# Bayesian Data Analysis

Eric M Reyes

Updated: 28 October 2023

# Table of contents

# Preface

Data is all around us. And, that data will be subject to variability; that is, measured characteristics will vary from one observation to the next. Learning to characterize that variability and make decisions in its presence is the idea behind statistics. The text emphasizes statistical literacy (interpretation and clear communication of statistical concepts, methods, and results) and statistical reasoning (defining the need for data to address questions, modeling variability in a process, and choosing the appropriate methodology to address a question of interest).

Specifically, this text introduces the Bayesian framework for statistical inference. Building from Bayes' Rule for probability computations, we develop a framework of estimation and hypothesis testing. We examine inference in several scenarios, including regression analysis. The heart of Bayesian inference is quantifying our beliefs about the data generating process prior to collecting data, and then using the observed data to update those beliefs. We discuss the construction of prior distributions given prior information about a parameter and give an introduction to computational tools for Bayesian inference, including Markov Chain Monte Carlo (MCMC) methods.

While we do work through derivations when introducing the fundamental elements of Bayesian inference, the text is applied. We therefore move quickly to computational approaches for the Bayesian approach. We focus on choosing an appropriate modeling strategy and interpreting the results of an analysis. Our aim is to provide a strong foundation in statistical ideas enabling readers to engage with research encountered in their field.

# Part I

# Unit I: Essential Probability

Probability is the field within mathematics that studies and models random processes. In contrast, Statistics is a discipline separate from mathematics that uses data to make inference on a population. Like many other disciplines (e.g., Engineering and the Sciences), while Statistics is a separate discipline, the theory underlying the discipline relies heavily on mathematics; for Statistics, probability plays a pivotal role. In fact, we once heard an author describe the Bayesian framework as "probability in action." The key components of a Bayesian approach to inference involve characterizing uncertainty and variability. And, Probability can be used to develop analytical models to describe that uncertainty and variability. A firm foundation in probability is necessary to approach inference from a Bayesian perspective.

This unit is not a replacement for a text in Probability; instead, it provides a brief review of key aspects of a Probability course that will be most frequently referenced in the remainder of the text. Our interest is in illustrating how Probability is applied to support statistical methodology. While we assume the reader has taken a course in Probability, we review key results as needed. As this is meant to be used in a Statistics course, our goals are much different than those of a mathematician. Instead of a rigorous treatment of Probability theory (axioms, etc.), our focus is on the application of Probability to Statistics.

# 1 Essential Probability

Probability is a vast field within Mathematics. However, the starting point for nearly every course in probability is the development of essential results (or "probability rules") based on the Axioms of Probability — an agreed upon mathematical framework for describing probability. While we will not make use of these results directly, it is helpful to review them as they lurk in the background of many more useful results.

## 1.1 Probability of an Event

Any process for which the outcome cannot be predicted with certainty is a random process. The collection of all possible results from this random process is known as the **sample space**, and elementary probability is centered on **events** (results of interest) within this sample space.

**Definition 1.1** (Sample Space)**.** The sample space for a random process is the collection of all possible results that we might observe.

**Definition 1.2** (Event)**.** A subset of the sample space that is of particular interest.

The Axioms of Probability are discussed in terms of such events.

**Definition 1.3** (Axioms of Probability)**.** Let $\mathcal{S}$ be the sample space of a random process. Suppose that to each event $A$ within $\mathcal{S}$, a number denoted by $Pr(A)$ is associated with $A$. If the map $Pr(\cdot)$ satisfies the following three axioms, then it is called a **probability**:

1. $Pr(A) \geq 0$
2. $Pr(\mathcal{S}) = 1$
3. If $\{A_1, A_2, \dots\}$ is a sequence of mutually exclusive events in $\mathcal{S}$, then

$$Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} Pr\left(A_i\right).$$

$Pr(A)$ is said to be the "probability of $A$" or the "probability $A$ occurs."

The first axiom states that probabilities cannot be negative. The second states that probabilities cannot exceed 1 and that something must result from a random process. The third states that if two events do not overlap, the probability of the combination of the events is found by adding up the individual probabilities. This third axiom begins to develop an idea of probability as an area. Figure 1.1 illustrates a hypothetical sample space $\mathcal{S}$ with two events $A$ and $B$ of interest. In the figure, the two events share some overlap. Variations of this graphic are used in probability courses to develop intuition for several probability rules. What we emphasize is that we are using the *area* of each event in the figure to represent probability. The applications of probability we will be studying continue to build on this idea of probability as an area.

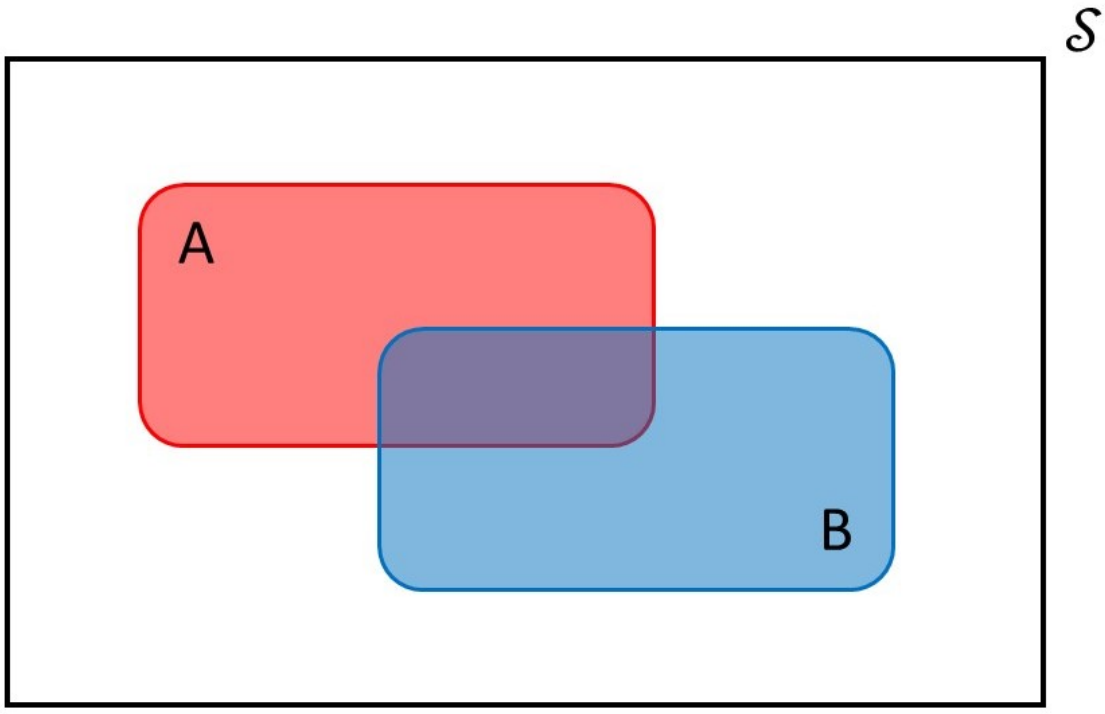


Figure 1.1: Venn-Diagram illustrating two events, $A$ and $B$, within a sample space $\mathcal{S}$.

Big Idea

Probability represents an area.

## 1.2 Essential Results

While the Axioms of Probability (Definition 1.3) set the foundation, we can combine these axioms to form a set of rules which can be employed to describe a myriad of scenarios. The first rule we review states that the probability of an event not occurring is equivalent to subtracting the probability it does occur from 1.

**Theorem 1.1** (Complement Rule)**.** *For any event $A$, the probability of its complement $A^c$ is given by*

$$Pr\left(A^c\right) = 1 - Pr(A).$$

Our interest is not in rigorously developing probability theory; so, we will offer many results without proof. However, to illustrate the connection to the axioms, note that the Complement Rule is a result of the second and third axioms. The second axiom tells us the probability of the sample space is 1, and the third axiom allows us to consider the probability of the union of two mutually exclusive events (which an event and its complement are by definition).

The second rule we consider generalizes the third axiom. The third axiom considers the union of mutually exclusive events, and the Addition Rule defines the probability for the union of arbitrary events.

**Theorem 1.2** (Addition Rule)**.** *Let $A$ and $B$ be arbitrary events, the probability of the union $A \cup B$ is given by*

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

*where $A \cap B$ represents the intersection of the two events.*

A very helpful technique in mathematical proofs is to "do nothing." This technique will be a recurring theme later in the text and manifests itself in adding nothing (adding and subtracting the same quantity to an expression) or multiplying by one (multiplying and dividing an expression by the same quantity).

**Theorem 1.3** (Total Probability Rule)**.** *Let $A$ and $B$ be arbitrary events. Then,*

$$Pr(A) = Pr(A \cap B) + Pr\left(A \cap B^c\right).$$

Though different than the proof you would likely encounter in a Probability text, we provide the proof below because it illustrates the "do nothing" technique that will be helpful later on.

*Proof.* Let $A$ and $B$ be arbitrary events. We note that $A \cap \mathcal{S}$ is the set $A$. And, since the intersection of any set with itself is itself (like multiplying by 1, or "doing nothing" to the set), we have

$$Pr(A) = Pr(A \cap \mathcal{S}).$$

Now, we recognize that an event and its complement together form the sample space; therefore, we can write

$$Pr(A \cap \mathcal{S}) = Pr\left(A \cap (B \cup B^c)\right).$$

Using a distributive law from set theory, we write this probability as

$$Pr\left(A \cap (B \cup B^c)\right) = Pr\left((A \cap B) \cup (A \cap B^c)\right).$$

We now recognize that the events $(A \cap B)$ and $(A \cap B^c)$ are mutually exclusive. Therefore, applying the third axiom of probability, we have that

$$Pr\left((A \cap B) \cup (A \cap B^c)\right) = Pr(A \cap B) + Pr\left(A \cap B^c\right)$$

giving the desired result.

□

We have described probability as an "area," and the above results describe various ways of computing that area. However, occasionally we are given additional information that changes the likelihood of an event. Suppose we are interested in the probability an individual makes a shot from the half-court line on a basketball court. Now, suppose we are told the individual plays for the NBA; our probability should reflect this additional knowledge. This is the idea of "conditional probability."

**Theorem 1.4** (Conditional Probability)**.** *Let $A$ and $B$ be arbitrary events. Then, the probability of $A$ given that $B$ will occur is given by*

$$Pr(A \mid B) = \frac{Pr(A \cap B)}{Pr(B)},$$

*where we read $A \mid B$ as "A given B."*

Conditional probability assumes $Pr(B) > 0$; it would not make sense to condition on an event that will not occur. While there are many other rules that are interesting and useful in application, the above rules suffice for our purposes.

## 1.3 Interpretation of Probability

Again, most probability courses are focused on the mathematics of probability; as a result, rarely is the *interpretation* of probability discussed. In fact, most individuals rarely think about what they mean by "the probability an event occurs." From a mathematical perspective, as long as we obey the Axioms of Probability (Definition 1.3), we have a probability; its meaning is irrelevant. But, for practitioners, the interpretation is critical. As it turns out, there are multiple interpretations of probability[1]. Two interpretations are of particular interest to us. To illustrate, consider the following scenario.

**Example 1.1** (Sugar Packets)**.** Restaurants can be sources of anxiety for small children. After placing their order, they must wait (for what seems like an eternity) for that food to arrive. This is different from their experience at home where they typically are not brought to the table until it is time to eat. Parents spend a lot of effort entertaining their children while waiting for their food to arrive. For parents who do not want to limit screen time, the following simple game is surprisingly effective:

> Take one of the sugar packets that is generally available at the table. Denote the side with the brand name as the "top side" and denote the side with the ingredient list as the "bottom side." The parent then takes the sugar packet and, hidden from view, tumbles the packet randomly in their hands. The packet is then placed on the table under the cover of the parent's hand. The child then declares which side of the packet is facing up by saying "top side" or "bottom side."

This is similar to flipping a coin, but who carries change with them these days? Consider one round of the above game; suppose the (covered) packet has been placed on the table and the child says "top side." The question we ask is then "what is the probability the child is correct?"

This simple example illustrates the two commonly applied interpretations of probability. Most people will say the probability the child is correct is 0.5. The reasoning is that there are two possibilities (the top of the sugar packet is face up; or, the bottom of the sugar packet is face up), and these two possibilities are equally likely (since it was randomly shuffled before being placed on the table). Therefore, the probability the child is correct is 0.5. **From a classical view of statistics, this interpretation is incorrect.** The complication here is what we believe probability is capturing.

From a classical (or "Frequentist") perspective, probability is capturing the likelihood of an event across repeated trials. From a Bayesian perspective, however, probability is used to quantify our uncertainty. That is, from the Bayesian perspective the phrase "the probability the child is correct is 0.5" is not actually quantifying the likelihood the child is correct — it is

[1] See the "Interpretations of Probability" entry in the Stanford Encyclopedia of Philosophy.

quantifying our uncertainty the child is correct. We are only "50% sure" the child is correct. This relies on the "subjective interpretation" of probability.

**Definition 1.4** (Subjective Interpretation of Probability)**.** In this perspective, the probability of $A$ describes the individual's uncertainty about event $A$.

Because the subjective interpretation is quantifying an individual's uncertainty, and since each individual may have different beliefs/information/expertise about the random process, each individual observing the same process may have a different probability. For example, consider asking the question "what is the probability that Netflix saves the latest television series dropped by ABC?" A casual viewer may have little information regarding this process and will rely solely on what they perceive the popularity of the show was among its fan base and news reports they have read online; they may quantify their uncertainty by saying the probability is 0.65. In contrast, an executive at Netflix who is deeply familiar with both the show, its fan base, its ratings in various markets, the interest of leadership to invest in a new series, and the amount they stand to earn by acquiring the property has a different set of knowledge; they may quantify their uncertainty by saying the probability is 0.05. The same process is viewed differently by different observers, leading to different answers.

Statisticians who adhere to the subjective interpretation of probability are known as Bayesians. Classically, statistical theory was developed under the frequentist interpretation, and statisticians who adhere to this perspective are known as Frequentists.

**Definition 1.5** (Frequentist Interpretation of Probability)**.** In this perspective, the probability of $A$ describes the long-run behavior of the event. Specifically, consider repeating the random process $m$ times, and let $f(A)$ represent the number of times the event $A$ occurs out of those $m$ replications. Then,

$$Pr(A) = \lim_{m \to \infty} \frac{f(A)}{m}.$$

The frequentist interpretation requires repeating a process infinitely often. When characterizing the probability of an event, the frequentist perspective leans on the future-oriented nature of probability. When we are characterizing the probability an event *will* occur (future-oriented), we are really thinking about repeating that process infinitely often and determining what fraction of the time the event occurs; we then apply that to the specific process we are about to observe. Of course, this does not always make sense in practice. For example, asking "what is the probability that Candidate A will win the upcoming election" is a one time event. The election cannot be held infinitely often; it will only be held once. In these cases, the frequentist interpretation still *imagines* infinitely many of these elections. For those who are fans of science fiction, you can think of the frequentist perspective as finding the limit over the infinitely many instances in the multiverse (the proportion of times Candidate A wins the election across all instances of the election in the multiverse). The frequentist

perspective is "objective" in the sense that it does not incorporate the observer's personal beliefs/information/expertise regarding the process.

Returning to Example 1.1, since the result has already occurred, probability does not make sense. Further, since the frequentist perspective does not quantify our uncertainty about the result (as the subjective perspective does), we are left saying that the probability that the child is correct is either 1 (they are correct) or 0 (they are not correct). Admittedly, this is unsatisfying, but we must remember that the frequentist interpretation is not interested in quantifying our uncertainty; it is only interested in the proportion of times the result will occur, and since the result is in the past, it either has occurred (proportion of 1) or it has not (proportion of 0).

This may seem like arguing over semantics, and admittedly, the importance of this discussion is not yet clear. But, we will see that how probability is interpreted impacts how we interpret the results of our statistical analyses.

> Big Idea
>
> The frequentist interpretation of probability quantifies the likelihood of an event in repeated observation, and the subjective interpretation quantifies our uncertainty of an event.

As this text focuses on the Bayesian perspective, we adopt the subjective interpretation of probability throughout. Note that this means that two analysts can approach the same problem in the same way and end up with a different conclusion if they have different beliefs!

# 2 Random Variables and Distributions

In Chapter 1, we discussed the probability of an "event." For statisticians, the events of interests center on measurements, or functions of those measurements, that we plan to take. In this chapter, we begin to connect probability to data analysis. Our goal is to reexamine concepts introduced in a probability course, relating them to their data-centric analogues, which will be discussed further in the next unit.

## 2.1 Random Variables

Consider collecting data; before the data is collected, we cannot predict with certainty what we will observe. Therefore, we can think of each observation as the result of a random process. These observations are recorded as variables in our dataset. In probability, a **random variable** is used to represent a measurement that results from a random process.

**Definition 2.1** (Random Variable)**.** Let $\mathcal{S}$ be the sample space corresponding to a random process; a random variable $X$ is a function mapping elements of the sample space to the real line.

Random variables represent a measurement that will be collected during the course of a study. Random variables are typically represented by a capital letter.

While for our purposes, it suffices to think of a random variable as a measurement, mathematically, it is a *function*. The image (or range) of this function is used to broadly classify random variables as **continuous** or **discrete**; we refer to this image as the **support** of the random variable.

**Definition 2.2** (Support)**.** The support of a random variable is the set of all possible values the random variable can take.

**Definition 2.3** (Continuous and Discrete Random Variable)**.** The random variable $X$ is said to be a discrete random variable if its corresponding support is countable. The random variable $X$ is said to be a continuous random variable if the corresponding support is uncountable (such as an interval or a union of intervals on the real line).

Discrete random variables are analogous to categorical (or qualitative) variables in data analysis; that is, discrete random variables are used to model the result of a random process which categorizes each unit of observation into a group. Continuous random variables are analogous to numeric (or quantitative) variables in data analysis; continuous random variables are used to model the result of a random process which produces a number for which arithmetic makes sense.

> **Warning**
>
> Whether we use a continuous or discrete random variable to represent a measurement is not always obvious. Suppose we consider recording the age of a student selected from a class at a university that typically enrolls "traditional" students (those coming directly from high school). Let the random variable $X$ denote the age of the student.
> If we record the student's age in years since birth, $X$ can take on only a finite number of values (most likely $\{18, 19, 20, 21, 22, 23\}$), making it a discrete random variable. However, if we record the student's age as the number of seconds since birth, we might well consider the support of $X$ to be a rather large interval, leading to a continuous random variable.

The goal of statistics is to use a sample to say something about the underlying population. Consider taking a sample of size $n$ and measuring a single variable on each unit of observation. Then, we might represent the measurements we will obtain (note the use of the future tense) as $X_1, X_2, \ldots, X_n$. While the majority of probability courses focus on a single, or maybe two, random variables, note that collecting data on a sample requires that we deal with at least $n$ random variables (one measurement for each of the observations in our sample).

## 2.2 Characterizing a Distribution

Again, the goal of statistics is to use a sample to say something about the underlying population. Consider the following research objective:

> Estimate the cost (in US dollars) of a diamond for sale in the United States.

For this research objective, our population of interest is all diamonds for sale in the United States. We would not expect every diamond for sale to have the same price; variability is inherent in any process. As a result, the sale price of diamonds has a distribution across this population. This is our primary use of probability theory in a statistical analysis — to model distributions.

Consider taking a sample of size 1 from the population; let $Y$ represent the cost of the diamond that is selected. Since we have not yet observed the cost of this diamond, $Y$ is a random variable. And, since this diamond is sampled from the population of interest, the support of $Y$ is determined by the cost of diamonds in the United States. Further, the likelihood that $Y$

falls within any interval is determined by the distribution of the cost across the population. That is, the distribution of $Y$ is the distribution of the population.

> **Big Idea**
>
> If a random variable $X$ represents a measurement for a single observation from a population, the distribution of the random variable corresponds to the distribution of the variable across the population.

A key realization in statistical analysis is that we will never fully observe the distribution of the population; however, we can posit a model for this distribution. In probability, the most common way to characterize the distribution of a random variable is through its density function.

**Definition 2.4** (Density Function)**.** A density function $f$ relates the values in the support of a random variable with the probability of observing those values.

Let $X$ be a continuous random variable, then its density function $f$ is the function such that

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

for any real numbers $a$ and $b$ in the support.

Let $X$ be a discrete random variable, then its density function $f$ is the function such that

$$Pr(X = u) = f(u)$$

for any real number $u$ in the support.

> **Properties of a Density Function**
>
> Let $X$ be a random variable with density function $f$ defined over support $\mathcal{S}$. Then,
>
> 1. $f(x) \geq 0$ for all $x \in \mathcal{S}$. That is, the density is non-negative for all values in the support.
> 2. If $X$ is a continuous random variable, then $\int_{\mathcal{S}} f(x)dx = 1$; similarly, if $X$ is a discrete random variable, then $\sum_{\mathcal{S}} f(x) = 1$. That is, $X$ must take a value in its support; so, $Pr(X \in \mathcal{S}) = 1$, similar to the second Axiom of Probability (Definition 1.3).
> 3. $f(x) = 0$ for all values of $x \notin \mathcal{S}$. The density takes the value of 0 for all values outside the support.

> **Note**
>
> In a probability course, there is often a distinction made between probability "density" functions (used for continuous random variables) and probability "mass" functions (used for discrete random variables). We do not make this distinction and instead rely on the context to determine whether we are dealing with a continuous or discrete random variable. Throughout, we will note when the operations differ between these two types of variables. Measure theory provides a unifying framework to these issues.

When working with a continuous random variable, the density function is a smooth function over some region, and the actual value of the function is not interpretable; instead, we get at a probability by considering the area under the curve. Again, drawing connections to data analysis, we can think of a density function as a mathematical formula representing a smooth histogram. The area under the curve for any region gives the proportion of the population which has a value in that region. That is, we get the probability that a random variable will be in an interval by integrating the density function over that interval.

Figure Figure 2.1 illustrates this idea; we have data from a sample of diamonds from the population of interest. The sample is summarized with a histogram; we have overlayed a posited density (with the corresponding mathematical function that describes this density) for the population. The sample (summarized with the histogram) is approximating the population (modeled using the density function).

You may recognize the particular form of the density function in Figure 2.1. The general form is

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma} \qquad \text{for } x > 0$$

where $\sigma$ is the *scale* parameter that defines the distribution (set at 4000 in Figure 2.1). This is known as the Exponential distribution with scale parameter $\sigma$. This illuminates another connection between probability and statistics.

Note that our research objective describe above is an ill-posed question as stated. The answer is "it depends" since each individual diamond in the population has a different value. Well-posed questions in statistics are centered on an appropriately chosen **parameter** (Definition 5.6).

In probability, the parameters are values that are tuned or set within a problem; we then work forward to compute the probability of an event of interest. In practice, however, when we posit a functional form for a density function to describe the distribution of the population, the parameters are unknown. We plan to use the data to estimate or characterize the parameter; but, the parameter itself will remain unknown. In both cases, however, the parameter is a *fixed quantity*, even if we are ignorant of that value.

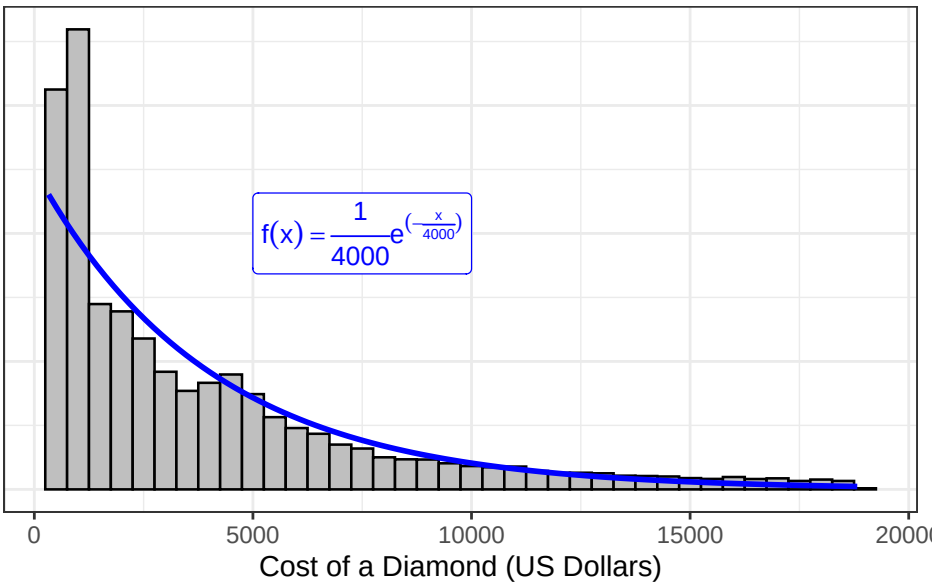


Figure 2.1: Illustration of a density function representing the posited distribution of the population alongside a histogram summarizing the cost of diamonds using a sample of 53940 diamonds.

Big Idea

When a probability model is specified for a population, it is generally specified up to some unknown parameter(s). Making inference on the unknown parameter(s) therefore characterizes the entire distribution.

### 2.2.1 Common Parameters

Most scientific questions are focused on the location or spread of a distribution. For example, we are interested in estimating the average cost of a diamond sold in the United States. Introductory statistics introduces summaries of location and spread within the sample (e.g., sample mean for location and sample variance for spread). Analogous summaries exist for density functions. As stated above, parameters are unknown constants that govern the form of the density function. Because they govern the form of the density, the parameters are also related to those summarizing the location or spread of the distribution.

**Definition 2.5** (Expected Value (Mean))**.** Let $X$ be a random variable with density function $f$ defined over the support $\mathcal{S}$. The expected value of a random variable, also called the mean and denoted $E(X)$, is given by

$$E(X) = \int_{\mathcal{S}} x f(x) dx$$

for continuous random variables and

$$E(X) = \sum_{\mathcal{S}} x f(x)$$

for discrete random variables.

Notice the similarity between the form of the sample mean and the population mean. A sample mean takes the sum of each value in the sample, weighting each value by $1/n$ (where $n$ is the sample size). Without information about the underlying population, the sample must treat each value observed as equally likely; values become more likely if they appear multiple times. In the population, however, when the form of $f$ is known, the density provides information about the likelihood of each value giving us a better weight than $1/n$. That is, the population mean is a sum of the values in the support, weighting each value by the corresponding value of the density function.

**Definition 2.6** (Variance)**.** Let $X$ be a random variable with density function $f$ defined over the support $\mathcal{S}$. The variance of a random variable, denoted $Var(X)$, is given by

$$Var(X) = E\left[X - E(X)\right]^2 = E\left(X^2\right) - E^2(X).$$

If we let $\mu = E(X)$, then this is equivalent to

$$\int_{\mathcal{S}} (x - \mu)^2 f(x) dx$$

for continuous random variables and

$$\sum_{\mathcal{S}} (x - \mu)^2 f(x)$$

for discrete random variables.

> **Warning**
>
> Pay careful attention to the notation. $E^2(X)$ represents the square of the expected value; that is,

$$E^2(X) = [E(X)]^2 .$$

However, $E(X)^2$ represents the expected value of the square of $X$; that is,

$$E(X)^2 = E\left(X^2\right).$$

The variance provides a measure of spread; in particular, it is capturing distance from the mean. Notice that the form of the variance involves taking the expectation of a squared term; in general, we will need to consider expectations of functions.

**Definition 2.7** (Expectation of a Function)**.** Let $X$ be a random variable with density function $f$ over the support $\mathcal{S}$, and let $g$ be a real-valued function. Then,

$$E\left[g(X)\right] = \int_{\mathcal{S}} g(x)f(x)dx$$

for continuous random variables and

$$E\left[g(X)\right] = \sum_{\mathcal{S}} g(x)f(x)$$

for discrete random variables.

> **Note**
>
> Definition 2.7 is sometimes referred to as the "Law of the Unconscious Statistician" in probability texts. We find the name somewhat insulting, as it suggests statisticians do not appreciate the mathematical underpinnings of their field. In reality, the expected value of a function is such a common operation in statistical theory that statisticians often present Definition 2.7 as the definition of an expectation (as we have done) instead of deriving it as a result of Definition 2.5 after applying a variable transformation (see Example 2.3 below). It is possible this is where the slight in the naming convention originated.

A result of Definition 2.7 is the following, very useful theorem, which states that expectations are linear operators.

**Theorem 2.1** (Expectation of a Linear Combination)**.** *Let $X$ be a random variable, and let $a_1, a_2, \ldots, a_m$ be real-valued constants and $g_1, g_2, \ldots, g_m$ be real-valued functions; then,*

$$E\left[\sum_{i=1}^{m} a_i g_i(X)\right] = \sum_{i=1}^{m} a_i E\left[g_i(X)\right].$$

The mean and variance play an important role in characterizing a distribution, especially within statistical theory (as we will see in future chapters). However, there is another set of parameters which are important.

**Definition 2.8** (Percentile for a Random Variable)**.** Let $X$ be a random variable with density function $f$. The $100k$ percentile is the value $q$ such that

$$Pr(X \leq q) = k.$$

**Example 2.1** (Parameters of Exponential Distribution)**.** Let $X$ be an Exponential distribution with scale parameter $\sigma$; that is, the density function $f$ is given by

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma} \qquad x > 0$$

where $\sigma > 0$. Compute the mean, variance, and median of this distribution, as a function of the unknown scale parameter.

The solution to this problem is particularly important as it illustrates a very useful technique when working with known distributions in statistical theory.

*Solution.* We note that the function

$$g(y) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta}$$

is a valid density function over the positive real line provided that $\alpha, \beta > 0$; in particular, this is known as a Gamma distribution. Since $g$ is a valid density function, then we know that

$$\int_0^{\infty} g(y)dy = 1$$

for all values of $\alpha, \beta > 0$.

Now, let $X$ be an Exponential random random variable with scale parameter $\sigma$. Then, the expected value of $X$ is given by

$$\begin{aligned} E(X) &= \int_0^{\infty} x \frac{1}{\sigma} e^{-x/\sigma} dx \\ &= \int_0^{\infty} \frac{1}{\sigma} x^{2-1} e^{-x/\sigma} dx \end{aligned}$$

where we have simply rewritten the exponent in the second line. Notice that expression within the integral shares a striking similarity to the form of the density function of a Gamma distribution; however, they are not exactly the same. To coerce the expression into that of the Gamma density function, we "do nothing" — multiplying and dividing the expression by the quantity $\sigma\Gamma(2)$. This gives

$$\begin{aligned}
E(X) &= \int_0^\infty x\frac{1}{\sigma}e^{-x/\sigma}dx \\
&= \int_0^\infty \frac{1}{\sigma}x^{2-1}e^{-x/\sigma}dx \\
&= \int_0^\infty \sigma\Gamma(2)\frac{1}{\sigma^2\Gamma(2)}x^{2-1}e^{-x/\sigma}dx \\
&= \sigma\Gamma(2)\int_0^\infty \frac{1}{\sigma^2\Gamma(2)}x^{2-1}e^{-x/\sigma}dx \\
&= \sigma\Gamma(2) \\
&= \sigma.
\end{aligned}$$

In line 3, we have multiplied and divided by $\sigma\Gamma(2)$, which does not change the problem. In line 4, we have pulled out the terms $\sigma\Gamma(2)$ since it is a constant with respect to the integral; what is left inside the integral is the form of the density function for a Gamma distribution where $\alpha = 2$ and $\beta = \sigma$. In line 5, we make use of the fact that the integral of any density function over the entire support for which it is defined must be 1. Finally, in line 6, we recognize that $\Gamma(k) = (k-1)!$ if $k$ is a natural number.

Applying the same process, we also have that

$$\begin{aligned}
E\left(X^2\right) &= \int_0^\infty x^2\frac{1}{\sigma}e^{-x/\sigma}dx \\
&= \sigma^2\Gamma(3)\int_0^\infty \frac{1}{\sigma^3\Gamma(3)}x^{3-1}e^{-x/\sigma}dx \\
&= 2\sigma^2.
\end{aligned}$$

Therefore,

$$Var(X) = E\left(X^2\right) - E^2(X) = 2\sigma^2 - \sigma^2 = \sigma^2.$$

Finally, the median is the value $q$ such that $Pr(X \leq q) = 0.5$; but, we recognize that

$$\begin{aligned} Pr(X \leq q) &= \int_0^q \frac{1}{\sigma} e^{-x/\sigma} dx \\ &= -e^{-x/\sigma}\Big|_0^q \\ &= -e^{-q/\sigma} + 1. \end{aligned}$$

Setting this expression equal to 0.5 and solving for $q$ yields $q = -\sigma \log(0.5)$, where $\log(\cdot)$ represents the *natural* logarithm.

> **Big Idea**
>
> Suppose the density $f$ is a function of the parameters $\theta$; then, the mean, variance, and median (as well as any other parameters of interest in a research objective) will be functions of $\theta$.

Example 2.1 highlighted a useful technique for simplifying integrals in statistical applications, which makes use of the "do nothing" strategy discussed in the previous chapter.

### 2.2.2 Kernels

One of the characteristics common to any density function that we noted above was that if we sum the density across the entire support, we get a value of 1. That is, if $X$ is a discrete random variable, then

$$\sum_{x \in \mathcal{S}_X} f(x) = 1,$$

and if $X$ is a continuous random variable, then

$$\int_{\mathcal{S}_X} f(x) dx = 1.$$

Any density function can be written as

$$f(x) = ak(x)$$

where $a > 0$ is a constant and $k(x)$ is a function of $x$. Specifically, if $X$ is a continuous random variable, then

$$a = \frac{1}{\int k(x) dx}$$

since $\int f(x)dx = 1$. We call $k(x)$ the **kernel** of the distribution. Kernels are helpful for quickly identifying distributions[1].

**Definition 2.9** (Kernel of a Distribution)**.** Let $k(x)$ be a non-negative function of $x$ over some region $\mathcal{S}_X$. Then, a valid density function $f$ over the support $\mathcal{S}_X$ can be constructed by taking

$$f(x) = ak(x)$$

where $a > 0$ is a suitably chosen scaling constant to ensure the density integrates (or sums) to 1 over the support. The function $k$ is known as the kernel of the distribution, and it can be used to identify the distributional family for a random variable.

**Example 2.2** (Kernel of an Exponential Random Variable)**.** In Example 2.1, we let $X$ be an Exponential random variable with scale parameter $\sigma$. Identify the kernel of this distribution.

*Solution.* Let $a = \sigma^{-1}$ and

$$k(x) = e^{-x/\sigma};$$

then, we have that $f(x) = ak(x)$. We note that $k(x)$ has no leading constants; therefore, the kernel for an Exponential distribution is

$$e^{-x/\sigma}.$$

As Example 2.1 illustrated, being able to identify a kernel can help us quickly evaluate an integral. In particular, notice that we immediately have that

$$\int e^{-x/\sigma}dx = \sigma$$

for any value of $\sigma > 0$ because we know that

$$\begin{aligned}\int e^{-x/\sigma}dx &= \sigma \int \frac{1}{\sigma}e^{-x/\sigma}dx \\ &= \sigma.\end{aligned}$$

The first line multiplies and divides by the appropriate scaling term so that the kernel becomes a valid density function. Once we have a valid density, we know it integrates to 1, simplifying the expression.

[1] A good table of common distributions is given in Casella and Berger, a popular text for statistical theory at the graduate level.

In addition to motivating the use of kernels in integration applications, the solution to Example 2.1 also shows that there is more than one way to characterize a distribution.

### 2.2.3 Distribution Function

Especially for visualization, the density function is the most common way of characterizing a probability model. However, computing the probability using the density is problematic due to the integration required. Many software address this by working with the cumulative distribution function (CDF).

**Definition 2.10** (Cumulative Distribution Function (CDF))**.** Let $X$ be a random variable; the cumulative distribution function (CDF) is defined as

$$F(u) = Pr(X \leq u).$$

For a continuous random variable, we have that

$$F(u) = \int_{-\infty}^{u} f(x)dx$$

implying that the density function is the derivative of the CDF. For a discrete random variable

$$F(u) = \sum_{x \leq u} f(x).$$

Working with the CDF improves computation because it avoids the need to integrate each time; instead, the integral is computed once (and stored internally in the computer) and we use the result to compute probabilities directly.

> Big Idea
>
> Density functions are the mathematical models for distributions; they link values of the variable with the likelihood of occurrence. However, for computational reasons, we often work with the cumulative distribution function which provides the probability of being less than or equal to a value.

## 2.3 Transformations of a Random Variable

Occasionally, we are interested in a transformation of a particular characteristic. That is, we have a model for the distribution of $X$, but we are interested in $Y = g(X)$. In this section, we examine one method for determining the density of $Y$ from the density of $X$. While relationships between many common distributions have been well studied[2], it is useful to know the process for addressing transformations.

While there various approaches to this problem, we find this method the most reliable. Further, it does not require the memorization of a formula, but instead builds on fundamental ideals. This is known as the **Method of Distribution Functions**.

**Definition 2.11** (Method of Distribution Functions)**.** Let $X$ be a continuous random variable with density $f$ and cumulative distribution function $F$. Consider $Y = h(X)$. The following process provides the density function $g$ of $Y$ by first finding its cumulative distribution function $G$.

1. Find the set $A$ for which $h(X) \leq t$ if and only if $X \in A$.
2. Recognize that $G(y) = Pr(Y \leq y) = Pr\left(h(X) \leq y\right) = Pr(X \in A)$.
3. If interested in $g(y)$, note that $g(y) = \frac{\partial}{\partial y} G(y)$.

When $h$ is a strictly monotone function (unique inverse exists), then step 1-2 is much easier because we can apply $h^{-1}$. In step 2 of the above process, the final expression is often left in terms of $F$, the CDF of $X$; then, when we find the density in step 3, we can apply the chain rule (avoiding the need to actually have an expression for $F$).

**Example 2.3** (Transformation of a Random Variable)**.** Previously, we posited the following model for the distribution of the cost of a diamond sold in the US:

$$f(x) = \frac{1}{\sigma} e^{-x/\sigma} \qquad x > 0$$

for some $\sigma > 0$. As cost is generally a heavily skewed variable, we may be interested in taking the (natural) logarithm before proceeding with an analysis. Find the density of $Y = \log(X)$; then, write an expression for $E(Y)$.

*Solution.* We note that $\log(x)$ is a strictly monotone function. Therefore, we have that

$$\begin{aligned} G(y) &= Pr(Y \leq y) \\ &= Pr(\log(X) \leq y) \\ &= Pr\left(X \leq e^y\right). \end{aligned}$$

[2] An excellent summary of the relationships between Distributions was developed by faculty at the College of William and Mary.

Just to place this within the method described above, since $\log(x) \leq y$ if and only if $x \leq e^y$, then $A = \{t : x \leq e^t\}$. Of course, we didn't really need to identify this because we were able to apply the inverse of $\log(x)$ directly within the probability expression. We now recognize that we have a probability of the form "$X$ less than or equal to something." And, this matches the form of the CDF of $X$. That is, we have that

$$G(y) = F\left(e^y\right).$$

This completes step 2 of the procedure; we have expressed the CDF of $Y$ as a function of the CDF of $X$. Now, to find the density, we apply the chain rule.

$$\begin{aligned} g(y) &= \frac{\partial}{\partial y} G(y) \\ &= \left[ \frac{\partial}{\partial x} F(x) \Big|_{x=e^y} \right] \frac{\partial}{\partial y} e^y \\ &= \left[ f(x)|_{x=e^y} \right] e^y \\ &= f\left(e^y\right) e^y \\ &= \frac{1}{\sigma} e^{-e^y/\sigma} e^y \end{aligned}$$

which will be valid for all real values of $y$; that is, the support of $Y$ is all real numbers. In line 2 above, we applied the chain rule to compute the derivative, avoiding the need to explicitly state the CDF of $X$.

Given the density of $Y$, we know (by Definition 2.5) that

$$E(Y) = \int y g(y) dy = \int y \frac{1}{\sigma} e^{-e^y/\sigma} e^y dy.$$

Letting $u = e^y$ (and therefore $du = e^y dy$) and performing a u-substitution, we have that

$$\begin{aligned} E(Y) &= \int y \frac{1}{\sigma} e^{-e^y/\sigma} e^y dy \\ &= \int \log(u) \frac{1}{\sigma} e^{-u/\sigma} du. \end{aligned}$$

Since the variable of integration is arbitrary, we recognize that this integral as what we defined (in Definition 2.7) as $E\left[\log(X)\right]$ where $X$ has density $f(x)$ defined in Example 2.3.

Warning

While mathematicians distinguish between a derivative $\frac{d}{dx}$ and a partial derivative $\frac{\partial}{\partial x}$, we do not make that distinction.

# Part II

# Unit II: Language of Data

Children learn the alphabet before tackling *The Odyssey*. Musicians become proficient in scales before playing in a symphony. And, chefs create world-class culinary experiences because they are experts at working with their ingredients. Similarly, working with statistical models benefits from understanding the language of data.

This first unit introduces the key components of any analysis — asking well-posed questions, collecting useful data, summarizing your data to tell a story. Once we are familiar with these ingredients, we can begin putting them together to address a range of interesting questions.

# 3 The Statistical Process

Is driving while texting as dangerous as driving while intoxicated? Is there evidence that my measurement device is calibrated inappropriately? How much force, on average, can our concrete blocks withstand before failing? 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.

> Big 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 3.1** (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 a person'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 study found that 79% of those in the neutral group, 42% of those in the opt-in group, and 82.0% of those in the opt-out group agreed to become donors.

The results of the study are presented in Figure 3.1. 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 claim the "opt-out" strategy is really superior compared to the "neutral" strategy in the broader population? The discipline of statistics provides a framework for addressing such ambiguity.

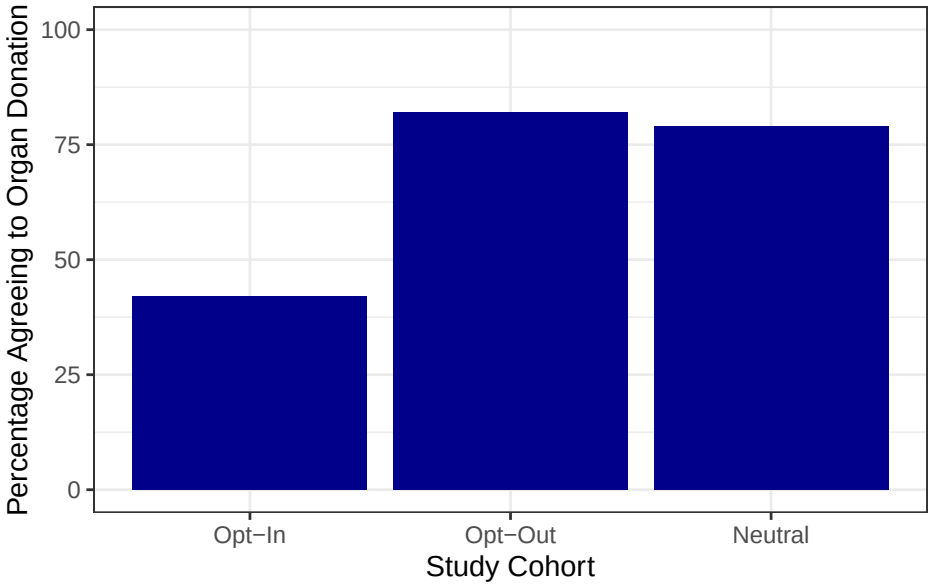


Figure 3.1: Summary of the responses for the Organ Donation Study described in Example 3.1.

