In this chapter, we introduce random variables, the fundamental objects of study in statistical modeling. We begin by defining them and describing some of their basic properties. We then discuss numerical methods of summarizing random variables, followed by graphical methods.

1.1 Preliminaries

We begin our study of statistical modeling with some preliminary definitions. The first of these is not really a technical term, but we did find it now because we will use it frequently throughout the text.

**Definition 1.1.1** A variable is a characteristic or property to be measured or observed. The result of measuring or observing a variable is called a value, even when it is not a number.

This definition is admittedly rather vague, but some examples will help to clarify it.

**Example 1.1.2** If we are studying trees, tree height might be one variable of interest. For each tree, observing tree height yields a numerical value.

**Example 1.1.3** If we are studying the relationship between hair color and eye
color in human beings (or some subpopulation thereof), both hair color and eye color are variables. For each human being, observing either one yields a value (such as “brown” or “blue”), although not a numerical one.

### Types of variables

There are many different ways to classify variables, but for our purposes there are a few particularly useful types of variables to distinguish. We introduce the first of these now.

**Definition 1.1.4** A variable is called **numerical** if the values that it can take on are all real numbers\(^1\) and if it makes sense to perform arithmetic with these values.

The requirement that a variable takes on values with which it makes sense to perform arithmetic is somewhat vague, but we now illustrate it (and the rest of the preceding definition) with some examples.

**Example 1.1.5** If we are studying trees as in Example 1.1.2, tree height would be considered to be a numerical variable. The values that it takes on are real numbers, and it makes sense to perform arithmetic operations with them. For example, the sum of the heights of two trees does tell us something meaningful about how tall those two trees are on average.

**Example 1.1.6** The zip code of a United States resident is a variable that takes on values that are all real numbers. However, it is not considered a numerical variable, since it doesn’t make sense to perform arithmetic operations with them. For example, what would the sum of two people’s zip codes mean? These numbers are simply labels placed on a finite list of postal delivery zones. They carry no arithmetic meaning.

**Example 1.1.7** Neither human hair color nor eye color, as considered in Example 1.1.3, would be considered numerical. The values these variables take on (such as “brown” or “blue”) are not real numbers.

We now introduce another important type of variable.

**Definition 1.1.8** A variable is called **categorical** if it can take on only finitely many different values and if the values that it can take on do not make sense to perform arithmetic with. A value that a categorical variable can take on is called a **level**.

Once again, some illustrations of this definition will be helpful. Note

\(^1\)Real numbers as distinguished from complex and other types of numbers.
first though that a variable cannot be both numerical and categorical because of the requirement that the values that a categorical variable can take on do not make sense to perform arithmetic with, while the values of a numerical variable must make sense to perform arithmetic with.

**Example 1.1.9** The gender of a human being is a categorical variable. This variable can take on only finitely many possible values (female and male), and it does not make sense to perform arithmetic with these values. For example, what would the sum of the genders of two human beings mean? To use the additional terminology that we have defined, this categorical variable has two levels, female and male.

**Example 1.1.10** The height of a tree is not a categorical variable for two reasons. First, its values are real numbers that can be used arithmetically (making it a numerical variable). Second, it could theoretically take on any of an infinite number of possible values. For example, there are infinitely many values even just between 2 meters and 3 meters that it might assume, since there are infinitely many real numbers between 2 and 3.

Of course, we would only measure the height of a tree with a certain degree of precision (maybe in meters to 3 decimal places or so), but for modeling purposes, we still usually conceptualize such a variable as being able to take on any real number, within a certain range.

**Example 1.1.11** If we are studying hair color in human beings as in Example 1.1.3, we would ordinarily treat this as a categorical variable. To do so though, we would first have to decide on a finite list of possible hair colors (such as: black, blond, brown, gray, and red) and agree upon how to classify each person’s hair as one element of the list. In this case, for modeling purposes it is usually not convenient to consider human hair color as having an infinite list of possible values, even if it does.

Once we have settled on a method of classifying hair color into a finite list of possibilities, we can consider hair color to be a categorical variable, since this variable can take on only finitely many values when defined this way, and since it makes no sense to perform arithmetic operations with these values. For example, it doesn’t make sense to try to add black and brown hair colors in an arithmetic way.

**Example 1.1.12** Suppose we define a variable that indicates the gender of a human being. Rather than using the labels female and male, we decide to define the variable’s value to equal 1 for a female and 2 for a male. In spite of the fact that all the values of this variable are real numbers, this variable would not be considered numerical because performing arithmetic operations with its values is not meaningful. For example, it wouldn’t make sense to ask about the arith-
metic sum of the genders of two people as given by this variable. Such a sum can be computed, but does not mean much. Rather, the numbers 1 and 2 are simply arbitrary labels for the levels of this categorical variable.

Not every variable can be classified as either numerical or categorical, but these two classes of variables are useful to be able to identify. Furthermore, within the class of numerical variables, two useful subclasses can be distinguished, continuous and discrete numerical variables. We begin with the subclass of continuous numerical variables.

**Definition 1.1.13** A numerical variable is called **continuous** if the set of its possible values forms an interval (possibly extending to $\pm\infty$) or collection of intervals in the real numbers.

Less illustrate this with some examples.

**Example 1.1.14** The height of a tree would ordinarily be considered a continuous numerical variable since the set of its possible values can be taken (for modeling purposes) to be an interval from 0 to some maximum tree height.

**Example 1.1.15** The number of squirrels observed in a particular area in a given time period is a numerical variable, but not a continuous numerical variable. Its possible values are all nonnegative integers and, as such, do not form an interval or collection of intervals.

Another type of numerical variable worth distinguishing is the discrete numerical variable.

**Definition 1.1.16** A numerical variable is called **discrete** if its possible values can be written out individually in a (possibly infinitely long) list.

Note that it is not possible to write out all of the elements of an interval in the real numbers. As such, no numerical variable can be both continuous and discrete. Not every numerical variable can be classified as continuous or discrete though. Some numerical variables are neither continuous nor discrete.

It is worth mentioning that any numerical variable with only finitely many possible values is discrete, since any finite set of values can be listed out.

We now give some examples of discrete numerical variables.

**Example 1.1.17** The number of squirrels observed in a particular place during a given time period is a discrete numerical variable. While the list of possible values

---

2 This may not be obvious, but it can be proved using an interesting argument known as Cantor's diagonalization method.
might be infinitely long, it is easy to see a form the list could take: 0, 1, 2, 3, 4, ... . In practice, there would be a maximum number of such squirrels, making the list finite. For modeling purposes though, we usually think of the list as being infinite.

Example 1.1.18 If we are thinking of the height of a tree as possibly taking on any nonnegative real number of meters, this numerical variable is not discrete because the set of its possible values contains an interval of real numbers. As remarked in Example 1.1.14, this numerical variable is actually continuous, so by the above remarks, we know it is not discrete anyway.

**Random variables**

We now examine another type of numerical variable that is particularly important in statistics, especially in linear regression analysis, the random variable.

**Definition 1.1.19** A random variable is a numerical variable whose value depends solely on the outcome of some random process. Being numerical variables, random variables can be continuous or discrete (or neither). We refer to the execution of a random process as a trial of the process. Conducting a trial of a random process to obtain a value of a random variable is called measuring or observing the random variable. The value of a random variable resulting from a trial of its associated random process is called a measurement or observation of the random variable.

This is really only an imprecise working definition, but it is the one we will use throughout the book. Rather than becoming sidetracked and bogged down in the details of all the terms used rather informally in this definition, we refer to the interested reader to a text on probability theory for a more careful and precise definition of a random variable. The definition we have given here will be adequate and suitable for our purposes in this book.

Implicit in this definition are two elements that are required to specify a random variable: the random process on which it is based and the rule for how the value is obtained from the outcome of the random process. We now give some examples to show how this works.

