distribution, we repeated the study. For each replication, we obtained a new sample that *had the same size as the original*. So, the sample size is baked into the sampling distribution. To see the impact of taking a larger sample, consider rolling a six-sided die 60 times instead of 20 times. When we build the sampling distribution, each replication will then involve repeating the process with 40 new rolls. Figure 6.6 shows the sampling distribution of the proportion of 60 rolls which result in a 1 using 50000 replications. Notice that the distribution is still centered on the true parameter $\theta = 1/6$. The primary difference between this figure and the last is that when we increased the sample size, the sampling distribution narrowed.

We all have this intuition that "more data is better." In truth, we should say "more *good* data is better." By "better," we mean that the statistic is less variable. Notice that we have to be careful here. We are not saying that the *sample* has less variability; we are saying the *statistic* has less variability. That is, we are more confident in our estimate because we do not expect it to change as much from one sample to the next. From Figure 6.6, we have that if we roll the die 60 times, we expect the proportion of 1's to be somewhere between 0.1 and 0.25 (somewhere between 6 and 15 ones out of the 60 show up). The proportion is varying much less

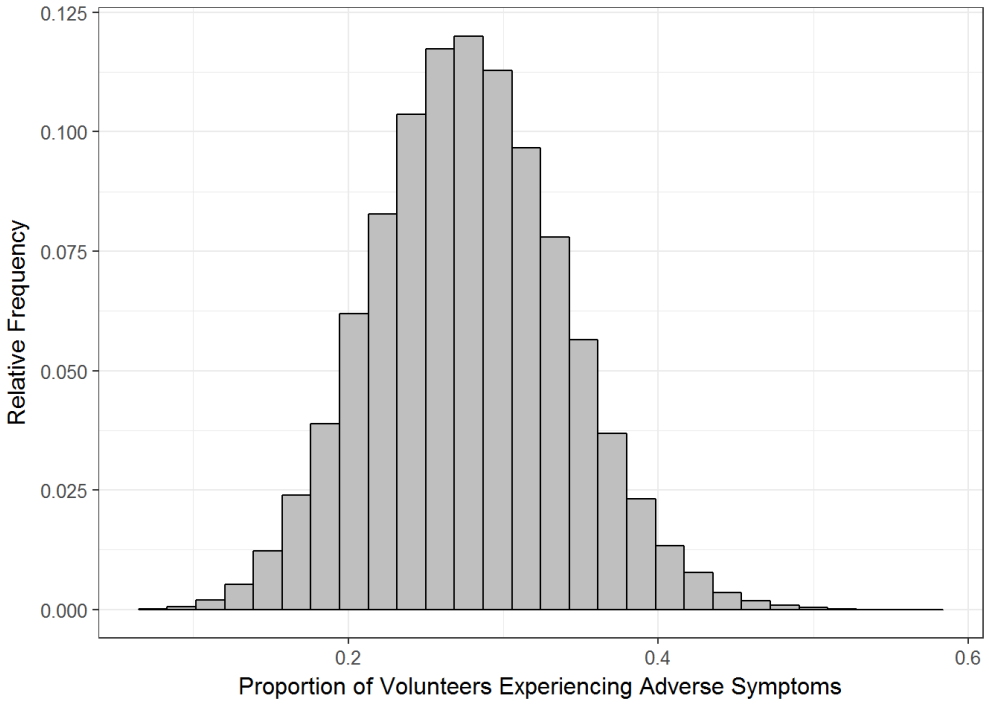


Figure 6.7: Sampling distribution for the proportion of volunteers assigned to wildlife who will develop adverse symptoms based on a sample of 54 volunteers.

from one sample to the next.

Key Idea: Larger samples result in *statistics* which are less variable. This shows itself in the sense that the sampling distribuiton is narrower.

Tip: Students often believe that a large sample reduces the variability in the data. That is not true; a large sample reduces the variability in the *statistic*.

6.3 Modeling the Sampling Distribution

Let's return to the Deepwater Horizon Case Study. In particular, suppose we are tyring to address the following question:

What proportion of volunteers assigned to clean wildlife will develop adverse respiratory symptoms?

We have an estimate for this proportion ($p = 0.278$) based on the observed sample. Based on the discussion in the previous section, we know the sampling distribution of this proportion can help us quantify the variability in the estimate. Figure 6.7 represents the sampling distribution of this proportion. From the graphic, we would not expect the proportion of volunteers who experience adverse respiratory symptoms to move much beyond 0.15 and 0.4 if we were to repeat the study; it would almost certainly not move beyond 0.1 and 0.5 if we were to repeat the study.

Now, you might ask "wait, where did this sampling distribution come from? There is no way you actually repeated the study 50000 times, right?" Right. In the previous section, we described building the sampling

distribution through repeated sampling. But, in practice, this is never practical; if it were, we would have just conducted a bigger sample to begin with. Generally, cost is the limiting factor in choosing a sample size; so, we only have a limited set of data to work with. So, the sampling distribution is critical to making inference, but we cannot take multiple samples to make it. Where does that leave us? The answer... modeling. Our goal is to construct a model of the sampling distribution that we can use to make inference.

There are three general techniques for modeling the sampling distribution of a statistic:

1. Build an empirical model.
2. Build an analytical model using probability theory.
3. Build an analytical model appealing to approximations.

We will focus on the first approach; the latter two approaches are discussed in the last unit of the text. The idea in constructing an empirical model is to mimic the discussion above regarding the construction of a sampling distribution. Our description references Figure 6.8 often. We are limited by our resources; because of time and money constraints, we cannot resample from the population (crossed off resamples). So, we pretend for a moment that our original sample (colored in green in the figure) is the population for a moment. Our idea is to randomly sample from this original sample, creating a *bootstrap resample* (colored in orange in the figure). Forgive the non-technical terms here, but since the orange "blob" is a random sample from the green "blob," then it is representative of the green blob. Therefore, if we construct an estimate $\widehat{\theta}^*$ from the orange blob (the star denotes a statistic from a resample), then it should be close to the statistic $\widehat{\theta}$ from the green blob; but, since this green blob is representative of the population, $\widehat{\theta}$ should be close to the true parameter θ. Therefore, we have that

$$\widehat{\theta}^* \approx \widehat{\theta} \approx \theta \Rightarrow \widehat{\theta}^* \approx \theta$$

That is, the bootstrap resamples produce statistics which are good estimates of the parameter from the underlying population. The benefit here is that the bootstrap resamples are constructed in the computer. And, given today's computing power, we are not limited by time or money (10000 bootstrap resamples can often be taken in a matter of seconds). If you want to see this process in action, we encourage you to check out the free online app located at http://www.lock5stat.com/StatKey/bootstrap_1_cat/bootstrap_1_cat.html.

Again, the idea is to mimic in the computer the resampling that we were unable to do in real life. This process is known as the **bootstrap** procedure.

Definition 6.3 (Bootstrap)**.** A method of modeling the sampling distribution by repeatedly resampling from the original data.

A couple of notes on the actual implementation of a bootstrap procedure:

1. Each resample is the same size as the original sample.
2. Each resample is taken *with replacement*; that means that values from the original sample can show up multiple times. This is like "catch and release" fishing.
3. Typically, between 3000 and 10000 bootstrap resamples are taken.

We will avoid actual computation throughout the text, but several resources are available for implementing the bootstrap procedure (and its many variants) in various computer programming languages and software packages.

> **Tip:** Students often believe that the bootstrap "creates more data." This is not true. Instead, the boostrap resamples from the existing data. This highlights the need to have a representative sample when performing analysis.

As an example, for the Deepwater Horizon Case Study, we performed the following steps to create Figure 6.7:

1. Select 54 volunteers at random (with replacement) from the original sample of 54 volunteers who had been assigned to clean wildlife.
2. For our resample, we computed the proportion of those individuals who had experienced adverse respiratory symptoms.

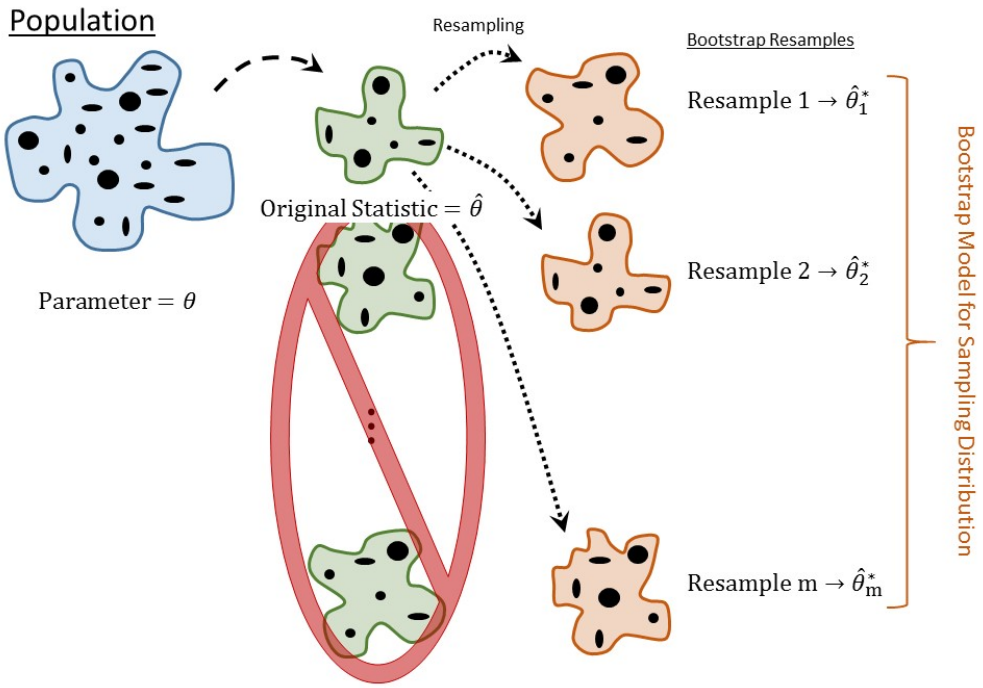


Figure 6.8: Illustration of modeling the sampling distribution via the Bootstrap.

3. We repeated steps 1 and 2 several thousand times, retaining the bootstrap statistics from each bootstrap resample.
4. We plotted the distribution of the bootstrap statistics.

6.4 Using a Model for the Sampling Distributions (Confidence Intervals)

From Figure 6.7, we observed that we would not expect the proportion of volunteers who had experienced adverse symptoms to move much beyond 0.15 to 0.4 if we were to repeat the study. How does this help us in performing inference? Remember that each value in the bootstrap model for the sampling distribution is an estimate of the underlying parameter. So, we can think of the above model as showing us what good estimates of the parameter look like. Another way of saying it: the model for the sampling distribution shows us the *reasonable* (or *plausbile*) values of the parameter. Here, by "reasonable," we mean values of the parameter for which the data is *consistent*. Consider the following statements (which are equivalent):

- Based on our sample of 54 volunteers, it is reasonable that the proportion of volunteers assigned to clean wildlife who would experience adverse respiratory symptoms is between 0.15 and 0.4.
- Our sample of 54 volunteers is consistent with between 15% and 40% of all volunteers assigned to clean wildlife experiencing adverse respiratory symptoms.

We have just conducted inference for "estimation" type questions. We are able to provide an estimate for the parameter which acknowledges that the data is not perfect and there is variability in sampling procedures. That variability incorporated itself into constructing an estimate that is an interval instead of a single point.

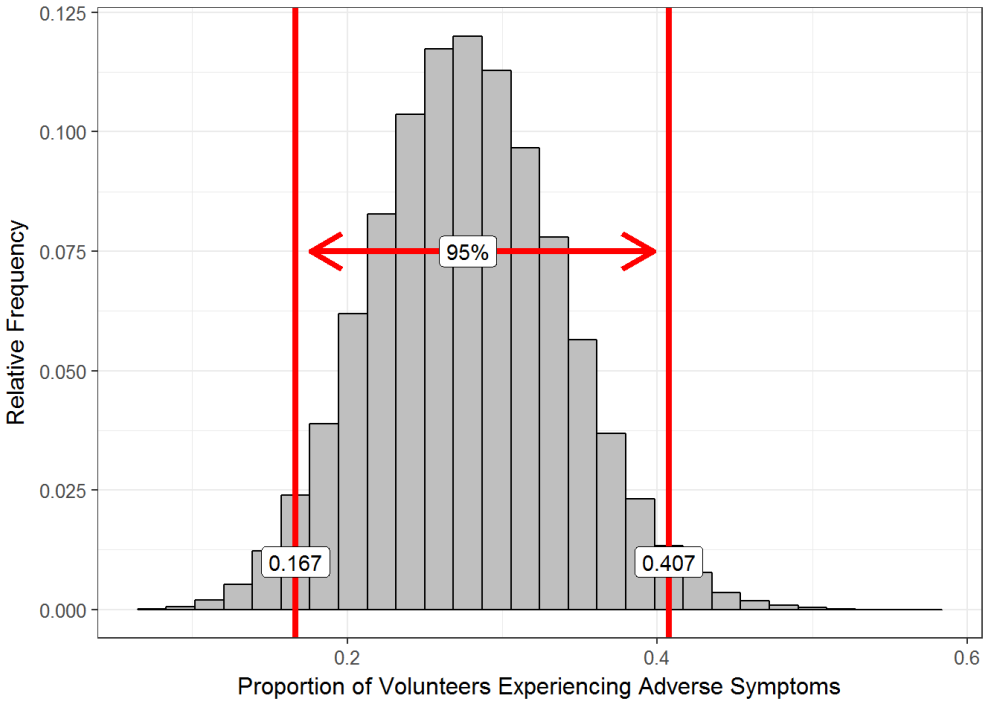
The above interval was chosen arbitrarily by just looking at the sampling distribution and capturing the peak of the distribution. If we want to be more formal, we might try to capture the middle 95% of values. This is

known as a **confidence interval**.

Definition 6.4 (Confidence Interval)**.** An interval (range of values) estimate of a parameter that incorporates the variability in the statistic. A k% confidence interval will contain the parameter of interest in k% of repeated studies.

If we were to capture the middle 95% of statistics, a 95% confidence interval, we would obtain an interval of (0.167, 0.407), as shown in Figure 6.4.

\begin{figure}



{

}

\caption{Construction of a 95% confidence interval via bootstrapping for the proportion of volunteers assigned to wildlife who will develop adverse symptoms based on a sample of 54 volunteers.} \end{figure}

Confidence intervals are often misinterpreted; this comes from their dependence on repeated sampling. When thinking about confidence intervals, think about playing a game of ring toss: you toss a ring in hopes of landing on top of a target. The target is the parameter characterizing the population. The confidence interval is like a ring. Since the confidence interval is constructed from a model of the sampling distribution, it changes with each sample; that is, the confidence interval itself is a statistic. Just like in ring toss, the ring moves with each toss, the confidence interval moves with each sample. However, the target stays fixed. Because of this, the following interpretations are *incorrect*:

- There is a 95% chance that the proportion of volunteers assigned to clean wildlife who will experience adverse symptoms is between 0.167 and 0.407.
- 95% of volunteers assigned to clean wildlife in our sample had a value between 0.167 and 0.407.

The first statement is incorrect because it treats the parameter as the thing that is moving. Once the data has been collected, the confidence interval is a fixed quantity; neither the estimate or the parameter is moving; so, there is no probability left. Again, think about tossing a ring; once the ring is tossed, you either captured the target or you did not. There is no "I captured the target with 95% probability."

The second statement is absurd in this case. A volunteer either had respiratory symptoms or they did not; so, saying they had a value between 0.167 and 0.407 is ridiculous. However, this is a common misconception with confidence intervals. They are describing reasonable values of the parameter, not values of the variable

in the sample or population. We recommend sticking to interpreting a confidence interval as specifying reasonable values for the parameter.

Tip: Confidence intervals *do not* provide a probability that the parameter is inside. Nor do they tell you anything about the individual values in a sample or population. They describe reasonable values of the parameter.

Key Idea: Confidence intervals specify *reaonable* values of the parameter based on the data observed.

It may seem like a good idea to make a 100% confidence interval to be sure we always capture the parameter. But, such intervals are not helpful in practice. For example, a 100% confidence interval for the proportion of volunteers experiencing adverse symptoms would be (0, 1). But, this is useless; it essentially says that the proportion has to be a number between 0 and 1, but we already knew that. Therefore, we must balance the confidence we desire with the amount of information the interval conveys.