## 3.1 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 3.1** (Population)**.** The collection of subjects we would like to say something about.

In the Organ Donation study (Example 3.1), 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 *all Americans who are of the age to consent to organ donation.*

In general, the subjects (or units of observation) in a population need not be people; in some studies, the population could be a collection of screws, cell phones, 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 3.2** (Sample)**.** The collection of subjects for which we actually obtain measurements (data).

> **Note**
>
> Some readers may associate "subjects" with people; to avoid this confusion, you may prefer "unit of observation" to subject. In this text, we use subject to mean any unit on which observations could be taken.

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** (illustrated in Figure 3.2).

**Definition 3.3** (Statistical Inference)**.** The process of using a sample to characterize some aspect of the underlying population.

## 3.2 Anatomy of a Dataset

Once we have our sample, we take measurements on each of the subjects within this sample. 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 social media 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 3.4** (Variable)**.** A measurement, or category, describing some aspect of the subject.

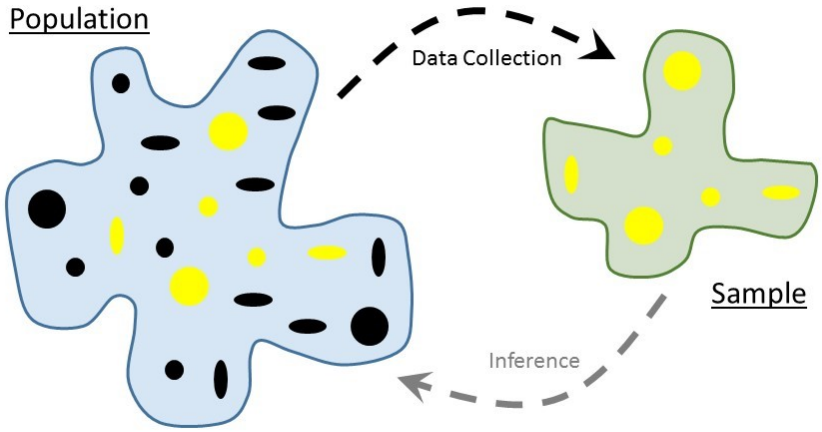


Figure 3.2: Illustration of the statistical process, using a sample to characterize some aspect of the underlying population.

Variables come in one of two flavors. **Categorical** variables are those which denote a grouping to which the subject belongs. Examples include marital status, manufacturer, and experimental treatment group. **Numeric** variables are those which take on values for which ordinary arithmetic (e.g., addition and 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 3.5** (Categorical Variable)**.** Also called a "qualitative variable," a measurement on a subject which denotes a grouping or categorization.

**Definition 3.6** (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.

Characteristics of Well-Structured Data

A well-structured dataset should adhere to the following characteristics:

- Each column contains a unique variable.

Table 3.1: Example of a common data structure which does not correspond to the characteristics of well-structured data we recommend. The 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 |

- 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 characteristics ensure the data is properly formatted for an analysis. Even unstructured data such as images or text files must be processed prior to performing a statistical analysis.

**Warning**

We note the above description eliminates a common method of storing data in engineering and scientific disciplines — storing each sample in a different column.

To illustrate the above description, 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 dataset like that in Table 3.1; the problem here is that the first record of the dataset contains information on two different units of observation. We have the lifetime from a battery of Brand A in the same row as the lifetime from a battery of Brand B. This violates the second characteristic of datasets described above.

In order to adhere to the characteristics of well-structured data outlined above, we can reformat the data in Table 3.1 to that shown in Table 3.2. 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.

Table 3.2: Example of a well-structured dataset. The 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 |

## 3.3 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 3.7** (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 alongside appropriate units for the measurements. Other codebooks give only an indication of what each 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.

# 4 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 screenshot 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, we might 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 oil). An article appearing in *The New England Journal of Medicine* (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 motivating example 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*

[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 4.1: *New York Times* coverage of the *Deepwater Horizon* oil spill.

*Fundamental Ideas of Inference*. We will also see that any statistical analysis moves 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.

# 5 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 (Chapter 4). It might seem natural to ask "if a volunteer cleans wildlife, will she develop adverse respiratory symptoms?" Let's consider the data. Of the 54 volunteers assigned to wildlife cleaning and rehabilitation, 15 reported experiencing adverse respiratory symptoms ("nose irritation, sinus problems, or sore throat"); while some volunteers developed symptoms, others did not. It seems the answer to our question is then "it depends" or "maybe." This is an example of an *ill-posed question*. Such questions exist because of **variability**, the fact that every subject in the population does not behave in exactly the same way; that is, the value of a variable potentially differs from one observation to the next. In our example not every volunteer had the same reaction when directly exposed to oil.

It is variability that creates a need for the discipline of 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 5.1** (Variability)**.** The notion that measurements differ from one observation to another.

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

## 5.1 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 about 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 probably want to say something about volunteers for any oil spill. The 54 volunteers in our dataset form the sample, a subset from

all volunteers who clean wildlife following an oil spill. Our population of interest is comprised of all volunteers who clean wildlife following an oil spill.

> **Note**
>
> When identifying the population of interest for a research question you have, be specific! Suppose you are trying to estimate the average height of trees. Are you really interested in *all* trees? 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 5.2** (Response)**.** The primary variable of interest within a study. This is the variable you would either like to explain or estimate.

**Definition 5.3** (Distribution)**.** The pattern of variability corresponding to a set of values.

Notice that in this case, the response is a categorical variable; describing the distribution of such a variable is equivalent to describing how individuals are divided among the possible groups. With a finite number of observations, we could present the number of observations, the **frequency**, 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 following an oil spill), reporting the frequencies is not appropriate. In this case, we report the fraction of observations, the **relative frequency**, falling within each group; this helps convey information about the distribution of this variable. That is, the relative frequencies give us a sense of which values of the variable are more or less common in the sample.

**Definition 5.4** (Frequency)**.** The number of observations in a sample falling into a particular group (level) defined by a categorical variable.

**Definition 5.5** (Relative Frequency)**.** Also called the "proportion," the fraction of observations falling into a particular group (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 5.6** (Parameter)**.** Numeric quantity which summarizes the distribution of a variable within the *population* of interest. Generally denoted by Greek letters in statistical formulas.

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. For example, the grade received on an exam varies from one student to another in a class; but, the *average exam grade* is a fixed number which summarizes the class as a whole. 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 generally fall into one of two categories:

- Estimation: what *proportion* of volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms?
- Hypothesis Testing: is it reasonable no more than 1 in 5 volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms; or, is there evidence more than 1 in 5 volunteers who clean wildlife following an oil spill will experience adverse respiratory symptoms?

**Definition 5.7** (Estimation)**.** Using the sample to approximate the value of a parameter from the underlying population.

**Definition 5.8** (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.

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 experience symptoms. However, we can determine what the data suggests about the population (that is what inference is all about).

Big Idea

Parameters are unknown values and can never, in general, be known.

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

Fundamental Idea I

A research question can often be framed in terms of a parameter that characterizes the population. Framing the question should then guide our analysis.

We now have a way of describing a well-posed question, a question which can be addressed using data. Well posed questions are about the population and can be framed in terms of a parameter which summarizes that population. We now describe how these questions are typically framed.

## 5.2 Framing the Question

In engineering and scientific applications, many questions fall under the second category of **hypothesis testing**, which is a form of model comparison in which 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. Present evidence to establish guilt, to the contrary of innocence (prosecution's responsibility).
3. Consider the weight of the evidence presented (jury's responsibility).
4. Make a decision. If the evidence is "beyond a reasonable doubt," the jury declares the defendant guilty; otherwise, the jury declares the defendant not guilty.

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

1. Assume the opposite of what we want the data to show (develop a working theory).
2. Gather data and compare it to the proposed model from step (1).
3. Quantify the likelihood of our data from step (2) 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 is one of several subtle points in hypothesis testing. 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*.

> **Note**
>
> This process may seem counter-intuitive; it is natural to ask "why can't we prove guilt directly?" However, when you disprove one statement, you are proving that statement's opposite — a technique known in mathematics as "proof by contradiction." So, our approach to proving a statement is to disprove all other possibilities. It is similar to the technique of the fictional detective Sherlock Holmes (Doyle 1890, pg. 92): "Eliminate all other factors, and the one which remains must be the truth."

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 statement that we would like to disprove is known as the **null hypothesis**; the opposite of this statement, called the **alternative hypothesis**, captures what we as the researchers would like to establish.

**Definition 5.9** (Null Hypothesis)**.** The statement (or theory) about the parameter that we would like to *disprove*. This is denoted $H_0$, read "H-naught" or "H-zero".

**Definition 5.10** (Alternative Hypothesis)**.** The statement (or theory) about the parameter 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 clean wildlife following an oil spill who experience adverse respiratory symptoms is no more than 0.20.
> $H_1$ : The proportion of volunteers assigned to clean 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.

> **Note**
>
> When framing your questions, be sure your null hypothesis and alternative hypothesis are exact opposites of one another, and ensure the "equality" component *always* goes in the null hypothesis.

We can now collect data and 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").

**Consistent vs. Evidence**

The term "consistent" and "reasonable" will be used interchangeably throughout the text; however, these terms differ substantially from the term "evidence," particularly in the Frequentist perspective The data is said to be consistent with a statement if the data is aligned with that statement. In general, evidence is a stronger statement.

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 $\theta$ represent the proportion of volunteers (assigned to clean 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, $\theta$ represents the parameter of interest; the value 0.20 is known as the **null value**.

**Definition 5.11** (Null Value)**.** The value associated with the equality component of the null hypothesis; it forms the threshold or boundary between the hypotheses. Note: not all questions of interest require a null value be specified.

**Big 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.

This section has focused on developing the null and alternative hypothesis when our question of interest is best characterized as one of comparing models or evaluating a particular statement. If our goal is estimation, a null and alternative hypothesis are not applicable. For example, we might have the following goal:

> Estimate the proportion of volunteers (assigned to clean wildlife following an oil spill) who experience adverse respiratory symptoms.

In this version of our research "question" there is no statement which needs to be evaluated. We are interested in estimation, not hypothesis testing and thus there is no corresponding null and alternative hypothesis.

**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 hypothesis the exact opposite statement (capturing what you want to provide evidence for).

# 6 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 (see Figure 6.1). This generally provokes unease in people, leading to a distrust of statistical results. In this section we attack that distrust head on.

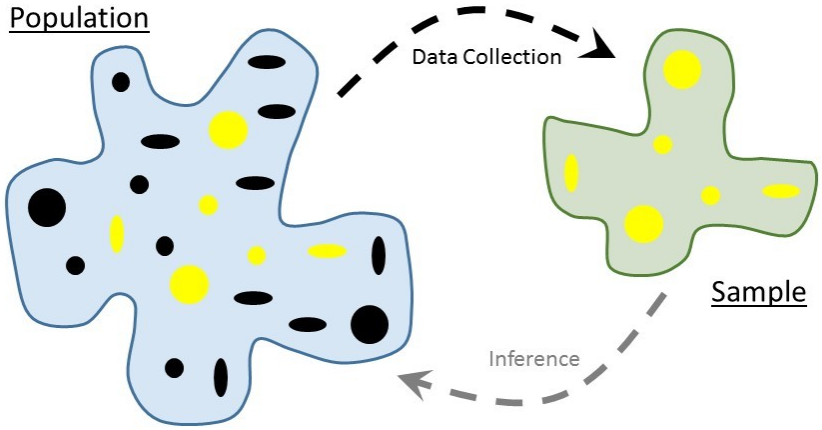


Figure 6.1: Illustration of the statistical process (reprint of Figure 3.2).

## 6.1 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 6.2) is a canvas painting depicting a pipe, below which the artist wrote the French phrase "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!


Figure 6.2: *The Treachery of Images* by René Magritte.

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

> **Big Idea**
>
> In order for a statistical analysis to be reliable, the sample must be *representative* of the population under study.

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 their depiction based on how they want to represent the object. Let's consider the following example.

**Example 6.1** (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

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

engineering school. There are many faculty at Rose-Hulman who choose to send their children to the institute. Suppose we were to ask 25 such faculty members who have a child that attended the institute to report the amount of student loans their children carried upon graduation from Rose-Hulman. Further, suppose we compile the responses and compute the average amount of debt. Using the data, we might report that based on our study, there is significant evidence 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 should we be hesitant to trust the results from our study?

Many objections to statistical results stem from a distrust of whether the data (the sample) is really representative of the population of interest. 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. There is a mismatch between the group we would like to study and the data we have collected.

This example 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 from Rose-Hulman. 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 of Rose-Hulman. However, the sample consists of only graduating seniors of Rose-Hulman *who have a parent employed by the institute*.

With regard to grade point average, the students in our sample are probably similar to all graduating seniors; the starting salary of the students in our sample is probably similar to all graduating seniors; 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 from the university.

> **Note**
>
> When thinking about whether a sample is representative, focus your attention to the characteristics specific to your research question or with regard to how you intend to generalize the results.

Does that mean the sample we collected in Example 6.1 is useless? Yes and no. The sample collected cannot be used to answer our initial question of interest since it is not representative of our population. 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. However, while the sample cannot be used to answer our initial question, it could be used to address a different question:

> 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 institute?

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 of our target population. Let's first look at what *not* to do.

## 6.2 Poor Methods of Data Collection

Example 6.1 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 students 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, which is what increased the ease of collection, 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 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 "middle of the road" opinions.

We could not possibly name all the poor methods for collecting a sample; but, poor 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 6.1** (Bias)**.** A set of measurements is said to be biased if they are *consistently* too high (or too low). Similarly, an estimate of a parameter is said to be biased if it is *consistently* too high (or too low).

To illustrate the concept of bias, consider shooting at a target as in Figure 6.3. We can consider the center of our target to be the parameter we would like to estimate within the population;

in this case, some measure of center. 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 Chapter 7) but about *consistently* shooting in the wrong direction.

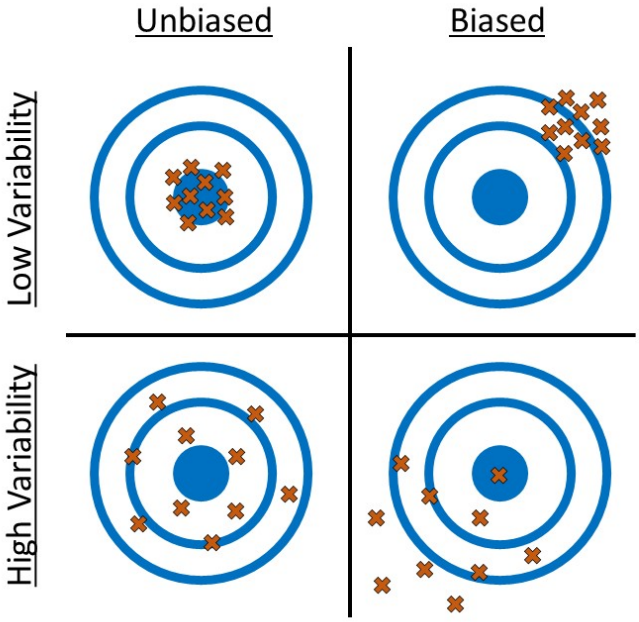


Figure 6.3: Illustration of bias and precision.

**Accuracy vs. Precision**

There is a difference between *accuracy* and *precision*. Generally, *accuracy* refers to location (and therefore relates to bias); we say a process is accurate when it is unbiased. *Precision* refers to the variability; data which is more precise has less variability.

**Big Idea**

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

The catch (there is always a catch) is that we will never *know* with certainty if a sample is actually representative or not. In practice, we critically examine the method in which the sample was collected, and we use summaries of the sample to make educated decisions on whether to generalize the results. Better, however, is to employ methods of data collection that help to minimize the bias in the sample.

## 6.3 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 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 (Example 6.1) again. Suppose instead of the data collection strategy described there, we had done the following:

> We constructed a list of all graduating seniors from the institute. We placed the name of each student on an index card; then, we thoroughly shuffled the cards and chose the top 25 cards. For these 25 individuals, we recorded the amount of debt in student loans each carried.

This essentially describes using a lottery to select the sample. This popular method is known as taking 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 6.2** (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.

> **i** Note
>
> It is convention to use $n$ to represent the sample size.

The primary benefit of a simple random sample is that it removes bias. More specifically, the *process* of simple random sampling is unbiased; that is, this process does *not* produce values which are *consistently* too high or low.

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, by chance alone, that in a simple random sample of 25 students none will be international. Instead of a simple random sample, we might consider taking a sample of 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 6.3** (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.

> ⚠ Warning
>
> Note that a stratified random sample essentially results in a representative sample *within* each strata. However, the combined sample may not be representative of the population. If there is interest in using the sample in its entirety, instead of comparing the strata in some way, advanced statistical methodology is required. See texts on analyzing "complex survey design" for a more thorough discussion. Our text will not consider such cases.

There are countless sampling techniques used in practice. The two described above can be very useful starting points for developing a custom method suitable for a particular application. Their benefit stems from their use of randomization as it limits researcher influence on the composition of the sample and therefore minimizes bias.

This section is entitled "Preferred Methods" because while these methods are ideal, they are not always practical. Consider the Deepwater Horizon Case Study described in Chapter 4; 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, any individual 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. While random selection is a nice tool, the goal is a sample which is *representative* of the population. While random sampling is helpful 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 participant demographics such as age, gender, etc. It is also why it is extremely

important for researchers to describe how observations were obtained so that readers may make the judgement for themselves whether the sample is representative.

# 7 Presenting the Evidence (Summarizing Data)

If you open any search engine and look up "data visualization," you will quickly be 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 III
>
> The use of data for decision making requires that the data be summarized and presented in ways that address the question of interest and represent the variability present.

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 chapter, we will consider various simple graphical and numerical summaries to help build a case for addressing the question of interest. The majority of the chapter is focused on summarizing a single variable; more complex graphics are presented in future chapters within a context that requires them.

## 7.1 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 5.3). Therefore, characterizing the underlying distribution is also the key to a good visualization or numeric summary. For the Deepwater Horizon Case Study described in Chapter 4, 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 the proportion of individual subjects that fall into each of the various groups defined by the categorical variable. Figure 7.1) 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

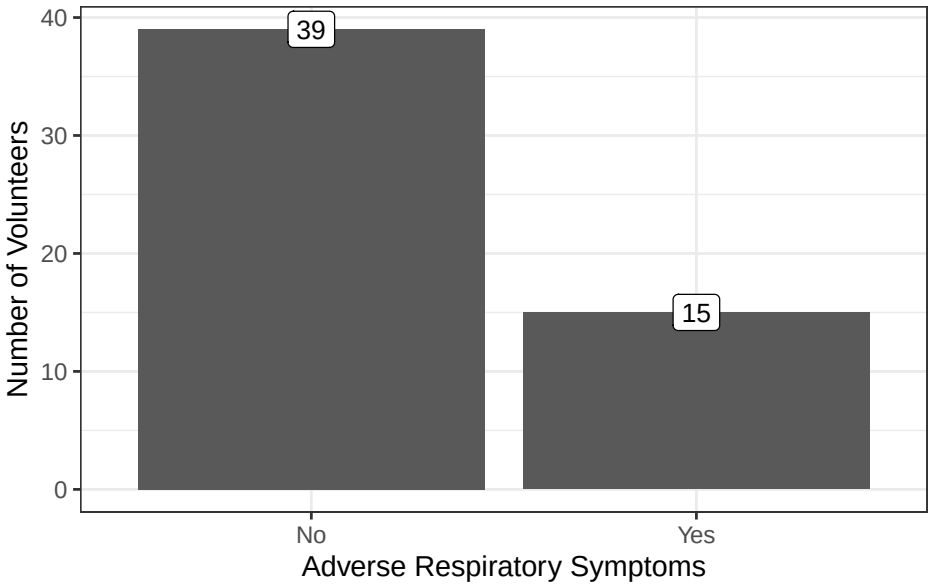


Figure 7.1: Frequency of adverse respiratory symptoms for volunteers cleaning wildlife following the Deepwater Horizon oil spill.

be provided with the figure, either as an annotation or within the caption. From the above graphic, we see that nearly 28% of volunteers assigned to wildlife experienced adverse respiratory symptoms; the graphic helps address our question, even if not definitively.

> **Note**
>
> When you are summarizing only categorical variables, a bar chart is sufficient. Statisticians tend to agree that bar charts are preferable to pie charts (see this whitepaper and this blog for further explanation).

While a single type of graphic (bar charts) are helpful for looking at categorical data, summarizing the distribution of a numeric variable requires a bit more thought. Consider the following example.

**Example 7.1** (Paper Strength)**.** While electronic records have become the predominant means of storing information, we do not yet 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,

Table 7.1: 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 |

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 7.1. The complete dataset 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 7.2 presents the breaking length for all 62 paper specimens in the sample through a *dot plot* in which the breaking length for each observed specimen is represented on a number line using a single dot.

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 7.1** (Percentile)**.** The $k$-th percentile is 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" and denoted $Q_1$).

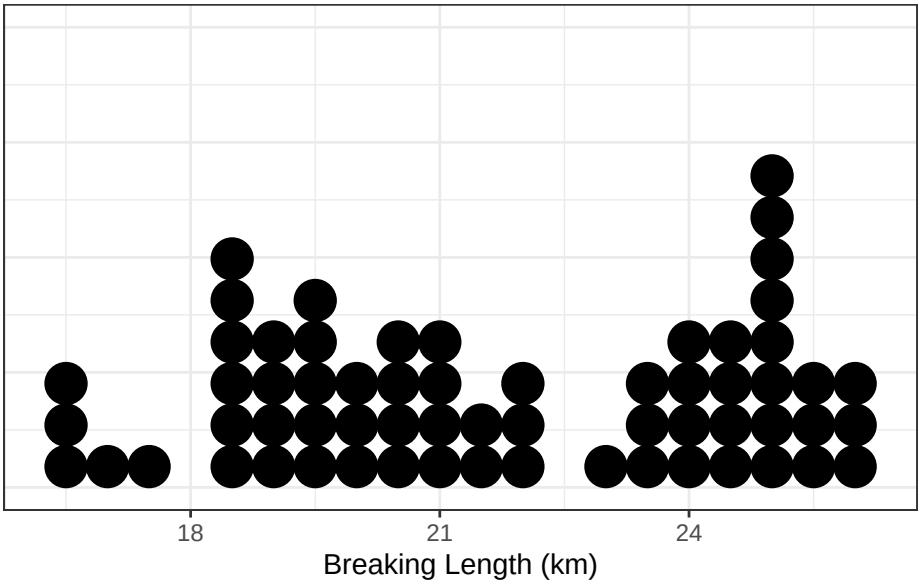


Figure 7.2: Breaking Length (km) for 62 paper specimens.

- 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" and denoted $Q_3$).

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 7.2** (Average)**.** Also known as the "mean," this measure of location represents the balance point for the distribution. If $x_i$ represents the $i$-th value of the variable $x$ in the sample, the sample mean is typically denoted by $\bar{x}$.

For a sample of size $n$, it is computed by

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

When referencing the average for a population, the mean is also called the "Expected Value," and is often denoted by $\mu$.

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 is 5.2 km.

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. Neither of these values has a natural interpretation; instead, larger values of these measures simply indicate a higher degree of variability in the data.

**Definition 7.3** (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 $n$ removes bias in the statistic.

The symbol $\sigma^2$ is often used to denote the variance in the population.

**Definition 7.4** (Standard Deviation)**.** A measure of spread, this is the square root of the variance.

**Definition 7.5** (Interquartile Range)**.** Often abbreviated as IQR, this is 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.

> **Note**
>
> The IQR is often incorrectly reported as the interval $(Q_1, Q_3)$. The IQR is actually the width of this interval, not the interval itself.

The measures we have discussed so far are illustrated in Figure 7.3. While some authors suggest the summaries you choose to report 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

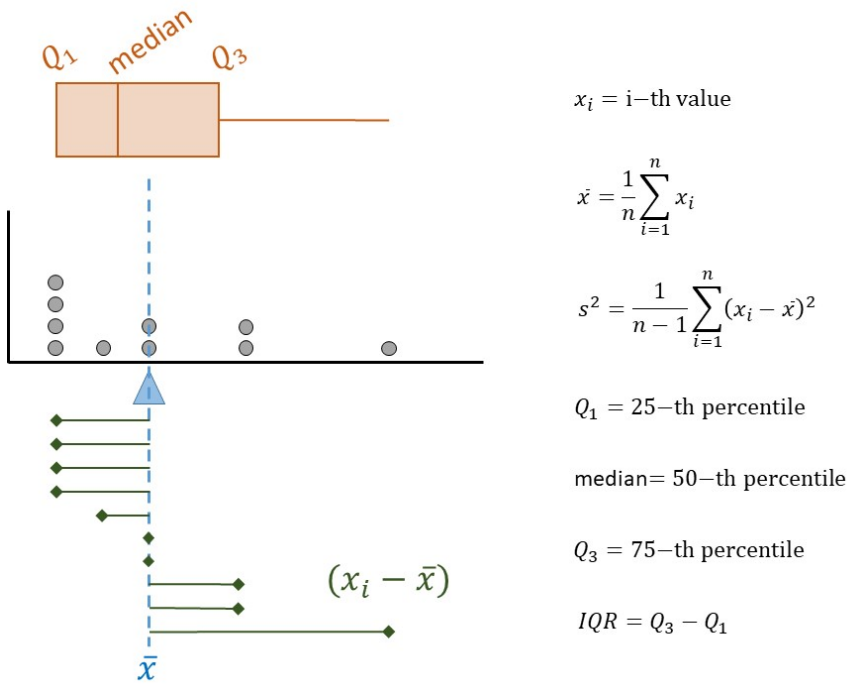


Figure 7.3: Illustration of measures of location and spread for a distribution of values.

"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 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 four components of what we call the *Distributional Quartet*.

**Definition 7.6** (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 numeric variable include:

- *jitter plot*: similar to a dot plot, each value observed is represented by a dot; the dots are "jittered" (shifted randomly) in order to avoid over-plotting when many subjects share the same value of the response.
- *box plot*: a visual depiction of five key percentiles; the plot includes the minimum, first quartile, median, third quartile, and maximum value observed. The quartiles are connected with a box, the median cuts the box into two components. Occasionally, **outliers** are denoted on the graphic.
- *histogram*: can be thought of as a grouped dot plot in which subjects are "binned" into groups of similar values. The height of each bin represents the number of subjects falling into that bin.
- *density plot*: a smoothed histogram in which the y-axis has been standardized so that the area under the curve has value 1. The y-axis is not interpretable directly, but higher values along the y-axis indicate that the corresponding values on along the x-axis are more likely to occur.

**Definition 7.7** (Outlier)**.** An individual observation which is so extreme, relative to the rest of the observations in the sample, that it does not appear to conform to the same distribution.

To illustrate these graphics, the breaking length for the Paper Strength example is summarized using various methods in Figure 7.4. 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.

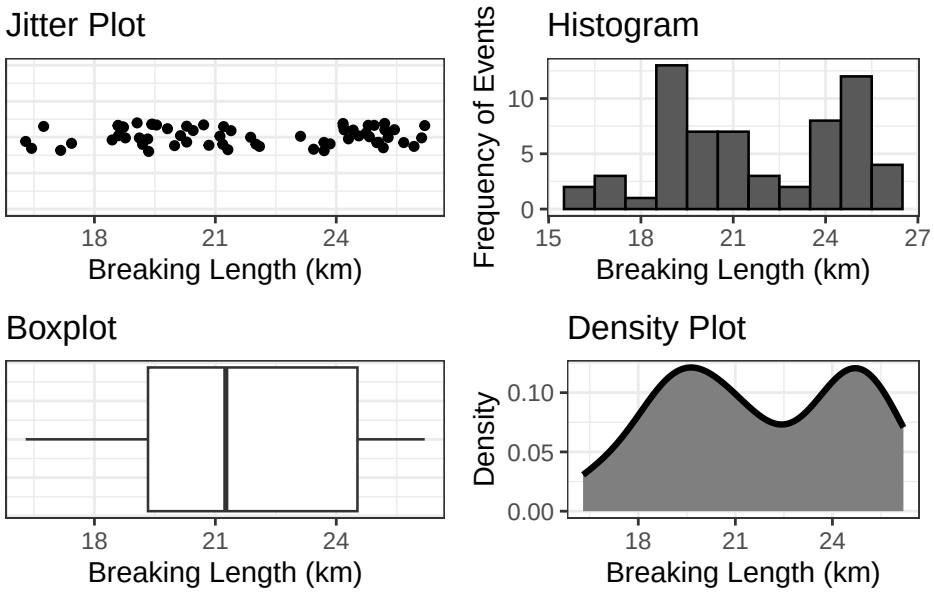


Figure 7.4: Four graphical summaries of the breaking length for the Paper Strength example.

**Definition 7.8** (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 long as the sample is representative, the distribution of the sample should reflect the **distribution of the population**; therefore, summaries of the sample should be close to the analogous summaries of the population (statistics estimate their corresponding parameters). 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 7.9** (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.

Statistics being a proxy for the corresponding parameter implies 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.

Big Idea

A representative sample reflects the population; therefore, we can use statistics as estimates of the population parameters.

Note

Notation in any discipline is both important and somewhat arbitrary. We can choose any symbol we want to represent the sample mean. However, it is convention that we never use $\bar{x}$ to represent a parameter like the mean of the population. The symbol $\bar{x}$ (or $\bar{y}$, etc.) represents observed values being averaged together. Since the values are observed, we must be talking about the sample, and therefore $\bar{x}$ represents a statistic. A similar statement could be made for $s^2$ (sample variance) compared to $\sigma^2$ (population variance). Again, in reality, the symbols themselves are not important. The importance is on their representation. Statistics are observed while parameters are not.

## 7.2 Summarizing Relationships

The summaries discussed above are nice for examining a single variable. In general, however, 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 as well as 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 described in #sec-caseDeepwater:

> 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 7.5; this is *not* a useful graphic. While it compares the number of volunteers with symptoms in each group, we cannot adequately address the question because the research question involves comparing the rates for the two groups; that is, we are lacking a sense of how many volunteers in each group did not report symptoms.

Instead, Figure 7.6 compares the rates within each group. Note that the graphic is still reporting frequency along the y-axis; that was not the primary problem with Figure 7.5. However, by reporting frequencies for both those with respiratory symptoms and those without, we get a sense of the relative frequency with which respiratory symptoms occur.

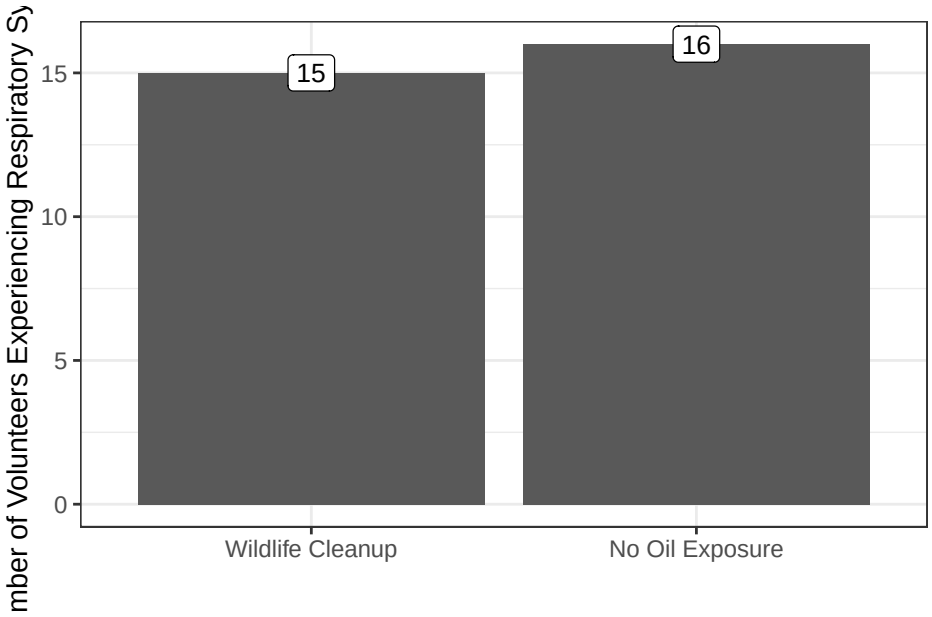


Figure 7.5: Illustration of a poor graphic; the graphic does not give us a sense of the rate within each group at which volunteers reported symptoms.

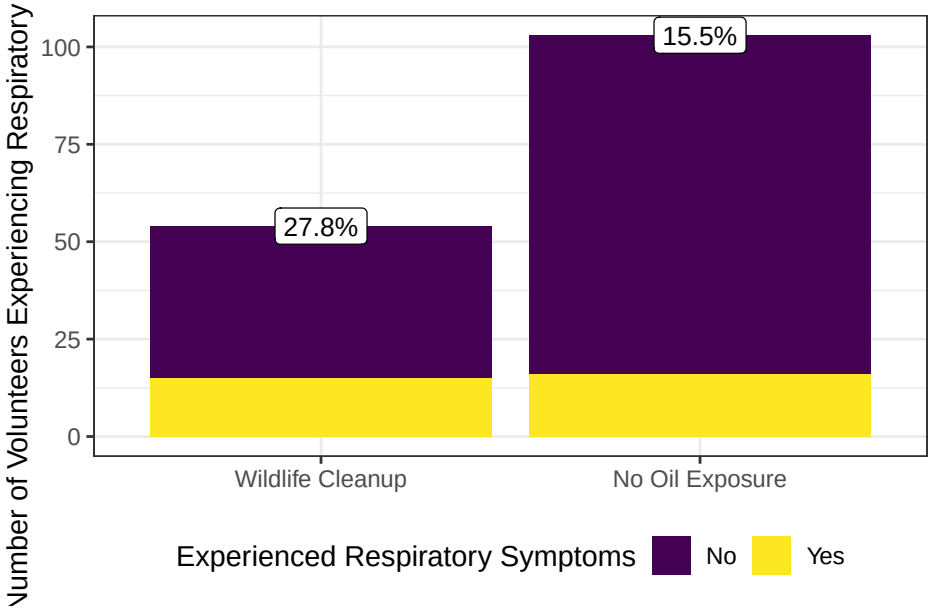


Figure 7.6: Comparison of the rate of adverse respiratory symptoms among volunteers assigned to different tasks.

From the graphic, it becomes clear that within the sample 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.

# Part III

# Unit III: Fundamentals of Bayesian Inference

Once we have data, we want to use it to say something about the underlying population. This is the process of "drawing inference." There are two large paradigms in the statistical community for defining the framework under which inference occurs. This unit introduces the fundamental components of the Bayesian paradigm. We focus on the mechanics in scenarios for which the process is analytically tractable by hand. The remainder of the text merely illustrates these principles in more complex scenarios.

# 8 Bayes Rule

In a probability course, Bayes' Rule is often presented as a neat trick for solving a particular type of problem involving two events. While this simple class of problems understates its true potential, it does serve as a way of highlighting the key idea behind the method.

**Example 8.1** (Disease Testing)**.** An enzyme-linked immunosorbent assay (ELISA) test is performed to determine if the human immunodeficiency virus (HIV) is present in the blood of individuals. Suppose that the ELISA test correctly indicates HIV 99% of the time, and it correctly indicates being HIV-free in 99.5% of cases. Finally, suppose that the prevalence of HIV among blood donors is known to be 1/10000. What is the probability an individual who testes positive is actually infected with HIV?

In this example, we are interested in the probability of an individual being infected with HIV *given* their test is positive. Recalling the definition of conditional probability (Theorem 1.4), we consider

$$Pr(\text{Infected with HIV} \mid \text{Tests Positive}) = \frac{Pr(\text{Infected with HIV} \cap \text{Tests Positive})}{Pr(\text{Tests Positive})};$$

however, these probabilities are not provided in the problem. What we actually have are the probability of testing positive given the patient is infected with HIV (0.99), the probability of testing negative given the patient is not infected with HIV (0.95), and the probability of having HIV (1/10000). This is the power of Bayes' Rule — it allows you to address problems by reversing the conditioning. That is, we can make a statement about the likelihood of $A$ given $B$ using information about the likelihood of $B$ given $A$! As we state Bayes Rule, we keep in mind that it is not the rule itself which is innovative but (a) what the rule implies and (b) how we apply it to solve problems in statistical inference, that make it valuable.

**Theorem 8.1** (Bayes Theorem for Two Events)**.** *Given events $A$ and $B$ such that $Pr(A), Pr(B) \neq 0$, then we have that*

$$Pr(A \mid B) = \frac{Pr(B \mid A)Pr(A)}{Pr(B \mid A)Pr(A) + Pr(B \mid A^c)Pr(A^c)}$$

Bayes' Rule is actually a convenient wrapper for an application of several other basic probability results combined:

- The numerator is an application of the definition of conditional probability; we can always write a joint probability ("and statement") as the product of a conditional probability and a marginal probability.
- The denominator is the result of the total probability rule; a marginal probability is computed by summing over mutually exclusive joint probabilities, which again are rewritten similarly to the numerator.

> **Big Idea**
>
> Bayes' Rule says that our information about $A$, *after* observing $B$, can be stated in terms of what we know about $B$ after observing $A$ and our belief about $A$ *prior* to seeing $B$.

The above is useful when talking about two events, but the majority of applications address random variables, not specific events. That is, we are interested in characterizing entire distributions, not just probabilities for specific events. Following the same logic as above, we are able to extend (from a total probability rule and from the definition of conditional probability) the above result to two random variables.

**Theorem 8.2** (Bayes Theorem for Two Random Variables)**.** *Let $X$ and $Y$ be two random variables; then, we have that*

$$f_{Y|X}(y \mid x) = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int_{S_Y} f_{X|Y}(x \mid y) f_Y(y) dy},$$

*where the integration is replaced by summation when necessary to account for a discrete random variable.*

> **Note**
>
> We will not distinguish between continuous and discrete random variables. For compactness, all results are presented assuming continuous random variables. When necessary, replace integration with summation (as summation is really just integration with respect to a specific measure).

As stated above, this result is really the application of basic definitions covered in a probability course. But, they are worth revisiting.

**Definition 8.1** (Conditional Density)**.** Let $X$ and $Y$ be two random variables; the conditional density of $X$ given $Y$ is

$$f_{X|Y}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

> **Note**
>
> Subscripts are to denote which random variable is being discussed; so, $f_X(x)$ refers to the density function of the random variable $X$ evaluated at $x$. We suppress the subscripts when the context makes it clear which random variable is being referenced.

Rearranging the terms in Definition 9.3, we are able to see that any joint density function $f_{X,Y}(x, y)$ can be written as the product of a conditional density and a marginal density. Similarly, we can write any marginal density by integrating over a joint density.

**Theorem 8.3** (Total Probability Rule for Random Variables)**.** *Let $X$ and $Y$ be two random variables with joint density $f_{X,Y}(x, y)$; then, the marginal density of $X$ is given by*

$$f_X(x) = \int_{\mathcal{S}_Y} f_{X,Y}(x, y)dy,$$

*where $\mathcal{S}_Y$ is the support of $Y$.*

To add a little more intuition to this result, let $X$ represent the grade in this course, and suppose we are interested in the event that $X$ takes the value "A." Well, this course is most certainly impacted by your other courses; so, let $Y$ take the value of the grade in the hardest class remaining on your schedule. Then, there are only a certain number of options:

- $X$ takes the value "A" while $Y$ takes the value "A" (whoo hoo!);
- $X$ takes the value "A" while $Y$ takes the value "B";
- $X$ takes the value "A" while $Y$ takes the value "C";
- $X$ takes the value "A" while $Y$ takes the value "D"; and,
- $X$ takes the value "A" while $Y$ takes the value "F" (let's hope not).

Each of these has some probability of occurring. Since this exhausts all possibilities for $Y$, then we can determine the probability of $X$ by summing the probability of each of these mutually exclusive events. The above lemma captures that we can do this for all values in the support of $X$ simultaneously. Essentially, we have a partition across the support of $X$ based on the value of $Y$; then, we compute the probability of $X$ by summing over the partition.

The above definition is sufficient for several applications. However, it is worth stating the theorem from the most general of perspectives. To do so, we need to define the concept of a random vector.

**Definition 8.2** (Random Vector)**.** Let $X_1, X_2, \ldots, X_n$ be $n$ random variables. Then, the vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ is a random vector of length $n$.

A random vector is essentially a vector comprised of random components. This will be necessary moving forward because we typically have samples of size $n > 1$.

**Theorem 8.4** (Bayes Theorem)**.** *Let* $\mathbf{X}$ *and* $\mathbf{Y}$ *be two random* vectors*. Then, we have that*

$$f_{\mathbf{Y}|\mathbf{X}}(\mathbf{y} \mid \mathbf{x}) = \frac{f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y})}{\int_{\mathcal{S}_{\mathbf{Y}}} f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x} \mid \mathbf{y}) f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y}}$$

*where the integral is now a multi-dimensional integral. Integration for any component is replaced by summation when needed.*

## 8.1 Tenants of the Bayesian Approach to Inference

The above results on their own may be interesting in a probability course. However, we are interested primarily in their application when we have observed a sample from a population which is not fully known. Before we delve into the mechanics, let's pause to reflect on how we intend to apply these results.

Recall that statistics is about using a sample to make inference on the population. Specifically, we will posit a model for the distribution of a response within the population; however, that model will be specified only up to some unknown parameters. There are two general statistical paradigms for performing inference, and these stem from two different questions we might ask:

- Given a hypothesis about the parameters is true, how likely is the observed data?
- Given the observed data, how likely is a particular hypothesis about the parameters?

The first question results in the classical Frequentist perspective (most statistical courses) and a frequentist interpretation of probability. The second results in the Bayesian perspective and a subjective interpretation of probability.

Prior to collecting data, we might have some belief about the the unknown parameters that govern our model for the population. Then, we collect a sample from the population; since this data is representative of the population, it must contain information about those parameters. Therefore, we want to update our belief about the parameters in light of this data. That is the Bayesian process in a nutshell.

**!** Tenants of the Bayesian Approach to Inference

Every analysis in this course is built on the following three tenants:

1. The Bayesian approach takes into account *prior* knowledge when making inference.
2. The Bayesian approach uses probability models to *quantify uncertainty* in the pa-

rameters.
3. The Bayesian approach *updates* our prior knowledge conditional on the observed data.

Throughout, we will rely on a subjective view of probability. That is, probability characterizes how sure you are of something. So, it does not make sense to say "how likely is it to rain tomorrow?" There is no one probability that answers this question. Instead, we will always have (even if not explicitly stated) a "how likely *do you believe*...'' element to our question. That is, we are always bringing in our personal (subjective) opinion. This can be very uncomfortable for some of us — the idea of there not being a single"right" answer. We will save this discussion for a future chapter.

# 9 Modeling Samples

Rarely is our data a single observation. Instead, we collect a sample of observations. As a result, we must be able to comfortably model a *collection* of random variables.

In a probability course, we are generally concerned with modeling a single random variable. The course may build into modeling the joint distribution of two random variables, but generally not further. However, if each random variable represents a single measurement taken on a unit observation, then taking measurements on a sample of $n$ observations is actually a collection of $n$ random variables. Part of quantifying our uncertainty in a parameter is first modeling the process that generated the data as a function of that parameter; that means, we need to model the distribution of the responses.

> **Note**
>
> We often talk about "modeling the data," but that is not precise language. The *data* is fixed; it is not modeled. We are modeling the *process* which generated that data — it is this process that produces random values which are then observed.