**Example 1.1.20** Suppose that we are studying trees on campus at the University of Puget Sound. The height of a randomly selected tree from the campus is

---

3 “Random” is not the same as “haphazard”. In this context, “random” is being used
a random variable. It takes on a numerical value based solely on the outcome of the random process of selecting a tree. The rule for determining the value is simply to measure the height of the selected tree. As mentioned in Example 1.1.21, this random variable is continuous.

**Example 1.1.21** Suppose that we are studying the gender of students currently enrolled in the introductory statistics course at the University of Puget Sound. The gender of a randomly selected student from this course is not a random variable because it is not numerical. However, there are many ways to define a random variable to indicate such a student’s gender. For example, working alphabetically, we might define a random variable that equals 1 if a randomly selected student is female and 2 if the randomly selected student is male. This gives a well-defined rule for determining the value of the variable based on the random process of randomly selecting a student enrolled in the class. This random variable is discrete, since it takes on only finitely many possible values.

Of course, we could have chosen other numbers instead of 1 and 2 (0 and 1 being another common choice), which is one reason why it is important to specify exactly which random variable is being used.

Now that we are familiar with the definition and basic classification of random variables, we turn our attention to an important concept in probability and statistics: the distribution of a random variable.

**Definition 1.1.22** For each measurable subset of the real numbers, there is a probability with which a given random variable assumes a value in that subset. The distribution of a random variable is the set of all such pairs of subsets and probabilities.

Since the set of all measurable subsets of the real numbers can’t even be enumerated, we need a better way to describe the distribution of a random variable than just listing out each possible subset and its associated probability. The resolution of this difficulty is easier for discrete random variables than for continuous ones, so we begin there.

**Definition 1.1.23** The probability mass function of a discrete random variable is a function \( \text{Prob} \) that gives the probability \( \text{Prob}(x) \) with which the random variable assumes each real number \( x \). Probability mass functions are often given as a table of values and probabilities, with the assumption to mean that each tree is equally likely to be selected. The interested reader might think about how to implement such a selection process in practice.

---

4 To be completely correct, we should specify “For each measurable subset” in this definition, but we will not get bogged down in such technicalities here.

5 Again, this should really be “all measurable subsets”.
that all values not listed are assumed with probability 0.

The probability mass function of a discrete random variable completely determines the distribution of the random variable (and vice versa of course). To see this, note that to determine the probability that the random variable will assume a value in a given subset of the real numbers, we need only sum the probabilities of all the values with nonzero probability in the subset. Consequently, the usual way of specifying the distribution of a discrete random variable is to give its probability mass function.

We now give perhaps the simplest example of a distribution (or of a probability mass function) of a discrete random variable.

**Example 1.1.24** Suppose that we are able to flip a coin in such a way that either side ("heads" or "tails") has probability 0.5 of landing face up. We can then define a random variable whose value is 1 when "heads" lands face up and 0 when "tails" lands face up. The distribution of this discrete random variable is given by the following probability mass function:

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.5</td>
</tr>
<tr>
<td>1</td>
<td>0.5</td>
</tr>
</tbody>
</table>

As usual, values not listed have probability 0. ♦

This example can be generalized slightly to give an important type of discrete random variable distribution.

**Definition 1.1.25** A Bernoulli distribution is the distribution whose probability mass function is given by the following table, for some real number p between 0 and 1:

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1 − p</td>
</tr>
<tr>
<td>1</td>
<td>p</td>
</tr>
</tbody>
</table>

A random process with exactly two possible outcomes is called a Bernoulli trial. These two possible outcomes are typically generically labeled success and failure, irrespective of whether or not these labels are apt in their non-technical senses. The probability of success is usually denoted by p, making the probability of failure equal to 1 − p.
Figure 1.1.27. A board game spinner, equally likely to land in any direction when spun.

A random variable taking on the value 0 when the outcome of some Bernoulli trial is failure and 1 when it is success has a Bernoulli distribution and is called a Bernoulli random variable.

The prototypical example of a Bernoulli trial is the flip of a coin, where “heads” is arbitrarily chosen as success, and where the probability of success is denoted by \( p \). (For a fair coin, \( p \) would equal 0.5.) Correspondingly, the prototypical example of a Bernoulli random variable is the one associated with this Bernoulli trial.

The distribution of a continuous random variable is more difficult to describe than that of a discrete random variable. To understand why we can’t simply use probability mass functions for continuous random variables as well, let’s consider the following example.

**Example 1.1.26** Suppose we have a board game spinner, meaning an arrow with its tail attached to the center of a circle to pivot on and its tip pointing radially outward through the circle, as pictured in Figure 1.1.27. Also suppose that if we spin this spinner, it is equally likely to point in any direction on the circle when it stops. If we make a reference marking somewhere on the circle and call it 0, then we can define a random variable (determined by the random process of spinning
the spinner) as how far around the circle counterclockwise from the reference point the spinner lands, expressed as a fraction between 0 and 1 (including 0 but not 1).

We might ask what the probability is that this random variable will (exactly) equal some given value, say 1/3. Well, according to our understanding of probability from the discrete case, since each outcome is equally likely and since there are infinitely many possible outcomes, the probability that this random variable will equal 1/3 is $1/\infty$, which is 0. Something is wrong with this picture! ♠

It turns out that the phenomenon illustrated by this example is true for continuous random variables in general: the probability that a continuous random variable will take on any particular value is always 0. Since this conveys essentially no helpful information, we need a different way to describe the distribution of a continuous random variable.

What is wrong here is that we are actually asking the wrong question. Instead of asking what the probability is that a continuous random variable will take on a particular value, we should ask what the probability is that the value of a continuous random variable will be between two given values.

For example, with the spinner, we get a much more informative answer if we ask: what is the probability that this random variable will be between 1/2 and 3/4? The answer to this question is 1/4, since the spinner is equally likely to stop in any direction and since 1/4 is the fraction of the circle between 1/2 and 3/4 of the way around it.

Not only is the right question to ask different from the discrete case, but the way we encode the answer to this question (and hence the notion of the distribution of a random variable) is different as well. For continuous random variables, the probability of being between two given values is encoded as the area underneath a particular curve. This brings us to the following definition.

**Definition 1.1.28** A **probability density function** of a continuous random variable is a function with the property that the area under its graph be-tween any two values on the $x$ axis is equal to the probability that this random variable will be between those two values.

The probability density function of a continuous random variable com-

---

6By "area under its graph", we mean what is usually meant in calculus. That is, we mean the area of the region lying between the graph of the function and the $x$ axis, where any part of the region below the $x$ axis contributes a negative amount to this area.
Chapter 1  Random variables

Figure 1.1.30. The graph of the probability density function for the spinner, shown by the dark black lines.

pletely determines the distribution of the random variable (and vice versa of course). However, this isn’t obvious or even easy to prove without calculus (not to mention some measure theory and so forth), so we will not prove it here. We will, however, use it. Consequently, the usual way that we will use to describe the distribution of a continuous random variable will be to give its probability density function.

We illustrate this definition with the spinner example that we have been considering.

Example 1.1.29 It is not difficult to verify that the probability density function $f$ of the random variable in the spinner example is given by:

$$f(x) = \begin{cases} 
1 & \text{if } 0 \leq x < 1 \\
0 & \text{otherwise.}
\end{cases}$$

This is depicted in Figure 1.1.30. Notice how the area under the graph of this function between any points on the $x$ axis is in fact the probability that the spinner will land in that range.

There are various things to note about probability density functions. First of all, how to find the probability density function of a given random
variable should not necessarily be at all obvious to the reader in any but
the simplest cases. Determining probability density functions on theoretical
grounds is included in the subject of probabilistic modeling. In addition, even if you have an idea of what the probability density function
should look like, it may take some calculus to fit it correctly. In this text
then, we will not dwell on how to find probability density functions but
will instead examine random variables whose probability density func-
tions are given to us.