Tip: If you want both a high level of confidence but also a narrow interval, increase the sample size. As the sample size increases, the variability in the statistic decreases leading to a narrower interval.

Tip: 95% confidence intervals are the most common in practice; however, 90%, 98%, and 99% intervals are also used. It is extremely rare to use less than a 90% CI.

6.5 Bringing it All Together

Consider the following question:

> Is there evidence that more than 1 in 5 volunteers assigned to clean wildlife will develop adverse respiratory symptoms?

Let's answer this question using a confidence interval. Based on the data obtained, we found that the 95% confidence interval (CI) for the proportion of volunteers experiencing adverse symptoms to be (0.167, 0.407). Is this data consistent with more than 1 in 5 volunteers developing adverse symptoms? Yes, since there are proportions within this interval which are larger than 0.2. But, *consistency* is not the same as *evidence*; remember, evidence is the idea of "beyond a reasonable doubt." After all, is this data *consistent* with less than 1 in 5 volunteers developing adverse symptoms? Yes, since there are proportions within this interval which are less than 0.2.

Confidence intervals specify reasonable values — those values of the parameter which are consistent with the data. This data is then consistent with proportions that are both less than 0.2 and greater than 0.2. So, what can we say then? We can say that there is *not* evidence that more than 1 in 5 volunteers assigned to clean wildlife will develop adverse respiratory symptoms, but the data is consistent with this claim.

More, we can say that there *is evidence* that the proportion of volunteers who will develop symptoms is less than 0.5; further, the proportion of volunteers who will develop symptoms is larger than 0.1. That is, the data provides evidence that more than 10% of volunteers will develop adverse symptoms, but this percentage will not be larger than 50%. How do we know? Because values less than 10% are not reasonble values of the parameter based on the 95% CI. Values like 0.1 are outside of the confidence interval and are therefore not reasonable. Similarly, values above 0.5 are outside the confidence interval and are therefore not reasonable.

The power of a model for the sampling distribution is that it allows us to determine which values of a parameter are reasonable and which values are not.

Chapter 7

Quantifying the Evidence (Rejecting Bad Models)

Again, the goal of statistical inference is to use the sample as a snapshot of the underlying population (Figure 7.1). Recall that there are essentially two categories of questions we ask when trying to perform inference:

- Estimation: what *proportion* of volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms?
- Model Consistency: is it reasonable that no more than 1 in 5 volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms?

In the previous chapter we addressed these questions through the use of confidence intervals — by specifying reasonable values of the parameters through a model of the sampling distribution. However, when working with questions of the second type (model consistency), there is a second approach; this latter approach is useful when confidence intervals cannot be constructed for the particular question of interest (see Unit 2).

Remember, assessing model consistency is similar to performing a trial in a court of law. After gathering the evidence, the jury is left with the following decision:

- Assuming the defendant is innocent, if the evidence is unlikely to have occurred (so is not consistent with innocence), then they vote "guilty."
- Assuming the defendant is innocent, if the evidence is reasonably likely to have occurred (so is consistent with innocence), then they vote "not guilty."

The goal in this section is to somehow quantify the evidence against a particular model to determine if we can say that the data is not consistent with the given model.

7.1 Some Subtleties

In a U.S. trial, there are some subtleties that we should be aware of, as they also creep up in statistical analyses and have implications for how we interpret statistical results. First, the jury weighs the evidence *under the assumption of innocence*. That is, they first develop a working hypothesis (the defendant is innocent). Then, the likelihood of the evidence *under this assumption* is determined. For example, if a defendant were innocent of murder, it is unlikely to have five eye witnesses stating the defendant was seen standing over the victim, holding the murder weapon, and screaming "I killed him!" Since that evidence does not jive with innocence, the jury convicts. If, however, the only evidence is that five eye witnesses place the defendant in the same city as the victim and the defendant matches the description of someone seen fleeing the crime scene, then the jury would not convict. Why? Because the evidence, while pointing toward guilt, is

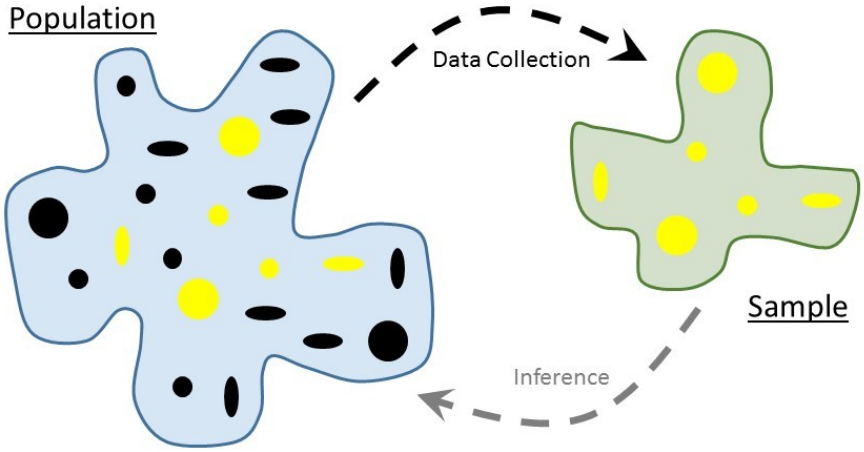


Figure 7.1: Illustration of the statistical process (reprinted from Chapter 1).

not overwhelming; these things could have happened by chance alone. Therefore, the evidence, while consistent with guilt does not provide evidence for guilt.

Also notice that a jury saying "not guilty" is not the same as saying "innocent." That is, a lack of evidence to convict does not imply the defendant is innocent. A lack of evidence is simply a lack of evidence. The defendant may still be guilty, but the evidence has just not proven it.

Similarly, when assessing model consistency, we will weigh the data *under the null hypothesis* (our working assumption). Then, the likelihood of our data occurring by chance alone *under this hypothesis* is determined. If that likelihood is small (data is not consistent with the null hypothesis), we can conclude the data supports the alternative hypothesis (guilty). If, however, that likelihood is large (data is consistent with the null hypothesis), we can only conclude that the data is consistent with the hypotheses. We are *not* able to say "supports the null" because that would be like saying a defendant is innocent. We can't prove innocence because we started by assuming it!

7.2 Assuming the Null Hypothesis

Consider the question we have been asking regarding the Deepwater Horizon Case Study:

> Is there evidence that more than 1 in 5 volunteers assigned to clean wildlife will develop adverse respiratory conditions?

Remember, we framed this question through statements about a parameter in Chapter 3:

> H_0 : the proportion of volunteers assigned to clean wildlife who develop adverse respiratory symptoms is no more than 0.20.
> H_1 : the proportion of volunteers assigned to clean wildlife who develop adverse respiratory symptoms exceeds 0.20.

Within the sample we observed that 27.8% of volunteers experienced adverse symptoms, which is certainly more than the 0.20 in the claim; therefore, the data is at least trending toward the alternative hypothesis. However, it is possible that we just have a strange sample. Remember in our discussion of sampling distributions in the previous chapter, however, that we expect the estimate to vary from one sample to another. Essentially, we need to know whether 27.8% of volunteers experiencing symptoms is a strong signal

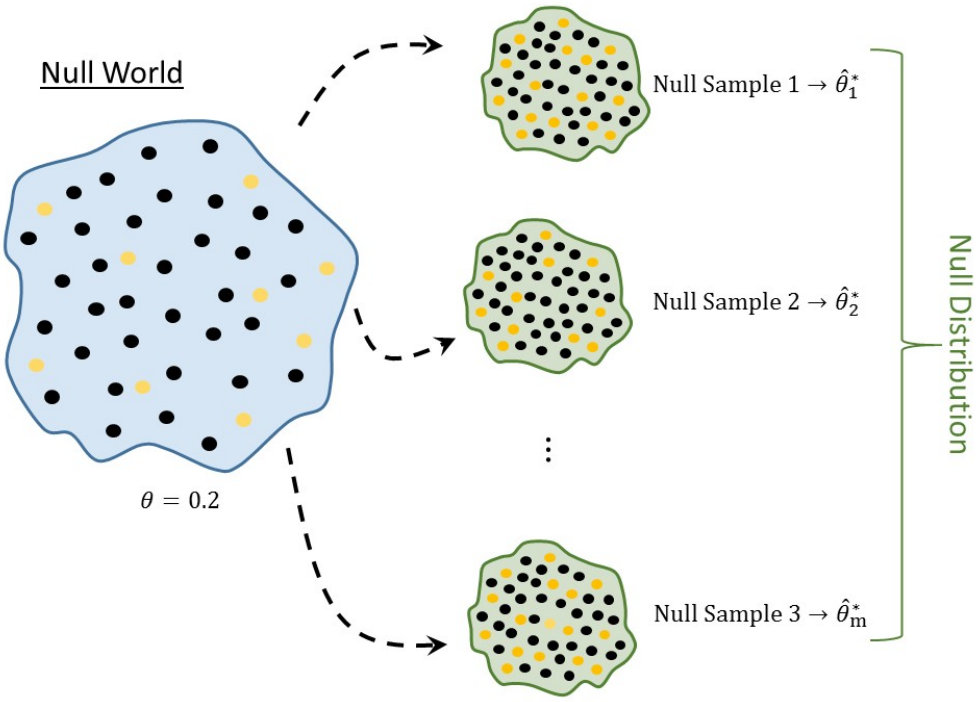


Figure 7.2: Illustration of constructing a null distribution. Notice the similarity to constructing the sampling distributon.

that the rate within the popoulation is larger than 0.2 (1 in 5) or 27.8% is simply a fluke that might happen due to sampling variability. While we are going to be attacking the question differently in this chapter than the previous, we see that the key is still variability in the estimate. That is, we are back to the *Fourth Fundamental Idea of Inference*. As stated above, in order to determine evidence for one statement (captured by the alternative hypothesis), we begin by assuming the opposite statement (captured by the null hypothesis) as our working assumption. That is, if we want to know if 27.8% of volunteers experiencing adverse symptoms is "evidence," we need to figure out what we *expect* to happen *if only 1 in 5 volunteers actually develop adverse respiratory symptoms.*

Consider this last statement. It is equivalent to saying "what type of evidence would we expect for an innocent person?" Only if we know what to expect can we determine if the evidence in front of us is extreme enough to convict. Only if we know what to expect can we determine if the observed sample provides evidence in favor of the alternative. So, we enter a fake world. . . a world in which exactly 1 in 5 volunteers actually develop respiratory symptoms. That is, we enter a world in which the null hypothesis is true. Now, in this world, how do we know what to expect? We construct the sampling distribution for the proportion under this assumption that the null hypothesis is true; this is known as the **null distribution**.

Definition 7.1 (Null Distribution)**.** The sampling distribution of a statistic *if* the null hypothesis is true.

To construct the null distribution, we do the following steps (illustrated in Figure 7.2:

1. Sample randomly from a fake population where the null hypothesis is true.
2. For each sample, compute the statistic of interest.
3. Repeat steps 1 and 2 several thousand times.
4. Plot the statistics retained from each sample.

Notice that these are the same steps as in constructing a sampling distribution with the exception that instead of sampling from the population of interest, we sample from a hypothetical population in which the null distribution is true. With today's computational power, we are able to make such samples possible

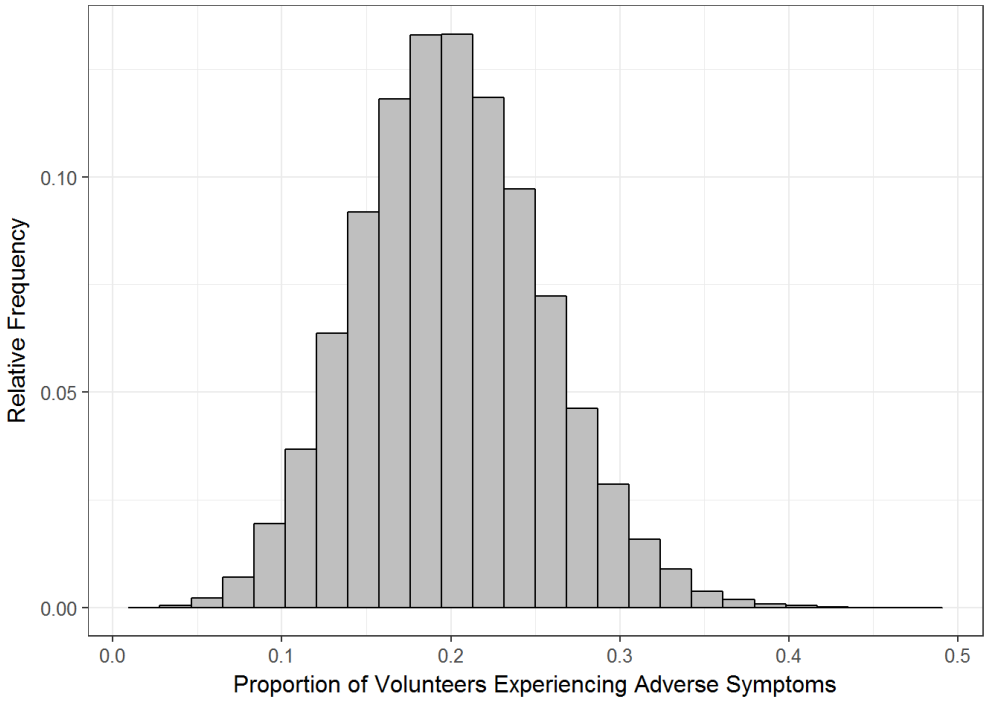


Figure 7.3: Null distribution for the proportion of volunteers assigned to clean wildlife experiencing adverse respiratory symptoms. Null hypothesis is that the proportion is 0.20; this is based on a sample of size 54.

similar to bootstrapping since we can make the null population in a virtual world and sample from it. That is, we are simulating what would happen if the null hypothesis were true. Figure 7.3 represents the null distribution for the proportion of volunteers in a sample of 54 assigned to clean wildlife which would develop adverse sympoms when the null hypothesis is that the proportion is 0.20.

7.3 Using the Null Distribution

From the figure, we see that *if the null hypothesis were true* — if only 1 in 5 volunteers assigned to clean wildlife experienced symptoms — then in a sample of 54 individuals, we would expect the proportion who experienced symptoms to be somewhere between 0.1 and 0.3. *If the null hypothesis were true*, it would be nearly impossible that half of the individuals experienced symptoms (since 0.5 is way off in the tail of the distribution). The further in the tail region, the more extreme the sample. The question is then how extreme is our sample? Again, the null distribution is just setting up expectations; now, we have to weigh the evidence against those expectations.