Let $X_1, X_2, \ldots, X_n$ represent $n$ observations of the same variable we intend to make (note the future tense). We group these in a random vector $\mathbf{X}$ of length $n$. Not only are we interested in how each *element* in $\mathbf{X}$ is distributed, we are also interested in modeling how they are interrelated.

**Definition 9.1** (Joint Density)**.** For a random vector $\mathbf{X}$, the function $f_{\mathbf{X}}(\mathbf{x})$ such that for any set $A \in \mathbb{R}^n$, we have

$$Pr(\mathbf{X} \in A) = \int \cdots \int_A f_{\mathbf{X}}(\mathbf{x}) dx_1 \cdots dx_n$$

is called the joint density function; this is also referred to as the *likelihood*. Integrals are replaced by sums when appropriate.

> **Big Idea**
>
> Probabilities involving multiple random variables involve integration over the joint density.

The joint density describes how the elements move together; if we are interested in only a single element, we consider the marginal density, which is accomplished by integrating (or summing) over all possible values for the *other* elements.

**Definition 9.2** (Marginal Density)**.** For a random vector $\mathbf{X}$, the marginal density of the first component $X_1$ (without loss of generality) is

$$f_{X_1}(u) = \int \cdots \int f_{\mathbf{X}}(\mathbf{x}) dx_2 \cdots dx_n.$$

Bayes Theorem, and therefore Bayesian data analysis, is primarily concerned with conditional densities, which also generalize to random vectors.

**Definition 9.3** (Conditional Density)**.** Let $\mathbf{X}$ be a random vector; without loss of generality, partition $\mathbf{X}$ such that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

where $\mathbf{X}_1$ represents the first $k$ components and $\mathbf{X}_2$ represents the remaining $n-k$ components. Then, the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2$ is

$$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1 \mid \mathbf{x}_2) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}_2}(\mathbf{x}_2)}.$$

It is worth noting that with respect to the components of interest $\mathbf{X}_1$, the denominator in the conditional density is just a constant scaling factor to ensure the density integrates/sums to 1; that is, *with respect to the variables of interest*, the denominator is a constant.

> **Note**
>
> When we are working with named distributions, using statistical software to compute probabilities is often superior to generic calculus software. This is because the algorithms for computing these probabilities are more stable for known distributions than general all-purpose numerical integration methods.

## 9.1 Independent and Identically Distributed

The above discussion, while accurate, is unrealistic in that it begins with a completely formed likelihood. In reality, we must posit models which correspond to the data generating process. Positing a model for the distribution of an individual observation (element of $\mathbf{X}$) often means choosing from among well-known named probability models. Regardless of whether a named model is used or a custom model constructed, the process always involves examining the context to determine an appropriate structure — the shape and support — of the distribution. We then allow the parameters of this distribution to remain unknown. This is where we turn from probability to statistics — suddenly, our models are only partly known, and there are some aspects (the parameters governing the behavior of the model) which are unknown. We will use data to make some statements about these parameters to address questions of interest which are framed in terms of these parameters.

Instead of trying to model the joint distribution of the observed data directly, we often model the variability in the individual observations. We then place additional conditions on the relationship between the observations in order to develop the joint distribution. One of the most popular conditions is that of independence.

**Definition 9.4** (Independence)**.** Random variables $X_1, X_2, \ldots, X_n$ are said to be mutually independent (or just "independent") if and only if

$$Pr\left(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n\right) = \prod_{i=1}^{n} Pr\left(X_i \in A_i\right),$$

where $A_1, A_2, \ldots, A_n$ are arbitrary sets. Perhaps more helpful, $X_1, X_2, \ldots, X_n$ are said to be mutually independent if and only if

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}\left(x_i\right).$$

> **Note**
>
> For those not familiar, $\prod_{i=1}^{n} a_i$ is the *product operator*. It is analogous to $\sum_{i=1}^{n} a_i$, but uses products instead of sums.

Essentially, a random variable $X$ is said to be independent of $Y$ if the likelihood that $X$ takes a particular value is the same regardless of the value $Y$ takes.

Assuming independence allows us to easily construct joint densities by taking the product of the marginal density for each observation. Independence is a powerful condition when constructing likelihoods. However, it cannot be blindly enforced; we should take caution when

assuming independence. This requires considering the method in which the data was obtained to determine if it is reasonable that the value of one observation does not affect the likelihood of any other observation.

Ideally, when we take a sample, each observation is representative of the same process. This is what allows us to use all observations in a sample in order to make inference — we believe that each observation is able to contribute information about the unknown parameter. Believing that each observation is representative of the same process is essentially assume that each corresponding random variable (prior to observing the data) has the same distribution.

**Definition 9.5** (Identically Distributed)**.** We say that random variables $X$ and $Y$ are identically distributed if $F_X(u) = F_Y(u)$ for all $u$. This is equivalent to saying the two random variables have the same density function $f$.

> ⚠ Warning
>
> Let $X$ and $Y$ be identically distributed random variables. This does not mean that $X = Y$. "Identically distributed" says the two random variables have the same distribution, not the same value. As a result, they share the same mean, variance, etc.

When the observations in our sample are both independent and identically distributed, we say we have a "random sample."

**Definition 9.6** (Random Sample)**.** A random sample of size $n$ refers to a collection of $n$ random variables $X_1, X_2, \ldots, X_n$ such that the random variables are mutually independent, and the distribution of each random variable is identical.

**Example 9.1** (Delivery by Cesarean Section (C-section))**.** It is sometimes necessary for babies to be delivered through a surgical procedure known as a Cesarean Section (C-section). As surgical procedures carry risk, a C-section is typically performed when a vaginal delivery would place the infant or mother in undue risk of complications. Suppose we are interested in characterizing the hospital experiences of mothers who have undergone a C-section at Union Hospital in Terre Haute, Indiana.

For this community health project, we would like to survey $n = 15$ mothers who have undergone a C-section. Of course, not every delivery is a C-section; let $X_i$ represent the number of vaginal deliveries that occur *between* the $i$-th C-section and the previous C-section we observe.

Suppose we are willing to believe that (absent any additional information on the pregnancy) each patient in the labor and delivery ward has the same probability of undergoing a C-section; further, whether one patient undergoes a C-section is independent of any other patient undergoing a C-section. Develop a model for the likelihood of the data to be observed.

*Solution.* We begin by thinking about the specific context. Note, for example, that $X_i$ is a non-negative integer; that is, $X_i \in \{0, 1, 2, \dots\}$. Let $\theta$ represent the probability that a patient undergoes a C-section (and therefore $1-\theta$ represents the probability of a vaginal birth). Since we believe the method of delivery for one patient is independent of the method of delivery for all other patients, and that each probability of a delivery by C-section is the same for each patient, then it is reasonable to state that

$$Pr\left(X_i = x\right) = \theta(1-\theta)^x \qquad x = 0, 1, 2, \dots \, .$$

That is, $X_i$ follows a Geometric distribution with parameter $\theta$. This distribution captures the idea that $x$ vaginal deliveries occur (each with probability $1-\theta$) before we see the $i$-th C-section (which occurs with probability $\theta$).

Further, since each birth is independent, we can consider $X_1, X_2, \dots, X_n$ to be a random sample. Letting $\mathbf{X} = (X_1, X_2, \dots, X_n)^\top$ be the random vector of observations, the likelihood is given by

$$\begin{aligned} f(\mathbf{x} \mid \theta) &= \prod_{i=1}^{n} f_{X_i}\left(x_i \mid \theta\right) \\ &= \prod_{i=1}^{n} \theta(1-\theta)^{x_i} \\ &= \theta^n (1-\theta)^{\sum_{i=1}^n x_i} \\ &= \theta^n (1-\theta)^{n\bar{x}}. \end{aligned} \tag{9.1}$$

Note that line (1) makes use of the independence to say the likelihood is the product of the marginal density functions of each observation. Line (2) makes use of that each observation is identically distributed; this means that each observation has the same functional family for the density and is governed by the same parameter. This still allows $x_i$ to differ from $x_j$, but the distribution is the same. Line (3) brings the product through the expression, with the product of exponentials with the same base resulting in adding the exponents. Line (4) simplifies the expression; for notational simplicity, we prefer $n\bar{x}$ to $\sum_{i=1}^n x_i$, though the two are equivalent.

We also note that the likelihood expressly acknowledges the dependence on the parameter $\theta$ by using $f(\mathbf{x} \mid \theta)$ instead of just $f(\mathbf{x})$.

> **Note**
>
> It is helpful to be in the habit acknowledging the dependence of the likelihood on the parameter.

Big Idea

By placing conditions on how the data is generated, we are able to model the joint distribution of the responses. This is sometimes referred to as the *likelihood* of the unknown parameters; we also refer to it as the model for the data generating process, as it explains the variability in the observed data.

# 10 Quantifying/Modeling Prior Information

Data contributes information to, and therefore impacts, our beliefs. But, prior to beginning a study, we generally have some established beliefs — based on previous studies, expert opinions, personal experience, etc. The Bayesian framework explicitly incorporates these established *prior* beliefs in the analysis. The beliefs just need to be quantified.

> Big Idea
>
> The Bayesian framework encodes any uncertainty through probability distributions.

Before we examine the technical aspects of quantifying the beliefs we have prior to the start of a study, we need to consider how this fits into the larger scope of performing inference. Recall that our primary aim is to make some statement about the population using a corresponding sample. Further, we have some model for the data generating process up to some unknown parameters (this was the focus of the previous chapter). When we collect data, it provides additional information about these unknown parameters. That is, the data impacts the beliefs we have about these unknown parameters. Similarly, any beliefs we have entering the study must relate to these unknown parameters.

Prior to beginning the study, we generally have some notion about the parameters that govern the data generating process. What is a typical GPA for a college student? How much does a member of the mathematics faculty earn each year, on average? While we use data to inform these beliefs (topic of the next chapter), even without data available, we have some idea of where we think the answer lies. Bayesians encode these beliefs into probability distributions. The beliefs we have prior to seeing the data are described by a "prior" distribution, since the beliefs were those we had *a priori*.

**Definition 10.1** (Prior Distribution)**.** A distribution quantifying our beliefs about uncertainty in the *parameter(s)* of the underlying sampling distribution *prior to* observing any data. This is often denoted by $\pi(\theta)$ where $\theta$ is the parameter vector.

- This relies on a *subjective* view of probability.
- As prior beliefs are subjective, there is no "one" prior, but each individual may have a unique prior.

Constructing a prior distribution is not all that different from constructing the likelihood. There are several aspects involved, but it is all about understanding the structure of the beliefs.

**Tips for Constructing a Prior**

The following considerations should be kept in mind when constructing the prior distribution.

- Identify the unknown parameter(s). That is, on what unknown value(s) does the *likelihood* (model for the data generating process) depend?
- Describe the support for the parameter(s).
- Use clear statements about our beliefs of the parameters to determine the **hyperparameters**.

**Definition 10.2** (Hyperparameter)**.** A constant term of a prior distribution that characterizes the family we are considering.

**Note**

It is sometimes said that a hyperparameter is a "parameter" of the prior distribution. You want to distinguish between "parameters" (constant terms that characterize the likelihood), which are unknown, and hyperparameters (constant terms that characterize a prior), which are known values chosen such that the prior distribution reflects our prior beliefs.

**Example 10.1** (A Naive Classification of College Students)**.** Rose-Hulman Institute of Technology (RHIT) and Indiana State University (ISU) are located in Terre Haute, IN. While both colleges cater to undergraduate students, they have different profiles. For the 2021-2022 academic year, Indiana State University reported having 5738 full-time undergraduate students, 3232 (56.3%) of which identified as female. For the same year, Rose-Hulman reported having 2058 full-time undergraduate students, 507 (24.6%) of which identified as female.

Suppose an individual sees a group of 10 college students hanging out at a coffee shop in Terre Haute; they are interested in determining which college the students attend. If the students attend ISU, then we would expect 56.3% to identify as female; if the students attend RHIT, then we would expect 24.6% to identify as female. However, since the coffee shop is located in downtown Terre Haute (which is near the ISU campus), the individual believes there is a 60% chance the students are from ISU.

Notice that in this example, no data has been collected — we have no information on how the students within the group identify. The belief about how likely the students are to attend

ISU is stated *prior* to seeing any data, and this belief can therefore be used to form a prior distribution. Again, notice the use of "a" when describing the prior instead of "the." While this prior will reflect the beliefs of this particular individual, if someone had a different set of beliefs, we would arrive at a different prior.

Let $Y$ represent the number of students who identify as female. Then, $Y \sim Bin(10, \theta)$, where $\theta$ is the probability that a student identifies as female. Notice we are modeling the data that we have not yet collected; that is, this represents the probability model for the likelihood. This likelihood depends on the unknown parameter $\theta$, which represents the probability a randomly selected student identifies as female. This is the first step in constructing a prior — constructing the likelihood and identifying any unknown parameters.

Now, we describe the support for $\theta$. Ordinarily, we might think that $\theta$ could be any value between 0 and 1 since it represents a probability. However, notice that the context we have here suggests there are really only two possible values: either $\theta = 0.563$, representing the gender diversity of ISU students; or $\theta = 0.246$ representing the gender diversity of RHIT students. Therefore, the support of $\theta$ in this particular context is the set $\{0.563, 0.246\}$. Since the support is countable, we will need a discrete distribution for $\theta$.

We are now ready to write a distribution that captures the individual's beliefs prior to observing the data. In this example, the individual is 60% sure the students are from ISU; we write this as

$$Pr(\theta = u) = \begin{cases} 0.4 & u = 0.246 \\ 0.6 & u = 0.563. \end{cases} \tag{10.1}$$

This says the individual is 60% sure that $\theta$ takes the value 0.563, and they are 40% sure that $\theta$ takes the value 0.246; the 0.6 is the *hyperparameter* that governs this distribution. It was chosen to correspond with the prior beliefs stated by the individual.

This is a completely acceptable way of writing the prior distribution. However, as we will later see, the prior distribution is much easier to work with when written in a compact form instead of piecewise notation. For example, we can rewrite Equation 10.1 as

$$\pi(\theta) = 0.4\delta(\theta - 0.246) + 0.6\delta(\theta - 0.563) \tag{10.2}$$

where $\delta(x)$ is the Dirac delta function.

**Definition 10.3** (Dirac Delta Function)**.** The Dirac delta function is the function (not in a rigorous sense) $\delta$ such that

$$\int_{-\infty}^{\infty} \delta(x)dx = 1$$

and

$$\int_{-\infty}^{\infty} f(x)\delta(x)dx = f(0)$$

for any real-valued function $f$.

The Dirac delta function allows us to describe a discrete distribution, which places mass at a single point, as a continuous function on the real line.

The above example offers a rather simplistic view of constructing a prior. In practice, nearly every problem will involve some numerical computation at some point. Rarely, perhaps never, are we simply provided with a complete prior distribution and asked to perform an analysis. Generally, we must convert statements from researchers into some type of distribution.

**Example 10.2** (C-Section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

While we do not know the probability of a C-section, we do have some external information (even before collecting data). Specifically, the March of Dimes has reported that in 2021, 30.4% of live births in Indiana were C-section deliveries. Suppose we have the following beliefs:

- On average, the rate of C-sections at Union Hospital equals the rate of C-sections in the state of Indiana.
- We feel fairly confident (90% sure) the rate of C-sections at Union Hospital is between 20% and 40%.

Develop a suitable prior distribution which captures these beliefs.

*Solution.* As is typical, the prior beliefs that have been provided to us are limited; that is, they do not come pre-packaged in the form of a prior distribution. So, we must develop a prior distribution that aligns with these beliefs. Let's begin by converting the above beliefs into statements about the unknown parameter.

The first belief specifies the average value of the parameter; specifically,

$$E(\theta) = 0.304.$$

The second belief conveys information about where the parameter is located; specifically,

$$Pr(0.2 < \theta < 0.4) = \int_{0.2}^{0.4} \pi(\theta)d\theta = 0.9.$$

Notice the use of the subjective interpretation of probability in capturing this belief. Unfortunately, these two statements alone do not define a unique distribution; this is extremely common as discipline experts do not typically think in probability distributions. Therefore, there is no one unique prior distribution (even for this set of beliefs); instead, we must make some decisions.

Notice that the unknown parameter $\theta$ is a probability; therefore, we know that $\pi(\theta)$ must have a support on the interval $(0, 1)$ since those are the only possible values for $\theta$. Without further guidance, it seems reasonable to select a common distributional family that shares this support; we suggest the Beta distribution. Therefore, we suggest that $\theta \sim Beta(a, b)$. We now must select the values of the hyperparameters $a$ and $b$ so that the prior distribution captures the above statements. That is, we want to choose the hyperparameters to satisfy the following system of equations:

$$\begin{aligned} 0.304 = E(\theta) &= \frac{a}{a+b} \\ 0.90 &= \int_{0.2}^{0.4} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1} d\theta. \end{aligned}$$

As we have two equations and two unknowns, this system can be solved (numerically). Solving this system results in $a = 17$ and $b = 39$ (approximately). Note that the choice of hyperparameters need not carry a lot of precision; these values get us extremely close to the prior beliefs. Therefore, we propose representing our prior beliefs with the prior distribution

$$\pi(\theta) = \frac{\Gamma(17+39)}{\Gamma(17)\Gamma(39)} \theta^{17-1}(1-\theta)^{39-1} \tag{10.3}$$

or equivalently $\theta \sim Beta(17, 39)$.

> **Big Idea**
>
> A prior distribution quantifies the uncertainty we have about a parameter prior to observing data.

# 11 Updating Prior Beliefs (Posterior Distributions)

The previous chapter addressed the construction of a prior distribution, a distribution which captures the uncertainty we have in the unknown parameters governing the data generating process prior to observing any data. Once we observe data, however, the data should update our beliefs about the parameters. Through an application of Bayes' Theorem, we derive the distribution of the parameters after observing the data, incorporating our prior beliefs. This is known as the posterior distribution.

**Definition 11.1** (Posterior Distribution)**.** A distribution quantifying our beliefs about the uncertainty in the parameter(s) of the underlying sampling distribution *after* observing data. This is often denoted by $\pi(\theta \mid \mathbf{y})$ where $\theta$ is the parameter vector and $\mathbf{y}$ the observe data.

Given the likelihood $f(\mathbf{y} \mid \theta)$ and a prior distribution on the parameters $\pi(\theta)$, the posterior distribution is computed using Bayes Theorem:

$$\pi(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta}.$$

A posterior distribution is the conditional distribution of the parameters given the observed data. This allows us to make statements like "given the data, how likely is it that the parameter is between $a$ and $b$." Just as a prior distribution depends on a subjective interpretation of probability, so too does a posterior distribution. With a posterior distribution, we have a way of quantifying our uncertainty in the parameters given the observed data!

> **i** Note
>
> Recall that there is no "one" prior distribution but instead a different prior distribution for each set of prior beliefs. Similarly, there is no "one" posterior distribution. When we say "the" posterior, we are referring to the posterior distribution corresponding to the chosen prior distribution and the data observed.

**Example 11.1** (Naive Classification of College Students, Cont.)**.** Consider Example 10.1 introduced in the previous chapter. Suppose that out of the 10 students, 3 identify as female.

Given this data, how sure is the individual that the college students are from ISU? How has the data observed impacted the individual's prior beliefs?

Recall that we had previously said that the likelihood could be modeled as a Binomial distribution. Specifically, letting $Y$ represent the number of college students in the group that identify as female, then $Y \sim Bin(10, \theta)$. That is,

$$f(y \mid \theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y}. \tag{11.1}$$

Further, based on our prior beliefs, we defined a prior distribution in Equation 10.2:

$$\pi(\theta) = 0.4\delta(\theta - 0.246) + 0.6\delta(\theta - 0.563).$$

For this likelihood and prior distribution, applying Bayes Theorem provides the corresponding posterior distribution. Specifically,

$$\pi(\theta \mid y) = \frac{\binom{10}{3}\theta^3(1-\theta)^{10-3}\left[0.4\delta(\theta - 0.246) + 0.6\delta(\theta - 0.563)\right]}{\int_0^1 \binom{10}{3}\theta^3(1-\theta)^{10-3}\left[0.4\delta(\theta - 0.246) + 0.6\delta(\theta - 0.563)\right] d\theta}. \tag{11.2}$$

Equation 11.2 is accurate, but it is not extremely useful in its current form; in particular, the form is daunting and makes it difficult to interpret directly. We can begin simplifying the expression by first simplifying the denominator. Note that

$$\begin{aligned}
\text{denom} &= \int_0^1 \binom{10}{3}\theta^3(1-\theta)^{10-3}(0.4)\delta(\theta - 0.246) \\
&\quad + \int_0^1 \binom{10}{3}\theta^3(1-\theta)^{10-3}(0.6)\delta(\theta - 0.563)d\theta \\
&= \binom{10}{3}(0.246)^3(0.754)^7(0.4) + \binom{10}{3}(0.563)^3(0.437)^7(0.6).
\end{aligned}$$

Notice that this denominator does not depend on the parameter (as the parameter was integrated out). The denominator is function only of the observed data.

> **!** Important
>
> Once the data is observed, the denominator in the posterior distribution is a constant.

We now use this computed denominator to simplify Equation 11.2. Plugging in, we have

$$\begin{aligned}\pi(\theta \mid y) &= \frac{\binom{10}{3}\theta^3(1-\theta)^{10-3}\left[0.4\delta(\theta-0.246)+0.6\delta(\theta-0.563)\right]}{\binom{10}{3}(0.246)^3(0.754)^7(0.4)+\binom{10}{3}(0.563)^3(0.437)^7(0.6)} \\ &= \theta^3(1-\theta)^{10-3}\left[\frac{\delta(\theta-0.246)}{(0.246)^3(0.754)^7+(0.563)^3(0.437)^7(3/2)}\right. \\ &\quad \left. + \frac{\delta(\theta-0.563)}{(0.246)^3(0.754)^7(2/3)+(0.563)^3(0.437)^7}\right],\end{aligned}$$

which simplifies to

$$\pi(\theta \mid y) = (0.7169)\delta(\theta-0.246)+(0.2831)\delta(\theta-0.563). \tag{11.3}$$

Given the data, the individual can be 71.69% sure the students attend RHIT. Notice that the data has reversed the individual's prior beliefs. Where they were 60% sure the students were from ISU, once they observed the data, they are now more than 70% sure the students are from RHIT. The data observed (3 out of 10 students identifying as female) could easily have come from either school; that is, it is entirely possible that we could sample 10 students at random from ISU and 3 identify as female. However, such a sample is more likely to occur if we sample our 10 students from the RHIT student body. Therefore, the individual's belief about where the students attend school was updated based on the data.

This example illustrates how Bayes Theorem can be used to update our beliefs given observed data. However, there are some additional observations that are worth noting. First, note that the support of the posterior matches the support of the prior.

> **Note**
>
> If the support of the likelihood does not depend on the parameter, then the support of the posterior matches the support of the chosen prior.
> If the support of the likelihood depends on the parameter, the data will further refine the support of the posterior.

If you go into a problem wholeheartedly believing something is not possible, then no amount of data will convince you otherwise; think of this as a core belief that is unshakable. That is, any parameter value that is excluded by the prior distribution will be excluded in the posterior distribution automatically. Data can only convince those who are open to believing something different!

Second, notice the hardest computational aspect of the above example was computing the integral in the denominator and then carrying the algebra through in order to determine a simplified form of the posterior distribution. While we could rely on a computer algebra system in order to perform these computations in simple settings, relying on these tools tends to fail in more complex problems encountered in practice. Moving forward, we will want a way of

Table 11.1: Hypothetical data representing the number of vaginal deliveries between consecutive C-sections.

| | | | | |
|---|---|---|---|---|
| 3 | 1 | 0 | 0 | 0 |
| 2 | 5 | 6 | 9 | 0 |
| 5 | 1 | 0 | 1 | 0 |

overcoming the integral in the denominator, especially in cases when the parameter vector grows to be high-dimensional. To begin emphasizing the need to find alternatives, consider the following observation: the denominator in the computation of the posterior is constant with respect to the parameter. Careful consideration of this observation allows us to move through computations more quickly.

**!** Applying Bayes Theorem in Practice:

The denominator in Bayes rule exists to ensure the distribution integrates to 1; it is just a scaling constant. That is,

$$\pi(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta)\pi(\theta).$$

This recognition allows us to quickly compute the *kernel* of the posterior, which in many cases is sufficient for identifying the posterior distribution.

Finally, we emphasize that the posterior distribution does *not* tell you the value of the unknown parameter — a parameter is unknown and will always remain so! The posterior distribution only tells you the beliefs you have about that parameter given the data you have observed *and* your prior beliefs.

**Example 11.2** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Using the likelihood developed in Example 9.1 and the prior developed in Example 10.2, derive the form of the posterior distribution given a sample of data $X_1, X_2, \ldots, X_n$. Then, suppose we observed the following data, how does it update our beliefs?

*Solution.* We first develop a general solution before substituting in the observed data. Recall that the likelihood (Equation 9.1) was given by

$$f(\mathbf{x} \mid \theta) = \theta^n (1-\theta)^{n\bar{x}},$$

and the prior (Equation 10.3) was given by

$$\pi(\theta) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},$$

where we have written the prior in its general form (not with the specific choices of the hyperparameter). Applying Bayes Theorem, we know that the posterior is proportional to the product of the likelihood and prior; that is,

$$\begin{aligned}\pi(\theta \mid \mathbf{x}) &\propto f(\mathbf{x} \mid \theta)\pi(\theta) \\ &= \theta^n(1-\theta)^{n\bar{x}}\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1} \\ &\propto \theta^{n+a-1}(1-\theta)^{n\bar{x}+b-1}.\end{aligned}$$

Observe that since we are simply trying to determine what the posterior is proportional to, we can drop any scaling constants (with respect to the parameter); doing this in line (3) allows us to drop the gamma terms, simplifying the expression greatly. In fact, we now note that our posterior distribution is proportional to the form $\theta^{\text{something}-1}(1-\theta)^{\text{something else}-1}$, which we recognize as the kernel of a Beta distribution. That is, the appropriate scaling term to ensure that the posterior integrates to 1 (and is therefore a valid density function) is

$$\frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)},$$

giving a posterior distribution of

$$\pi(\theta \mid \mathbf{x}) = \frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)}\theta^{n+a-1}(1-\theta)^{n\bar{x}+b-1}, \tag{11.4}$$

or $\theta \mid \mathbf{x} \sim Beta(n+a, n\bar{x}+b)$. Of course, $a$ and $b$ are known values (chosen in the derivation of the prior), and once we observe the data, $n$, $\bar{x}$ are also known. Therefore, the posterior distribution is fully specified. Specifically, substituting in these known values given the data observed, we have that $\theta \mid \mathbf{x} \sim Beta(32, 72)$.

Figure 11.1 compares the prior and posterior densities given the data in Example 11.2. Notice the two distributions are similar. Both have the same support (the interval $(0, 1)$), and both tend to have a mode (peak) at roughly the same location. However, the posterior distribution has less variability (notice most of its mass is condensed around a tighter interval). This suggests that the data has increased our confidence in the value of the unknown parameter. However, notice that we did not "solve" for the value of $\theta$; in fact, the posterior distribution highlights that we are not certain about the value of $\theta$. Instead, the posterior is simply telling us how likely we feel the parameter is within any particular interval given the observed data.

For example, since

$$\int_{0.2}^{0.4} \pi(\theta \mid \mathbf{x}) d\theta = \int_{0.2}^{0.4} \frac{\Gamma(32+72)}{\Gamma(32)\Gamma(72)} \theta^{32-1}(1-\theta)^{72-1} d\theta = 0.971,$$

given the data observed, we are now 97.1% sure that the rate of C-sections at the hospital is between 20% and 40%; this is an increase from what we believed prior to observing the data.

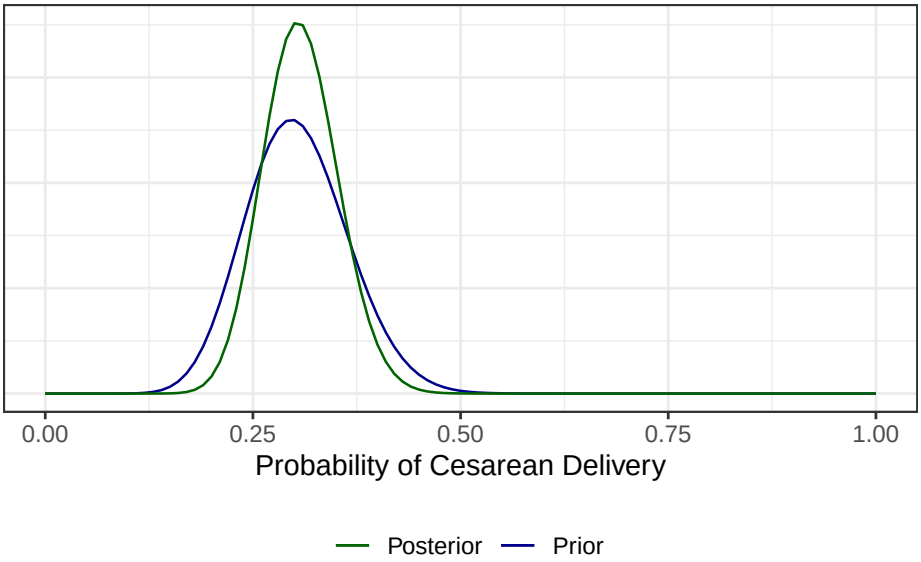


Figure 11.1: Comparison of the prior distribution and posterior distribution for the C-section Deliveries example given the observed data.

# 12 Point Estimation

Everything we could ever want to know about a parameter, given the data we have observed, is contained in the posterior distribution. For those very comfortable with probability theory, we may not shy away from having an entire distribution presented to us as a way of summarizing the information available about the parameter. However, we know there are ways of summarizing a distribution, and often practitioners prefer to be presented with a summary of the posterior distribution. In this chapter, we examine methods for estimating the parameter of interest using the posterior distribution.

> Big Idea
>
> Estimating the parameter of interest is typically done by summarizing the *location* of the posterior distribution.

A course in probability introduces us to the idea of using the mean, the median, and/or the mode in order summarize the location of a distribution. We can apply the same techniques to summarize the location of the posterior distribution. The three common point estimates for a parameter in the Bayesian framework are the posterior mean, the posterior median, and posterior mode.

**Definition 12.1** (Posterior Mean)**.** The posterior mean is the average value of the parameter, given the data:

$$E\left[\theta \mid \mathbf{y}\right] = \int \theta \pi(\theta \mid \mathbf{y}) d\theta.$$

We note that as the dimension of the parameter vector increases, this could be a very difficult integral to compute. We will address this issue in the next unit. For now, we will typically work with known distributions and therefore known expressions for the posterior mean.

**Definition 12.2** (Posterior Median)**.** We are 50% sure, given the data, the parameter falls below the posterior median. Formally, the posterior median is the value $q$ such that

$$0.5 = \int_{-\infty}^{q} \pi(\theta \mid \mathbf{y}) d\theta.$$

While closed-form solutions may exist for the posterior mean, even with known distributions the posterior median must often be computed numerically. This is not problematic; we are simply acknowledging that in statistics, numerical solutions are common and are not viewed as inferior.

**Definition 12.3** (Posterior Mode)**.** We think of the posterior mode as the most likely value of the parameter, given the data. If the posterior distribution is continuous, the posterior mode is the value of the parameter that maximizes the posterior distribution. Formally, the posterior mode is given by

$$\arg\max_{\theta} \pi(\theta \mid \mathbf{y}).$$

> **Note**
>
> The posterior mode only makes sense as an estimate if the posterior distribution is unimodal.

One might ask which of the three estimates is "best." It depends. The mean and median may not be representative of a "typical" value. The mean is more sensitive to extreme values; the median tends to be more stable. However, many software packages default to reporting the posterior mean, making it a popular choice out of simplicity. Again, regardless of which value we choose to report, we should not neglect that we have access to the entire posterior distribution; therefore, we are not limited by a single estimate but can provide a much richer summary of the posterior distribution.

> **Warning**
>
> It is common for those first learning the Bayesian framework to confuse the parameter being estimated with the method of estimation used. We can use the posterior mode to estimate the mean response. We can use the posterior mean to estimate the variance of the response. The method of estimation (posterior mean, posterior median, or posterior mode) is *not* linked to the parameter (mean response, variance of the response, etc.).

We close this chapter by considering two examples. First, consider Example 11.1. Since the support of the posterior includes only two values (0.246 and 0.563), the posterior mean would necessarily take a value not in the support. The posterior median suffers from the same limitation. The posterior mode is 0.246, since it is the more likely value, given the data and the individual's prior beliefs. Note, however, that we lose information by only reporting the point estimate; we have a much richer conclusion when we report the entire posterior distribution: we are 71.69% sure the students are from RHIT and 28.31% sure the students are from ISU.

**Example 12.1** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Example 11.2 found the posterior distribution to be

$$\theta \mid \mathbf{x} \sim Beta\left(n + a, n\bar{x} + b\right)$$

where $a = 17, b = 39, n = 15$ and $n\bar{x} = 33$ given the observed data. Estimate the rate of C-sections at the hospital given the observed data.

*Solution.* While it may seem obvious, our estimate is based on the *observed data*; different data would lead to a different estimate. Hidden in that statement is that our estimate is also based (at least partially) on our prior beliefs; different prior beliefs may also lead to a different estimate.

Since the distributional family of the posterior is well-studied (a Beta distribution), we can make use of established properties in computing our point estimate. In particular, we immediately have that

$$\begin{aligned} E\left(\theta \mid \mathbf{x}\right) &= \frac{n + a}{n + a + n\bar{x} + b} = 0.308 \\ \text{Posterior Mode} &= \frac{n + a - 1}{n + a + n\bar{x} + b - 2} = 0.304. \end{aligned}$$

While we must compute it numerically, the posterior median is also readily available and is given by 0.306. All three estimates are extremely similar. Given the data, we estimate the C-section rate at the hospital to be just over 30% (very near the rate in the state of Indiana).

# 13 Interval Estimation

The previous chapter considered a single point estimate for the parameter of interest. However, this ignores the fact that there is variability in these estimates. The distribution itself tells us the parameter is more likely to fall in some regions than others. In response, we consider providing a range of plausible values for the parameter and quantifying our belief that the parameter falls in that range. This is the contrast between point and interval estimation.

**Definition 13.1** (Point Estimation)**.** Point estimation is the process of estimating a parameter with a single statistic. This is like trying to hit an infinitesimally small target with a dart.

**Definition 13.2** (Interval Estimation)**.** Interval estimation is the process of estimating a parameter with a range of values. This is like trying to capture a target with a ring.

Regardless of which method we use, both are estimates, and both depend on the posterior distribution. That is, both are statements about the parameter given the observed data and our prior beliefs. As there were various techniques for constructing a point estimate, there are various techniques for an interval estimate; the most common of these is the credible interval.

**Definition 13.3** (Credible Interval)**.** A $100c\%$ credible interval is an interval $(a, b)$ such that

$$Pr(a \leq \theta \leq b \mid \mathbf{y}) = \int_a^b \pi(\theta \mid \mathbf{y}) d\theta = c.$$

> ⚠ Warning
>
> For those who have had a previous statistics course taught from the classical Frequentist perspective, this seems to mirror a confidence interval, but the interpretation is completely different. Since probability is used to quantify subjective beliefs, notice that the credible interval allows us to say that we are $100c\%$ sure the parameter falls in this range, given the data.
> Since we are working from a subjective interpretation of probability, we do not need to appeal to repeated sampling (like a Frequentist would). In fact, since a parameter is a fixed, unknown quantity, any probability statement is illogical from a Frequentist perspective. However, from the Bayesian perspective, the posterior quantifies our uncer-

tainty about the parameter, and therefore the credible interval is simply summarizing this uncertainty. We can now say, based on the data observed (however much or little we have), we are $100c\%$ sure the parameter falls in this interval.

**Note**

There is no one unique credible interval for a parameter.

Since there are infinitely many regions which contain $100c\%$ of the posterior distribution, there are infinitely many credible intervals we could provide. In order to provide some level of continuity between applications, we tend to gravitate to one of two types of intervals which have nice properties.

**Definition 13.4** (Equal-Tailed Credible Interval)**.** The equal-tailed credible interval, which is probably the most commonly used in practice, chooses endpoints such that

$$Pr(\theta < a \mid \mathbf{y}) = \frac{1-c}{2} = Pr(\theta > b \mid \mathbf{y}).$$

As the name implies, an equal-tailed credible interval places the same probability in each tail; we are taking the middle $100c\%$ of the posterior distribution.

An equal-tailed interval is easy, but it may not always be the most intuitive interval. Figure 13.1 compares two potential 90% credible intervals for a hypothetical posterior distribution. Observe that the equal-tailed interval removes the bottom 5% of the distribution; while this band is narrow, it represents values which correspond to the highest posterior density values. It seems intuitive that we would want to choose the narrowest credible interval which still retains the same area under the curve, as illustrated in the second panel of Figure 13.1.

**Definition 13.5** (Highest Density Interval)**.** The highest density interval, often called an HDI or HPD (for highest posterior density), chooses the endpoints such that the interval is as short as possible.

When the density is unimodal, this can be accomplished by choosing the endpoints $a$ and $b$ such that

$$\pi(\theta \mid \mathbf{y}) \mid_{\theta=a} = \pi(\theta \mid \mathbf{y}) \mid_{\theta=b}$$

and

$$\int_a^b \pi(\theta \mid \mathbf{y} d\theta = c.$$

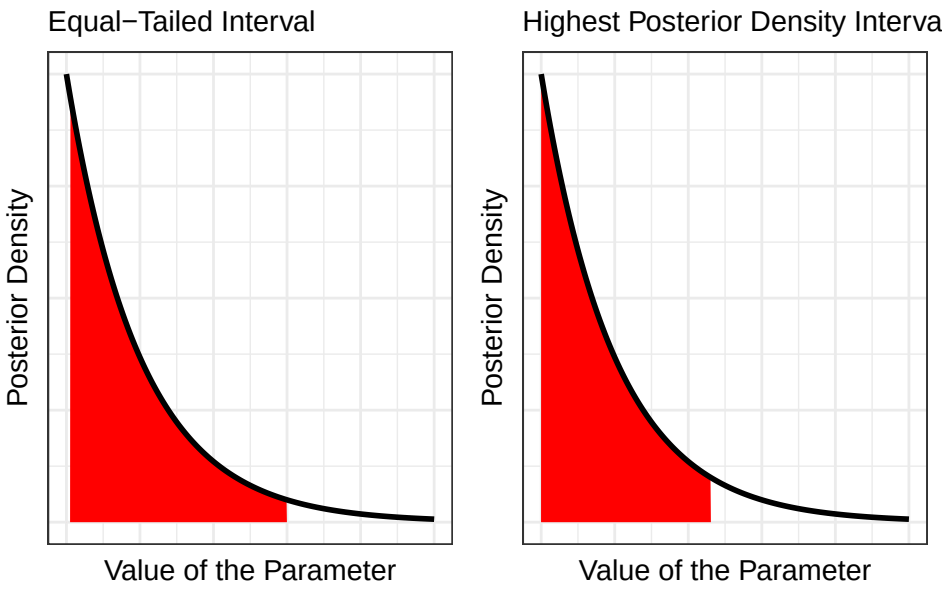


Figure 13.1: Comparison of two 90% credible intervals for a hypothetical posterior distribution.

> **Note**
>
> If the posterior distribution is multimodal, then the highest density interval is actually a *region* as it will likely involve two disjoint intervals.

> **Warning**
>
> Most software that computes an HDI assumes the poterior distribution is unimodal.

Suppose we have a $100c\%$ credible interval $(a, b)$ for some parameter $\theta$, but we are interested in a transformation of the parameter $\eta = g(\theta)$. We can develop a $100c\%$ credible interval for $\eta$ by applying the same transformation to each endpoint of the interval for $\theta$. That is, $(g(a), g(b))$ will be a $100c\%$ credible interval for the parameter $\eta$ given the data.

While we can guarantee that $(g(a), g(b))$ is a $100c\%$ credible interval, it will in general **not** be the HDI for $\eta$, even if $(a, b)$ is the HDI for $\theta$.

**Example 13.1** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Example 11.2 found the posterior distribution to be

$$\theta \mid \mathbf{x} \sim Beta\left(n + a, n\bar{x} + b\right)$$

where $a = 17, b = 39, n = 15$ and $n\bar{x} = 33$ given the observed data. Estimate the rate of C-sections at the hospital given the observed data.

*Solution.* Example 12.1 developed point estimates of the unknown parameter given the observed data (and prior beliefs). Here, we consider an interval estimate. To compute an equal-tailed interval, we must choose the values of $s$ and $t$ such that

$$0.025 = \int_0^s \frac{\Gamma(32+72)}{\Gamma(32)\Gamma(72)}\theta^{32-1}(1-\theta)^{72-1}d\theta$$
$$0.025 = \int_t^1 \frac{\Gamma(32+72)}{\Gamma(32)\Gamma(72)}\theta^{32-1}(1-\theta)^{72-1}d\theta.$$

Solving this system numerical gives the interval $(0.223, 0.399)$. We are 95% sure, given the data, the rate of C-sections at the hospital is between 22.3% and 39.9%. Given that the posterior distribution is unimodal and nearly symmetric, we would expect the HDI to be very similar to the equal-tailed interval.

# 14 Prediction

Previous chapters have focused on making a statement about the parameter given the data. However, researchers are often interested in using the data to predict what might occur in the future. Hopefully, by now you realize that within the Bayesian framework, we are never interested *solely* in a point estimate. So, when we say we are interested in "prediction," we mean we are interested in characterizing our uncertainty in a future value given the data we have observed. As this statement is not directly about the parameter, the posterior distribution does **not** contain the relevant information; but, it can be used to derive the distribution that is relevant.

Imagine we have not yet collected data. Given only our prior beliefs, how might we characterize a future observation (one not yet observed)? We may not know what value a future observation will take, but we have some sense of the process that will generate it if we have posited a likelihood to describe the data generating process.

In probability, we routinely use the distribution to characterize future values each time we answered a question like "what is the probability we will observe a value between $a$ and $b$?" It is therefore intuitive that we might turn toward the likelihood $f(\mathbf{y} \mid \theta)$ to describe the variability in data that has not yet been observed. Of course, this highlights the difference between probability and statistics — in probability, we always knew the value of $\theta$, but in statistics, we do not. Without a value of $\theta$ to plug into the density, we are unable to use it to compute probabilities about future observations (or, more precisely, any probabilities we computed would be a function of the unknown parameter).

This is where our prior beliefs come into play. While we do not know the value of $\theta$, we do have some prior beliefs about it, and these are captured in the prior distribution. The Bayesian framework proposes marginalizing out the parameter — essentially taking a weighted average over all possible values it could be. What results is not a single value, but a distribution of values known as the prior predictive distribution.

**Definition 14.1** (Prior Predictive Distribution)**.** The prior predictive distribution is the marginal distribution of the response(s) prior to observing any data:

$$m(\mathbf{y}) = \int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta.$$

The distribution marginalizes the parameter out of the likelihood using the beliefs from the prior distribution.

> **Note**
>
> The prior predictive distribution is the denominator in Bayes Theorem.

While rarely used directly, the prior predictive distribution provides a way of characterizing our uncertainty in future observations based solely on our beliefs about $\theta$ prior to observing any data.

> **Big Idea**
>
> We can describe our beliefs about future values of the response by marginalizing the parameter out of the likelihood.

