Introduction to Statistical Theory

Eric M Reyes

Updated: 26 August 2024

Table of contents

Preface

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. The key step in any statistical analysis is characterizing variability; this is what allows a statistician to make decisions using data. And, Probability can be used to develop analytical models to describe that variability. While we do not need proficiency with probability to be practitioners of statistical methodology, a foundation in probability models is necessary for developing statistical theory and helps us see common threads in statistical modeling.

This is not a Probability text; instead, this text acts as a supplement to an Introductory Statistics course. 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.

Our goal is that this supplement provides a richer experience for those students entering Statistics with exposure to Probability by leveraging the connections between the two disciplines.

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.

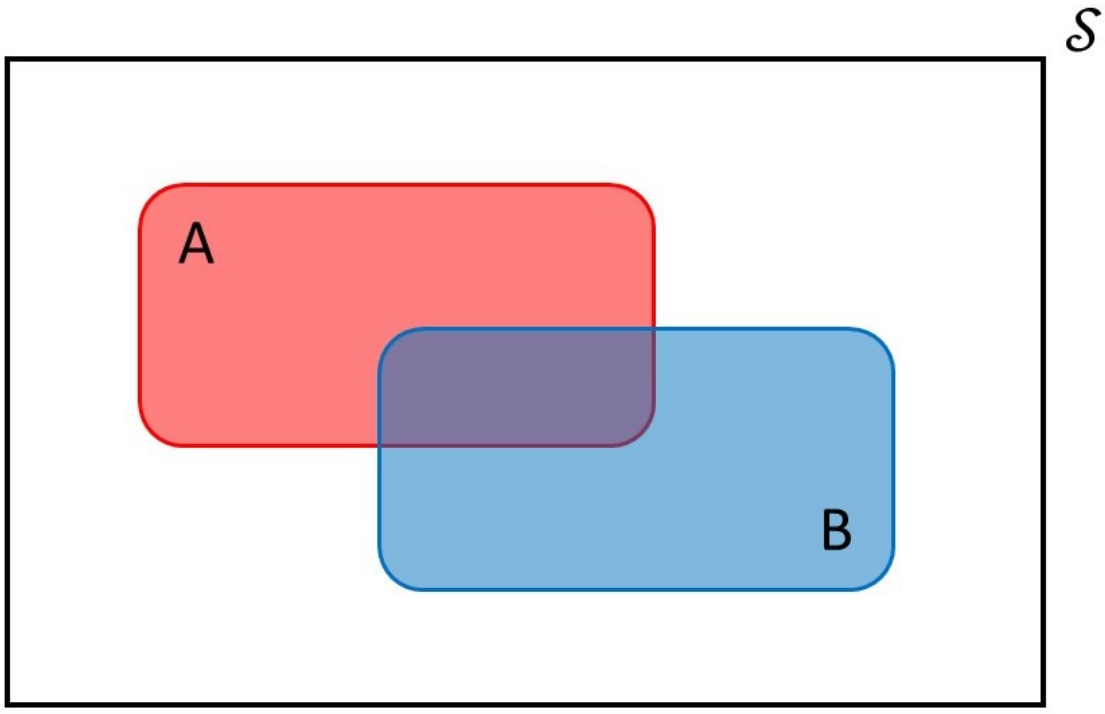

Figure 1.1: Venn-Diagram illustrating two events, A and B, within a sample space $\mathcal{S}$.

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. □

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.

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

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. **This is incorrect from the lens of this course.** The complication here is that while we are ignorant of the result (whether the child is correct or not), in reality, the result has already been determined.

Probability is the study of random processes; in particular, it seeks to quantify the likelihood that an event *will* occur. Note the use of the future tense. Once an event has occurred, it does not make sense to describe the likelihood of the result. Returning to Example 1.1, when a person says that the probability the child is right is 0.5, what they are really quantifying is not the likelihood of the child being correct but instead their uncertainty about that event. This is 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. While this interpretation leads to a very interesting development of statistical theory (we encourage everyone to take a course in Bayesian Data Analysis), this is not the predominant interpretation. 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 does not quantify our uncertainty about an event; it quantifies the likelihood of an event in repeated observation.

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.

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.

With few exceptions, we will be working with continuous random variables. As a result, 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 σ 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 σ. 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 2.5 (Parameter)**.** Numeric quantity which summarizes the distribution of a variable within the *population* of interest. Generally denoted by Greek letters in statistical formulas.

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

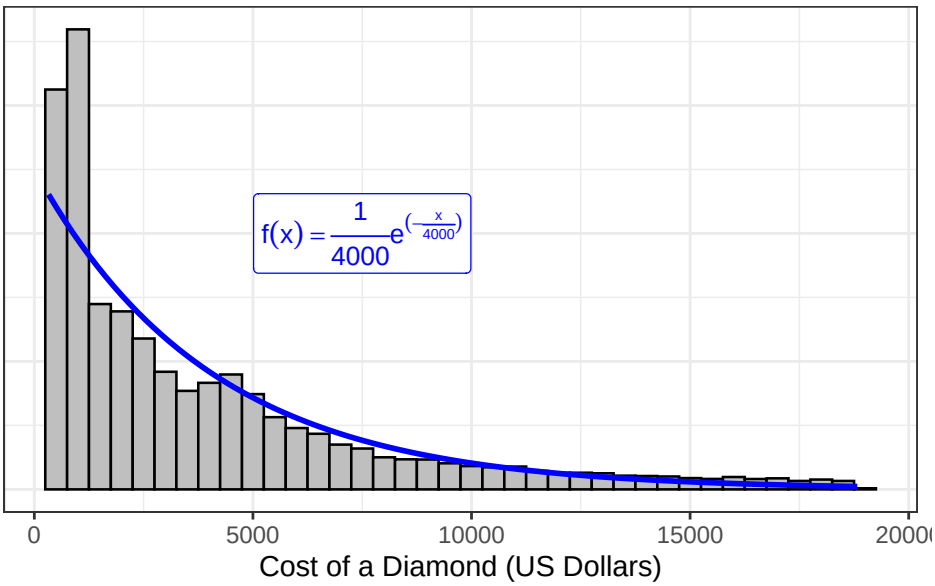

Figure 2.1: Illustration of a density function representing the posited distribution of the population alongside a histogram summarizing the cost of diamonds using a sample of 53940 diamonds.

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.

> 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 2.6 (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 2.7 (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 2.8 (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.

A result of Definition 2.8 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 2.9 (Percentile)**.** 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 σ; 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 σ. 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 θ; then, the mean, variance, and median (as well as any other parameters of interest in a research objective) will be functions of θ.

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. The solution also shows that there is more than one way to characterize a distribution.

2.2.2 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 Common Probability Distributions

While we could posit any non-negative function as a model for a density function (properly scaled of course), there are some models that are very common. While the following list is not exhaustive, it does include the most commonly used distributions that we will encounter in the text.

When a response is binary (assumes one of two values), it is a Bernoulli distribution. In order to make use of this distribution, we typically define one of the two possible outcomes as a "success" and the other as a "failure." For example,

$$X = \begin{cases} 1 & \text{if a success is observed} \\ 0 & \text{if a success is not observed.} \end{cases}$$

Definition 2.11 (Bernoulli Distribution)**.** Let X be a discrete random variable taking the value 0 or 1. X is said to have a Bernoulli distribution with density

$$f(x) = \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1\},$$

where $0 < \theta < 1$ is the probability that X takes the value 1.

- $E(X) = \theta$
- $Var(X) = \theta(1-\theta)$

We write $X \sim Ber(\theta)$, which is read "X follows a Bernoulli distribution with probability θ."

We can generalize the Bernoulli distribution to count the number of successes out of n independent trials.

Definition 2.12 (Binomial Distribution)**.** Let X be a discrete random variable taking integer values between 0 and n, inclusive. X is said to have a Binomial distribution with density

$$f(x) = \binom{n}{x} \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1, \ldots, n\},$$

where $0 < \theta < 1$ is the probability of a success on an individual trial.

- $E(X) = n\theta$
- $Var(X) = n\theta(1-\theta)$

We write $X \sim Bin(n, \theta)$, which is read "X follows a Binomial distribution with parameters n and θ."

While there are many other discrete distributions that play important roles in categorical data analyses, the majority of our text will focus on quantitative response variables. So, we list several key distributions for continuous random variables.

Definition 2.13 (Normal (Gaussian) Distribution)**.** Let X be a continuous random variable. X is said to have a Normal (or Guassian) distribution if the density is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty,$$

where μ is any real number and $\sigma^2 > 0$.

- $E(X) = \mu$
- $Var(X) = \sigma^2$

We write $X \sim N(\mu, \sigma^2)$, which is read "X follows a Normal distribution with mean μ and variance σ^2." This short-hand implies the density above. When $\mu = 0$ and $\sigma^2 = 1$, this is referred to as the Standard Normal distribution.

This model is a bell-shaped distribution centered at the mean μ. While this is a common model, it should not be assumed by default. In future chapters, we will consider methods for assessing whether, given a sample, assuming the population follows a Normal distribution is reasonable.

Definition 2.14 (Gamma Distribution)**.** Let X be a continuous random variable. X is said to have a Gamma distribution if the density is given by

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \qquad x > 0,$$

where $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter.

- $E(X) = \alpha\beta$
- $Var(X) = \alpha\beta^2$

We write $X \sim Gamma\,(\alpha, \beta)$, which is read "X follows a Gamma distribution with shape α and scale β." This short-hand implies the density above. When $\alpha = 1$, we refer to this as the Exponential distribution with scale β.

We note that, in general, there is no closed form solution for $\Gamma(\alpha)$, but

- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
- $\Gamma(k) = (k - 1)!$ for non-negative integer k

The Gamma distribution is useful for modeling time-to events.

> **Warning**
>
> We have presented the Gamma (and Exponential) distribution in terms of the *scale* parameter. It is sometimes easier to parameters the distribution in terms of the *rate* parameter, where the rate is the inverse of the scale. When consulting tables of distributions[1], be sure to note the parameterization of the distribution provided.

> **Note**
>
> The Exponential distribution being a special case of the Gamma distribution is not the only relationship between common distributions. There are many relationships[2] that are useful; we will describe these as needed.

The (standardized) t-distribution is a bell-shaped distribution, similar to the Normal distribution but with wider tails. It has a single parameter, known as the degrees of freedom. Note that unlike many other distributions, this parameter (the degrees of freedom) is not associated with the location of the distribution. Instead, it governs the spread (but is not equivalent to the variance).

Definition 2.15 (t-Distribution)**.** Let X be a continuous random variable. X is said to have a (standardized) t-distribution, sometimes called the Student's t-distribution, if the density is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad -\infty < x < \infty$$

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

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

where $\nu > 0$ is the degrees of freedom.

We write $X \sim t_\nu$, which is read "X follows a t-distribution with ν degrees of freedom."

> **Note**
>
> As the degrees of freedom approach infinity, the density function of the t-distribution approaches that of a Standard Normal random variable.

The Chi-Square distribution is a skewed distribution (looks like a giant slide). It has a single parameter, known as the degrees of freedom. The degrees of freedom for this distribution characterize both the location and spread simultaneously.

Definition 2.16 (Chi-Square Distribution)**.** Let X be a continuous random variable. X is said to have a Chi-Square distribution if the density is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1}e^{-x/2} \qquad x > 0,$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim \chi^2_\nu$, which is read "X follows a Chi-Square distribution with ν degrees of freedom." The Chi-Square distribution is a special case of the Gamma distribution where $\alpha = \nu/2$ and $\beta = 2$.

The F-distribution is a skewed distribution. It has two parameters, known as the numerator and denominator degrees of freedom. While neither variable is directly the mean or variance, together these two parameters characterize both the location and the spread.

Definition 2.17 (F-Distribution)**.** Let X be a continuous random variable. X is said to have an F-distribution if the density is given by

$$f(x) = \frac{\Gamma((r+s)/2)}{(\Gamma(r/2)\Gamma(s/2))}(r/s)^{(r/2)}x^{(r/2-1)}(1+(r/s)x)^{-(r+s)/2} \qquad x > 0,$$

where $r, s > 0$ are the numerator and denominator degrees of freedom, respectively.

We write $X \sim F_{r,s}$, which is read "X has an F-distribution with r numerator degrees of freedom and s denominator degrees of freedom."

The formulas above are ugly, but we will not be working with them directly. Instead, statistical software have these distributions embedded. The key idea here is that when we know the model for a distribution, we can make use of several results about this distribution.

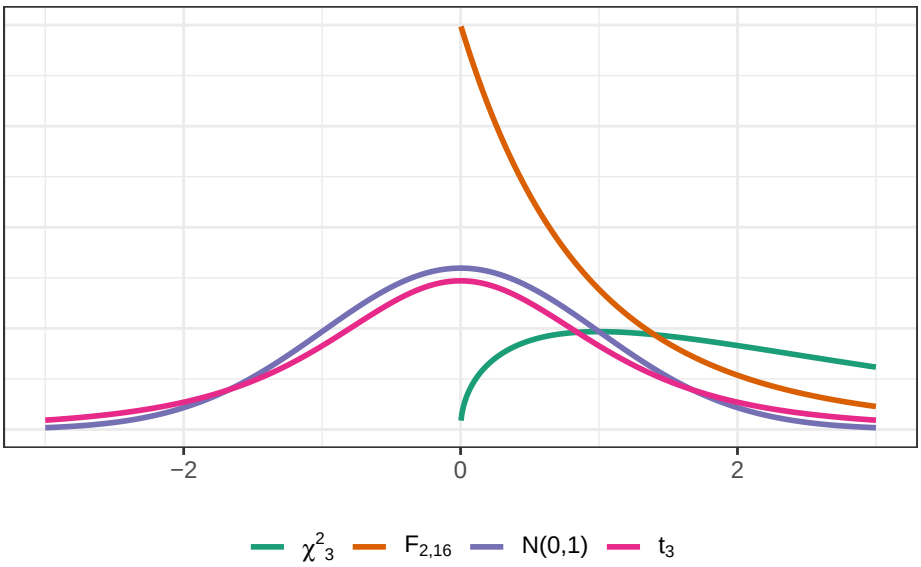

Figure 2.2: Comparison of various common distributions for continuous random variables.

Big Idea

Some probability models occur so frequently that we give them names for easy reference. Some models are common for modeling the population, in which case they are defined in terms of unknown parameters to be estimated. Some models are used, not to model a population, but to model other distributions that occur in statistical practice.

2.4 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 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 2.18 (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 2.2 (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 $Y = \log(X)$.

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

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

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.

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

3 Sampling Distribution of a Statistic

In the previous chapter, we related probability concepts associated with random variables to their statistical counterparts in data collection. We introduced the idea of using a density function as a model of the distribution of a variable within the population. This led to the observation that the population parameters govern the shape of these density functions. Further, these parameters are the focal point of research objectives. In a statistical analysis, the parameters would be estimated using statistics; in this chapter, we explore how probability theory helps us to characterize the variability in these statistics — which is the key component of statistical inference.

3.1 Statistics and Expectations

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.

No researcher would believe that measuring the cost of a single diamond would be sufficient to address the above research objective. Instead, we would consider taking a sample of n diamonds and measuring the cost of each. We can represent the cost we will record (note the use of the future tense) as $X_1, X_2, \ldots, X_n$, where X_i is the cost of the i-th diamond we will measure. 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).

In almost every analysis, we compute a numerical summary of the data. For example, the sample mean takes the form

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

> **Note**
>
> While a capital letter denotes a random variable (not yet observed), the corresponding lowercase letter is used to denote a past observation (which is no longer random).

These numerical summaries, statistics, are functions of the data. So, prior to when we collect that data, our statistics are functions of random variables.

Definition 3.1 (Statistic)**.** A statistic is a numerical summary of a sample; it is a function of the data alone. Prior to collecting data, a statistic is a function of the data to be collected.

Definition 3.1 eliminates the possibility of the statistic being a function of the underlying *parameters*; certainly, the behavior of the statistic is determined by the parameters, but the computation of a statistic should not require knowledge of the parameter once the data is collected.