In our sample, we observed 27.8% of volunteers who experienced symptoms. Since 0.278 is towards the center of the distribution, we would say that it is not an extreme sample. In order to quantify how extreme (or not extreme) its, we find out what fraction of values are more extreme (larger than in this case) than the value observed; that is, what fraction of values appear in the right tail of the distribution. Figure 7.4 illustrates this computation. Based on the null distribution, there is a 10.6% chance that *if the null hypothesis were true* — only 1 in 5 volunteers actually experienced symptoms — that in a random sample of 54 volunteers we would obtain data this extreme or moreso by chance alone. Essentially, this tail area is quantifying the strength of the evidence. The smaller this area, the further in the tail region our data is; that is, our data is more unexpected. Therefore, small areas indicate that the data (our evidence) does not jive with our expectations under the null (innocence), forcing us to conclude the data provides evidence *against* the null hypothesis. In our case, since the area is relatively large, our data is completely consistent with what we

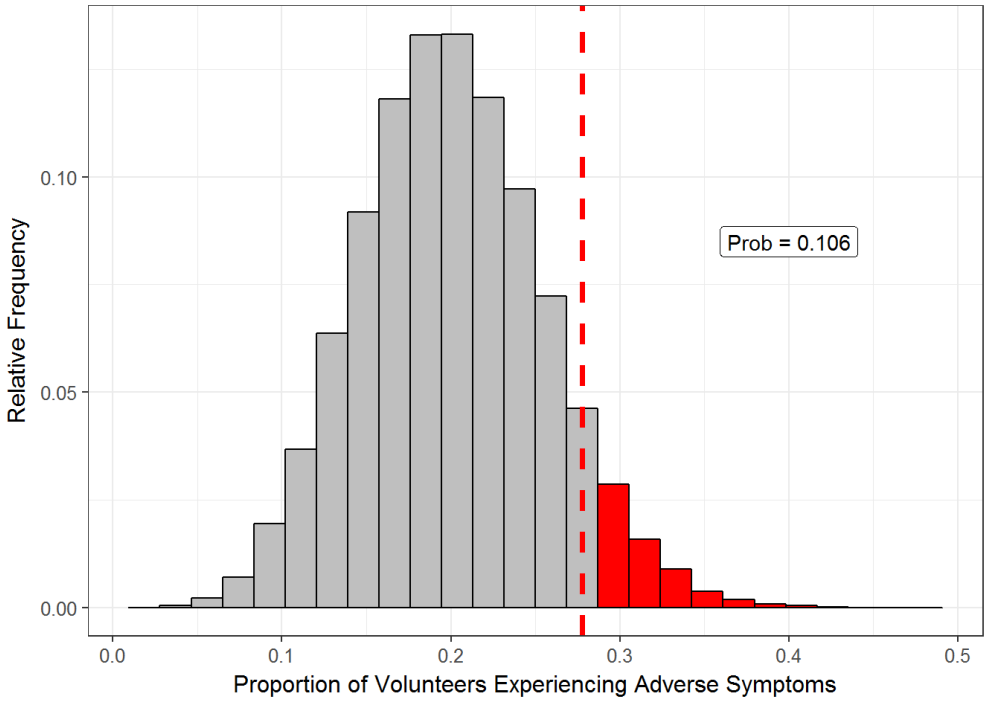


Figure 7.4: Likelihood of obtaining a sample as extreme or moreso as that of the original sample when the parameter of interest is the proportion of volunteers assigned to clean wildlife experiencing adverse respiratory symptoms. Null hypothesis is that the proportion is 0.20; this is based on a sample of size 54.

might expect if the null were true. Therefore, in this case, we conclude that there is no evidence that the rate of those experiencing symptoms exceeds 1 in 5. This area is known as the **p-value**.

Definition 7.2 (P-Value)**.** The probability, assuming the null hypothesis is true, that we would observe a statistic, by chance alone, as extreme or moreso as that observed in our sample. This quantifies the strength of evidence against the null hypothesis. Smaller values indicate stronger evidence.

It is natural to ask "how small does the p-value need to be to prove a statement?" Like a trial, the weight of the evidence presented depends on the context. In some studies, a p-value less than 0.01 may be strong evidence while in other studies a p-value less than 10^{-6} is required. And, as in a trial, it is not only the strength of the evidence but the type of evidence presented (DNA evidence may be stronger than fingerprint evidence). In statistics, it is important to consider the effect size as well as the p-value. That is, consider whether the difference between the estimate and the null value is actually large; this is always based on subject-matter expertise. It is often helpful to report a confidence interval alongside a p-value.

Tip: While what constitutes "significant" may vary from discipline to discipline, the list below is a good rule of thumb:

- $p \geq 0.1$: no evidence against the null hypothesis.
- $0.05 \leq p < 0.1$: weak evidence against the null hypothesis.
- $0.01 \leq p < 0.05$: some evidence against the null hypothesis.
- $0.001 \leq p < 0.01$: evidence against the null hypothesis.
- $p < 0.001$: strong evidence against the null hypothesis.

As with any rule of thumb, this should not be considered binding and may vary depending on the application.

Like confidence intervals, p-values are often misinterpreted. In fact, they have become so abused that some

researchers argue against their use. It is our opinion that the p-value can be a useful tool once it is appropriately understood; so, let's dispell some of these misconceptions. Consider these *incorrect* statements regarding the p-value obtained for the Deewater Horizon Case Study computed above:

- There is a 10.6% chance that only 1 in 5 volunteers assigned to clean wildlife will experience adverse symptoms.
- Since the p-value is large, there is evidence (or the data supports the claim) that 1 in 5 volunteers assigned to clean wildlife will experience adverse symptoms.

The first statement incorrectly assumes that there is some chance that the null hypothesis is true. Remember, our two hypotheses are statements about the parameter. One is true and other is not. Our ignorance does not change this; therefore, it does not make sense to talk about the probability of the null being true or false. Instead, our job is to quantify the likelihood of the data *assuming the null is true*. The p-value is about the likelihood of the data under a particular model (the null hypothesis).

The second statement makes the common mistake that a lack of evidence for the alternative is evidence in favor of the null. A lack of evidence is like a "not guilty" verdict. It simply means we were not convinced. However, it does not mean that the defendant is innocent. All we are saying with the large p-value in this case is that the data is *consistent* with only 1 in 5 volunteers getting adverse symptoms; unfortunately, it is also *consistent* with more than 1 in 5 volunteers getting adverse symptoms. This may be an unsatisfying conclusion, but it is still a conclusion nonetheless. Our conclusion was based on assessing the variability of a the statistic under a particular model. This is captured in our last of the *Five Fundamental Ideas of Inference*:

Fundamental Idea: **Fundamental Idea V**: With a model for the distribution of a statistic, we can quantify the error in our estimate and the likelihood of a sample under a proposed model. This allows us to draw conclusions about the corresponding parameter, and therefore the population, of interest.

7.4 Sampling Distributions vs. Null Distributions

Clearly the sampling distribution and null distribution of a statistic are closely related. The difference is that the null distribution is created under a proposed model while the sampling distribution lets the data speak for itself. It is worth taking just a moment to highlight the differences in the use of these two components of the *Distributional Quartet*.

The sampling distribution is centered on the true value of the parameter; the null distribution is centered on the null value. Once we assume the null hypothesis is true, we have a value for the parameter; as a result, we expect the sampling distribution under this assumption (that is, the null distribution) to be centered on this hypothesized value. So, null distributions are *always* centered on the null value.

Sampling distributions lead to confidence intervals by specifying reasonable values of the parameter.

Null distributions lead to p-values by quantifying the likelihood of our data under a proposed model.

Tip: Model the sampling distribution to construct a confidence interval; to assess a hypothesis the null value is overlayed on the sampling distribution. Extreme values of the distribution are unreasonable values for the parameter.
Model the null distribution to compute a p-value; to assess a hypothesis, the statistic from the sample is overlayed on the null distribution. Extreme values of the distribution are values which provide evidence against the null hypothesis.

Chapter 8

Using the Tools Together

In this unit, we have introduced the key components in both the language and logic of statistical inference. In fact, with a firm grasp of the concepts in this unit, you should be able to read and interpret key statistical findings. All statistical analyses make use of the *Five Fundamental Ideas of Inference* and alternate between the members of the *Distributional Quartet*. The context of each problem differs, but the logic remains the same. In this chapter, we present another analysis based on the Deepwater Horizon Case Study, annotating it along the way to see how these elements work together fluidly to reach a conclusion. Specifically, we are interested in the following question:

> Are volunteers assigned to clean wildlife at higher risk of developing adverse respiratory symptoms compared to those volunteers who do not come into direct contact with oil? If so, estimate the increased risk.

8.1 Framing the Question (Fundamental Idea I)

We are really interested in whether the rate of respiratory symptoms in one group of volunteers is larger than that in a second group. Therefore, our working assumption is that there is no difference in the rate of respiratory symptoms between these two groups. That is, we have

> H_0 : the rate of adverse respiratory symptoms is similar between volunteers assigned to clean wildlife and those assigned to tasks which do not involve direct exposure to oil.
> H_1 : the rate of adverse respiratory symptoms is greater for volunteers assigned to clean wildlife and those assigned to tasks which do not involve direct exposure to oil.

We can also state this more formally with mathematical notation as follows:

> Let θ_1 be the rate of developing adverse respiratory symptoms for volunteers assigned to clean wildlife.
> Let θ_2 be the rate of developing adverse respiratory symptoms for volunteers assigned to tasks without direct exposure to oil.
> $H_0 : \theta_1/\theta_2 \leq 1$
> $H_1 : \theta_1/\theta_2 > 1$

The ratio θ_1/θ_2 is known as the *risk ratio* as it captures the increased risk for one group compared to another.

Notice that this is a well-posed question as it centers on parameters which characterize the population. Therefore, it can be answered with appropriate data.

> Distribution of the Population: Our questions of interest are about the population and therefore focus on characterizing this distribution.

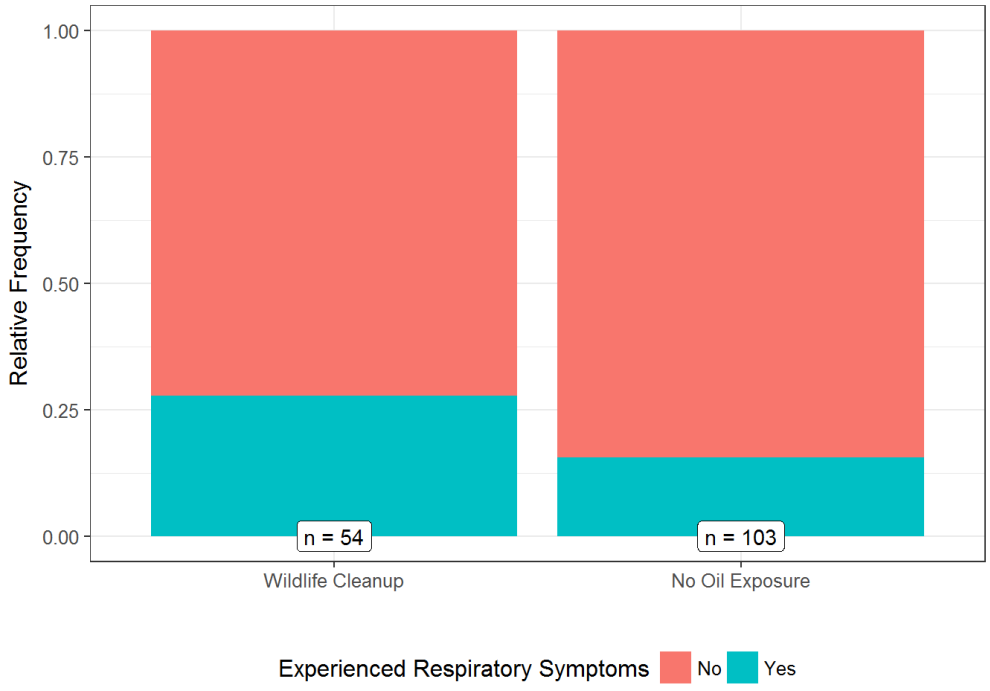


Figure 8.1: The risk of developing adverse respiratory symptoms for volunteers assigned to clean wildlife and those volunteers assigned to tasks which do not have direct exposure to oil.

8.2 Getting Good Data (Fundamental Idea II)

As we are working with previously collected data, we are unable to design a good sampling scheme. The only thing we can do at this point is critique the sample we have. The key question to ask ourselves is whether there is any reason that these group of volunteers differs systematically from other volunteers working oil spills. For example, this oil spill occurred in the Gulf of Mexico; the majority of volunteers were then naturally residents of Gulf states. It is possible that these residents are somehow fundamentally different with respect to their risk of developing adverse respiratory symptoms compared to the remainder of the United States. If that is the case, the results of this study would not generalize to oil spills occuring in the Atlantic. However, it is probably reasonable to say that these results would apply to future oil spills in the Gulf.

Also note that this was not a controlled experiment. Volunteers were not randomly allocated to their assignments that we know of. Therefore, our results could be somewhat limited. The two groups should be compared regarding other attributes (this data is unavailable to us currently) in order to determine if they are similar with respect to other variables which may potentially confound the results.

8.3 Presenting the Data (Fundamental Idea III)

The heart of this question is comparing the rate of adverse events in each group. Figure 8.1 makes this comparison.

As seen in the figure, the rate of adverse respiratory symptoms is larger in the group of volunteers assigned to wildlife cleanup. The rate of respiratory symptoms is 1.79 times higher in the volunteers assigned to clean wildlife compared to those assigned to tasks with no direct oil exposure.

Notice that we reported the relative risk comparing the two groups as it is directly tied to how we specified

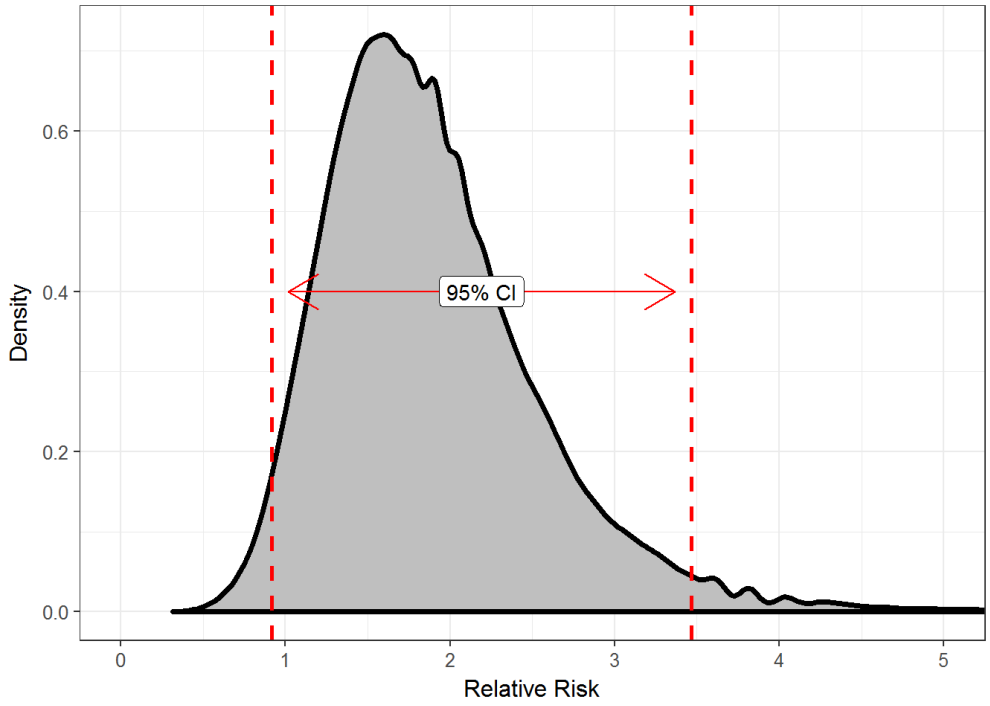


Figure 8.2: Model of the sampling distribution for the relative risk comparing volunteers assigned to clean wildlife to volunteers assigned to tasks not involving oil exposure. The model was developed via bootstrapping using 50000 replications.

the hypotheses above.

Distribution of the Sample: graphics and numerical summaries characterize this distribution, informing us about the underlying population. This is possible as long as the sample is representative of the population.

8.4 Quantifying the Variability in the Estimate (Fundamental Idea IV)

While we have an estimate for the increased risk of adverse respiratory symptoms for those volunteers assigned to clean wildlife, the estimate has not taken into account the variability in the sample. In order to quantify this variability, we use a bootstrap procedure to model the sampling distribution of the risk ratio. Observe that we focus on the sampling distribution of the statistic that estimates the parameter of interest.

Recall that the bootstrap mimics the process for generating a sampling distribution. In this case, "repeating the study" involves collecting data from not one, but two groups. So, we must resample both from the 54 volunteers who were assigned to clean wildlife and the 103 volunteers assigned to tasks not involving direct oil exposure. Each time we resample, we ensure that we select 54 volunteers who clean wildlife and 103 who do not. We need the process of the original study to be maintained. Each time we resample from these groups, we compue the relative risk and retain this value. Figure 8.2 shows the sampling distribution for the relative risk comparing these two groups. Again, it is important to note that we are not generating *new* data; we are *resampling/reusing* the original sample.

