6. PROBABILITY

R symbols, operators, and functions introduced in this chapter: c(), sample(), rnorm(), pnorm(), for(i in 1:n){}, and print().

In the last four chapters, we have analyzed data for the purpose of (i) estimating causal effects with both randomized experiments and observational data, (ii) inferring population characteristics via survey research, and (iii) making predictions. Thus far, we have focused our attention on identifying systematic relationships and ignored the noise in the data. Real-world data, however, contain a nontrivial amount of noise, or irrelevant variation, which adds uncertainty to our conclusions. In the next chapter, we will learn how to quantify the degree of statistical uncertainty in our empirical findings. First, though, we need to learn about probability and how we can use it to model variation.

6.1 WHAT IS PROBABILITY?

There are two different ways of interpreting probability: frequentist and Bayesian.

According to the **frequentist** interpretation, probabilities represent proportions of specific events occurring over infinitely many identical trials.

According to the **frequentist** interpretation, probabilities represent proportions of specific events occurring over a large number of identical trials. Specifically, the probability of an event is the proportion of its occurrence among infinitely many identical trials.

RECALL: We define the proportion of observations that meet a criterion as:

$$\frac{\text{number of observations that meet criterion}}{\text{total number of observations}}$$

To interpret a proportion as a percentage, we multiply the decimal value by 100.

Think of a coin flip, for example. What is the probability of getting heads when flipping a coin? Imagine flipping the coin a large number of times. The probability of getting a head can be approximated by the number of heads realized over the number of coin flips. If the coin is fair, as we increase the number of flips, the proportion of heads should approach 0.5, meaning that about 50% of the coin flips should result in heads.

According to the **Bayesian** interpretation, probabilities represent personal, subjective beliefs about the relative likelihood of events.

In contrast, according to the **Bayesian** interpretation, probabilities represent one's subjective beliefs about the relative likelihood of events. For example, when we state that the probability of rain today is about 80%, we are not describing the frequency of rain events over multiple days. We are simply describing how certain we are about the event occurring. A probability of 1, or 100%, indicates certainty that the event will occur. A probability of 0, or 0%, indicates certainty that the event will not occur.

Critics of the frequentist interpretation argue that it is impossible to conduct an infinite number of identical trials. (For example, when flipping a coin, it would be difficult to maintain the same launch angle and speed.) Critics of the Bayesian interpretation argue that personal, subjective beliefs should not play a role when analyzing data. Fortunately, despite their differences, the two interpretations rely on the same mathematical rules. For the rest of the chapter, we focus on these common rules.

6.2 AXIOMS OF PROBABILITY

Probability axioms are the basic rules upon which the entire probability theory rests. Before we learn about the axioms of probability, we need to define some concepts:

A **trial** is an action or set of actions that produces outcomes of interest. An **outcome** is the result of a trial. An **event** is a set of outcomes. **Mutually exclusive events** are events that do not share any outcomes. The **sample space** is the set of all possible outcomes produced by a trial. An **event** is said to **occur** if any one of the possible outcomes included in the event is realized.

- A **trial** is an action or set of actions that produces outcomes of interest. For example, rolling a die can be considered a trial.
- An **outcome** is the result of a trial. Rolling a die produces one of six possible outcomes: 1, 2, 3, 4, 5, or 6.
- An **event** is a set of outcomes. In the example at hand, one possible event is *rolling a number less than 3*, which includes two possible outcomes, 1 and 2. Note that events may include any number of outcomes. For example, another possible event is *rolling a 3*, which includes only one outcome, 3.
- **Mutually exclusive events** are events that do not share any outcomes. The two events defined above, *rolling a number less than 3* and *rolling a 3*, for example, are mutually exclusive events since they have no outcomes in common.
- The **sample space**, denoted as Ω (the Greek letter Omega), is the set of all possible outcomes produced by a trial. Since it is a set of outcomes, the sample space is also considered an event. In the case of rolling a die, $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- An **event** is said to **occur** if any one of the possible outcomes included in the event is realized. For example, if we roll a 1, we would consider that the event *rolling a number less than 3* has occurred and so has the event defined by the sample space.

There are three axioms of probability. Remarkably, we can derive the entire probability theory from these three basic rules.

1. The first axiom states that the probability of any event A is non-negative. In mathematical notation, we can write this axiom as:

$$P(A) \geq 0$$

where P stands for "the probability of" and A represents the event.

This means that probabilities can be either zero or positive. For example, the probability of *rolling a 3* cannot be negative.

2. The second axiom states that the probability of the sample space is always 1. In mathematical notation:

$$P(\Omega) = 1$$

where Ω represents the sample space, that is, the set of all possible outcomes produced by a trial.

For example, when rolling a die, the sample space, Ω, is {1, 2, 3, 4, 5, 6}. Recall that an event is said to occur if any one of the possible outcomes included in the event occurs. In this case, $P(\Omega)$ represents the probability that any one of the six possible outcomes occurs. In mathematical notation:

$$P(\Omega) = (1 \text{ or } 2 \text{ or } 3 \text{ or } 4 \text{ or } 5 \text{ or } 6) = 1$$

3. The third axiom states that, if events A and B are mutually exclusive (that is, they cannot occur at the same time), then the probability that either A or B occurs equals the probability that A occurs plus the probability that B occurs. In mathematical notation:

$$P(A \text{ or } B) = P(A) + P(B)$$
if A and B are mutually exclusive events

For example, the probability of either *rolling a number less than 3* or *rolling a 3* equals the probability of *rolling a number less than 3* plus the probability of *rolling a 3*, since the two events are mutually exclusive.

These axioms together imply that probabilities range from 0 to 1 and that the probabilities of all possible outcomes produced by a trial must add up to 1.

Consider flipping a coin once. This trial can result in two possible outcomes: *getting a head* or *getting a tail*. The sample space in this case is, then: $\Omega = \{head, tail\}$.

The probability of *getting a head* and the probability of *getting a tail* must both be between 0 and 1, and together they must add up to 1 since they constitute the sample space (that is, no other outcome is possible).

Let's see how we arrive at this conclusion using mathematical notation. We can start with axiom 2:

$$P(head \text{ or } tail) = 1$$

Since *getting a head* and *getting a tail* are two mutually exclusive events, according to axiom 3:

$$P(head \text{ or } tail) = P(head) + P(tail) = 1$$

Therefore, the probabilities of the two possible outcomes produced by flipping a coin must add up to 1.

6.3 EVENTS, RANDOM VARIABLES, AND PROBABILITY DISTRIBUTIONS

We can categorize most things that occur in our lives as **events**. The fact that you are reading this book is an event, and so is your height, the color of your eyes, your political party preference, and your choice to attend or not to attend college.

An **event** is a set of outcomes that occur with a particular probability. A **random variable** assigns a numeric value to each mutually exclusive event produced by a trial. The **probability distribution** of a random variable characterizes the likelihood of each possible value the random variable can take.

As soon as we assign a number to an event, we create what is known as a **random variable**. A random variable assigns a numeric value to each mutually exclusive event produced by a trial. In fact, we have been dealing with random variables throughout this book. We have just been calling them variables.

For example, if we assign a 1 to the event of attending college and a 0 to the event of not attending college, then the binary random variable *college*, as defined below, would capture these events.

POSSIBLE EVENTS PRODUCED BY A TRIAL	RANDOM VARIABLE *college*	PROBABILITY DISTRIBUTION OF *college*
- attending college	- $college_i$=1 if individual i attends college	- $P(college=1)$
- not attending college	- $college_i$=0 if individual i does not attend college	- $P(college=0)$

Each random variable has a **probability distribution**, which characterizes the likelihood of each value the variable can take. By definition, all probabilities in a distribution must add up to 1.

In mathematical notation, we can write the probability that the random variable X takes the value x as:

$$P(X=x) = p$$

TIP: We use uppercase letters to refer to random variables (such as X, Y, and Z) and lowercase letters to refer to fixed values the random variables may take (such as x, y, and z).

where:

- X is the random variable
- x is the fixed value the random variable X may take
- p is the probability that X takes the value x.

For example, the distribution of the random variable *college* above represents the probability of attending college, $P(college{=}1)$, and the probability of not attending college, $P(college{=}0)$, since those are the only two possible values the variable can take.

6.4 PROBABILITY DISTRIBUTIONS

In this book, we focus on two types of probability distributions: (i) the Bernoulli distribution, which is the probability distribution of a binary variable; and (ii) the normal distribution, which is the probability distribution we commonly use as a good approximation for many non-binary variables. Within the normal distribution, we will pay special attention to the standard normal distribution.

As we saw in chapter 3, functions such as mean, median, standard deviation, and variance can be used to summarize numerically the main traits of the probability distribution of a random variable. In this section, we focus on (i) the center of the distributions as measured by the mean, and (ii) the spread of the distributions as measured by the variance (which, as you may recall, is equivalent to the standard deviation squared).