Definition 3.1 highlights the need to study functions and combinations of random variables. If you recall, Theorem 2.1 introduced the idea of expectation as a linear operator. The result, for a single random variable, is nearly intuitive. If expectation is associated with integration (as defined in Definition 2.6), then expectations should adopt the properties of integration, including linearity. The generalization of this result is extremely important within statistics.

Theorem 3.1 (Linearity of Expectations)**.** *For random variables* $X_1, X_2, \ldots, X_n$*, constants* $a_1, a_2, \ldots, a_n$ *and real-valued functions* $g_1, g_2, \ldots, g_n$*, we have that*

$$E\left[\sum_{i=1}^{n} a_i g\left(X_i\right)\right] = \sum_{i=1}^{n} a_i E\left[g\left(X_i\right)\right].$$

The importance of Theorem 3.1 is seen in determining the expectation of the sample mean.

Example 3.1 (Mean of the Sample Mean)**.** Let $X_1, X_2, \ldots, X_n$ be random variables representing observations from a sample that will be collected. Define the sample mean of that future sample to be

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Determine $E\left(\bar{X}\right)$.

Before addressing the prompt in Example 3.1, we note that *before collecting the data*, statistics, like the sample mean, is a function of random variables and is therefore itself a random variable! And, all random variables have distributions, and they have parameters that govern the shape of that distribution. Example 3.1 is focused on the mean of the distribution.

Solution. Applying Theorem 3.1, we have

$$\begin{aligned} E\left(\bar{X}\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E\left(X_i\right) \end{aligned}$$

If we assume that each X_i is taken from the same population, then $E\left(X_i\right) = \mu$, the population mean, for some constant μ. Therefore, we have that

$$E\left(\bar{X}\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu.$$

Theorem 3.1 discusses how expectations work with linear combinations; we have a similar result for variances. However, there is a particular caveat.

Theorem 3.2 (Variance of the Sum of Independent Random Variables)**.** *Let $X_1, X_2, \ldots, X_n$ be independent random variables, and define constants $a_1, a_2, \ldots, a_n$ and real-valued functions $g_1, g_2, \ldots, g_n$. Then,*

$$Var\left[\sum_{i=1}^{n} a_i g\left(X_i\right)\right] = \sum_{i=1}^{n} a_i^2 Var\left[g\left(X_i\right)\right].$$

Comparing Theorem 3.1 to Theorem 3.2, we see that while the expectation always move through sums, the variance only does so when the random variables are independent.

Definition 3.2 (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.

> **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.

Consider taking a sample of n diamonds to address our above research objective. It seems reasonable that the cost of one diamond does not depend on the cost of another. That allows us to assume the values we collect will be independent; that is, the random variables we use to represent these costs are independent of one another. It also seems natural to assume that each diamond we select comes from the same underlying population. These two attributes together form the basis of what we mean by a "random sample."

Definition 3.3 (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.

We say $X_1, X_2, \ldots, X_n$ are independent and identically distributed, abbreviated IID. We might also write this as $X_i \overset{\text{IID}}{\sim} f$ for some density f.

> **Warning**
>
> Let X and Y be identically distributed random variables. This does not mean that $X = Y$. That is, the two random variables need not take on the same value. Instead, identically distributed means the density function of the two random variables are the same. As a result, they share the same mean, variance, etc. That is, their distributions are the same, not their value.

Example 3.2 (Variance of the Sample Mean)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample of size n from a population with a variance of σ^2. Define the sample mean of that future sample to be

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Determine $Var\left(\bar{X}\right)$.

Solution. Since $X_1, X_2, \ldots, X_n$ form a random sample; we know that they are mutually independent. Therefore, Theorem 3.2 gives

$$
\begin{aligned}
Var\left(\bar{X}\right) &= Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 \\
&= \sigma^2/n.
\end{aligned}
$$

> **Warning**
>
> It is important to remember that Example 3.1 and Example 3.2 describe the mean and variance of the sample mean *prior to* collecting data. Once the data is collected, the sample mean has no distribution. Once the data is collected, a statistic is simply a number.

There is one additional result for independent random variables that we should keep in mind.

Definition 3.4 (Expectation of a Product of Independent Random Variables)**.** Let $X_1, X_2, \ldots, X_n$ be independent random variables, then

$$
E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E\left(X_i\right).
$$

3.2 Sampling Distribution of Sample Mean

Together, Example 3.1 and Example 3.2 characterize the center and spread of the distribution of the sample mean. Since each statistic is a random variable, it has a distribution, and we call that distribution a **sampling distribution**.

Definition 3.5 (Sampling Distribution)**.** The distribution of a statistic across repeated samples.

Note that Definition 3.5 makes use of the idea of repeated sampling. Once again, this leans on the frequentist interpretation of probability (Definition 1.5). We are thinking of the distribution of a statistic as the result of the different values that could potentially be observed if we were to repeatedly take samples of the same size.

While Theorem 3.1 and Theorem 3.2 allowed us to characterize the sampling distribution of the sample mean, these results did not provide information about the shape of the sampling

distribution, much less provide a functional form of the density. To make progress on this front, we return to another tool introduced in a probability course — the **moment-generating function**.

Definition 3.6 (Moment-Generating Function (MGF))**.** For a random variable X, let $M_X(t)$ be defined as

$$M_X(t) = E\left(e^{tX}\right).$$

If $M_X(t)$ is defined for all values of t in some interval about 0, then $M_X(t)$ is called the moment-generating function (MGF) of X.

> **Note**
>
> When we are working with multiple random variables, it is common to use a subscript to denote the random variable we are referencing. For example, F_X may represent the CDF of the random variable X, f_Y denote the density function of the random variable Y, and $M_Z(t)$ denote the MGF of the random variable Z.

First, we note that the MGF is a function, but not a function of the random variable. That is, $M_X(t)$ is not a random variable since it is the result of taking the expected value of a random variable. Second, the definition does not guarantee the existence of the MGF for any particular random variable. That is, there are distributions for which the MGF is not defined. The power of the MGF is summarized in **?@thm-mgf-properties**.

3.3 Properties of the Moment-Generating Function

Let X be a random variable with moment-generating function $M_X(t)$. Then, we have that

1. $E\left(X^k\right) = M_X^{(k)}(0)$ for all integers k, where $M_X(k)(0)$ is the k-th derivative of $M_X(t)$ evaluated at $t = 0$.
2. The MGF uniquely defines the random variable. That is, if two random variables have the same MGF, then those random variables have the same density function.

Consider again our sample of n diamonds; let's assume it is a random sample from the population. Suppose we are interested in the total number of diamonds within this sample that have a "princess" cut. Define the random variable

$$Y_i = \begin{cases} 1 & \text{if i-th diamond has a princess cut} \\ 0 & \text{otherwise.} \end{cases}$$

If we are intersted in the total number of diamonds within the sample that have a princess cut, we are interested in the statistic $\sum_{i=1}^{n} Y_i$. We can use **?@thm-mgf-properties**, together with our previous results about expectations to derive the sampling distribution of this statistic.

Example 3.3 (Sampling Distribution of a Sum of Bernoulli Random Variables)**.** Let $Y_1, Y_2, \ldots, Y_n$ be IID random variables from a Bernoulli distribution with probability θ. Define

$$Z = \sum_{i=1}^{n} Y_i,$$

the total number of "successes" in the random sample. Determine the sampling distribution of Z.

Solution. In order to determine the distribution of Z, we find its MGF.

$$\begin{aligned} M_Z(t) &= E\left(e^{tZ}\right) \\ &= E\left(e^{t\sum_{i=1}^{n} Y_i}\right) \\ &= E\left(\prod_{i=1}^{n} e^{tY_i}\right), \end{aligned}$$

where the third line results from recognizing that the product of exponential terms is the exponential of the sum of the powers. Now, applying **?@thm-product-expectation**, we have

$$\begin{aligned} M_Z(t) &= \prod_{i=1}^{n} E\left(e^{tY_i}\right) \\ &= \prod_{i=1}^{n} M_{Y_i}(t) \\ &= \prod_{i=1}^{n} M_{Y_1}(t) \end{aligned}$$

where the last line is the result of the random variables being identically distributed. Since they have the same distribution, they must have the same MGF's; therefore, $M_{Y_i}(t) = M_{Y_1}(t)$ for each i. Again, we are *not* saying $Y_i = Y_1$; we are saying the moment-generating functions of these random variables is equivalent.

Consulting a table to determine the MGF of a Bernoulli random variable, we have that

$$M_{Y_1}(t) = (1 - \theta) + \theta e^t.$$

Thus, we have that

$$\begin{aligned} M_Z(t) &= \prod_{i=1}^{n} M_{Y_1}(t) \\ &= \left[M_{Y_1}(t)\right]^n \\ &= \left[(1-\theta) + \theta e^t\right]^n. \end{aligned}$$

But, consulting a table of common distributions, we recognize $M_Z(t)$ as the moment-generating function of a Binomial distribution with parameters n and θ. Since moment-generating functions uniquely define a distribution when they exist, we have that $Z \sim Bin(n, \theta)$.

The next chapter will illustrate how we can capitalize on this information. For this chapter, we simply note that we are able to characterize how the sum of Bernoulli random variables behaves. We note that this sampling distribution depends on the unknown parameter θ that also governs the underlying population. In addition, it depends on the sample size.

> **Big Idea**
>
> The sampling distribution of a statistic depends on both the parameters from the underlying population as well as the sample size.

While the previous example illustrates the process, we admit that it simply reiterates a fact that we already knew (and noted in Definition 2.12). The utility of the next result may be a bit more apparent.

Example 3.4 (Sampling Distribution of the Sample Mean from a Normal Population)**.** Let $Y_1, Y_2, \ldots, Y_n \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$. Define

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

the sample mean. Determine the sampling distribution of $\bar{Y}$.

Solution. In order to determine the distribution of $\bar{Y}$, we find its MGF.

$$\begin{aligned} M_{\bar{Y}}(t) &= E\left(e^{t\bar{Y}}\right) \\ &= E\left(e^{tn^{-1}\sum_{i=1}^{n} Y_i}\right) \\ &= E\left(\prod_{i=1}^{n} e^{tn^{-1}Y_i}\right) \end{aligned}$$

where the third line results from recognizing that the product of exponential terms is the exponential of the sum of the powers. Now, applying **?@thm-product-expectation**, we have

$$\begin{aligned} M_{\bar{Y}}(t) &= \prod_{i=1}^{n} E\left(e^{tn^{-1}Y_i}\right) \\ &= \prod_{i=1}^{n} M_{Y_i}(t/n) \\ &= \prod_{i=1}^{n} M_{Y_1}(t/n) \end{aligned}$$

where the last line is the result of the random variables being identically distributed, and the second line makes use of the definition of the MGF, which can be evaluated at any value, including t/n. Since they have the same distribution, they must have the same MGF's; therefore, $M_{Y_i}(t) = M_{Y_1}(t)$ for each i and all t. Again, we are *not* saying $Y_i = Y_1$; we are saying the moment-generating functions of these random variables is equivalent.

Consulting a table to determine the MGF of a Normal random variable, we have that

$$M_{Y_1}(t) = e^{\mu t + (1/2)\sigma^2 t^2}.$$

Thus, we have that

$$\begin{aligned} M_{\bar{Y}}(t) &= \prod_{i=1}^{n} M_{Y_1}(t/n) \\ &= \left[M_{Y_1}(t/n)\right]^n \\ &= \left[e^{\mu t/n + (1/2)\sigma^2 t^2/n^2}\right]^n \\ &= e^{\mu t + (1/2)(\sigma^2/n)t^2}. \end{aligned}$$

But, consulting a table of common distributions, we recognize $M_{\bar{Y}}(t)$ as the moment-generating function of a Normal distribution with a mean of μ and a variance of σ^2/n. Since moment-generating functions uniquely define a distribution when they exist, we have that $\bar{Y} \sim N\left(\mu, \sigma^2/n\right)$.

We note a difference between Example 3.3 and Example 3.4. In Example 3.3, the statistic we examined did not estimate a parameter of interest. While there is nothing wrong with examining the total number of diamonds in a sample, the statistic itself is dependent on the sample size — we would expect a larger value with larger samples. In Example 3.4, however, the sample mean is a common estimate of the population mean. It turns out the sampling

distributions of most estimators (statistics chosen to estimate a parameter) tend to share similar characteristics.

Common Characteristics of Sampling Distributions

While not guaranteed, the sampling distribution of many statistics tend to have the following characteristics:

1. The sampling distribution is centered on the corresponding parameter of interest, or the center approaches the corresponding parameter as the sample size increases.
2. The spread of the sampling distribution is smaller than within the population, and the spread decreases as the sample size increases.
3. The shape of the sampling distribution differs from that of the underlying population, and the sampling distribution becomes more bell-shaped as the sample size increases.

Of the three characteristics above, the third is the one most likely to be broken.

Considering Example 3.4, we see that all three characteristics hold. First, we see (as we also saw in Example 3.1) that the expected value of the sample mean is the population mean. When this occurs, we say the estimator is **unbiased**.

Definition 3.7 (Unbiased)**.** An estimator (statistic) $\hat{\theta}$ is said to be unbiased for the parameter θ if

$$E\left(\hat{\theta}\right) = \theta.$$

While being unbiased is a good quality in an estimator, it is not required. For example, the sample standard deviation is not an unbiased estimator of the population standard deviation, yet it is still a preferred estimator.

In Example 3.4, we see that variance of the sample mean is smaller than the variance of the population by a factor of n; therefore, as the sample size increases, the variability of the sample mean decreases. This implies that the sample mean of a large sample will tend not to stray as far from the population mean as that of a smaller sample. This is where the notion of "more data is better" comes from.

Big Idea

Larger samples result in more reliable statistics.

Finally, we see in Example 3.4 that the sampling distribution of the sample mean is a Normal distribution, which is bell-shaped. Again, being bell-shaped is not necessarily more advantageous than any other distribution, but it reinforces the idea that the statistic tends to be near the parameter of interest across repeated sampling.

> **Warning**
>
> Remember, these discussions are about the distribution of a statistic across repeated samples, and so they apply prior to collecting data. Once we have a sample, the statistic does not have a distribution.

Example 3.4 is a really nice result because it tells us the behavior of a popular statistic; unfortunately, it only applies to a sample from a population which follows a Normal distribution. More, it only applies when the population variance is known, which rarely happens in practice. Theorem 3.3 generalizes the results to the case when the population variance is unknown.

Theorem 3.3 (Student's Theorem)**.** *Let* $Y_1, Y_2, \ldots, Y_n \overset{IID}{\sim} N(\mu, \sigma^2)$*. Define*

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

to be the sample mean and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$$

to be the sample variance. Then,

$$\frac{\sqrt{n}\left(\bar{Y} - \mu\right)}{S} \sim t_{n-1}.$$

To see the impact of estimating the population variance, we recognize that Example 3.4 gave us that when the population variance is known

$$\bar{Y} \sim N\left(\mu, \sigma^2/n\right),$$

which can be rewritten as

$$\frac{\sqrt{n}\left(\bar{Y} - \mu\right)}{\sigma} \sim N(0, 1).$$

Therefore, when the population variance is estimated (using the typical sample variance), we have that the sampling distribution of this standardized ratio follows a t-distribution instead of a Standard Normal distribution.

Theorem 3.3 highlights that we often characterize the distribution of some "standardized statistic" (instead of the statistic that estimates the parameter directly). Exact results like Example 3.4 and Theorem 3.3 are quite rare. It is more common for us to rely on approximations to the sampling distribution, the most famous of which is the Central Limit Theorem.

Theorem 3.4 (Central Limit Theorem (CLT)). *Let $X_1, X_2, \ldots, X_n$ be IID random variables such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all i. As n approaches infinity, the ratio*

$$\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{S}$$

behaves like a Standard Normal random variable. That is,

$$Pr\left(\frac{\sqrt{n}\left(\bar{X} - \mu\right)}{S} \leq q\right) \to Pr(Z \leq q) \qquad as\ n \to \infty$$

where q is any real number, $Z \sim N(0,1)$, $\bar{X}$ is the sample mean and S^2 is the sample standard deviation.