Volunteers assigned to clean wildlife have are 1.79 times (95% CI = (0.92, 3.47)) more likely to experience adverse respiratory symptoms compared to those volunteers assigned to tasks not requiring direct exposure

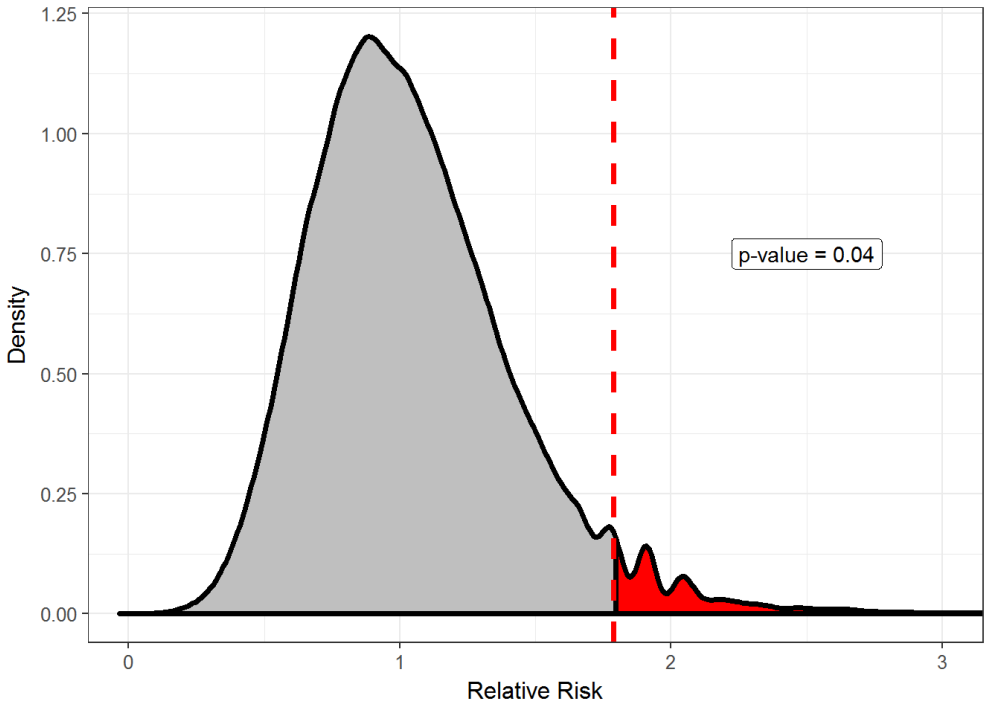


Figure 8.3: Null distribution for the relative risk comparing volunteers assigned to clean wildlife to volunteers assigned to tasks not involving oil exposure. The null hypothesis assumed the two groups of volunteers had a similar risk. The nul distribution was developed via bootstrapping using 50000 replications.

to oil. Our data suggests that our data is consistent with the two groups having a similar risk but tends toward volunteers assigned to clean wildlife being at increased risk..

Sampling Distribution: allows us to quantify the variability in the statistic and provide an interval estimate for the paraemter which incorporates this variability.

8.5 Quantifying the Evidence (Fundamental Idea V)

In order to quantify the departure of the data from our working assumption that the risk is similar between the two groups, we rely on the null distribution and compute a p-value.

There is some evidence ($p = 0.04$) to suggest that volunteers exposed to oil have an increased risk of developing adverse respiratory symptoms. Given the estimated level of this increased risk, we believe this is something health officials should investigate further. It would be worth investigating further what aspects of the oil exposure may have caused the increased risk to determine if it can be avoided in the future.

Note we are careful to not claim that the assignments have caused an increase in the risk as this data is not from a controlled experiment. This is one of the limitations of this analysis. However, if we are able to assume the two groups are fairly similar with respect to other attributes — that is, there is no reason why people prone to respiratory symptoms would become assigned to wildlife cleaning — then we may have some reason to believe the results are causal. We will wrestle more with these types of conclusions in the next unit.

Null Distribution: allows us to quantify the level of evidence against a particular claim; how strongly do the data disagree with the working assumption.

8.6 Summary

Notice that our analysis moved through the *Five Fundamental Ideas*, and in doing so made use or referenced each of the four components of the *Distributional Quartet*. As we move through the remainder of the text, we will explore how these frameworks are used in various other analysis scanarios. As we do, we reveal additional concepts that underly statistical modeling.

We admit that there are several other questions that may be raised by the above analysis. This unit is meant to introduce the big concepts of inference. We will concern ourselves more with the details as we progress through the text.

Part II

Unit II: Comparing the Average Response Across Groups

Chapter 9

Case Study: Organic Foods and Superior Morals

"You are what you eat" is a common phrase dating back to at least the 1820's used to suggest that if you want to be fit, you must eat healthy foods. However, does the phrase extend to our personality as well as our physique? Recent research has suggested that specific tastes (sweet vs. disgusting, for example) can influence moral processing. That is, certain foods may lead us to be nicer to those around us or lead us to be more judgemental. Organic foods are often marketed using phrases like "pure" or "honest" (Jessica Alba's Honest Company, for example); is there some relationship between the consumption of organic foods and moral behavior?

Dr. Eskine of the Department of Psychological Sciences at Loyola University sought to answer this question (Eskine 2013). He conducted a study to investigate whether exposure to certain types of food had an effect on a person's moral processing. Specifically, he randomized 62 Loyola University undergradates to one of three food types: organic, comfort, and control. Each participant received a packet containing pictures of four food items from the assigned category:

- Organic Foods: apple, spinach, tomato, carrot
- Comfort Foods: ice cream, cookie, chocolate, brownie
- Control Foods: oatmeal, rice, mustard, beans

The control foods are those which are pre-packaged and are generally considered staple items; organic foods are those which are associated with a healthy diet; and, comfort foods were sweets. After viewing the images for a set period of time, each participant received a packet containing six counter-balanced moral transgressions. An example of such a transgression is produced below:

> Bob was at a family gathering when he met Ellen, a second cousin of his that he had seen once or twice before. Bob found Ellen very attractive and he asked her out on a date. Ellen accepted and they began to have a romantic and sexual relationship. They often go on weekend trips to romantic hotels in the mountains.

Participants were then asked to rate the morality of the scenario on a 7-point scale (1 = "not at all morally wrong" to 7 = "very morally wrong"). The average of the morality scores across the six scenarios was used as an overall measure of their moral expectations. A higher value indicates high moral expectations (very strict) and a lower value indicates lower moral expectations (very lenient).

Dr. Eskine's analysis revealed that there was strong evidence ($p = 0.001$) that participants' moral judgments differed, on average, across the various food exposure groups. In particular, those exposed to organic foods had higher moral expectations (an average mean moral judgment of 5.58) compared to those experiencing comfort foods (average mean moral judgment of 4.89) or control foods (average mean moral judgment of 5.08). He therefore concluded that exposure to organic food did lead to higher moral expectations.

Table 9.1: Subset of data from study characterizing moral behavior following exposure to various food categories.

Participant	Food Condition	Response (Avg of Moral Questions)
18	organic	5.500
20	organic	5.500
21	organic	6.333
1	comfort	6.000
2	comfort	3.500
3	comfort	6.167
4	control	5.167
10	control	7.000
12	control	6.833

Understandably, Dr. Eskine's work caught the interest of various media outlets and researchers. Two researchers within the Department of Psychology at Domincan University in Illinois sought to replicate Dr. Eskine's work (Moery and Calin-Jageman 2016). There were several components to their research, but the first phase included a replication of Dr. Eskine's initial study with minor variants. They enrolled 124 college students into their study. The participants were presented with the same food images as in Eskine's study with the exception that celery was used instead of an apple for organic food. The same moral dilemmas were given to participants. As in the original study, the average score from the six moral dilemmas was the primary response for this study. A subset of the collected data, showing three participants from each treatment group (type of food shown), is presented below. The full dataset[1] has been made available by the researchers at the following website: https://osf.io/atkn7/wiki/home/

[1]There were multiple phases to their research. The direct replication of Dr. Eskine's work was Study 1, which is the dataset being considered in this text.

Chapter 10

Framing the Question

"Does exposure to various food types lead to different moral expectations?" The primary question from the Organic Food Case Study is primarily about the relationship between two variables: the response (see Definition 3.2; moral expectations) and the **factor** of interest (food type).

Definition 10.1 (Factor)**.** Also referred to as the "treamtent," a categorical variable used to explain/predict a response.

The majority of interesting research questions involve identifying or quantifying the relationship between two variables. Despite the complexity of the analyses sometimes employed to address these questions, the basic principles are the same as those studied in Unit 1. To begin, asking good questions involves defining the population of interest and characterizing the variable(s) at the population level through well-defined parameters.

The question of the Organic Food Case Study, as stated above, is ill-posed. Almost certainly, there are individuals for which exposure to organic foods may result in higher moral expectations compared to exposure to comfort foods. However, there are almost certainly individuals for which the effect is reversed — higher moral expectations are expected following exposure to comfort foods compared with organic foods. That is, we expect there to be *variability* in the effect of food types on the resulting moral expectations. The question needs to be refined.

While the study was conducted using college students, the original question seems quite broad (we discuss this discrepancy in more detail in the next chapter). Notice that the original question is not predicated on *consuming* various foods but simply *exposure* to various foods. The question itself is not limited to only those individuals which purchase a specific type of food but concerns all individuals. More, we really see that there are three groups of interest — those which are exposed to organic foods, those exposed to comfort foods, and those exposed to the control foods. We can think of actually three distinct populations:

1. All individuals exposed to organic foods.
2. All individuals exposed to comfort foods.
3. All individuals exposed to control foods.

We now work to characterize the response within each of these three populations. Since the response of interest is a numeric variable (taking values between 1 and 7 with higher values indicating higher moral expectations), summarizing the variable using the mean is reasonable. That is, we might ask "does exposure to various food types lead to different moral expectations, *on average*?" Our question now compares the mean response across the groups. In particular, our question is looking for some type of difference in this mean response across the groups; our working hypothesis is then that the groups are all equivalent, on average. This could be framed in the following hypotheses:

H_0 : the average moral expectations are the same following exposure to each of the three types of food.

H_1 : the average moral expectations following exposure to food differ for at least one of the three types.

This is equivalent to expressing the hypotheses in terms of a relation between the two variables:

H_0 : there is no association between the type of food an individual is exposed to and their moral expectations, on average.
H_1 : there is is an association between the type of food an individual is exposed to and their moral expectations, on average.

We can represent these hypotheses mathematically as

$H_0 : \mu_{\text{comfort}} = \mu_{\text{control}} = \mu_{\text{organic}}$
H_1 : At least one μ differs from the others

where μ_{comfort} is the mean moral expectations for individuals exposed to comfort foods, etc. The question is now well-posed — it is centered on the population and captured through parameters.

For this particular setting, there is an alternative way of thinking about the population. You might argue that there are not three distinct populations; instead, there is only a single population (all individuals) and three different exposures (organic, comfort and control foods). This is a reasonable way of characterizing the population. The hypotheses remain the same:

$H_0 : \mu_{\text{comfort}} = \mu_{\text{control}} = \mu_{\text{organic}}$
H_1 : At least one μ differs from the others

The difference is in our interpretation of the parameters. We would describe μ_{comfort} as the mean moral expectations when an individual is exposed to comfort foods. The distinction, while subtle is to place emphasis on switching an individual from one group to another instead of the groups being completely distinct. In fact, this latter way of thinking is more in line with how the study was conducted. Individuals were allocated to one of the exposure groups, suggesting that exposure is something that could be changed for an individual.

From an analysis perspective, there is little difference between these two ways of describing the population. The difference is primarily in our interpretation. In many cases, we can envision the population either way; however, there are a few instances where that is not possible. Suppose we were comparing the average number of offspring of mice compared to rats (a lovely thought, I know). It does not make sense to think about changing a mouse into a rat; here, it only makes sense to think about two distinct populations being compared on some metric. How we describe the population is often related to the question we are asking.

Tip: How we describe the population is often connected to the study design we implement. In a controlled experiment, we envision a single population under various conditions. For an observational study, we generally consider distinct populations.

10.1 General Setting

This unit is concerned with comparing the mean response of a numeric variable across k groups. Let $\mu_1, \mu_2, \ldots, \mu_k$ represent the mean response for each of the k groups. Then, we are primarily interested in the following hypotheses:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
H_1 : At least one μ differs from the others

When there are only two groups ($k = 2$), then this can be written as

$H_0 : \mu_1 = \mu_2$
$H_1 : \mu_1 \neq \mu_2$

Tip: When there are two groups, it makes sense to say the means are equal or not. While tempting to do something when there are more than two groups, it is not possible. The opposite of "all groups equal" is *not* "all groups differ." The opposite of "all groups equal" is "at least one differs," which is what we are capturing with the above hypotheses. Keep it simple and do not try to get fancy with the notation.

Here we are writing things in the mathematical notation, but let's not forget that every hypothesis has a context. Throughout this unit, we are looking for some signal in the *location* of the response across the groups. Our working assumption then states that the groups are all similar, *on average*. This may not be the only comparison of interest to make in practice. For example, it may not be the location that is of interest but the spread of a process. In some applications, managers would prefer to choose the process that is the most precise. These questions are beyond the scope of this unit, but the concepts are similar to what we discuss here.

Chapter 11

Study Design

Chapter **??** discussed the impact that the design of the study has on interpreting the results. Recall that the goal of any statistical analysis is to use the sample to say something about the underlying population. Observational studies are subject to confounding. In order to use the available data in order to make causal statements that apply within the population, we need to address the confounding. There are two ways of doing this:

1. Conduct a controlled experiment. While we do not limit our discussion to controlled experiments in this unit, our discussion will emphasize the elements of a well designed experiment.
2. Use observational data and account for confounders. This will be the emphasis of the discussion in the subsequent unit.

As discussed in Chapter **??**, controlled experiments balance the groups being compared relative to the potential confounders. As a result, such studies permit causal conclusions to be drawn.

11.1 Aspects of a Well Designed Experiment

Generally speaking, there are three components to a well-designed study: replication, randomization, and comparative groups.

As we have stated repeatedly, variability is inherit in any process. We know there is variability in the population; not every subject will respond exactly the same to each treatment. Therefore, our questions do not seek to answer statements about individuals but about general trends in the population. In order to establish these general trends, we must allow that subject-to-subject variability be present within the study itself. This is accomplished through **replication**, obtaining data on multiple subjects from each group. Each subject's response would be expected to be similar, with variability within the group due to the inherit variability in the data-generating process.

Definition 11.1 (Replication)**.** Taking measurements on different subjects, for which you expect the results to be similar. That is, any variability is due to nautral variability within the population.

When we talk about gathering "more data," we typically mean obtaining a larger number of replicates. Ideally, replicates will be obtained through *randomly selecting* from the underlying population to ensure they are representative. The subjects are then *randomly allocated* to a particular level of the factor under study (randomly allocated to a group). This random allocation breaks the link between the factor and any potential confounders, allowing for causal interpretations. However, if a link exists between the factor and the response, that is preserved. These are the two aspects of **randomization**.

Definition 11.2 (Randomization)**.** Refers to the random *selection* of subjects which minimizes bias and random *allocation* of subjects which permits causal interpretation.

> **Tip:** While students can typically describe random selection vs. random allocation, they often confuse their purpose. Random selection is to ensure the sample is representative. Random allocation balances the groups with respect to confounders.

We now have two sources of variability. That is, we have two reasons the response will differ from one subject to another. Subjects assigned to different groups may differ because of an effect due to the group; this is a signal that we are tyring to identify with our hypotheses. Subjects within the same group will differ due to natural variability.

Random allocation ensures the groups are balanced with respect to confounders. However, there may still be a lot of variability within each group. The more variability present, the more difficult it is to detect a signal. The study will have more **power** to detect the signal if the groups are similar. This is the idea of having **comparative groups**.

Definition 11.3 (Power)**.** Refers to the probability that a study will find a signal when one really exists in the data generating process. This is like saying "the probability a jury will declare a defendant guilty when he actually committed the crime."

Definition 11.4 (Comparative Groups)**.** The idea that the treatment groups (levels of the factor under study) should be as similar as possible to reduce external variability in the process.

It is tempting to manually adjust the treatment groups to achieve what the researcher views as balance. This temptation should be avoided as balancing one feature of the subjects may lead to an imbalance in other features. We want to rely on randomization. However, when there is a particular feature which we would like to balance, we can employ specialized randomization techniques. For example, if we would like an equal number of males and females in a study, we can use stratified random sampling (see Definition 4.3) to ensure equal representation. During the random allocation, we can employ **blocking**, in which the random allocation to treatments happens within a secondary feature.

Definition 11.5 (Blocking)**.** One way of minimizing variability contributed by an inherit characteristic. All observations that are linked through the characteristic are grouped together and random allocation occurs *within* the block.

Example 11.1 (Overseeding Golf Greens)**.** Golf is a major pasttime, especially in southern states. Each winter, the putting greens need to be overseeded with grasses that will thrive in cooler weather. This can affect how the ball rolls along the green. Dudeck and Peeacock (1981) reports on an experiment that involved comparing the ball roll for greens seeded with one of five varieties of rye grass. Ball roll was measured by the mean distance (in meters) that five balls traveled on the green. In order to induce a constant initial velocity, each ball was rolled down an inclined plane.