Of course, we will eventually collect data, and we would like to take this knowledge into account. After observing the data, the parameter remains unknown; however, our beliefs about the parameter are updated (and are captured in the posterior distribution). We want to marginalize the parameter out of the likelihood while accounting for these updated beliefs.

We consider swapping out the role of the prior distribution in Definition 14.1 with the posterior distribution. The result is the posterior predictive distribution.

**Definition 14.2** (Posterior Predictive Distribution)**.** Let $\mathbf{Y}^*$ represent a collection of $m$ *future* observations. The distribution of these future observations given the observed data $\mathbf{Y}$ (of length $n$), called the posterior predictive distribution, is given by

$$\pi\left(\mathbf{y}^* \mid \mathbf{y}\right) = \int f\left(\mathbf{y}^* \mid \theta\right) \pi(\theta \mid \mathbf{y}) d\theta.$$

While this definition is correct, its derivation requires some additional constraints on the data generating process. We present the derivation below primarily to combat any misconceptions about what is happening in the integration above.

## 14.1 Derivation of the Posterior Predictive

Let $\mathbf{Y}^*$ denote a collection of $m$ future (or new) observations not yet observed. This is distinguished from the collection of $n$ observations we have already made $\mathbf{Y}$. We impose the following two conditions/assumptions on the data generating process:

- Given the value of the parameter, the likelihood of $\mathbf{Y}^*$ has the same form as the likelihood of of the observed data $\mathbf{Y}$.
- Given the value of the parameter, the observed data $\mathbf{Y}$ is *independent* of the new observations $\mathbf{Y}^*$.

The first condition essentially states the data generated under one process should only be used to predict data generated from the same process. Intuitively, when we collect data, it can only inform us about the process from which it was generated. Therefore, our future observations are always related in some way to the likelihood, as that models the data generating process of interest.

The second condition extends the concept of independence presented in a typical probability course. This is *conditional independence*.

**Definition 14.3** (Conditional Independence)**.** Two random variables $X$ and $Y$ are said to be independent, conditional on (or "given") $Z$ if, and only if,

$$f_{(X,Y)|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z).$$

Conditional independence is common in statistical theory. Two random quantities are somehow related, but given an additional piece of information become independent. That is, all the information about the relationship between $X$ and $Y$ is contained in the random variable $Z$.

Returning to the stated condition, we are saying that the only thing the new and old observations have in common is the data generating process; once we know the quantities that govern this process (the parameters), then we can gain no further knowledge about the new observations from the old observations. That is, if someone told you what the parameters were, there would be no need to collect data — you would know everything possible for predicting a future observation. So, the data observed is only useful in that it informs our beliefs about the unknown parameters.

> Big Idea
>
> The data observed informs our beliefs about the parameters in the data generating process. It is only through what the data tells us about the parameters that the data is useful in predicting a future observation.

We are now prepared to derive the posterior predictive distribution. Recall that a marginal distribution can be constructed by integrating over the other elements of a joint distribution. For example,

$$f\left(\mathbf{y}^*\right) = \int f\left(\mathbf{y}^*, \theta\right) d\theta.$$

Here, we have considered the joint distribution of the new data $\mathbf{Y}^*$ and the parameter $\theta$, then integrated over $\theta$. This strategy works even when we are carrying a conditional term through. That is, we have that

$$\pi\left(\mathbf{y}^* \mid \mathbf{y}\right) = \int f\left(\mathbf{y}^*, \theta \mid \mathbf{y}\right) d\theta. \tag{14.1}$$

Here, we have considered the joint distribution of the new data $\mathbf{Y}^*$ and the parameter $\theta$ (conditional on the observed data), then integrated over $\theta$. This statement would have been true for any choice of random variable, but using the parameter allows us to make use of the information we have collected through the observed data.

We now recall that any joint distribution can be written as the product of a conditional distribution and a marginal distribution; this is true even if we are already conditioning on another random variable. This gives

$$f(\mathbf{y}^*, \theta \mid \mathbf{y}) = f(\mathbf{y}^* \mid \mathbf{y}, \theta)\pi(\theta \mid \mathbf{y}). \tag{14.2}$$

Substituting Equation 14.2 into Equation 14.1, we now have that the posterior predictive distribution is given by

$$\pi\left(\mathbf{y}^* \mid \mathbf{y}\right) = \int f(\mathbf{y}^* \mid \mathbf{y}, \theta)\pi(\theta \mid \mathbf{y})d\theta. \tag{14.3}$$

We now make use of conditional independence. We now consider the term $f(\mathbf{y}^* \mid \mathbf{y}, \theta)$ inside the integral. By the definition of a conditional density, we have

$$f(\mathbf{y}^* \mid \mathbf{y}, \theta) = \frac{f(\mathbf{y}^*, \mathbf{y} \mid \theta)}{f(\mathbf{y} \mid \theta)}.$$

However, if we are willing to assume that $\mathbf{Y}^*$ is independent of $\mathbf{Y}$ given $\theta$, then the numerator becomes the product of $f(\mathbf{y}^* \mid \theta)$ and $f(\mathbf{y} \mid \theta)$, meaning we have that $f(\mathbf{y}^* \mid \mathbf{y}, \theta) = f(\mathbf{y}^* \mid \theta)$ under conditional independence. Substituting in this expression into Equation 14.3 gives the posterior predictive distribution in Definition 14.2.

## 14.2 Summary

Once you have the posterior predictive distribution, you have everything there is to know about future observations given the data observed. Further, the posterior predictive distribution can be summarized just like any other distribution. Summarizing the location (mean, median, mode) would result in point estimates for future observations. Alternatively, we can construct interval estimates by defining a range for which the future observations would fall with some known probability.

> **Warning**
>
> Keep in mind that we have switched our focus. We are now focused on a possible data point, not a parameter. Therefore, the support of the posterior predictive need not be the same as the support of the posterior distribution.

**Example 14.1** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Example 11.2 found the posterior distribution to be

$$\theta \mid \mathbf{x} \sim Beta\left(n + a, n\bar{x} + b\right)$$

where $a = 17, b = 39, n = 15$ and $n\bar{x} = 33$ given the observed data. Suppose we are interested in adding a 16-th patient to our survey; predict the number of vaginal deliveries we should expect before we observe a C-section.

*Solution.* We are interested in predicting a new *response* given the observed data. Following the discussion in the derivation of the posterior predictive distribution, it makes sense that we would assume the density of this new response follows the same distribution as the observed data. In particular, if $Y$ represents the future observation, then

$$f(y \mid \theta) = \theta(1-\theta)^y$$

since each observed value (previously denoted by $X_i$) was modeled as a random variate from a Geometric distribution. Applying Definition 14.2, we have that the posterior predictive distribution is given by

$$\begin{aligned}
\pi(y \mid \mathbf{x}) &= \int f\left(y \mid \theta\right) \pi(\theta \mid \mathbf{x}) d\theta \\
&= \int \theta(1-\theta)^y \frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)} \theta^{n+a-1}(1-\theta)^{n\bar{x}+b-1} d\theta \\
&= \frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)} \int \theta^{n+a+1-1}(1-\theta)^{n\bar{x}+b+y-1} d\theta.
\end{aligned}$$

Notice that the integrand is the kernel of a Beta distribution; therefore, we can multiply and divide by the appropriate scaling terms. This gives

$$\begin{aligned}
\pi(y \mid \mathbf{x}) &= \frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)} \int \theta^{n+a+1-1}(1-\theta)^{n\bar{x}+b+y-1} d\theta \\
&= \frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)} \frac{\Gamma(n+a+1)\Gamma(n\bar{x}+b+y)}{\Gamma(n+a+n\bar{x}+b+1+y)} \\
&\quad \cdot \int \frac{\Gamma(n+a+n\bar{x}+b+1+y)}{\Gamma(n+a+1)\Gamma(n\bar{x}+b+y)} \theta^{n+a+1-1}(1-\theta)^{n\bar{x}+b+y-1} d\theta \\
&= \frac{\Gamma(n+a+n\bar{x}+b)}{\Gamma(n+a)\Gamma(n\bar{x}+b)} \frac{\Gamma(n+a+1)\Gamma(n\bar{x}+b+y)}{\Gamma(n+a+n\bar{x}+b+1+y)}.
\end{aligned}$$

We are not meant to recognize this distribution. However, we can work with it. We must keep in mind the warning, however, given just prior to this example — the support of this distribution is non-negative integers, not the interval $(0, 1)$.

Again, prediction is not about saying how many vaginal births we *will* see before the next C-section; it is really about quantifying our uncertainty in the various possibilities. For example, given the data observed, we are 30.8% sure that the very next birth will be a C-section, since

$$Pr(Y = 1 \mid \mathbf{x}) = \frac{\Gamma(32+72)}{\Gamma(32)\Gamma(72)} \frac{\Gamma(32+1)\Gamma(72+0)}{\Gamma(32+72+1+0)} = 0.308.$$

Similarly, we are 88.3% sure that we will not experience more than 5 vaginal births before the next C-section, since

$$Pr(Y \leq 5 \mid \mathbf{x}) \sum_{u=0}^{5} Pr(Y = u \mid \mathbf{x}) = 1 - \sum_{u=0}^{5} \frac{\Gamma(32+72)}{\Gamma(32)\Gamma(72)} \frac{\Gamma(32+1)\Gamma(72+u)}{\Gamma(32+72+1+u)} = 0.883.$$

However, if we would like to provide a single point estimate instead of probabilities of specific responses, we might report that, given the data, *on average*, we expect to see 2.32 vaginal deliveries before the next C-section since

$$E(Y \mid \mathbf{x}) = \sum_{u=0}^{\infty} u Pr(Y = u \mid \mathbf{x}) = 2.32.$$

# 15 Hypothesis Testing

We have considered both estimation and prediction at this point. The third type of question often asked by researchers is which model (out of some pre-defined set) is most supported by the data. As with previous estimation and prediction, the Bayesian framework seeks to characterize the evidence for each model, given the data.

> **Warning**
>
> For those who are familiar with a classical Frequentist approach to hypothesis testing, you will note that the above language is fundamentally different than the Frequentist perspective. The Frequentist perspective does not allow us to characterize the "evidence" for the null hypothesis, and it would certainly never allow us to quantify the probability of a hypothesis being true.

You may recall that for the Naive Classification of College Students example (Example 10.1), we derived the following posterior distribution (see Equation 11.3):

$$\pi(\theta \mid y) = (0.7169)\delta(\theta - 0.246) + (0.2831)\delta(\theta - 0.563).$$

This posterior distribution followed from learning 3 of the 10 students identify as female combined with the prior belief that there was a 60% chance the students were from ISU.

This example is nice for illustrating hypothesis testing because baked into the problem were essentially two hypotheses:

$$H_0 : \theta = 0.246 \qquad \text{vs.} \qquad H_1 : \theta = 0.563$$

where the null hypothesis represents the belief that the students are from Rose-Hulman (where 24.6% of students identify as female) and the alterntative hypothesis represents the belief that the students are from ISU (where 56.3% of students identify as female). In this example, each hypothesis specifies a single value for the parameter; in general, each hypothesis could specify a region for the parameter. That is, we are generally interested in testing

$$H_0 : \theta \in \Theta_0 \qquad \text{vs.} \qquad H_1 : \theta \in \Theta_1.$$

While not a requirement, in practice $\Theta_0$ and $\Theta_1$ are typically mutually exclusive sets which form a partition of the parameter space.

> **Note**
>
> Nothing prohibits us from having more than two hypotheses in this framework, and nothing requires the hypotheses form a partition of the parameter space.

Under the Bayesian framework, it is straightforward to compute probabilities of the form "the probability $H_j$ is true given the data observed."

> **Note**
>
> Under the classical Frequentist perspective, the probability a hypothesis is true is nonsensical; however, under the Bayesian approach, we are using probability simply to characterize our uncertainty about the statement.

For the Naive Classification of College Students, the posterior naturally allows us to address the hypothesis. Notice that

$$\begin{aligned} Pr\left(H_0 \mid y\right) &= Pr(\theta = 0.246 \mid y) = 0.7169 \\ Pr\left(H_1 \mid y\right) &= Pr(\theta = 0.563 \mid y) = 0.2831. \end{aligned}$$

That is, after observing the data, we believe it is much more likely the the students are from Rose-Hulman than from ISU.

> **Note**
>
> Again, contrasting the conclusion in the Bayesian framework with a classical Frequentist framework, we are saying that the data is convincing us to believe the null hypothesis over the alternative. The idea of "accepting" the null hypothesis in the classical Frequentist framework is heretical.

In addition to computing the probability of each hypothesis, we can compute how much more likely we are to believe the students are from Rose-Hulman compared to believing they are from ISU:

$$\frac{Pr\left(H_0 \mid y\right)}{Pr\left(H_1 \mid y\right)} = 2.53;$$

that is, given the data, we believe it is 2.53 times more likely that the students are from Rose-Hulman than from ISU. This is known as the posterior odds.

**Definition 15.1** (Posterior Odds)**.** Let $H_j$ denote the hypothesis that $\theta \in \Theta_j$ for some region $\Theta_j$. Then, the posterior odds *in favor of* $H_j$ is given by

$$\frac{Pr\left(\theta \in \Theta_j \mid \mathbf{y}\right)}{Pr\left(\theta \notin \Theta_j \mid \mathbf{y}\right)}.$$

> **Note**
>
> When there are only two hypotheses, and they partition the parameter space, then the posterior odds captures how strongly we favor one hypothesis over the other given the data.

The posterior odds is useful, but it only suggests how we feel after we have observed the data. We may be more interested in how much the data *impacted* our prior beliefs. For example, if the data simply confirmed our beliefs, that is not nearly as extraordinary as the data completely reversing our beliefs. The Bayes Factor captures this impact.

**Definition 15.2** (Bayes Factor)**.** A measure of how the observed data *alters* your prior beliefs about a hypothesis. Let $H_j$ denote the hypothesis that $\theta \in \Theta_j$ for some region $\Theta_j$. The Bayes Factor *in favor of* $H_j$ is the ratio of the posterior odds in favor of $H_j$ to the prior odds in favor of $H_j$:

$$BF_j = \left(\frac{Pr\left(\theta \in \Theta_j \mid \mathbf{y}\right)}{Pr\left(\theta \notin \Theta_j \mid \mathbf{y}\right)}\right)\left(\frac{Pr\left(\theta \notin \Theta_j\right)}{Pr\left(\theta \in \Theta_j\right)}\right).$$

> **Note**
>
> Some prefer to report the logarithm of the Bayes Factor, as it is a quantity that is easier to work with in theoretical derivations.

Keep in mind, the Bayesian framework is about quantifying our uncertainty about the unknown parameters. The Bayes Factor measures how that uncertainty has been impacted by the observed data. Two independent papers made recommendations for how to interpret a Bayes Factor; however, we should remember that these are simply "rules of thumb."

Table 15.1: Rules of thumb for interpreting the Bayes Factor as suggested by Jeffreys and Kass and Raftery.

| Strength of Evidence | Jeffreys' Scale | Kass and Raftery Scale |
|---|---|---|
| Weak | $0 \leq \log_{10}(BF) < 0.5$ | $0 \leq \log(BF) < 1$ |
| Substantial | $0.5 \leq \log_{10}(BF) < 1$ | $1 \leq \log(BF) < 3$ |

| Strength of Evidence | Jeffreys' Scale | Kass and Raftery Scale |
|---|---|---|
| Strong | $1 \leq \log_{10}(BF) < 2$ | $3 \leq \log(BF) < 5$ |
| Decisive | $2 \leq \log_{10}(BF)$ | $5 \leq \log(BF)$ |

**Warning**

Take caution when interpreting a Bayes Factor; it quantifies the degree to which the data altered your prior beliefs. It is possible to have a really small Bayes Factor in favor of a hypothesis and yet simultaneously believe overwhelmingly in that hypothesis given the data; the small Bayes Factor simply implies that you believed in that hypothesis before collecting the data as well. Conversely, it is possible to have a really large Bayes Factor in favor of a hypothesis and yet simultaneously not distinguish between that hypothesis and another given the data; the large Bayes Factor simply implies that your beliefs about that hypothesis have dramatically changed.

For the Classification of College Students example (Example 10.1), our Bayes Factor, in favor of students being from Rose-Hulman, is given by

$$BF_0 = \left(\frac{0.7169}{0.2831}\right)\left(\frac{0.6}{0.4}\right) = 3.80.$$

The log-Bayes Factor is then 1.33, which falls under the "substantial evidence" category in the Kass and Raftery scale. This Bayes Factor is capturing the idea that we are still somewhat divided on the issue, but we have switched from favoring that the students are from ISU to favoring that the students are from Rose-Hulman. The data led to a shift in our beliefs.

## 15.1 Point-Null Hypotheses

Technically speaking, hypothesis testing in the Bayesian framework requires little; the posterior distribution allows us to quantify our uncertainty about parameters (or corresponding hypotheses) given the data. However, there is a common scenario which has a potential pitfall worth discussion. Consider testing the following hypotheses:

$$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0$$

where here $\Theta_0 = \{\theta_0\}$ is a singleton set. This is known as a "point-null hypothesis," and it poses a problem for many prior distributions. To illustrate, consider computing the prior odds in favor of $H_1$:

$$\frac{Pr\left(\theta \neq \theta_0\right)}{Pr\left(\theta = \theta_0\right)} = \frac{\int_{\theta \neq \theta_0} \pi(\theta) d\theta}{\int_{\theta = \theta_0} \pi(\theta) d\theta}. \tag{15.1}$$

For any continuous prior distribution, the numerator in Equation 15.1 will be 1 and the denominator will be 0! For any continuous distribution, the probability the random variable takes a specific value is 0; therefore, a continuous prior distribution is incompatible with a point-null hypothesis. The continuous prior distribution communicates you are infinitely more likely to believe the parameter takes any value other than $\theta_0$. There is a misalignment of beliefs; if you are truly interested in testing a point-null hypothesis, you must believe the null hypothesis has some probability of describing reality. Therefore, the prior distribution must incorporate the belief that

$$Pr\left(H_0\right) > 0$$

.

> **Big Idea**
>
> We can only test a hypothesis if, a priori, we believe there is some chance the hypothesis is true. If we go into a study believing something is impossible, no amount of data will convince us otherwise.

One way of incorporating our prior beliefs about a point-null hypothesis is specifying a mixture distribution.

**Definition 15.3** (Mixture Distribution)**.** Let $X$ be a random variable and $f(x)$ and $g(x)$ be valid density functions defined on a common support. Then,

$$h(x) = wf(x) + (1 - w)g(x),$$

where $0 < w < 1$, is known as a mixture distribution.

Appropriately applied, a mixture distribution allows us to place mass on the value of $\theta_0$ in a point-null hypothesis and spread out the remaining probability along the support.

> **Mixture Prior for Point-Null Hypotheses**
>
> Let $\theta$ be a parameter which has support $\Theta$, and consider testing the hypotheses
>
> $$H_0 : \theta = \theta_0 \qquad \text{vs.} \qquad H_1 : \theta \neq \theta_0$$
>
> for some $\theta_0 \in \Theta$. Suppose, a priori, we believe $Pr\left(\theta = \theta_0\right) = u$ for some $0 < u < 1$. Then,

$$\pi(\theta) = u\delta\left(\theta - \theta_0\right) + (1-u)\pi_0(\theta)$$

is a suitable family of prior distributions, where $\pi_0(\theta)$ is any continuous distribution on the support $\Theta$.

**Note**

A mixture prior with a point mass is not a continuous distribution, nor is it a discrete distribution.

**Example 15.1** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Suppose the hospital is interested in testing

$$H_0 : \theta = 0.304 \qquad \text{vs.} \qquad H_1 : \theta \neq 0.304.$$

The null hypothesis represents the C-section rate at the hospital being equivalent to the rate across the state of Indiana. Suppose we believe, prior to observing any data, either hypothesis is equally likely. Combining this belief with those expressed in Example 10.2, a reasonable prior distribution has the form

$$\pi(\theta) = 0.5\delta(\theta - 0.304) + 0.5\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\theta^{a-1}(1-\theta)^{b-1},$$

where $a = 10.3$ and $b = 23.6$. Note that the mass at $\theta_0 = 0.304$ led to different values of hyperparameters in $\pi_0(\theta)$ compared to Example 10.2. The resulting posterior will have the form

$$\pi(\theta \mid \mathbf{x}) = w\delta(\theta - 0.304) + (1-w)\frac{\Gamma(a+b+n+n\bar{x})}{\Gamma(a+n)\Gamma(b+n\bar{x})}\theta^{a+n-1}(1-\theta)^{b+n\bar{x}-1},$$

where

$$w = \frac{(0.5)(0.304)^n(1-0.304)^{n\bar{x}}}{(0.5)(0.304)^n(1-0.304)^{n\bar{x}} + (0.5)\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\frac{\Gamma(a+n)\Gamma(b+n\bar{x})}{\Gamma(a+b+n+n\bar{x})}},$$

the sample size $n$ and sample mean $\bar{x}$ are given by the data, and the hyperparameters $a$ and $b$ were defined above.

Using this posterior distribution, describe our belief about the hypotheses.

*Solution.* Notice that the form of the posterior distribution directly encodes our belief about the null hypothesis given the data. Specifically,

$$Pr\left(\theta = \theta_0 \mid \mathbf{x}\right) = w = 0.6098.$$

The posterior odds in favor of the null hypothesis is

$$\frac{Pr\left(\theta = \theta_0 \mid \mathbf{x}\right)}{Pr\left(\theta \neq \theta_0 \mid \mathbf{x}\right)} = 1.56,$$

and the Bayes Factor is the same (since the prior odds were 1). The data strengthened our belief in the null hypothesis, but not by much.

## 15.2 Model Comparison

Hypothesis testing is a special case of model comparison, where in the Bayesian framework the "model" consists of *both* the likelihood and the prior.

> **Big Idea**
>
> A Bayesian model consists of the choice for the likelihood as well as the choice for the prior.

Consider the Naive Classification of College Students (Example 10.1); we can reframe the problem as a choice between two models:

$$\begin{aligned}
\text{Model 0}: \quad & \pi(\theta) = \delta(\theta - 0.246) \\
& f(y \mid \theta) = \binom{n}{\theta} \theta^y (1-\theta)^{n-y} \\
\text{Model 1}: \quad & \pi(\theta) = \delta(\theta - 0.563) \\
& f(y \mid \theta) = \binom{n}{\theta} \theta^y (1-\theta)^{n-y},
\end{aligned}$$

where we believed, a priori, that $Pr(\text{Model 1}) = 0.4$ and $Pr(\text{Model 2}) = 0.6$. In this case, the models differed only in their choice of prior (which here simplifies further to which value of the parameter to select). This simplification allowed us to choose between these two models by working only with the posterior distribution.

We would like to generalize this process to allow the model to alter both the likelihood and the prior distribution, and for us to place some prior probability on the entirety of the model.

We are then interested in using the data to quantify the evidence for each model. To do this, we essentially consider the model $\mathcal{M}$ to be a parameter. This is known as a *hierarchical model* because the priors on the parameter are conditional on the model, and then a further prior is placed on the model itself (it is a multi-level model).

**!** Model Comparison

Let $\mathcal{M}_j$ represent the $j$-th potential model for a data generating process. Reflect the likelihood and prior as a function of the model. For example, write

$$\begin{aligned} \text{Model 0}: \quad & f_0(\mathbf{y} \mid \theta_0, \mathcal{M}_0) \\ & \pi_0(\theta_0 \mid \mathcal{M}_0) \\ & \pi(\mathcal{M}_0) = Pr(\mathcal{M}_0) \\ \text{Model 1}: \quad & f_1(\mathbf{y} \mid \theta_1, \mathcal{M}_1) \\ & \pi_1(\theta_1 \mid \mathcal{M}_1) \\ & \pi(\mathcal{M}_1) = Pr(\mathcal{M}_1). \end{aligned}$$

Notice we (potentially) allow the form of the likelihood, the parameters governing that likelihood, and the form of the prior to differ for each model. Our prior beliefs about the parameter are captured within each prior, but we also have prior beliefs about the model itself — how likely each model is. We are interested in determining $Pr(\mathcal{M}_j \mid \mathbf{Y})$ for each $j$.

An iterated application of Bayes Theorem allows us to write the likelihood of each model, given the data, as

$$\begin{aligned} Pr\left(\mathcal{M}_j \mid \mathbf{Y}\right) &= \frac{f_j(\mathbf{y} \mid \mathcal{M}_j) Pr(\mathcal{M}_j)}{\sum_j f_j(\mathbf{y} \mid \mathcal{M}_j) Pr(\mathcal{M}_j)} \\ f_j(\mathbf{y} \mid \mathcal{M}_j) &= \int f_j(\mathbf{y} \mid \theta_j, \mathcal{M}_j) \pi_j(\theta_j \mid \mathcal{M}_j) d\theta_j. \end{aligned}$$

**Definition 15.4** (Evidence for a Model)**.** Under the Model Comparison framework defined above, the evidence for model $\mathcal{M}_j$ is defined as

$$f_j(\mathbf{y} \mid \mathcal{M}_j) = \int f_j(\mathbf{y} \mid \theta_j, \mathcal{M}_j) \pi_j(\theta_j \mid \mathcal{M}_j) d\theta_j.$$

The evidence for a model is a number; since it is a function only of observed data, once we observe the data, the evidence is a constant. We can also think of the evidence as the prior predictive distribution under a particular model evaluated at the observed data.

> **Warning**
>
> It may seem strange to use the prior predictive distribution when defining the evidence instead of the posterior predictive, but keep in mind that within model comparison, the *model itself* is the parameter of interest. Changing either the likelihood or the prior will impact the evidence.

The evidence can also be used to compute the Bayes Factor for one model over another.

**Definition 15.5** (Bayes Factor for Model Comparison)**.** The Bayes Factor, in favor of Model 1, is

$$\begin{aligned} BF_1 &= \left(\frac{Pr(\mathcal{M}_1 \mid \mathbf{y})}{Pr(\mathcal{M}_0 \mid \mathbf{y})}\right)\left(\frac{Pr(\mathcal{M}_0)}{Pr(\mathcal{M}_1)}\right) \\ &= \left(\frac{f_1(\mathbf{y} \mid \mathcal{M}_1)Pr(\mathcal{M}_1)}{f_0(\mathbf{y} \mid \mathcal{M}_0)Pr(\mathcal{M}_0)}\right)\left(\frac{Pr(\mathcal{M}_0)}{Pr(\mathcal{M}_1)}\right) \\ &= \frac{f_1(\mathbf{y} \mid \mathcal{M}_1)}{f_0(\mathbf{y} \mid \mathcal{M}_0)}. \end{aligned}$$

That is, the Bayes Factor is a ratio of the evidence for each model.

Model comparison simply extends our framework by the inclusion of another parameter. That parameter, the model itself, just happens to be discrete. Our goal is to use our machinery to quantify the uncertainty in each model given the data observed.

# 16 Constructing Prior Distributions

The selection of a prior distribution is integral to the Bayesian framework; it is also the most criticized component. There is rarely sufficient prior information to determine an exact prior distribution; that is, rarely do we know for certain the family which represents the distribution as well as the exact parameters. Instead, we make some modeling assumptions, as with any analysis. In this chapter, we examine some common paths when constructing prior distributions and the implications of allowing the prior distribution to vary across analysts.

## 16.1 Elicitation from Experts

Ideally, the prior distribution would not be arbitrary but guided by experts. Example 10.2, for example, illustrated the use of statements from experts to form a parametric approximation to the prior information. We elicited information about the uncertainty in order to determine values for the hyperparameters — those values that determine the specific shape of the prior distribution.

For Example 10.2, the prior distribution chosen was a conjugate prior.

**Definition 16.1** (Conjugate Prior)**.** A prior distribution chosen such that the posterior distribution belongs to the same family as the prior distribution, with the (hyper)parameters that govern the family updated based on the observed data.

In Example 10.2, we chose a Beta distribution to represent the prior, and we found in Example 11.2, the posterior distribution also belonged to the Beta family. The choice to use a conjugate prior was often done historically in order to simplify computation in an era where computing power was limited. In the era of higher-speed computing, this is no longer necessary.

One argument for the use of conjugate priors is that the form is invariant to the data; that is, the data is restricted in what it can say about the unknown parameter. The data can update our beliefs, but it cannot update the family which encodes those beliefs.

While we will not go into details here, it is *almost* always possible to construct a conjugate prior. And, if chosen carefully, that prior can approximate nearly any prior information given.

> **Note**
>
> The posterior distribution is always a combination of the prior distribution and the likelihood. Conjugate priors make that very clear. As the sample gets large, the prior distribution is swamped by the data; and, as the sample size increases, Bayesian inference agrees with Frequentist inference.

Again, Example 10.2 illustrated combining the expert opinions provided with a parametric family in order to construct a prior distribution. When we are unable to determine a suitable parametric approximation for the prior beliefs provided, a "histogram approach" is possible.

**Definition 16.2** (Histogram Approach to Constructing a Prior)**.** Using expert information, attach probability to various intervals for the parameter. Specifically,

- Define $m$ intervals $\left(\theta_{j-1}, \theta_j\right)$ for $j = 1, 2, \ldots, m$ that partition the parameter space; define $\theta_0$ as the lower bound of the support for the parameter, and define $\theta_m$ as the upper bound of the support for the parameter.
- Eliciting expert opinions, assign probability $\pi_j$ to each interval: $\pi_j = Pr\left(\theta_{j-1} < \theta < \theta_j\right)$ for each $j = 1, 2, \ldots, m$.
- Set the prior $\pi(\theta)$ to be the piecewise distribution over this interval where $\sum_{j=1}^{m} \pi_j = 1$.

There have been critiques of eliciting information from experts. Estimates given may be biased, due to the current availability of data on which the experts are making their informed decisions. We tend to be overconfident in our opinions or go with our initial reaction instead of allowing our beliefs to be updated. We also tend to want to create the prior only after observing the data, despite the fact that the prior should capture our beliefs prior to observing the data.

The experts may not actually represent a reasonable sample to capture widespread prior belief. How does one determine who is expertly qualified to speak on a particular topic? How do you rank levels of expertise?

We mention these critiques because more important than the choices we make is that those choices be clearly documented. It is okay to construct work that others critique; that is how science develops. No study is perfect, and being able to identify and own the limitations of our study and analysis is critical to the development of knowledge.

## 16.2 Mixture Priors

Chapter 15 introduced the idea of a mixture distribution (Definition 15.3) for constructing a prior. While we discussed its use for a particular case, mixture distributions have wider applicability. Suppose we would like to work with a parametric approximation, but we cannot

find a parametric family which captures the structure suggested by the prior information. In these cases, combining multiple distributions may be appropriate.

As an example, suppose we have a parameter on the support $(0, 1)$. For such a parameter, it is natural to consider the Beta distribution for a prior. However, suppose our prior beliefs suggest a multimodal distribution similar to that in Figure 16.1; it is impossible to choose hyperparameters for a Beta distribution that would result in such a prior. Instead, we could achieve such a prior by mixing two Beta distributions.

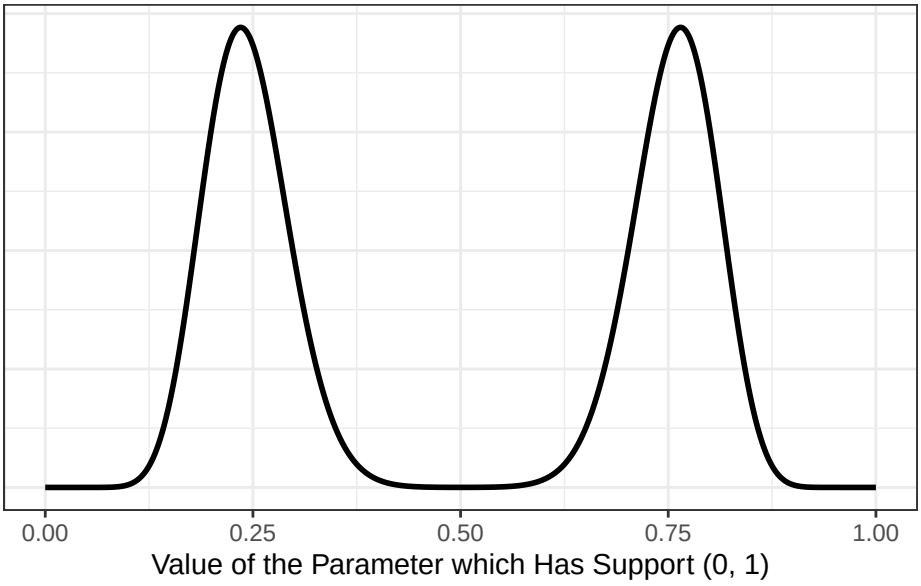


Figure 16.1: Illustration of a mixture prior for a parameter on the interval (0, 1).

**Definition 16.3** (General Mixture Distribution)**.** Let $\theta$ be a parameter with support $\Theta$, and let $\pi_k(\theta)$ be a valid distribution on the support, for $k = 1, 2, \ldots, K$. Then,

$$\pi(\theta) = \sum_{k=1}^{K} w_k \pi_k(\theta)$$

is a valid prior distribution provided $\sum_{k=1}^{K} w_k = 1$.

It turns out that if each component of the mixture distribution $\pi_k(\theta)$ is a member of the conjugate family, the entire prior will be conjugate (a weighted average of distributions). This was illustrated in Example 15.1.

Nothing requires that the individual components of a mixture distribution be of the same family. For example, we might choose to mix a Normal distribution with a t-distribution in order to capture the presence of some outliers.

> **Note**
>
> While we have described the use of a mixture distribution for the prior distribution, nothing prevents the use from using a mixture distribution for the likelihood.

It has been shown that any distribution can be approximated by some mixture distribution. That is, if we choose $K$ to be large enough, we can approximate any distributional shape with a mixture distribution.

## 16.3 Chains

Within this unit, we have developed the fundamental concepts of Bayesian inference in a general setting. We have avoided a litany of examples and instead opted to illustrate the concepts with a single unifying example throughout the text (Example 9.1). In both this example and the exposition in the text, we have acted as though there is a single parameter $\theta$ governing the likelihood. Many interesting questions, however, involve models for the data that depend upon multiple parameters. These types of problems often necessitate the need for numerical solutions, which we address in the next unit. Here, we simply discuss a common tool for constructing priors over multiple parameters.

When $\theta$ is a parameter *vector*, then $\pi(\theta)$ is actually a *joint* density across all parameters. Therefore, one key decision that must be made is whether, a priori, we believe these parameters to be independent of one another.

**Example 16.1** (Heights of Children)**.** During early development, children are regularly benchmarked against national growth charts. One such chart traces a child's height as they grow. However, these charts were developed using the entire population of "healthy" children. Suppose I am interested in developing a growth chart for children with Hispanic parents, as I believe they tend to be a bit shorter, on average. It is typical to model heights using a Normal distribution, which has two unknown parameters (which govern the location and spread of the distribution).

We consider developing a likelihood and prior for this process.

The likelihood for the above example is readily available if we are willing to assume a random sample of $n$ children (of the same age) born to Hispanic parents:

$$\begin{aligned} f(\mathbf{y} \mid \mu, \tau) &= \prod_{i=1}^{n} \frac{\tau^{1/2}}{\sqrt{2\pi}} e^{-\tau/2(y_i-\mu)^2} \\ &= \frac{\tau^{n/2}}{(2\pi)^{n/2}} e^{-(\tau/2)\sum_{i=1}^{n}(y_i-\mu)^2} \end{aligned}$$

where we have defined the likelihood in terms of the mean $\mu$ and the *precision* $\tau$, which is the inverse of the variance. The likelihood was simplified by assuming the height of one child is independent of the height of any other child. Suppose we are further willing to believe the *parameters* are independent of one another; then, it is reasonable to propose the prior distributions independently. Choosing a Normal prior for $\mu$ and a Gamma prior for $\tau$, we could then propose

$$\begin{aligned}
\pi(\mu) &= \frac{\sqrt{b}}{\sqrt{2\pi}} e^{-b/2(\mu-a)^2} \\
\pi(\tau) &= \frac{s^r}{\Gamma(r)} \tau^{r-1} e^{-s\tau} \\
\Rightarrow \pi(\mu, \tau) &= \pi(\mu)\pi(\tau).
\end{aligned}$$

The joint prior across the parameters is easy to specify because of the independence assumption. Of course, nothing requires we assume the parameters are independent of one another; this was a modeling assumption. A different set of beliefs would lead to a different structure for the prior. For example, the prior given by

$$\begin{aligned}
\pi(\tau) &= \frac{s^2}{\Gamma(r)} \tau^{r-1} e^{-s\tau} \\
\pi(\mu \mid \tau) &= \frac{\sqrt{\tau}}{\sqrt{2\pi}} e^{-\tau/2(\mu-a)^2} \\
\Rightarrow \pi(\mu, \tau) &= \pi(\mu \mid \tau)\pi(\tau)
\end{aligned}$$

suggests a Gamma prior for $\tau$ and a Normal prior for $\mu$ *conditional* on the value of $\tau$. This hierarchical structure allows the mean $\mu$ to depend on the precision $\tau$. The joint distribution of the parameters (prior to seeing the data) is then the product of the marginal distribution of $\tau$ and the conditional distribution of $\mu \mid \tau$.

This process of defining a prior in stages, each stage conditioning on parameters for which a prior distribution is specified, is known as "chaining."

Regardless of whether we form a prior through assuming independence or through chaining, the prior predictive distribution (the denominator in Bayes Theorem) will have the form

$$\int \int f(\mathbf{y} \mid \mu, \tau)\pi(\mu, \tau) d\mu d\tau.$$

The more parameters we have, the more complex the integration in the denominator; this is what motivates the computational methods we examine in the next unit.

## 16.4 Non-Informative Priors

Each of the above sections assumes that we have prior information that needs to be encoded into a distribution; we now consider the case when we have very little (or no) prior information. In such a setting, we must determine how we encode "ignorance."

**Definition 16.4** (Laplace Prior)**.** The Laplace prior, also known as a "flat" prior, considers the form

$$\pi(\theta) = 1 \qquad \forall \theta \in \Theta.$$

> ⚠ Warning
>
> For any unbounded support $\Theta$, the Laplace prior will be *improper*; that is,
>
> $$\int \pi(\theta) d\theta = \infty.$$
>
> In such settings the Laplace prior is not actually a valid density function. This seems like it is breaking all the rules, and to some degree it is, but it is still commonly used.

> ⚠ Warning
>
> The Bayes Factor should never be computed when you have an improper prior as the prior odds are not defined since there is no valid probability of each hypothesis a priori.

The Laplace prior is a common default prior when no (or little) prior information is available. However, a flat prior cannot represent total ignorance. When the parameter space is unbounded, notice that a flat prior essentially says that no matter how large of a value $q$ you imagine, $Pr(\theta > q) = \infty$ — that is, there is always infinite probability that $\theta$ is larger than you can imagine.

> 💡 Big Idea
>
> Flat priors are chosen because they are easily overwhelmed by the observed data.

The benefit of a flat prior is that the posterior distribution is proportional to the likelihood. The idea here is to make the Bayesian framework dependent solely on the data, similar to a Frequentist approach (though the two are still not guaranteed to give the same results).

Flat priors are a subset of a larger class of priors that continues to be an active area of research; this larger class of priors is determined solely by the form of the likelihood.

**Definition 16.5** (Noninformative Prior)**.** A prior distribution which is derived solely based on the form of the likelihood.

A noninformative prior seems like a compromise between Bayesians and those who dislike the subjective nature of a prior distribution. So, why is this not the standard? On the one hand, true Bayesians argue that we should make use of prior information; we should not seek to make use of only the data available in that single study. This allows the information from one study to become the prior information for a follow-up study instead of beginning from scratch. Second, there is a potential pitfall when using noninformative priors when they are improper — it is possible for the posterior distribution to be improper (which is a nice way of saying it is not a distribution at all)! If the posterior distribution is improper, it cannot yield any valid inference on the parameters.

> **Warning**
>
> Valid inference cannot be made when the posterior distribution is improper.

An improper prior *can* lead to an improper posterior; however, a proper prior will *always* lead to a proper posterior. The danger is that software which automates Bayesian analyses currently have no way of checking if a posterior is proper; so, this must be done manually. As computing the posterior can be difficult (the entire reason for the next unit), this often involves bounding the integral in some way — a job for true mathematicians.

Fear around improper priors often leads to what are known as vague priors. This is taking a parametric family and choosing the hyperparameters to result in a massive variance so that the prior distribution, while proper, appears flat over the parameter space. The idea here is to allow the data to easily overwhelm any prior information.

> **Big Idea**
>
> Noninformative priors try to make it easy for the likelihood to dominate the prior distribution in the computation of the posterior distribution.

# Part IV

# Unit IV: Numerical Approaches to Bayesian Computations

The previous unit discussed the fundamental components of the Bayesian paradigm. We began by specifying the likelihood, a model describing the data generating process as a function of unknown parameters, and a prior distribution, which captures our beliefs about the unknown parameters prior to observing data. Using Bayes Theorem, we were able to characterize our beliefs about the unknown parameter after observing the data in the posterior distribution. We know that the posterior distribution is proportional to the product of the likelihood and the prior; that is,

$$\pi(\theta \mid \mathbf{x}) \propto f(\mathbf{x} \mid \theta)\pi(\theta).$$

However, to determine the exact form of the posterior distribution, we need to determine the scaling constant given by

$$\frac{1}{\int f(\mathbf{x} \mid \theta)\pi(\theta)d\theta}.$$

Unfortunately, this integral can be intractable, especially as the dimension of $\theta$ grows. In this unit, we consider numerical approaches for common Bayesian quantities, such as point and interval estimation. With these techniques, we can address more intricate problems.

# 17 Monte Carlo Integration

Integration plays an integral part (pun completely intended) in Bayesian inference. Integration is necessary to determine the scaling term for the posterior distribution; it is necessary to perform point and interval estimation; it is necessary for computing posterior probabilities needed for hypotheses testing; integration is at center of all Bayesian computations. The integrals we need to compute are also typically analytically intractable. This chapter introduces a numerical integration technique popular within the statistical community that serves as the foundation for modern Bayesian computational approaches we will consider in the remainder of this unit.

Example 17.1 illustrates the need for numerical integration techniques in practice.

**Example 17.1** (Pharmacokinetics)**.** When modeling the rate at which a drug is broken down by the body (known as pharmacokinetics), it is often of interest to know the logarithm of the ED50 value (the time at which 50% of the drug has been metabolized by the body). Suppose it is known that for a particular drug, the ED50 value $T$ for the patient population can be modeled using a Gamma distribution with a shape parameter $a = 4$ and a rate parameter of $b = 2$. That is, the density of $T$ is given by

$$f(t) = \frac{b^a}{\Gamma(a)} t^{a-1} e^{-bt}$$

on the interval $t > 0$, where $a = 4$ and $b = 2$. We are interested in the average logarithm of the ED50 value, which is given by

$$\int \log(t) f(t) dt.$$

While this example is contrived (the parameters governing the distribution of the ED50 value are known explicitly), integrals like this arise in practice regularly. The integral of interest does not yield an analytical solution; a numerical procedure must be used. While there are several numerical methods for integration which are developed and studied in mathematics, many are restricted to low dimensional problems. We need a technique which is applicable in high dimensions. As a foundation, consider the following suggested procedure:

- Let $T_1, T_2, \ldots, T_n$ be a random sample such that $T_i \sim f(t)$ for all $i$.

- Define $Y_i = \log(T_i)$ for all $i$.
- Compute $\bar{Y}$.