> **Note**
>
> "The" Central Limit Theorem is a misnomer; there are actually several Central Limit Theorems which differ in their assumptions. The one most commonly presented in texts uses the population standard deviation σ instead of the sample standard deviation S as we have presented it. The more common presentation is easier to prove, but it is far less useful in practice (as the population variance is rarely known). The proof of the version we have presented is beyond the scope of the course but is more useful in practice.

Known as a "limit" (or "asymptotic") result, Theorem 3.4 provides an approximation to the sampling distribution. That is, the CLT states that as the sample size gets large, the Standard Normal distribution is a good approximation to the true sampling distribution of this standardized statistic. Of course, that begs the question, "how good is the approximation?" as well as "how large of a sample is large enough?" While we can never address these questions with certainty, there are some graphical techniques for assessing these questions in practice.

The huge draw of the CLT is that it applies under vary weak conditions — the underlying population has a finite mean and variance. For nearly any population, we have an approximation for the sampling distribution of the sample mean.

> **Note**
>
> The above methods describe the actual sampling distribution. Of course, because these describe how a statistic behaves across repeated samples, this is not something we get to observe directly. Instead, the sampling distribution must be modeled. This is often done by replacing the parameters in the sampling distribution with the corresponding estimates from the sample. Therefore, the *model* for the sampling distribution based on a given sample is really capturing the shape and spread of the sampling distribution.

3.4 Bootstrapping

In the above sections, we have discussed analytical methods (both exact and approximation through limit theorems) for the sampling distribution. Given a set of data, we can also construct a model for the sampling distribution empirically. As there is no single Central Limit Theorem, there is no single bootstrapping algorithm. Instead, "bootstrapping" refers to the idea of using resampling methods to model a sampling distribution of a statistic, but the easiest algorithm is defined below.

Definition 3.8 (Case-Resampling Bootstrap)**.** Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from an underlying population, and let θ represent a parameter of interest characterizing the underlying population. Further, define $\widehat{\theta} = h(\mathbf{Y})$ be a statistic which estimates the parameter. The case-resampling bootstrap algorithm proceeds as follows:

1. Take a random sample, with replacement, from the set $\{Y_1, Y_2, \ldots, Y_n\}$ of size n. Call these values $Y_1^*, Y_2^*, \ldots, Y_n^*$. This is known as a bootstrap resample.
2. Compute $\widehat{\theta}^* = h(\mathbf{Y}^*)$ and store this value. This is known as a bootstrap statistic.
3. Repeat steps 1-2 m times, for some large value of m (say $m = 5000$). Denote $\widehat{\theta}_j^*$ to be the bootstrap statistic from the j-th bootstrap resample.

The empirical distribution of $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_m^*$ will approximate the shape and spread of the sampling distribution of the statistic $h(\mathbf{Y})$.

While the proof of the efficacy of a bootstrap algorithm is beyond the scope of this text, we can gain some intuition regarding the process. Let's start by characterizing the distribution from which the algorithm resamples — the distribution of the sample. When we sample, with replacement, from the original sample, only n values are possible $(Y_1, Y_2, \ldots, Y_n)$. And, each value will be selected with probability $1/n$. That is, we have that

$$Pr(Y_i^* = u) = \frac{1}{n} \qquad u \in \{Y_1, Y_2, \ldots, Y_n\}.$$

The mean of this distribution is represented by

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i,$$

and the variance of this distribution is

$$S_n^2 = \frac{1}{n}\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2.$$

We divide by n instead of $n-1$ because since we are treating the original sample as a population from which to be sampled, we rely on the formulas from Chapter 2.

If $\hat{\theta} = \bar{Y}$, then we have that the sampling distribution of $\hat{\theta}^* = n^{-1}\sum_{i=1}^{n} Y_i^*$ will have a mean of

$$E\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^*\right] = \bar{Y}$$

and a variance of

$$Var\left[\frac{1}{n}\sum_{i=1}^{n} Y_i^*\right] = \frac{S_n^2}{n}.$$

These are the direct application of Example 3.1 and Theorem 3.2. And, if n is large enough, we would expect the resulting empirical distribution to approximate that of a Normal distribution as a result of the CLT. This highlights that *models* of the sampling distribution tend to have similar characteristics to that of sampling distributions.

Common Characteristics of Models for Sampling Distributions

While not guaranteed, the model of a sampling distribution of many statistics tend to have the following characteristics:

1. The model of the sampling distribution is centered on the statistic from the original sample, or the center of the model approaches the statistic from the original sample as the sample size increases.
2. The spread of the model of the sampling distribution is smaller than within the sample, and the spread decreases as the sample size increases.
3. The shape of the model of the sampling distribution differs from that of the sample, and the model of the sampling distribution becomes more bell-shaped as the sample size increases.

Of the three characteristics above, the third is the one most likely to be broken.

3.5 Application

Consider the following research objective:

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

As stated, this is an ill-posed objective as it is not centered on a parameter. We might refine it to be

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

We have a large ($n = 53940$) sample of diamonds at our disposal that can be used to address this research objective. A plot of the sample is shown in Figure 3.1.

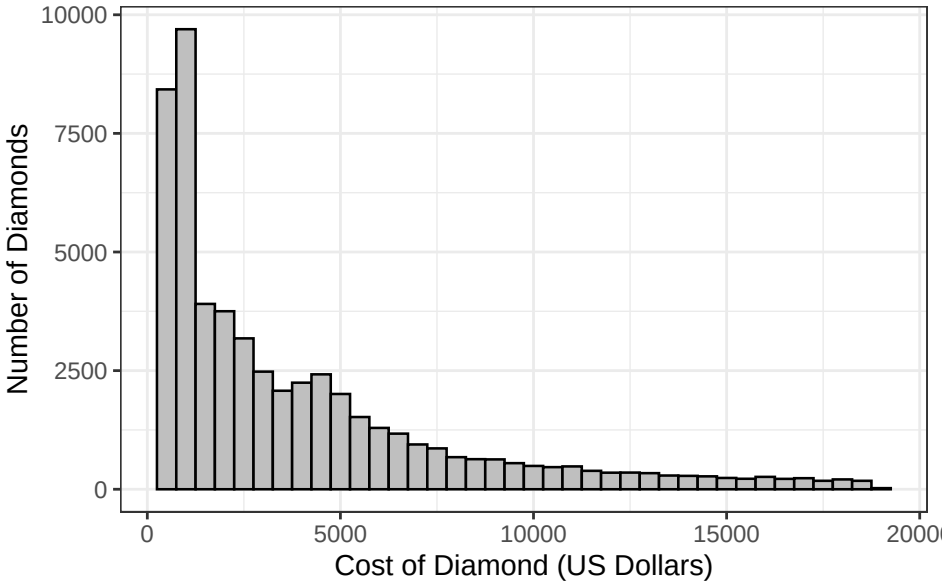

Figure 3.1: Distribution of the cost of a diamond sold in the United States.

Using the sample, we conduct 5000 bootstrap replications, each time computing the sample mean. The bootstrap sampling distribution is shown in Figure 4.3. We have overlayed the model suggested by the CLT as well. Observe that even though the sample was skewed to the right, the model for the sampling distribution is bell-shaped. The center of our model is the observed sample mean of 3933, and the spread of the sampling distribution is much smaller

than that observed in the sample. With a large sample size, we see that the empirical model of the sampling distribution is very similar to that suggested by the CLT.

> **Note**
>
> Bootstrapping can be used to qualitatively assess whether the CLT is appropriate for a particular sample. Of course, if we have gone through the effort of constructing an empirical model, we would likely rely on the empirical model.

Figure 4.3 does not include the results from Theorem 3.3; the rationale for excluding this result is that a quick glance at Figure 3.1 is enough to convince us that the underlying population does not follow a Normal distribution; therefore, those results are inappropriate.

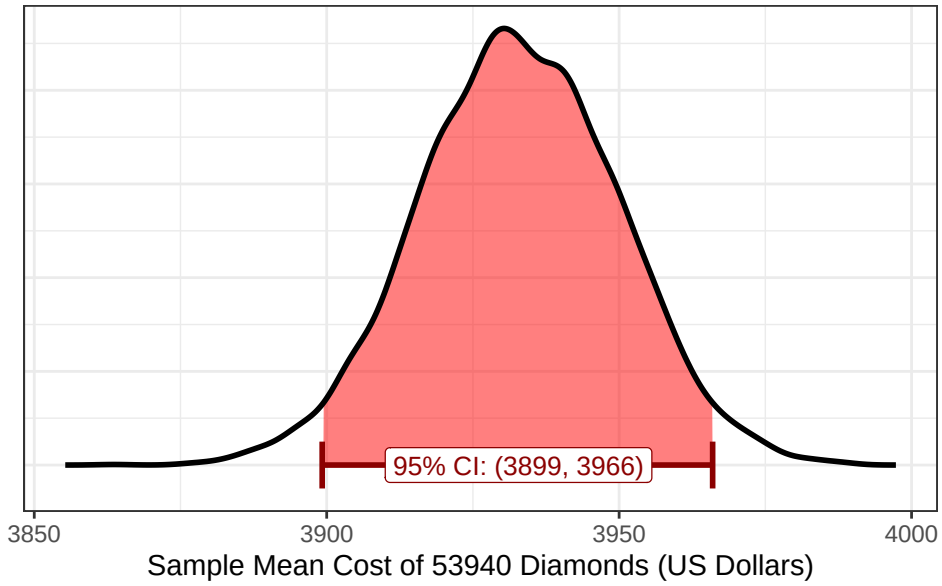

Figure 3.2: Bootstrap sampling distribution of the sample mean cost of a diamond based on the available data. 5000 bootstrap replicates were used.

The next chapter considers ways of using the above tools to perform inference.

4 Inference for a Population Mean

In the previous chapter, we examined properties of sampling distribution, with particular attention paid to the sampling distribution of the sample mean. We also discussed both analytical and empirical methods for modeling the sampling distribution of a statistic. In this chapter, we build on those results to conduct inference on a parameter.

4.1 Confidence Intervals

Sampling distributions characterize the variability in a statistic across repeated samples. That is, we expect the statistic to differ from one sample to another. However, what the sampling distribution reveals is that the degree to which these statistics vary, and the degree to which they vary from the corresponding parameter, is quantifiable. This is how sampling distributions are used in a probability course — given an underlying population, determine the probability that the sample mean exceeds a particular value. In statistics, however, we want to go in the other direction — given a set of data, determine a suitable interval of estimates for the parameter.

Figure 4.1 accompanies the "forward" direction that would accompany a probability problem. Given the sampling distribution, we would be 95% sure the statistic would fall within d units of the parameter θ, as illustrated by the shaded region.

Given a sample, we want to "reverse engineer" the problem. That is, the sampling distribution tells us that the statistic would likely not fall more than d units from the parameter. Therefore, if we take a sample from the underlying population and compute a statistic, we would expect the parameter to be within d units of this statistic. This is illustrated in Figure 4.2, where the difference is that the model for the sampling distribution is centered on the statistic. We call this a **confidence interval**.

Definition 4.1 (Confidence Interval)**.** Consider repeatedly taking samples $\mathbf{Y}$ of size n from a population characterized by the parameter θ. The interval $(h_1(\mathbf{Y}), h_2(\mathbf{Y}))$ is said to be a $100c\%$ confidence interval if

$$Pr\left(h_1(\mathbf{Y}) \leq \theta \leq h_2(\mathbf{Y})\right) = c.$$

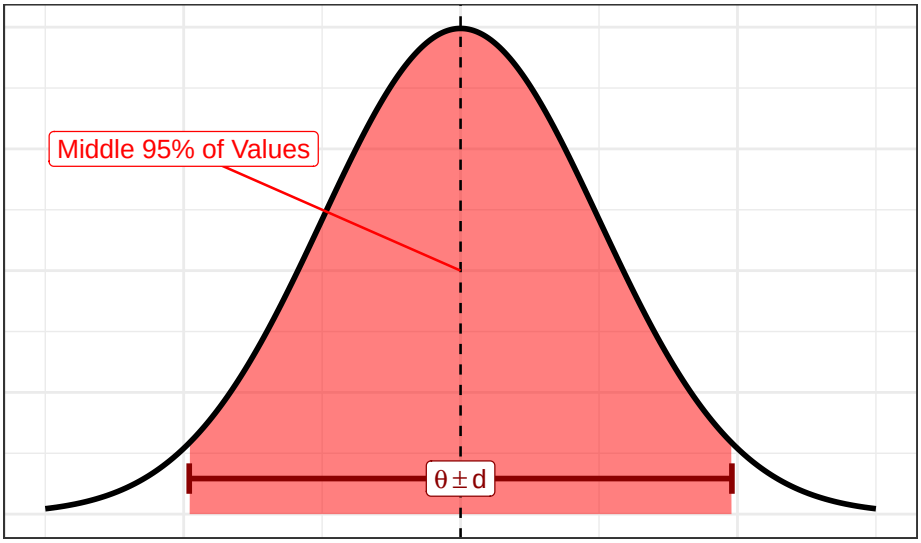

Figure 4.1: Illustration of using a sampling distribution to determine the values of a statistic we would likely observe in a new sample.

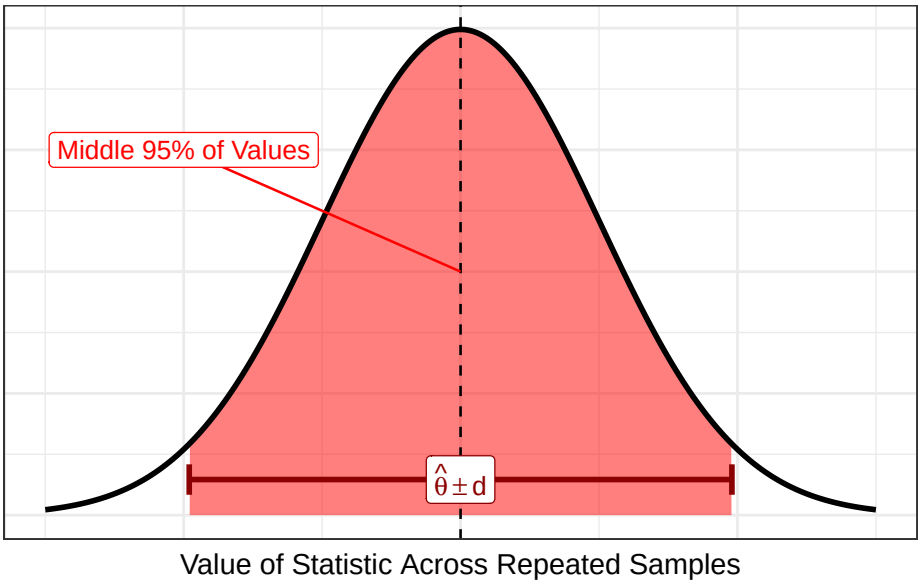

Figure 4.2: Illustration of using a model for the sampling distribution to determine the values of the parameter that are consistent with the observed sample.

> **Warning**
>
> It is important to note that in the definition of a confidence interval, the statistics $h_1(\mathbf{Y})$ and $h_2(\mathbf{Y})$ are the random variables; the parameter θ is fixed. That is, it is the *interval* that is moving across the repeated samples, not the parameter. Instead of saying "the probability the parameter is within the interval," we would say "the probability the interval captures the parameter." While this may seem subtle, it is important for correctly interpreting the interval.

Notice that in the definition of a confidence interval depends on repeated sampling. Once we have a sample, probability no longer makes sense; the interval either captures the parameter or it does not, but our ignorance of the result does not warrant a probability statement. This is a direct result of our interpretation of probability (Definition 1.5).

Example 4.1 (CI for Sample Mean Using CLT)**.** Let $Y_1, Y_2, \ldots, Y_n$ be a large random sample from a population with a finite mean μ and variance σ^2. Develop a $100c\%$ confidence interval for the population mean μ.