Because the distance a ball rolls is influenced by the slope of the green, 20 greens were placed into four groups in such a way that the five greens in the same group had a similar slope. Then, within each of these four groups, one green was randomly assigned to be overseeded with one of the five types of Rye grass. The average ball roll was recorded for each of the 20 greens.

The data for Example 11.1 is shown in Table 11.1.

It would have been easy to simply assign 4 greens to each of the Rye grass varieties; the random allocation would have balanced the slope of the greens across the five varieties. However, an additional layer was added to the design in order to control some of that additional variability. In particular, greens with similar slopes were grouped together; then, the random allocation to Rye grass varieties happened *within* groups of greens. As a result, what we see is that there is one green of each type of slope for each Rye grass variety. This has the effect of reducing variability due to nuisance characteristics of the subjects.

> **Tip:** Blocking is often a way of gaining additional power when limited resources require your study to have a small sample size.

The extreme case of blocking occurs when you have repeatedly measure the response on the same subject

Table 11.1: Data from Overseeding Golf Greens example.

Rye Grass Variety	Slope of Green Grouping	Mean Distance Traveled (m)
A	1	2.764
B	1	2.568
C	1	2.506
D	1	2.612
E	1	2.238
A	2	3.043
B	2	2.977
C	2	2.533
D	2	2.675
E	2	2.616
A	3	2.600
B	3	2.183
C	3	2.334
D	3	2.164
E	3	2.127
A	4	3.049
B	4	3.028
C	4	2.895
D	4	2.724
E	4	2.697

under different treatment conditions. For example, a pre-test/post-test study is an example of a study which incorporates blocking. In this case, the blocks are the individual subjects. The subjects then undergo each of the possible treatment options. The rationale here is to use every subject as his or her own control. The treatment groups are then as similar as possible.

We do note that blocking, while a powerful aspect of a design, has an impact on the type of analysis that can be conducted. Specifically, we must account for the blocking when conducting the analysis. We will discuss this in Chapter **??**.

How did the design of the Organic Food Case Study incorporate these aspects? First, we notice that random allocation was utilized. Each of the 124 participants was randomly assigned to one of three treatment groups (type of food to which the particpant was exposed). The random allocation allows us to make causal conclusions from the data as any confounder should be balanced across the three foods. For example, subjects who adhere to a strict diet for religious purposes would naturally tend toward organic foods and higher moral expectations. However, for each subject like this exposed to organic foods, there is someone like this (on average) who was assigned to the comfort foods (on average). We also note that there is replication. Instead of assigning only one subject to each of the three treatment groups, we have several subjects within each group. This allows us to evaluate the degree to which the results vary within a particular treatment group.

The study does not make use of blocking. There are a couple of potential reasons for this; first, with such a large sample size, the researchers may not thought it necessary. Second, it could be that there was a restriction on timee. For example, researchers may have considered having students be exposed to each of the three types of food and answering different scenarios after each. However, this would take a longer amount of time to collect data. Third, it could be that researchers were not concerned about any identifiable characteristics that would generate additional variability. Regardless, the study is not worse off because it did not use blocking; the design is still a very reliable design.

While it is clear that random allocation was utilized in the design, random selection was not. Students participating in the study are those from a particular lecture hall. As a result, these students were not randomly sampled from all college students (or even from the university student body). As a result, we must

really consider whether the conclusions drawn from this study would apply to all college students within the United States. Having additional information on their demographics may help determine this, but in general, this is not something that can be definitively answered. It is an assumption we are either willing to make or not.

11.2 Collecting Observational Data

An inability to conduct a controlled experiment does not mean we neglect study design. Random sampling is still critical to ensuring that the data is representative of the population. Similarly, ensuring there are a sufficient number of replications to capture the variability within the data is an important aspect of conducting an observational study. When collecting observational data, one of the most important steps is constructing a list of potential confounders and then collecting data on these variables. This will allow us to account for these confounders in our analyses, as we will discuss in the next unit.

Chapter 12

Presenting the Data

When a research question involves the relationship between two or more variables, such as comparing the mean response across levels of a factor, successful presentations of the data which address the question of interest *partition the variability*. This key idea is essential to both the data presentation and the data analysis.

We have already argued that variability makes addressing questions difficult. If every subject had the same response to a particular exposure, there would be no need for statistics. We would simply evaluate one subject and determine which treatment to give. Statistics exists because of the ambiguity created by variability in the responses. In response to this variability, our statistical graphics and models distinguish (partition) the various sources of variability. That is, with any analysis, we try to answer the question "why aren't all the values the same? What are the reason(s) for the difference we are observing?"

From the Organic Food Case Study, consider the primary question of interest:

> Is there evidence of a relationship between the type of food a person is exposed to and their moral expectations, on average, following exposure?

What we are really asking is "does the food exposure help explain the differences in the moral expectations of individuals?" We know that there are differences in moral expectations between individuals. But, are these differences solely due to natural variability (some people are just inherently, possibly due to how they were raised, more or less liberal with moral beliefs); or, is there some systematic component that explains at least a portion of the differences between individuals. We are thinking about partitioning the "why the responses differ" (the variability).

A good graphic must then tease out how much of the differences in the moral expectations is from subject-to-subject variability and how much is due to the food exposure. First, consider a common graphic which is **not** useful in this situation (Figure 12.1).

To determine an appropriate graphic, we need to remember that we want to partition the variability. So, we must not only compare the differences between the groups but also allow the viewer to get a sense of the variability within the group. A common way of doing this within the engineering and sciences is to construct side-by-side boxplots, as illustrated in Figure 12.2.

From the graphic, we see that the moral expectation scores seem to have nearly the same pattern in each of the exposure groups. More, the center of each of the groups is roughly the same. That is, there does not appear to be any evidence that the type of food to which a subject is exposed is associated with moral expectations, on average.

Side-by-side boxplots can be helpful in comparing large samples as they summarize the location and spread of the data. When the sample is smaller, it can be helpful to overlay the raw data on the graphic in addition to the summary provided by the boxplot. We might also consider adding additional information, like the mean within each group. An alternative to boxplots is to use violin plots which emphasize the shape of the distribution instead of summarizing it like boxplots. Yet another option is to construct density plots which

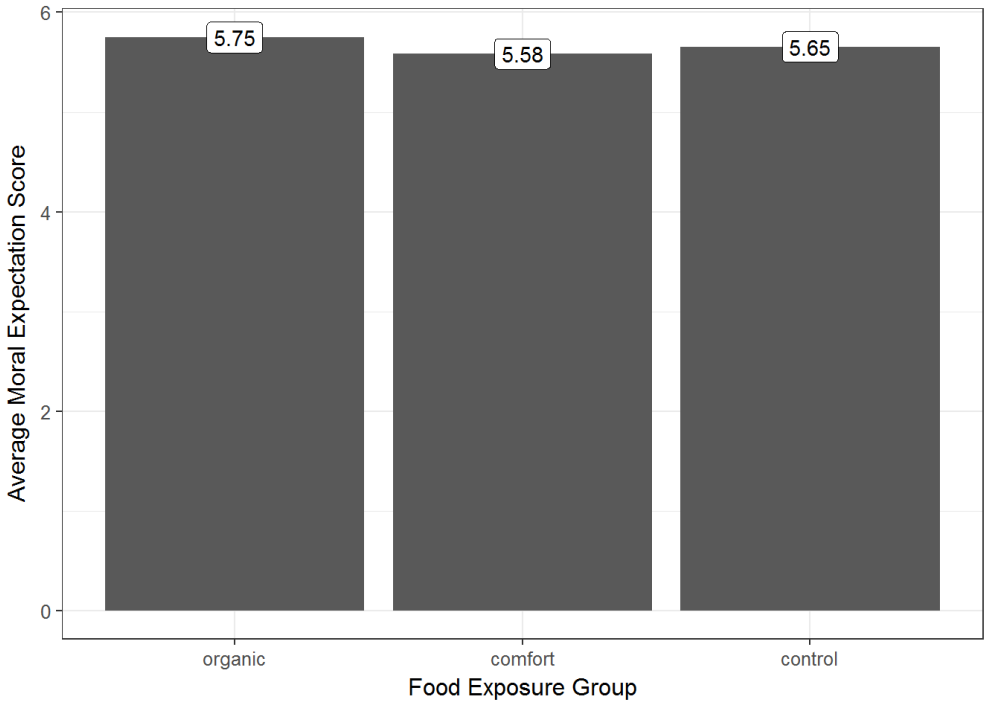


Figure 12.1: Illustration of a poor graphic using the Organic Food Case Study; the graphic does not give us a sense of variability. As a result, it is not clear how different these means really are.

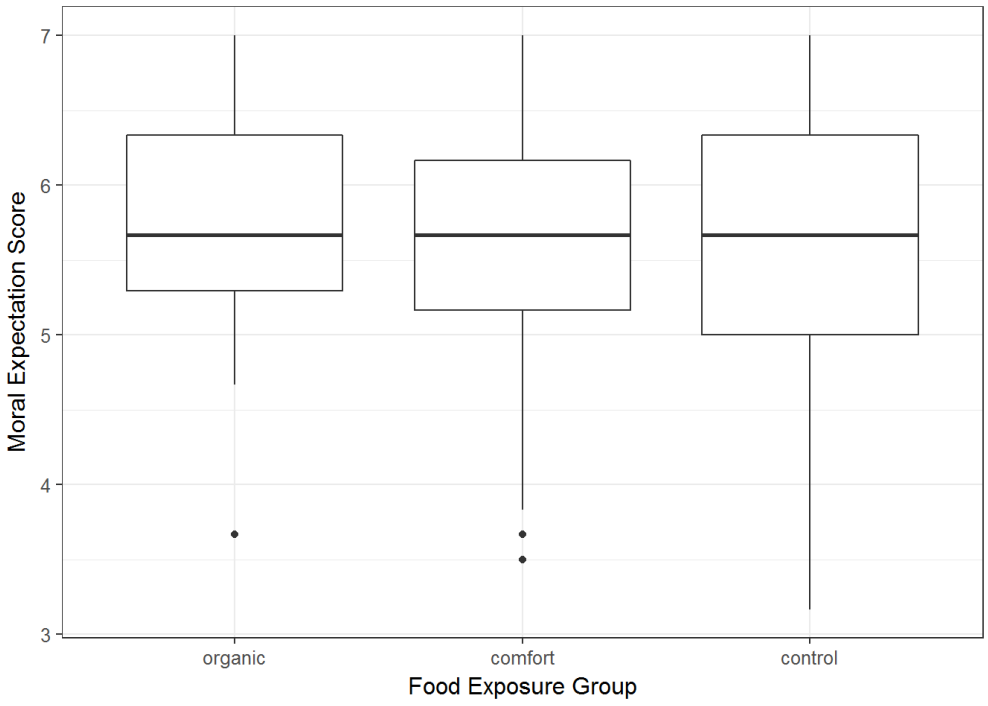


Figure 12.2: Comparison of the moral expectations for college students exposed to different types of food.

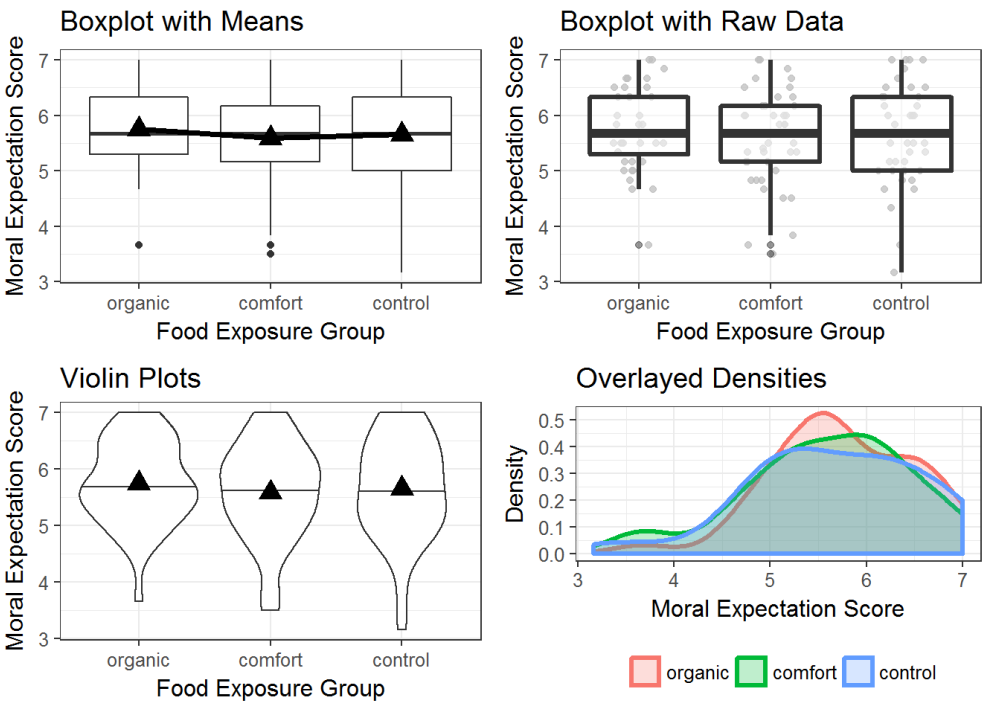


Figure 12.3: Multiple ways to effectively compare the response across multiple groups.

are overlayed on one another. This works when there are only a small number of groups; if the number of groups is large, then placing the distributions side-by-side is much more effective. A comparison of these approaches is in Figure **??**.

Each of these plots is reasonable. What makes them useful in addressing the question is that in each plot, we can compare the degree to which the groups differ relative to the variability within a group. That is, we partition the variability. With each plot, we can say that one of the reasons the groups differ is because of exposure to different food types; however, this difference is extremely small relative to the fact that regardless of which food group you were exposed to, the variability in moral expectations with that group is quite large. Since the predominant variability in the moral exposure is the variability within the groups, we would say there is no signal here. That is, there is no evidence that the average scores differ across food exposure groups.

The key to a good summary is understanding the question of interest and building a graphic which addresses this question through a useful characterization of the variability.

Chapter 13

Quantifying the Evidence

Figure 13.1 displays a numeric response across three groups for two different datasets. Consider the following question:

> For which dataset is there *stronger* evidence that the response is associated with the grouping variable?

Nearly everyone will say that Dataset A provides stronger evidence of a relationship between the grouping variable and the response. We generated these data such that the mean for Groups I, II and II are 5, 6 and 7, respectively, *for both Datasets A and B*. While there is a difference, on average, in the response across the groups in both cases, it is correct that Dataset A provides stronger evidence for that relationship. The real question is "what is it that leads everyone to make the same conclusion when we have not yet discussed how to analyze this data?" When we ask students why they feel Dataset A provides stronger evidence, we typically hear that it is because the "gaps" between the groups "look bigger." In essence, that is exactly right!

13.1 Partitioning Variability

Subconsciously, when we are deciding whether there is a difference between the groups, we are partitioning the variability in the response. We are essentially describing two sources of variability: the variability in the response caused by subjects belonging to different groups and the variability in the response within a group (Figure 13.2). In both Datasets A and B from Figure 13.1, the **between-group variability** is the same; the difference in the means from one group to another is the same in both cases. However, the **within-group variability** is much smaller for Dataset A compared to Dataset B.

Definition 13.1 (Between Group Variability)**.** The variability in the average response from one group to another.

Definition 13.2 (Within Group Variability)**.** The variability in the response within a particular group.

Figure 13.1 then illustrates the larger the variability between groups *relative to* the variability within groups, the stronger the signal. Quantifying the strength of a signal is then about quantifying the ratio of these two sources of variability. Let this sink in because it is completely counter-intuitive. We are saying that in order to determine if there is a difference in the mean response across groups, we have to examine variability. Further, a signal in data is measured by the variability it produces. For this reason, comparing a quantitative response across a categorical variable is often referred to as Analysis of Variance (ANOVA).

Key Idea: Consider the ratio of the variability between groups to the variability within groups. The larger this ratio, the stronger the evidence of a signal provided by the data.