6.4.1 THE BERNOULLI DISTRIBUTION

The **Bernoulli distribution** is the probability distribution of a binary variable. Since binary variables can take only two values (1 or 0), the Bernoulli distribution characterizes two probabilities: the probability that the variable equals 1 and the probability that the variable equals 0.

The **Bernoulli distribution** is the probability distribution of a binary variable. It is characterized by one parameter, p, which is the probability that the binary random variable takes the value of 1. Consequently, $1-p$ is the probability that the binary random variable takes the value of 0. The mean of a Bernoulli distribution is p and the variance is $p(1-p)$.

By definition, the sum of all probabilities in a Bernoulli distribution must equal 1. If we use p to denote the probability that the binary variable equals 1, the probability that the binary variable equals 0 is $1-p$ (notice that $p+(1-p)=1$).

Consider again the flip of a coin. The action of flipping a coin can result in only one of two possible events: heads or tails. If we assign 1 to heads and 0 to tails, we can create a binary random variable with the results. The distribution of this random variable—a Bernoulli distribution—represents the probability that we get heads as well as the probability that we get tails. The definition of this binary random variable and its distribution are:

$$flip_i = \begin{cases} 1 \text{ if coin flip } i \text{ lands on heads;} & P(flip{=}1) = p \\ 0 \text{ if coin flip } i \text{ lands on tails;} & P(flip{=}0) = 1-p \end{cases}$$

The mean of a Bernoulli distribution is equal to p, that is, the probability that the binary variable equals 1, and the variance of a Bernoulli distribution is equal to $p(1-p)$. (We will see examples shortly.)

RECALL: As we saw in chapter 1, the mean of a binary variable is equivalent to the proportion of the observations that have the characteristic identified by the variable. In other words, the mean of a binary variable is the probability that the variable equals 1, denoted as p.

To approximate the value of p, we could flip the coin many times and calculate the proportion of heads among the multiple flips. For illustration purposes, see the example below, where we hypothetically flip a coin 12 times and calculate the corresponding proportions.

REALIZED EVENTS

REALIZED VALUES OF A RANDOM VARIABLE

$$flip = \{1, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0\}$$

APPROXIMATE PROBABILITY DISTRIBUTION

$$P(flip{=}1) \approx \frac{\text{number of heads}}{\text{number of flips}} \approx 8/12 = 0.67$$

$$P(flip{=}0) \approx \frac{\text{number of tails}}{\text{number of flips}} \approx 4/12 = 0.33$$

In the example above, the proportion of heads is 67%, and the proportion of tails 33%. These proportions are far from 50% because we flipped the coin only 12 times. As we increase the number of flips, if the coin is fair, the proportion of heads, p, and the proportion of tails, $1-p$, should approach 50%.

TIP: In mathematical notation, the symbol $\approx$ stands for "approximately equal to."

To get a better sense of this, we can use R to simulate flipping a fair coin 1 million times and then calculate the proportions of heads and tails. We start by listing the two possible values that might result from each coin flip using the function c(), which stands for "combine values into a vector." The following code creates an object named *possible_values* with a vector containing a 1 and a 0, where 1 stands for heads and 0 for tails:

c() combines values into a vector (a collection of elements, each identified by an index). The values to be combined should be specified inside the parentheses and separated by commas. Example: c(1, 2, 3).

```
## create a vector with possible values
possible_values <- c(1, 0) # 1 for heads, 0 for tails
```

TIP: The code used in this chapter can be found in the "Probability.R" file.

Now we can ask R to choose one of these two values at random 1 million times, where the probability of choosing 1 and the probability of choosing 0 are each 0.5. To accomplish this, we can use

`sample()` randomly samples from a set of values. The only required argument is a vector with the set of values to draw from. By default, this function samples values without replacement. To specify the number of draws, we use the argument `size`. To draw with replacement, which allows the same value to be sampled more than once, we set the argument `replace` to `TRUE`. To specify the probabilities of selecting each value, we set the argument `prob` to equal a vector containing the probabilities of each value. Examples: `sample(c(1, 2, 3))` and `sample(c(0, 1), size=1000000, replace=TRUE, prob=c(0.2, 0.8))`.

TIP: When writing code, do not use a comma to indicate thousands or millions. Commas in R are reserved for separating arguments. Example: write `size=1000000`, not size=1,000,000.

RECALL: `prop.table()` converts a frequency table into a table of proportions. The only required argument is the output of the function `table()` with the code identifying the variable inside the parentheses. Example: `prop.table(table(data$variable))`.

RECALL: `mean()` calculates the mean of a variable, and `var()` calculates the variance. Examples: `mean(data$variable)` and `var(data$variable)`.

the function `sample()`, which stands for "randomly sample from a set of values." Inside the parentheses, we first specify a vector with the set of values from which we want to sample. In this case, we use *possible_values* for this vector. Then, we specify that (i) we want 1 million draws by setting the argument `size` to equal `1000000`, (ii) the draws should be with replacement—meaning that we allow the same value to be sampled more than once—by setting the argument `replace` to `TRUE`, and (iii) the probabilities of selecting each value are both 0.5 by setting the argument `prob` to equal the vector `c(0.5, 0.5)`, where the first number identifies the probability of selecting a 1 (the first value in *possible_values*) and the second number identifies the probability of selecting a 0 (the second value in *possible_values*).

```
## randomly sample from possible_values
flip <- sample(possible_values, # vector to draw from
               size=1000000, # 1 million times
               replace=TRUE, # with replacement
               prob=c(0.5, 0.5)) # from a fair coin
```

The variable *flip* contains the results of simulating 1 million flips of a fair coin. (It contains 1 million observations of 1s and 0s.) To calculate the proportion of 1s (heads) and 0s (tails), we can use the function `prop.table()` in conjunction with the function `table()`.

```
prop.table(table(flip)) # creates table of proportions
## flip
##        0        1
## 0.499933 0.500067
```

As we can see in the output above, once we simulate flipping a fair coin 1 million times, the proportion of heads (p) and the proportion of tails ($1-p$) both approximate 0.5. (Note that the values you will see in your console after running the code above will likely not be the exact values shown here. Because we created *flip* via a random process, it will contain slightly different values each time it is created. In fact, all computations in this chapter rely on random processes, and therefore, you should expect slight differences throughout between the outputs we show and what you see in your console.)

Now we can calculate the mean and variance of *flip* by running:

```
mean(flip) # calculates the mean
## [1] 0.500067

var(flip) # calculates the variance
## [1] 0.2500002
```

The mean of *flip* is about 0.5, which is what we expected given that the mean of a Bernoulli distribution is equivalent to p, the probability that the binary random variable equals 1 (here p=0.5).

We can interpret the mean of *flip*, then, as indicating that the probability of the coin landing on heads is 50% (0.5×100=50%).

Finally, the variance of *flip* is 0.25, which is also what we expected given that the variance of a Bernoulli distribution is equal to $p(1-p)$ (in this case: $0.5(1-0.5)=0.25$).

6.4.2 THE NORMAL DISTRIBUTION

The **normal distribution** is a well-known symmetric, bell-shaped distribution, commonly used as an approximation for the distribution of many non-binary variables.

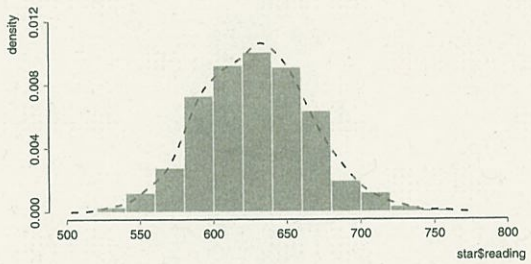
The **normal distribution** is the distribution of a normal random variable. It is characterized by two parameters: mean (μ, pronounced mu) and variance (σ^2, pronounced sigma-squared). In mathematical notation, we write a normal random variable X as:

$$X \sim N(\mu, \sigma^2)$$

For an illustration, let's return to the dataset we analyzed in chapter 2 and create the density histogram of the reading test scores from Project STAR:

```
star <- read.csv("STAR.csv") # reads and stores data

hist(star$reading, freq=FALSE) # creates density histogram
```



RECALL: Before loading the dataset, we need to set the working directory. If the DSS folder is saved directly on your Desktop, to set the working directory, you must run `setwd('~/Desktop/DSS')` if you have a Mac and `setwd('C:/user/Desktop/DSS')` if you have a Windows computer (where *user* is your own username). If the DSS folder is saved elsewhere, please see subsection 1.7.1 for instructions on how to set the working directory.

If we focus on the shape of the histogram demarcated by the height of the bins (shown by the dashed line we added to the density histogram above), we can see that the probability distribution of *reading* is more or less symmetric and bell-shaped. We can approximate the distribution of this non-binary variable using a normal distribution.