Two basic properties of probability density functions are:

1. Probability density functions are nonnegative, since negative values
would lead to negative areas and hence negative probabilities.

2. The total area under a graph of a probability density function (going
all the way to $-\infty$ on the left and $\infty$ on the right) equals 1.

These properties can actually be used to characterize or define proba-
bility density functions.

**Working with distributions**

Probability mass functions and probability density functions are not the
only way to specify distributions of random variables. Another common
method of doing so is with the *cumulative distribution function*.

**Definition 1.1.31** The **cumulative distribution function** $F$ of a random
variable $X$ is defined by:

$$F(x) = \text{Prob}(X \leq x).$$

The cumulative distribution function and the probability density func-
tion both encode the same information (namely what the distribution of
the random variable is). In terms of the probability density function $f$, the
cumulative distribution function $F$ can be written as:

$$F(x) = \text{area under the graph of } f \text{ to the left of } x = \int_{-\infty}^{x} f(u) \, du.$$ 

The last expression above is not necessary to know, but is included for
those familiar with some calculus.
The probability density function can be written in terms of the cumulative distribution function, although this requires some calculus:

\[ f(x) = \frac{dF}{dx}. \]

For those not familiar with calculus, it is enough to know that the probability density function can be written in terms of the cumulative distribution function. Since both functions can be written in terms of each other, the both contain the same information about the random variable, just in different ways.

**Example 1.1.32** For the random variable associated with the board game spinner discussed in Example 8, we found that the probability density function is given by the step function discussed in Example 10. By direct inspection of the probability that this random variable will be less than an arbitrary value \( x \), we also find that the cumulative distribution function is given by

\[
F(x) = \begin{cases} 
0 & \text{if } x \leq 0 \\
x & \text{if } 0 < x \leq 1 \\
1 & \text{if } x > 1.
\end{cases}
\]

The graph of the cumulative distribution function is depicted in Figure 1.1.33. ♦

Most statistical computer programs have functions to compute the more common cumulative distribution functions. We will see in the next chapter how cumulative distribution functions are often the easiest way to compute probabilities of interest, which is an advantage they have over probability density functions. The main disadvantage of cumulative distribution functions, however, is that their graphs are not as simple or intuitive as those of probability density functions when it comes to conveying information about probability distributions.

Another function that is closely related to the cumulative distribution function is the quantile function.

**Definition 1.1.34** The quantile function \( F \) of a continuous random variable \( X \) is defined for any \( 0 < p < 1 \) by:

\[ F^{-1}(p) = \text{the real number } x \text{ such that } \text{Prob}(X \leq x) = p. \]

The value of \( F^{-1}(p) \) is not defined for any \( p \) not satisfying \( 0 < p < 1 \).

The number \( F^{-1}(p) \) is called the random variable \( p \)-quantile of \( X \).
The quantile function $F^{-1}$ can also be defined similarly for discrete random variables, although one must be careful when the quantiles fall on values that have nonzero probability. We will not get sidetracked with such details, since we will usually use the quantile function for continuous random variables here. Rather, we simply note that the definition of the quantile function is conceptually the same for discrete random variables.

By definition, the quantile function $F^{-1}$ is the inverse of the cumulative distribution function $F$, in the sense that:

$$F^{-1}(p) = x \quad \text{if and only if} \quad F(x) = p.$$ 

This explains the notation $F^{-1}$ that we use for the quantile function, and it also tells us that the graph of the quantile function should be the region of the graph of the cumulative distribution function with $0 < y < 1$, reflected across the line $x = y$. Since the graph of the quantile function isn’t of much importance for our purposes, we will not go into this relationship any further.

We will, however, explore other aspects of the quantile function in greater detail in the next chapter when there is good reason to do so. For the present, we merely note the definition and that most statistical com-
puter programs have the quantile functions of the most commonly used distributions built in.

Samples

Random variables and their distributions are theoretical constructs, models of real-world situations. Consequently, we have no direct way of determining them. The only way we have to get at them is by collecting observations. That is, we conduct trials of the random procedure that generates an outcome that determines a value for the random variable, and by analyzing the values that result in this manner, we can draw probable conclusions about different aspects of the model.

In other words, we essentially never handle the random variables and the distributions themselves. We must get at them by analyzing real data. In this section, we introduce some basic techniques to do exactly that.

We begin by defining samples of random variables.

**Definition 1.1.35** A sample of a random variable is a finite collection of independent observations of that random variable. The number of observations in the collection is called the size of the sample. Also, for each value in a given sample, there is a proportion of observations in the sample that are equal to that value. The distribution of a sample is the set of all such pairs of values and proportions.

Note that the observations must be independent, in the sense that the random procedure producing one observation is not allowed to affect the outcome of the random procedure producing any other observation of the random variable in the collection.

**Example 1.1.36** As an example of this definition, suppose that the random variable of interest is the number of “heads” landing face up in three flips of a particular coin. To generate a sample of size 5 of this random variable, we would repeat 5 times the procedure of flipping the coin 3 times. Denoting “heads” by $H$ and “tails” by $T$, the list of outcomes of the random procedure might look something like: \{HHT, HTH, HTH, HHH, HTH\}. This would then give the following list of values of the random variable: \{2, 2, 2, 3, 2\}, which would be our sample.

Notice that the sample is the collection of values of the random variable resulting from the trials, not the collection of outcomes of the trials themselves.

**Definition 1.1.37** The empirical probability mass function of a sample is
the function $\text{Prop}$ that gives for each real number $x$ the proportion $\text{Prop}(x)$ of the observations in the sample that are equal to $x$. Empirical probability mass functions are often given as a table of values and probabilities, with the assumption that all values not listed are assumed with probability 0.

Another way to think about the empirical probability mass function is that it is the probability mass function for the random variable whose value is determined by selecting a single value from the sample, with each of the observations of the sample being equally likely. The probability of choosing any given value then is simply the proportion of observations in the sample equal to that value.

The empirical probability mass function gives a complete description of the distribution of a sample. The various other methods that we have encountered for encoding information about the distribution of a random variable are also defined for samples, and work analogously.

**Definition 1.1.38** The empirical cumulative distribution function of a sample gives the proportion of observations in the sample that are less than or equal to the given input value.

Conceptually, the empirical quantile function of a sample gives for any $0 < p < 1$ the real number $x$ such that the proportion of observations in the sample that are less than or equal to $x$ equals $p$. For a more precise definition though, additional care is needed to establish what to do about values with multiple occurrences in the sample. We will go into such details only for certain cases and in a later section.

The output of the empirical quantile function with input $p$ is called the empirical $p$-quantile.

These functions are used in similar ways to their random variable counterparts.

### 1.2 Numerical summaries

Since a full description of the distribution of a random variable is often either not available or too complicated to be of much help, we frequently use numerical summaries to describe certain aspects of the distribution. Similarly, we often use numerical summaries to describe samples of random variables as well. We now describe some of the most commonly used summaries of random variable and their samples.
Numerical summaries of discrete random variables

The two most useful numerical descriptions of random variable distributions for our purposes here are the expected value, which is a measure of the center of a distribution, and the random variable standard deviation, which is a measure of the spread of the distribution. We begin with the expected value.

Definition 1.2.1 The expected value (also called the random variable mean) \( \mu[X] \) of a discrete random variable \( X \) that takes on values \( x_1, x_2, x_3, \ldots \) with probabilities \( p_1, p_2, p_3, \ldots \) is defined to be

\[
\mu[X] = p_1x_1 + p_2x_2 + p_3x_3 + \cdots .
\]

The following example illustrates this definition.