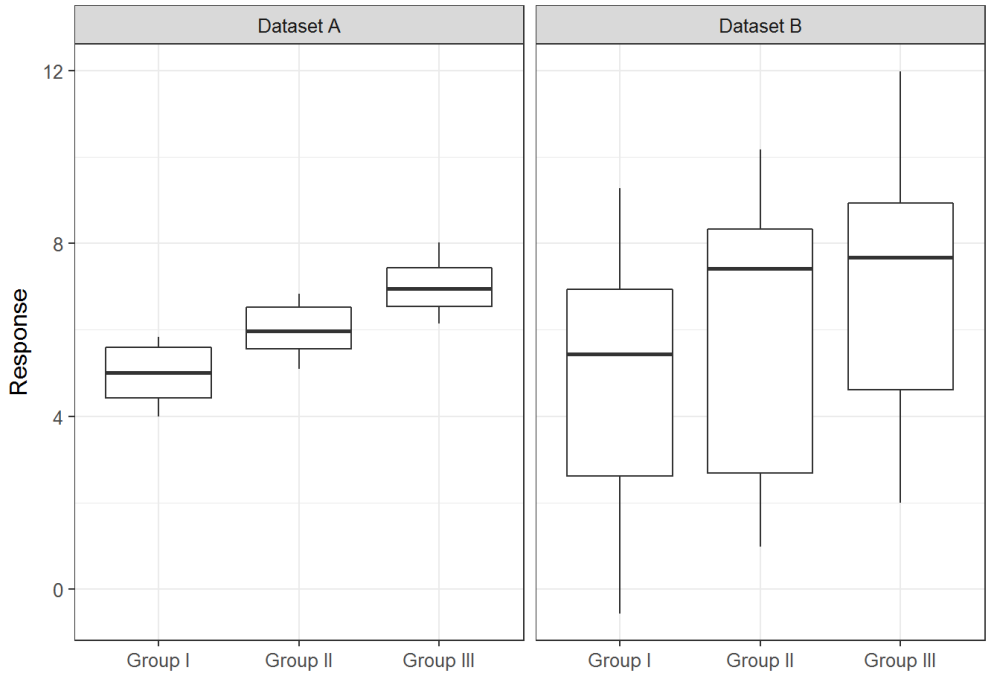


Figure 13.1: Simulated data illustrating that signal strength is determined by partitioning variability. There is a clear signal (difference in the location across groups) for Dataset A but not for Dataset B.

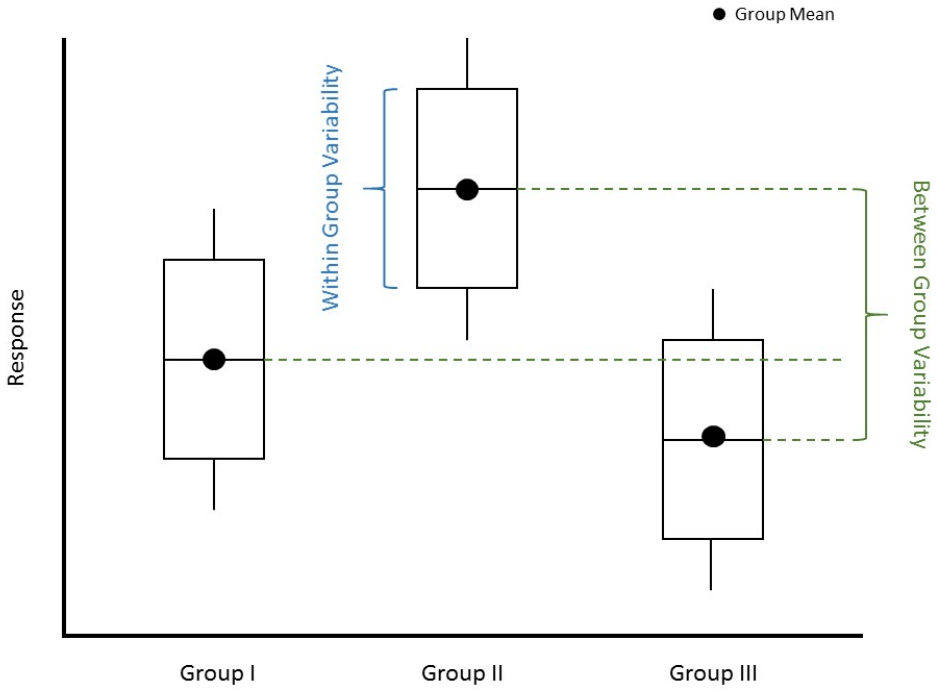


Figure 13.2: Illustration of partitioning the variability in the response to assess the strength of a signal.

13.2 Forming a Standardized Test Statistic

As we stated above, quantifying the strength of a signal is equivalent to quantifying the ratio of two sources of variability. Such ratios are known as **standardized test statistics**.

Definition 13.3 (Standardized Test Statistic)**.** A ratio of two sources of variability, or a signal-to-noise ratio. The larger the test statistic, the stronger the evidence of a signal; said another way, the larger the test statistic, the stronger the evidence against the null hypothesis.

Based on our observations above, the standardized test statistic for comparing the mean response across multiple groups has the general form

$$T = \frac{\text{(Between Group Variability)}}{\text{(Within Group Variability)}} \tag{13.1}$$

The question we then have before us is the following: how do we measure these sources of variability? Consider again the hypothesis of interest for the Oranic Food Case Study:

$H_0 : \mu_{\text{comfort}} = \mu_{\text{control}} = \mu_{\text{organic}}$
H_1 : At least one μ differs from the others

In order to form the standardized test statistic, let's again think about what constitutes evidence *against* the null hypothesis. The more the means differ from one another, the stronger the evidence. But, in the previous unit, we had a measure for how different values were from one another — variance. That is, the *between-group* variability can be measured by the variance of the means; we call this the **Mean Square for Treatment (MSTrt)**.

Definition 13.4 (Mean Square for Treatment (MSTrt))**.** This captures the between-group variability in an Analysis of Variance; it is a weighted variance among the sample means from the various groups. It represents the signal.

Since we do not know the means for each groups (remember, each μ is a parameter), we assess the between group variability within the sample using the estimates for these parameters — the sample means. This is our signal. The larger this variance, the further apart the means are from one another (agreeing with the alternative hypothesis); the smaller this variance, the closer the means are (agreeing with the null hypothesis).

While the numerator provides some measure of the size of the signal, we need again need to consider how much noise is within the data. Again, in Figure 13.1, the variability between the means is identical for the two datasets; the signal is stronger for Dataset A because this variability is larger *with respect to the noise*. In order to capture the *within-group* variability, we pool the variances for each group; this is called the **Mean Square for Error (MSE)**.

Definition 13.5 (Mean Square for Error (MSE))**.** This captures the within-group variability; it is a pooled estimate of the variance within the groups. It represents the noise.

Our test statistic in Equation (13.1) is then refined to

$$T = \frac{MSTrt}{MSE} \tag{13.2}$$

Tip: Consider testing the hypotheses $> H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$
$> H_1$: At least one μ differs from the others
The standardized test statistic of interest is

$$T = \frac{MSTrt}{MSE}$$

where

$$MSTrt = \frac{1}{k-1}\sum_{j=1}^{k} n_j \left(\bar{y}_j - \bar{y}\right)^2$$

$$MSE = \frac{1}{n-k}\sum_{j=1}^{k} \left(n_j - 1\right) s_j^2$$

and n_j represents the sample size for the j-th group, $\bar{y}_j$ represents the sample mean for the j-th group, $\bar{y}$ represents the overall mean response across all groups, and s_j^2 represents the sample variance for the j-th group.

We note that while mathematical formulas have been provided to add some clarity to those who think algebraically, our emphasis is *not* on the computational formulas as much as the idea that we are comparing two sources of variability.

13.3 Obtaining a P-value

Standardized test statistics quantify the strength of a signal, but they do not allow for easy interpretation. However, with a standardized test statistic, we are able to compute a p-value to quantify how unlikely our particular sample is. That is, we need to construct the null distribution for the standardized test statistic. We need to know what type of signal we would expect if the null hypothesis were true. Conceptually, this is no different than it was in Unit I. We consider running the study again in a world in which all the groups are the same; for the Organic Food Case Study, this would involve - Obtaining a new sample of students. - Randomizing each student to one of the three groups at random, all showing the same foods. - Having each student answer a questionnaire regarding moral dilemmas. - Summarize the data by computing a standardized test statistic.

Notice the difference in step 2 above compared to what actually happened in the real study. In the real study, each group had a different set of foods. This was to answer the question about whether there is a difference in the groups. However, in order to construct the *null distribution*, we need to force all groups to be the same. This could be accomplished by showing every group the same set of foods. The primary difference in this unit is that the strength of the signal is measured through a standardized test statistic. After repeating the above steps over and over again, we determine how often the recorded standardized test statistics exceeded the value we obtained in our actual sample.

Figure 13.3 represents the null distribution of the standardized test statistic. Again, these are values of the standardized test statistic we would expect if there were no relationship between the food categories to which the students were exposed and their moral score. We are then interested in finding out if the observed dataset is consistent with these expectations.

Notice that in our data, we observed a standardized test statistic of 0.41; based on the null distribution, we would expect a signal this strong or stronger about 66.9% of the time *when no signal existed at the population* (by chance alone). That is, our data is quite consistent with what we would expect under the null hypothesis. There is no evidence of a relationship between the type of food a student is exposed to and their moral expectations, on average.

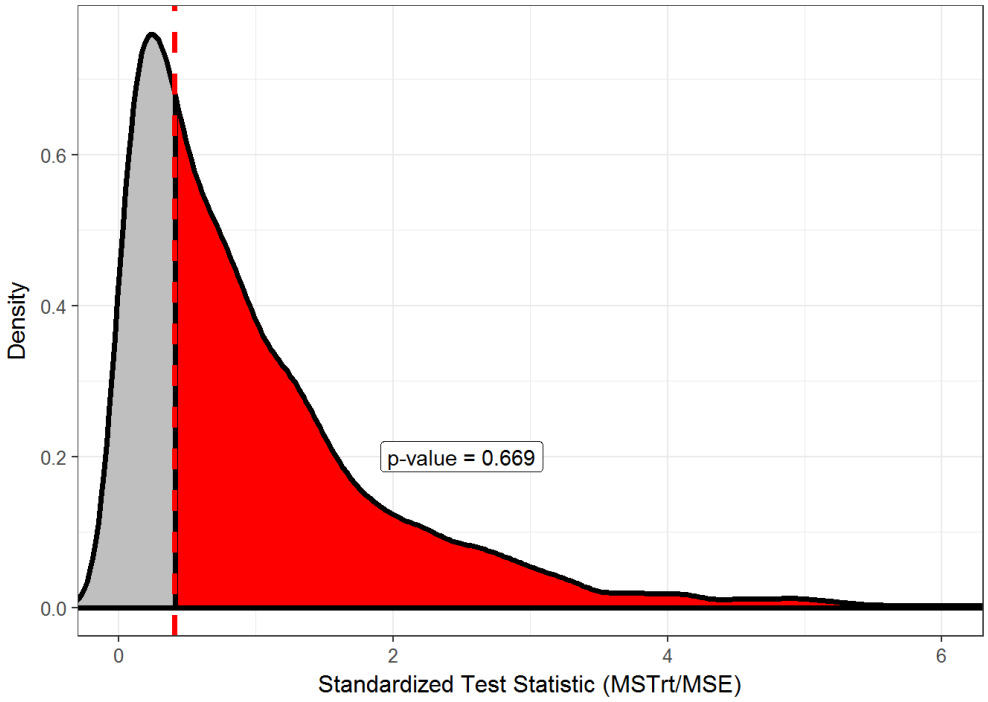


Figure 13.3: Computation of the p-value for the Organic Food Case Study by simulating the null distribution. The null distribution is based on 5000 replications.

Again, conceptually, this is similar to what we saw in the previous unit. We are simply determining how likely our data is under the null hypothesis. However, unlike the previous unit, it may not be clear how we actually model this null distribution. If we cannot physically redo the study, how can we construct this model? In order to understand this, we must consider a different model — that for the data generating process. This is the topic of the next chapter.

Chapter 14

Building the Statistical Model

The numerical summaries of any study are subject to sampling variability. That is, if we were to repeat the study with new subjects, the statistics we compute would almost certainly change to some degree. The key to feeling confident in our results is to quantify the variability in our estimates; this was the argument made in Chapters 6 and 7. The goal of any statistical analysis is then to develop a model for the sampling (or null) distribution of a statistic. Often times, this requires modeling the data-generating process as a precursor. As in any other discipline, statistical models simplify the process being modeled by making certain assumptions. In this chapter, we develop a model that will help us make inference about the mean of several populations.

14.1 General Formulation

Consider dropping a tennis ball from the top of a 50-meter building and recording the time required before the ball hits the ground. Applying the principles learned in a first course in physics, we would be able to compute the time precisely using the formula

$$\text{time} = \sqrt{\frac{2(\text{distance})}{9.8}}$$

where $9.8 m/s^2$ is the acceleration due to gravity; further, this formula works regardless of the mass of the object. Plugging 50 meters into the equation yields a time of 10.2 seconds. If we were to drop a second tennis ball from the same building, the formula tells us that it will also take 10.2 seconds to hit the ground below. This is known as a **deterministic** system since entering a constant input always results in the same output.

Definition 14.1 (Deterministic Process)**.** One which is completely determined by the inputs. That is, entering the same input twice will always result in the same output with certainty.

This is a model; it simplifies extremely complex processes involving the gravitational pull between objects and works reasonably well. However, it does not always match reality. If we were to repeatedly drop tennis balls from the same 50-meter building and record the time before hitting the ground, we might find that the time differs slightly from one ball to the next. There are several reasons why our observed responses do not line up directly with those predicted by the above equation; for example, our device for measuring time may be subject to some measurement error, a strong gust of wind could alter the results (while the above equation assumes no air resistance), or the person dropping the ball may have inadvertantly increased the initial velocity of the ball. These reasons, and others, contribute to the observations not lining up with the model. That is, there is associated noise in the resulting measurements. A model which incorporates this noise might be written as

$$\text{time} = \sqrt{\frac{2(\text{distance})}{9.8}} + \text{noise}$$

where the noise is not a known quantity. As a result, this is a **stochastic** model as the same value for distance may result in different outputs each time.

Definition 14.2 (Stochastic Process)**.** One which has an element of randomness. That is, the resulting output of the system cannot be predicted with certainty.

This leads us to our general formulation for a statistical model:

$$\text{Response} = f(\text{variables, parameters}) + \text{noise} \tag{14.1}$$

The response we observe is the result of two components:

- A deterministic component which takes the form of a function of variables and unknown parameters. It is often this component on which we would like to make inference.
- A stochastic component which captures the unexplained variability in the data generating process.

Since the noise is a random element, it has a distribution. We often make additional assumptions on the structure of this distribution to enable inference on the deterministic component of the model. We discuss this later in the chapter.

This general model adheres to the idea of partitioning the variability in the response. It says that a part of the reason the responses differ between subjects is because they have different variables (remember, parameters are fixed for all subjects in a population); part of the reason is unexplained noise. The overall goal of a statistical model is to give an explanation for why the data is what it is. How did it come to be? What process generated the values I have observed? Our statistical model says that these values have some deterministic component plus some additional noise we cannot explain. We now turn towards employing this model in the case of comparing the mean response for multiple groups.

14.2 Statistical Model for A Quantitative Response and a Categorical Predictor

For the Organic Food Case Study, we are comparing the moral expectations (quantitative response) for different food exposures (levels of a categorical variable). Our model for the data-generating process is best understood in light of the graphic we used to display the data (see Figure 14.1).