Solution. Define $z_{0.5(1+c)}$ to be the value such that

$$Pr\left(Z \leq z_{0.5(1+c)}\right) = 0.5(1+c)$$

for any $0 < c < 1$ where $Z \sim N(0,1)$. Then, we know that

$$c = Pr\left(-z_{0.5(1+c)} \leq Z \leq z_{0.5(1+c)}\right)$$

because of the symmetry of the Normal distribution. Defining $\bar{Y}$ and S to be the sample mean and standard deviation from the sample, from the CLT, we know that the ratio $\frac{\sqrt{n}(\bar{Y}-\mu)}{S}$ can be approximated by a Standard Normal distribution. That is, probability statements about this ratio are equivalent to probability statements about a Standard Normal random variable. Therefore, we have

$$c = Pr\left(-z_{0.5(1+c)} \leq \frac{\sqrt{n}\left(\bar{Y}-\mu\right)}{S} \leq z_{0.5(1+c)}\right).$$

We now rearranging terms, we have

$$
\begin{aligned}
c &= Pr\left(-z_{0.5(1+c)} \leq \frac{\sqrt{n}\left(\bar{Y}-\mu\right)}{S} \leq z_{0.5(1+c)}\right) \\
&= Pr\left(-z_{0.5(1+c)}\frac{S}{\sqrt{n}} \leq \left(\bar{Y}-\mu\right) \leq z_{0.5(1+c)}\frac{S}{\sqrt{n}}\right) \\
&= Pr\left(\bar{Y}-z_{0.5(1+c)}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y}+z_{0.5(1+c)}\frac{S}{\sqrt{n}}\right).
\end{aligned}
$$

Defining $h_1(\mathbf{Y}) = \bar{Y} - z_{0.5(1+c)}\frac{S}{\sqrt{n}}$ and $h_2(\mathbf{Y}) = \bar{Y} + z_{0.5(1+c)}\frac{S}{\sqrt{n}}$, we have identified an interval $(h_1(\mathbf{Y}), h_2(\mathbf{Y}))$ that satisfies the definition of a $100c\%$ confidence interval.

Consider the following research objective:

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

In the previous chapter, we saw that the CLT could be used to model the sampling distribution of the sample mean using the available data. Applying the result of Example 4.1, we have that a 95% confidence interval for the average cost of a diamond is given by

$$
\begin{aligned}
&\bar{y} \pm z_{0.925}\left(s/\sqrt{n}\right) \\
&3932.8 \pm (1.96)(3989.4/\sqrt{53940}) \\
&(3899,\ 3966).
\end{aligned}
$$

While Example 4.1 provided a formula for a confidence interval using the CLT, we can also develop a confidence interval in the same spirit from an empirical model. Simply define $h_1(\mathbf{Y})$ and $h_2(\mathbf{Y})$ to be the 2.5-th and 97.5-th percentiles from the empirical sampling distribution. Then, we will have a valid 95% confidence interval. **?@fig-inference-bootstrap** illustrates the 95% CI using the same data.

Not surprisingly, since we saw in the last chapter that the empirical model and the CLT were extremely similar, the 95% CI from the empirical model matches the 95% CI given by the CLT.

4.2 Null Distributions and P-Values

The disciplines of probability and statistics blend together most naturally when performing a hypothesis test. In Chapter 3, we examined sampling distributions of a statistic, with particular attention paid to the sample mean. We had to transition to modeling these sampling distributions using data since the population parameters, which also characterize the sampling distribution of the statistic, are unknown. When we are performing a hypothesis test, however, the null distribution specifies a particular value of the unknown parameter(s); this then allows us to make use of our models directly! This is known as a **null distribution**.

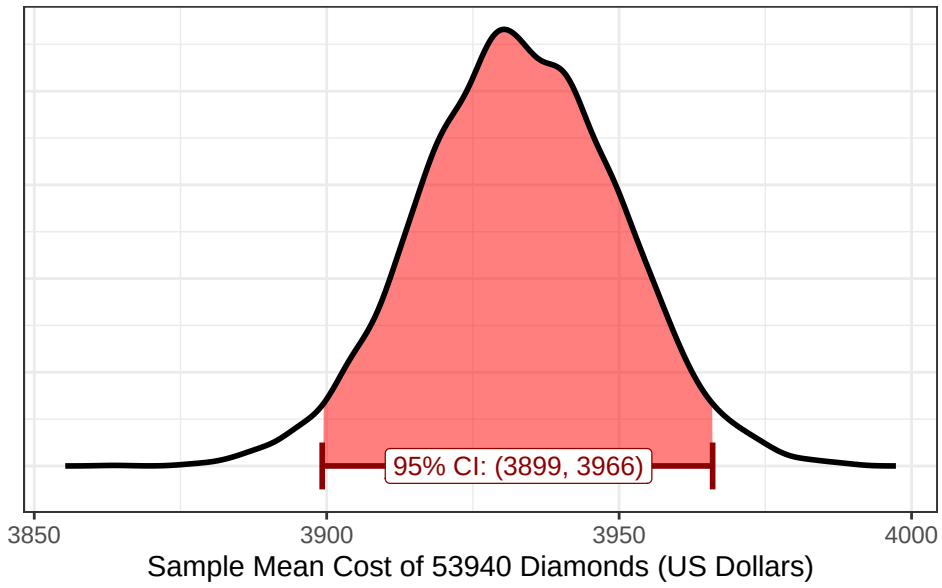

Figure 4.3: Bootstrap sampling distribution of the sample mean cost of a diamond based on the available data. 5000 bootstrap replicates were used.

Definition 4.2 (Null Distribution)**.** Distribution of a statistic under a hypothesized value of the population parameter(s).

Let's return to our investigation of diamond prices. Suppose we are interested in the following research question:

> Is it reasonable that the average cost (in US dollars) of a diamond is \$3900? Or, does the sample provide evidence that the average cost of a diamond exceeds \$3900?

Letting θ represent the average cost (in US dollars) of a diamond, the above research question can be captured using the following hypotheses:

$$H_0 : \theta \leq 3900 \qquad \text{vs.} \qquad H_1 : \theta > 3900.$$

From the CLT, we have that the sampling distribution of the ratio

$$\frac{\sqrt{n}\left(\bar{Y} - \theta\right)}{S}$$

can be approximated by a Standard Normal distribution, where $\bar{Y}$ and S represent the sample mean and standard deviation, respectively. In the above ratio, θ is unknown. However, *if we are willing to assume the above null hypothesis is true*, then we know that $\theta = 3900$.

> **Note**
>
> When working with a one-sided hypothesis, "assuming the null hypothesis" always considers the boundary of the interval. The idea here is that if we can establish evidence against this boundary, then we can establish evidence against any other value represented in the null hypothesis.

That is, *if the null hypothesis is true*, then we have that the sampling distribution of the ratio

$$\frac{\sqrt{n}\left(\bar{Y} - 3900\right)}{S}$$

can be approximated by a Standard Normal distribution. Essentially, we are using the CLT combined with the knowledge provided by the null hypothesis about the parameters. That is, we really are discussing the *null distribution*. Now, we can use this distribution to make probabilistic statements. For example, we might ask the following question:

> Assuming the null hypothesis is true, what is the probability that the ratio $\frac{\sqrt{n}(\bar{Y}-3900)}{S}$ would meet or exceed 1.909?

This is a straight-forward probability problem. Observe that

$$\begin{aligned} Pr\left(\frac{\sqrt{n}\left(\bar{Y} - 3900\right)}{S} \geq 1.909\right) &= Pr(Z \geq 1.909) \\ &= 0.0281 \end{aligned}$$

where $Z \sim N(0, 1)$. That is, assuming the mean cost of a diamond is \$3900, there is only a 2.81% chance that we would observe a ratio of at least 1.909. However, using the data we obtained, we find that the ratio we observed in our sample was 1.909, where we use the observed sample mean and standard deviation in the computation. That is, if the average cost of a diamond is \$3900, if we were to repeat the study, there is only a 2.81% chance that we would observe data that would provide a ratio as extreme or more so than that observed. Further, larger values of this ratio are more consistent with the alternative hypothesis (average costs larger than \$3900). So, if the average cost of a diamond is \$3900, there is only a 2.81% chance that we would observe data as consistent or more so with the alternative hypothesis as that observed. This is known as a **p-value**.

Definition 4.3 (P-value)**.** The probability, assuming the null hypothesis is true, that we would observe a statistic, by chance alone, as extreme or more so than that observed in the sample.

> **Note**
>
> Often times, our analytic results imply a sampling distribution (or an approximation of the sampling distribution) for a ratio instead of the statistic of interest. This ratio is often called the "standardized (test) statistic" in hypothesis testing because the ratio, when evaluated with the observed data, provides a metric quantifying the difference between our expectations and our observations in the sample.

5 Inference for Regression Models

The previous chapters introduced the primary components of inference. Particular attention was paid to making inference for the mean of a quantitative variable. Most interesting questions involve examining a relationship between two variables. For example, consider the following question:

> Does the average cost of a diamond (in US dollars) tend to change as the carat (a measure of the size) changes?

The above question depends on the relationship between the cost of a diamond and the carat of the diamond. In this chapter, we examine some of the probabilistic components associated with such problems, known as **regression**.

Definition 5.1 (Regression)**.** Allowing the parameters characterizing the distribution of a random variable to depend, through some specified function, on the value of additional variables.

5.1 Simple Linear Regression Model

In previous chapters, we considered distributional models for the population of the form

$$Y_1, Y_2, ..., Y_n \overset{\text{Ind}}{\sim} N\left(\mu, \sigma^2\right),$$

for some unknown mean μ and variance σ^2. Note that under such a distributional model, if we knew the value of the parameters μ and σ, we know everything we need to know to fully characterize the variability in the response Y. However, knowing the parameters does not explain *why* there is variability in the population. Why doesn't every unit share the exact same value of the response? We know that variability is inherent in any process; perhaps all of the variability in this response is due to measurement error. If that were true, then if we had infinitely precise measuring tools, then the variability would be eliminated.

It is more likely that the various units are not meant to be completely identical. That is, there are additional characteristics that might explain why each unit has a different response. For example, in our sample of diamonds, the units (the diamonds) are not identical. Some diamonds are larger than others; some have superior color or clarity. These differences in

the characteristics of the diamonds might explain their different prices. Unfortunately, the distributional model above does not incorporate additional characteristics.

The research question posed above suggested that the average response (cost) might depend upon another variable (carat), a predictor. The simple linear regression model allows the average response to depend upon the predictor through a linear function.

Definition 5.2 (Classical Simple Linear Regression)**.** Let $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$ be observations made on a sample of n units. Under the classical simple linear regression model, the distributional model for the response among the population is given by

$$Y_1, Y_2, \ldots, Y_n \overset{\text{Ind}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

where β_0, β_1, and σ^2 are unknown parameters.

> **Note**
>
> In Definition 5.2, note that we only consider the response to be a random variable; the predictor is considered fixed (hence the use of a lowercase x instead of a capital X). This is consistent with the idea of a designed experiment for which the values of the predictor can be determined in advance by the researchers.
> However, for observational studies, the values of the predictor cannot be fixed. That is, in practice, the predictor is also unknown in advance and is therefore a random variable as well. In such cases, we can proceed in the same manner, considering the *conditional* distribution of the response *given* the predictor.

Notice that Definition 5.2 simply extends the form of the distributional model we have previously considered. It makes it clear that

$$E(Y_i) = \beta_0 + \beta_1 x_i$$

for each unit. Notice that we do not say that the observations are "identically distributed," because they are not! The mean response differs (at least potentially) for each observation as it depends on the corresponding value of the predictor.

While the above presentation connects the process for the inference of a single mean with that of regression, it is not the common presentation. Instead, the model is traditionally presented as saying that

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where $\varepsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2)$, where now we can make use of the "identically distributed" language. In this presentation, we have introduced a new random variable, ε. Since the expression $\beta_0 + \beta_1 x_i$ does not contain a random variable, it is deterministic in nature. Therefore, the distribution of Y_i is determined because we are simply shifting the distribution of ε_i.

Whenever we posit a distributional model for the population, as we do in Definition 5.2, we are making a fairly strong assumption. In practice, we may not feel comfortable positing the form of the distribution. We can reduce the conditions on ε and still retain some of the key characteristics of the model.

Regardless of the conditions we impose on ε, we are essentially specifying a model for the data generating process — the set of statements we are willing to make regarding the variability in the response. We know that since the response is a random variable it has some distribution. The model for the data generating process is really a set of statements about that distribution. We may only be characterizing the mean of the distribution of the response; we may be willing to characterize the mean and the variance; or, we may be willing to fully characterize the distributional form.

5.2 Least Squares Estimation

While the previous section defines a model for the data generating process, it does not specify how to estimate the unknown parameters. Prior to this section, we were able to estimate the parameters by making analogous computations within the sample. As we replace a singular parameter with a function, corresponding computations are no longer obvious. A general process for estimating the unknown parameters that define the mean response is the **method of least squares**.

Definition 5.3 (Method of Least Squares)**.** The least squares estimates of the parameters β_0 and β_1 in a simple linear regression model are the values that minimize the quantity

$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 .$$

The estimates are often denoted $\hat{\beta}_0$ and $\hat{\beta}_1$.

Theorem 5.1 (Least Squares Estimates)**.** *Let $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$ be observations made on a sample of n units. The least squares estimates for the simple linear regression model relating the response and predictor are given by*

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Proof. By definition, the least squares estimates are the values of the parameters that minimize the quantity

$$\sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 .$$

To minimize this quantity, we can consider taking partial derivatives with respect to each quantity:

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)$$
$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 x_i)^2 = -2 \sum_{i=1}^n x_i (Y_i - \beta_0 - \beta_1 x_i) .$$

Setting each derivative equal to zero, we have

$$0 = n\bar{Y} - n\beta_0 - n\beta_1 \bar{x}$$
$$0 = \sum_{i=1}^n x_i Y_i - n\beta_0 \bar{x} - \beta_1 \sum_{i=1}^n x_i^2.$$

We now have two equations and two unknown terms. Solving the first equation, we have that

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}.$$

Plugging this into the second expression, we have

$$\begin{aligned} 0 &= \sum_{i=1}^n x_i Y_i - n\bar{Y}\bar{x} + n\beta_1 \bar{x}^2 - \beta_1 \sum_{i=1}^n x_i^2 \\ &= \sum_{i=1}^n x_i \left(Y_i - \bar{Y}\right) - \beta_1 \sum_{i=1}^n (x_i^2 - \bar{x}^2) \\ &= \sum_{i=1}^n (x_i - \bar{x} + \bar{x})\left(Y_i - \bar{Y}\right) - \beta_1 \sum_{i=1}^n \left(x_i^2 - \bar{x}^2\right), \end{aligned}$$

where the second line rearranges the terms by bringing things inside the sums, and the third line "does nothing" by adding and subtracting the same term, $\bar{x}$. Note that the second term can be written as

$$\begin{aligned}\beta_1 \sum_{i=1}^{n} (x_i^2 - \bar{x}^2) &= \beta_1 \sum_{i=1}^{n} \left[(x_i - \bar{x})^2 + 2x_i\bar{x} - 2\bar{x}^2 \right] \\ &= \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \beta_1 2n\bar{x}^2 - 2n\bar{x}^2 \\ &= \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 .\end{aligned}$$

And, the first term simplifies to

$$\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right) + \bar{x} \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right) = \sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right).$$

Putting these components together, we have that the least squares estimate of β_1 must satisfy the equality

$$0 = \sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right) - \beta_1 \sum_{i=1}^{n} (x_i - \bar{x})^2 .$$

This gives us that

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Of course, we have simply found the limits; to establish these are actually minimums, we must consider the Hessian matrix, which consists of second-order partial derivatives. For our model, we have

$$\begin{aligned}\frac{\partial^2}{\partial \beta_0^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 &= 2n \\ \frac{\partial^2}{\partial \beta_1^2} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 &= 2 \sum_{i=1}^{n} x_i^2 \\ \frac{\partial^2}{\partial \beta_1 \partial \beta_0} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2 &= 2n\bar{x}.\end{aligned}$$

The determinant of the Hessian matrix is then

$$\begin{aligned}
4n\sum_{i=1}^{n} x_i^2 - 4n\bar{x}^2 &= 4n\left[\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right] \\
&= 4n\left\{\sum_{i=1}^{n}\left[(x_i-\bar{x})^2 + 2x_i\bar{x} - \bar{x}^2\right] - \bar{x}^2\right\} \\
&= 4n\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + n\bar{x}^2 - \bar{x}^2\right] \\
&= 4n\left[\sum_{i=1}^{n}(x_i-\bar{x})^2 + (n-1)\bar{x}^2\right] \\
&> 0.
\end{aligned}$$

Since the determinant is always positive, we have that our least squares estimates minimize the quantity of interest. □

> **Note**
>
> In the proof above, we essentially derive a very helpful relation that pops up often in statistical proofs:
>
> $$\sum_{i=1}^{n}(x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$$

The method of least squares can be used to estimate the parameters in the mean model, but this is not *statistics*; it is *mathematics*. In particular, the method of least squares is just one of several optimization problems we might have defined to estimate the parameters. Statistics is the discipline of characterizing variability. Characterizing the uncertainty in these estimates — characterizing their sampling distribution — that is where we move into a statistical analysis.