Example 1.2.2 As an example of computing a expected value, consider the random variable \( X \) that is the number of “heads” resulting from 10 flips of a coin for which the probability of “heads” on any flip equals 0.65. Although we have not shown it, the distribution of this random variable turns out to be given by:

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000028</td>
</tr>
<tr>
<td>1</td>
<td>0.00051</td>
</tr>
<tr>
<td>2</td>
<td>0.0043</td>
</tr>
<tr>
<td>3</td>
<td>0.021</td>
</tr>
<tr>
<td>4</td>
<td>0.069</td>
</tr>
<tr>
<td>5</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>0.24</td>
</tr>
<tr>
<td>7</td>
<td>0.25</td>
</tr>
<tr>
<td>8</td>
<td>0.18</td>
</tr>
<tr>
<td>9</td>
<td>0.073</td>
</tr>
<tr>
<td>10</td>
<td>0.013</td>
</tr>
</tbody>
</table>

(Note that the probabilities don’t add up to exactly 1 because of the way that they have been rounded.) Reading directly from this table, we find that:

\[
\mu[X] = 0.000028 \cdot 0 + 0.00051 \cdot 1 + 0.0043 \cdot 2 + 0.021 \cdot 3 + 0.069 \cdot 4 + 0.15 \cdot 5 + 0.24 \cdot 6 + 0.25 \cdot 7 + 0.18 \cdot 8 + 0.073 \cdot 9 + 0.013 \cdot 10
\]

\[= 6.5.\]
The expected value has a nice geometric interpretation: if we were to attach a small ball at each value \(x_i\) to a real number line, with mass proportional to \(p_i\), then the expected value \(\mu[X]\) would be the point on the number line at which the line, masses and all, would balance perfectly, either when placed on a fulcrum below or when hung by a string from above. For example, Figure 1.2.3 shows one way to illustrate the expected value for the random variable discussed above that is the number of “heads” resulting from 10 flips a coin whose probability of “heads” in any flip equals 0.65.

Here the expected value is indicated by the \(\times\), and we have drawn a fulcrum on which the line could pivot as well. Also, the size of the dot is designed to indicate the relative mass of the ball that is at that point.

It should be noted that there are other measures of center that can be used, such as the random variable median. We will postpone our discussion of these until we describe summaries of continuous random variables.

The second main numerical description of the random variable distribution that we use here is the random variable standard distribution, which we now introduce along with the closely related concept of the random variable variance.

**Definition 1.2.4** The random variable variance \(\sigma^2[X]\) of a discrete random variable \(X\) that takes on values \(x_1, x_2, x_3 \ldots\) with probabilities \(p_1, p_2, p_3 \ldots\) is defined to be

\[
\sigma^2[X] = p_1(x_1 - \mu[X])^2 + p_2(x_2 - \mu[X])^2 + p_3(x_3 - \mu[X])^2 + \cdots
\]

The random variable standard deviation \(\sigma[X]\) is defined to be the non-negative square root of the random variable variance:

\[
\sigma[X] = \sqrt{\sigma^2[X]}
\]
As usual, we illustrate this definition with an example.

**Example 1.2.5** As an example of computing a random variable standard deviation, again consider the random variable $X$ that is the number of “heads” resulting from 10 flips of a coin for which the probability of “heads” on any flip equals 0.65. We computed the expected value $\mu[X]$ for this random variable to be 6.5 above. This allows us to compute its random variable variance $\sigma^2[X]$ as:

\[
\sigma^2[X] = 0.000028(0 - 6.5)^2 + 0.00051(1 - 6.5)^2 + 0.0043(2 - 6.5)^2 \\
+ 0.021(3 - 6.5)^2 + 0.069(4 - 6.5)^2 + 0.15(5 - 6.5)^2 \\
+ 0.24(6 - 6.5)^2 + 0.25(7 - 6.5)^2 + 0.18(8 - 6.5)^2 \\
+ 0.073(9 - 6.5)^2 + 0.013(10 - 6.5)^2 \\
= 2.275.
\]

(Note that 2.275 comes from using unrounded values for the probabilities, which we have not written out above. With the rounding as written above, the result would be 2.273.)

From this, we compute that the random variable standard deviation $\sigma[X]$ equals:

\[
\sigma[X] = \sqrt{\sigma^2[X]} = \sqrt{2.275} = 1.508. \blacklozenge
\]

As with the definition of the expected value, the definition of the random variable standard deviation may appear to be somewhat arbitrary. As it turns out, there is some arbitrariness to it, and its form is partly just for mathematical convenience. On the other hand, we can provide some motivation for it as follows.

The reason for the term “$x_i - \mu[X]$” is apparent, since this term measures the deviation of the value $x_i$ from the “center” $\mu[X]$, but why square it? Well, some of the points are to the left of the center, and some of the points are to the right of the center, yet both types of points contribute to the overall deviation from the center. We wouldn’t want deviations to the left to cancel (even in part) deviations to the right. One way to avoid such cancellation is to square each term, as we have in the definition. (Another way to avoid such cancellation would be to take the absolute value of each term, but that turns out to be mathematically less convenient in a variety of ways.)

Why multiply each of these square terms by a probability though? This can be motivated in terms of game theory and economics. In such terms, the random variable variance (or the random variable standard deviation)
can be thought of as a measure of volatility. From this point of view, only a small contribution to the overall volatility of a random variable is made when a value that is far from the expected value occurs with only a very small probability. On the other hand, a large contribution to the overall volatility is made when such values occur with higher probability. Volatility results from the random variable “jumping around”. If the random variable is fairly stable and does not have highly deviant values occur very often, then it should have a small standard deviation. If the random variable “jumps around” a lot and takes on highly deviant values with high probability, then it should have a large standard deviation. This helps to explain why it makes sense to multiply each term by a probability in the definition of the standard deviation.

**Numerical summaries of continuous random variables**

The numerical summaries that we used for discrete random variables, namely the expected value (Definition 1.2.1 on page 16), random variable variance (Definition 1.2.4 on page 17), and random variable standard deviation (also Definition 1.2.4 on page 17), translate directly to continuous random variables. However, the definitions for the continuous case involve integrals (which are studied in calculus) rather than sums, so we do not dwell on the precise definitions here. Rather, we give them here primarily for the sake of completeness. For those who are familiar with such things, we remark that the continuous versions are arrived at from the discrete versions by replacing the probabilities $p_i$ of $x_i$ by the “infinitesimal probability” $f(x) \, dx$ at $x$ and changing the sums to integrals.

**Definition 1.2.6** Let $X$ be a continuous random variable whose values lie in the interval from $a$ to $b$ (where $a$ might be $-\infty$ and $b$ might be $\infty$) and whose probability density function is $f$. The **expected value** $\mu[X]$ (also called the **random variable mean**) of $X$ is defined to be:

$$\mu[X] = \int_a^b x f(x) \, dx.$$ 

The **random variable variance** $\sigma^2[X]$ of $X$ is given by:

$$\sigma^2[X] = \int_a^b (x - \mu[X])^2 f(x) \, dx.$$
The random variable standard deviation is defined as the non-negative square root of the random variable variance, exactly as in the discrete case:

\[ \sigma[X] = \sqrt{\sigma^2[X]} . \]

For our purposes here, the exact definition of these summaries for continuous random variables are not so important. However, it is important to know that the interpretation of these numerical summaries is essentially the same for continuous random variables as it is for discrete random variables. That is, the expected value is still a balance point, only now we have to imagine balancing the graph of a probability density function (with the mass at each point corresponding to the height of the graph at the point), and the random variable variance and random variable standard deviation are still both measures of how widely spread a distribution is (about its expected value).

We now introduce the random variable median, an another measure of center for continuous distributions. The definition of random variable median can also be adapted for use with discrete random variables, although we will not go into the details here.

**Definition 1.2.7** The random variable 0.5-quantile is called the random variable median of \( X \).

Since the total area under a distribution equals 1, the 0.5-quantile has the property that exactly half of the area under the distribution of \( X \) lies in the region to left of it and exactly half to the right.