The theoretical normal distribution is a family of probability distributions that (i) characterize random variables that can take any value on the real line (from negative infinity to infinity), and (ii) follow a symmetric bell curve that has a very specific shape determined by the formula given in detail below. We refer to random variables that follow a normal distribution as normal random variables.

RECALL: A density histogram provides a visualization of the (probability) distribution of a (random) variable. The relative height of the bins implies the relative likelihood of the values. The areas of all the bins in a density histogram must add up to 1. In R, the function `hist()` creates a density histogram when we set the optional argument `freq` to `FALSE`. Example: `hist(data$variable, freq=FALSE)`.

FORMULA IN DETAIL

The probability density function of the normal probability distribution is determined by the following formula:

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

where:

- μ is the mean of the random variable
- σ is the standard deviation and σ^2 is the variance of the random variable
- x is any value on the real line (from negative infinity to infinity) that the random variable may take
- π is the constant pi, which is approximately 3.1416
- e is the constant known as Euler's number, which is approximately 2.7183.

The **probability density function** of the normal distribution represents the likelihood of each possible value the normal random variable can take (from negative infinity to infinity). The relative height of the curve provides the relative likelihood of the values. The total area underneath the curve of a probability density function equals 1.

The **probability density function** of the normal distribution (the formula in detail above) provides the height of the density curve for each value of x. (See figure 6.1.)

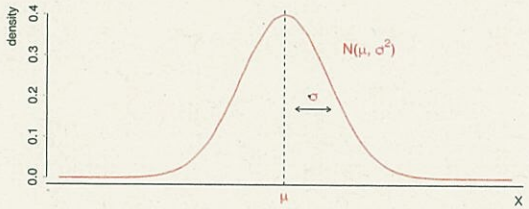


FIGURE 6.1. Probability density function of a normal random variable:

$$X \sim N(\mu, \sigma^2)$$

RECALL: The variance of a variable is the square of the variable's standard deviation. If σ is the standard deviation, then σ^2 is the variance.

The shape of the curve of the probability density function depends on the values of two parameters:

- μ (the Greek letter mu), which stands for the mean of the random variable and determines the center of the distribution
- σ^2 (the Greek letter sigma, squared), which stands for the variance of the random variable and determines the spread of the distribution.

In mathematical notation, if a random variable X follows a normal distribution, we write:

$$X \sim N(\mu, \sigma^2)$$

where:

- X is the name of the random variable
- the symbol $\sim$ stands for "distributed according to"
- N stands for "normal distribution"
- μ is the mean and σ^2 is the variance of the variable.

For example, to state that X is distributed according to a normal distribution with mean 3 and variance 4, we write: $X \sim N(3, 4)$

To visualize the shape of the probability density function of $N(3, 4)$, we can use the formula in detail above. Alternatively and more conveniently, we can ask R to simulate it for us.

Using R, we can randomly draw 1 million observations from the normal distribution we are interested in. Then, we can create the density histogram of the drawn observations. A sample of 1 million observations is large enough that its distribution should approximate the distribution from which the observations are drawn.

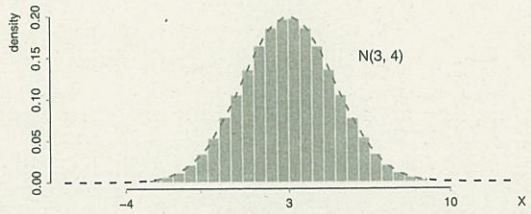
We can start by using the function rnorm(), which stands for "randomly sample from a normal distribution." The one required argument is the number of observations we want to sample. By default, this function samples from the normal distribution with mean 0 and variance 1. If we want to sample from another normal distribution, we can specify a different mean with the optional argument mean, and a different standard deviation with the optional argument sd. (Note that we need to specify the standard deviation, σ, not the variance, σ^2, of the normal distribution we want to sample from.) For example, to draw 1 million observations from $N(3, 4)$ and save them as a variable named X, we run:

```
## randomly sample from distribution N(3, 4)
X <- rnorm(1000000, # sample size
           mean=3, # mean
           sd=2) # standard deviation
```

rnorm() randomly samples from a normal distribution. The only required argument is the number of observations we want to sample. By default, this function samples from the normal distribution with mean 0 and variance 1 (known as the standard normal distribution). To sample from a different normal distribution, we can specify a different mean with the optional argument mean, and a different standard deviation with the optional argument sd. Examples: rnorm(100) and rnorm(100, mean=3, sd=2).

Now, to visualize the probability distribution of X, we can create its density histogram.

```
hist(X, freq=FALSE) # creates density histogram
```



The shape of the probability density function of $N(3, 4)$ is demarcated by the height of the bins of the density histogram (shown by the dashed line we added to the histogram above).

To find out the mean and variance of X, we run:

```
mean(X) # calculates the mean
## [1] 2.998545

var(X) # calculates the variance
## [1] 3.998968
```

Based on the outputs above, the distribution of X is centered at about 3 with variance of about 4. This confirms that the sample of 1 million observations approximately follows the same distribution as the one from which the observations were drawn.

Now, we can follow the same procedure a few times to get a better sense of how the shape of normal distributions varies when the two defining parameters change. For example, figure 6.2 shows the probability density functions of three different normal distributions: $N(0, 1)$, $N(2, 1)$, and $N(0, 4)$.

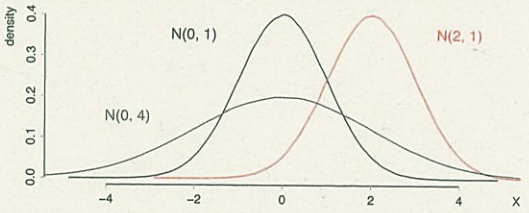


FIGURE 6.2. Probability density functions of three normal distributions: $N(0, 1)$, $N(2, 1)$, and $N(0, 4)$.

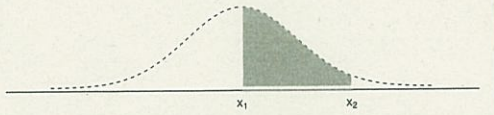
As shown above, both $N(0, 1)$ and $N(0, 4)$ are centered at 0, but $N(0, 4)$ is flatter and more spread out than $N(0, 1)$ because of its larger variance. The spread and height of $N(0, 1)$ and $N(2, 1)$ are the same because they have the same variance, but $N(2, 1)$ is centered at 2, whereas $N(0, 1)$ is centered at 0.

How can we use a probability density function to compute probabilities? We can use the area underneath the curve of the probability density function to compute what are often referred to as *cumulative* probabilities, that is, the probability that a normal random variable takes a value *within a given range*. For example, the area under the curve between x_1 and x_2 equals the probability that the normal random variable takes a value between x_1 and x_2. (Since all probabilities in a distribution must add up to 1, the total area underneath the curve of a probability density function equals 1.)

TIP: The area under the probability density function from negative infinity to x equals the cumulative probability that the normal random variable takes a value less than or equal to x: $P(X \leq x)$. The function that produces this probability is known as the cumulative distribution function.

FOR PROBABILITY DENSITY FUNCTIONS

$P(x_1 \leq X \leq x_2)$ = area under the curve between x_1 and x_2



TIP: The height of the density curve for a particular value of x is not equivalent to the probability of x. There are infinitely many values a normal random variable, X, can take, and the probability that X takes a value *equal to* any specific value, x, is zero. As discussed here, however, we can use the area under the curve of a probability density function to compute the probability that X takes a value within a specific range.

This property of probability density functions enables us to figure out relative probabilities. For example, take a look at the probability density function of X shown in figure 6.3.

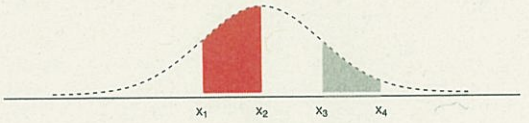


FIGURE 6.3. Probability density function of X where the area under the curve between x_1 and x_2 (shaded in red) is greater than the area under the curve between x_3 and x_4 (shaded in gray). Therefore, the probability that X takes a value between x_1 and x_2 is greater than the probability that X takes a value between x_3 and x_4.

The area under the curve between x_1 and x_2 (shaded in red) is larger than the area under the curve between x_3 and x_4 (shaded in gray). This means that the probability that X takes a value between x_1 and x_2 is greater than the probability that X takes a value between x_3 and x_4. In mathematical notation, we can state:

$$P(x_1 \leq X \leq x_2) > P(x_3 \leq X \leq x_4)$$

6.4.3 THE STANDARD NORMAL DISTRIBUTION

The **standard normal distribution** is the normal distribution with mean 0 (μ=0) and variance 1 (σ^2=1). Since the square root of 1 is 1, the standard deviation of the standard normal distribution is also 1 (σ=1).

The **standard normal distribution** is the normal distribution with mean 0 and variance 1. In mathematical notation, we refer to the standard normal random variable as Z and write it as:

$$Z \sim N(0, 1)$$

In mathematical notation, we usually refer to the standard normal random variable as Z and write it as:

$$Z \sim N(0, 1)$$

(Note that Z here has nothing to do with the Z we used to denote confounding variables in chapter 5.)