Let's consider how the value 3.67, highlighted red in Figure 14.1, was generated. As discussed previously, there are two sources of variability in the moral expectation scores (two reasons that the values are not all the same). One source is the fact that different subjects had different exposures. That is, one reason the value 3.67 differs from others observed is because this subject belongs to the organic group and not the comfort or control exposure groups. As this is something we can explain, it goes into the deterministic portion of the model; it is a function of known variables (group exposure). Let the function $f(\cdot)$ be such that the input is the group exposure for the i-th subject and the output is the mean moral expectation score for that group; this can be represented as a piecewise function:

$$f\left((\text{Food Exposure Group})_i\right) = \begin{cases} \mu_1 & \text{if i-th subject exposed to organic foods} \\ \mu_2 & \text{if i-th subject exposed to control foods} \\ \mu_3 & \text{if i-th subject exposed to comfort foods} \end{cases}$$

Notice that $f(\cdot)$ involves both a variable of interest as well as parameters of interest — the mean response μ_1, μ_2, μ_3 for each of the three groups. This function is perfectly acceptable, but it is cumbersome to write in a shortened form. Notice how the function works: it receives an input regarding which group, and it directs you to the appropriate parameter as an output. We can write this in a compact way as

$$f\left((\text{Food Exposure Group})_i\right) = \sum_{j=1}^{3} \mu_j \mathbb{I}\left(\text{i-th subject in food exposure group j}\right)$$

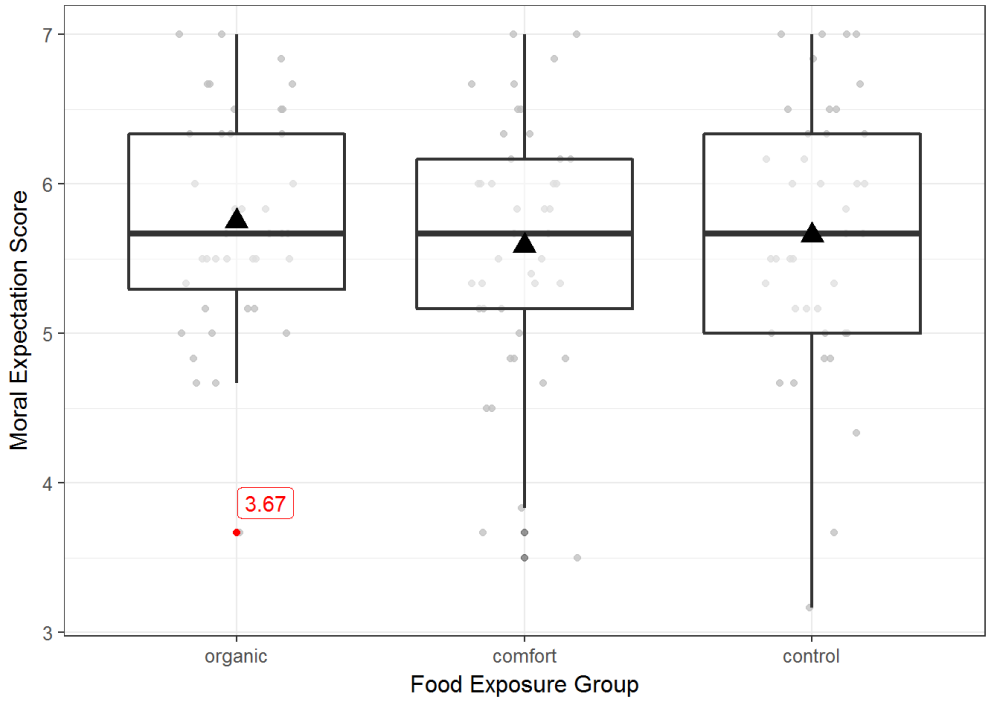


Figure 14.1: Moral expectation scores for students following exposure to various food types.

where $\mathbb{I}(\cdot)$ is the indicator function taking value 1 if the event occurs and 0 otherwise.

Key Idea: The deterministic component of a statistical model incorporates the parameters which govern the question of interest. It is built to explain differences in the response based on differences in group membership or other characteristics of the subjects.

This is the deterministic part of the model, as inputing the same group always results in the same output — the unknown parameter characterizing the mean response for the group. This, however, only captures one reason we feel the responses differ across subject. This deterministic component says that every single person exposed to the same food group should have the same moral expectations. It does not explain why subjects within the organic group do not all share the average moral expectation score. This source of variability is something we cannot fully explain but attribute to natural variability in this group or measurement error in how we obtained the response. In order to capture this, we add noise to the system, and we allow this noise to be a random variable which is unique to each subject within the population. Letting ϵ_i represent the noise accompanying the response of the i-th subject, we can now extend the model in Equation (14.1) to accommodate these two sourses of variability and obtain

$$(\text{Moral Expectation Score})_i = \sum_{j=1}^{3} \mu_j \mathbb{I}(\text{i-th subject in food exposure group j}) + \epsilon_i$$

This may be written in shorthand (suppressing the parameters and noise) as

$$\text{Moral Expectation Score} \sim \text{Food Exposure Group}$$

Key Idea: The stochastic component of a statistical model captures the unexplained variability due to natural variability in the population or measurement error in the response.

Tip: In general, given a quantitative response variable y, our model for the data generating process comparing this variable across several levels of a factor is

$$y_i = \sum_{j=1}^{k} \mu_j \mathbb{I}(\text{i-th subject in factor level j}) + \epsilon_i$$

In general, students struggle with the fact that we have two different models floating around. Currently, we are modeling the data-generating process. This model is used to develop a secondary model of the sampling distribution (or null distribution) of a statistic of interest. It is this secondary model that is actually necessary in order to conduct inference; the model for the data-generating process is simply a stepping stone to the model of interest.

14.3 Conditions on the Error Distribution

In our model for the data-generating process we incorporated a component ϵ to capture the noise within each group. Since the error is a random variable (stochastic element), we know it has a distribution. We typically assume a certain structure to this distribution. The more assumptions we are willing to make, the easier the analysis, but the less likely our model is to be applicable to the actual data-generating process we have observed. The conditions we make dictate how we conduct inference (the computation of a p-value or confidence interval).

The first condition we consider is that the noise attributed to one observed individual is **independent** of the noise attributed to any other individual observed. That is, the amount of error in any one individual's response is unrelated to the error in any other response observed. It is easiest to understand this condition by examining a case when the condition would not hold.

Definition 14.3 (Independence)**.** Two variables are said to be independent when the likelihood that one variable takes on a particular value does not depend on the value of the other variable.

Example 14.1 (Programming Speed)**.** Suppose we are conducting a study to compare the speed required to complete a particular programming task in two different languages: Python and R. We obtain a sample of 100 programmers previously exposed to Java but neither Python nor R. We ask each programmer to complete a programming exercise in Python and record the time required to successfully complete the task. Then, we ask each programmer to perform the same task in R and record the time required to successfully complete the task.

The model for the data generating process would be

$$(\text{Time})_i = \mu_1 \mathbb{I}(\text{i-th task programmed in Python}) + \mu_2 mathbbI(\text{i-th task programmed in R}) + \epsilon_i$$

Given the method in which the data was collected, it would not be reasonable to assume the errors are independent of one another. Some programmers are naturally faster than others. A programmer with a below average (negative ϵ) time in Python will most likely have a below average (negative ϵ) time in R on the same task. Therefore, there is a relationship between the errors for some of the observations taken. This violates the independence condition.

The second condition that is typically placed on the distribution of the errors is that the variability of the responses is similar within each group. This assumption is known as **homoskedasticity**.

Definition 14.4 (Homoskedasticity)**.** Also known as "constant variance," this assumption states that the variability of error terms for individuals within a group is the same across all groups.

Practically, this means that the responses in one group are not dramatically more variable than any other group (the width of the box portion of a boxplot should be roughly the same across groups). This condition

ensures that the precision of the measurements is roughly similar. In fact, we made use of this assumption in the construction of our standardized test statistic

$$T = \frac{MSTrt}{MSE}$$

since MSE was a pooled estimate of the variability. If we were not willing to assume that the variabilities were similar, we would not construct a pooled estimate. This also highlights that the MSE is an estimate of the variability of observations within any group when this condition is satisfied.

14.4 Simulating the Null Distribution

We note that this section is a bit more technical than other sections. We want to give the reader a feel for the computational aspect of simulating the null distribution. However, understanding conceptually that we are repeating the study in a world in which the null hypothesis is true is sufficient for interpreting a p-value.

Under the above conditions, we can model the null distribution of our standardized test statistic. The key here is to lean on our data generating process. Consider the Organic Food Case Study. *If the null hypothesis is true*, then we have that

$$\mu_{\text{organic}} = \mu_{\text{comfort}} = \mu_{\text{control}}$$

Let's define this common mean to be μ; we do not know what this value is, but it is common to all groups. Therefore, *if the null hypothesis is true*, we have that the data generating process reduces to

$$(\text{Moral Expectation Score})_i = \mu + \epsilon_i \tag{14.2}$$

Therefore, we can generate data according to this model. We can replace μ by our best estimate — the sample mean response across all observations regardless of their group. It simply remains to determine how to approximate a random variable from the noise distribution. In order to do this, we need estimates of the errors, known as **residuals**.

Definition 14.5 (Residual)**.** The difference between the observed response and the predicted response (estimated deterministic portion of the model). Residuals approximate the noise in the data-generating process.

The deterministic model gives a way of predicting the response. For example, consider the Organic Food Case Study; the data is reproduced in Figure **??**. Based on the data available, if a subject were to be exposed to organic foods, we would expect their moral expectation score to be NA; this is the average observed among individuals randomized to this treatment within our study.

That is, we can define the **predicted value** for i-th observation in our study as

$$\widehat{y}_i = \sum_{j=1}^{3} \bar{y}_j \mathbb{I}(\text{i-th subject in food exposure group j})$$

and the corresponding residual as

$$e_i = y_i - \widehat{y}_i$$

Let's not get lost in the mathematical notation; the residual here is simply the difference between the response of the subject and the average response for their corresponding group.

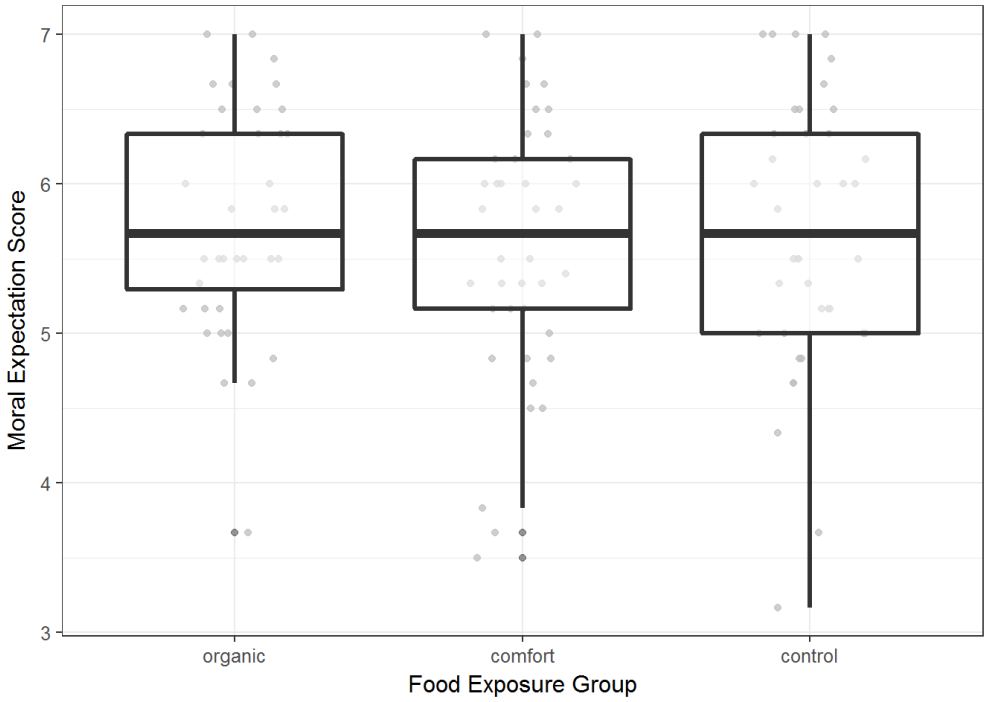


Figure 14.2: Comparison of the moral expectations for college students exposed to different types of food.

The key idea here is that residuals approximate the unseen error. Therefore, if we take this error and perturb it, we can generate new data. A new dataset, generate under the null hypothesis, can then be constructed as

$$y_i^* = \bar{y} + e_i^*$$

where y_i^* is then a new observation constructed by taking a mean and adding a perturbed version of the residual for that observation. Notice that each newly generated response has the same mean (so that the null is true). We then take this new dataset and compute the standardized test statistic as before and record it. Then, we repeat this process over and over again until we have constructed the null distribution. This gives us a sense of the p-value.

14.5 Recap

We have covered a lot of ground in this chapter, and it is worth taking a moment to summarize the big ideas. In order to construct a model for the null distribution of the standardized test statistic, we took a step back and modeled the data generating process. Such a model consists of two components: a deterministic component explaining the differences between groups and a stochastic component capturing the noise in the system.

Certain conditions are placed on the distribution of the noise in our model. Using these assumptions, we can generate data which adheres to the null hypothesis. Therefore, we can obtain an empirical model that suggests what values of a test statistic we might expect.

Chapter 15

Assessing Modeling Assumptions

In the previous chapter, we introduced a model for how a quantitative response being generated across multiple groups. For the Organic Food Case Study, this is essentially

$$\text{(Moral Expectation Score)}_i = \sum_{j=1}^{3} \mu_j \mathbb{I}(\text{i-th subject in food exposure group j}) + \epsilon_i$$

Further, we added two conditions to the distribution of the error term: 1. The error in the moral expectation score for one individual is independent of the error in the moral expectation score for all other individuals. 2. The variability in the error for the moral expectation score within a group is similar for any food exposure group.

Unfortunately, we cannot just state that these are the conditions we hope hold for the data generating process and move on our merry way. Since the p-value was computed assuming these conditions hold, the p-value is only meaningful if the data is consistent with these conditions. If any of these conditions is violated, then the p-value is meaningless.

Key Idea: Residuals, since they are estimates of the noise in the data-generating process, provide a way of assessing the modeling conditions placed on the distribution of the error term.

In this section, we discuss how to use residuals to assess these conditions qualitatively.

15.1 Assessing Independence

Generally, independence is assessed through the context of the data collection scheme. By carefully considering the manner in which the data was collected, we can typically determine whether it is reasonable that the errors in the response are independent of one another. Some key things to consider when examining the data collection process: - Are there repeated observations made on the same subject? This often suggests some type of relationship between the responses and therefore would not be consistent with errors being independent. - Is the response measured over time (time-series) such as daily temperature over the course of a month? Time-series data often exhibits strong period-to-period relationships suggesting the errors are not independent. For example, if it is hot today, it will probably be hot tomorrow as well. - Is there a learning curve in how the data was collected? Learning curves again suggest some dependence from one observation to the next.

Random sampling and random assignment allow us to confidently state that the errors are independent of one another. One additional pitfall to watch out for when collecting your own data is whether there is some

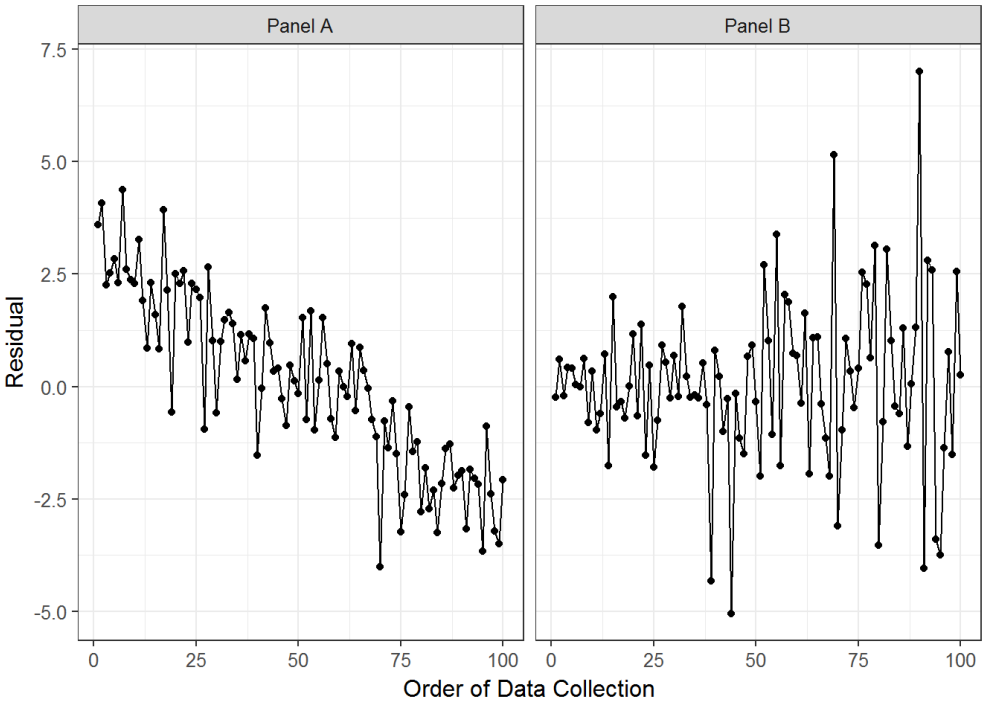


Figure 15.1: Examples of trends in a time-series plot of the residuals. Such trends indicate the data is not consistent with the condition that the errors are independent of one another.

type of systematic error in the measurement device. - Measurement devices which are failing over time will introduce a dependence from one observation to the next. Imagine a bathroom scale that begins to add an additional pound each day. Then, being above average weight one day will most likely lead to an above average weight the next, due primarily to the measurement device.