Random variable medians often require calculus to compute without a computer, although they can often be guessed or approximated by inspection of the graph of the distribution. Here is one such example.

**Example 1.2.8** Inspecting the graph of the uniform distribution given in Figure 1.1.30 on page 10, we can see that the random variable median in this case is equal to 0.5, since the region under the distribution to the left of 0.5 is a rectangle of height 1 and width 0.5.

While the random variable median is another measure of center for a distribution, the first and third quartiles give us some information about the spread of a distribution.

**Definition 1.2.9** The random variable 0.25-quantile is called the random variable first quartile of \( X \), and the 0.75-quantile is called the random variable third quartile of \( X \). The random variable interquartile range is
defined to be the random variable third quartile minus the random variable first quartile.

We now examine this definition applied to a very simple example.

**Example 1.2.10** Again inspecting the graph of the uniform distribution given in Figure 1.1.30 on page 10, we can see that the random variable first quartile in this case is equal to 0.25, since the region under the distribution to the left of 0.25 is a rectangle of height 1 and width 0.25 and hence area 0.25. Similarly, we see that that random variable third quartile is equal to 0.75, so the random variable interquartile range is 0.5.

We will investigate random variable quartiles further in the next chapter, in the context of normal and $t$ distributions.

**Numerical summaries of samples**

As with distributions of discrete random variables, the two main numerical summaries that we use here for distributions of samples of discrete random variables are: the **sample mean**, which is a measure of the center of the distribution of the sample, and the **sample standard deviation**, which is a measure of the spread of the distribution of the sample. The definitions of these are similar to those for distributions of discrete random variables, but there are some small differences.

**Definition 1.2.11** The **sample mean** $\bar{x}$ of a sample of size $n$ taking on distinct values $v_1, v_2, \ldots, v_m$ in proportions $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m$ is defined to be:

$$\bar{x} = \hat{p}_1 v_1 + \hat{p}_2 v_2 + \cdots + \hat{p}_m v_m.$$ 

If the complete list of (not necessarily distinct) values in the sample is $x_1, x_2, \ldots, x_n$, then another way to write this formula (which is more useful when computing sample means by hand) is:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$ 

The first form of the definition makes the parallel between the sample mean and the random variable mean (or expected value) clear, which is why we have given it first. It shows that the sample mean can also be interpreted as a “balance point” of the sample. However, when computing the sample mean by hand, the second form of the definition is generally more useful. A little bit of algebraic manipulation shows that the two forms of the definition are the same, as we illustrate with an example.
EXAMPLE 1.2.12 To see that the two forms really are the same, we continue with the example in which the random variable is the number of “heads” resulting from three flips of a particular penny, using the sample \{2, 2, 1, 3, 0, 3, 1, 1, 1, 2\}. Starting with the second form of the sample mean, we have:

\[
\bar{x} = \frac{2 + 2 + 1 + 3 + 0 + 3 + 1 + 1 + 1 + 2}{10} = \frac{0 + 1 + 1 + 1 + 2 + 2 + 2 + 3 + 3}{10} = \frac{1 \cdot 0 + 4 \cdot 1 + 3 \cdot 2 + 2 \cdot 3}{10} = \frac{1}{10} \cdot 0 + \frac{4}{10} \cdot 1 + \frac{3}{10} \cdot 2 + \frac{2}{10} \cdot 3 = (0.1) \cdot 0 + (0.4) \cdot 1 + (0.3) \cdot 2 + (0.2) \cdot 3,
\]

which is the first form of the sample mean. Either way, of course, the result is that \(\bar{x} = \frac{16}{10} = 1.6\).

There are a couple of important things to note about the sample mean. First of all, the sample mean does not have to be a value that can be assumed by the random variable. It also does not have to be a value that occurs anywhere in the sample, as this example shows. In addition, different samples (even of the same size) of the same random variable will generally have different sample means. The sample mean is related to the random variable mean (or expected value), but we will discuss that in a later section.

The definitions of the sample variance and sample standard deviation have a small extra complication in them, as compared to the corresponding definitions of the random variable variance and the random variable standard deviation.

**Definition 1.2.13** For a sample of size \(n\) consisting of distinct values \(v_1, v_2, \ldots, v_m\), define numbers \(\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_m\) by:

\[
\hat{p}_i = \frac{\text{number of times } x_i \text{ occurs in the sample}}{n - 1}.
\]

Then the sample variance \(s^2\) is defined to be:

\[
s^2 = \hat{p}_1(v_1 - \bar{x})^2 + \hat{p}_2(v_2 - \bar{x})^2 + \cdots + \hat{p}_m(v_m - \bar{x})^2.
\]
If the complete list of (not necessarily distinct) values in the sample is \(x_1, x_2, \ldots, x_n\), then this formula can also be written as:

\[
s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_m - \bar{x})^2}{n - 1}.
\]

As before, the **sample standard deviation** \(s\) is simply the nonnegative square root of the sample variance:

\[
s = \sqrt{s^2}.
\]

The strange aspect of these definitions is the \(n - 1\) in the denominators. If it were \(n\) in the definition of \(\hat{p}_i\) instead of \(n - 1\), then the result would be simply proportion \(\hat{p}_i\) of the value \(v_i\) in the sample, and the definitions of the sample variance and sample standard deviation would look exactly as you would expect them to judging from the corresponding definitions of the random variable variance and the random variable standard deviation, only substituting “proportion” for “probability”. However, there are various technical reasons for having \(n - 1\) instead of \(n\) which we will not go into here. Suffice it to say that when the sample size \(n\) is large, it makes little difference whether we divide by \(n - 1\) or \(n\).

**Example 1.2.14** To continue with the penny-flipping example, we compute the variance using the second form of the definition (although we put the \(n - 1\) denominator in front for layout reasons):

\[
s^2 = \frac{1}{10 - 1} \left( (2 - 1.6)^2 + (2 - 1.6)^2 + (1 - 1.6)^2 + (3 - 1.6)^2 + (0 - 1.6)^2 + (3 - 1.6)^2 + (1 - 1.6)^2 + (1 - 1.6)^2 + (2 - 1.6)^2 \right)
\]

\[
= .933,
\]

meaning that the standard deviation is \(s = \sqrt{.933} = .966\). The fact that the first form of the definition is equivalent can be seen using the same method used in showing that the two forms of the definition of the sample mean were equivalent, so we will not write it out here. ♦

The definition of the sample median is almost exactly like that of the random variable median, except that a little more care needs to be taken in finding the “middle” of the sample because samples are necessarily discrete. The sample median is, roughly speaking, the value for which half the given sample is smaller and half larger than it. In other words, the
Chapter 1  Random variables

The sample median is obtained by first listing all the values in the sample in order from lowest to highest and then selecting the value halfway down the list.

**Definition 1.2.15** The empirical $0.5$-quantile is called the **sample median**. To compute the sample median for a sample of sample size $n$, list the values in the sample in order from lowest to highest, allowing values that occur multiple times to be listed multiple times and ties to be listed in any order. The **sample median** of the sample is defined to be the value in the $(n + 1)/2$ position in the ordered list if $n$ is odd and the sample mean of the values in the $n/2$ and $(n/2) + 1$ positions in the ordered list if $n$ is even.

We should note that the adjectives *sample* and *empirical* mean the same thing in this context, and you will often see them used interchangeably. We will try to employ the more widely-used term in each case here, even though that makes for some inconsistency in notation.

The sample median requires lots of words to define, but it is conceptually straightforward and not difficult to compute, as we illustrate with the following example.

**Example 1.2.16** To compute the sample median of a sample \{2, 0.4, −3, 11.2, 11.2\}, we first order the values in the sample from smallest to largest:

\[-3, 0.4, 2, 11.2, 11.2.\]

The sample median then is the middle value in the list, which is 2 in this case.