Two properties of the standard normal distribution are particularly useful. First, because the distribution is symmetric and centered at 0, the probability that Z takes a value less than or equal to -z is the same as the probability that Z takes a value greater than or equal to z (where z is defined as $z \geq 0$). This is true because the area under the curve between negative infinity and -z is the same as the area under the curve between z and infinity.

FOR THE STANDARD NORMAL RANDOM VARIABLE, Z

$$P(Z \leq -z) = P(Z \geq z)$$

(where $z \geq 0$)

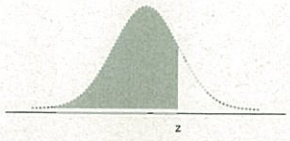
RECALL: As we mentioned in chapter 3, one of the distinct characteristics of normal distributions is that about 95% of the observations fall within two standard deviations from the mean (that is, are between the mean minus two standard deviations and the mean plus two standard deviations). Here, since the standard deviation equals 1, about 95% of the observations are between -2 and 2 (or more precisely, between -1.96 and 1.96).

Second, in the standard normal distribution, about 95% of the observations are between -2 and 2, or more precisely, between -1.96 and 1.96.

FOR THE STANDARD NORMAL RANDOM VARIABLE, Z

$$P(-1.96 \leq Z \leq 1.96) \approx 0.95$$

Let's learn how to calculate probabilities of normal random variables in R so that we can better understand these two properties.



To calculate probabilities of normal random variables, we can use the function pnorm(), which stands for "the cumulative probability of a normal random variable from negative infinity to x." By default, this function calculates the probability that the standard normal random variable takes a value *less than or equal to* the number specified inside the parentheses. (See figure in the margin.) For example, to calculate the probability that Z takes a value less than or equal to -1.96, we run:


```
## probability of Z less than or equal to -1.96
pnorm(-1.96)
## [1] 0.0249979
```

pnorm() calculates the probability that the standard normal random variable, Z, takes a value *less than or equal to* the number specified inside the parentheses. To calculate probabilities of a different normal random variable, we can specify a different mean with the optional argument mean and a different standard deviation with the optional argument sd. Examples: pnorm(0) and pnorm(0, mean=3, sd=2).

Based on the output, we can state that the probability that Z takes a value less than or equal to -1.96 is about 2.5% (0.025×100=2.5%).

If we are interested in the probability that Z takes a value *greater than or equal to* a value, z, we can calculate the probability that Z takes a value *less than or equal to* z and then compute 1 minus the resulting probability. (This is true for all normal random variables because all probabilities in a distribution must add up to 1.)

FOR A NORMAL RANDOM VARIABLE, X

$$P(X \geq x) = 1 - P(X \leq x)$$

TIP: For a normal random variable, X:

$$P(X=x) = 0$$

Therefore:

$$P(X \geq x) = P(X > x)$$
$$P(X \leq x) = P(X < x)$$

For example, if we want to calculate the probability that Z takes a value greater than or equal to 1.96, we run:

```
## probability of Z greater than or equal to 1.96
1 - pnorm(1.96)
## [1] 0.0249979
```

Based on the output, we can state that the probability that Z takes a value greater than or equal to 1.96 is also about 2.5%. This confirms that the probability that Z takes a value less than or equal to -1.96 is the same as the probability that Z takes a value greater than or equal to 1.96.

Now, if we are interested in the probability that Z takes a value between z_1 and z_2, we can calculate the probability that Z takes a value less than or equal to z_2 minus the probability that Z takes a value less than or equal to z_1. (This is also true for all normal random variables because, again, all probabilities in a distribution must add up to 1.)

TIP: For a normal random variable, X:

$$P(X=x) = 0$$

Therefore:

$$P(x_1 \leq X \leq x_2) = P(x_1 < X < x_2)$$
$$P(X \leq x) = P(X < x)$$

FOR A NORMAL RANDOM VARIABLE, X

$$P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1)$$

For example, if we are interested in the probability that Z takes a value between -1.96 and 1.96, we can compute:

$$P(-1.96 \leq Z \leq 1.96) = P(Z \leq 1.96) - P(Z \leq -1.96)$$

So, to calculate the probability that Z takes a value between -1.96 and 1.96, we can run in R:

```
## probability of Z between -1.96 and 1.96
pnorm(1.96) - pnorm(-1.96)
## [1] 0.9500042
```

This output confirms that in the standard normal distribution, about 95% of the observations are between -1.96 and 1.96.

As we will soon see, it is helpful to know these two properties of the standard normal distribution because we can transform any normal random variable into the standard normal random variable. All we need to do is standardize it, that is, subtract the mean from the original normal random variable, and then divide the result by the standard deviation. Graphically, this transformation shifts the center and adjusts the spread of the distribution.

HOW TO TRANSFORM
A NORMAL RANDOM VARIABLE INTO
THE STANDARD NORMAL RANDOM VARIABLE

$$\text{if } X \sim N(\mu, \sigma^2), \quad \frac{X-\mu}{\sigma} \sim N(0, 1)$$

where:

- μ is the mean of X
- σ^2 is the variance of X
- σ is the standard deviation of X ($\sigma=\sqrt{\sigma^2}$).

FORMULA 6.1. Formula to transform a normal random variable, X, into the standard normal random variable, Z.

The resulting standardized variable is commonly referred to as the z-scores of the original random variable. (These are the same z-scores as the ones we used in chapter 3 when computing the correlation coefficient between two variables.)

Take, for example, the normal random variable, X, we created in the subsection above. The variable X has mean 3 and variance 4. Given formula 6.1, $(X-3)/2$ should follow the standard normal distribution. (Note that to standardize a normal random variable, we use the standard deviation, σ, not the variance, σ^2, in the denominator. To compute the standard deviation, we take the square root of the variance. In this case, the variance is 4, and therefore the standard deviation is 2.)

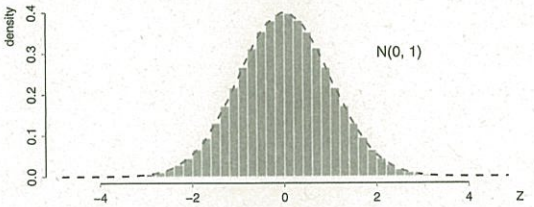
$$\text{if } X \sim N(3, 4), \quad \frac{X-3}{2} \sim N(0, 1)$$

To confirm this, we can ask R to create a new random variable, Z, equivalent to the variable X standardized:

```
## create new random variable
Z <- (X - 3) / 2 # standardized X
```

Then, we can create the density histogram of Z to visualize its probability distribution:

```
hist(Z, freq=FALSE) # creates density histogram
```



As we can see in the density histogram above, Z closely follows the standard normal distribution (centered at 0 and with a variance of 1). To verify this, we calculate the mean and variance of Z by running:

```
mean(Z) # calculates the mean
## [1] -0.0007277148

var(Z) # calculates the variance
## [1] 0.999742
```

As we expected, the distribution of Z is centered at more or less 0, and its variance is approximately 1.

To summarize, while X was a random variable distributed as $N(3, 4)$, after standardization (subtracting the mean and then dividing the result by the standard deviation), the resulting random variable is distributed as $N(0, 1)$. (See figure 6.4, which shows the probability density functions of both X and Z.)

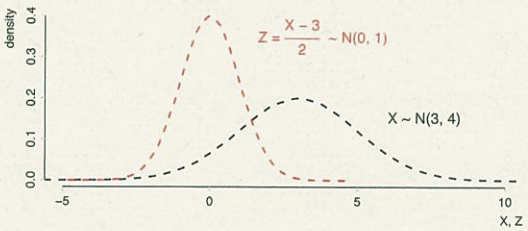


FIGURE 6.4. Probability density functions of X and Z, where Z is equivalent to standardized X.

Why is the transformation of a normal random variable into the standard normal random variable helpful? Because we can use the properties of the standard normal distribution to compute probabilities for all the other types of normal distributions.

Imagine we want to know the range of values that contains 95% of the observations of the normal random variable X above, which again follows a $N(3, 4)$ distribution. Now that we know how to transform X into the standard normal distribution, we can use as our starting point the fact that in the standard normal distribution, about 95% of the observations are between -1.96 and 1.96.

$$P(-1.96 \leq Z \leq 1.96) \approx 0.95$$

$$P(-1.96 \leq \frac{X-3}{2} \leq 1.96) \approx 0.95 \quad (\text{since } Z = \frac{X-3}{2})$$

$$P(-0.92 \leq X \leq 6.92) \approx 0.95 \quad (\text{after multiplying by 2 and adding 3 to each term})$$

After substituting Z and isolating X, we arrive at the conclusion that 95% of the observations of X are between -0.92 and 6.92.