The value of $\bar{Y}$ will estimate the integral! As an initial justification, observe that on average, the estimate is correct. That is, consider taking the expected value of $\bar{Y}$; specifically,

$$\begin{aligned} E(\bar{Y}) &= E\left(n^{-1}\sum_{i=1}^{n} Y_i\right) \\ &= E\left(n^{-1}\sum_{i=1}^{n} \log(T_i)\right) \\ &= n^{-1}\sum_{i=1}^{n} E(\log(T_i)), \end{aligned} \tag{17.1}$$

where the last line is a result of the expectation being a linear operator. Since each $T_i$ is identically distributed (we have a random sample from the population), the expectation is the same for each $i$; specifically, we have (see Definition 2.7)

$$E(\log(T_i)) = \int \log(t) f(t) dt.$$

Since the right-hand side is not indexed by $i$, substituting back into Equation 17.1, we have that

$$E(\bar{Y}) = \int \log(t) f(t) dt.$$

That is, the expected value of $\bar{Y}$ is the desired integral. Of course, this just states that the distribution of $\bar{Y}$ (across many samples of size $n$) will center on the true value of the integral; it does not guarantee the average we compute for any specific $n$ will be accurate. However, the Law of Large numbers gives a stronger result.

## 17.1 Law of Large Numbers

Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed random variables with density $f(x)$. Consider a real valued function $g$. Then, for any $\epsilon > 0$ we have that

$$Pr\left(|\frac{1}{n}\sum_{i=1}^{n} g(X_i) - E[g(X)]| > \epsilon\right) \to 0$$

as $n \to \infty$, assuming $E\left[g(X)\right]$ exists.

The Law of Large Numbers essentially says that the sample mean can be made arbitrarily close to the expectation it approximates given a large enough sample size. That is, as the sample size increases, the sample mean is really close to the true mean. Using mathematical notation, this means

$$\frac{1}{n}\sum_{i=1}^{n} g\left(X_i\right) \approx \int g(x)f(x)dx.$$

> **Big Idea**
>
> The Law of Large Numbers allows us to approximate integrals, in the form of expectations of random variables, using a corresponding sample mean.

In practice, limited resources (namely, time and money) limit the size of the sample we can consider in our study, which in turn limits the applicability of the Law of Large Numbers. However, numerical integration makes use of a computational study (all the "data" is generated within the computer), where speed and cost are greatly reduced. Therefore, the Law of Large Numbers is more applicable to such computational studies.

The following pseudo-code illustrates the use of the Law of Large Numbers in order to compute the integral that corresponds to the average logarithm of the ED50 value in Example 17.1:

```
let m = 10000;

for i from 1 to m do;
  x[i] = random_gamma(shape = 4, rate = 2);
  y[i] = log(x[i]);
end;

ybar = mean(y[1:m]);
```

Since we have a large sample size, we know that the sample mean computed in the last step will be a good approximation to the integral of interest.

Upon first inspection, it may seem that the Law of Large Numbers is limited to estimating means. However, nearly every quantity of interest can be written in terms of an integral and therefore approximated using this technique. This includes probabilities, means, quantiles (and therefore credible intervals), variances, and even the evidence in favor of a model. That is, the Law of Large Numbers provides a technique for performing integration numerically, allowing us to compute summaries for the posterior distribution. Because of its dependence on random processes, this integration technique is known as Monte Carlo (or MC) Integration.

**Definition 17.1** (Monte Carlo Integration)**.** Consider an integral of the form

$$\int_{\mathcal{S}} g(x)f(x)dx$$

where $f(x)$ is a valid density function for a random variable $X$ with support $\mathcal{S}$. Then, the following algorithm, known as Monte Carlo (or MC) Integration, gives a numerical approximation to the integral:

1. Take a random sample $X_1, X_2, \ldots, X_m$ such that $X_i \sim f(x)$ for all $i$, where $m$ is large.
2. Compute $m^{-1} \sum_{i=1}^{m} g(X_i)$.

By the Law of Large Numbers,

$$\frac{1}{m} \sum_{i=1}^{m} g(X_i) \approx \int_{\mathcal{S}} g(x)f(x)dx.$$

> Big Idea
>
> We can construct Bayesian estimators using only a random sample from the posterior distribution.

**Example 17.2** (Estimating a Posterior Probability)**.** Assume that our beliefs about an unknown parameter $\theta$ (with support on the real line), given an observed sample $\mathbf{x}$, is characterized by the posterior distribution $\pi(\theta \mid \mathbf{x})$. Derive a Monte Carlo Integration technique for estimating the probability

$$Pr(\theta > q \mid \mathbf{x}) = \int_{q}^{\infty} \pi(\theta \mid \mathbf{x})d\theta.$$

*Solution.* First, we recognize that the probability of interest is naturally an integral. As currently written, however, the integral is not over the entire support of $\theta$. Note, however, that we can rewrite the integral to be over the entire support. Specifically, observe that

$$\begin{aligned} Pr(\theta > q \mid \mathbf{x}) &= \int_{q}^{\infty} \pi(\theta \mid \mathbf{x})d\theta \\ &= \int_{-\infty}^{q} 0 \cdot \pi(\theta \mid \mathbf{x})d\theta + \int_{q}^{\infty} 1 \cdot \pi(\theta \mid \mathbf{x})d\theta \\ &= \int_{-\infty}^{\infty} \mathbb{I}(\theta > q)\pi(\theta \mid \mathbf{x})d\theta, \end{aligned}$$

where $\mathbb{I}(u)$ is the "indicator function" taking the value 1 if $u$ occurs and 0 otherwise. Specifically, in this case

$$\mathbb{I}(q < \theta) = \begin{cases} 1 & \text{if } \theta > q \\ 0 & \text{if } \theta \leq q. \end{cases}$$

Having rewritten the integral of interest, we now recognize that

$$Pr(\theta > q \mid \mathbf{x}) = E\left[\mathbb{I}(\theta > q) \mid \mathbf{x}\right].$$

That is, the posterior probability of interest is the posterior mean of the quantity $\mathbb{I}(\theta > q)$.

Applying the Law of Large Numbers, our MC Integration technique is defined by the following algorithm:

1. Take a random sample $\theta_1^*, \theta_2^*, \ldots, \theta_m^*$ from the posterior distribution, $\theta_i^* \sim \pi(\theta \mid \mathbf{x})$ for all $i$ where $m$ is large.
2. Compute $m^{-1} \sum_{i=1}^{m} \mathbb{I}\left(\theta_i^* > q\right)$.

This sample mean (which essentially computes the proportion of the $\theta^*$ values which exceed $q$) approximates the integral of interest. Further, according to the Law of Large Numbers, $m$ can be chosen to make the approximation as precise as desired.

In practice, there are multiple unknown parameters; therefore, the posterior distribution is a *joint density*. As a result, when applying MC Integration in practice, we must often generate random variates from a joint distribution. In general, such generation is difficult to perform directly; however, the process is much simpler if we can create a set of chained expressions to guide the generation. Suppose $(X, Y)$ is a vector of two random variables with joint density $f(x, y)$; recall that

$$f(x, y) = f(x \mid y) f(y);$$

that is, the joint distribution is the product of a conditional distribution and a marginal distribution. This suggests a procedure for generating from a joint distribution.

**Simulating from a Joint Distribution**

Let the random vector $(X, Y)$ be distributed according to the joint density $f(x, y)$. The following procedure can be used to simulate observations $(X_1, Y_1), (X_2, Y_2), \ldots, (X_m, Y_m)$ from the joint density:

1. Generate $m$ variates $Y_1, Y_2, \ldots, Y_m$ from $f(y)$, the marginal distribution of $Y$.
2. For each $Y_i$, generate a single $X_i$ from $f(x \mid y)$, the conditional distribution of $X$

given $Y$.

The resulting pairs will have the desired joint distribution. Further, the column of $X$'s will behave according to the density $f(x)$, the marginal distribution of $X$.

The Law of Large Numbers guarantees that we can obtain approximations of Bayesian estimators; further, it guarantees these approximations can be made arbitrarily good by choosing a large enough sample size $m$. Of course, in practice we will specify the value of $m$; and, since MC Integration relies on random processes, it is reasonable to ask how good the resulting approximation is. Similarly, it is natural to ask how many random samples are needed for an approximation with a specific amount of precision. This is addressed via a version of the Central Limit Theorem.

**Theorem 17.1** (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots, X_m$ be independent and identically distributed random variables. Consider a real-valued function $g$ such that $E[g(X)]$ and $Var[g(X)]$ exist. Then, we have that the quantity*

$$\frac{\sqrt{m}\left[m^{-1}\sum_{i=1}^{m} g(X_i) - E[g(X)]\right]}{\sqrt{\widehat{Var}[g(X)]}}$$

*behaves like a standard normal random variable as $m \to \infty$.*

The Central Limit Theorem provides a way of quantifying the error in the Monte Carlo approximation.

**Definition 17.2** (Monte Carlo Error)**.** Also called the standard error for an approximation of the form $m^{-1}\sum_{k=1}^{m} g(X_k)$, the MC error is given by

$$\sqrt{\frac{1}{m(m-1)}\sum_{k=1}^{m}\left[g(X_k) - \frac{1}{m}\sum_{j=1}^{m} g(X_j)\right]^2}$$

which is the sample standard deviation of the generated variates divided by the square root of the number of replications.

We close this chapter by revisiting Example 12.1; instead of an analytical solution, we apply the techniques discussed in this chapter to provide a numeric solution to estimating the unknown parameter.

**Example 17.3** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Example 11.2 found the posterior distribution to be

$$\theta \mid \mathbf{x} \sim Beta\left(n + a, n\bar{x} + b\right)$$

where $a = 17, b = 39, n = 15$ and $n\bar{x} = 33$ given the observed data.
Example 12.1 showed that the posterior mean, estimating the rate of C-sections at the hospital given the observed data, is given by 0.308. First, use MC Integration to compute this estimate, and then estimate the *odds* of a C-section at the hospital given the observed data.

*Solution.* The posterior mean is given by

$$E\left(\theta \mid \mathbf{x}\right) = \int \theta\pi(\theta \mid \mathbf{x})d\theta.$$

Now, as seen in Example 12.1, there is a closed-form expression for this quantity. However, here, we rely on MC Integration technique. The following pseudo-code illustrates the process:

```
let m = 10000;

for i from 1 to m do;
  x[i] = random_beta(shape1 = (15 + 17), shape2 = (33 + 39));
end;

xbar = mean(x[1:m]);
```

The quantity `xbar` is an estimate of the posterior mean. This code requires we have access to a function `random_beta()` which generates a (pseudo) random variate from a Beta distribution with specified shape parameters. Given a function `standard_deviation()` which computes the sample standard deviation of a vector of values, we can compute the MC error in the approximation with

```
MCerrorx = sqrt(standard_deviation(x[1:m]) / m);
```

One implementation of this algorithm gave an estimate of 0.308 with an MC error of 0.0021. While this computation simply illustrated the process, the next is computation greatly simplifies our burden of work.

We now consider estimating the odds of a C-section at this hospital, given the data. The odds of an event with rate $\theta$ are defined as $\theta/(1-\theta)$. Appealing to Definition 2.7, the posterior mean of the odds is then

$$E\left[\frac{\theta}{1-\theta} \mid \mathbf{x}\right] = \int \frac{\theta}{1-\theta}\pi(\theta \mid \mathbf{x})d\theta.$$

A closed-form solution for this integral is not readily available; however, a slight modification of our previous pseudo-code allows us to numerically compute this integral:

```
let m = 10000;

for i from 1 to m do;
  x[i] = random_beta(shape1 = (15 + 17), shape2 = (33 + 39));
  y[i] = x[i] / (1 - x[i]);
end;

ybar = mean(y[1:m]);
MCerrory = sqrt(standard_deviation(y[1:m] / m));
```

The quantity `ybar` is an estimate of the posterior mean for the odds of interest. One implementation of this algorithm gave an estimate of 0.451 with an MC error of 0.0031.

> **Big Idea**
>
> Given a large random sample from the posterior distribution, we approximate posterior quantities of interest using the corresponding sample statistic.

# 18 Markov Chain Monte Carlo (MCMC)

The previous chapter laid the foundation for Bayesian computations: given a random sample from the posterior distribution, we can approximate any Bayesian estimator, and the level of approximation is only limited by the size of the sample we can take. When the posterior distribution has the form of a known distribution, there are often built-in functions for obtaining a (pseudo) random sample. The majority of applications, however, involve posterior distributions which cannot be derived analytically. In such cases, Markov Chain Monte Carlo (MCMC) techniques are used to sample from the posterior distribution. Once we have a sample from the posterior distribution, we apply the MC Integration techniques from the previous chapter for computing estimates. In this chapter, we give a brief overview of MCMC techniques.

The technical details of MCMC can be overwhelming; we begin by discussing the conceptual goal of the algorithm, and we do that through Example 18.1.

**Example 18.1** (Relative Wealth (MCMC Concepts))**.** Suppose you are attending a large gathering of your relatives. Given your current status as a student (without no income), you would like to network with these relatives in hopes that they will write you into their will, ensuring you a sizable inheritance. Naturally, you would like to be strategic about how much time you spend with each person, dividing your time with each person proportionally according to their wealth (spend more time with wealthier individuals). However, you face a couple of problems.

- Problem 1: you are uncertain about how many people are actually attending the party.
- Problem 2: you have no way of determining the wealth of each person.

However, while no one is willing to share their actual wealth, each person is willing to share a bit of information with you. If you show interest in speaking with them, the individual will let you know their wealth relative to that of the person you are currently speaking with ("I am half as rich as the person you are speaking with," for example).

We want an algorithm for determining who to speak with, and for how long to speak with them.

While Example 18.1 is a toy problem, it illustrates the obstacles we are trying to overcome with MCMC methods. We want to take a sample from the posterior distribution, but we do not have the form of the posterior distribution. Instead, we only know the kernel:

$$\pi(\theta \mid \mathbf{y}) \propto f(\mathbf{y} \mid \theta)\pi(\theta).$$

The first "problem" in Example 18.1 reflects our uncertainty about where we should focus our attention within the posterior. We have a sense of the support of the posterior, but there may be large areas of the posterior which have essentially probability 0 of occurring. The second "problem" is that we do not have the value of the posterior; instead, we are only able to compute the posterior up to some scalar constant. Therefore, while we are not able to compute the value of the posterior, we can accurately determine the ratio of the posterior between two points. Observe that

$$\pi(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{m(\mathbf{y})},$$

where $m(\mathbf{y})$ represents the prior predictive distribution (see Definition 14.1). Now, we have that the ratio of the posterior evaluated at $\theta = a$, relative to the posterior evaluated at $\theta = b$ is given by

$$\begin{aligned}\frac{\pi(\theta = a \mid \mathbf{y})}{\pi(\theta = b \mid \mathbf{y})} &= \frac{(1/m(\mathbf{y}))f(\mathbf{y} \mid \theta = a)\pi(\theta = a)}{(1/m(\mathbf{y}))f(\mathbf{y} \mid \theta = b)\pi(\theta b)} \\ &= \frac{f(\mathbf{y} \mid \theta = a)\pi(\theta = a)}{f(\mathbf{y} \mid \theta = b)\pi(\theta = b)}.\end{aligned}$$

That is, the ratio of the posterior evaluated at two points can be determined using only the kernel of the distribution. We now consider an algorithm for talking to your relatives that addresses Example 18.1.

*Solution.* Consider the following scheme. You randomly choose a person to begin speaking with (your "partner"). After a fixed period of time, you will flip a coin. If the coin is "heads up," you will consider speaking with the person to your partner's right; if the coin is "tails up," you will consider speaking with the person to your partner's left. This determines your "candidate."

You ask the candidate about their wealth relative to your current partner. If your candidate is wealthier than your partner, you will definitely move to them and make them your new partner. If your candidate is less wealthy than your partner, however, you will only move to them with a probability equivalent to the candidate's wealth relative to your current partner's. That is, if the candidate is half as wealthy as your current partner, you will move to the candidate with probability 0.5.

You then repeat this process many times.

This process summarizes the idea behind MCMC methods. We generate a candidate value of the parameter; if the value of the posterior at the candidate value is higher than our current position, then we choose to move to the candidate point. Otherwise, we move to the candidate point with a probability equal to the ratio of the posterior for the candidate relative to our current position. We move through the parameter space in this fashion until we have generated a large sample from the posterior (say 3000 replicates).

The above thought exercise illustrates one of the simplest algorithms for generating from an unknown density, known as the Metropolis Algorithm. While in practice this algorithm is rarely implemented directly, it forms the basis of several more complex algorithms, and it illustrates the basic properties of all MCMC methods.

**Definition 18.1** (Metropolis Algorithm)**.** Suppose we want to generate random variates from the density $\pi(\theta \mid \mathbf{y})$. We perform the following steps:

1. Generate an initial value $\theta^{(0)}$.

2. At the $k$-th step, generate $\theta^*$ (a candidate) according to a symmetric proposal density $q\left(\theta \mid \theta^{(k-1)}\right)$.

3. Compute $A\left(\theta^*, \theta^{(k-1)}\right)$ where

$$A\left(\theta^*, \theta^{(k-1)}\right) = \frac{\pi\left(\theta^* \mid \mathbf{y}\right)}{\pi\left(\theta^{(k-1)} \mid \mathbf{y}\right)} = \frac{f\left(\mathbf{y} \mid \theta^*\right)\pi\left(\theta^*\right)}{f\left(\mathbf{y} \mid \theta^{(k-1)}\right)\pi\left(\theta^{(k-1)}\right)}.$$

4. Generate $U \sim Unif(0,1)$. If $U \leq A\left(\theta^*, \theta^{(k-1)}\right)$, then set $\theta^{(k)} = \theta^*$; else, set $\theta^{(k)} = \theta^{(k-1)}$.
5. Repeat Steps 2-4 $m$ times, for some large $m$.

When generating an initial value, $\theta^{(0)}$, we could choose $\theta^{(0)} \sim \pi(\theta)$ if the prior is easy to generate from. While it is common to choose $q(\cdot)$ to be a Normal distribution with mean $\theta^{(k-1)}$, it is not a requirement to do so; when a Normal distribution is used, it can be difficult to determine a reasonable variance (too large, and you drift too far away; too small, and you do not move at all).

The Metropolis Algorithm is typically run several thousand times, resulting in what is known as a Markov Chain.

**Definition 18.2** (Markov Chain)**.** A sequence of random vectors $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ is a Markov Chain with stationary transition probabilities if for any set $A$ and any $k \leq n$

$$\begin{aligned} Pr\left(\theta^{(k)} \in A \mid \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k-1)}\right) &= Pr\left(\theta^{(k)} \in A \mid \theta^{(k-1)}\right) \\ &= \int_A q\left(\theta^{(k)} \mid \theta^{(k-1)}\right) d\theta^{(k)} \end{aligned}$$

where $q$ is called the transition kernel.

In general, Markov chains are not required to have stationary transition probabilities, but it is a nice simplifying assumption that is applied in Bayesian computing methods. Markov chains are a topic of interest in and of themselves in probability theory and are beyond the scope of this text. We primarily focus on the fact that the probability that a value is in some region depends *only* on the previous state; the remaining "history" (previous states in the chain) is unimportant.

The Markov Chain is essentially a sample; of course, for it to be useful, we need to establish that the chain is also representative of the appropriate target distribution. That is, we need to know the Markov Chain represents the posterior distribution. This target distribution has a name in Markov Chain literature — the "stationary distribution."

**Definition 18.3** (Stationary Distribution)**.** Let $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ be a Markov Chain. The stationary distribution of the Markov Chain is the distribution $p(\theta)$ such that

$$Pr\left(\theta^{(k)} \in A\right) = \int_A p(\theta) d\theta.$$

The stationary distribution is where the Markov Chain settles down so that the probability that any variate is in some region is computed using that same distribution $p$ (as opposed to the transition kernel). You can think of it as the limit of the transition kernel as $k$ gets large.

The idea is to choose a proposal density $q$ in the Metropolis Algorithm such that the stationary distribution of the resulting Markov Chain, if one exists, will be the posterior distribution of interest. So, we want to pick a proposal density $q$ that is easy to generate from, is symmetric, and so that eventually, the values we are generating behave as if they were drawn from the posterior distribution. The machinery needed to prove such results is beyond the scope of our text, but it can be shown that the Metropolis Algorithm produces a Markov Chain for which the stationary distribution is the same as the posterior distribution, provided that the proposal density is symmetric. That is, as $k$ increases, the values generated by this algorithm behave as if they were drawn from the posterior distribution. More, we can show that this is true regardless of the choice of the starting value $\theta^{(0)}$.

> **i** Note
>
> It can be shown that the the Metropolis Algorithm has the posterior distribution as the stationary distribution of the Markov Chain, but the Metropolis Algorithm is not always the most efficient algorithm for generating from the posterior distribution. In practice, other algorithms, such as the Gibbs sampler or Hamiltonian Monte Carlo, improve on the efficiency by implementing modifications to the above Metropolis Algorithm.

We now have a way of generating values which have properties similar to random variates drawn from the posterior distribution. The language here is chosen with care because, as always, there is a catch: **the values generated in the Markov Chain are identically distributed, but are not independent.**

In practice, analysts often state that they have obtained a random sample from the posterior using MCMC methods; and, we can often proceed as if that were true. However, when we use MCMC methods, the resulting sample is not truly a "random sample" in the sense of being IID. The points are identically distributed, but since each variate in the sample was generated based on the value of the previous variate, there is a dependence between the variates. Fortunately, this dependence between values generated consecutively (known as autocorrelation) is often negligible in practice.

The real issue is that the theory introduced in Chapter 17 relied on having a random sample (IID variates). Therefore, the fact that the Markov Chain generated by the Metropolis Algorithm does not result in independent observations seems to imply that we are unable to rely on MC Integration. Fortunately, however, there is a Law of Large Numbers-type result for Markov Chains which says

$$\lim_{m\to\infty} \frac{1}{m} \sum_{k=1}^{m} g\left(\theta^{(k)}\right) = \int g(\theta)p(\theta)d\theta$$

where $g(\cdot)$ is some real-valued function and $p(\theta)$ is the stationary distribution of the Markov Chain, if it exists. This result says that even though the points are related, as we take a large number of them, we can approximate integrals (therefore, Bayesian estimators) using sample averages. That is, we can apply MC Integration techniques to Markov Chains.

> Big Idea
>
> For a sufficiently large number of replications, the Markov Chain resulting from an MCMC algorithm will behave as a random sample from the posterior distribution.

## 18.1 Hamiltonian Monte Carlo

The trick to MCMC methods is to choose a transition kernel that is efficient and can handle the myriad of situations encountered in practice. The Metropolis Algorithm, while simplistic, is not very efficient, and can be quite difficult to implement when the dimension of the parameter vector increases. A commonly implemented alternative is known as the Gibbs sampler. This is implemented in the popular Bayesian software packages BUGS (Bayesian inference Using Gibbs Sampling) and JAGS (Just Another Gibbs Sampler). BUGS is a standalone software package while JAGS is implemented in other computing languages (like R and Python). These

software packages provide a myriad of algorithms based on the Gibbs sampler which address hierarchical models in a nice way. However, in some complex models, these algorithms can be inefficient or fail to produce variates from the posterior. Stan implements a Hamiltonian Monte Carlo (HMC) algorithm which can succeed in these situations. While the details of the algorithm are beyond the scope of this text, we discuss the ways in which HMC improves upon the Metropolis Algorithm discussed above.

The Metropolis Algorithm can be summarized in the following two statements:

- Choose the candidate point using a symmetric proposal distribution centered on the current point.
- Favor points with a larger corresponding posterior density, moving to candidate points with lower posterior density probabilistically.

The key distinction between the Metropolis Algorithm and HMC is to allow the proposal distribution to be dependent upon our current location.

> Big Idea
>
> HMC uses proposal distributions which favor moving toward the posterior mode.

The idea is illustrated in Figure 18.1, created by John Kruschke (Kruschke 2015), which shows how proposals are generated for two different initial values. Note, the end result in this graphic is not a sample from the posterior but the distribution of potential next steps.

To illustrate how this works, consider two different current positions within a posterior distribution. The Metropolis Algorithm would simply say to generate proposals which are symmetric about the current position. HMC generates proposals that are closer to the posterior mode (as evidenced by the bottom part of the figure where the majority of proposals are near the mode). In order to determine where to move from the current position, the HMC algorithm considers the *potential* of the position, defined through the negative log-density. The potential gives an idea of how far we might want to travel (the potential of the position to change).

**Definition 18.4** (Potential)**.** The potential of a value $\theta$ is the negative logarithm of the posterior evaluated at $\theta$. In practice, we need only know the potential up to a constant. That is, it suffices to define the potential as

$$\text{Potential}(\theta) = -\log\left[f(\mathbf{y} \mid \theta)\pi(\theta)\right].$$

While we have described the potential as a value, since it exists for all $\theta$ in the support, we can think of the potential as a function (row 2 of Figure 18.1). Now, imagine the current position is a ball on the potential; the proposed position is determined by flicking the ball randomly. This random "flick" is done by selecting a random variable from a Standard Normal distribution, which determines both the magnitude and direction of the flick (negative values move the

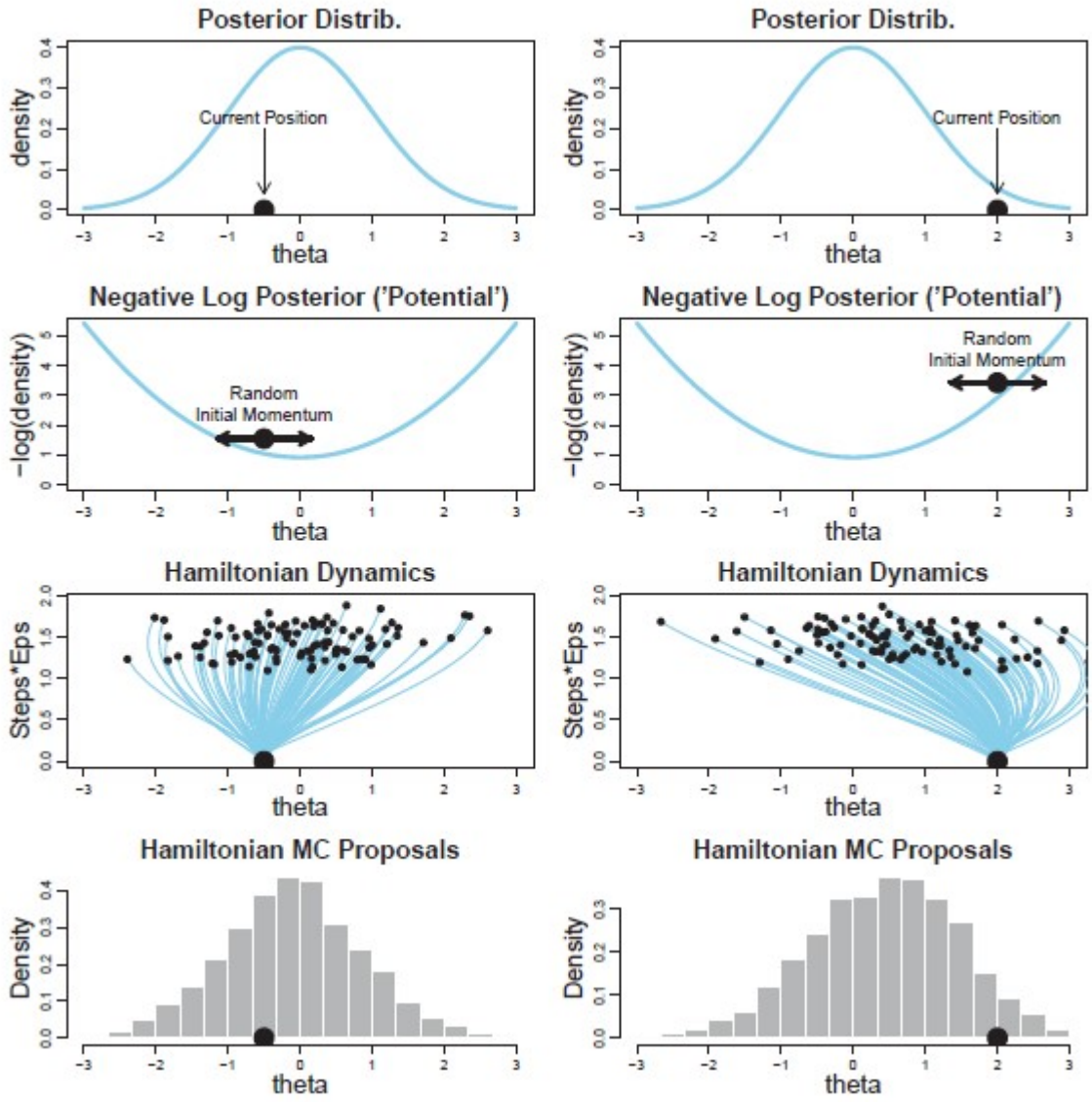


Figure 18.1: Examples of Hamiltonian Monte Carlo proposal distributions. Two columns show two different current parameter values, marked by a large dot. The first row shows the posterior distribution. The second row shows the potential energy, with a random impulse given to the dot. The third row shows trajectories, which are the theta value (x-axis) as a function of time (y-axis marked Steps*Eps). The fourth row shows histograms of the proposals.

ball to the left, and positive values move the ball to the right). We then watch the ball roll around for a while. Wherever the ball stops is the proposed position. This is illustrated in the third row of graphics in Figure 18.1 that show how the ball moves over time to the proposed position.

> **Note**
>
> The sum of potential and kinetic energy is known as the Hamiltonian (hence the name of this procedure). The total energy should be conserved at each point in the algorithm.

As the ball rolls around on the potential, it will naturally be drawn to lower points on this surface. That is, candidate points will tend to be drawn from regions with lower potential, corresponding to regions with a higher posterior density. Notice that when the current position is near the posterior mode, the potential positions are nearly symmetric about the current location as in the Metropolis Algorithm. But, if the current position is far from the posterior mode, the potential positions are drawn from regions closer to the posterior mode and the potential positions are far from the current position.

> **Big Idea**
>
> The HMC algorithm generates proposals which tend to have lower potential.

We emphasize that these are just *candidate* positions. Once a candidate position is identified, we must decide whether to move there or remain in the current position, just as we do in the Metropolis Algorithm.

> **Decision Rule for HMC:**
>
> Generate $U \sim Unif(0,1)$ and $A\left(\theta^*, \theta^{(k-1)}\right)$ where
>
> $$A\left(\theta^*, \theta^{(k-1)}\right) = \frac{f\left(\mathbf{y} \mid \theta^*\right)\pi\left(\theta^*\right)\omega\left(\theta^*\right)}{f\left(\mathbf{y} \mid \theta^{(k-1)}\right)\pi\left(\theta^{(k-1)}\right)\omega\left(\theta^{(k-1)}\right)}$$
>
> and $\omega(\cdot)$ is the momentum. If $U \leq A\left(\theta^*, \theta^{(k-1)}\right)$, then we move to the new position; otherwise, we remain in the same position.

The momentum can be thought of as how much speed the ball has when you reach the candidate position (remember, we stop the ball not when it comes to rest but after some fixed amount of time). Recall that we apply a random momentum to the current location of the ball. The aspect we want to emphasize here is that the decision rule is quite similar to the Metropolis Algorithm.

We have described this process as letting the "ball roll around" for some fixed set of time. In practice, we emulate this by taking some predefined number of steps of a certain size based

on the gradient (much like numeric function minimization). Both the step size and number of steps require some tuning. The step size is tuned to balance how far away from the current position we move and the degree of approximation. If we take small steps, we approximate the curve quite nicely, but we do not get anywhere. If we take large steps, we move away from our current position, but the approximation suffers. The total duration (the number of steps taken) is tuned to ensure we do not overshoot or make a u-turn. If we let the ball roll for too long, it could overshoot the posterior mode by a large degree; or, we may end up stopping the ball when it has rolled back to where it started. Figure 18.2, created by John Kruschke (Kruschke 2015), illustrates the impact of allowing the "time" (number of steps and length of step size) to be too large; notice the difference in the distribution of candidate points in Figure 18.2 compared to Figure 18.1.

In addition to these tuning parameters, we must determine the standard deviation of the symmetric distribution used to apply the momentum to the current position. This choice needs to balance variety with accuracy. Too small of a standard deviation (like nudging the ball) means it will not roll far from where it started, and every candidate is essentially the same (leading to a higher likelihood of acceptance/rejection). Too large of a standard deviation, and a high degree of candidates will be rejected. Figure 18.3, created by John Kruschke (Kruschke 2015), illustrates the impact of standard deviation in the proposal distribution; notice the difference in the distribution of candidate points between the two columns in Figure 18.3.

Proper tuning ensures that the algorithm is efficient and a majority of the variates are useful in representing the posterior. These are handled internally by the software, but it is important to have an understanding of what is happening in the background.

With MCMC methods, we can address a multitude of more complex problems. We do note the one limitation of Stan is that it does not currently support discrete parameters directly. This is because the HMC algorithm needs a smooth function in order to compute the gradient. Not supporting discrete parameters is not as limiting as it might seem, but it does prohibit automatic model comparison within Stan and eliminates the ability to put a point mass in the prior distribution.

> **Warning**
>
> While you could write custom implementations of the HMC (or any MCMC) algorithm, software like Stan does the hard work for you. However, in order to make use of those tools, you must specify the Bayesian model (the likelihood and the prior) in addition to providing the data. This can require your learning a new "probability language" (as opposed to a computing language) for specifying such models. Some software packages have pre-built functions/interfaces for commonly specified models allowing you to get started more quickly.

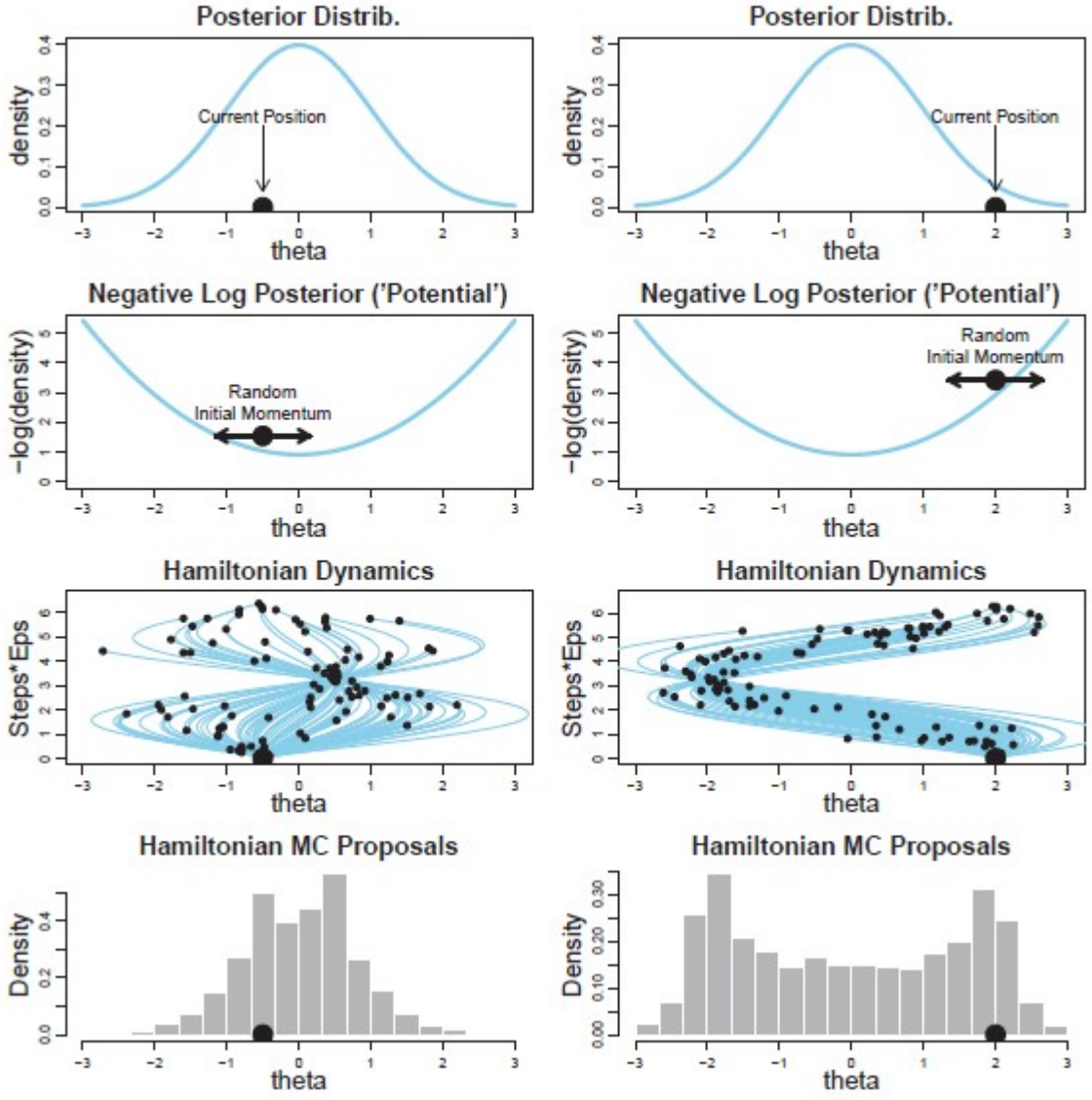


Figure 18.2: Examples of Hamiltonian Monte Carlo proposal distributions for two different current parameter values, marked by the large dots, in the two columns. For this figure, a large range of random trajectory lengths (Steps*Eps) is sampled.

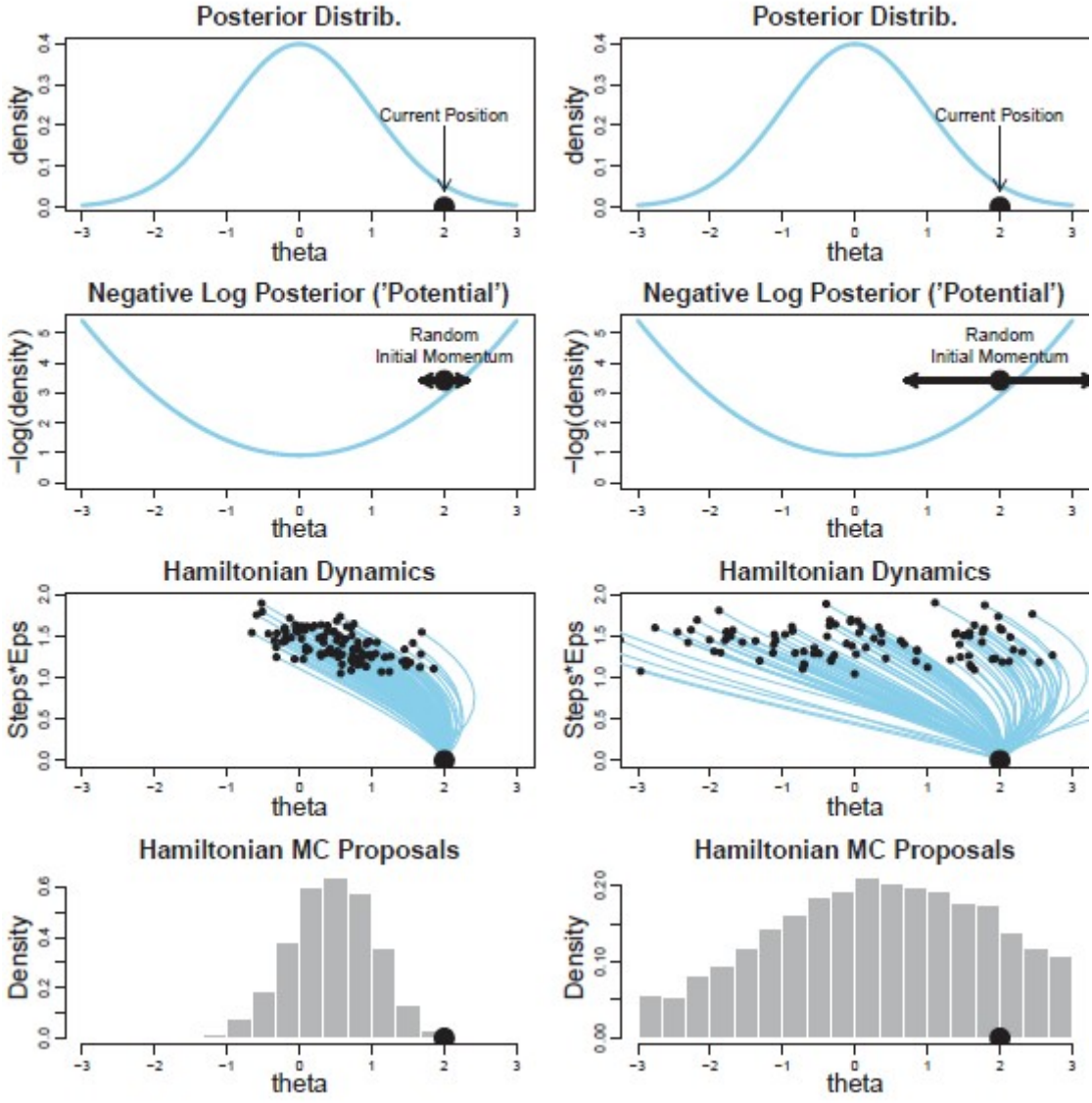


Figure 18.3: Examples of a Hamiltonian Monte Carlo proposal distributions for two different variances of the initial random momentum, indicated in the second row.

# 19 Assessing MCMC Samples

The reliability of any statistical analysis depends on the quality of the data obtained; as a result, any good analysis requires that we give some thought to the data we have obtained. Similarly, we must consider the derivation of our prior distribution and the reasonableness of our model for the likelihood. When our analysis also includes the use of an MCMC algorithm, we should, at a minimum, also investigate that certain assumptions about the resulting sample are reasonable before proceeding. This chapter briefly discusses some checks that are done on the posterior sample to determine its suitability for answering questions.

There are essentially four considerations when examining the output of any MCMC algorithm.

**! Assessment of an MCMC Algorithm**

Before using a sample from an MCMC algorithm, the following should be considered:

1. The posterior distribution is proper.
2. The resulting Markov Chains converged.
3. Sensitivity of the algorithm to starting values.
4. The correlation between generated variates is negligible.

Software which implements MCMC algorithms typically generate output for assessing the reliability of the resulting sample. In this chapter, we focus on navigating this output for assessment.

An improper posterior distribution cannot provide valid inference. Unfortunately, an MCMC algorithm will generate a sequence of values, even if the target distribution is improper. If the combination of the likelihood and prior specified would result in an improper posterior, the resulting sample from the MCMC algorithm is useless. In general, software is unable to determine if the target distribution is improper; therefore, it is up to the analyst to analytically determine if the posterior distribution is proper. The simplest way to ensure that we have a proper posterior distribution is to use a proper prior distribution.

**i Ensuring a Proper Posterior**

If we use a proper prior, we are guaranteed a proper posterior.

**Example 19.1** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

While MCMC methods are not required for this example, as Example 11.2 illustrates, it is possible to use and MCMC algorithm to address this question. Before even writing MCMC code, explain why we can be confident that the posterior is proper in this case.

*Solution.* Our prior information was represented using a Beta distribution, which is a proper distribution. Therefore, we know the posterior will also be proper.