To compute the sample median of a sample \{−1, 2.2, 1.1, 4\}, we again first order the values in the sample from smallest to largest:

\[-1, 1.1, 2.2, 4.\]

Since the sample size is even, there is no middle value in the list, so the median equals the sample mean of the middle two values in the list, meaning that the median equals:

\[
\frac{1.1 + 2.2}{2} = 1.65. \blacklozenge
\]

The sample median is another way besides the sample mean to measure where the “center” of the distribution of a sample is. The biggest difference between these two measures of center is that the sample median does not take into account the magnitude of the differences between values in the sample; instead, it only uses which values are larger and smaller.
Consequently, if the largest value in a sample were made much larger, the mean would increase greatly, but the sample median would remain unchanged because the order of the values would remain unchanged. If a sample has a value which is much larger (or much smaller) than the other values in the sample, that value will have very little effect on the sample median but will make the sample mean much larger (or smaller) than it otherwise would have been. In statistical terms, values that are much larger or much smaller than other values in the sample are a type of outlier, a term which we will later define. Because of this terminology, we say that the sample mean is sensitive to outliers, and the sample median is (highly) insensitive to outliers.

Similarly to the way we define the sample median, or halfway point in the sample, we can also define other points in the sample, specifically the quarter point and the three-quarter point. Roughly speaking, the sample first quartile is the value for which one quarter of the values in the sample are smaller and three quarters are larger than it. Also roughly speaking, the sample third quartile is the value for which three quarters of the values in the sample are smaller and one quarter are larger than it.

**Definition 1.2.17** The empirical 0.25-quantile is called the **sample first quartile**. More precisely, the sample first quartile is the sample median of the half of the sample consisting of the smallest values (removing the middle value for odd sample sizes).

The empirical 0.75-quantile is called the **sample third quartile**. More precisely, the sample third quartile is the sample median of the half of the sample consisting of the largest values (again removing the middle value for odd sample sizes).

The **sample interquartile range** is defined to be the sample third quartile minus the sample first quartile.

As with the sample median, the adjectives sample and empirical are interchangeable in this context, and we are simply using the more common terms in each case.

We now compute these in the context of the previous example.

**Example 1.2.18** To compute the sample first and third quartiles of a sample \{2, 0.4, −3, 11.2, 11.2\}, we first order the values in the sample from smallest to largest:

\[−3, 0.4, 2, 11.2, 11.2\].

Since the sample size is odd, we remove the middle value of 2. The sample first
quartile then is the sample median of the lower half \{-3, 0.4\}, which is \(-1.3\), and the sample third quartile is the sample median of the upper half \{11.2, 11.2\}, which is 11.2. (Note that we really do mean the sample \textit{median} of each half here. It only looks like the sample mean in this case because these halves consist of only two values.)

The sample interquartile range is the sample third quartile minus the sample first quartile, which is \(11.2 - (-1.3) = 12.5\) for this sample.

The sample interquartile range is a measure of how widely spread a sample is. Unlike the sample variance (and sample standard deviation), which also measures how widely spread a sample is, the sample interquartile range is highly insensitive to outliers. Just as with the sample median, the sample interquartile range remains unchanged if the largest observation is increased, for example. On the other hand, because the sample variance involves the square of the difference between values and the sample mean, it is very sensitive to outliers.

1.3 Graphical summaries

Having discussed numerical methods of summarizing distributions of random variables and samples, we now turn our attention to graphical methods. With the ready availability of powerful computers, graphical methods have become increasingly more accessible and useful in summarizing distributions of random variables and samples.

Graphical summaries of random variables

The main graphical method of displaying the distribution of a discrete random variable doesn’t really have a standard name in the literature, so we will simply call it a \textit{linegram}. A \textit{linegram} is constructed by placing a vertical line segment at each point on the real number line where the probability of the random variable assuming that value is nonzero. The height of this vertical line segment should equal the probability that the random variable will assume that value.

\textbf{Example 1.3.1} The distribution of the random variable that is the number of “heads” resulting from 10 flips of a coin for which the probability of “heads” on any flip equals 0.65 is given in the table in Example 1.2.1 on page 16. We give the table again here, and to the right of which is a linegram of the distribution.
Section 1.3 Graphical summaries

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.000028</td>
</tr>
<tr>
<td>1</td>
<td>0.00051</td>
</tr>
<tr>
<td>2</td>
<td>0.0043</td>
</tr>
<tr>
<td>3</td>
<td>0.021</td>
</tr>
<tr>
<td>4</td>
<td>0.069</td>
</tr>
<tr>
<td>5</td>
<td>0.15</td>
</tr>
<tr>
<td>6</td>
<td>0.24</td>
</tr>
<tr>
<td>7</td>
<td>0.25</td>
</tr>
<tr>
<td>8</td>
<td>0.18</td>
</tr>
<tr>
<td>9</td>
<td>0.073</td>
</tr>
<tr>
<td>10</td>
<td>0.013</td>
</tr>
</tbody>
</table>

Figure 1.3.2. The distribution and corresponding linegram of the random variable that is 10 flips of a fair coin.

Of course, for very small distributions it is probably easier to interpret the associated table of numbers than the linegram. However, for large distributions with many outcomes, it may be nearly impossible to interpret the table directly, whereas the linegram may be very informative.

Just about the only method of displaying the distribution of a continuous random variable is simply to graph the probability density function, so there isn’t anything new to discuss here about graphical displays of continuous random variables.

Graphical summaries of samples

The main graphical tool for displaying samples of discrete random variables is simply a linegram whose line segment heights are determined by the proportion of the sample that assumes each possible value. While this will in general not be identical to the linegram of the distribution of the random variable itself, it introduces no new concepts. Since, in addition, we will not have much use for it in this text, we will not dwell on this topic further here.

The graphical display of samples of continuous random variables is,
however, highly useful but quite different. Because of the continuous range of possible values, a linegram is not suitable for displaying the distribution of a continuous random variable. By the same reasoning we used in the spinner example (Example ?? on page ??) earlier in this chapter, the probability of obtaining exactly the same value more than once in a sample of the continuous random variable is 0. In practice though, we only have a certain amount of precision in our data collection process, so it does happen sometimes that we have the same value more than once in a sample, but still not enough to make for a useful data display in a linegram.

However, if we extend the rounding process to a somewhat larger scale, we do obtain a useful display. The precise method of extending this rounding process is given in the following definition.

**Definition 1.3.3** Suppose we divide the real number line into intervals of equal length called bins. A histogram is a graphical display similar to a linegram except that each vertical bar of a histogram has a width that spans the entire bin. Also, the height of each bar of a histogram can be given either by frequency (the number of observations in the sample with values in each bin) or proportion (scaled so that the area of the bar — width times height — equals the proportion of observations in the sample with values in each bin). In this text, unless we specifically refer to a frequency histogram, all histograms will display proportion as the area of the bars.

Of course, the best way to learn about histograms is to see some rather than to have them described, so we give two in Figure 1.3.4. These histograms illustrate one of the fundamental difficulties with histograms, namely that there is no single “best” way to determine the set of bins (including the bin width) for a histogram. In fact, there may not be a single “best” set of bins for a histogram of a given sample. However, we should note that thin bins will generally lead to a “bumpier” histogram, and wide bins will generally lead to a “flatter” histogram. Consequently, if the bins are too thin, interesting features of the distribution may be lost in the bumpiness; if the bins are too wide, interesting features of the distribution may be lost in the flatness. Some experimentation is often necessary, although as a rule of thumb, having approximately 10 to 12 bins is often a good place to start.

In using a histogram to gain information about the distribution of a continuous random variable, there are four main things to look for.
Section 1.3  Graphical summaries

**Figure 1.3.4.** Two histograms with different bin widths, both displaying the same sample of size 462 of a random variable whose value is the mass of an (approximately) randomly selected Plain M&M.