Example 5.1 (Expected Value of the Intercept)**.** The least squares estimate of the intercept is an unbiased estimator. Establish this, assuming that the least squares estimate of the slope is also an unbiased estimator.

Solution. To establish that this statement is correct, observe that

$$
\begin{aligned}
E\left(\hat{\beta}_0\right) &= E\left(\bar{Y} - \hat{\beta}_1\bar{x}\right) \\
&= E\left(n^{-1}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right) - \hat{\beta}_1\bar{x}\right) \\
&= E\left(n^{-1}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i + \varepsilon_i\right)\right) - \bar{x}E\left(\hat{\beta}_1\right) \\
&= n^{-1}\sum_{i=1}^{n}\left(\beta_0 + \beta_1 x_i + E\left(\epsilon_i\right)\right) - \bar{x}\beta_1 \\
&= \beta_0 + \beta_1\bar{x} - \bar{x}\beta_1 \\
&= \beta_0.
\end{aligned}
$$

Note that in line 2, we insert the value of Y_i from the model for the data generating process; in line 3, we make use of the linearity of expectations; in line 4, we assume that the least squares estimate of the slope is an unbiased estimator. Finally, in line 5, we assume that the error term has an average of 0.

Note that in our solution, the only assumption we needed on the error term was that it had an average of 0. While the other conditions on the error term typically assumed (classical regression model) are helpful in characterizing the sampling distribution of the least squares estimates, only this one condition is needed to establish where that sampling distribution is centered.

Theorem 5.2 (Sampling Distribution for Least Squares Estimators)**.** *Under the conditions of the Classical Regression Model (Definition 5.2), we have that*

$$
\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim t_{n-2}
$$

where

$$
\begin{aligned}
Var\left(\hat{\beta}_0\right) &= \hat{\sigma}^2\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right) \\
Var\left(\hat{\beta}_1\right) &= \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}
\end{aligned}
$$

and

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)^2$$

is our estimate of the unknown population variance σ^2*.*

Once we have a model for the sampling distribution, we have the ability to make inference — confidence intervals or p-values.

6 More on the Classical Model

In the previous chapter, we introduced the classical simple linear regression model (Definition 5.2). For a sample $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$, the distributional model for the response under the classical model is

$$Y_1, Y_2, \ldots, Y_n \overset{\text{Ind}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$$

where β_0, β_1, and σ^2 are unknown parameters. As we have seen, this embeds four conditions on the distribution — that the mean is correctly specified, that the observations are independent of one another, that the variance is constant, and that the Normal distribution is appropriate. While each of these plays a role in determining the sampling distribution of the parameter estimates, the impact of the Normality assumption is unique. The moment we remove the assumption of Normality, the sampling distribution is no longer tractable analytically.

In this chapter, we examine some additional results associated with the Normal distribution that are instrumental in classical theory of the linear model.

6.1 Linear Combinations

In general, combining random variables does not yield a predictable result; the Normal distribution is a notable exception. The following theorem builds on an example seen earlier in the text.

Theorem 6.1 (Linear Combination of Independent Normal Random Variables)**.** *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables such that*

$$Y_1, Y_2, \ldots, Y_n \overset{Ind}{\sim} N(\mu_i, \sigma_i^2).$$

Let $a_1, a_2, \ldots, a_n \in \mathbb{R}$ be known constants. Then,

$$\sum_{i=1}^{n} a_i Y_i \sim N\left(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i \sigma_i^2\right).$$

Proof. Let $Z = \sum_{i=1}^{n} a_i Y_i$. Then, observe that

$$\begin{aligned} M_Z(t) &= E\left(e^{tZ}\right) \\ &= E\left(e^{t\sum_{i=1}^{n} a_i Y_i}\right) \\ &= E\left(\prod_{i=1}^{n} e^{ta_i Y_i}\right) \\ &= \prod_{i=1}^{n} E\left(e^{ta_i Y_i}\right), \end{aligned}$$

where the last line is a result of the random variables being independent of one another. Note that we do not get a further simplification here because we are *not* assuming that the random variables are identically distributed. Looking up the form of the moment generating function for a Normal random variable and substituting that here, we have

$$\begin{aligned} M_Z(t) &= \prod_{i=1}^{n} E\left(e^{ta_i Y_i}\right) \\ &= \prod_{i=1}^{n} e^{\mu_i(ta_i)+\sigma_i^2(ta_i)^2/2} \\ &= \exp\left\{\sum_{i=1}^{n} \mu_i t a_i + \frac{1}{2}\sum_{i=1}^{n} \sigma^2 t^2 a_i^2\right\} \\ &= \exp\left\{t\left(\sum_{i=1}^{n} a_i \mu_i\right) + \frac{t^2}{2}\left(\sum_{i=1}^{n} a_i^2 \sigma^2\right)\right\}. \end{aligned}$$

We now recognize this as the form of an MGF of a Normal distribution with a mean of $\sum_{i=1}^{n} a_i \mu_i$ and a variance of $\sum_{i=1}^{n} a_i^2 \sigma^2$. By the uniqueness of MGF's, we have the result. □

Why is Theorem 6.1 so important? It turns out that many of the statistics of interest that result from applying least squares are linear combinations of the response. And, under the classical simple linear regression model, these responses are independent variates from a Normal distribution!

We have already seen one of these results in an earlier chapter when examining the sampling distribution of the sample mean from a Normal distribution.

Example 6.1 (Form of Slope from Simple Linear Regression)**.** Recall that the least squares estimate for the slope in a simple linear regression model is given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

Show that this can be written as a linear combination of the responses.

Solution. Consider expanding the numerator to observe that

$$\begin{aligned}
\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right) &= \sum_{i=1}^{n} (x_i - \bar{x}) Y_i - \sum_{i=1}^{n} (x_i - \bar{x}) \bar{Y} \\
&= \sum_{i=1}^{n} (x_i - \bar{x}) Y_i - n\bar{x}\bar{Y} + n\bar{x}\bar{Y} \\
&= \sum_{i=1}^{n} (x_i - \bar{x}) Y_i.
\end{aligned}$$

Now, substituting this into the definition of $\hat{\beta}_1$, we are able to rewrite the least squares estimator of the slope as

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) \left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n} (x_i - \bar{x}) Y_i}{\sum_{i=1}^{n} (x_i - \bar{x})^2}.
\end{aligned}$$

Recognizing that the denominator is a constant with respect to the sum in the numerator, we can define

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

Then, we have that

$$\hat{\beta}_1 = \sum_{i=1}^{n} a_i Y_i.$$

The real power of recognizing that the least squares estimate of the slope is a linear combination of the responses is that we can apply Theorem 6.1. Immediately, we have the sampling distribution of the least squares estimator of the slope:

$$\hat{\beta}_1 \sim N\left(\sum_{i=1}^{n} a_i \left(\beta_0 + \beta_1 x_i\right), \sigma^2 \sum_{i=1}^{n} a_i^2\right)$$

where a_i was defined in the solution to Example 6.1. This reduces to

$$\hat{\beta}_1 \sim N\left(\beta_1, \sigma^2 \sum_{i=1}^{n} a_i^2\right).$$

Note

Notice that since $\sum_{i=1}^{n} (x_i - \bar{x})\, q = 0$ for any constant q, we have that $\sum_{i=1}^{n} a_i = 0$ and $\sum_{i=1}^{n} a_i x_i = 1$. Further, we have that

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

We know from probability if $X \sim N(\mu, \sigma^2)$, then $(X - \mu)/\sigma) \sim N(0, 1)$. So, it would seem that Example 6.1, and the resulting discussion, would suggest that

$$\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 \sum_{i=1}^{n} a_i^2}} \sim N(0, 1),$$

but in the previous chapter, we suggested using the t-distribution. The difference is that in the above, it is assumed that σ^2 is known. Estimating the parameter σ^2 impacts the sampling distribution.

Big Idea

When the sampling distribution of an estimator depends on unknown nuisance parameters, the estimation of those nuisance parameters impacts the sampling distribution of the estimator.

6.2 Results for Squares

The previous section considered a linear combination of Normal random variables. In this section, we explore a related result.

Theorem 6.2 (Sum of Squared Normal Random Variables)**.** *Let* $Y_1, Y_2, \ldots, Y_n \overset{IID}{\sim} N(0,1)$. *Then,*

$$\sum_{i=1}^{n} Y_i^2 \sim \chi_n^2.$$

As with previous proofs, this is best addressed through a moment generating function argument.

Proof. Let $Z = \sum_{i=1}^{n} Y_i^2$; then, observe that

$$\begin{aligned} M_Z(t) &= E\left(e^{tZ}\right) \\ &= E\left(e^{t\sum_{i=1}^{n} Y_i^2}\right) \\ &= E\left(\prod_{i=1}^{n} e^{tY_i^2}\right) \\ &= \prod_{i=1}^{n} E\left(e^{tY_i^2}\right) \\ &= \left[M_{Y_1^2}(t)\right]^n \end{aligned}$$

where line 4 is a result of independence and line 5 is a result of the random variables being identically distributed. Unfortunately, we do not know the MGF of Y_1^2; so, we must first determine its distribution before proceeding. This is best done through a transformation. Observe that

$$\begin{aligned} F_{Y^2}(y) &= Pr\left(Y^2 \leq y\right) \\ &= Pr\left(-\sqrt{y} \leq Y \leq \sqrt{y}\right) \\ &= F_Y(\sqrt{y}) - F_Y(-\sqrt{y}). \end{aligned}$$

And, since $Y \sim N(0,1)$, we know the density function of Y; therefore,

$$\begin{aligned} f_{Y^2}(y) &= \frac{\partial}{\partial y} F_{Y^2}(y) \\ &= \frac{\partial}{\partial y} F_Y(\sqrt{y}) - \frac{\partial}{\partial y} F_Y(-\sqrt{y}) \\ &= f_Y(\sqrt{y})(1/2)y^{-1/2} + f_Y(-\sqrt{y})(1/2)y^{-1/2} \\ &= y^{-1/2} f_Y(\sqrt{y}) \end{aligned}$$

where the last line is based on the symmetry of the Standard Normal distribution. Thus, we have that the density of Y^2 is given by

$$\frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-\frac{1}{2}y^2} = \frac{1}{2^{1/2}\Gamma(1/2)} y^{1/2-1} e^{-(1/2)y^2},$$

where we have made use of the fact that $\Gamma(1/2) = \sqrt{\pi}$. We recognize this as the density function of a $Gamma(1/2, 2)$ or equivalently a χ^2_1 distribution. We therefore use the MGF for this distribution to continue our previous derivation, giving

$$\begin{aligned} M_Z(t) &= \left[M_{Y_1^2}(t)\right]^n \\ &= \left[(1-2t)^{-1/2}\right]^n \\ &= (1-2t)^{-n/2} \end{aligned}$$

which we recognize as the MGF of a χ^2_n distribution. By the uniqueness of moment generating functions, we have the result. □

6.3 Relation to the F Distribution

While this chapter has been focused on the Normal distribution, its presence in statistical analysis is often associated with other distributions (notably, the t-distribution, the Chi-Squared distribution, and the F-distribution). We have already seen one way in which the Normal distribution is associated with the Chi-Squared distribution. In this section, we establish the link between Chi-Squared distributions and the F-distribution.

Theorem 6.3 (Relationship Between Chi-Squared Distribution and the F-Distribution)**.** *Let U and V be independent Chi-Squared random variables with ν and η degrees of freedom, respectively. Then, $\frac{U/\nu}{V/\eta} \sim F_{\nu,\eta}$.*

Unlike previous results that relied on the MGF, this theorem requires that we perform a transformation.

Proof. Let $X = U/V$ and let $Y = V$. We will find the joint density of X and Y, and then integrate out Y to derive the density of X. Observe that

$$\begin{aligned} F_{X,Y}(x,y) &= Pr(X \leq x, Y \leq y) \\ &= Pr(U/V \leq x, V \leq y) \\ &= Pr(U \leq Vx, V \leq y) \\ &= \int_0^y \int_0^{vx} f_{U,V}(u,v)dudv. \end{aligned}$$

Since U and V are independent random variables, their joint density is easily given. However, we are particularly interested in the joint density of X and Y. Therefore, we need only consider the derivative. Observe that

$$\begin{aligned} f_{X,Y}(x,y) &= \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x,y) \\ &= f_{U,V}(yx,y)y \\ &= \frac{1}{2^{\nu/2}\Gamma(\nu/2)}(yx)^{\nu/2-1}e^{-yx/2}\frac{1}{2^{\eta/2}\Gamma(\eta/2)}(y)^{\eta/2-1}e^{-y/2}y. \end{aligned}$$

Therefore, we have the joint density of X and Y. We now integrate out y to obtain the marginal density of X. We have

$$\begin{aligned} f_X(x) &= \int_0^\infty f_{X,Y}(x,y)dy \\ &= x^{\nu/2-1}\frac{1}{\Gamma(\nu/2)\Gamma(\eta/2)2^{(\eta+\nu)/2}}\int_0^\infty y^{(\nu+\eta)/2-1}e^{-y(1+x)/2}dy \\ &= \frac{\left(x^{\nu/2-1}\right)\Gamma((\nu+\eta)/2)}{\Gamma(\nu/2)\Gamma(\eta/2)2^{(\eta+\nu)/2}((1+x)/2)^{(\nu+\eta)/2}}\int_0^\infty \frac{((1+x)/2)^{(\nu+\eta)/2}}{\Gamma((\nu+\eta)/2)}y^{(\nu+\eta)/2-1}e^{-y(1+x)/2}dy. \end{aligned}$$

We recognize that the integral is the density function of a Gamma distribution with shape parameter $(\nu+\eta)/2$ and rate parameter $(1+x)/2$. Since we are integrating over the entire support, the integral will go to 1. This means the density of X is given by

$$\frac{\left(x^{\nu/2-1}\right)\Gamma((\nu+\eta)/2)}{\Gamma(\nu/2)\Gamma(\eta/2)2^{(\eta+\nu)/2}((1+x)/2)^{(\nu+\eta)/2}}.$$

We can rewrite this density to be

$$\frac{\Gamma((\nu+\eta)/2)}{\Gamma(\nu/2)\Gamma(\eta/2)}x^{\nu/2-1}\left(1+x\right)^{-(\nu+\eta)/2}.$$

Therefore, the density of $Z = \frac{U/\nu}{V/\eta} = (U/V)(\eta/\nu)$ is given by

$$\begin{aligned} f_Z(z) &= \frac{\partial}{\partial z} F_Z(z) \\ &= \frac{\partial}{\partial z} Pr(U/V \leq (\nu/\eta)z) \\ &= \frac{\partial}{\partial z} F_X((\nu/\eta)z) \\ &= (\nu/\eta) f_X((\nu/\eta)z). \end{aligned}$$

Finally, substituting in our expression for the density of X, we have

$$\frac{\Gamma((\nu+\eta)/2)}{\Gamma(\nu/2)\Gamma(\eta/2)} \left(\frac{\nu}{\eta}\right)^{\nu/2} x^{\nu/2-1} \left(1 + x\frac{\nu}{\eta}\right)^{-(\nu+\eta)/2},$$

which we recognize as the density of the F-distribution with ν numerator and η denominator degrees of freedom. □

7 Location Scale Families

As we have seen, the Normal distribution plays a large role in classical statistical theory. However, it is just one example of a location-scale family. In this chapter, we briefly discuss a key property of location-scale families — probability plots.