Recall that we "seed" an MCMC algorithm with an initial value. Only after a large number of iterations is the algorithm generated values from the stationary distribution — the distribution to which the process essentially converges. Therefore, we must ensure we have allowed the algorithm to run long enough that the process has converged to the stationary distribution — that the variates generate behave as if they are drawn from the posterior distribution. A *trace plot* showing the value of the generated variates at each step of the algorithm can be used to visually assess convergence.

You are essentially looking for whether the chain eventually settles in a particular region of the parameter space; this should not be confused with the chain reaching a specific value. As we are looking for a stationary distribution, we expect the variates generated to bounce around (according to the stationary distribution); however, if the chain has some signal to it (trending in location or spread), that is unexpected. It should eventually look like noise around some central point.

Often, we eliminate the early part of a chain, the discarded portion known as the "warm-up" (or "burn-in") period. That is, we might remove the first 1000 variates from the resulting Markov chain because we expect the chain is still moving toward the stationary distribution during this time period. The variates after the burn-in period behave more like a sample from the stationary distribution and are retained for analysis.

> **i** Graphically Assessing Convergence of MCMC Chains
>
> Create a plot of the values generated (after eliminated the values from the burn-in period) against the order in which they were generated. This is known as a "trace plot." If the chain has converged, the plot should not have any trends in the location or spread over time.

**Example 19.2** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

We used Stan to implement a Hamiltonian Monte Carlo to obtain a sample from the posterior distribution. We generated three chains (each seeded with a different initial value); each chain

Table 19.1: Summary of results from MCMC algorithm used to estimate the rate of C-sections at a hospital. The model was fit with Stan; 3 chains were generated, each with 5000 iterations and a burn-in of 2000 for a total of 9000 post-burn-in variates. The 95% credible interval reported is an equal-tailed interval.

| Posterior Mean | Posterior Median | 95% Credible Interval | ESS | Shrink Ratio |
|---|---|---|---|---|
| 0.307 | 0.306 | (0.224, 0.397) | 3495.544 | 1.001 |

had 5000 iterations, with the first 2000 representing a burn-in. This provided a sample of size 9000 from the posterior distribution after combining the three chains.

Table 19.1 summarizes the sample generated by the MCMC algorithm. Figure 19.1 is a trace plot for each of the three chains. Comment on the convergence of each chain.

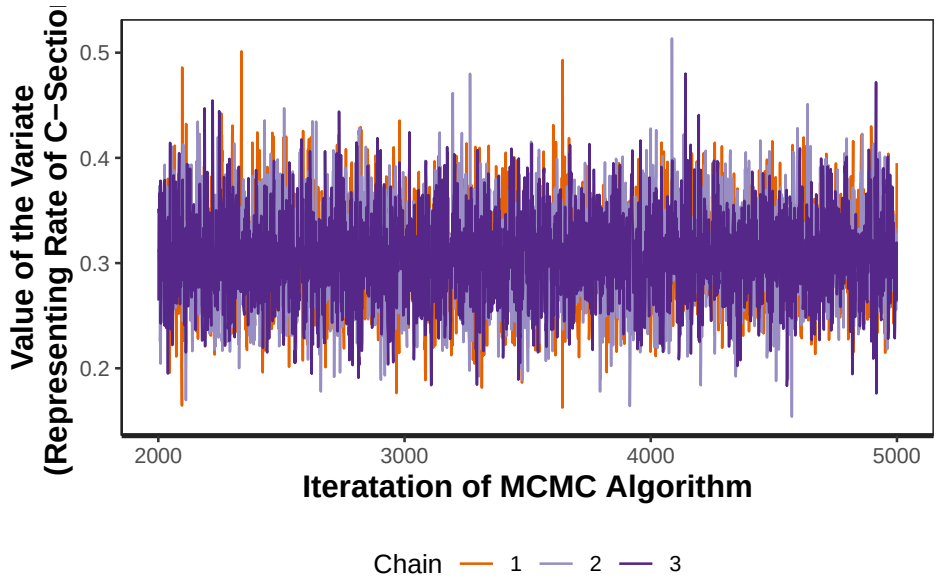


Figure 19.1: Traceplot from an example estimating the rate of C-sections at a hospital. The burn-in of 2000 iterates is removed from the graphic.

*Solution.* We do not notice any trends in the location or spread of any of the three chains after the burn-in period. Each of the chains seems to bounce around the value 0.3, and the spread stays relatively constant (values generated tend to be between 0.2 and 0.4).

As there are no trends in the location or spread, the samples generated by the algorithm are consistent with what we would expect if they had reached the stationary distribution.

In theory, the stationary distribution is the posterior distribution, we just need to run the algorithm long enough to get there. Practically, however, the distribution to which the algorithm converges could depend on the value used to seed the process. If that happens, then any results based on the sample are potentially biased. Therefore, we want to determine if the MCMC algorithm is sensitive to the chosen starting (initial) value. To do so, we seed the algorithm with multiple starting values, resulting in multiple chains. It is generally sufficient to consider three chains. Overlaying the trace plot from each chain allows us to assess whether the chains "mix" well. If the various chains are distinct, this suggests that the stationary distribution suggested by the algorithm varies according to the starting value, which means a stationary distribution was not really obtained.

In addition to the visual check, we can compute the "shrink factor." Also known as the "potential scale reduction factor," this is the ratio of the between-chain variability to the within-chain variability. If the chains are well mixed, this ratio should be near 1. If the ratio gets much larger than 1.1, it indicates a serious problem with the mixing.

**Assessing Sensitivity of MCMC Algorithm to Starting Value**

Seed the algorithm with three initial values. If the trace plots of the resulting chains are distinct — occupy different aspects of the parameter space — then your results are sensitive to the starting value. If the chains mix well, the chains are then combined to form a single sample for estimation.
Alternatively, a shrink factor above 1.1 indicates sensitivity to the starting value.

**Example 19.3** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Revisiting Example 19.2 above, assess the sensitivity of the algorithm to the initial value.

*Solution.* Notice that in Figure 19.1, the three chains overlap completely; in fact, it is difficult to distinguish one chain from another. This suggests the chains are mixing well as they occupy the same part of the parameter space.

Reported in Table 19.1, the shrink factor was estimated to be 1, which is consistent with our observations in the graphic above. There is no evidence the algorithm is sensitive to the initial value, and it seems reasonable to combine the variates from the three chains.

Recall that we expect our Markov chain to result in correlated variates. As a result, each variate does not contain as much unique information as we may believe. The "effective sample size" gives a crude measure of how much *independent* information there is within the chain. For example, we may generate 5000 iterates, but if the variates are highly dependent, the effective sample size may suggest we act as if only 100 iterates were generated.

**Definition 19.1** (Effective Sample Size)**.** The effective sample size (ESS) is given by

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} ACF(k)}$$

where ACF is the auto-correlation function of degree $k$.

> **Warning**
>
> When discussing MCMC procedures, it is common to use "sample size" to actually refer to the number of variates generated during the MCMC algorithm. We should not confuse the number of variates generated from the posterior with the number of observations in our data set.

If you want to measure something that is toward the center of the distribution (mean/median), the ESS need not be large. But, if you want to compute a tail probability (such as for a credible interval), you need a much larger ESS. Some texts recommend near 10000 variates from the posterior in order to reliably compute a highest posterior density interval, for example.

> **Assessing Independence of Variates**
>
> The effective sample size (ESS) takes into account the correlation between the variates and gives you an indication of how precise your results are. A small ESS suggests high correlation between the variates and indicates that your results are less reliable.

**Example 19.4** (C-section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital.

Revisiting Example 19.2, assess the independence of the variates.

*Solution.* We see that while we generated 9000 (post burn-in) variates, the effective sample size is estimated to be just under 3500. This suggests that there is a relatively strong correlation between the observations; only every third variate generated is independent. However, the effective sample size still remains fairly large; so, we are confident in both our point estimates and any credible intervals generated.

While these four checks are somewhat universal, there is an additional check that can be helpful when using a Hamiltonian Monte Carlo (HMC) algorithm. Recall that the total energy should be conserved during the HMC procedure. A *divergent transition* occurs when the simulated Hamiltonian departs from the true value (as measured at the initial point). Divergent transitions (after warm-up) indicate the results will be biased. A "pairs plot" allows us to

visualize when this occurs. If the amount of error (divergence) is larger than the median, it can often be fixed by increasing the target acceptance rate. If not, then this indicates that the posterior may be very difficult to sample from for this algorithm.

> **Assessing Divergent Transitions**
>
> For the HMC algorithm only, a pairs plot allows us to determine if the amount of divergence is larger than expected. If there are issues, try increasing the target acceptance rate.

With regard to Example 19.2, no divergent iterations were noted among the 9000 variates generated.

Posterior checks of the MCMC samples are necessary in order to prevent making conclusions that are unreasonable. We have only discussed how to identify problems. Fixing the problems is often dependent upon the specific application. There are several subtleties with each model that take time to learn in order to understand where the algorithm might get hung-up. The fix is often a clever reparameterization of the likelihood or prior in order to help the algorithm. With enough computation, we can overcome many of the problems faced.

# Part V

# Unit V: Hierarchical Models for Comparing Groups

Unit III introduced the fundamentals of the Bayesian framework. While the text introduced these ideas in a general setting, the running example through that unit considered modeling a single response variable. In this unit, we consider comparing a response across groups. While the framework remains the same, we discuss additional considerations to be made in these settings.

# 20 Elements of Good Study Design

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.

## 20.1 Two Types of Studies

In order to illustrate how study design can impact the results, consider the following example.

**Example 20.1** (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. 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.

Within this context, 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 Kangaroo 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 situations 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 groups being compared 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 an example of a **controlled experiment**.

**Definition 20.1** (Observational Study)**.** A study in which each subject "self-selects" into one of groups being compared in the study. The phrase "self-selects" is used very loosely here and can include studies for which the groups are defined by an inherent characteristic or are chosen haphazardly.

**Definition 20.2** (Controlled Experiment)**.** A study in which each subject is *randomly* assigned to one of the groups being compared in the study.

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

To understand the impact of random allocation, suppose that we had conducted a study using design (A); further, the results of the study 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 are given Kangaroo Care cry for less time, on average, following the Vitamin K shot.

(2) It turns out that those nurses who choose to implement Kangaroo Care (remember, they have a choice under design (A) whether they implement the method) are also the nurses with a gentler bedside manner. Therefore, these nurses tend to be very gentle when giving the Vitamin K shot whereas the nurses who choose not to implement Kangaroo Care tend to jab the needle when giving the shot. 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 correct under study design (A). 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 an *association* 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 (that is, the reduced crying time may be due to something else, like the bedside manner of the nurse). In this hypothetical scenario, the nurse's bedside manner is called a **confounder**.

**Definition 20.3** (Confounding)**.** When the effect of a variable on the response is misrepresented due to the presence of a third, potentially unobserved, variable known as a confounder.

> **Note**
>
> While both result in estimates we may not trust, confounding is not equivalent to a biased sample.

Confounders can mask the relationship between the factor under study and the response. Did you know there is a documented association between ice cream sales and the risk of shark attacks? As ice cream sales increase, the risk of a shark attack also tends to increase. 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 20.1 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, thereby increasing the risk of a shark attack. The two variables, ice cream sales and shark attacks, appear to be related as a result of the confounder of temperature.

> **Big Idea**
>
> 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

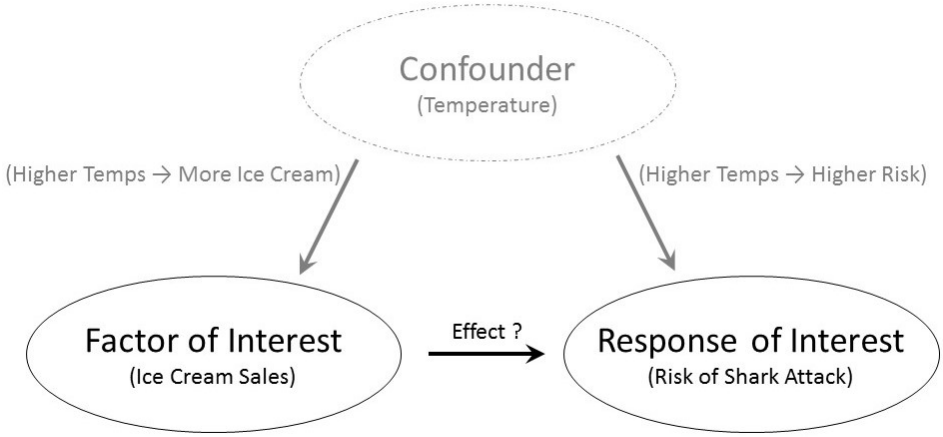


Figure 20.1: Illustration of a confounding variable. The confounder, related to both the response and the factor of interest (or treatment) can make it appear as though there is a causal relationship when none exists.

from the data. Controlled experiments allow for causal interpretations because the random allocation to the levels of the factor removes the impact of confounders. Let's return to the hypothetical Vitamin-K study in Example 20.1. Suppose there are nurses with a gentle bedside manner and those who are a little less gentle. If we randomly determine which infants receive Kangaroo Care, then for every gentle nurse who is told to implement Kangaroo Care while giving the shot, there tends to be a gentle nurse who is told to not implement Kangaroo Care. Similarly, for every less-gentle nurse who is told to implement Kangaroo Care while giving a shot, there tends to be a less-gentle nurse who is told to not implement Kangaroo Care. This is illustrated in Figure 20.2. For an observational study, the treatment groups can be unbalanced; for example, Figure 20.2 illustrates a case in which 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 tend to be balanced with respect to this confounder; 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, since we have eliminated all other differences between the groups (other than the treatment they receive), any differences we observe between the groups *must* be due to the grouping and not an underlying confounding variable.

**Big Idea**

Randomly assigning subjects to groups balances the groups with respect to any confounders; that is, the groups being compared are similar. Therefore, any differences between the two groups can be attributed to the grouping itself, leading to cause-and-

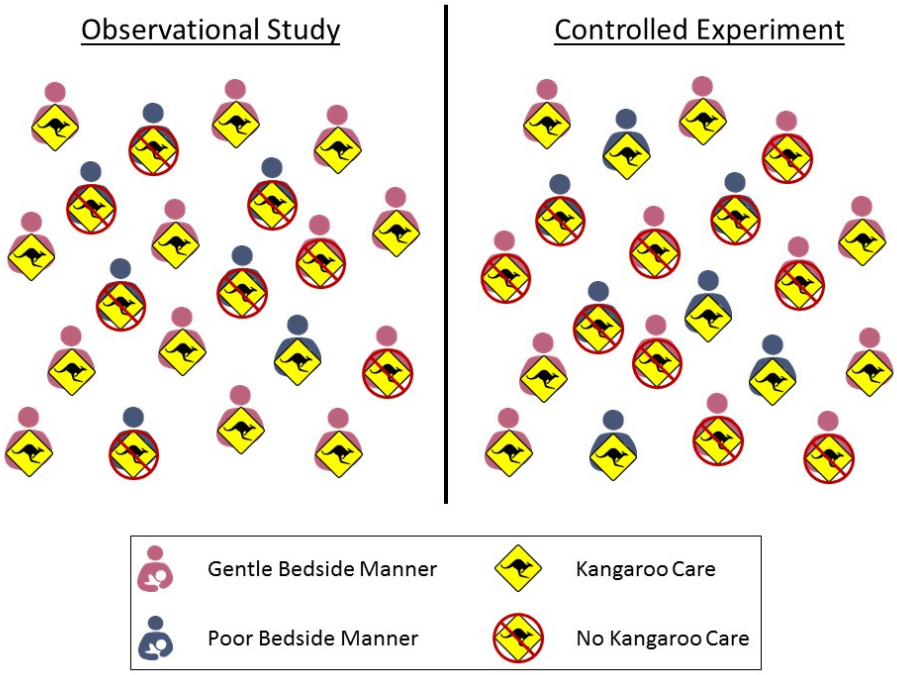


Figure 20.2: Illustration of the impact of random assignment in study design. For the observational study, the treatment groups are unbalanced. For the controlled experiment, the treatment groups are balanced.

effect conclusions.

While controlled experiments are a fantastic study design, we should not discount the use of observational studies. Consider the Deepwater Horizon Case Study described in Chapter 4; 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 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 acknowledge, and potentially address, the potential confounding when performing the analysis and discussing the results.

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

## 20.2 Aspects of a Well-Designed Study

While controlled experiments are a useful tool, there are many aspects to consider when designing a study. Generally speaking, there are three components to a well-designed study: replication, randomization, and reduction of extraneous noise.

**Warning**

A study is not poor just because it lacks one of these elements. That is, a study can provide meaningful insights even if it did not make use of each of these elements; every study is unique and should be designed to address the research objective. These elements are simply helpful in creating study designs.

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 inherent variability in the data generating process.

**Definition 20.4** (Replication)**.** Replication results from taking measurements on different units (or subjects), for which you expect the results to be similar. That is, any variability across the units is due to natural variability within the population.

> ⚠ Warning
>
> The term "replication" is also used in the context of discussing whether the results of a study are replicable. While our use of the term is about replicating a measurement process within a study, this does not downplay the importance of replicating an entire study.

When we talk about gathering "more data," we typically mean obtaining a larger number of replicates. Ideally, replicates will be obtained through *random selection* 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) when performing a controlled experiment. This random allocation breaks the link between the factor and any potential confounders, allowing for causal interpretations. However, random allocation preserves any link between the factor and the response, if a link exists. These are the two aspects of **randomization**.

**Definition 20.5** (Randomization)**.** Randomization can refer to random *selection* or random *allocation*. Random selection refers to the use of a random mechanism (e.g., a simple random sample, Definition 6.2, or a stratified random sample, Definition 6.3) to select units from the population. Random selection minimizes bias.

Random allocation refers to the use of a random mechanism when assigning units to a specific treatment group in a controlled experiment (Definition 20.2). Random allocation eliminates confounding and permits causal interpretations.

> ℹ Note
>
> While those new to study design can typically describe random selection and 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.

It is tempting to manually adjust the treatment groups to achieve what the researcher views as balance between the groups. This temptation should be avoided as balancing one feature of the subjects may lead to an imbalance in other features. Remember, random allocation leads

to balance. Of course, random allocation does not guarantee any particular sample is perfectly balanced; however, any differences are due to chance alone. As the sample size increases, these differences due to chance are minimized.

Even with random allocation providing balance between the groups, there will still be variability within each group. The more variability present, the more difficult it is to detect a signal — to discern a difference in the mean response across groups. The study will more easily detect the signal if the groups are similar. This leads to the third component of a well-designed study — the **reduction of noise**.

**Definition 20.6** (Reduction of Noise)**.** Reducing extraneous sources of variability can be accomplished by fixing extraneous variables or blocking (Definition 20.7). These actions reduce the number of differences between the units under study.

> **i** Tension between Lab Settings and Reality
>
> Scientists and engineers are trained to control unwanted sources of variability (or sources of error in the data generating process). This creates a tension between what is observed in the study (under "lab" settings) and what is observed in practice (in "real-world" settings). This tension always exists, and the proper balance depends on the goals of the researchers.

Fixing the value of extraneous variables can reduce variability in a study. For example, in Example 20.1, we might choose to conduct the study at a single hospital. This choice impacts the value of an extraneous variable. It is likely each hospital has its own training process and protocols. The choice to only conduct the study at a single hospital reduces the "noise" in how infants respond due to different nurse behavior that reflects hospital training/protocol. However, note that this decision also potentially limits the scope of the study. It may not longer be appropriate to apply these results to all hospitals.

An additional tool for reducing noise is **blocking**, in which observations which are dependent on one another because of a shared characteristic are grouped together.

**Definition 20.7** (Blocking)**.** Blocking is a way of minimizing the variability contributed by an inherent characteristic that results in dependent observations. In some cases, the blocks are the unit of observation which is sampled from a larger population, and multiple observations are taken on each unit. In other cases, the blocks are formed by grouping the units of observations according to an inherent characteristic; in these cases that shared characteristic can be thought of having a value that was sampled from a larger population.

In both cases, the observed blocks can be thought of as a random sample; within each block, we have multiple observations, and the observations from the same block are more similar than observations from different blocks.

Table 20.1: Data from the 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 |

**Example 20.2** (Overseeding Golf Greens)**.** Golf is a major pastime, especially in southern states. Each winter, the putting greens need to be overseeded with grasses that will thrive in cooler weather. This overseeding 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, each of the five greens 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 from Example 20.2 are shown in Table 20.1.

It would have been easy to simply assign 4 greens to each of the Rye grass varieties; the

random allocation would have balanced any confounders 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* the grouped greens. The blocks in this study are the "slope groups." Each block represents greens with a different slope. Certainly, there are more than 4 potential slopes that a green might have; yet, we observed 4 such groups in our study. We can think of these 4 observed slopes as a sample of all slopes that might exist on a putting green.

Within each block, we have five units of observations; these five units were randomized to the five treatment groups (the five Rye grass varieties). Notice the random allocation strategy ensures that each variety appears exactly once within each slope grouping. This study design will allow us to compare the impact of the Rye grass variety while minimizing the extraneous variability due to the slope of the green, which is a nuisance characteristic. To see how we capitalize on blocking in the analysis, we refer you to Unit IV of the text.

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

An extreme case of blocking occurs when you repeatedly measure the response on the same subject 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 unit of observation. The response is then observed on the subject both prior to the intervention (the "test") and following the intervention. The rationale here is to use every subject as his or her own "control." This reduces extraneous noise because the two treatment groups (the pre-test group and the post-test group) are identical.

> **Big Idea**
>
> A block is a secondary grouping variable present during the data collection that records a nuisance characteristic. While it reduces extraneous noise in the sample, the block must be accounted for appropriately during the analysis of the data (as described in Chapter 22).

## 20.3 Collecting Observational Data

An inability to conduct a controlled experiment does not mean we neglect study design. Random sampling is still helpful in ensuring that the data is representative of the population. And, even if random sampling is not feasible, we should still aim to minimize bias and have a sample that is representative of our 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 as well. This will allow us to account for these confounders in our analysis (see Chapter 24); we cannot model what we do not collect. Finally, observational studies may still permit the blocking of subjects and accounting for this additional variability in our analysis.

# 21 Considerations when Comparing Independent Groups

While the approaches we have described in the text have been general, we have generally assumed we were modeling a single variable from a single population. That is, we were interested in characterizing the data generating process for a response $Y$; specifically, we assumed the density $f(y \mid \theta)$ depended on unknown parameters $\theta$. We would use a random sample $Y_1, Y_2, \ldots, Y_n$ from this population to make inference on those parameters. Many questions do not fit into that mold.

**Example 21.1** (Distracted Driving)**.** Nearly every state in the US has a restriction on cell phone use while driving. Some states prohibit texting while driving, while states like Indiana only permit hands-free use of a phone. The goal of these regulations is to reduce distractions while driving and thereby improve reaction time.

Does cell phone use reduce reaction time? Does it increase the likelihood of a collision? These questions seek to compare a response (the reaction time) across groups formed by the predictor (cell phone use or not). There are two ways to conceptually think about this comparison. First, we could think of there being two independent populations — those who use cell phones while driving, and those who do not use cell phones while driving. In this perspective, we are interested in comparing these two populations, and we imagine sampling from each population separately. Alternatively, we could think of there being a single population (those who drive) exposed to two different treatments (using cell phones and not using cell phones). In this perspective, we are interested in comparing the impact of the treatments on the response, and we imagine sampling from the single population and then each unit being allocated one of the treatment groups. While the details are different between these two perspectives, mathematically, they are equivalent.

Let's first consider the perspective of two independent populations. Specifically, let $Y_{j,i}$ be the response for the $i$-th observation in population $j$, $i = 1, 2, \ldots, n_j$, where $n_j$ is the number of subjects observed from population $j$ and $j = 1, 2, \ldots, J$, where $J$ is the number of populations being compared. Further, we assume $Y_{j,i} \overset{IID}{\sim} f_j\left(y \mid \theta_j\right)$; that is, we have a random sample from each population. Notice that we are potentially allowing the form of the data generating process $f$ to vary for each population; and, we certainly allow the parameters $\theta$ to vary for each population. Since we consider each sample independent of one another, we have that the likelihood has the form

$$f(\mathbf{y} \mid \theta) = \prod_{j=1}^{J} \prod_{i=1}^{n_j} f_j \left( y_{j,i} \mid \theta_j \right). \tag{21.1}$$

It now remains to specify a prior on each $\theta_j$. Of course, if the groups are independent of one another, it is also reasonable to assume the parameters are independent of one another. This leads to

$$\pi(\theta) = \prod_{j=1}^{J} \pi_j(\theta_j). \tag{21.2}$$

Now let's consider the perspective of a single population exposed to two treatments. Specifically, let $Y_i$ be the response for the $i$-th observation in the population, and we let $x_i$ be a variable that indicates to which treatment the $i$-th observation has been exposed; that is, $x_i = j$ if the $i$-th observation has been exposed to the $j$-th treatment. We then have that $Y_i \overset{Ind}{\sim} f\left(y \mid \theta_{x_i}\right)$. Notice that we do not say that the responses are identically distributed; while the family of distributions is the same (all have the same $f$), the parameters on which each distribution depends is allowed to differ (based on the value of $x_i$). Therefore, we are only willing to say the observations are independent. The likelihood then has the form

$$f(\mathbf{y} \mid \theta) = \prod_{i=1}^{n} f\left(y \mid \theta_{x_i}\right). \tag{21.3}$$

It now remains to specify a prior on each $\theta_j$. If we are willing to assume the treatment groups are independent of one another, it is also reasonable to assume the parameters are independent of one another. This leads to

$$\pi(\theta) = \prod_{j=1}^{J} \pi_j(\theta_j). \tag{21.4}$$

Equation 21.1 and Equation 21.3 are equivalent models for the likelihood; similarly, Equation 21.2 and Equation 21.4 are equivalent models for the prior.

> **Note**
>
> When comparing groups, whether you view the groups as separate populations or a single population exposed to different treatments, the inference is the same provided the groups are independent.

Big Idea

In order to extend inference from a single variable to comparing groups, we essentially allow our parameters to be based on the group from which the data was taken. That is, we add a second layer to our model.

**Example 21.2** (C-Section Deliveries Continued)**.** Example 9.1 introduced a study, a component of which includes estimating the probability of a mother undergoing a C-section delivery at a particular hospital. That example focused on estimating the C-section rate at a particular hospital. However, there are two hospitals in Terre Haute, Indiana. Suppose we are now interested in extending the study to compare the rate of C-sections at the two hospitals.

Develop a suitable model for addressing this new research objective.

*Solution.* Let $Y_{j,i}$ denote the number of vaginal deliveries at hospital $j$ prior to the $i$-th C-section we observe at hospital $j$. Further, let $\theta_j$ denote the rate of C-sections at hospital $j$; our interest is in comparing $\theta_1$ and $\theta_2$. Following the discussion of Example 9.1, we have that

$$Pr\left(Y_{j,i} = y\right) = \theta_j \left(1 - \theta_j\right)^y \qquad y = 0, 1, 2, \ldots \, .$$

Further, it is reasonable to assume that whether a mother undergoes a C-section at one hospital is unrelated to whether a mother undergoes a C-section at the other hospital. Therefore, we can consider $Y_{1,i}$ to be independent of $Y_{2,k}$ for all choices of $i$ and $k$. And, since each birth is independent, we can consider $Y_{j,1}, Y_{j,2}, \ldots, Y_{j,n}$ to be a random sample from the $j$-th hospital. Note that while we have set $n$ to be the same for both hospitals, this is not a requirement.

Letting $\mathbf{Y} = \left(Y_{1,1}, Y_{1,2}, \ldots, Y_{1,n}, Y_{2,1}, Y_{2,2}, \ldots, Y_{2,n}\right)^\top$ be the random vector of observations, the likelihood is given by

$$\begin{aligned} f(\mathbf{y} \mid \theta) &= \prod_{j=1}^{2} \prod_{i=1}^{n} f_{Y_{j,i}}\left(y_{j,i} \mid \theta_j\right) \\ &= \prod_{j=1}^{2} \prod_{i=1}^{n} \theta_j \left(1 - \theta_j\right)^{y_{j,i}} \\ &= \prod_{j=1}^{2} \theta_j^n \left(1 - \theta_j\right)^{\sum_{i=1}^{n} y_{j,i}} \\ &= \theta_1^n \left(1 - \theta_1\right)^{n\bar{y}_1} \theta_2^n \left(1 - \theta_2\right)^{n\bar{y}_2}, \end{aligned} \tag{21.5}$$

where $\bar{y}_j$ represents the observed sample mean for hospital $j$. We note that the likelihood expressly captures both populations and therefore acknowledges the dependence on the both parameters.

We now consider developing a prior distribution. Since $\theta_j$ is a probability, it seems reasonable to choose a Beta distribution for each parameter. Further, if we believe the parameters are independent, the prior distribution is

$$\pi(\theta) = \frac{\Gamma(a_1 + b_1)}{\Gamma(a_1)\Gamma(b_1)}\theta_1^{a_1-1}(1-\theta_1)^{b_1-1}\frac{\Gamma(a_2 + b_2)}{\Gamma(a_2)\Gamma(b_2)}\theta_2^{a_2-1}(1-\theta_2)^{b_2-1}.$$

Notice that our choice of prior allows the hyperparameters to potentially differ for each hospital.

## 21.1 Bridge Sampling

In Chapter 15 we introduced the idea of model comparison. This is especially useful when comparing groups. Consider a simple case in which we have two groups. Specifically, suppose

$$\begin{aligned} Y_{1,i} &\overset{IID}{\sim} f(y \mid \theta_1) \\ Y_{2,i} &\overset{IID}{\sim} f(y \mid \theta_2) \\ Y_{1,i} &\perp\!\!\!\perp Y_{2,j} \quad \forall i, j \end{aligned}$$

where $\perp\!\!\!\perp$ refers to independence. In this example, the data generating process for each group is distinguished only by a potentially different value of the parameter. It would be natural in such settings to consider the hypotheses

$$H_0 : \theta_1 = \theta_2 \qquad \text{vs.} \qquad H_1 : \theta_1 \neq \theta_2.$$

If we place a continuous prior on the parameters, then the probability of $H_0$ must be 0. One way of addressing this problem is to consider a clever prior which places mass along the *line* equating the two parameters. A more natural solution, however, is to think of this as a model comparison problem:

$$\begin{aligned} \mathcal{M}_0 : \quad & Y_{j,i} \overset{IID}{\sim} f(y \mid \theta) \\ & \theta \sim \pi_0(\theta) \\ \mathcal{M}_1 : \quad & Y_{j,i} \overset{Ind}{\sim} f(y \mid \theta_j) \\ & \theta_j \overset{Ind}{\sim} \pi_j(\theta_j) \end{aligned}$$

for $j = 1, 2$. Notice that under model $\mathcal{M}_0$, there is only a single parameter, capturing the null hypothesis $\theta_1 = \theta_2$. And, model $\mathcal{M}_1$ allows the flexibility for two parameters, capturing the alternative hypothesis. We would therefore be interested in computing a Bayes Factor (see

Definition 15.2) comparing these two models. The problem is, the Bayes Factor depends on the evidence

$$f(\mathbf{y} \mid \mathcal{M}_j) = \int f(\mathbf{y} \mid \theta, \mathcal{M}_j)\pi(\theta \mid \mathcal{M}_j)d\theta.$$

Unfortunately, this integral is quite difficult to estimate. Using the prior distribution and performing MC Integration techniques does not typically result in reliable estimation; this results from the prior distribution often being too vague. Instead, the evidence is estimated from the posterior distribution using **bridge sampling**.

**Definition 21.1** (Bridge Sampling)**.** The bridge sampling estimator of the marginal likelihood $m(\mathbf{y})$ is given by

$$\begin{aligned} m(\mathbf{y}) &= \int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta \\ &= \frac{E_g\left[h(\theta)f(\mathbf{y} \mid \theta)\pi(\theta)\right]}{E_\pi\left[h(\theta)g(\theta)\right]} \\ &\approx \frac{m^{-1}\sum_{j=1}^{m} h\left(\tilde{\theta}_j\right) f\left(\mathbf{y} \mid \tilde{\theta}_j\right) \pi\left(\tilde{\theta}_j\right)}{m^{-1}\sum_{i=1}^{m} h\left(\theta_j^*\right) g\left(\theta_j^*\right)} \end{aligned}$$

where $h(\theta)$ is called the bridge function and $g(\theta)$ is the proposal distribution. Here, $\tilde{\theta}$ denotes a random variate from the proposal distribution and $\theta^*$ a random variate from the posterior; $E_g$ denotes taking an expectation with respect to the proposal distribution and $E_\pi$ denotes taking an expectation with respect to the posterior distribution.

> **Note**
>
> While not required, it is typical to use a Normal distribution for the proposal distribution $g(\theta)$.

Bridge sampling is available in some software packages (such as R, for example). While the details of bridge sampling are beyond the scope of this text, we do want to note that the definition comes from making the following observation:

$$\begin{aligned} 1 &= \frac{\int f(\mathbf{y} \mid \theta)\pi(\theta)g(\theta)h(\theta)d\theta}{\int f(\mathbf{y} \mid \theta)\pi(\theta)g(\theta)h(\theta)d\theta} \\ \Rightarrow m(\mathbf{y}) &= m(\mathbf{y})\frac{\int f(\mathbf{y} \mid \theta)\pi(\theta)g(\theta)h(\theta)d\theta}{\int f(\mathbf{y} \mid \theta)\pi(\theta)g(\theta)h(\theta)d\theta}. \end{aligned}$$

We are able to multiply both sides by the prior predictive distribution because it will be non-negative on its support. Next, recognize that the integral is with respect to $\theta$; therefore, we can move the marginal distribution inside the integral in the denominator to get

$$\begin{aligned} m(\mathbf{y}) &= \frac{\int f(\mathbf{y} \mid \theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int \frac{f(\mathbf{y}|\theta)\pi(\theta)}{m(\mathbf{y})}h(\theta)g(\theta)d\theta} \\ &= \frac{\int f(\mathbf{y} \mid \theta)\pi(\theta)h(\theta)g(\theta)d\theta}{\int \pi(\theta \mid \mathbf{y})h(\theta)g(\theta)d\theta}, \end{aligned}$$

where the last equality makes use of the definition of the posterior distribution from Bayes Theorem. Now, the top integral can be viewed in terms of an expectation over a random variable $\theta \sim g(\theta)$, and the denominator can be viewed as an expectation over a random variable which follows the posterior distribution.

Big Idea

Bridge sampling is an efficient algorithm for estimating the evidence of a model, allowing for computation of the Bayes Factor.

# 22 Considerations when Comparing Related Groups

The previous chapter was a natural extension of the framework we had developed in earlier portions of the text. Assuming each group is independent, we were essentially able to model each group separately; the likelihood was then the product of the likelihoods from each group, and the prior was the product of the priors from each group. This independence will actually carry through into the posterior distribution. That is, the independence allows us to model the groups separately and then combine afterwards. However, independence between the groups is not always reasonable.

Recall that one of the aspects of a good study design is comparative groups — treatment groups which are similar except for the treatment being applied. The benefit of this is that it reduces extraneous variability. We also saw that blocking (Definition 20.7) is a useful strategy for reducing extraneous variability that groups together observations which share some inherent characteristic. A very common example of blocking is a pre/post test. In such settings, participants are given a baseline assessment. Then, each participant is exposed to some form of treatment; following this, participants take an another assessment. Interest is typically on quantifying the change from baseline.

In this case, the two groups (pre-treatment and post-treatment) are not only similar, they are identical! Intuitively, this is a good design because we are allowing every individual to act as their own control. We have eliminated all other external sources of variability allowing us to focus on the treatment of interest. Here, the individual participants act as the blocks. In such cases, we believe the variability among observations between blocks is greater than the variability among observations within a block. Unfortunately, this causes a relationship among the observations, meaning it is no longer reasonable to assume the observations are independent of one another.

> Big Idea
>
> When subjects can be blocked (or pooled) into similar groups, independence is no longer reasonable.

When we believe there are clusters of related observations, we lean on the hierarchical nature of the data generating process, and this allows us to construct the likelihood.

**Example 22.1** (Final Exams)**.** Common final exams are typical for multiple sections of the same course at a university. For example, there may be four instructors, each teaching two sections of Calculus; all eight sections will receive the same final exam. Suppose each exam is graded out of 100 points and we are interested in modeling the exam score for students taking the exam.

We might start off by assuming a common distribution for all students. That is,

$$\begin{aligned} Y_i \mid \theta &\overset{IID}{\sim} f(y \mid \theta) \\ \theta &\sim \pi(\theta). \end{aligned}$$

where $Y_i$ is the exam score for the $i$-th subject. However, this model is imposing a particular assumption — there are no differences among instructors that would impact the scores of the students. Essentially, every instructor delivers the same content in the same way suggesting the students might has well have been taught by the same instructor. While this is a nice ideal, it is typically not the case. Autonomy in the classroom means that different instructors approach the material differently — emphasizing different topics and presenting the material in different ways. As a result, it is possible that students who share an instructor are more likely to perform well on the same types of problems and also more likely to make the same types of mistakes (compared with students who do not share the same instructor). The additional variability introduced by the differences across instructors is being ignored in this model.

We might try to correct for this by assuming the instructors are completely independent of one another, following the approach of the previous chapter. This would say

$$\begin{aligned} Y_{j,i} \mid \theta_j &\overset{IID}{\sim} f(y \mid \theta_j) \\ Y_{r,s} &\perp\!\!\!\perp Y_{t,u} \quad \forall r, s, t, u \\ \theta_j &\overset{Ind}{\sim} \pi_j(\theta_j). \end{aligned}$$

However, this model has very limited utility in practice. If we wanted to predict the exam score of a student taking calculus, our model would only be applicable if they were taking it with one of *these* four instructors! Given another student, we would need to know which of these instructors they were taking the course with in order to know which of the four parameters to lean on in predicting their score. That is, this model inherently compares instructors; but, we do not actually care about comparing Instructor 1 and Instructor 2. Further, the next time the course is offered, there are likely to be four completely new instructors, and we want our model to be accommodate this. Perhaps more importantly, we do not actually think the instructors are completely unrelated (they are teaching the same major content after all…at least, we would hope so). So, this extreme perspective also seems to ignore the structure in the problem.

Big Idea

How we model the relationship between groups, if one exists at all, must be based on the context of the problem.

It seems natural to think of these four instructors as a representative sample from a population of all potential instructors. That is, we are adding a layer to the data generating process. First, instructors are chosen to teach the course; then, students are placed with an instructor, impacting their performance on the final exam. The way these instructors impact the data generating process is through the formation of the parameters $\theta_j$. Therefore, we incorporate this into the model. Specifically, consider

$$\begin{aligned} Y_{j,i} \mid \theta_j &\overset{IID}{\sim} f(y \mid \theta_j) \\ \theta_j \mid \eta &\overset{IID}{\sim} \pi(\theta \mid \eta) \\ \eta &\sim \pi(\eta). \end{aligned}$$

This model has three layers:

- Layer 1: Describes how students *within* an instructor perform; the parameters are unique to each instructor but shared across students with the same instructor.
- Layer 2: Describes the variability *across* instructors; treating the instructors as a random sample of all instructors, we allow the parameters to move across instructors according to some overall model.
- Layer 3: Describes our prior beliefs about the shared parameters for the instructor-level model.

This hierarchical model bridges the gap between ignoring the blocks and treating the blocks as independent groups. This allows us to pool information similar across blocks while allowing unique properties to exist for each block as well.

Hierarchical Approach to Addressing Dependence

When the dependence in data is the result of a hierarchical data generating process, we can incorporate that additional variability by modeling the hierarchical structure directly. At a minimum, this has the following three layers:

- Layer 1: Describes how observations *within* a block behave; with parameters unique to each block.
- Layer 2: Describes how the parameters move *across* blocks; this views the blocks as a random sample.
- Layer 3: Describes our prior beliefs on the common parameters for the model in Layer 2.

It is natural to question the difference between a block and a factor (or the grouping comparison that we considered in the previous chapter). The difference is primarily in how we address them in the modeling. If we are interested in making group-to-group comparisons, and if the groups would remain the same when repeating the study, the variable should be treated as a factor of interest. If the groups primarily capture an additional source of variability and can be viewed as a random sample from some larger population, the variable should be treated as a block.

Of course, we can combine the two ideas in a single study. Suppose further that we were interested in comparing students majoring in mathematics and those not majoring in mathematics. The first layer of the model must still describe what happens within a block (or within an instructor in our example). The assumptions you are willing to make determine how complex this becomes. For example, assuming that the difference between mathematics majors and non-mathematics majors is similar regardless of the instructor implies that mathematics majors share some common parameter across instructors. However, allowing the difference to vary across instructors implies that there is no common parameter. This is addressed within the first two layers of the model. If the effect is held constant across the blocks, a single set of "global" parameters enters into Layer 1 for which priors are described in Layer 3. If the effect is allowed to vary, then the way in which the parameters vary is defined in Layer 2. To broaden our model above, we might have something of the form

$$\begin{aligned} Y_{j,i} \mid \theta_j, \phi &\overset{IID}{\sim} f(y \mid \theta_j, \phi) \\ \theta_j \mid \eta &\overset{IID}{\sim} \pi(\theta \mid \eta) \\ \eta &\sim \pi(\eta) \\ \phi &\sim \pi(\phi). \end{aligned}$$

In this formulation, $\phi$ represents parameters which are constant across the blocks; so, they are separated out from $\theta_j$ as they do not vary across blocks. We place a prior directly on these parameters in Layer 3.

Note that we have been intentionally vague about the modeling structure by leaving everything in terms of $f$ instead of defining a specific model. The form of the model will change depending on the data generating process. Further, which parameters in that family may be altered. For example, suppose we considered a Normal distribution for the data generating process; we could allow the mean response to vary across groups, the variability to vary across groups, both the mean response and variability to vary across groups, or hold both the mean response and variability to be similar across groups. This is true for both the blocks as well as the treatment groups! That is, our modeling should be specific to the question at hand and the data generating process being considered.

# Part VI

# Unit VI: Introduction to Regression Modeling

The heart of this text (Unit III) introduced the fundamentals of the Bayesian framework for performing inference. We then allowed the distribution of the response to vary across a finite number of groups by allowing the unknown parameters that govern the likelihood to potentially differ for each group (Unit IV). We now consider those unknown parameters that govern the distribution of the response to depend on one or more predictors through some functional form.

This unit provides a brief overview of considerations for such "regression" models. We take a very general approach, considering both quantitative and categorical response variables, quantitative and categorical predictors, and various functional forms for how these predictors impact the parameters governing the likelihood.

# 23 Regression Models for a Quantitative Response

Recall that continuous random variables are analogous to quantitative variables in a study; similarly, discrete random variables are analogous to categorical (or qualitative) variables in a study. In this chapter, we are interested in allowing the (continuous) distribution of the random variable that represents the response to depend on one or more predictors. Let's begin with an example to illustrate the range of questions we might ask in such a setting.

**Example 23.1** (Rehabilitation Therapy)**.** Physical therapy is a vital part of the recovery from any surgery, particularly surgery which impacts motor function, such as knee or hip surgery. A researcher at a local rehabilitation center was interested in the recovery time of individuals who have undergone corrective knee surgery. Specifically, the researcher is interested in examining the relationship between the physical fitness (of persons undergoing corrective knee surgery) prior to surgery and the time required in physical therapy until successful rehabilitation ("recovery time"). Patient records in the rehabilitation center were examined, and 24 male subjects ranging in age from 18 to 30 years who had undergone similar corrective knee surgery during the past year were selected for the study. The following information was obtained from the case record of each patient:

- Recovery Time: number of days of physical therapy until the subject has met their physical therapy goals.
- Prior Activity Level: physical fitness status (low, moderate, high) prior to surgery; this is based on a questionnaire asking about the frequency of various physical activity over the course of a typical week.
- Age: age (years, rounded to nearest tenth).