RECALL: To calculate probabilities of a different normal distribution than the standard normal with `pnorm()`, we specify the optional arguments `mean` and `sd`. Example: `pnorm(0, mean=3, sd=2)`.

To confirm this result, we can calculate the probability that $N(3, 4)$ takes a value between -0.92 and 6.92, by running:

```
## probability of N(3, 4) between -0.92 and 6.92
pnorm(6.92, mean=3, sd=2) -
  pnorm(-0.92, mean=3, sd=2)
## [1] 0.9500042
```

Based on the output above, 95% of the observations of $X \sim N(3, 4)$ are indeed between -0.92 and 6.92.

6.4.4 RECAP

In this book, we focus on two types of random variables: binary and normal. Their probability distributions are summarized below.

IF X IS A BINARY RANDOM VARIABLE	IF X IS A NORMAL RANDOM VARIABLE
X has a Bernoulli distribution, characterized by one parameter: p	X has a normal distribution, characterized by two parameters: μ and σ^2
mean $= p$	mean $= \mu$
variance $= p(1-p)$	variance $= \sigma^2$

Now that we are familiar with what probability is and the probability distributions of binary and normal random variables, let's clarify the distinction between population parameters and sample statistics.

6.5 POPULATION PARAMETERS VS. SAMPLE STATISTICS

When analyzing data, we are usually interested in the value of a parameter at the population level. For example, we might be interested in the level of support for a particular political candidate among the population of all voters in a country. Typically, however, we have access to statistics from only a small sample of observations drawn from the target population. For example, we may know only the proportion of supporters among the voters who responded to a survey. In this section, we see how to use sample statistics to learn about the corresponding population parameters.

TIP: Parameters are unknown quantities of interest (often at the population level). Statistics are based on the sample of data observed; that is, they are sample-specific.

POPULATION PARAMETERS OF RANDOM VARIABLE X	SAMPLE STATISTICS OF n OBSERVATIONS OF X
mean $= \mathbb{E}(X)$ (expectation of X)	mean $= \overline{X}$ (sample mean of X)
variance $= \mathbb{V}(X)$ (population variance of X)	variance $= var(X)$ (sample variance of X)

The **sample mean of X**, $\overline{X}$, refers to the average value of X in a particular sample, while the **expectation of X**, $\mathbb{E}(X)$, refers to the population mean of the random variable X. The **sample variance of X**, $var(X)$, refers to the variance of X in a particular sample, while the **population variance of X**, $\mathbb{V}(X)$, refers to the population variance of the random variable X.

To distinguish the sample statistics from the corresponding parameters at the population level, we use different terms to refer

to them. The **sample mean of X**, denoted as $\overline{X}$, refers to the average value of X in a particular sample, while the **expectation of X**, denoted as $\mathbb{E}(X)$, refers to the population mean of the random variable X. The **sample variance of X**, denoted as $var(X)$, refers to the variance of X in a particular sample, while the **population variance of X**, denoted as $\mathbb{V}(X)$, refers to the population variance of the random variable X.

RECALL: The mean of a binary variable is equivalent to the proportion of observations that have the characteristic identified by the variable.

In the current example, we can define *support* as a binary variable that identifies whether individual i supports the candidate of interest (1=support, 0=no support). The sample mean of *support* would be the proportion of supporters among survey respondents, and the expectation of *support* would be the proportion of supporters among all the individuals in the target population.

Are the population-level parameters identical to the sample-level statistics? They are generally not the same unless the sample is the entire population.

Sampling variability refers to the fact that the value of a statistic varies from one sample to another because each sample contains a different set of observations drawn from the target population. Smaller sample size generally leads to greater sampling variability.

The sample statistics differ from the population parameters because the sample contains noise. The noise comes from **sampling variability**. If we randomly draw multiple samples from the same population, each sample will contain different observations. As a result, each sample will yield different values of sample statistics, even if all the observations are drawn using random sampling. In the running example, different surveys will show varying levels of support for the candidate because they contain different respondents. This will be true even when the surveys use exactly the same method to select their respondents.

Smaller sample size generally leads to greater sampling variability. Conversely, as sample size increases, sampling variability decreases. This is why when we get to extremely large sample sizes, such as 1 million observations, the sample statistics approximate the population parameters well. In the survey example, as the sample size increases, we expect the sample proportion of supporters to approach the population proportion of supporters.

When drawing conclusions about the population from a sample, we need to take into consideration the noise in the data introduced by sampling variability. The two large sample theorems we discuss in this section—the law of large numbers and the central limit theorem—help us do just that by clarifying the relationship between population parameters and sample statistics.

6.5.1 THE LAW OF LARGE NUMBERS

The law of large numbers states that as the sample size increases, the sample mean of X approximates the population mean of X, also known as the expectation of X.

THE LAW OF LARGE NUMBERS

$$\text{as } n \text{ increases, } \quad \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \approx \quad \mathbb{E}(X)$$

where:

- n is the sample size
- X is the original random variable
- $\overline{X}$ is the sample mean of X
- $\mathbb{E}(X)$ is the population mean of X

To illustrate the law of large numbers, we can use R to draw random samples of different sizes from the same distribution and compare the sample means to the population mean. To show the general applicability of this theorem, we will do this exercise twice, once with an original random variable that is binary and once with an original random variable that is normal.

EXAMPLE WITH A BINARY RANDOM VARIABLE

Suppose we are interested in the binary variable *support* as defined above (1=support, 0=no support). Since this is a binary variable, it has a Bernoulli distribution.

Further suppose that 60% of the voters in the country support the political candidate of interest. The probability that *support* equals 1, then, is 0.60 (p=0.60), which is equal to the population mean of *support* ($\mathbb{E}(support)$=0.60). (As we saw earlier, the mean of a Bernoulli distribution is equivalent to the probability that the binary variable equals 1, denoted as p.)

Now we can use the function `sample()` to draw three random samples from this particular binary variable, each of a different size. Note that in this case, we set the argument `prob` to equal `c(0.6, 0.4)` because the probability of 1 is 0.6 (p=0.60) and the probability of 0 is 0.4 ($1-p$=1−0.6=0.4).

```
## draw random samples from binary variable
support_sample_1 <- sample(c(1, 0), # possible values
                           size=10, # n=10
                           replace=TRUE, # with replacement
                           prob=c(0.6, 0.4)) # probabilities

support_sample_2 <- sample(c(1, 0),
                           size=1000, # n=1,000
                           replace=TRUE,
                           prob=c(0.6, 0.4))
```

RECALL: `sample()` randomly samples from a set of values. The only required argument is a vector with the set of values to draw from. By default, this function samples values without replacement. To specify the number of draws, we use the argument `size`. To draw with replacement, which allows the same value to be sampled more than once, we set the argument `replace` to `TRUE`. To specify the probabilities of selecting each value, we set the argument `prob` to equal a vector containing the probabilities of each value. Examples: `sample(c(1, 2, 3))` and `sample(c(0, 1), size=1000000, replace=TRUE, prob=c(0.2, 0.8))`.

```
support_sample_3 <- sample(c(1, 0),
                           size=1000000, # n=1,000,000
                           replace=TRUE,
                           prob=c(0.6, 0.4))
```

As we can see in the code above, the first sample contains 10 observations, the second contains 1,000 observations, and the third contains 1 million observations. To calculate the mean for each of the three samples, we run:

```
## calculate sample means
mean(support_sample_1) # in n=10 sample
## [1] 0.8

mean(support_sample_2) # in n=1,000 sample
## [1] 0.62

mean(support_sample_3) # in n=1,000,000 sample
## [1] 0.599957
```

RECALL: The mean of a binary variable is interpreted as the proportion of observations that have the characteristic identified by the variable (after multiplying the number by 100).

The proportion of support among respondents varies across the three samples. In the first sample, 80% of respondents support the candidate; in the second, 62% of respondents support the candidate; and in the third, close to 60% of respondents support the candidate. The sample with the largest number of observations, sample 3 with 1 million observations, produces the proportion of support that is closest to the true proportion of support in the population. This finding is consistent with the fact that as the sample size increases, the sample mean tends to be closer to the population mean, which in this case is 60%.

EXAMPLE WITH A NORMAL RANDOM VARIABLE

Now suppose we are interested in the height, measured in inches, of each person in a population. We assume that the corresponding random variable, *height*, follows a normal distribution. Further suppose that we know that the mean of this normal distribution is 67 inches and the variance is 14 inches.

RECALL: rnorm() randomly samples from a normal distribution. The only required argument is the number of observations we want to sample. By default, this function draws observations from the standard normal distribution (mean=0 and sd=1). To sample from a different normal distribution, we can specify a different mean with the optional argument mean and a different standard deviation with the optional argument sd. Examples: rnorm(100) and rnorm(100, mean=3, sd=2).