Definition 7.1 (Location-Scale Family)**.** For $a \in \mathbb{R}$ and $b > 0$, let $X = a + bZ$ for some random variable Z. X is said to be the location-scale family associated with the distribution of Z with location parameter a and scale parameter b.

While technically we can create a location-scale family for any distribution, we generally think of those distributions for which the support remains unchanged. Many such distributions we have seen belong to this class.

Example 7.1 (Scale Family: Exponential Distribution)**.** The Exponential distribution is a scale family (the location parameter is 0). Let X follow an Exponential distribution with scale parameter λ. Show that aX also follows an Exponential distribution with scale parameter $a\lambda$.

Solution. Let $Y = aX$, then the CDF of Y is given by

$$\begin{aligned} F_Y(y) &= Pr(Y \leq y) \\ &= Pr(aX \leq y) \\ &= Pr(X \leq y/a) \\ &= F_X(y/a) \\ &= 1 - e^{-\frac{y}{a\lambda}}, \end{aligned}$$

which we recognize as the CDF of an Exponential distribution with scale parameter $a\lambda$.

The above example illustrates a very helpful result regarding location-scale families.

Theorem 7.1 (CDF for Location-Scale Families)**.** *Let Z be a random variable with CDF $F_Z(z)$, and define $X = aZ + b$ for $a \in \mathbb{R}$ and $b > 0$. Then,*

$$F_X(x) = F_Z\left(\frac{x-a}{b}\right).$$

That is, the CDF for a location-scale family is a function of the CDF of the "standard" distribution, the location parameter, and the scale parameter. Throughout a course in probability, we often make use of this property when computing probabilities of a Normal distribution (this is essentially what every Normal table in the back of a probability textbook is relying on). However, we can exploit this feature of location-scale families for much more than creating a table to enable "by-hand" computations of probabilities.

Example 7.2 (Quantiles of a Location-Scale Family)**.** Let $X \sim F_X$ be a random variable belonging to a location-scale family with location parameter $a \in \mathbb{R}$ and scale parameter $b > 0$. Let $Z \sim F_Z$ be the "standard" distribution generating this location-scale family. Find an expression for the q-th quantile of X as a function of the q-th quantile of Z.

Solution. The q-th quantile of X, call it X_q, is the value such that

$$q = F_X\left(X_q\right).$$

By Theorem 7.1, we have that

$$q = F_Z\left(\frac{X_q - a}{b}\right).$$

This implies that the q-th quantile of Z, call it Z_q is given by

$$\frac{X_q - a}{b}.$$

Therefore,

$$X_q = Z_q b + a.$$

Example 7.2 reveals that the quantiles of a location-scale distribution are linearly related to the quantiles of the corresponding "standard" distribution.

In order to see how this can be useful in a statistical analysis, we must first consider how we define a sample quantile. Our first pass at a definition might go something like the following:

> The q-th sample quantile is the value k such that $100q\%$ of observed vales in the sample are no more than k.

While this makes intuitive sense, we recognize that it does not result in a unique value. For example, consider a the simple sample $\{1, 2, 3, 4, 5\}$. In this sample, 60% of observations in the sample have a value no more than 3; so, it would seem that 3 represents the 0.6 quantile. But, we also know that 60% of observations have a value no more than 3.5; so, 3.5 is also the 0.6 quantile. More, we would generally consider 3 to be the median (0.5 quantile) of the sample! It seems our working definition of a sample quantile seems a bit off. In order to create a more rigorous definition, we first need the concept of an order statistic.

Definition 7.2 (Order Statistic)**.** Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. The j-th order statistic, denoted $X_{(j)}$ is the j-th smallest value in the sample. Special cases include

- $X_{(1)}$, the sample minimum, and
- $X_{(n)}$, the sample maximum.

Definition 7.3 (Sample Quantile)**.** Let $x_{(j)}$ denote the j-th observed order statistic in a sample. Then, the q-th quantile of the sample is given by $x_{(j)}$ for $\frac{j-1}{n} < q \leq \frac{j}{n}$ and $x_{(1)}$ if $q = 0$, for $j = 1, 2, \ldots, n$.

Let's apply the above definition to our sample $\{1, 2, 3, 4, 5\}$. For the median, we have $q = 0.5$, which falls between $\frac{2}{5}$ and $\frac{3}{5}$; therefore, we take the third observation in the dataset, which is 3.

> **Note**
>
> There are several accepted definitions for computing a sample quantile. In fact, the default definition used by the major statistical software packages (SAS, R, SPSS) do not agree. In large sample sizes, the results are nearly identical. For smaller samples, the results can differ dramatically; however, this is not cause for alarm in practice. If quantiles are the focal point of an analysis, be sure to check the definition being used by your software.

Let's consider applying the above to a sample from a Normal distribution. In particular, consider taking a random sample of 15 observations from a Normal distribution with a *known* mean of μ and a *known* variance of σ^2. For these 15 observations, note that we have that

$$Q\left(\frac{j - 0.5}{15}\right) = X_{(j)}$$

where $Q(p)$ represents the p-th quantile for $j = 1, 2, \ldots, 15$. The choice of the these 15 quantiles is somewhat arbitrary but provides 15 evenly spaced points over the interval $(0, 1)$. Additionally, since each observation is from a known Normal distribution, we can apply the results of Example 7.2 to recognize that

$$X_{(j)} = Z_{(j-0.5)/15}\sigma + \mu,$$

where

$$Z_{(j-0.5)/15} = \begin{cases} -1.83 & j = 1 \\ -1.28 & j = 2 \\ \dots & \dots \\ 1.83 & j = 15 \end{cases}$$

are known values since $Z \sim N(0, 1)$. Therefore, a plot of the observed order statistics against the corresponding quantiles of a Standard Normal distribution will fall along a line with a slope of σ and an intercept of μ.

Of course, in practice, we do not know μ and σ, but we can rely on this same idea because it should work for *any* choice of μ and σ.

Assessing Normality

Given a sample of size n, a plot of the observed quantiles against the quantiles from a Standard Normal distribution will be roughly linear if the sample is from a Normal distribution. The slope of the relationship will be an estimate of the standard deviation, and the intercept of the relationship will be an estimate of the average.

This graphic that is constructed is known as a "Quantile-Quantile" plot; it is also known as a "Probability Plot" as it is sometimes presented equivalently with different scales. Again, while the choice of which quantiles should be considered when constructing the plot is somewhat arbitrary, we want to select n evenly spaced points across the interval $(0, 1)$, making use of all the unique information in the sample. One such choice was illustrated above, choosing quantiles to correspond to $\frac{j-0.5}{n}$ for $j = 1, 2, \dots, n$.

8 Matrix View of Regression

Definition 5.2 introduced us to a model for the mean response as a function of a single predictor. Specifically, we considered $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$ to be observations made on a sample of n units. Under the classical simple linear regression model, the distributional model for the response among the population is given by

$$Y_1, Y_2, \ldots, Y_n \overset{\text{Ind}}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

where β_0, β_1, and σ^2 are unknown parameters.

This model is simplistic in that it only considers a single predictor. In reality, we often want the mean response to depend on several variables.

> **Predictors vs. Covariates**
>
> A statistical model posits a relationship between a response and one or more variables. Depending on the discipline or the use of the statistical model, we might refer to the variables in the model as "predictors," "covariates," "factors," or "treatments." We tend to refer to quantitative variables as "predictors" and categorical variables as "factors" when they appear in the mean model. However, these terms may be used differently in different disciplines.

On one hand, incorporating additional predictors into the model is straight-forward; for example, we might consider Y_i to be the response measured on the i-th observation in a sample, and $x_{1,i}, x_{2,i}, \ldots, x_{p,i}$ to be the 1st, 2nd, ..., p-th predictor measured on the i-th observation (for $i = 1, 2, \ldots, n$). Then, we might posit the following model for the response among the population:

$$Y_1, Y_2, \ldots, Y_n \overset{\text{Ind}}{\sim} N\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i}, \sigma^2\right)$$

where $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ and σ^2 are unknown parameters.

> **Note**
>
> As in Definition 5.2, note that we only consider the response to be a random variable; the predictors are considered fixed (hence the use of a lowercase x instead of a capital X). This is consistent with the idea of a designed experiment for which the values of the predictor can be determined in advance by the researchers.
> However, for observational studies, the values of the predictor cannot be fixed. That is, in practice, the predictor is also unknown in advance and is therefore a random variable as well. In such cases, we can proceed in the same manner, considering the *conditional* distribution of the response *given* the predictor.

While we can continue to extend results from Chapter 5 to the case of p predictors, the notation can become tedious. It helps to express our model using matrices.

8.1 Expressing Linear Combinations

You might have noticed that we wrote our mean model as a linear combination of predictors

$$\sum_{j=1}^{p} \beta_j x_{j,i}.$$

Actually, while it may not be clear at this stage, it is better to recognize the mean model as a linear combination of *parameters*, written as

$$\beta_0 + \sum_{j=1}^{p} \beta_j x_{j,i} = 1\beta_0 + x_{1,i}\beta_1 + \cdots + x_{p,i}\beta_p.$$

Recall that a linear combination can be written as a dot product of two vectors. Therefore, we can express this linear combination as

$$1\beta_0 + x_{1,i}\beta_1 + \cdots + x_{p,i}\beta_p = \mathbf{x}_i^\top \beta,$$

where

$$\mathbf{x}_i = \begin{pmatrix} 1 \\ x_{1,i} \\ x_{2,i} \\ \vdots \\ x_{p,i} \end{pmatrix}$$

and

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}.$$

> **Note**
>
> All vectors are by default column vectors, which corresponds to how they are stored in statistical programming languages.

Notice we maintain the i subscript, because the linear combination $\mathbf{x}_i^\top \beta$ will differ for each observation in our sample (the values of the variables for one observation will differ from the values of the variables for other observations). We might then re-express distributional model for the response as

$$Y_i \overset{\text{Ind}}{\sim} N\left(\mathbf{x}_i^\top \beta, \sigma^2\right).$$

8.2 Multiple Regression

Often termed "multiple regression," we write out the model for the response when we allow the mean to be a linear function of several predictors.

Definition 8.1 (Classical Regression Model)**.** Let Y_i represent the response for the i-th observation, and let $\mathbf{x}_i$ represent the vector of observed predictors for the i-th observation in a sample of n units, including the intercept term. Under the classical linear regression model, the distributional model for the response among the population is given by

$$Y_i \overset{\text{Ind}}{\sim} N\left(\mathbf{x}_i^\top \beta, \sigma^2\right).$$

> **Note**
>
> We have been assuming that the first element in $\mathbf{x}_i$ is a 1 to capture the intercept. It is possible to express a model without an intercept term in which case the first element of $\mathbf{x}_i$ is $x_{1,i}$.

Notice that Definition 8.1 simply extends the form of the distributional model we have previously considered. It makes it clear that

$$E\left(Y_i\right) = \mathbf{x}_i^\top \beta$$

for each unit. While this presentation connects the process for the inference of a single mean with that of regression, it is not the common presentation. Instead, the model is traditionally presented as saying that

$$Y_i = \mathbf{x}_i^\top \beta + \varepsilon_i$$

where $\varepsilon_i \overset{\text{IID}}{\sim} N\left(0, \sigma^2\right)$, where now we can make use of the "identically distributed" language. In this presentation, we have introduced a new random variable, ε. Since the expression $\mathbf{x}_i^\top \beta$ does not contain a random variable, it is deterministic in nature. Therefore, the distribution of Y_i is determined because we are simply shifting the distribution of ε_i. As with the simple linear regression model, we can relax the conditions we place on the distribution of ε_i.

Regardless of the conditions we impose on ε, we are essentially specifying a model for the data generating process — the set of statements we are willing to make regarding the variability in the response. We know that since the response is a random variable it has some distribution. The model for the data generating process is really a set of statements about that distribution. We may only be characterizing the mean of the distribution of the response; we may be willing to characterize the mean and the variance; or, we may be willing to fully characterize the distributional form.

Just as we did with the simple linear regression model, we can use the method of least squares to estimate the parameters in the model. Specifically, we choose the parameter vector β to minimize the quantity

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \beta\right)^2. \tag{8.1}$$

The corresponding estimates are denoted by the vector $\hat{\beta}$. Unlike the simple linear regression case, we now have $p+1$ parameters (assuming an intercept term) and therefore minimizing this quantity requires that we take $p+1$ partial derivatives and simultaneously solve the $p+1$ resulting equations. This is where the power of matrix algebra makes the computations simpler.

Let $\mathbf{Y}$ denote the vector of responses; that is,

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}.$$

And, let $\mathbf{X}$ denote the **design matrix**

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{2,1} & \cdots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \cdots & x_{p,2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{2,n} & \cdots & x_{p,n} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}.$$

> **Note**
>
> The first column being a column of 1's captures the intercept term; if there is no intercept in the model, this column is omitted.

Now, we can express the least squares objective function (Equation 8.1) as

$$(\mathbf{Y} - \mathbf{X}\beta)^\top (\mathbf{Y} - \mathbf{X}\beta). \tag{8.2}$$

> **Tip**
>
> Before proceeding, convince yourself that this algebra makes sense. A dot product of any vector with itself is simply a sum of squared terms ("sum of squares"), and the i-th element of each vector is simply the i-th component of the sum in Equation 8.1.

Least squares estimation is now about choosing the parameter vector β such that we minimize Equation 8.2. Taking a derivative (with respect to the *vector* β) results in choosing the parameter vector β such that

$$(\mathbf{X}^\top \mathbf{X})\, \beta = \mathbf{X}^\top Y, \tag{8.3}$$

which are known as the **normal equations**. Again, we have not changed the original problem; we are just expressing it in matrices; there are still $p+1$ equations with $p+1$ unknowns (assuming an intercept term). This leads to a compact expression for the least squares estimator:

$$\widehat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\, \mathbf{X}^\top Y. \tag{8.4}$$

> **Warning**
>
> In practice, the inverse of the matrix $\mathbf{X}^\top\mathbf{X}$ is never taken directly. There are much more efficient algorithms for computing the least squares estimates.

Theorem 8.1 (Least Squares Estimates)**.** *Let Y_i represent the response for the i-th observation, and let $\mathbf{x}_i$ represent the vector of observed predictors for the i-th observation in a sample of n units, including the intercept term. The least squares estimates for the multiple linear regression model relating the response and predictor are given by*

$$\hat{\beta} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y.$$

where $\mathbf{Y}$ is the vector of responses and

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{pmatrix}.$$

is known as the design matrix.