> **Note**
>
> We note that this study has potential bias. The initial question was framed in terms of all patients undergoing corrective knee surgery, but the sample consists of only male patients. Unfortunately, for a number of years, women were highly underrepresented in medical research (with the exception of studies involving cosmetic products), and individuals who do not identify as male or female continue to be underrepresented in medical studies.

By all accounts, this is a small study, recording only 3 variables on 24 subjects. Yet, even with only a few variables, there are a myriad of potential questions that require more complex models than we have previously discussed. Consider the following:

- Overall, is the recovery time, on average, linearly related to the age of patients?
- For subjects of the same age, is there a relationship between the recovery time, on average, and the prior activity level of patients?
- Does the relationship between the recovery time, on average, and the age of a patient depend on the patient's prior activity level?
- Does the variability in the recovery time depend on the age of the patient?

Any one of these questions could be the focus of a researcher. Notice that each, though, is slightly different. While the first question involves a relationship between a quantitative response and a quantitative variable, the second involves a quantitative response and multiple predictors (one quantitative and one categorical) with a desire to essentially "isolate" the impact of one of the predictors. The third question looks at how the relationship between the response and predictor might be modified by a third variable; and, the fourth looks at a relationship impacting the variability instead of the mean response. Each of these questions require us to move beyond the methods we have previously discussed. And, each of these questions shares a similar structure.

We might be tempted to force these questions into the "comparison of groups" approach studied previously. To examine the difference in recovery time across age, for example, we might categorize subjects as "in their teens" or "in their 20's." However, this approach prevents us from examining smooth trends across age. Such categorizations can also lead to categories with a very small number of subjects, meaning our priors will have undue influence on the results. We will consider an alternative approach.

Big Idea

Broadly speaking, there are three types of scientific questions that regression models allow us to address:

- Marginal relationships: captures the "overall" relationship between two variables, ignoring any other contributing factors.
- Adjusted relationships: captures the effect of a variable after "isolating" it from other contributing factors.
- Effect Modification: captures how the effect of one variable on the response is modified by a second contributing factor.

Of course, regardless of these types of scientific questions we might address, we can also use our model to predict a future observation given a specific set of contributing factors.

## 23.1 Developing a Model

In many introductory statistics courses, statistical models are typically introduced in the following generic form:

$$\text{Response} = \text{Signal} + \text{Noise}.$$

The "response" is the variable we would like to explain or predict. The "signal" represents the part of the data generating process we can explain; it is the deterministic portion of the model and is a function of the predictor(s). The "noise" represents the stochastic portion capturing the variability observed beyond what can be explained by the deterministic portion alone. The common starting point for such a model is the simple linear regression model:

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor})_i + \varepsilon_i.$$

This model views the deterministic portion of the model as a line. Notice that any two subjects with the same value of the predictor would have the same value for the deterministic portion of the model; however, the two subjects will not necessarily have the same response as the noise $\varepsilon_i$ can differ from one subject to the next. Of course, this model is too vague to be helpful; so, we place additional constraints on the stochastic portion. In particular, we place conditions on the *distribution* of $\varepsilon_i$. For example, we might assume that $E(\varepsilon_i) = 0$ or that the $\varepsilon_i$'s are identically distributed.

Notice that this approach, while reasonable, does not align with the approach we have taken thus far in the course. Instead of thinking of the response as a signal plus noise and then placing conditions on the distribution of the noise, we have considered fully specifying the distributional form of the data generating process. That is, we specified the likelihood $f(\mathbf{y} \mid \theta)$. This approach is preferred in the Bayesian perspective (and in the classical perspective in our opinion) because it generalizes more easily and provides a unifying framework for inference. It is within this context of fully specifying the likelihood that we consider developing regression models.

**Definition 23.1** (Regression)**.** A regression model is one for which the parameter(s) governing the data generating process depends on one or more predictors. "Parametric" regression models do this through specifying a functional form for the dependence of the parameter(s) on the predictor(s).

To illustrate the scope of Definition 23.1, let's consider several potential models we might consider in the Bayesian framework. Consider

$$(\text{Response})_i \mid \beta_0, \beta_1, \sigma^2 \overset{\text{Ind}}{\sim} N\left(\beta_0 + \beta_1(\text{Predictor})_i, \sigma^2\right). \tag{23.1}$$

Notice that Equation 23.1 allows the *mean* response to depend on the predictor through the form

$$\beta_0 + \beta_1(\text{Predictor})_i.$$

That is, instead of a single mean response $\mu$, the mean response is determined only after first specifying the value of the predictor. As the mean response differs for each subject, depending on their value of the predictor, the responses are *not* identically distributed. However, we have retained the assumption that the responses are independent of one another (given the unknown parameters). Further, it is *only* the mean response that is impacted by the predictor. The variance of the response is not impacted but remains constant across all observations. This model also fully specifies the distributional form of the response — it is a Normal distribution.

In contrast to Equation 23.1, consider

$$(\text{Response})_i \mid \alpha, \gamma_0, \gamma_1 \overset{\text{Ind}}{\sim} Gamma\left(\alpha(\text{Predictor 1})_i^2, \gamma_0 e^{\gamma_1(\text{Predictor 2})_i}\right), \tag{23.2}$$

which says the responses follow a Gamma distribution where the shape parameter depends on "Predictor 1" through a quadratic relationship; and, the rate parameter depends on "Predictor 2" through an exponential function. Again, while the responses are not identically distributed, they remain independent. Both Equation 23.1 and Equation 23.2 extend common distributional models by allowing the parameters to depend on predictors. The model that is appropriate should be driven by the research objectives and discipline expertise.

To get a sense of how model construction is related to the research objectives, let's return to Equation 23.1. Notice that by replacing the mean with a functional form, we have introduced new parameters into the distribution: $\beta_0$ and $\beta_1$. These parameters govern the mean response. As a result, their interpretation is tied to the mean response:

- $\beta_0$ represents the mean response when the value of the predictor is 0.
- $\beta_1$ represents the change in the mean response when the predictor is increased by 1 unit.

Notice that the interpretation of the parameters could be related to specific scientific questions. If we are interested in how changes in the predictor relate with changes in the response, we are interested in $\beta_1$.

The interpretation of $\alpha$ in Equation 23.2 is not as clear. It represents the shape parameter when the first predictor takes a value of 1; since the shape parameter is not directly interpretable in terms of the response (like the mean or variance is), this interpretation is less satisfying. But, we know that $\frac{\alpha}{\gamma_0}$ would represent the average response when the first predictor takes a value of 1 and the second predictor takes a value of 0. That is, through a combination of the parameters, we have some sense of the mean response.

Of course, there are infinitely many other models we could specify. The distributional family could vary (Normal, Gamma, Beta, a mixture, etc.); the functional form could vary (linear, exponential, sinusoidal, etc.); the parameters impacted could vary (mean, shape, rate, etc.). What is common to each of these potential models is that we specify the distribution of the response allowing key aspects of the distribution to vary as functions of the predictor(s).

Specifying the distribution of the response is only a portion of the model under the Bayesian framework. Returning to the distributional family specified in Equation 23.1, the assumption of independence allows us to easily construct the likelihood, given by

$$\begin{aligned} f\left(\textbf{Response} \mid \beta_0, \beta_1, \sigma^2\right) &= \prod_{i=1}^{n} f\left(\left(\text{Response}\right)_i \mid \beta_0, \beta_1, \sigma^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\left(\text{Response}\right)_i - \beta_0 - \beta_1(\text{Predictor})_i\right)^2\right\}. \end{aligned}$$

We present this to show that the likelihood can quickly become difficult to work with as the complexity of the distributional model grows.

## 23.2 Simple Extensions

As previously stated, the distributional model in Equation 23.1 represents a common model for introducing regression. However, there are two simple extensions that are worth considering. First, we consider the inclusion of multiple predictors. For example, given $p$ predictors, we might posit that

$$\left(\text{Response}\right)_i \mid \beta, \sigma^2 \overset{\text{Ind}}{\sim} N\left(\beta_0 + \sum_{j=1}^{p}\beta_j(\text{Predictor } j)_i, \sigma^2\right) \tag{23.3}$$

for $j = 1, 2, \ldots, p$, giving a likelihood of

$$f\left(\textbf{Response} \mid \beta, \sigma^2\right) = \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(\left(\text{Response}\right)_i - \beta_0 - \sum_{j=1}^{p}\beta_j(\text{Predictor } j)_i\right)^2\right\}.$$

This model allows several predictors to impact the mean response.

> **Note**
>
> While it is common to have the predictors enter the model such that the mean response is linear *in the parameters*, this is not a requirement. We could easily allow the predictors to impact the mean response through some nonlinear function.

Second, nothing requires that the response follow a Normal distribution. For example, we might posit that

$$(\text{Response})_i = \beta_0 + \beta_1(\text{Predictor})_i + \varepsilon_i,$$

where $\varepsilon_i \overset{\text{IID}}{\sim} t_\nu$. While we have specified this in a "signal plus noise" form, notice that we are simply saying the response can be modeled by a t-distribution which has been shifted to be centered on $\beta_0 + \beta_1(\text{Predictor})_i$.

Both of these extensions are complex from a classical perspective, requiring different machinery to be able to conduct inference. However, from the Bayesian perspective, we have specified a different model, but the process for performing inference remains exactly the same: specify a distribution to capture the prior information (see Chapter 25) and then compute the posterior distribution (typically using MCMC methods).

## 23.3 Fixed vs. Random Predictors

You may have noticed that Equation 23.1 only specified the distribution of the response variable as a function of the predictor; it omitted the distribution of the predictor itself. For a designed experiment in which the values of all predictors are determined in advance by the researchers, the predictors are constants. As such, the notation of Equation 23.1 is appropriate. However, in many situations, the predictors are not fixed in advanced but observed during the data collection; that is, the predictors can also be viewed as observed values of random variables which vary across individuals in the population. Consider the Rehabilitation example (Example 23.1) above; the age of each patient is unknown prior to the study. We expect the age to vary across individuals who have undergone knee replacement; therefore, age has a distribution across the population that should be modeled.

It is common in a regression analysis to *condition* on the predictors when making inference. Consider Equation 23.1; if we believe the predictor is not determined in advance, but we condition on the value of the predictor, then our model would be expressed as

$$(\text{Response})_i \mid (\text{Predictor})_i, \beta_0, \beta_1, \sigma^2 \overset{\text{Ind}}{\sim} N\left(\beta_0 + \beta_1(\text{Predictor})_i, \sigma^2\right). \tag{23.4}$$

The likelihood for this model is equivalent to what we had before; in fact, nothing about our analysis changes. However, conceptually, we are saying that our model applies only after knowing the value of the predictor. As a result, the model does not specify the distribution of the predictor, and all interpretations assume we have access to the predictor prior to making a statement about the response.

While it is common to condition on the predictors when making inference, it is not a requirement. Including the distribution of the predictors is simply a modeling exercise when developing the likelihood. Since Equation 23.4 already specifies the conditional distribution of the response given the predictor, adding a statement about the marginal distribution of the predictor leads to a hierarchical model that fully specifies the distribution of all observed variables. For example, we might consider the model

$$\begin{aligned} (\text{Response})_i \mid (\text{Predictor})_i, \beta_0, \beta_1, \sigma^2 &\overset{\text{Ind}}{\sim} N\left(\beta_0 + \beta_1(\text{Predictor})_i, \sigma^2\right) \\ (\text{Predictor}_i \mid \gamma, \eta^2 &\overset{\text{IID}}{\sim} N\left(\gamma, \eta^2\right). \end{aligned} \tag{23.5}$$

While Equation 23.5 has the same predictor as in Equation 23.1; however, it fully specifies the distribution of all observed variables. The likelihood of the observed data is then

$$\begin{aligned} f\left(\textbf{Data} \mid \beta_0, \beta_1, \sigma^2, \gamma, \eta^2\right) &= \prod_{i=1}^{n} f\left((\text{Response})_i \mid (\text{Predictor})_i, \beta_0, \beta_1, \sigma^2\right) g\left((\text{Predictor})_i \mid \gamma, \eta^2\right) \\ &= \left(2\pi\sigma^2\right)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left((\text{Response})_i - \beta_0 - \beta_1(\text{Predictor})_i\right)^2\right\} \\ &\quad \cdot \left(2\pi\eta^2\right)^{-n/2} \exp\left\{-\frac{1}{2\eta^2}\sum_{i=1}^{n}\left((\text{Predictor})_i - \gamma\right)^2\right\}. \end{aligned}$$

While clearly a more complex model, the benefit is that we are able to simultaneously predict the value of the predictor and response for a future observation.

> **Note**
>
> While it is common to model the response conditional on the predictors and not specify the distribution of the predictors, it is possible to fully specify the likelihood of all observed variables, if desired.

## 23.4 Interpreting the Predictors

As we have already stated, the range of potential models is infinitely large; as a result, there is no one interpretation we can provide for a parameter in the model. However, it is common that

the predictors in a regression model govern the mean response. For example, Equation 23.3 considers multiple predictors, but each impacts the mean response while assuming the variance $\sigma^2$ is constant across the population. In this model, $\beta_j$ describe the relationship between the response and the $j$-th predictor. Specifically,

> $\beta_j$ is the change in the average response given a 1-unit increase in the $j$-th predictor, *holding all other predictors fixed.*

We can also provide an interpretation for the intercept $\beta_0$:

> $\beta_0$ is the average value of the response when *all* predictors take the value 0.

While not governing the mean response, we should not ignore the interpretation of $\sigma^2$:

> $\sigma^2$ is the variability in the response for a *fixed set of predictors.*

These interpretations seem straight forward but are hiding a lot of complexity. Working backward, notice the conditional/cross-sectional nature of the interpretation of $\sigma^2$. It is *not* the overall variability in the response; it is the variability of the response for any fixed set of values for the predictors. That is, the model is specifying the distribution of the response for a specified level of the predictors.

> **Warning**
>
> Recall that a marginal distribution and a conditional distribution of a random variable are distinct distributions. The distributional model of the response in a regression setting is conditional on the values of the predictors, and the form of the marginal distribution of the response need not have the same form.

Next, we notice that the interpretation of $\beta_0$ may not always make sense in context. For example, suppose our response is the weight of individuals (in pounds) and our predictor is their height (in inches). It does not make sense to talk about an individual with a height of 0 inches. This is the result of extrapolation.

**Definition 23.2** (Extrapolation)**.** Extrapolation occurs when we use a model to predict outside of the region for which data is available.

The danger with extrapolation is that without scientific justification, we have no reason to believe the model we have observed in one region of the support will continue to hold for all regions of the support. For example, it is possible the relationship between the mean response and the predictor is linear on the interval $(a, b)$, but it becomes quadratic outside of this interval. If we fit the model on the interval $(a, b)$ and then predict outside this range, our predictions will be biased. This is what can lead to senseless interpretation of $\beta_0$. We should always be cautious of extrapolation.

Finally, and we cannot emphasize the benefit of this enough, the interpretation of $\beta_j$ measures the effect of the $j$-th predictor *holding the value of other predictors in the model fixed*. This means that in a regression model, we are naturally isolating the effect of the predictor. This provides a unique interpretation to a hypothesis of the form

$$H_0 : \beta_j = 0 \qquad \text{vs.} \qquad H_1 : \beta_j \neq 0.$$

This hypothesis is really asking whether the $j$-th predictor is linearly associated with the mean response *after accounting for the impact of the other predictors in the model*.

**Example 23.2** (Rehabilitation Therapy Continued)**.** Example 23.1 described a study to investigate recovery time among patients who have undergone a corrective knee surgery. Suppose we are willing to believe that the mean recovery time is linearly related to the age of a patient. Further, suppose that we believe that given the age of the patient, the recovery time will follow an Exponential distribution. Write down the likelihood for the observed data conditional on the age of the patient.

*Solution.* Recall that for an exponential distribution, the mean response is given by the scale parameter; therefore, the mean response fully characterizes the distribution. If we are willing to assume that given a patient's age, the recovering time of one patient is unrelated to the recovery time of any other patient, then based on the description, we have

$$(\text{Recovery Time})_i \mid (\text{Age})_i, \beta_0, \beta_1 \overset{\text{Ind}}{\sim} Exp\left(\beta_0 + \beta_1(\text{Age})_i\right),$$

where the Exponential distribution is parameterized by its scale parameter. This results in a likelihood of

$$\begin{aligned} f\left(\textbf{RecoveryTime} \mid \textbf{Age}, \beta_0, \beta_1\right) &= \prod_{i=1}^{n} \frac{1}{\beta_0 + \beta_1(\text{Age})_i} \exp\left\{-\frac{(\text{Recovery Time})_i}{\beta_0 + \beta_1(\text{Age})_i}\right\} \\ &= \left[\prod_{i=1}^{n} \frac{1}{\beta_0 + \beta_1(\text{Age})_i}\right] \exp\left\{-\sum_{i=1}^{n} \frac{(\text{Recovery Time})_i}{\beta_0 + \beta_1(\text{Age})_i}\right\}. \end{aligned} \tag{23.6}$$

The likelihood does not reduce to a simple form.

**i** Reparameterization

Consider the proposed solution to Example 23.2. Recall that the scale parameter of an Exponential distribution must be positive; however, nothing in the above specification requires the linear predictor

$$\beta_0 + \beta_1(\text{Age})_i$$

be positive for all ages. Often times, this is not a problem; the region of reasonable values of the parameter will result in a positive value of the scale parameter within the range of our data (again, extrapolation could be problematic). However, we might be in a case where reasonable values of the parameter lead to negative mean responses; or, it could be that the MCMC algorithm wanders into such regions creating numerical instability in the algorithm itself. Regardless, if we would like to enforce the constraint that the mean response be positive, we have two options.

First, we could address the constraint by ensuring that both $\beta_0$ and $\beta_1$ are restricted to be positive. This could be done by choosing priors which have support on the positive real line, for example. Alternatively, we could reparameterize the model to force $\beta_0$ and $\beta_1$ to be positive. That is, we write

$$(\text{Recovery Time})_i \mid (\text{Age})_i, \gamma_0, \gamma_1 \overset{\text{Ind}}{\sim} Exp\left(e^{\gamma_0} + e^{\gamma_1}(\text{Age})_i\right),$$

where we have substituted $\beta_0 = e^{\gamma_0}$ and $\beta_1 = e^{\gamma_1}$. Notice that $\beta_0$ and $\beta_1$ will be positive for any value of $\gamma_0$ and $\gamma_1$; therefore, we now have unconstrained parameters $\gamma_0$ and $\gamma_1$ and yet have constrained the mean response to be positive.

There is yet a third option. We could reparameterize the mean response directly. That is, we consider

$$(\text{Recovery Time})_i \mid (\text{Age})_i, \beta_0, \beta_1 \overset{\text{Ind}}{\sim} Exp\left(e^{\beta_0 + \beta_1(\text{Age})_i}\right).$$

In this specification, there are no constraints on $\beta_0$ and $\beta_1$, but we have ensured that the mean response is positive. This has come at a cost, however; in this specification, the mean recovery time is no longer linearly related to age.

Reparameterization is a helpful tool to consider to enforce constraints in a numerically stable way.

# 24 Extensions to the Linear Model

It is difficult to develop a general theory of regression models since each model can be uniquely constructed, allowing the predictors to impact the mean response, the variability of the response, or both through a shape parameter, for example. However, regardless of the regression model being considered, there are some common modeling techniques that can be helpful.

Let $\theta = g(\beta, \textbf{Predictors})$ be some parameter characterizing the distribution of the response conditional on some predictors. For example, $\theta$ might represent the mean response given the predictors. The previous chapter introduced the idea of $g(\cdot)$ being a linear function of the parameter vector $\beta$; that is, the predictors impact $\theta$ through a linear function like

$$\theta_i = \beta_0 + sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i. \tag{24.1}$$

It turns out that while it may not be immediately obvious, this modeling structure is quite flexible. In this chapter, we consider some common extensions that can be framed in terms of this linear structure.

## 24.1 Including Categorical Predictors

When our collection of predictors only consists of quantitative variables, applying Equation 24.1 is straightforward. However, when one of the predictors is a categorical variable (for example, the prior activity level of a patient in Example 23.1), it may not be obvious how to incorporate these into Equation 24.1. This is actually related to the models we considered in Chapter 21; we revisit the idea from a regression perspective.

**Example 24.1** (Anxiety Among College Students)**.** Suppose we conduct a small survey to determine whether the distribution of the anxiety level of college students differs depending on their class standing (freshman, sophomore, junior, senior). Let $\theta_i$ represent the parameter of interest governing the distribution of anxiety for the $i$-th student. We want to let $\theta_i$ depend on class standing within the form illustrated in Equation 24.1.

We approached this problem in Chapter 21 essentially by saying

$$(\text{Anxiety})_i \mid \theta \overset{\text{Ind}}{\sim} f\left((\text{Anxiety})_i \mid \theta_j\right),$$

where $\theta_1$ is the parameter for freshman students, $\theta_2$ the parameter for sophomores, $\theta_3$ the parameter for juniors, and $\theta_4$ the parameter for seniors. Illustrating the impact of class standing on the distribution of anxiety, we might define a new variable

$$(\text{Class})_i = \begin{cases} 1 & \text{if i-th student is a freshman} \\ 2 & \text{if i-th student is a sophomore} \\ 3 & \text{if i-th student is a junior} \\ 4 & \text{if i-th student is a senior} \end{cases}$$

and then proceeded to say

$$(\text{Anxiety})_i \mid (\text{Class})_i, \theta \overset{\text{Ind}}{\sim} f\left((\text{Anxiety})_i \mid \theta_{(\text{Class})_i}\right). \tag{24.2}$$

Again, this was essentially our approach in Chapter 21, and it illustrates that we can allow the parameter to depend upon the predictor (we are doing regression). It just feels disjoint from the approach we have taken in Chapter 23. At first glance, we might be tempted to say

$$\begin{aligned} (\text{Anxiety})_i \mid (\text{Class})_i, \theta_i &\overset{\text{Ind}}{\sim} f\left((\text{Anxiety})_i \mid \theta_i\right) \\ \theta_i &= \beta_0 + \beta_1 (\text{Class})_i \end{aligned} \tag{24.3}$$

making use of the numeric variable $(\text{Class})_i$ we defined above. However, this imposes additional structure on the model. In particular, it suggests that the parameter $\theta$ changes *linearly* as we move between class standing (freshman to sophomore, junior to senior). Notice that the additional structure is captured by a reduced number of parameters. The original model in Equation 24.2 had four parameters: $\theta_1, \theta_2, \theta_3$ and $\theta_4$. The approach in Equation 24.3, however, only has two parameters: $\beta_0$ and $\beta_1$. These two models are clearly *not* equivalent. However, that does not mean that we cannot embed categorical predictors into the linear framework of Equation 24.1; we just need a different approach, and that approach involves indicator variables.

**Definition 24.1** (Indicator Variable)**.** An indicator variable is a binary variable (takes on the value 0 or 1), taking the value 1 when a specific event occurs. A collection of $k-1$ indicator variables can be used to capture a categorical variable with $k$ levels in a regression model.

- The "reference group" (or reference level) is the group (level) defined by setting all indicator variables in a regression model to 0.

For the Anxiety example, we define $4 - 1 = 3$ indicator variables to capture class status:

$$(\text{Sophomore})_i = \begin{cases} 1 & \text{if i-th student is a Sophomore} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Junior})_i = \begin{cases} 1 & \text{if i-th student is a Junior} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Senior})_i = \begin{cases} 1 & \text{if i-th student is a Senior} \\ 0 & \text{otherwise.} \end{cases}$$

We can then place these into the model following the format of Equation 24.1; specifically, we have

$$\begin{aligned} (\text{Anxiety})_i \mid \theta_i &\overset{\text{Ind}}{\sim} f\left((\text{Anxiety})_i \mid \theta_i\right) \\ \theta_i &= \beta_0 + \beta_1(\text{Sophomore})_i + \beta_2(\text{Junior})_i + \beta_3(\text{Senior})_i. \end{aligned} \tag{24.4}$$

It may at first seem that this model neglects the freshman class; however, the freshman class serves as the reference group (not sophomores or juniors or seniors) so the parameter for these students is represented by $\beta_0$. Comparing the modeling strategy in Equation 24.4 to that in Equation 24.2, we note that

$$\begin{aligned} \theta_1 &= \beta_0 \\ \theta_2 &= \beta_0 + \beta_1 \\ \theta_3 &= \beta_0 + \beta_2 \\ \theta_4 &= \beta_0 + \beta_3. \end{aligned}$$

Our revised model incorporates the categorical predictor while maintaining the general structure/complexity (we have not reduced the number of parameters).

> **Big Idea**
>
> A model which is linear in the parameters can accommodate categorical predictors through the inclusion of indicator variables.

**Example 24.2** (Rehabilitation Therapy Continued)**.** Example 23.1 described a study to investigate recovery time among patients who have undergone a corrective knee surgery. Suppose we are willing to believe that the mean recovery time is linearly related to the age of a patient, but we also believe that the mean recovery time may differ depending on the patient's level of activity prior to the surgery.

Further, suppose that we believe that given the age of the patient and their prior activity level, the recovery time will follow an Exponential distribution. Write down the likelihood for the observed data conditional on the age of the patient and their prior activity level.

*Solution.* Generalizing our solution in Example 23.2, we can say that

$$(\text{Recovery Time})_i \mid (\textbf{Predictors})_i, \beta \overset{\text{Ind}}{\sim} Exp\left(\theta_i\right)$$

where $\theta_i$, the scale term, will depend on the age and prior activity level of the patient, giving a likelihood of

$$\begin{aligned} f\left(\textbf{RecoveryTime} \mid \textbf{Predictors}, \beta\right) &= \prod_{i=1}^{n} \frac{1}{\theta_i} \exp\left\{-\frac{(\text{Recovery Time})_i}{\theta_i}\right\} \\ &= \left[\prod_{i=1}^{n} \frac{1}{\theta_i}\right] \exp\left\{-\sum_{i=1}^{n} \frac{(\text{Recovery Time})_i}{\theta_i}\right\}. \end{aligned}$$

It just remains to define the relationship between the predictors and the scale parameter $\theta_i$. If we adopt the linear structure of Equation 24.1, we have

$$\theta_i = \beta_0 + \beta_1(\text{Age})_i + \beta_2(\text{Moderate})_i + \beta_3(\text{High})_i,$$

where

$$(\text{Moderate})_i = \begin{cases} 1 & \text{if i-th patient has a moderate amount of activity prior to surgery} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{High})_i = \begin{cases} 1 & \text{if i-th patient has a high amount of activity prior to surgery} \\ 0 & \text{otherwise,} \end{cases}$$

leaving the "low activity" group as the reference group.

## 24.2 Curvature

Equation 24.1 is typically referred to as a "linear model." So, it may seem at first glance that such a model restricts us to only consider cases in which the predictor linearly impacts the parameter of interest (the mean response being linearly related to the predictor, for example). However, Equation 24.1 is linear *in the parameters*; and therefore, it is flexible enough to capture curvature.

While there are more sophisticated approaches, to simply illustrate the idea, we can incorporate curvature through higher-order terms. For example, the following model fits nicely within the linear framework of Equation 24.1:

$$\theta_i = \beta_0 + \beta_1(\text{Predictor 1})_i + \beta_2(\text{Predictor 1})_i^2.$$

This is linear in the parameters, even though it is not linear in the predictor. Essentially, we can simply define

$$(\text{Predictor 2})_i = (\text{Predictor 1})_i^2$$

and then it aligns directly with Equation 24.1.

> **Note**
>
> We need not restrict ourselves to polynomial terms. For example, it is possible to use cubic splines in regression models, and it has even been shown that carefully chosen splines can approximate nearly any form of curvature.

Of course, we are not restricted to only considering functions which are linear in the parameters; they tend to be commonly used for their simplicity of interpretation when scientific models do not suggest a particular form for the model. However, we could consider models like

$$\theta_i = \beta_0 e^{\beta_1(\text{Predictor})_i},$$

which are not linear in the parameters. These are also valid models and easy to make inference on in the Bayesian framework.

# 25 Default Priors in Regression Models

As the number of predictors we take into account grows, the number of unknown parameters governing the distribution of the response grows. Ideally, the prior distribution for each parameter would be elicited from the beliefs of discipline experts. However, with so many parameters, it can be difficult to elicit enough information from experts to form a joint prior distribution across all parameters. With so many parameters, there is often a demand for what to do in "no information" settings.

A popular approach at one time was to use a "spike and slab" prior that places a point mass at 0 mixed with a relatively flat distribution on the real line. For example, the prior

$$\pi\left(\beta_j\right) = 0.5\delta\left(\beta_j - 0\right) + 0.5\frac{1}{1000\sqrt{2\pi}}\exp\left\{-\frac{1}{1000^2}\left(\beta_j - 0\right)^2\right\}$$

mixes a point mass at 0 (with probability 0.5) with a Normal distribution centered at 0 with a standard deviation of 1000 (with probability 0.5). The large variance within the Normal distribution component of the prior distribution spreads the mass thinly across the real line.

These priors were once popular because they mixed the typical null distribution $\left(\beta_j = 0\right)$ with a vague prior that allowed the parameter to take nearly any value. However, there have been some recent recommendations against such priors.

The authors of the Stan programming language (which implements the Hamiltonian Monte Carlo approach) make the following suggestions regarding default priors.

1. Do not use vague priors.
2. Flat, bounded, priors are helpful when you have some idea of the range.
3. To conduct an analysis that is robust to outliers, use a Cauchy distribution for location parameters and a Half-Cauchy for scale parameters.
4. Given enough data, it is possible to use flat priors for default or sensitivity analyses.

The first suggestion comes from the idea that vague priors put a lot of weight on values we often do not believe are reasonable. However, others argue that vague priors are safer than unbounded flat priors because each puts weight on unreasonable values, but unbounded flat priors run the risk of producing an improper posterior distribution.

The second suggestion really highlights that often we have *some* information. If really pressed, we are often able to at least suggest unreasonable values for a parameter — an upper or lower bound, for example. This then suggests the possibility of a flat prior over a closed interval.

The Cauchy distribution is a bell-shaped distribution similar to a Normal distribution. However, while it is unimodal (with mode 0), it does not have a finite mean. Essentially, the distribution is so variable that it has no moments. This acts similarly to the idea behind a vague prior, placing most mass in a reasonable range (near 0), but still having a mass that extends out in both directions. A Half-Cauchy distribution limits the support to the positive real line.

A sensitivity analysis allows us to see how dependent our results are on the choice of prior. If our results would change dramatically depending on the choice of prior, that at a minimum, warrants a discussion.

**Example 25.1** (Rehabilitation Therapy Continued)**.** Example 23.1 described a study to investigate recovery time among patients who have undergone a corrective knee surgery. In Example 24.2, we developed a model for the distribution of the response given the age and prior activity level of the patient:

$$(\text{Recovery Time})_i \mid (\textbf{Predictors})_i, \beta \overset{\text{Ind}}{\sim} Exp\left(\theta_i\right)$$

where $\theta_i$, the scale term, will depend on the age and prior activity level of the patient, through the function

$$\theta_i = \beta_0 + \beta_1(\text{Age})_i + \beta_2(\text{Moderate})_i + \beta_3(\text{High})_i,$$

where

$$(\text{Moderate})_i = \begin{cases} 1 & \text{if i-th patient has a moderate amount of activity prior to surgery} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{High})_i = \begin{cases} 1 & \text{if i-th patient has a high amount of activity prior to surgery} \\ 0 & \text{otherwise.} \end{cases}$$

Specify reasonable default priors for each of the unknown parameters.

*Solution.* As discussed in Example 23.2, the mean response for the Exponential distribution should be positive. One way to ensure a positive mean response is to ensure each parameter is positive. Therefore, a reasonable default prior could be a Half-Cauchy distribution.

# 26 QR Factorization

As a regression model grows in complexity, we need to consider the computational efficiency of the algorithms used to fit the model. QR factorization is a well-known computational step that can increase the efficiency of an MCMC algorithm in a regression model.

Consider a regression model in which the parameter $\theta$ is allowed to vary according to a linear function of the predictors:

$$\theta_i = \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

We have eliminated the "intercept" term $\beta_0$, but this is done without loss of generality as we could view "Predictor 1" as the value 1 for all subjects, resulting in $\beta_1$ acting as the intercept. We can represent this model in matrix notation as

$$\theta = \mathbf{X}\beta,$$

where $\beta = \left(\beta_1, \ldots, \beta_p\right)^\top$ is the parameter vector; and, the $j$-th column of $\mathbf{X}$ is the vector of length $n$ containing the values of the $j$-th predictor for the $n$ subjects. That is,

$$\mathbf{X}_{i,j} = (\text{Predictor } j)_i \qquad j = 1, 2, \ldots, p.$$

The matrix $\mathbf{X}$ is known as the design matrix.

As long as the number of observations $n$ exceeds the number of predictors $p$ in the model, we can decompose the design matrix $\mathbf{X}$ as

$$\mathbf{X} = \mathbf{QR}$$

where $\mathbf{Q}$ is an orthogonal $n \times p$ matrix and $\mathbf{R}$ is an upper-triangular $p \times p$ matrix. Then, we can write the linear predictor for the mean response as

$$\mathbf{X}\beta = \mathbf{QR}\beta.$$

Therefore, the regression is conducted with the "new" design matrix $\mathbf{Q}$ with parameters $\eta = \mathbf{R}\beta$. These parameters are then transformed back into the parameters of interest $\beta$ by acknowledging that

$$\beta = \mathbf{R}^{-1}\eta.$$

Admittedly, this feels like only algebraic manipulation. The reason this works is that the columns of the "new" design matrix $\mathbf{Q}$ are orthogonal. This allows the MCMC algorithm to move more easily through the parameter space because changing one column has no effect on the other columns in the optimization routine. The columns of this new design matrix are also on the same scale; that is, the impact of one variable (like yearly income) taking on extreme values while another (like an indicator variable) taking on smaller units is reduced. Having variables on the same scale allows the MCMC algorithm to move around the parameter space with a small number of large steps. As a result, the numerical accuracy of the algorithm is improved.

> Big Idea
>
> QR factorization improves the computational efficiency of the regression and MCMC algorithms. There is no change to the actual distribution of the parameters.

# 27 Assessment for Regression Models for the Mean

While we have tried to emphasize the flexibility of regression models, the most common regression model is one of the form

$$\begin{aligned} (\text{Response})_i \mid (\text{Predictors})_i, \beta, \theta &\overset{\text{Ind}}{\sim} f(\mu_i, \theta) \\ \mu_i &= \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i, \end{aligned} \tag{27.1}$$

where $\mu_i$ represents the (conditional) mean response and $\theta$ captures additional parameters that do not depend on the predictors (often scale parameters). Equation 27.1 represents a model focused on the mean response.

It is reasonable to ask if the model we have constructed is reasonable for the data we have observed. When our models are of the form described in Equation 27.1, a large assumption is that we have correctly specified the mean response. This can be assessed graphically using *residuals*.

**Definition 27.1** (Residual)**.** A residual is the difference between an observed response and the predicted mean response for that same individual:

$$(\text{Residual})_i = (\text{Response})_i - (\text{Fitted Value})_i,$$

where

$$(\text{Fitted Value})_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j (\text{Predictor} j)_i.$$

> **Note**
>
> In a classical introductory statistics course, the least squares estimates are used when computing the residuals. However, from the Bayesian framework, we have explored various point estimates. It is common to use the posterior mean for each parameter when

computing the residuals; however, nothing prohibits the use of the posterior median or another point estimate. Transparency is critical; you should be clear about the analysis you have conducted.

In order to assess that the form of the mean model is correctly specified, it is common to construct a graphic of the residuals against the fitted values. If the form of the mean model is correct, there should not be any distinguishable pattern in the *location* of the graphic. Trends in the location suggest the form of the mean model is incorrectly specified (see Figure 27.1).

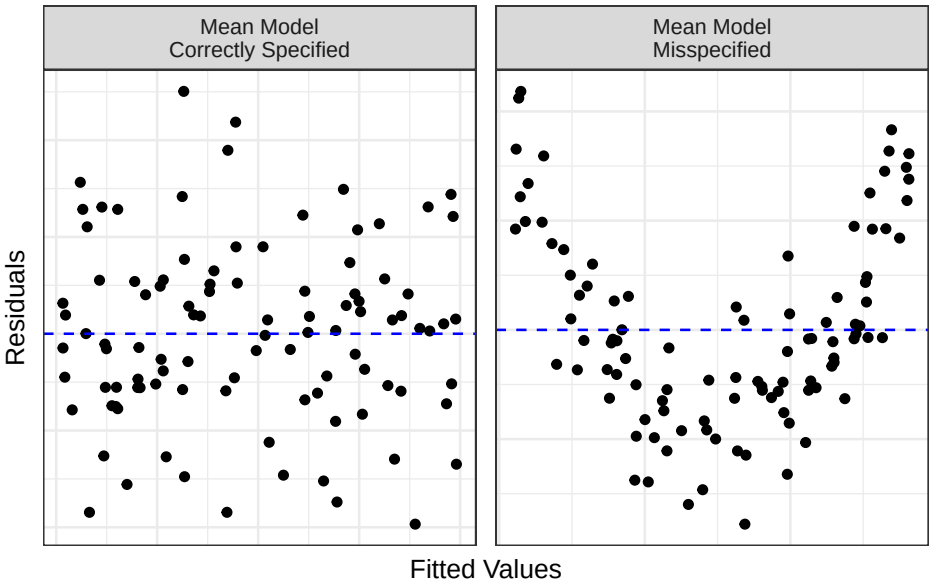


Figure 27.1: Plot of the residuals against the fitted values for two hypothetical models. One illustrates what we would expect under a correctly specified mean model; the second is an example of a graphic indicating the mean model is misspecified.

**i** Assessing Specification of Mean Response

If the mean response is correctly specified, we would expect the residuals to balance around 0, regardless of the estimated mean response. When examining a plot of the residuals against the fitted values, any trends in the location suggest the functional form of the mean response has been incorrectly specified.

**Example 27.1** (Rehabilitation Therapy Continued)**.** Example 23.1 described a study to investigate recovery time among patients who have undergone a corrective knee surgery. Suppose we are willing to believe that the mean recovery time is linearly related to the age of a patient. Combining the model for the likelihood suggested in Example 23.2 and the advice on default priors specified in Chapter 25, consider the following model:

$$
\begin{aligned}
(\text{Recovery Time})_i \mid (\text{Age})_i, \beta &\overset{\text{Ind}}{\sim} Exp\left(\theta_i\right) \\
\theta_i &= \beta_0 + \beta_1 (\text{Age})_i \\
\beta_0 &\sim Unif(0, 25) \\
\beta_1 &\sim Unif(0, 5).
\end{aligned}
$$

This model was fit using an MCMC algorithm with 3 chains; a burn-in of 2000 was applied to each of the chains, and a total of 5000 samples were generated for each chain (for a total of 9000 variates after the burn-in period). The posterior mean was used to estimate each of the unknown parameters. Figure 27.2 presents the plot of the residuals against the fitted values for this model. Comment on the assumption that the mean response is properly specified.

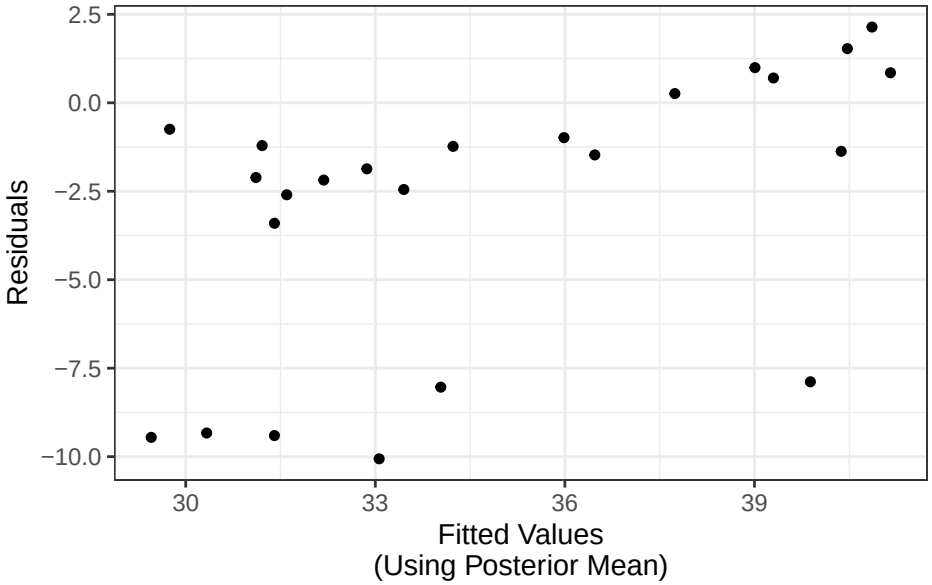


Figure 27.2: Assessment of the mean response model for the Therapy example.

*Solution.* If the mean response model is correctly specified, we would expect the residuals to balance around 0 for all fitted values. Notice that the residuals are centered below for the majority of the graphic, and only balance around 0 for large fitted values. This trend in the location of the residuals suggests the mean response model was not correct specified.

In particular, this slight upward trend suggests that perhaps we were incorrect in forcing the intercept to be positive (remember, our prior distribution forced the support for the intercept to be positive). We had done this because the mean response for an exponential distribution must always be positive. However, because of extrapolation, forcing this to be the case at an age of 0 seems to be problematic. There are two approaches we could consider in addressing this:

- We could consider a different prior that allows the intercept to be negative, accepting that it is nonsensical and that the model will not predict well for small values of age.
- We could center the age variable (by subtracting the average observed age from each observation). Center the age variable does not impact the slope, but it changes the interpretation of the intercept. In particular, the intercept would represent the average recovery time for a patient of average age. This avoids the problem of extrapolation when interpreting the intercept and might address the problems we are seeing above.

The other primary assumption that we make when fitting a regression model is that the conditional distribution of the response is appropriate. In Example 27.1, for example, we are assuming that the Exponential distribution for the response (conditional on the age) is appropriate, as opposed to a Normal distribution, for example. One technique for assessing whether this distributional assumption is appropriate is to compare the posterior predictive distribution with the observed distribution. As we have seen, the posterior predictive distribution (Definition 14.2) can be challenging to derive; fortunately, it is easily simulated using a sample from the posterior distribution. With the general model of Equation 27.1 in mind, we can generate the posterior predictive distribution as follows:

- Obtain a sample of size $M$ from the posterior distribution of each unknown parameter: $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(M)}$ and $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$.
- For each sample from the posterior, generate a new *sample* of $n$ responses according to $(\text{Predicted Response})_i^{(m)} \sim f\left(\mu_i^{(m)}, \theta^{(m)}\right)$ for $m = 1, 2, \ldots, M$. If we are conditioning on the predictors, they are taken to be those from the original sample.

The above two steps produces $M$ new samples; each generated sample will produce a unique distribution of the responses across the $n$ observations. We can summarize each of these $M$ distributions using a density plot, overlaid on the same graphic. Then, we can overlay the density from the observed response to get a sense of how they compare. Figure 27.3 gives an example of what this plot might look like.

> **Note**
>
> Due to the computational intensity of this graphic, it is common to do this for a random sample of variates instead of all $M$ variates generated by the MCMC algorithm.