Things to look for in a histogram

1. **Shape.** Is the histogram symmetric, or the same when reflected across some vertical line? How many modes, or values where the histogram reaches the top of a large hill, does the histogram have? (Small bumps do not count as large hills. It is rare for distributions to have more than 2 or 3 modes.) A distribution having exactly one mode is called unimodal, a distribution having exactly two modes is called bimodal, etc.

2. **Center.** Approximately where are the mean and median of the distribution?

3. **Spread.** Is the distribution widely spread out about its center (corresponding to a large sample variance) or only narrowly spread out about its center (corresponding to a small sample variance)?

4. **Outliers.** Are there any values that don’t seem to fit the overall pattern? Such values are called outliers.
30 Chapter 1 Random variables

While histograms are very good for displaying the distribution of a single sample, they are difficult to compare among different samples because it is hard to line them up next to each other, especially when the bins are not the same (and different bins might be useful for different samples). For comparing multiple samples, we have another method of graphical display.

**Definition 1.3.5** A boxplot is a type of diagram made up of a box, two vertical lines called whiskers extending from the top and bottom of the box, and possibly some dots above and/or below the whiskers. The width of the box is arbitrary and is chosen for clarity of display. The x axis is labeled with the name of the random variable whose sample is being displayed. The y axis is a scale to display the values of the random variable in the sample (with the appropriate units). The bottom and top of the box are placed at the sample first and third quartiles. A horizontal line is drawn through the box at the sample median. There are variations on how far the whiskers should extend. We will use the convention that they should extend to the value in the sample that is most nearly less than 1.5 times the sample interquartile range above the top of the box (on top) and below the bottom of the box (on the bottom). Any points in the sample that lie above the top whisker or below the bottom whisker should be plotted individually as dots.

As with histograms, the best way to understand boxplots is to look at some. For this, we plot two of them side by side in Figure 1.3.6. The figure depicts the times to complete a particular exam for students in two sections of an introductory statistics course at the University of Puget Sound in 2008. Note that although these times are for all the actual students in the classes, they are thought of as being samples of a random variable that gives the test time for a student chosen at random from among a theoretical (fictitious) population of possible students in the two sections.

There are several things to notice in this picture. First, it is simple to display two or more boxplots side by side, and doing so greatly facilitates comparison of the distributions of the various samples. In this particular example, it appears that students in Section A seem to have been taking longer on this test than those in Section B. Second, each boxplot displays very clearly what is known as the **five number summary** of the sample: the minimum value, the sample first quartile, the sample median, the sample third quartile, and the maximum value. This is a handy feature of the
Figure 1.3.6. Boxplots displaying samples of a random variable whose value is the mass of an (approximately) randomly selected Plain M&M of a given color, sample sizes from left to right being 87, 86, 92, 75, 58, and 64.

The main weakness of boxplots is that they don’t work as well for samples whose distributions are not unimodal. In such cases, much of the relevant information is lost inside the box, where the points are not plotted individually.

For fairly small sample sizes, there is another graphical display that gets around some of the shortcomings of boxplots, namely the stripchart.

**Definition 1.3.7** A **stripchart** is a graphical display in which all the values in a sample are plotted in a (usually vertical) line. The position of each point plotted is determined by the value corresponding to it in the sample, as on a (vertical) real number line.

Stripcharts can be useful for examining individual samples, but they are most commonly used to compare several samples side-by-side. As
with the other graphical displays, a picture to illustrate the definition, as given on the left in Figure 1.3.8, is helpful. Boxplots of the same samples are given on the right for the purpose of comparison. The sample sizes here allow for both stripcharts and boxplots. Smaller sample sizes would probably make stripcharts preferable; larger sample sizes would probably make boxplots preferable.

1.4 Multiple random variables

Now that we have considered how to investigate single random variables, we turn our attention to investigating more than one random variable at a time. To keep the presentation simple, we will only state the definitions and results for two random variables. However, all of the definitions and results in this section can be readily extended to more than two random variables. We will often do so without further elaboration throughout the text.

Also, while we will use multiple continuous random variables later in the text, it will be preferable in introducing the upcoming new concepts to give examples only with discrete random variables. This will reduce the number of additional complications for the present.
Joint random variables

When considering pairs of random variables simultaneously, we must take into account all possible pairs of values that they might assume. This leads us to the following definition.

Definition 1.4.1 The joint random variable built from two random variables has as its possible values the set of all ordered pairs of values assumed by the two random variables. The random process underlying the joint random variable consists of the two random processes underlying the individual random variables.

The joint distribution of two random variables is defined to be the distribution of the joint random variable built from them. For continuous random variables, the probability density function of the joint random variable can also be referred to as the joint probability density function of the two continuous random variables.

The joint random variable is aptly named, since it is used to consider more than one random variable jointly. We now give a simple example to illustrate this definition.

Example 1.4.2 Suppose that $X$ is a random variable whose value is equal to the result of the roll of a fair 6-sided die. Also, suppose that $Y$ is a Bernoulli random variable based on the flip of a fair coin with “heads” being defined as success. The distributions of $X$ and $Y$ considered individually are given by the following tables:

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/6</td>
</tr>
<tr>
<td>2</td>
<td>1/6</td>
</tr>
<tr>
<td>3</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>1/6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Value</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>1</td>
<td>1/2</td>
</tr>
</tbody>
</table>

The term ordered refers to the fact that the first value comes from the first random variable, and the second value comes from the second random variable, with the order never being switched.
Suppose further that the roll of the die and the flip of the coin are independent random processes. By elementary probability theory, this means that the probability of both a particular value of the die roll and a particular result from the coin flip (such as a 2 on the die and a “heads” for the coin) is 1/12, the product of the two individual probabilities.

The joint random variable built from $X$ and $Y$ has as its possible values the set of ordered pairs of values of $X$ and $Y$. The easiest way to list these values and their probabilities is in a table in which the rows and columns indicate the pairs of values being considered, and the probability of each pair of values is given by the corresponding entry in the table. The joint distribution of $X$ and $Y$ can be given as:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y</td>
<td>0</td>
<td>1/12</td>
<td>1/12</td>
<td>1/12</td>
<td>1/12</td>
<td>1/12</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1/12</td>
<td>1/12</td>
<td>1/12</td>
<td>1/12</td>
<td>1/12</td>
</tr>
</tbody>
</table>

As an example of how to read this table, the entry in the upper right corner indicates that the probability that both $X = 6$ and $Y = 0$ equals 1/12.

We could instead encode this table as a function $f$ of the $X$ and $Y$ values. We have:

$$f(x, y) = \frac{1}{12} \text{ where } x \in \{1, 2, 3, 4, 5, 6\} \text{ and } y \in \{0, 1\}.$$ 

This function point of view carries over to the case of continuous random variables more readily than the table notation, since the table notation requires lists of the possible values of $X$ and $Y$. ♦

This example illustrates the case of two discrete random variables. For continuous random variables, the term “distribution” in the definition of a joint random variable refers to a probability density function instead of a table. The joint probability density function is very similar to the function given at the end of the example. For those familiar with calculus, we should add that since the probability density function gives probability density instead of probability, we would have to multiply by $dx\,dy$ to obtain an actual “infinitesimal probability”.

**Marginal distributions of random variables**

Marginal distributions of random variables are not conceptually difficult. They are simply the distributions of random variables considered individually rather than jointly. When working with multiple random variables,
we just need a name for the distribution of each individual random variable, ignoring the other random variables.

**Definition 1.4.3** In the context of a joint random variable built on a pair of discrete random variables $X$ and $Y$, the **marginal distribution** of $X$ (or $Y$) is the distribution of $X$ (or $Y$) ignoring the other random variable.

If we know the joint distribution of two discrete random variables, we can determine the marginal distributions of the individual random variables, as the following example demonstrates.