Proof. By definition, the least squares estimate of the parameter vector β is the vector that minimizes the quantity

$$Q(\beta) = \left(\mathbf{Y} - \mathbf{X}\beta\right)^\top \left(\mathbf{Y} - \mathbf{X}\beta\right).$$

Define $\hat{\beta}$ to be the vector

$$\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top Y.$$

Observe that

$$\begin{aligned} Q(\beta) &= \left(\mathbf{Y} - \mathbf{X}\beta\right)^\top \left(\mathbf{Y} - \mathbf{X}\beta\right) \\ &= \left(\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\beta\right)^\top \left(\mathbf{Y} - \mathbf{X}\hat{\beta} + \mathbf{X}\hat{\beta} - \mathbf{X}\beta\right) \end{aligned}$$

where the second line is obtained by adding and subtracting the term $\mathbf{X}\hat{\beta}$ from each vector. We then expand the terms and obtain

$$
\begin{aligned}
Q(\beta) = &\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^\top \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) \\
&+ \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^\top \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right) \\
&+ \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right)^\top \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) \\
&+ \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right)^\top \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right).
\end{aligned} \tag{8.5}
$$

Let's consider the second line of Equation 8.5; observe that

$$
\begin{aligned}
\left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^\top \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right) &= \mathbf{Y}^\top\mathbf{X}\hat{\beta} - \mathbf{Y}^\top\mathbf{X}\beta \\
&\quad - \hat{\beta}^\top\mathbf{X}^\top\mathbf{X}\hat{\beta} + \hat{\beta}^\top\mathbf{X}^\top\mathbf{X}\beta \\
&= \mathbf{Y}^\top\mathbf{X}\hat{\beta} - \mathbf{Y}^\top\mathbf{X}\beta \\
&\quad - \left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right]^\top\mathbf{X}^\top\mathbf{X}\hat{\beta} \\
&\quad + \left[\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{Y}\right]^\top\mathbf{X}^\top\mathbf{X}\beta \\
&= \mathbf{Y}^\top\mathbf{X}\hat{\beta} - \mathbf{Y}^\top\mathbf{X}\beta \\
&\quad - \mathbf{Y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\beta} \\
&\quad + \mathbf{Y}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{X}\beta \\
&= \mathbf{Y}^\top\mathbf{X}\hat{\beta} - \mathbf{Y}^\top\mathbf{X}\beta \\
&\quad - \mathbf{Y}^\top\mathbf{X}\hat{\beta} + \mathbf{Y}^\top\mathbf{X}\beta \\
&= 0.
\end{aligned}
$$

That is, the cross product terms cancel out. Further, since line 3 of Equation 8.5 is simply the transpose of line 2, we also have that line 3 is equivalent to 0. That is, we rewrite Equation 8.5 as

$$
\begin{aligned}
Q(\beta) &= \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right)^\top \left(\mathbf{Y} - \mathbf{X}\hat{\beta}\right) \\
&\quad + \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right)^\top \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right) \\
&= Q\left(\hat{\beta}\right) + \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right)^\top \left(\mathbf{X}\hat{\beta} - \mathbf{X}\beta\right).
\end{aligned} \tag{8.6}
$$

We note that both terms are sums of squares and therefore must be non-negative. Further, since only the second term is a function of β, minimizing $Q(\beta)$ is equivalent to minimizing the second term. Since the second term in Equation 8.6 is non-negative, we can minimize it by equating it to 0, which happens if and only if $\beta = \hat{\beta}$. This establishes the result. □

Example 8.1 (Least Squares Estimate in the Intercept Only Model)**.** Suppose we have only an intercept in the model, determine the least squares estimates of the intercept starting with the matrix representation.

Solution. While we have already determined the least squares estimates for the simple linear model in Chapter 5, and we could use that result to determine this estimate. However, we establish this as a special case of the multiple regression model as well. Specifically, note that the least squares estimates are given by

$$\hat{\beta} = (\mathbf{X}^t op\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}.$$

Now, we know that for an intercept-only model, the design matrix is given by

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Therefore, we have that

$$\mathbf{X}^\top\mathbf{X} = n$$

and

$$(\mathbf{X}^\top\mathbf{X})^{-1} = \frac{1}{n}.$$

We also have that

$$\mathbf{X}^\top\mathbf{Y} = \sum_{i=1}^{n} y_i = n\bar{y}.$$

Therefore, the least squares estimate is given by

$$\hat{\beta}_0 = \frac{1}{n}n\bar{y} = \bar{y}$$

the sample mean response.

Theorem 8.2 (Sampling Distribution for Least Squares Estimators)**.** *Under the conditions of the Classical Regression Model (Definition 8.1), we have that, holding all other predictors fixed*

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var\left(\hat{\beta}_j\right)}} \sim t_{n-p-1}$$

where $Var\left(\hat{\beta}_j\right)$ *is the* (j, j)*-th element of the matrix*

$$\hat{\sigma}^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$$

and

$$\hat{\sigma}^2 = \frac{1}{n-p-1} \sum_{i=1}^{n} \left(Y_i - \mathbf{x}_i^\top \hat{\beta}\right)^2$$

is our estimate of the unknown population variance σ^2*.*

Once we have a model for the sampling distribution, we have the ability to make inference — confidence intervals or p-values.

> **!** Important
>
> While subtle, the phrase "holding all other predictors fixed" is critical in the sampling distribution of the least squares estimates.

9 Hierarchical Models

We have primarily addressed models where the only random variable was the response. All terms in the model were either fixed predictors or fixed (albeit unknown) parameters. In this model, we consider a hierarchical models.

9.1 Motivating Example

Consider obtaining the weight of an individual. We might randomly select the individual from a population; then, we place them on a scale, which is subject to measurement error. Since the scale is subject to measurement error, we do not observe their actual weight; instead, we see some jittered version of their weight. So, we might want to take multiple weight readings. As a result, the j-th weight we observe Y_j on the subject is the sum of the individual's true weight θ and the measurement error ε_j in the scale on the j-th reading:

$$Y_j = \theta + \varepsilon_j.$$

This looks a lot like a linear model with only an intercept. However, the primary difference is that this is for multiple observations from a single unit! We would of course not only measure the weight of one person from the population but a sample of size n from the population. Therefore, our model would become

$$Y_{i,j} = \theta_i + \varepsilon_{i,j} \tag{9.1}$$

where

- $Y_{i,j}$ represents the j-th weight measurement taken on the i-th person (remember, we record each person's weight multiple times),
- θ_i represents the true weight for person i, and
- $\varepsilon_{i,j}$ represents the measurement error in the j-th weight measurement taken on the i-th person.

And, since the measurement error is random, we might posit a model for its distribution. For example, we might say that

$$\varepsilon_{i,j} \overset{\text{IID}}{\sim} N\left(0, \sigma^2\right).$$

Notice that unlike our previous models, the unknown terms $\theta_1, \theta_2, \ldots, \theta_n$ change with each person. More, since the individuals are from some underlying population, we might think about their weights as coming from some underlying distribution. That is, we might posit that

$$\theta_i \overset{\text{IID}}{\sim} N\left(\mu, \eta^2\right)$$

for some unknown parameters μ and η^2. Here, μ represents the average weight of individuals in the population.

We now have two distributional models; putting this together, we can write

$$\begin{aligned} Y_{i,j} \mid \theta_j &\overset{\text{Ind}}{\sim} N\left(\theta_j, \sigma^2\right) \\ \theta_j &\overset{\text{IID}}{\sim} N\left(\mu, \eta^2\right). \end{aligned}$$

That is, the distributional model for the observed weight is *conditional* on knowing the true weight of the individual, and this true weight has its own distribution. This is an example of a **hierarchical model**.

Definition 9.1 (Hierarchical Model)**.** A distributional model for a response that is constructed in layers. The top-most layer is conditional on components that are expressed in lower layers.

Hierarchical models rely on the notion of conditional probability.

9.2 Quick Review of Conditional Probability

Definition 9.2 (Conditional Probability)**.** Let A and B be events with non-zero probability. Then, the probability that event A occurs given that event B occurs (written $A \mid B$) is given by

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

The notion of conditional probability constrains the sample space to consider only that region where the event B does occur. We are essentially looking for the proportional area that A takes up within B. We generalize this definition to random variables as well. In order to do so, we need the notion of a joint density function.

Definition 9.3 (Joint Density)**.** Let X and Y be continuous random variables. Then, the joint density function of X and Y, written $f_{X,Y}(x, y)$ is the function such that

$$Pr(a < X < b, c < Y < d) = \int_c^d \int_a^b f_{X,Y}(x, y)dxdy.$$

While we have defined a joint density function for two continuous random variables, we could create an analogous definition for two discrete random variables or a mix of the two. We can now define a conditional density.

Definition 9.4 (Conditional Density)**.** Let X and Y be continuous random variables. Then, the density function of X given the value of Y, written as $f_{X|Y}(x \mid y)$ is given by

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

where $f_{X,Y}(x, y)$ is the joint density function.

This is always a challenging part of a course in probability theory. Here, we are saying that for each value the random variable Y takes, a new *distribution* for X is created. Knowing information about one variable informed the distribution of the other.

Just as we talk about how a random variable "varies" (the variability of the distribution), we can talk about how two random variables "co-vary."

Definition 9.5 (Covariance)**.** Let X and Y be random variables. The covariance of X and Y, written $Cov(X, Y)$ is defined as

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

9.3 Properties for the Simple Hierarchical Model

Consider the simple hierarchical model from our motivating example:

$$Y_{i,j} = \theta_i + \varepsilon_{i,j}.$$

Since the response is a sum of random variables, we trivially have that

$$E\left(Y_{i,j}\right) = E\left(\theta_i\right) + E\left(\varepsilon_{i,j}\right) = \mu.$$

So, the mean response across all observed observations is the average weight in the population. Since we assume that the θ_i terms are independent of the $\varepsilon_{i,j}$ terms, we also readily obtain that

$$Var\left(Y_{i,j}\right) = Var\left(\theta_i\right) + Var\left(\varepsilon_{i,j}\right) = \eta^2 + \sigma^2.$$

That is, the variability in the response is the result of the variability across individuals as well as the variability due to measurement error. And, if the measurement error did not exist ($\sigma^2 = 0$), then we would have that $Y_{i,1} = Y_{i,2} = \cdots = Y_{i,m}$.

Example 9.1 (Covariance in Hierarchical Model)**.** Consider the simple hierarchical model

$$Y_{i,j} = \theta_i + \varepsilon_{i,j}$$

where $\theta_i \overset{\text{IID}}{\sim} N\left(\mu, \eta^2\right)$, $\varepsilon_{i,j} \overset{\text{IID}}{\sim} N\left(0, \sigma^2\right)$, and θ_i independent of $\varepsilon_{i,j}$ for all i and j. Determine the covariance between two observations made on the same unit, and determine the covariance between two observations made on different units.

First, consider two observations made on different units. Without loss of generality, we consider $Y_{1,j}$ and $Y_{2,j}$. We are interested in finding the covariance between $Y_{1,j}$ and $Y_{2,j}$. By definition

$$Cov\left(Y_{1,j}, Y_{2,j}\right) = E\left(Y_{1,j}Y_{2,j}\right) - E\left(Y_{1,j}\right)E\left(Y_{2,j}\right).$$

We have already established the mean of each observation, μ. Plugging in the form of the linear model for $Y_{1,j}$ and $Y_{2,j}$ and expanding, we have that

$$\begin{aligned} Cov\left(Y_{1,j}, Y_{2,j}\right) &= E\left(Y_{1,j}Y_{2,j}\right) - E\left(Y_{1,j}\right)E\left(Y_{2,j}\right) \\ &= E\left(\theta_1\theta_2 + \theta_1\varepsilon_{2,j} + \theta_2\varepsilon_{1,j} + \varepsilon_{1,j}\varepsilon_{2,j}\right) - \mu^2 \\ &= E\left(\theta_1\right)E\left(\theta_2\right) + E\left(\theta_1\right)E\left(\varepsilon_{2,j}\right) + E\left(\theta_2\right)E\left(\varepsilon_{1,j}\right) + E\left(\varepsilon_{1,j}\right)E\left(\varepsilon_{2,j}\right) - \mu^2 \end{aligned}$$

where the expectations separate due to the independence between the θ and ε terms. Now, plugging in, we have that

$$Cov\left(Y_{1,j}, Y_{2,j}\right) = \mu^2 + 0 + 0 + 0 - \mu^2 = 0.$$

We now consider two observations from the same unit; without loss of generality, we consider $Y_{i,1}$ and $Y_{i,2}$ (the same i ensures they are from the same subject). Following a similar process as before, we have

$$\begin{aligned} Cov\left(Y_{i,1}, Y_{i,2}\right) &= E\left(Y_{i,1}Y_{i,2}\right) - E\left(Y_{i,1}\right)E\left(Y_{i,2}\right) \\ &= E\left(\theta_i\theta_i + \theta_i\varepsilon_{i,1} + \theta_i\varepsilon_{i,2} + \varepsilon_{i,1}\varepsilon_{i,2}\right) - \mu^2 \\ &= E\left(\theta_i^2\right) + E\left(\theta_i\right)E\left(\varepsilon_{i,1}\right) + E\left(\theta_i\right)E\left(\varepsilon_{i,2}\right) + E\left(\varepsilon_{i,1}\right)E\left(\varepsilon_{i,2}\right) - \mu^2 \\ &= E\left(\theta_i^2\right) - \mu^2 \end{aligned}$$

where the expectations separate due to the independence. However, we still have one term left to resolve. Remembering the definition of variance, we have that

$$Cov\left(Y_{i,1}, Y_{i,2}\right) = Var\left(\theta_i\right) + E^2\left(\theta_i\right) - \mu^2 = \eta^2 + \mu^2 - \mu^2 = \eta^2.$$

That is, observations from the same subject are correlated with one another.

9.4 Repeated Measures as a Hierarchical Model

The repeated measures ANOVA model views the mean response as a function of a factor of interest and a block term. Specifically, we write

$$Y_i = \sum_{j=1}^{k} \mu_j(\text{Group } j)_i + \sum_{b=1}^{m} \beta_b(\text{Block } b)_i + \varepsilon_i$$

where

$$(\text{Group } j)_i = \begin{cases} 1 & \text{if i-th observation comes from group j} \\ 0 & \text{otherwise} \end{cases}$$

$$(\text{Block } b)_i = \begin{cases} 1 & \text{if i-th observation comes from block b} \\ 0 & \text{otherwise} \end{cases}$$

are indicator variables. It is common to assume the block effects, $\beta_1, \beta_2, \ldots, \beta_b$ are random variables that are independent of the treatment groups and the error in the response ε_i; further, it is common to assume that

$$\beta_j \overset{\text{IID}}{\sim} N\left(0, \sigma_b^2\right).$$

That is, the repeated measures ANOVA model is a hierarchical model. The distribution of the response can only be specified if we know the block effect, and the block effect itself is a random variable.

Viewing it as a hierarchical model allows us to see where the correlation structure comes from. All observations that share a similar value of β_b are associated in some way.

10 Autocorrelation

A critical assumption throughout the text has been that of independence. For example, we might assume that

$$Y_i = \mu + \varepsilon_i$$

where the ε_i are independent random variables. For a more concrete example, suppose we say that $\mu = 0$ and $\varepsilon \sim N(0, 1)$, again, all independent. And, suppose we take a sample of size 100. Figure 10.1 illustrates several possible samples, where the observations are made sequentially. That is, Figure 10.1 shows several time-series plots (plot of the observations over time).

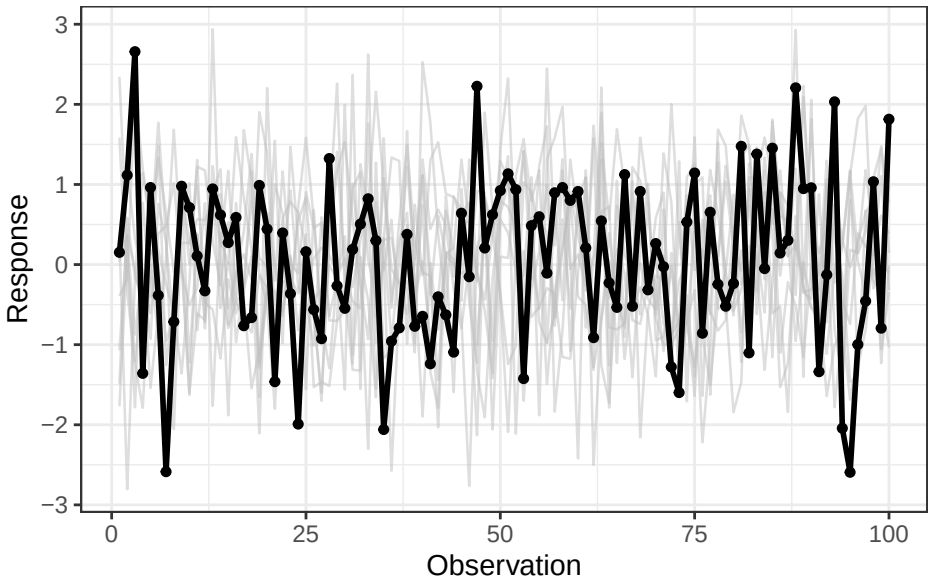

Figure 10.1: The time-series plot for several random samples of size 100 taken from a Standard Normal distribution. One sample is highlighted for illustration. No visible patterns are present.