This last point illustrates a particular deviation from our condition of independence in which two observations collected close together in time are related. When we know the order in which the data was collected, we can assess whether the data is consistent with independence or tends to deviate in this manner. This is done graphically through a **time-series plot** of the *residuals*. If two errors were unrelated, then the value of one residual should tell us nothing about the value of the next residual. Therefore, a plot of the residuals over time should look like noise (since residuals are supposed to be estimates of noise). If there are any trends, then it suggests the data is not consistent with independence.

Definition 15.1 (Time Series Plot)**.** Plot of a variable over time. This plot allows us to assess some deviations from independence. A trend in the location or spread of the points over time suggests a deviation from independence.

As an example, consider the time-series plots shown in Figure 15.1, both representing hypothetical datasets. In Panel A, the residuals display a trend in the location over time. Knowing that a response was below average suggests the next response will also be below average. In Panel B, the results deplay a trend in the spread over time. This suggests that measurements taken later in the study were less precise. Both panels are then examples of patterns which would suggest the data is not consistent with the condition of independence.

Instead, if the data were consistent with the condition of independence on the error terms, we would expect to see a plot as in Figure 15.2. Notice there are no trends in the location or spread of the residuals.

For the Organic Food Case Study, participants were assessed simultaneously within a large lecture. Therefore, there is no ordering in time to be concerned about. Further, since students worked individually on the questionnaire, it is reasonable to assume that the errors in the moral expectation score are unrelated to one another.

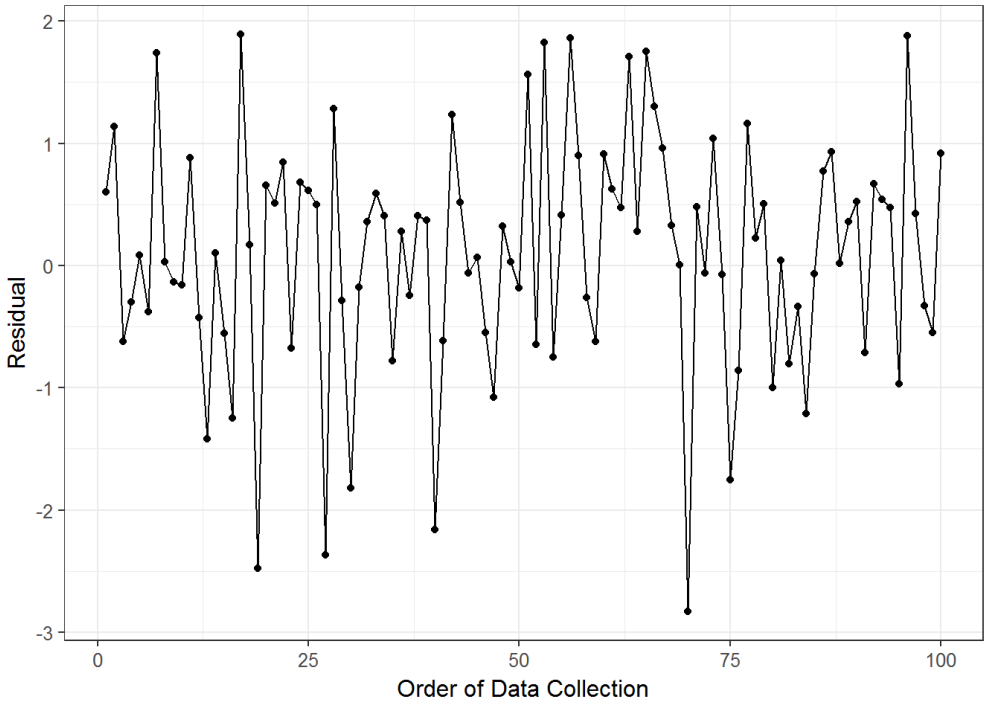


Figure 15.2: Example of a time-series plot of residuals which shows no trends in location or spread. This is consistent with what we would expect if the condition of independence among errors were satisfied.

15.2 Assessing Homoskedasticity

We want the variability in the errors within a group to be the same across the groups. This corresponds to the spread of the response within each group is the same. This implication leads to a simple way of assessing this assumption. Examining the side-by-side boxplots (or jitter plots, etc.) of the response allows us to get a sense of the variability within each group. Figure 15.3 shows the moral expectation score for each individual across the various groups. Notice that the boxes for each group are roughly the same size; that is, the interquartile ranges are similar. This suggests that the variability within each group is similar from one group to the next. That is, the data is consistent with this condition.

15.3 Assessing Normality

Assessing whether observations adhere to a particular distribution is a large area in statistical research. Many methods have been developed for this purpose. We emphasize a single graphical summary known as a **probability plot**. The construction of the plot is beyond the scope of this text, but the concepts underlying its construction actually tie in nicely to the big themes of the course. Recall that if a sample is representative, then it should be a snapshot of the underlying population. Therefore, if we believe the underlying population has some particular distribution, we would expect the properties of this distribution to be apparent in the sample as well.

If we believe the errors follow a Normal distribution, then it is reasonable that the residuals should maintain some of those properties. For example, the 10-th percentile of the residuals should roughly equate to the 10-th percentile expected from a Normal distribution. Mapping up the percentiles that we observe to those that we expect is the essence of a probability plot.

Definition 15.2 (Probability Plot)**.** Graphic for comparing a theoretical probability model for the distribution

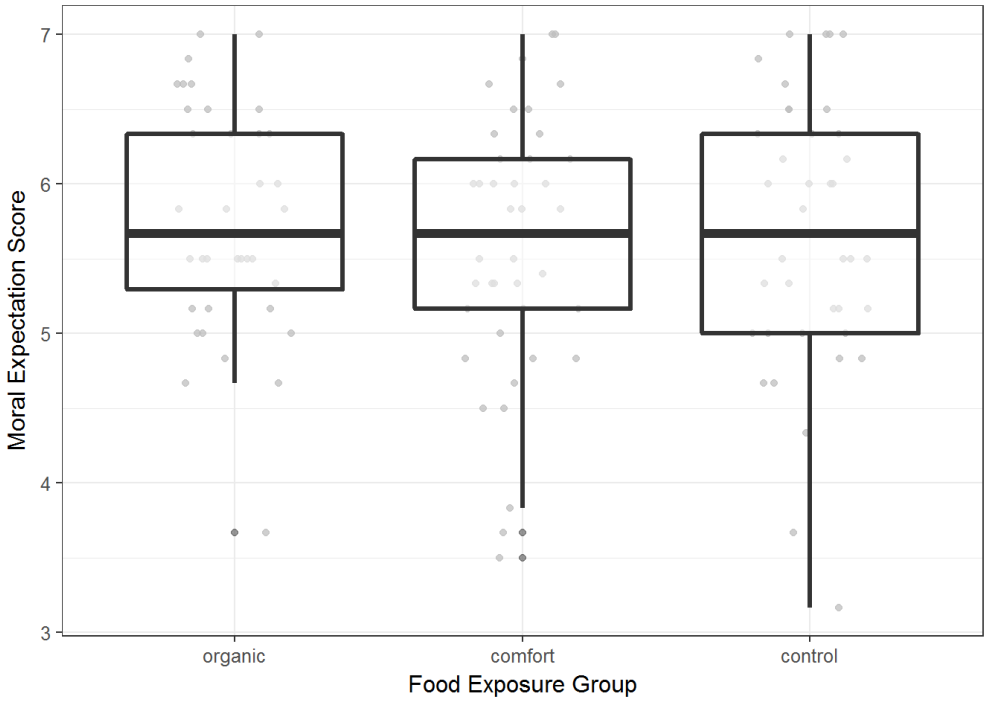


Figure 15.3: Comparison of the moral expectations for college students exposed to different types of food.

an underlying population with the distribution of the sample. Sample points should follow a straight line. If points deviate from this linear trend, that suggests the points do not align with the proposed model.

While a probability plot can be used for a host of probability distributions, the most common is the normal probability plot. Since we expect the percentiles to line up directly, we would expect a one-to-one linear relationship to be exhibited in the plot. Trends away from a linear relationship suggest the proposed Normal distribution is not a reasonable model for the distribution of the errors.

Figure 15.4 shows the probability plot for the residuals from the Organic Food Case Study.

Overall, the points do tend to follow a straight line. There are some deviations from a linear relationship at each end of the plot, but the deviations are not extreme. We argue that these residuals are consistent with the errors having a Normal distribution.

For comparison, Figure 15.5 illustrates a hypothetical dataset for which the residuals suggest the condition of the errors following a Normal distribution is violated.

15.4 General Tips for Assessing Assumptions

Each of the methods presented here are qualitative assessements, which means they are subjective. That is okay. As the analyst, it is up to you to determine which assumptions you are willing to make. You need to determine whether you feel the data is consistent with the assumptions. Here are two overall things to keep in mind.

First, do not spend too much time examining residual plots. If you stare at a plot too long, you can convince yourself there is pattern in anything. We are looking for glaring evidence that the data is not consistent with the conditions we have imposed on our model. This is especially true when we have only a few observations. In these settings, reading plots can be very difficult. Again, it is about what you are comfortable assuming; how much faith do you want to place in the results?

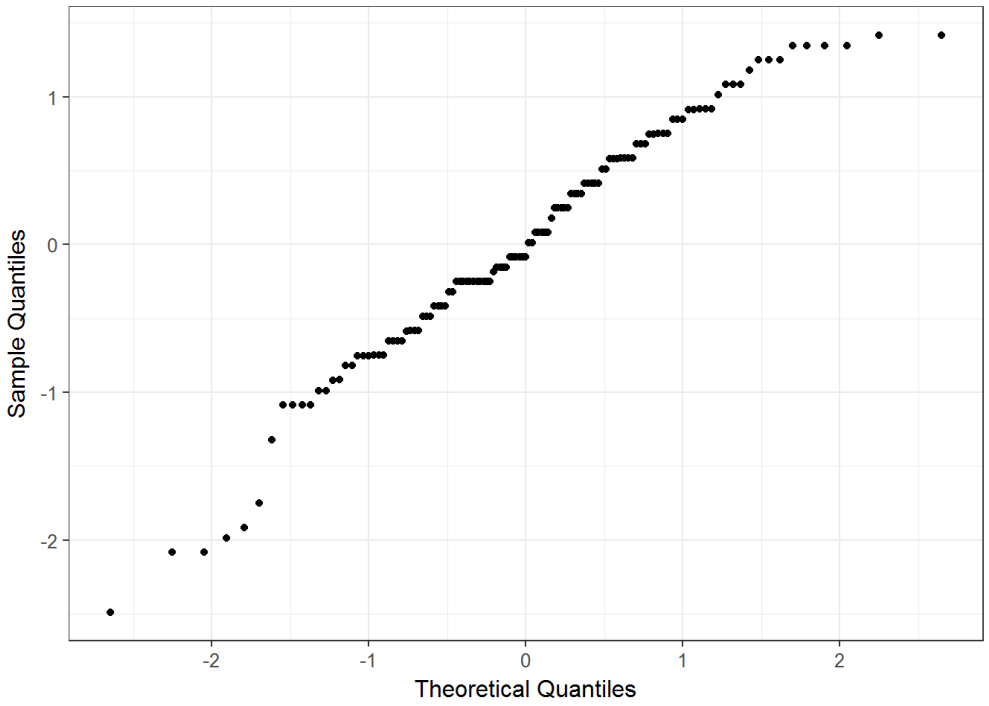


Figure 15.4: Probability plot of the residuals for the Organic Food Case Study. If the errors follow a Normal distribution, we would expect the residuals to fall along a straight line.

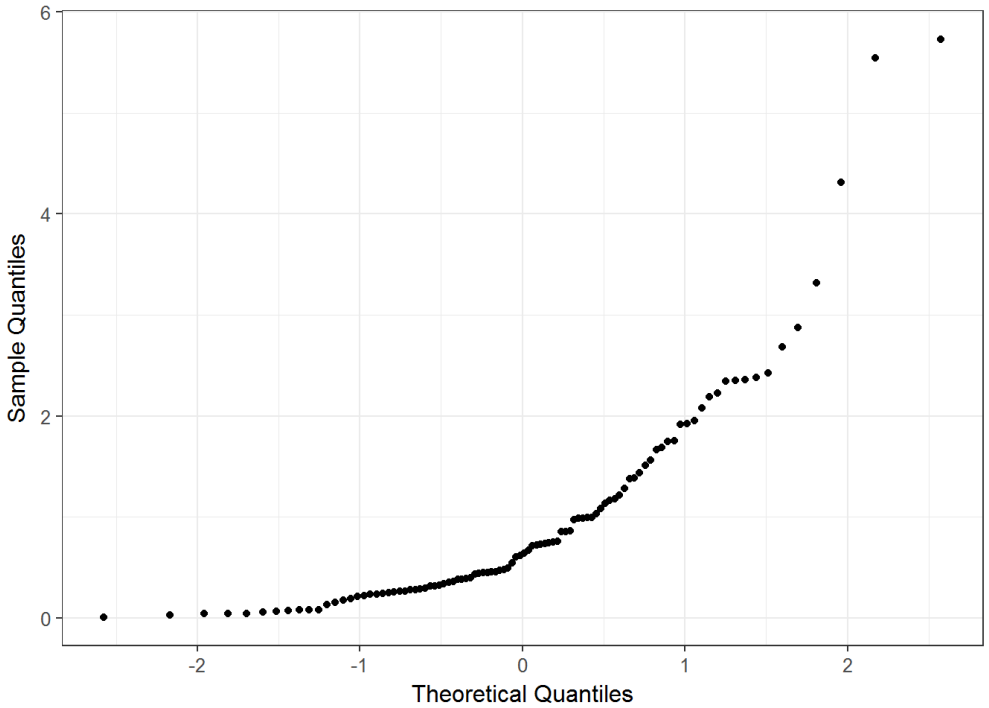


Figure 15.5: Probability plot of residuals for a hypothetical dataset. The trend away from a straight line suggests assuming the errors follow a Normal distribution would be unreasonable.

Second, we have chosen the language carefully throughout this chapter. We have never once stated that a condition was satisfied. When we perform an analysis, we are making an assumption that the conditions are satisfied. We can never prove that they are; we can only show that the data is consistent with a particular condition. We can, however, provide evidence that a condition is violated. When that is the case, we should be wary of trusting the resulting p-values and confidence intervals. This is not unlike hypothesis testing; just as we can never prove the null hypothesis is true, we cannot prove that a condition is satisfied.

Dudeck, A E, and C H Peeacock. 1981. "Effects of Several Overseeded Ryegrasses on Turf Quality, Traffic Tolerance and Ball Roll." In *Proceedings of the Fourth International Turfgrass Research Conference*, edited by R W Sheard, 75–81.

Eskine, Kendall J. 2013. "Wholesome Foods and Wholesome Morals? Organic Foods Reduce Prosocial Behavior and Harshen Moral Judgments." *Social Psychological and Personality Science* 4 (2): 251–54.

Goldstein, Bernard D, Howard J Osofsky, and Maureen Y Lichtveld. 2011. *The Gulf Oil Spill. The New England Journal of Medicine*. Vol. 364. doi:10.1056/NEJMra1007197.

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science* 302: 1338–9.

Lee, J. 1992. "Relationships Between Properties of Pulp-Fibre and Paper."

Moery, Eileen, and Robert J Calin-Jageman. 2016. "Direct and Conceptual Replications of Eskine (2013): Organic Food Exposure Has Little to No Effect on Moral Judgments and Prosocial Behavior." *Social Psychological and Personality Science* 7 (4): 312–19.

Tintle, Nathan, Beth L Chance, A J Rossman, S Roy, T Swanson, and J VanderStoep. 2015. *Introduction to Statistical Investigations*. Wiley.