We can use the function rnorm() to draw three random samples from this normal random variable, each of a different size:

```
## draw random samples from normal distribution
height_sample_1 <- rnorm(10, # n=10
                         mean=67, # population mean=67
                         sd=sqrt(14)) # population variance=14

height_sample_2 <- rnorm(1000, # n=1,000
                         mean=67,
                         sd=sqrt(14))
```



```
height_sample_3 <- rnorm(1000000, # n=1,000,000
                         mean=67,
                         sd=sqrt(14))
```

RECALL: When using the function rnorm(), to change the spread of the normal distribution, we need to specify the standard deviation, not the variance. Here, given that the variance is 14, the standard deviation is sqrt(14).

As in the previous example, the first sample contains 10 observations, the second contains 1,000 observations, and the third contains 1 million observations. To calculate the sample mean for each of the samples, we run:

```
## calculate sample means
mean(height_sample_1) # in n=10 sample
## [1] 65.21607

mean(height_sample_2) # in n=1,000 sample
## [1] 66.81554

mean(height_sample_3) # in n=1,000,000 sample
## [1] 66.99905
```

The average height varies across the three samples. It is about 65.22 inches in the first sample, 66.82 inches in the second, and 67 inches in the third. Here, too, as the sample size increases, the sample mean tends to approach the population mean of the original random variable, which in this case is 67 inches.

6.5.2 THE CENTRAL LIMIT THEOREM

The **central limit theorem** states that as the sample size increases, the standardized sample mean of X can be approximated by the standard normal distribution.

The central limit theorem states that as the sample size increases, the standardized sample mean of X can be approximated by the standard normal distribution.

THE CENTRAL LIMIT THEOREM

$$\text{as } n \text{ increases, } \frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}} \overset{\text{approx.}}{\sim} N(0, 1)$$

where:

- n is the sample size
- X is the original random variable, and $\overline{X}$ is the sample mean, a random variable containing the sample means from multiple large samples of X
- $\mathbb{E}(X)$ is the population mean of X, and $\mathbb{V}(X)$ is the population variance of X
- $\overset{\text{approx.}}{\sim}$ stands for "approximately distributed according to," and $N(0, 1)$ is the standard normal distribution.

Let's see how we arrive at this theorem so that we can understand it better.

First, we can think of the sample mean of X as a random variable because it varies from one sample to another. As is the case with all random variables, the sample mean of X has its own distribution.

Second, the central limit theorem implies that as the sample size increases, the distribution of the sample mean of X approaches the normal distribution.

as n increases, $\overline{X}$ is approximately distributed as normal

Third, as you may recall, the normal distribution is characterized by two parameters: mean and variance. To figure out the mean and variance of the sample mean of X, we need to rely on the properties of expectations and variances. (See the formulas in detail below.)

FORMULA IN DETAIL

Some properties of expectations:

- $\mathbb{E}(aX) = a\,\mathbb{E}(X)$ where a is a constant and X is a random variable
- $\mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$ where X_1 and X_2 are random variables.

Given the properties above, what is the population mean or expectation of the sample mean of X, $\mathbb{E}(\overline{X})$?

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \quad \text{because } \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \quad \text{because } \mathbb{E}(aX) = a\,\mathbb{E}(X)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) \quad \text{because } \mathbb{E}(X_1 + X_2) = \mathbb{E}(X_1) + \mathbb{E}(X_2)$$

$$= \frac{1}{n} \times n\,\mathbb{E}(X) \quad \text{because } \sum_{i=1}^{n}\mathbb{E}(X_i) = n\,\mathbb{E}(X)$$

$$= \mathbb{E}(X)$$

FORMULA IN DETAIL

Some properties of variances:

- $\mathbb{V}(aX) = a^2\,\mathbb{V}(X)$ where a is a constant and X is a random variable
- $\mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2)$ where X_1 and X_2 are random variables that are independent of each other (that is, the values of one variable cannot be used to infer the values of the other).

Given the properties above, what is the population variance of the sample mean of X, $\mathbb{V}(\overline{X})$?

$$\mathbb{V}(\overline{X}) = \mathbb{V}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) \quad \text{because } \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \left(\frac{1}{n}\right)^2 \mathbb{V}\left(\sum_{i=1}^{n} X_i\right) \quad \text{because } \mathbb{V}(aX) = a^2\,\mathbb{V}(X)$$

$$= \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n}\mathbb{V}(X_i) \quad \text{because } \mathbb{V}(X_1 + X_2) = \mathbb{V}(X_1) + \mathbb{V}(X_2)$$

$$= \frac{1}{n^2} \times n\,\mathbb{V}(X) \quad \text{because } \sum_{i=1}^{n}\mathbb{V}(X_i) = n\,\mathbb{V}(X)$$

$$= \frac{\mathbb{V}(X)}{n}$$

As shown in detail above:

- the population mean of the sample mean of X equals the population mean of X:

$$\mathbb{E}(\overline{X}) = \mathbb{E}(X)$$

- the population variance of the sample mean of X equals the population variance of X divided by the sample size:

$$\mathbb{V}(\overline{X}) = \frac{\mathbb{V}(X)}{n}$$

Fourth, now that we know the population mean and variance of the sample mean, we can standardize the sample mean of X by subtracting the population mean and dividing the result by the population standard deviation (see formula 6.1). The standardized sample mean is then:

$$\frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}}$$

According to the central limit theorem, as the sample size increases, the standardized sample mean of X (as defined above) can be approximated by the standard normal distribution:

$$\text{as } n \text{ increases, } \frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}} \overset{\text{approx.}}{\sim} N(0, 1)$$

Remarkably, this theorem holds when the original random variable X follows a Bernoulli distribution. In fact, it holds when the original random variable follows almost any of the distributions we use in statistics. This is important because we rarely know the probability distribution that generates the data of interest.

To illustrate this theorem, we can use R to (i) draw multiple, large random samples from the same distribution, (ii) compute the mean of each sample, (iii) standardize the mean of each sample applying formula 6.1 using the population mean and population variance of the sample mean, (iv) save the standardized sample means as a new variable, and (v) examine the distribution of the standardized sample means. If the samples are large enough, the standardized sample means should approximately follow the standard normal distribution.

Here we go over only one example, one in which the original random variable is binary. If the original random variable is normal, we do not need the central limit theorem. In this case, the standardized sample mean follows *exactly* the standard normal distribution, which makes the large sample approximation unnecessary.

EXAMPLE WITH A BINARY RANDOM VARIABLE

Let's return to the binary random variable *support*, which, as we discussed above, follows a Bernoulli distribution.

We continue to suppose that 60% of the voters in the country support the political candidate of interest. So, in this case: p=0.60.

Given the properties of Bernoulli distributions, the original random variable, *support*, should be centered at 0.60 and have a variance of 0.24.

$$\mathbb{E}(support) = p = 0.60$$
$$\mathbb{V}(support) = p(1-p) = 0.6 \times (1-0.6) = 0.24$$

`c()` combines values into a vector. If no main argument is provided, this function creates an empty vector that can be used to store outputs. Example: `c()`.

To start the simulation, we need to create an empty vector where we will store the standardized means of the samples from the random variable *support*. For this purpose, we can use the function `c()`, which creates an empty vector when no arguments are specified inside.


```
## create an empty vector to store standardized sample means
sd_sample_means <- c()
```

Now, we can use R to draw 10,000 random samples from *support*, each one containing 1,000 observations, and save the standardized mean of each sample in the vector we have just created. Because we do not want to write the code to draw a random sample 10,000 times, we can use what is known as a `for loop`. A for loop executes a given code repeatedly, for as many times as indicated. (For a more detailed explanation on how for loops work, please see the appendix near the end of this chapter.)

`for(i in 1:n){}` is the basic syntax of a for loop. A for loop executes a given code repeatedly, for each i in the sequence from 1 to n (meaning for i={1, . . . , n}, using one i at a time, starting with i=1 and ending with i=n). The code to be executed repeatedly should be specified inside the curly brackets. Example: `for(i in 1:3){print(i)}` displays the value of i from i=1 until i=3.

For example, by running the code below, we are asking R, for each i in the sequence from i=1 until i=10,000 (so, 10,000 times in total), to:

- draw a random sample of 1,000 observations from a binary random variable with p=0.6
- calculate the standardized sample mean, which in this case is:

$$\frac{\overline{support_i} - \mathbb{E}(support)}{\sqrt{\mathbb{V}(support)/n}} = \frac{\overline{support_i} - 0.60}{\sqrt{0.24/1000}}$$

- store the standardized sample mean in observation i of the empty vector *sd_sample_means*.