There are no visible trends in the location or spread of the responses as we move across the graphic. While any one time-series plot *may* show some small trend, overall, across repeated

samples, no trend is observed. This lack of trend is the result of the independence between observations. Any one observation does not help us predict the location of the next.

In contrast, suppose we continue to take $\mu = 0$, but we say that $\varepsilon_1 \sim N(0, 1)$ and

$$\varepsilon_i \mid \varepsilon_{i-1} \sim N\left(\varepsilon_{i-1}, 1\right)$$

for $i = 2, 3, \ldots, n$. Notice that the distribution of one error term depends on the value of the previous error. We again consider taking a sample of size 100. Figure 10.2 shows several time-series plots for this updated situation.

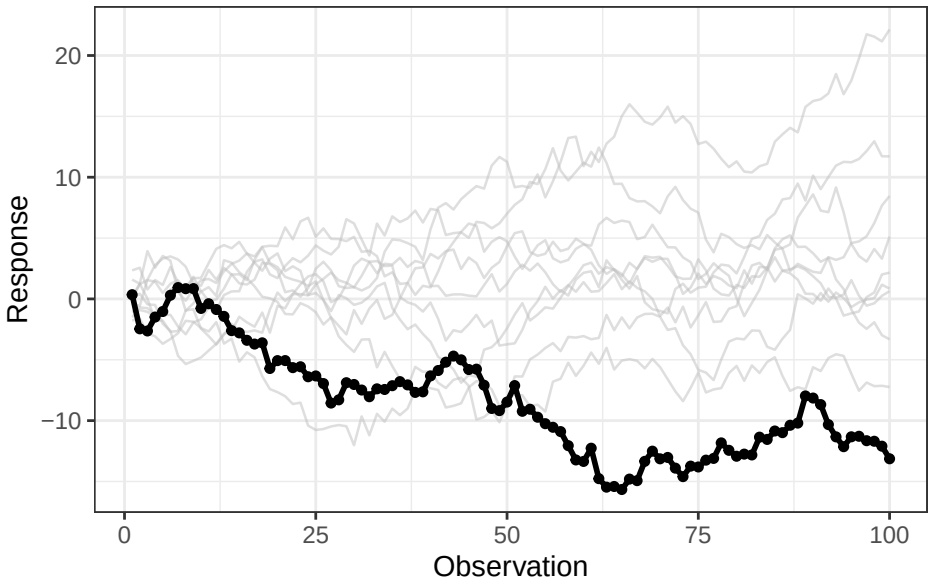

Figure 10.2: The time-series plot for several samples of size 100 taken sequentially. The first observation is taken from a Standard Normal distribution; each subsequent observation is taken from a Normal distribution centered on the previous observation with a variance of 1. One sample is highlighted for illustration.

Unlike Figure 10.1, we notice a clear trend in the location of the series highlighted in Figure 10.2. The location of the response tends to decrease over time. The lack of independence is revealing itself in a trend in the location of the response over time.

This is known as auto-correlation. While the specifics of this phenomena would be studied in a course on regression modeling, what we see is that the relationship between one observation and the next is time-dependent. That is, observations close together in time (indices that are near one another) are related to one another, while observations further apart in time are less related to one another.

Example 10.1 (Covariance in Autocorrelation Model)**.** Consider the simple autocorrelation model

$$Y_i = \varepsilon_i$$

where $\varepsilon_1 \sim N(0,1)$ and

$$\varepsilon_i \mid \varepsilon_{i-1} \sim N\left(\varepsilon_{i-1}, 1\right)$$

for $i > 1$. Determine the covariance between two observations.

Without loss of generality, consider Y_1 and Y_2. Observe that

$$Cov\left(Y_1, Y_2\right) = Cov\left(\varepsilon_1, \varepsilon_2\right) = E\left(\varepsilon_1 \varepsilon_2\right) - E\left(\varepsilon_1\right) E\left(\varepsilon_2\right).$$

In order to evaluate these expectations, we need an interim result: for any two random variables X and Y, we have that $E(X) = E(E(X \mid Y))$. That is, in the inner expectation, we take the conditional expectation of X given Y; the result will be only a function of the random variable Y. In the outer expectation, we take the expectation with respect to Y.

Returning to our problem at hand, we know that $E\left(\varepsilon_1\right) = 0$. Applying our latest result, we have

$$\begin{aligned} E\left(\varepsilon_2\right) &= E\left[E\left(\varepsilon_2 \mid \varepsilon_1\right)\right] \\ &= E\left[\varepsilon_1\right] \\ &= 0. \end{aligned}$$

We also have that

$$\begin{aligned} E\left(\varepsilon_1 \varepsilon_2\right) &= E\left[E\left(\varepsilon_1 \varepsilon_2 \mid \varepsilon_1\right)\right] \\ &= E\left[\varepsilon_1 E\left(\varepsilon_2 \mid \varepsilon_1\right)\right] \\ &= E\left[\varepsilon_1^2\right] \\ &= Var\left(\varepsilon_1\right) + E^2\left(\varepsilon_1\right) \\ &= 1. \end{aligned}$$

Line 2 comes from recognizing that if we are "given" ε_1, then it is constant in terms of the inner expectation and comes out of the expectation. Now, we have shown that

$$Cov\left(Y_1, Y_2\right) = 1$$

meaning that there is a correlation between terms that are next to one another.

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.

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 $\{A_1, A_2, \dots\}$ 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."

Bernoulli Distribution (Definition 2.11) Let X be a discrete random variable taking the value 0 or 1. X is said to have a Bernoulli distribution with density

$$f(x) = \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1\},$$

where $0 < \theta < 1$ is the probability that X takes the value 1.

- $E(X) = \theta$
- $Var(X) = \theta(1-\theta)$

We write $X \sim Ber(\theta)$, which is read "X follows a Bernoulli distribution with probability θ."

Binomial Distribution (Definition 2.12) Let X be a discrete random variable taking integer values between 0 and n, inclusive. X is said to have a Binomial distribution with density

$$f(x) = \binom{n}{x} \theta^x (1-\theta)^{1-x} \qquad x \in \{0, 1, \dots, n\},$$

where $0 < \theta < 1$ is the probability of a success on an individual trial.

- $E(X) = n\theta$
- $Var(X) = n\theta(1-\theta)$

We write $X \sim Bin(n, \theta)$, which is read "X follows a Binomial distribution with parameters n and θ."

Case-Resampling Bootstrap (Definition 3.8) Let $Y_1, Y_2, \ldots, Y_n$ be a random sample from an underlying population, and let θ represent a parameter of interest characterizing the underlying population. Further, define $\widehat{\theta} = h(\mathbf{Y})$ be a statistic which estimates the parameter. The case-resampling bootstrap algorithm proceeds as follows:

1. Take a random sample, with replacement, from the set $\{Y_1, Y_2, \ldots, Y_n\}$ of size n. Call these values $Y_1^*, Y_2^*, \ldots, Y_n^*$. This is known as a bootstrap resample.
2. Compute $\widehat{\theta}^* = h(\mathbf{Y}^*)$ and store this value. This is known as a bootstrap statistic.
3. Repeat steps 1-2 m times, for some large value of m (say $m = 5000$). Denote $\widehat{\theta}_j^*$ to be the bootstrap statistic from the j-th bootstrap resample.

The empirical distribution of $\widehat{\theta}_1^*, \widehat{\theta}_2^*, \ldots, \widehat{\theta}_m^*$ will approximate the shape and spread of the sampling distribution of the statistic $h(\mathbf{Y})$.

Chi-Square Distribution (Definition 2.16) Let X be a continuous random variable. X is said to have a Chi-Square distribution if the density is given by

$$f(x) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, x^{\nu/2-1} e^{-x/2} \qquad x > 0,$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim \chi^2_\nu$, which is read "X follows a Chi-Square distribution with ν degrees of freedom." The Chi-Square distribution is a special case of the Gamma distribution where $\alpha = \nu/2$ and $\beta = 2$.

Classical Regression Model (Definition 8.1) Let Y_i represent the response for the i-th observation, and let $\mathbf{x}_i$ represent the vector of observed predictors for the i-th observation in a sample of n units, including the intercept term. Under the classical linear regression model, the distributional model for the response among the population is given by

$$Y_i \overset{\text{Ind}}{\sim} N\left(\mathbf{x}_i^\top \beta, \sigma^2\right).$$

Classical Simple Linear Regression (Definition 5.2) Let $(Y_1, x_1), (Y_2, x_2), \ldots, (Y_n, x_n)$ be observations made on a sample of n units. Under the classical simple linear regression model, the distributional model for the response among the population is given by

$$Y_1, Y_2, \ldots, Y_n \overset{\text{Ind}}{\sim} N\left(\beta_0 + \beta_1 x_i, \sigma^2\right)$$

where β_0, β_1, and σ^2 are unknown parameters.

Conditional Density (Definition 9.4) Let X and Y be continuous random variables. Then, the density function of X given the value of Y, written as $f_{X|Y}(x \mid y)$ is given by

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

where $f_{X,Y}(x, y)$ is the joint density function.

Conditional Probability (Definition 9.2) Let A and B be events with non-zero probability. Then, the probability that event A occurs given that event B occurs (written $A \mid B$) is given by

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

Confidence Interval (Definition 4.1) Consider repeatedly taking samples $\mathbf{Y}$ of size n from a population characterized by the parameter θ. The interval $(h_1(\mathbf{Y}), h_2(\mathbf{Y}))$ is said to be a $100c\%$ confidence interval if

$$Pr\left(h_1(\mathbf{Y}) \leq \theta \leq h_2(\mathbf{Y})\right) = c.$$

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).

Covariance (Definition 9.5) Let X and Y be random variables. The covariance of X and Y, written $Cov(X, Y)$ is defined as

$$Cov(X, Y) = E\left[(X - E(X))(Y - E(Y))\right] = E(XY) - E(X)E(Y).$$

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.

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

Expectation of a Function (Definition 2.8) 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.

Expectation of a Product of Independent Random Variables (Definition 3.4) Let $X_1, X_2, \ldots, X_n$ be independent random variables, then

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E\left(X_i\right).$$

Expected Value (Mean) (Definition 2.6) 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.

F-Distribution (Definition 2.17) Let X be a continuous random variable. X is said to have an F-distribution if the density is given by

$$f(x) = \frac{\Gamma((r+s)/2)}{(\Gamma(r/2)\Gamma(s/2))}(r/s)^{(r/2)}x^{(r/2-1)}(1+(r/s)x)^{-(r+s)/2} \qquad x > 0,$$

where $r, s > 0$ are the numerator and denominator degrees of freedom, respectively.

We write $X \sim F_{r,s}$, which is read "X has an F-distribution with r numerator degrees of freedom and s denominator degrees of freedom."

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,

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

Gamma Distribution (Definition 2.14) Let X be a continuous random variable. X is said to have a Gamma distribution if the density is given by

$$f(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \qquad x > 0,$$

where $\alpha > 0$ is the shape parameter and $\beta > 0$ is the scale parameter.

- $E(X) = \alpha\beta$
- $Var(X) = \alpha\beta^2$

We write $X \sim Gamma\,(\alpha, \beta)$, which is read "X follows a Gamma distribution with shape α and scale β." This short-hand implies the density above. When $\alpha = 1$, we refer to this as the Exponential distribution with scale β.

We note that, in general, there is no closed form solution for $\Gamma(\alpha)$, but

- $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$
- $\Gamma(k) = (k-1)!$ for non-negative integer k

Hierarchical Model (Definition 9.1) A distributional model for a response that is constructed in layers. The top-most layer is conditional on components that are expressed in lower layers.

Independence (Definition 3.2) 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.

Joint Density (Definition 9.3) Let X and Y be continuous random variables. Then, the joint density function of X and Y, written $f_{X,Y}(x, y)$ is the function such that

$$Pr(a < X < b, c < Y < d) = \int_c^d \int_a^b f_{X,Y}(x, y) dx dy.$$

Location-Scale Family (Definition 7.1) For $a \in \mathbb{R}$ and $b > 0$, let $X = a + bZ$ for some random variable Z. X is said to be the location-scale family associated with the distribution of Z with location parameter a and scale parameter b.

Method of Distribution Functions (Definition 2.18) 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 $g(y) = \frac{\partial}{\partial y} G(y)$.

Method of Least Squares (Definition 5.3) The least squares estimates of the parameters β_0 and β_1 in a simple linear regression model are the values that minimize the quantity

$$\sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 x_i\right)^2.$$

The estimates are often denoted $\hat{\beta}_0$ and $\hat{\beta}_1$.

Moment-Generating Function (MGF) (Definition 3.6) For a random variable X, let $M_X(t)$ be defined as

$$M_X(t) = E\left(e^{tX}\right).$$

If $M_X(t)$ is defined for all values of t in some interval about 0, then $M_X(t)$ is called the moment-generating function (MGF) of X.

Normal (Gaussian) Distribution (Definition 2.13) Let X be a continuous random variable. X is said to have a Normal (or Guassian) distribution if the density is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad -\infty < x < \infty,$$

where μ is any real number and $\sigma^2 > 0$.

- $E(X) = \mu$
- $Var(X) = \sigma^2$

We write $X \sim N\left(\mu, \sigma^2\right)$, which is read "X follows a Normal distribution with mean μ and variance σ^2." This short-hand implies the density above. When $\mu = 0$ and $\sigma^2 = 1$, this is referred to as the Standard Normal distribution.

Null Distribution (Definition 4.2) Distribution of a statistic under a hypothesized value of the population parameter(s).

Order Statistic (Definition 7.2) Let $X_1, X_2, \ldots, X_n$ be a random sample of size n. The j-th order statistic, denoted $X_{(j)}$ is the j-th smallest value in the sample. Special cases include

- $X_{(1)}$, the sample minimum, and
- $X_{(n)}$, the sample maximum.

P-value (Definition 4.3) The probability, assuming the null hypothesis is true, that we would observe a statistic, by chance alone, as extreme or more so than that observed in the sample.

Parameter (Definition 2.5) Numeric quantity which summarizes the distribution of a variable within the *population* of interest. Generally denoted by Greek letters in statistical formulas.

Percentile (Definition 2.9) Let X be a random variable with density function f. The $100k$ percentile is the value q such that

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

Random Sample (Definition 3.3) 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.

We say $X_1, X_2, \ldots, X_n$ are independent and identically distributed, abbreviated IID. We might also write this as $X_i \overset{\text{IID}}{\sim} f$ for some density f.

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.

Regression (Definition 5.1) Allowing the parameters characterizing the distribution of a random variable to depend, through some specified function, on the value of additional variables.

Sample Quantile (Definition 7.3) Let $x_{(j)}$ denote the j-th observed order statistic in a sample. Then, the q-th quantile of the sample is given by $x_{(j)}$ for $\frac{j-1}{n} < q \leq \frac{j}{n}$ and $x_{(1)}$ if $q = 0$, for $j = 1, 2, \ldots, n$.

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

Sampling Distribution (Definition 3.5) The distribution of a statistic across repeated samples.

Statistic (Definition 3.1) A statistic is a numerical summary of a sample; it is a function of the data alone. Prior to collecting data, a statistic is a function of the data to be collected.

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.

Unbiased (Definition 3.7) An estimator (statistic) $\hat{\theta}$ is said to be unbiased for the parameter θ if

$$E\left(\hat{\theta}\right) = \theta.$$

Variance (Definition 2.7) 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.

t-Distribution (Definition 2.15) Let X be a continuous random variable. X is said to have a (standardized) t-distribution, sometimes called the Student's t-distribution, if the density is given by

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad -\infty < x < \infty$$

where $\nu > 0$ is the degrees of freedom.

We write $X \sim t_\nu$, which is read "X follows a t-distribution with ν degrees of freedom."