> **Warning**
>
> It is important to remember that comparing the posterior predictive distribution to that of the observed distribution is combining multiple conditions/assumptions together: the complete form of the distribution as well as how the individual observations vary compared to how the aggregate dataset varies. While our modeling is conditioned, the density of the observed response marginalizes across the predictors. Care must be taken

not to over-interpret this graphic as proving we have the correct model.

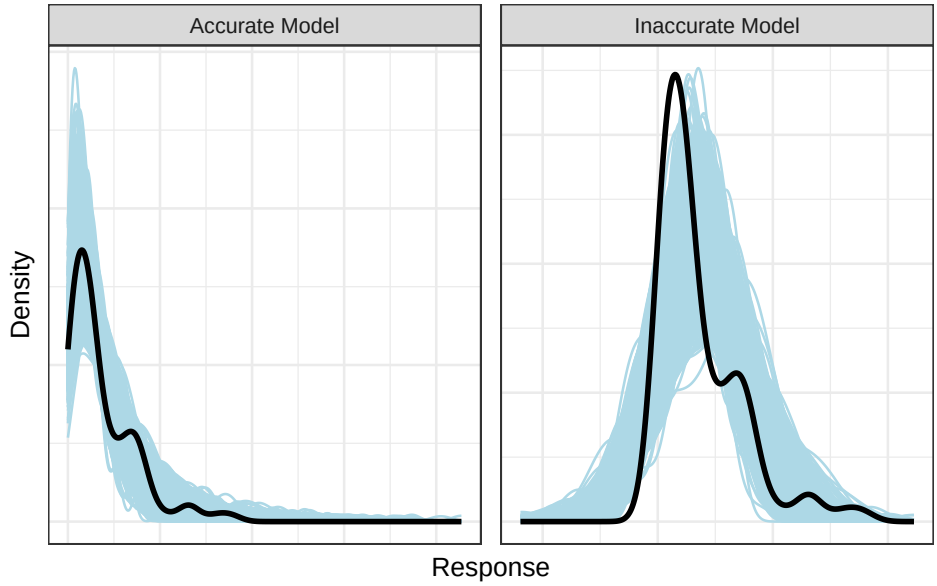


Figure 27.3: Two plots of the posterior predictive distributions from a regression model with the observed distribution. One illustrates what we would expect when the distribution accurately captures the process; the second is an example of a graphic indicating the distributional assumptions are incorrect.

**Assessing the Likelihood**

If the model for the likelihood is correctly specified, then the marginal distribution of the observed response should be similar to the posterior predictive distribution given the observed data. For a regression model, this is done by generating several *samples* of the same size given the posterior variates; if the distributional model is appropriate, a density plot of the observed response should line up with the density plots of those samples generated from the posterior variates. Any major differences in these shapes would suggest *some* aspect of the model (including the distributional form) is incorrect.

**Example 27.2** (Rehabilitation Therapy Continued)**.** Example 27.1 presented a model for the study described in Example 23.1. Figure 27.4 is a plot of the posterior predicted distribution of the responses (using 250 randomly selected posterior variates) against the observed distribution of the response. Comment on the assumption that the distributional model specified is appropriate.

*Solution.* The observed marginal distribution of the response is very different than the pre-

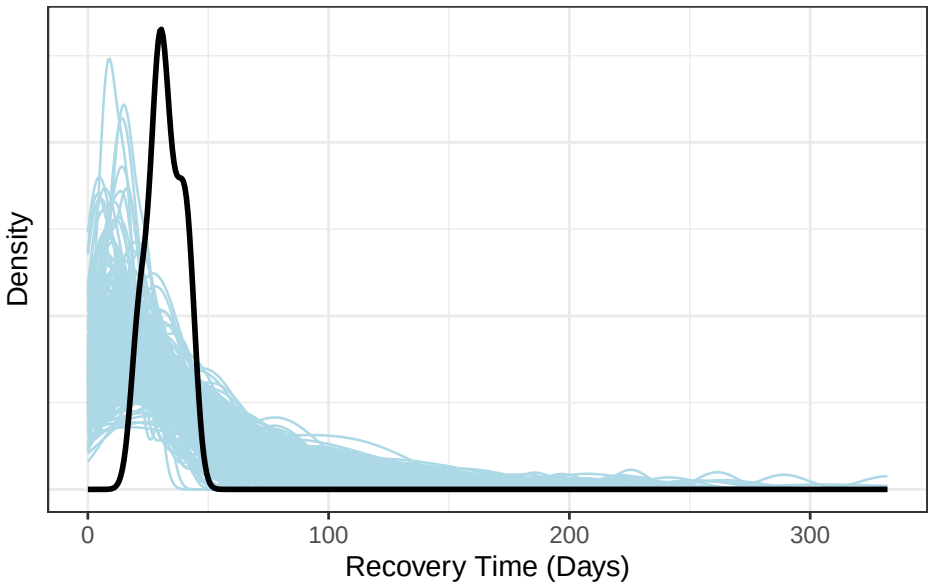


Figure 27.4: Assessment of the distributional assumptions for the Therapy example.

dicted distributions. While this could be due in part to the misspecification of the mean response model (as noted in Example 27.1), the level of departure here suggests the distributional assumption on the likelihood is also incorrect. If we wanted to be sure, we should refit the model using a different mean response function first, and then repeat this process.

# 28 Regression Models for Categorical Responses

On the one hand, modeling a categorical response (where the response follows a discrete distribution) is no different than modeling a quantitative response (where the response follows a continuous distribution). In both cases, we specify a distribution, and we allow one or more parameters to vary across individuals based on their predictors through some pre-specified function. On the other hand, the modeling is quite different as there are often more pitfalls to be aware of. In particular, we are no longer in a position to think of the model as

$$(\text{Response})_i = (\text{Signal})_i + (\text{Noise})_i.$$

As an example, if the response is binary, what type of noise could be added to "jitter" the response? A binary response must always take the value 0 or 1, which implies that thinking of the response as a "jittered" signal seems somehow inauthentic. The key to extending the regression model to categorical response variables is to view regression as an extension as a more complex specification of the conditional response.

## 28.1 Considerations for a Binary Response

Suppose our response is binary, taking the value 1 when an event of interest occurs (a "success") and taking the value 0 when the event does not occur (a "failure"). A starting point for this model is

$$(\text{Response})_i \overset{\text{Ind}}{\sim} Ber\left(\theta_i\right),$$

where we are allowing the probability of success $\theta_i$ to potentially vary from one observation to the next. This model continues to assume the response from one individual is independent of the response from any other individual. Our initial attempt at generalizing to a regression setting may be to borrow the linear model structure of the previous chapters and consider

$$\theta_i = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

Unfortunately, this can be a poor strategy. Notice that there is nothing ensuring that this linear function produces values of $\theta_i$ which are consistent with its support (this is the same problem we encountered in Example 23.2). That is, we know that since $\theta_i$ is a probability, its support is the interval $(0, 1)$. The linear function, however, could quite easily produce values which are negative or exceed 1; instead, we desire a function that is always bounded between 0 and 1, similar to that represented in Figure 28.1.

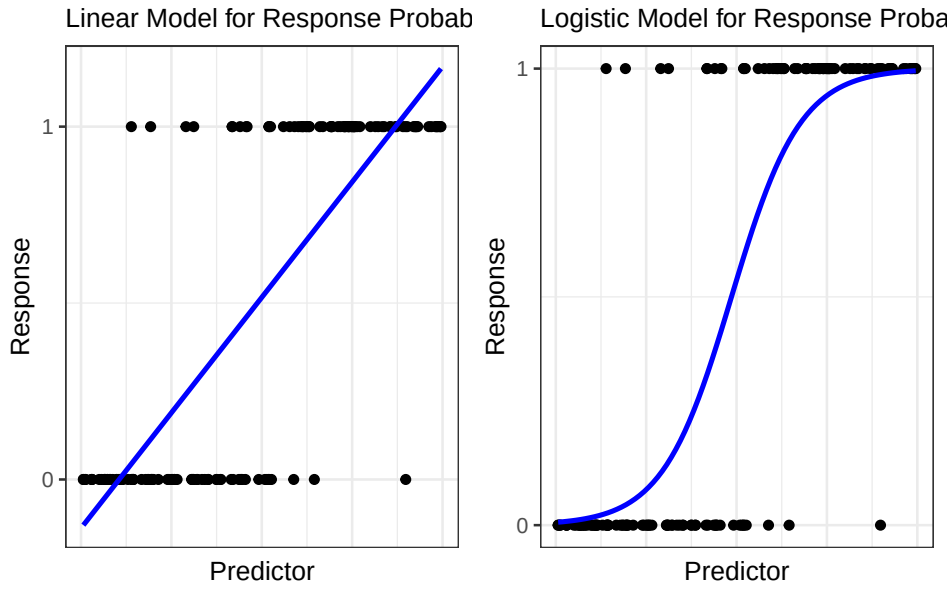


Figure 28.1: Comparison of two functional forms of a single predictor for the success probability in a regression for a binary response. The first indicates the problems with a linear functional form, while the second illustrates a function that is bounded on the correct support.

One of the most common choices for the functional form is the *logistic function*.

**Definition 28.1** (Logistic Regression)**.** Given a binary response; logistic regression assumes that

$$(\text{Response})_i \mid (\text{Predictors})_i, \beta \overset{\text{Ind}}{\sim} Ber\left(\theta_i\right)$$
$$\theta_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}.$$

This is known as logistic regression because the functional form of the probability matches that of the CDF of the Standard Logistic Distribution:

$$\frac{e^x}{1+e^x}.$$

> **Note**
>
> When performing regression with a binary response, any CDF could be used for the functional form relating the predictors to the success probability; however, the Standard Logistic ("logistic regression") and Standard Normal ("probit regression") distributions are most common.

Using a CDF for the functional form ensures that for any choice of $\beta$ and the predictors, the function will always take a value between 0 and 1. Notice that our choice of $\theta_i$ is *nonlinear* in the parameters $\beta$; however, it still has that feel of a linear model because of the *linear predictor*

$$\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

Nothing prohibits us from considering a nonlinear function instead of the linear predictor; it is just common practice to use the linear predictor, as we have already seen it is flexible enough to accommodate categorical predictors and curvature.

For a logistic regression, the $j$-th coefficient $\beta_j$ is the log odds ratio of the response occurring when the $j$-th predictor is increased by one unit compared to its current value, holding all other predictors fixed. For probit regression, the $j$-th coefficient $\beta_j$ has no intuitive interpretation (hence the popularity of logistic regression).

## 28.2 Considerations for Count Data

As in the binary response setting, we take care to ensure that the functional form relating the predictors to key parameters in the conditional distribution of the response enforces constraints on the support.

Consider a response which counts the number of successes out of a fixed number of trials. We might consider a model of the form

$$\begin{aligned} (\text{Response})_i \mid (\text{Predictors})_i, \beta &\overset{\text{Ind}}{\sim} Bin\left(m_i, \theta_i\right) \\ \log\left(\frac{\theta_i}{1-\theta_i}\right) &= \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i. \end{aligned}$$

While we have written this in a slightly different form, we are using the same functional form for linking the response probability to the predictors; here, we have written it in terms of the *link function*.

**Definition 28.2** (Link Function)**.** The functional form "linking" the linear predictor

$$\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i$$

to the mean response of the model in a regression model. In particular, it is the function $g$ such that

$$g(\theta_i) = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

Common link functions include:

- Identity link: $g(\theta_i) = \theta_i$
- Logit link: $g(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right)$
- Log link: $g(\theta_i) = \log(\theta_i)$
- Inverse link: $g(\theta_i) = \frac{1}{\theta_i}$
- Negative inverse link: $g(\theta_i) = -\frac{1}{\theta_i}$
- Inverse squared link: $g(\theta_i) = \frac{1}{\theta_i^2}$

The logistic distribution function is chosen since $\theta_i$ should be constrained to the interval $(0, 1)$. Notice that using the logit link for the success probability in the Binomial distribution on the response does not specify the mean response directly; instead, it is simply used to link the predictors to the mean response such that

$$E[(\text{Response}) \mid (\text{Predictors}), \beta] = m \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)}},$$

where $m$ is the number of trials.

> **Note**
>
> Any distribution for which the mean response is defined through the "success" probability (Bernoulli, Binomial, Geometric) could potentially make use of the logit link.

The Poisson distribution is common for modeling the number of rare events in a large population, or for arbitrary counts that have no upper bound. Extending this into a regression model generally takes the form

$$(\text{Response})_i \mid (\text{Predictors})_i, \beta \overset{\text{Ind}}{\sim} Poisson\left(\lambda_i\right)$$
$$\log\left(\lambda_i\right) = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

The log link function is chosen to ensure that $\lambda > 0$ for all possible choices of the parameter vector $\beta$ and predictors.

In each of the cases considered in this chapter, the model specifies the mean response; however, it also specifies the variability in the response. For example, the mean of a Poisson distribution is $\lambda_i$, and the variance is also given by $\lambda_i$. These distributional models are unique in that knowing the mean response uniquely determines the variability in the response. Occasionally, we come across data for which the response behaves *nearly* like one of these distributions; however, additional variability is present. There are ways of generalizing these models to account for this additional dispersion.

# References


Doyle, Sir Arthur Conan. 1890. *The Sign of the Four*. Spencer Blackett.

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.

Goldstein, Bernard D, Howard J Osofsky, and Maureen Y Lichtveld. 2011. "The Gulf Oil Spill." *The New England Journal of Medicine* 364: 1334–48. https://doi.org/10.1056/NEJMra1007197.

Johnson, Eric J, and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science* 302: 1338–39.

Kruschke, John K. 2015. *Doing Bayesian Data Analysis: A Tutorial with r, JAGS, and Stan*. 2nd ed. Elsevier.

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

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

# A Glossary

The following key terms were defined in the text; each term is presented with a link to where the term was first encountered in the text.

**Alternative Hypothesis (Definition 5.10)** The statement (or theory) about the parameter 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.

**Average (Definition 7.2)** Also known as the "mean," this measure of location represents the balance point for the distribution. If $x_i$ represents the $i$-th value of the variable $x$ in the sample, the sample mean is typically denoted by $\bar{x}$.

For a sample of size $n$, it is computed by

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

When referencing the average for a population, the mean is also called the "Expected Value," and is often denoted by $\mu$.

**Axioms of Probability (Definition 1.3)** Let $\mathcal{S}$ be the sample space of a random process. Suppose that to each event $A$ within $\mathcal{S}$, a number denoted by $Pr(A)$ is associated with $A$. If the map $Pr(\cdot)$ satisfies the following three axioms, then it is called a **probability**:

1. $Pr(A) \geq 0$
2. $Pr(\mathcal{S}) = 1$
3. If $\{A_1, A_2, \ldots\}$ is a sequence of mutually exclusive events in $\mathcal{S}$, then

$$Pr\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} Pr\left(A_i\right).$$

$Pr(A)$ is said to be the "probability of $A$" or the "probability $A$ occurs."

**Bayes Factor (Definition 15.2)** A measure of how the observed data *alters* your prior beliefs about a hypothesis. Let $H_j$ denote the hypothesis that $\theta \in \Theta_j$ for some region $\Theta_j$. The Bayes Factor *in favor of* $H_j$ is the ratio of the posterior odds in favor of $H_j$ to the prior odds in favor of $H_j$:

$$BF_j = \left(\frac{Pr\left(\theta \in \Theta_j \mid \mathbf{y}\right)}{Pr\left(\theta \notin \Theta_j \mid \mathbf{y}\right)}\right)\left(\frac{Pr\left(\theta \notin \Theta_j\right)}{Pr\left(\theta \in \Theta_j\right)}\right).$$

**Bayes Factor for Model Comparison (Definition 15.5)** The Bayes Factor, in favor of Model 1, is

$$\begin{aligned} BF_1 &= \left(\frac{Pr(\mathcal{M}_1 \mid \mathbf{y})}{Pr(\mathcal{M}_0 \mid \mathbf{y})}\right)\left(\frac{Pr(\mathcal{M}_0)}{Pr(\mathcal{M}_1)}\right) \\ &= \left(\frac{f_1(\mathbf{y} \mid \mathcal{M}_1)Pr(\mathcal{M}_1)}{f_0(\mathbf{y} \mid \mathcal{M}_0)Pr(\mathcal{M}_0)}\right)\left(\frac{Pr(\mathcal{M}_0)}{Pr(\mathcal{M}_1)}\right) \\ &= \frac{f_1(\mathbf{y} \mid \mathcal{M}_1)}{f_0(\mathbf{y} \mid \mathcal{M}_0)}. \end{aligned}$$

That is, the Bayes Factor is a ratio of the evidence for each model.

**Bias (Definition 6.1)** A set of measurements is said to be biased if they are *consistently* too high (or too low). Similarly, an estimate of a parameter is said to be biased if it is *consistently* too high (or too low).

**Blocking (Definition 20.7)** Blocking is a way of minimizing the variability contributed by an inherent characteristic that results in dependent observations. In some cases, the blocks are the unit of observation which is sampled from a larger population, and multiple observations are taken on each unit. In other cases, the blocks are formed by grouping the units of observations according to an inherent characteristic; in these cases that shared characteristic can be thought of having a value that was sampled from a larger population.

In both cases, the observed blocks can be thought of as a random sample; within each block, we have multiple observations, and the observations from the same block are more similar than observations from different blocks.

**Bridge Sampling (Definition 21.1)** The bridge sampling estimator of the marginal likelihood $m(\mathbf{y})$ is given by

$$\begin{aligned} m(\mathbf{y}) &= \int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta \\ &= \frac{E_g\left[h(\theta)f(\mathbf{y} \mid \theta)\pi(\theta)\right]}{E_\pi\left[h(\theta)g(\theta)\right]} \\ &\approx \frac{m^{-1}\sum_{j=1}^{m} h\left(\tilde{\theta}_j\right) f\left(\mathbf{y} \mid \tilde{\theta}_j\right) \pi\left(\tilde{\theta}_j\right)}{m^{-1}\sum_{i=1}^{m} h\left(\theta_j^*\right) g\left(\theta_j^*\right)} \end{aligned}$$

where $h(\theta)$ is called the bridge function and $g(\theta)$ is the proposal distribution. Here, $\tilde{\theta}$ denotes a random variate from the proposal distribution and $\theta^*$ a random variate from the posterior; $E_g$ denotes taking an expectation with respect to the proposal distribution and $E_\pi$ denotes taking an expectation with respect to the posterior distribution.

**Categorical Variable (Definition 3.5)** Also called a "qualitative variable," a measurement on a subject which denotes a grouping or categorization.

**Codebook (Definition 3.7)** Also called a "data dictionary," these provide complete information regarding the variables contained within a dataset.

**Conditional Density (Definition 9.3)** Let $X$ and $Y$ be two random variables; the conditional density of $X$ given $Y$ is

$$f_{X|Y}(y \mid x) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

**Conditional Density (Definition 9.3)** Let $\mathbf{X}$ be a random vector; without loss of generality, partition $\mathbf{X}$ such that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$$

where $\mathbf{X}_1$ represents the first $k$ components and $\mathbf{X}_2$ represents the remaining $n-k$ components. Then, the conditional density of $\mathbf{X}_1$ given $\mathbf{X}_2$ is

$$f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1 \mid \mathbf{x}_2) = \frac{f_{\mathbf{X}}(\mathbf{x})}{f_{\mathbf{X}_2}(\mathbf{x}_2)}.$$

**Conditional Independence (Definition 14.3)** Two random variables $X$ and $Y$ are said to be independent, conditional on (or "given") $Z$ if, and only if,

$$f_{(X,Y)|Z}(x, y \mid z) = f_{X|Z}(x \mid z) f_{Y|Z}(y \mid z).$$

**Confounding (Definition 20.3)** When the effect of a variable on the response is misrepresented due to the presence of a third, potentially unobserved, variable known as a confounder.

**Conjugate Prior (Definition 16.1)** A prior distribution chosen such that the posterior distribution belongs to the same family as the prior distribution, with the (hyper)parameters that govern the family updated based on the observed data.

**Continuous and Discrete Random Variable (Definition 2.3)** The random variable $X$ is said to be a discrete random variable if its corresponding support is countable. The random variable $X$ is said to be a continuous random variable if the corresponding support is uncountable (such as an interval or a union of intervals on the real line).

**Controlled Experiment (Definition 20.2)** A study in which each subject is *randomly* assigned to one of the groups being compared in the study.

**Credible Interval (Definition 13.3)** A $100c\%$ credible interval is an interval $(a, b)$ such that

$$Pr(a \leq \theta \leq b \mid \mathbf{y}) = \int_a^b \pi(\theta \mid \mathbf{y}) d\theta = c.$$

**Cumulative Distribution Function (CDF) (Definition 2.10)** Let $X$ be a random variable; the cumulative distribution function (CDF) is defined as

$$F(u) = Pr(X \leq u).$$

For a continuous random variable, we have that

$$F(u) = \int_{-\infty}^{u} f(x) dx$$

implying that the density function is the derivative of the CDF. For a discrete random variable

$$F(u) = \sum_{x \leq u} f(x).$$

**Density Function (Definition 2.4)** A density function $f$ relates the values in the support of a random variable with the probability of observing those values.

Let $X$ be a continuous random variable, then its density function $f$ is the function such that

$$Pr(a \leq X \leq b) = \int_a^b f(x) dx$$

for any real numbers $a$ and $b$ in the support.

Let $X$ be a discrete random variable, then its density function $f$ is the function such that

$$Pr(X = u) = f(u)$$

for any real number $u$ in the support.

**Dirac Delta Function (Definition 10.3)** The Dirac delta function is the function (not in a rigorous sense) $\delta$ such that

$$\int_{-\infty}^{\infty} \delta(x)dx = 1$$

and

$$\int_{-\infty}^{\infty} f(x)\delta(x)dx = f(0)$$

for any real-valued function $f$.

The Dirac delta function allows us to describe a discrete distribution, which places mass at a single point, as a continuous function on the real line.

**Distribution (Definition 5.3)** The pattern of variability corresponding to a set of values.

**Distribution of the Population (Definition 7.9)** The pattern of variability in values of a variable at the population level. Generally, this is impossible to know, but we might model it.

**Distribution of the Sample (Definition 7.6)** The pattern of variability in the observed values of a variable.

**Effective Sample Size (Definition 19.1)** The effective sample size (ESS) is given by

$$ESS = \frac{N}{1 + 2\sum_{k=1}^{\infty} ACF(k)}$$

where ACF is the auto-correlation function of degree $k$.

**Equal-Tailed Credible Interval (Definition 13.4)** The equal-tailed credible interval, which is probably the most commonly used in practice, chooses endpoints such that

$$Pr(\theta < a \mid \mathbf{y}) = \frac{1-c}{2} = Pr(\theta > b \mid \mathbf{y}).$$

**Estimation (Definition 5.7)** Using the sample to approximate the value of a parameter from the underlying population.

**Event (Definition 1.2)** A subset of the sample space that is of particular interest.

**Evidence for a Model (Definition 15.4)** Under the Model Comparison framework defined above, the evidence for model $\mathcal{M}_j$ is defined as

$$f_j(\mathbf{y} \mid \mathcal{M}_j) = \int f_j(\mathbf{y} \mid \theta_j, \mathcal{M}_j)\pi_j(\theta_j \mid \mathcal{M}_j)d\theta_j.$$

**Expectation of a Function (Definition 2.7)** Let $X$ be a random variable with density function $f$ over the support $\mathcal{S}$, and let $g$ be a real-valued function. Then,

$$E\left[g(X)\right] = \int_{\mathcal{S}} g(x)f(x)dx$$

for continuous random variables and

$$E\left[g(X)\right] = \sum_{\mathcal{S}} g(x)f(x)$$

for discrete random variables.

**Expected Value (Mean) (Definition 2.5)** Let $X$ be a random variable with density function $f$ defined over the support $\mathcal{S}$. The expected value of a random variable, also called the mean and denoted $E(X)$, is given by

$$E(X) = \int_{\mathcal{S}} xf(x)dx$$

for continuous random variables and

$$E(X) = \sum_{\mathcal{S}} xf(x)$$

for discrete random variables.

**Extrapolation (Definition 23.2)** Extrapolation occurs when we use a model to predict outside of the region for which data is available.

**Frequency (Definition 5.4)** The number of observations in a sample falling into a particular group (level) defined by a categorical variable.

**Frequentist Interpretation of Probability (Definition 1.5)** In this perspective, the probability of $A$ describes the long-run behavior of the event. Specifically, consider repeating the random process $m$ times, and let $f(A)$ represent the number of times the event $A$ occurs out of those $m$ replications. Then,

$$Pr(A) = \lim_{m\to\infty} \frac{f(A)}{m}.$$

**General Mixture Distribution (Definition 16.3)** Let $\theta$ be a parameter with support $\Theta$, and let $\pi_k(\theta)$ be a valid distribution on the support, for $k = 1, 2, \ldots, K$. Then,

$$\pi(\theta) = \sum_{k=1}^{K} w_k \pi_k(\theta)$$

is a valid prior distribution provided $\sum_{k=1}^{K} w_k = 1$.

**Highest Density Interval (Definition 13.5)** The highest density interval, often called an HDI or HPD (for highest posterior density), chooses the endpoints such that the interval is as short as possible.

When the density is unimodal, this can be accomplished by choosing the endpoints $a$ and $b$ such that

$$\pi(\theta \mid \mathbf{y}) \mid_{\theta=a} = \pi(\theta \mid \mathbf{y}) \mid_{\theta=b}$$

and

$$\int_a^b \pi(\theta \mid \mathbf{y} d\theta = c.$$

**Histogram Approach to Constructing a Prior (Definition 16.2)** Using expert information, attach probability to various intervals for the parameter. Specifically,

- Define $m$ intervals $\left(\theta_{j-1}, \theta_j\right)$ for $j = 1, 2, \ldots, m$ that partition the parameter space; define $\theta_0$ as the lower bound of the support for the parameter, and define $\theta_m$ as the upper bound of the support for the parameter.
- Eliciting expert opinions, assign probability $\pi_j$ to each interval: $\pi_j = Pr\left(\theta_{j-1} < \theta < \theta_j\right)$ for each $j = 1, 2, \ldots, m$.
- Set the prior $\pi(\theta)$ to be the piecewise distribution over this interval where $\sum_{j=1}^{m} \pi_j = 1$.

**Hyperparameter (Definition 10.2)** A constant term of a prior distribution that characterizes the family we are considering.

**Hypothesis Testing (Definition 5.8)** 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.

**Identically Distributed (Definition 9.5)** We say that random variables $X$ and $Y$ are identically distributed if $F_X(u) = F_Y(u)$ for all $u$. This is equivalent to saying the two random variables have the same density function $f$.

**Independence (Definition 9.4)** Random variables $X_1, X_2, \ldots, X_n$ are said to be mutually independent (or just "independent") if and only if

$$Pr\left(X_1 \in A_1, X_2 \in A_2, \cdots, X_n \in A_n\right) = \prod_{i=1}^{n} Pr\left(X_i \in A_i\right),$$

where $A_1, A_2, \ldots, A_n$ are arbitrary sets. Perhaps more helpful, $X_1, X_2, \ldots, X_n$ are said to be mutually independent if and only if

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} f_{X_i}\left(x_i\right).$$

**Indicator Variable (Definition 24.1)** An indicator variable is a binary variable (takes on the value 0 or 1), taking the value 1 when a specific event occurs. A collection of $k-1$ indicator variables can be used to capture a categorical variable with $k$ levels in a regression model.

- The "reference group" (or reference level) is the group (level) defined by setting all indicator variables in a regression model to 0.

**Interquartile Range (Definition 7.5)** Often abbreviated as IQR, this is 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.

**Interval Estimation (Definition 13.2)** Interval estimation is the process of estimating a parameter with a range of values. This is like trying to capture a target with a ring.

**Joint Density (Definition 9.1)** For a random vector $\mathbf{X}$, the function $f_{\mathbf{X}}(\mathbf{x})$ such that for any set $A \in \mathbb{R}^n$, we have

$$Pr(\mathbf{X} \in A) = \int \cdots \int_{A} f_{\mathbf{X}}(\mathbf{x}) dx_1 \cdots dx_n$$

is called the joint density function; this is also referred to as the *likelihood*. Integrals are replaced by sums when appropriate.

**Kernel of a Distribution (Definition 2.9)** Let $k(x)$ be a non-negative function of $x$ over some region $\mathcal{S}_X$. Then, a valid density function $f$ over the support $\mathcal{S}_X$ can be constructed by taking

$$f(x) = ak(x)$$

where $a > 0$ is a suitably chosen scaling constant to ensure the density integrates (or sums) to 1 over the support. The function $k$ is known as the kernel of the distribution, and it can be used to identify the distributional family for a random variable.

**Laplace Prior (Definition 16.4)** The Laplace prior, also known as a "flat" prior, considers the form

$$\pi(\theta) = 1 \qquad \forall \theta \in \Theta.$$

**Link Function (Definition 28.2)** The functional form "linking" the linear predictor

$$\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i$$

to the mean response of the model in a regression model. In particular, it is the function $g$ such that

$$g(\theta_i) = \beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i.$$

Common link functions include:

- Identity link: $g(\theta_i) = \theta_i$
- Logit link: $g(\theta_i) = \log\left(\frac{\theta_i}{1-\theta_i}\right)$
- Log link: $g(\theta_i) = \log(\theta_i)$
- Inverse link: $g(\theta_i) = \frac{1}{\theta_i}$
- Negative inverse link: $g(\theta_i) = -\frac{1}{\theta_i}$
- Inverse squared link: $g(\theta_i) = \frac{1}{\theta_i^2}$

**Logistic Regression (Definition 28.1)** Given a binary response; logistic regression assumes that

$$(\text{Response})_i \mid (\text{Predictors})_i, \beta \overset{\text{Ind}}{\sim} Ber(\theta_i)$$
$$\theta_i = \frac{e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}{1 + e^{\beta_0 + \sum_{j=1}^{p} \beta_j (\text{Predictor } j)_i}}.$$

**Marginal Density (Definition 9.2)** For a random vector $\mathbf{X}$, the marginal density of the first component $X_1$ (without loss of generality) is

$$f_{X_1}(u) = \int \cdots \int f_{\mathbf{X}}(\mathbf{x}) dx_2 \cdots dx_n.$$

**Markov Chain (Definition 18.2)** A sequence of random vectors $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ is a Markov Chain with stationary transition probabilities if for any set $A$ and any $k \leq n$

$$Pr\left(\theta^{(k)} \in A \mid \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(k-1)}\right) = Pr\left(\theta^{(k)} \in A \mid \theta^{(k-1)}\right)$$
$$= \int_A q\left(\theta^{(k)} \mid \theta^{(k-1)}\right) d\theta^{(k)}$$

where $q$ is called the transition kernel.

**Method of Distribution Functions (Definition 2.11)** Let $X$ be a continuous random variable with density $f$ and cumulative distribution function $F$. Consider $Y = h(X)$. The following process provides the density function $g$ of $Y$ by first finding its cumulative distribution function $G$.

1. Find the set $A$ for which $h(X) \leq t$ if and only if $X \in A$.
2. Recognize that $G(y) = Pr(Y \leq y) = Pr\left(h(X) \leq y\right) = Pr(X \in A)$.
3. If interested in $g(y)$, note that $g(y) = \frac{\partial}{\partial y} G(y)$.

**Metropolis Algorithm (Definition 18.1)** Suppose we want to generate random variates from the density $\pi(\theta \mid \mathbf{y})$. We perform the following steps:

1. Generate an initial value $\theta^{(0)}$.

2. At the $k$-th step, generate $\theta^*$ (a candidate) according to a symmetric proposal density $q\left(\theta \mid \theta^{(k-1)}\right)$.

3. Compute $A\left(\theta^*, \theta^{(k-1)}\right)$ where

$$A\left(\theta^*, \theta^{(k-1)}\right) = \frac{\pi\left(\theta^* \mid \mathbf{y}\right)}{\pi\left(\theta^{(k-1)} \mid \mathbf{y}\right)} = \frac{f\left(\mathbf{y} \mid \theta^*\right)\pi\left(\theta^*\right)}{f\left(\mathbf{y} \mid \theta^{(k-1)}\right)\pi\left(\theta^{(k-1)}\right)}.$$

4. Generate $U \sim Unif(0,1)$. If $U \leq A\left(\theta^*, \theta^{(k-1)}\right)$, then set $\theta^{(k)} = \theta^*$; else, set $\theta^{(k)} = \theta^{(k-1)}$.
5. Repeat Steps 2-4 $m$ times, for some large $m$.

When generating an initial value, $\theta^{(0)}$, we could choose $\theta^{(0)} \sim \pi(\theta)$ if the prior is easy to generate from. While it is common to choose $q(\cdot)$ to be a Normal distribution with mean $\theta^{(k-1)}$, it is not a requirement to do so; when a Normal distribution is used, it can be difficult to determine a reasonable variance (too large, and you drift too far away; too small, and you do not move at all).

**Mixture Distribution (Definition 15.3)** Let $X$ be a random variable and $f(x)$ and $g(x)$ be valid density functions defined on a common support. Then,

$$h(x) = wf(x) + (1-w)g(x),$$

where $0 < w < 1$, is known as a mixture distribution.

**Monte Carlo Error (Definition 17.2)** Also called the standard error for an approximation of the form $m^{-1}\sum_{k=1}^{m} g(X_k)$, the MC error is given by

$$\sqrt{\frac{1}{m(m-1)}\sum_{k=1}^{m}\left[g(X_k)-\frac{1}{m}\sum_{j=1}^{m}g(X_j)\right]^2}$$

which is the sample standard deviation of the generated variates divided by the square root of the number of replications.

**Monte Carlo Integration (Definition 17.1)** Consider an integral of the form

$$\int_{\mathcal{S}} g(x)f(x)dx$$

where $f(x)$ is a valid density function for a random variable $X$ with support $\mathcal{S}$. Then, the following algorithm, known as Monte Carlo (or MC) Integration, gives a numerical approximation to the integral:

1. Take a random sample $X_1, X_2, \ldots, X_m$ such that $X_i \sim f(x)$ for all $i$, where $m$ is large.
2. Compute $m^{-1}\sum_{i=1}^{m} g(X_i)$.

By the Law of Large Numbers,

$$\frac{1}{m}\sum_{i=1}^{m} g(X_i) \approx \int_{\mathcal{S}} g(x)f(x)dx.$$

**Noninformative Prior (Definition 16.5)** A prior distribution which is derived solely based on the form of the likelihood.

**Null Hypothesis (Definition 5.9)** The statement (or theory) about the parameter that we would like to *disprove*. This is denoted $H_0$, read "H-naught" or "H-zero".

**Null Value (Definition 5.11)** The value associated with the equality component of the null hypothesis; it forms the threshold or boundary between the hypotheses. Note: not all questions of interest require a null value be specified.

**Numeric Variable (Definition 3.6)** Also called a "quantitative variable," a measurement on a subject which takes on a numeric value *and* for which ordinary arithmetic makes sense.

**Observational Study (Definition 20.1)** A study in which each subject "self-selects" into one of groups being compared in the study. The phrase "self-selects" is used very loosely here and can include studies for which the groups are defined by an inherent characteristic or are chosen haphazardly.

**Outlier (Definition 7.7)** An individual observation which is so extreme, relative to the rest of the observations in the sample, that it does not appear to conform to the same distribution.

**Parameter (Definition 5.6)** Numeric quantity which summarizes the distribution of a variable within the *population* of interest. Generally denoted by Greek letters in statistical formulas.

**Percentile (Definition 7.1)** The $k$-th percentile is 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" and denoted $Q_1$).
- 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" and denoted $Q_3$).

**Percentile for a Random Variable (Definition 2.8)** Let $X$ be a random variable with density function $f$. The $100k$ percentile is the value $q$ such that

$$Pr(X \leq q) = k.$$

**Point Estimation (Definition 13.1)** Point estimation is the process of estimating a parameter with a single statistic. This is like trying to hit an infinitesimally small target with a dart.

**Population (Definition 3.1)** The collection of subjects we would like to say something about.

**Posterior Distribution (Definition 11.1)** A distribution quantifying our beliefs about the uncertainty in the parameter(s) of the underlying sampling distribution *after* observing data. This is often denoted by $\pi(\theta \mid \mathbf{y})$ where $\theta$ is the parameter vector and $\mathbf{y}$ the observe data.

Given the likelihood $f(\mathbf{y} \mid \theta)$ and a prior distribution on the parameters $\pi(\theta)$, the posterior distribution is computed using Bayes Theorem:

$$\pi(\theta \mid \mathbf{y}) = \frac{f(\mathbf{y} \mid \theta)\pi(\theta)}{\int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta}.$$

**Posterior Mean (Definition 12.1)** The posterior mean is the average value of the parameter, given the data:

$$E\left[\theta \mid \mathbf{y}\right] = \int \theta\pi(\theta \mid \mathbf{y})d\theta.$$

**Posterior Median (Definition 12.2)** We are 50% sure, given the data, the parameter falls below the posterior median. Formally, the posterior median is the value $q$ such that

$$0.5 = \int_{-\infty}^{q} \pi(\theta \mid \mathbf{y}) d\theta.$$

**Posterior Mode (Definition 12.3)** We think of the posterior mode as the most likely value of the parameter, given the data. If the posterior distribution is continuous, the posterior mode is the value of the parameter that maximizes the posterior distribution. Formally, the posterior mode is given by

$$\arg\max_{\theta} \pi(\theta \mid \mathbf{y}).$$

**Posterior Odds (Definition 15.1)** Let $H_j$ denote the hypothesis that $\theta \in \Theta_j$ for some region $\Theta_j$. Then, the posterior odds *in favor of* $H_j$ is given by

$$\frac{Pr\left(\theta \in \Theta_j \mid \mathbf{y}\right)}{Pr\left(\theta \notin \Theta_j \mid \mathbf{y}\right)}.$$

**Posterior Predictive Distribution (Definition 14.2)** Let $\mathbf{Y}^*$ represent a collection of $m$ *future* observations. The distribution of these future observations given the observed data $\mathbf{Y}$ (of length $n$), called the posterior predictive distribution, is given by

$$\pi\left(\mathbf{y}^* \mid \mathbf{y}\right) = \int f\left(\mathbf{y}^* \mid \theta\right) \pi(\theta \mid \mathbf{y}) d\theta.$$

**Potential (Definition 18.4)** The potential of a value $\theta$ is the negative logarithm of the posterior evaluated at $\theta$. In practice, we need only know the potential up to a constant. That is, it suffices to define the potential as

$$\text{Potential}(\theta) = -\log\left[f(\mathbf{y} \mid \theta)\pi(\theta)\right].$$

**Prior Distribution (Definition 10.1)** A distribution quantifying our beliefs about uncertainty in the *parameter(s)* of the underlying sampling distribution *prior to* observing any data. This is often denoted by $\pi(\theta)$ where $\theta$ is the parameter vector.

- This relies on a *subjective* view of probability.
- As prior beliefs are subjective, there is no "one" prior, but each individual may have a unique prior.

**Prior Predictive Distribution (Definition 14.1)** The prior predictive distribution is the marginal distribution of the response(s) prior to observing any data:

$$m(\mathbf{y}) = \int f(\mathbf{y} \mid \theta)\pi(\theta)d\theta.$$

The distribution marginalizes the parameter out of the likelihood using the beliefs from the prior distribution.

**Random Sample (Definition 9.6)** A random sample of size $n$ refers to a collection of $n$ random variables $X_1, X_2, \ldots, X_n$ such that the random variables are mutually independent, and the distribution of each random variable is identical.

**Random Variable (Definition 2.1)** Let $\mathcal{S}$ be the sample space corresponding to a random process; a random variable $X$ is a function mapping elements of the sample space to the real line.

Random variables represent a measurement that will be collected during the course of a study. Random variables are typically represented by a capital letter.

**Random Vector (Definition 8.2)** Let $X_1, X_2, \ldots, X_n$ be $n$ random variables. Then, the vector $\mathbf{X} = (X_1, X_2, \ldots, X_n)^\top$ is a random vector of length $n$.

**Randomization (Definition 20.5)** Randomization can refer to random *selection* or random *allocation*. Random selection refers to the use of a random mechanism (e.g., a simple random sample, Definition 6.2, or a stratified random sample, Definition 6.3) to select units from the population. Random selection minimizes bias.

Random allocation refers to the use of a random mechanism when assigning units to a specific treatment group in a controlled experiment (Definition 20.2). Random allocation eliminates confounding and permits causal interpretations.

**Reduction of Noise (Definition 20.6)** Reducing extraneous sources of variability can be accomplished by fixing extraneous variables or blocking (Definition 20.7). These actions reduce the number of differences between the units under study.

**Regression (Definition 23.1)** A regression model is one for which the parameter(s) governing the data generating process depends on one or more predictors. "Parametric" regression models do this through specifying a functional form for the dependence of the parameter(s) on the predictor(s).

**Relative Frequency (Definition 5.5)** Also called the "proportion," the fraction of observations falling into a particular group (level) of a categorical variable.

**Replication (Definition 20.4)** Replication results from taking measurements on different units (or subjects), for which you expect the results to be similar. That is, any variability across the units is due to natural variability within the population.

**Residual (Definition 27.1)** A residual is the difference between an observed response and the predicted mean response for that same individual:

$$(\text{Residual})_i = (\text{Response})_i - (\text{Fitted Value})_i,$$

where

$$(\text{Fitted Value})_i = \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j (\text{Predictor}j)_i.$$

**Response (Definition 5.2)** The primary variable of interest within a study. This is the variable you would either like to explain or estimate.

**Sample (Definition 3.2)** The collection of subjects for which we actually obtain measurements (data).

**Sample Space (Definition 1.1)** The sample space for a random process is the collection of all possible results that we might observe.

**Simple Random Sample (Definition 6.2)** 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.

**Standard Deviation (Definition 7.4)** A measure of spread, this is the square root of the variance.

**Stationary Distribution (Definition 18.3)** Let $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(n)}$ be a Markov Chain. The stationary distribution of the Markov Chain is the distribution $p(\theta)$ such that

$$Pr\left(\theta^{(k)} \in A\right) = \int_A p(\theta) d\theta.$$

**Statistic (Definition 7.8)** Numeric quantity which summarizes the distribution of a variable within a *sample*.

**Statistical Inference (Definition 3.3)** The process of using a sample to characterize some aspect of the underlying population.

**Stratified Random Sample (Definition 6.3)** 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.

**Subjective Interpretation of Probability (Definition 1.4)** In this perspective, the probability of $A$ describes the individual's uncertainty about event $A$.

**Support (Definition 2.2)** The support of a random variable is the set of all possible values the random variable can take.

**Variability (Definition 5.1)** The notion that measurements differ from one observation to another.

**Variable (Definition 3.4)** A measurement, or category, describing some aspect of the subject.

**Variance (Definition 7.3)** Let $X$ be a random variable with density function $f$ defined over the support $\mathcal{S}$. The variance of a random variable, denoted $Var(X)$, is given by

$$Var(X) = E\left[X - E(X)\right]^2 = E\left(X^2\right) - E^2(X).$$

If we let $\mu = E(X)$, then this is equivalent to

$$\int_{\mathcal{S}} (x-\mu)^2 f(x) dx$$

for continuous random variables and

$$\sum_{\mathcal{S}} (x-\mu)^2 f(x)$$

for discrete random variables.

**Variance (Definition 7.3)** 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

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

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 $n$ removes bias in the statistic.

The symbol $\sigma^2$ is often used to denote the variance in the population.