TIP: Here i is not representing the position of the observation but rather the number of the iteration of the for loop. In the first iteration, i=1. In the last iteration, i=n (10,000 in this case).

```
## for loop with 10,000 iterations
for(i in 1:10000){
  ## draw a random sample of 1,000 observations
  ## from binary random variable with p=0.6
  support_sample <- sample(c(1, 0), # possible values
                           size=1000, # n=1,000
                           replace=TRUE, # with replacement
                           prob=c(0.6, 0.4)) # probabilities
  ## calculate and store the standardized sample mean
  sd_sample_means[i] <-
      (mean(support_sample) -0.60) / sqrt(0.24 / 1000)
}
```

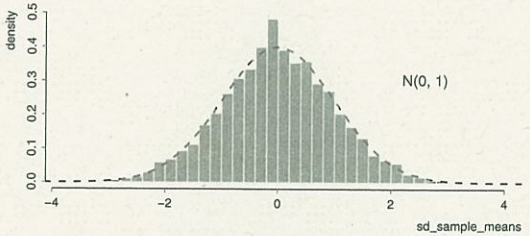
`[]` is the operator used to extract a selection of observations from a vector. To its left, we specify the vector we want to subset. Inside the square brackets, we specify the criterion of selection. For example, we can specify the position of the observation `i` to be extracted. Example: `vector[i]`.

TIP: Make sure to run this piece of code all at once, starting with `for(i in 1:n){` and all the way until `}`. Otherwise, R will not be able to execute it and will give you an error message.

After running the code above, *sd_sample_means* will contain the standardized sample means of 10,000 samples from the binary random variable we are interested in (that is, p=0.6).

Now we can visualize the distribution of the standardized sample means by creating the density histogram:

```
## create density histogram
hist(sd_sample_means, freq=FALSE)
```



As we expected, even though the samples were drawn from a binary random variable, the standardized sample means approximately follow the standard normal distribution.

6.5.3 SAMPLING DISTRIBUTION OF THE SAMPLE MEAN

TIP: The central limit theorem states:

$$\frac{\overline{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)/n}} \overset{\text{approx.}}{\sim} N(0,1)$$

Using formula 6.1 in the opposite direction, we can conclude:

$$\overline{X} \overset{\text{approx.}}{\sim} N\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

Thanks to the central limit theorem, we know that if we drew multiple large samples of a random variable X, with mean $\mathbb{E}(X)$ and variance $\mathbb{V}(X)$, the sample means would approximately follow a normal distribution with mean $\mathbb{E}(X)$ and variance $\mathbb{V}(X)/n$.

POPULATION

random variable X
population mean $= \mathbb{E}(X)$
population variance $= \mathbb{V}(X)$

↓

SAMPLE 1 | SAMPLE 2 | ... | SAMPLE K

$\overline{X}_{\text{sample1}}$ | $\overline{X}_{\text{sample2}}$ | | $\overline{X}_{\text{sampleK}}$

↓

SAMPLE MEANS FROM MULTIPLE LARGE SAMPLES

$$\overline{X} = \{\overline{X}_{\text{sample1}}, \overline{X}_{\text{sample2}}, \ldots, \overline{X}_{\text{sampleK}}\}$$

BASED ON THE CENTRAL LIMIT THEOREM

$$\overline{X} \overset{\text{approx.}}{\sim} N\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$$

The **sampling distribution of the sample mean** characterizes how much the sample means vary from one sample to another due to sampling variability.

The distribution above is known as the **sampling distribution of the sample mean**. It characterizes how much the sample means vary from one sample to another due to sampling variability.

6.6 SUMMARY

This chapter introduced us to probability. First, we learned about the frequentist and Bayesian interpretations of probability and the axioms of probability. Next, we saw the probability distributions of two types of random variables, binary and normal, paying special attention to the standard normal distribution. Then, we learned two large sample theorems—the law of large numbers and the central limit theorem—and specifically how they help us understand the relationship between population parameters and sample statistics. Finally, we ended the chapter by using the central limit theorem to derive the sampling distribution of the sample mean. In the next chapter, we will learn how to use this knowledge to quantify the level of uncertainty in our population-level conclusions.

6.7 APPENDIX: FOR LOOPS

RECALL: `for(i in 1:n){ }` is the basic syntax of a for loop. A for loop executes a given code repeatedly, for each i in the sequence from 1 to n (meaning for $i=\{1,\ldots,n\}$, using one i at a time, starting with $i=1$ and ending with $i=n$). The code to be executed repeatedly should be specified inside the curly brackets. Example: `for(i in 1:3){print(i)}` displays the value of i from i=1 until i=3.

`print()` displays in the R console the argument specified inside the parentheses. Example: `print("this")`.

Let's start by looking at a simple example to understand how for loops work in R. Go ahead and run the following code:

```
for(i in 1:3){
  print(i) # print the value of i
}
## [1] 1
## [1] 2
## [1] 3
```

The first line of code, `for(i in 1:3){`, can be interpreted as: "For each i in the sequence from 1 to 3 (meaning for $i=\{1, 2, 3\}$, using one i at a time, starting with $i=1$ and ending with $i=3$), execute the following code."

The second line is the code that will be executed repeatedly, in sequence, from $i=1$ until $i=3$. In this case, the code simply asks R to print the value of i using the function `print()`.

Finally, the third line of code closes the parentheses we started in the first line, to indicate the end of the code to be executed repeatedly.

After running the three lines of code all together, R provides us three outputs, one for each i between 1 and 3. The first output is 1, since that was the value of i in the first iteration and so on.

Now, let's modify the for loop above to change the code we want R to execute repeatedly. Go ahead and run:

```
## for loop with 3 iterations
for(i in 1:3){
  ## draw a random sample of 1,000 observations
  ## from binary variable with p=0.5
  flip <- sample(c(1, 0), # possible values
                 size=1000, # n=1,000
                 replace=TRUE, # with replacement
                 prob=c(0.5, 0.5)) # probabilities
  ## print sample mean
  print(mean(flip))
}
## [1] 0.495
## [1] 0.505
## [1] 0.502
```

Because we didn't modify the first line of code, the for loop includes only three iterations. Hence, running the code produces three outputs. Each is the sample mean of simulating 1,000 flips from a fair coin (where 1 stands for heads and 0 stands for tails). Notice that each sample mean is slightly different. As we discussed earlier, these differences are due to sampling variability.

Now, we can modify the for loop so that instead of printing the sample means, R stores them into a vector. We start by creating an empty vector to store the sample means:

```
## create an empty vector to store sample means
sample_means <- c()
```

RECALL: `c()` combines values into a vector. If no main argument is provided, this function creates an empty vector that can be used to store outputs. Example: `c()`.

Then, we need to modify the code to be executed repeatedly. Instead of `print(mean(flip))`, we write `sample_means[i] <- mean(flip)`. This code saves each sample mean as a new observation in the vector *sample_means*. The `[i]` following the name of the vector on the left hand side of the assignment operator subsets the vector to the observation i. As a result, the first sample mean is saved as the first observation of the vector, and so on.

RECALL: `[]` is the operator used to extract a selection of observations from a vector. To its left, we specify the vector we want to subset. Inside the square brackets, we specify the criterion of selection. For example, we can specify the position of the observation `i` to be extracted. Example: `vector[i]`.

```
## for loop with 3 iterations
for(i in 1:3){
  ## draw a random sample of 1,000 observations
  ## from binary variable with p=0.5
  flip <- sample(c(1, 0), # possible values
                 size=1000, # n=1,000
                 replace=TRUE, # with replacement
                 prob=c(0.5, 0.5)) # probabilities
  ## store sample mean
  sample_means[i] <- mean(flip)
}
```

After running the code above, *sample_means* should contain the sample means of three random samples. To confirm this, run the name of the object so that R provides its contents:

```
sample_means # shows contents of object
## [1] 0.481 0.531 0.485
```

Finally, if we wanted to draw 10,000 samples instead of three, we would modify the first line of the for loop. Instead of `for(i in 1:3)`, we would write `for(i in 1:10000)`.

```
## for loop with 10,000 iterations
for(i in 1:10000){
  ## draw a random sample of 1,000 observations
  ## from binary variable with p=0.5
  flip <- sample(c(1, 0), # possible values
                 size=1000, # n=1,000
                 replace=TRUE, # with replacement
                 prob=c(0.5, 0.5)) # probabilities
  ## store sample mean
  sample_means[i] <- mean(flip)
}
```

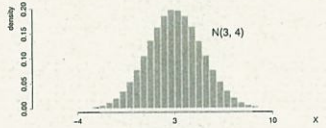
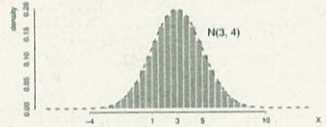
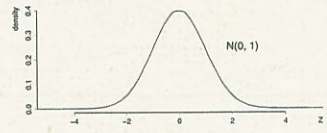
6.8 CHEATSHEETS

6.8.1 CONCEPTS AND NOTATION

concept/notation	description	example(s)
frequentist interpretation of probabilities	probabilities represent proportions of specific events occurring over infinitely many identical trials	when flipping a coin, the probability of heads is the proportion of heads observed over infinitely many identical flips
Bayesian interpretation of probabilities	probabilities represent personal, subjective beliefs about the relative likelihood of events; a probability of 1, or 100%, indicates certainty that the event will occur; a probability of 0, or 0%, indicates certainty that the event will not occur	when stating that the probability of rain today is 80%, we are describing how certain we are about the rain event occurring; we are not describing the frequency of rain events over multiple days
trial	action or set of actions that produces outcomes of interest	rolling a die
outcome	the result of a trial	rolling a die produces one of six possible outcomes: 1, 2, 3, 4, 5, or 6
event	a set of outcomes; an event is said to occur if any one of the possible outcomes included in the event is realized	*rolling a number less than 3* is one of the potential events that may occur when rolling a die; if we roll a 1, we would consider that the event *rolling a number less than 3* has occurred
mutually exclusive events	events that do not share any outcomes	*rolling a number less than 3* and *rolling a 3* are mutually exclusive events when rolling a die
sample space (Ω)	denoted by the Greek letter Omega; the set of all possible outcomes produced by a trial; considered an event in itself	in the case of rolling a die: $\Omega = \{1, 2, 3, 4, 5, 6\}$
random variable	assigns a numeric value to each mutually exclusive event produced by a trial	we can create a random variable, *flip*, to capture the results of flipping coins, where 1s mean heads and 0s mean tails: $flip_i = \begin{cases} 1 \text{ if coin flip } i \text{ lands on heads} \\ 0 \text{ if coin flip } i \text{ lands on tails} \end{cases}$
probability distribution	characterizes the likelihood of each possible value a random variable can take; all probabilities in a distribution must add up to 1	the probability distribution of the random variable *flip* above could be: $P(flip{=}1) = p = 0.67$ $P(flip{=}0) = 1{-}p = 1{-}0.67 = 0.33$
Bernoulli distribution	probability distribution of a binary variable it is characterized by one parameter, p, which is the probability that the binary random variable takes the value of 1; consequently, $1{-}p$ is the probability that the binary random variable takes the value of 0 the mean of a Bernoulli distribution is p and the variance is $p(1{-}p)$	flipping a coin can result in only one of two possible events: heads or tails; if we assign 1 to heads and 0 to tails, we can create a binary random variable with the results of multiple coin flips; this binary random variable will follow a Bernoulli distribution $P(flip{=}1) = p$ $P(flip{=}0) = 1{-}p$

continues on next page...

6.8.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
normal distribution	distribution of a normal random variable we write a normal random variable X as: $X \sim N(\mu, \sigma^2)$ it is characterized by two parameters: - μ (the Greek letter mu), which stands for the mean of X - σ^2 (the Greek letter sigma, squared), which stands for the variance of X two useful properties of X: - $P(X \geq x) = 1 - P(X \leq x)$ - $P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1)$	the density histogram of the normal random variable, X, with mean 3 and variance 4 is: 
probability density function of the normal distribution	determined by the following formula: $\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$ represents the likelihood of each possible value the normal random variable can take (from negative infinity to infinity); the relative height of the curve provides the relative likelihood of the values the area under the curve between x_1 and x_2 equals the probability that the normal random variable takes a value between x_1 and x_2; the total area underneath the curve equals 1	the shape of the probability density function is demarcated by the height of the bins of the density histogram; below is the density histogram of $X \sim N(3, 4)$ with the probability density function shown as a dashed line:  the area under the curve between -4 and 1 is smaller than the area under the curve between 1 and 5; therefore, the probability that X takes a value between -4 and 1 is lower than the probability that X takes a value between 1 and 5
standard normal distribution	normal distribution with mean 0 and variance 1 in mathematical notation, we refer to the standard normal random variable as Z and write it as: $Z \sim N(0, 1)$ two useful properties of Z: - $P(Z \leq -z) = P(Z \geq z)$ (where $z \geq 0$) - $P(-1.96 \leq Z \leq 1.96) \approx 0.95$ to transform a normal random variable X into the standard normal random variable Z, we subtract the mean and then divide the result by the standard deviation: if $X \sim N(\mu, \sigma^2)$, $\frac{X - \mu}{\sigma} \sim N(0, 1)$	the probability density function of Z is:  if $X \sim N(3, 4)$, then: $\frac{X - 3}{2} \sim N(0, 1)$

continues on next page...

6.8.1 CONCEPTS AND NOTATION (CONTINUED)

concept/notation	description	example(s)
expectation of X or expected value of X or $\mathbb{E}(X)$	mean of the random variable X at the population level	if the true support for a candidate at the population level is 40% and *support* is a binary variable that identifies the support for this candidate, then: $\mathbb{E}(\text{support})=p=0.40$
sample mean of X or $\overline{X}$	average value of X in a particular sample	if the support for a candidate among a sample of individuals from the population above is 35%, then: $\overline{\text{support}}=0.35$
population variance of X or $\mathbb{V}(X)$	variance of the random variable X at the population level	in the example above, because *support* is a binary variable, the population variance of *support* is: $\mathbb{V}(\text{support})=p(1-p)=0.4(1-0.4)=0.24$
sample variance of X or $var(X)$	variance of a sample of observations of X	in the example above, the variance of the variable *support* in the sample is: $var(\text{support})=0.35(1-0.35)=0.23$
sampling variability	refers to the fact that the value of a statistic varies from one sample to another because each sample contains a different set of observations drawn from the target population; smaller sample size generally leads to greater sampling variability	if we conduct 100 surveys, each containing a representative sample of 1,000 individuals from a population of millions, the results will differ from one survey to another because of sampling variability
law of large numbers	states that as the sample size increases, the sample mean of X approximates the population mean of X	the mean of a sample of 100,000 observations of X is likely to be closer to the population mean of X than the mean of a sample of 100 observations of X
central limit theorem	states that as the sample size increases, the standardized sample mean of X can be approximated by the standard normal distribution	if we were to draw multiple large samples of X, the standardized sample means will follow the standard normal distribution, regardless of how X is distributed
sampling distribution of the sample mean	characterizes how much the sample means vary from one sample to another due to sampling variability $\overline{X} \overset{\text{approx.}}{\sim} N\left(\mathbb{E}(X), \frac{\mathbb{V}(X)}{n}\right)$	if we drew multiple samples of 1,000 observations of a random variable X, with mean 2 and variance 4, the sample means would approximately follow a normal distribution with mean 2 and variance 0.004 (4/1000=0.004)

6.8.2 R SYMBOLS AND OPERATORS

code	description	example(s)
for(i in 1:n){}	basic syntax of a for loop; a for loop executes a given code repeatedly, for each i in the sequence from 1 to n, meaning for $i=\{1,\ldots,n\}$, using one i at a time, starting with $i=1$ and ending with $i=n$; the code to be executed repeatedly should be specified inside the curly brackets	for(i in 1:3){ print(i) } # displays the value of i from i=1 until i=3

6.8.3 R FUNCTIONS

function	description	required argument(s)	example(s)
c()	combines values into a vector (a collection of elements, each identified by an index)	values to be combined, separated by commas; if no main argument is provided, this function creates an empty vector that can be used to store outputs	c(1, 2, 3) c() # creates an empty vector
sample()	randomly samples from a set of values; by default, it samples without replacement	the vector with the set of values to draw from optional argument size: specifies the number of draws optional argument replace: if set to TRUE, the function draws with replacement (allowing the same value to be drawn more than once) optional argument prob: specifies the probabilities of selecting each value in the vector; we set this argument to equal a vector containing the probabilities of each value	sample(c(1, 2, 3)) # randomly draws one observation at a time from the vector c(1; 2, 3), without replacement sample(c(0, 1), size=1000, replace=TRUE, prob=c(0.2, 0.8)) # randomly draws 1,000 observations of 0s and 1s, with replacement, where the probability of a 0 is 20% and the probability of a 1 is 80%

rnorm()	randomly samples from a normal distribution; by default, this function samples from the standard normal distribution	the number of observations we want to sample
		optional argument mean: specifies the mean of the normal distribution to sample from (if different than the default of 0)
		optional argument sd: specifies the standard deviation of the normal distribution to sample from (if different than the default of 1)
			rnorm(100)
			# randomly draws 100 observations from the standard normal distribution
			rnorm(100, mean=3, sd=2)
			# randomly draws 100 observations from the normal distribution with mean 3 and standard deviation 2

pnorm()	calculates the probability that a normal random variable takes a value *less than or equal to* the number specified inside the parentheses; by default, it calculates probabilities of the standard normal random variable	the number for which we want to calculate the probability
		optional argument mean: specifies the mean of the normal random variable we want to compute probabilities of (if different than the default of 0)
		optional argument sd: specifies the standard deviation of the normal random variable we want to compute probabilities of (if different than the default of 1)
			pnorm(0)
			# computes the probability that the standard normal random variable takes a value less than or equal to 0
			pnorm(0, mean=3, sd=2)
			# computes the probability that the normal random variable with mean 3 and standard deviation 2 takes a value less than or equal to 0

print()	displays in the R console the specified argument	what we want to have displayed in the R console	print("this")