**Example 1.4.4** Continuing with Example 1.4.2, suppose that we are given the joint distribution and we would like to determine the marginal distribution of $X$. Well, what is the probability that $X = 0$? It is the probability that $X = 0$ and $Y = 1$ plus the probability that $X = 0$ and $Y = 2$ plus the probability that $X = 0$ and $Y = 3$, and so on through all the possible values of $Y$. That is, to find the probability that $X = 0$, we simply sum all of the probabilities in the $X = 0$ row to obtain $6/12 = 1/2$. Similarly, to find the probability that $X = 1$, we sum all of the probabilities in the $X = 1$ row to obtain $6/12 = 1/2$.

Also, to find the probability that $Y$ is equal to any particular value, we sum over all the probabilities in the corresponding column, which always yields $2/12 = 1/6$ in this example.

This gives us a way to determine the marginal distributions of $X$ and $Y$ in general, namely by summing over the corresponding row or column of the table describing their joint distribution.

On the other hand, knowing the marginal distributions of each of two random variables is not enough to determine their joint distribution, as the following example shows.

**Example 1.4.5** Suppose $X_1$ and $Y_1$ are the number of “heads” and the number of “tails” out of three independent flips of a fair coin. Also suppose that $X_2$ is the number of “heads” out of three independent flips of a fair coin, and that $Y_2$ is the number of “tails” out of three independent flips of a different and independent fair coin. Then $X_1$ and $X_2$ have the same probability distribution, as do $Y_1$ and $Y_2$. Stated differently, the marginal distributions of the individual random variables considering $X_1$ and $Y_1$ are the same as the corresponding marginal distributions of the individual random variables considering $X_2$ and $Y_2$.

However, the joint distribution of $X_1$ and $Y_1$ is not the same as the joint distribution of $X_2$ and $Y_2$. One way to see this would be to write out joint distribution tables for both pairs, but we can see that the joint distributions are not the same more easily than that. For example, notice that the probability that $X_1 = 1$ and $Y_1 = 1$ is 0, since the total number of “heads” and “tails” in three flips of a fair
coin must be 3. However, the probability that \( X_2 = 1 \) and \( Y_2 = 1 \) is certainly not 0, since it is quite possible to flip one coin and obtain exactly 1 “heads” while flipping a second coin and obtaining exactly 1 “tails”. Therefore the joint probability distribution of \( X_1 \) and \( Y_1 \) is not the same as that for \( X_2 \) and \( Y_2 \).

While these examples have used discrete random variables, the definition and results hold without modification for continuous random variables.

**Conditional random variables**

The conditional distribution of one random variable given the value of another random variable is exactly what the terminology suggests.

**Definition 1.4.6** Let \( X \) and \( Y \) be random variables. For any possible value \( x \) of \( X \), the **conditional random variable** \( Y|X = x \) (which reads “\( Y \) given that \( X = x \)”) is defined to be the random variable whose values are those of \( Y \) that occur when \( X = x \), with the associated probabilities with which they occur when \( X = x \). (Note that these associated probabilities may not be the same as the probabilities with which \( Y \) assumes these values irrespective of the value of \( X \).)

We illustrate this definition with an example using discrete random variables for simplicity.

**Example 1.4.7** Suppose we have two fair coins that we flip independently. For simplicity, suppose that we have labeled “tails” as 0 and “heads” as 1 on each coin. Let \( X \) be the random variable that is the sum of the numbers on the two coins, and let \( Y \) be the random variable that is the absolute value of the difference of the numbers on the two coins. With a little calculation and some elementary probability theory, the joint distribution of \( X \) and \( Y \) can be shown to be given by the following table:

<table>
<thead>
<tr>
<th></th>
<th>( X )</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y )</td>
<td>0</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0</td>
<td>1/2</td>
<td>0</td>
</tr>
</tbody>
</table>

How do we compute the distribution of \( Y|X = 0 \)? Or, phrased in terms of what we are modeling, how do we compute the distribution of the absolute value of the difference of the numbers on the two coins given that the sum of the numbers on the two coins equals 0? Well, if the sum of the numbers equals 0, then both
Section 1.4 Multiple random variables

coins must have landed with 0 face up, so the absolute value of the difference of the numbers must also equal 0. This means that $Y|X = 0$ takes on the value 0 with probability 1, which gives the distribution.

What about the distribution of $X|Y = 0$? Or, phrased in terms of what we are modeling, how do we compute the distribution of the sum of the numbers on the two coins given that the absolute value of the difference of the numbers equals 0? Well, if the absolute value of the difference equals 0, then the two numbers are the same. This means that either both coins are 0 or both coins are 1, and according to the joint probability table, these two cases happen equally often. Therefore $X|Y = 0$ takes on the value 0 with probability 1/2 and the value 2 with probability 1/2, which gives the distribution.

So where did those two probabilities of 1/2 in the distribution of $X|Y = 0$ come from? In short, we restricted our attention to the $Y = 0$ row in the table and looked at the relative probabilities of occurrence of the various possibilities for $X$ (1/4 for $X = 0$ and 1/4 for $X = 2$ in this case). These probabilities are only relative because they refer to probabilities in the full, unrestricted case, rather than the conditional case that we are interested in. To compute the actual probabilities from the relative probabilities though, we need only divide by the total of the relative probabilities in the row, in order to ensure that the actual probabilities in the restricted setting sum to 1 (instead of whatever the summed to before). In this case, we should divide each of the relative probabilities (which are both 1/4) by 1/2 (the row total), which gives 1/2 for each of the two probabilities in the distribution of $X|Y = 0$. ♦

The line of reasoning used in this example works in general. To compute the distribution of a conditional discrete random variable, simply restrict attention to the row or column of interest and divide the probabilities from the joint probability distribution by the row or column total (in order to ensure that the results sum to 1). For continuous random variables, replace the rows and columns with horizontal and vertical lines in the $xy$-plane, and replace the sums of probabilities with integrals of probability density functions.

Conditional random variables lead naturally to another concept: independence of random variables.

**Definition 1.4.8** Two random variables $X$ and $Y$ are independent if any of the following three equivalent conditions are satisfied:

1. For every value $x$ that $X$ assumes, the distribution of $Y|X = x$ is the same as the marginal distribution of $Y$.

2. For every value $y$ that $Y$ assumes, the distribution of $X|Y = y$ is the
same as the marginal distribution of $X$.

3. For every value $x$ that $X$ assumes and $y$ that $Y$ assumes, the probability that both $X = x$ and $Y = y$ is equal to the probability that $X = x$ times the probability of $Y = y$.

It should not be at all obvious that these three conditions are equivalent, but it can be shown that they are.

The first of these conditions can be interpreted as asserting that information about the value of $X$ conveys no additional information about the value of $Y$. The second of these conditions can be interpreted as asserting that information about the value of $Y$ conveys no additional information about the value of $X$. Both of these interpretations give some motivation for why we use the term “independent”.

The third condition is not as easy to interpret, but it turns out to be very useful in checking independence and computing with independent random variables.

**Example 1.4.9** Because we modeled the die roll and the coin flip as independent processes in Example 1.4.2, the random variables based on them are independent. A quick check of the joint probability table indicates that the third condition in the definition of independence is satisfied (essentially by design).

On the other hand, the random variables based on the pair of coin flips in Example 1.4.7 are not independent, as a quick check of any of the three conditions reveals. For example, we computed that the distribution of $Y|X = 0$ had only the single value 0 with probability 1. However, the marginal distribution of $Y$ has the two values 0 and 1, each with probability 1/2. This shows that the first condition is not satisfied. ♦
1.5 Problems

1. Use the pennies data set for this problem. As always, be sure to read the description of the data set.

   (a) Compute the sample mean, sample median of the mass of these pennies.

   (b) Compute the sample standard deviation and sample variance of the mass of these pennies.

   (c) Compute the five-number summary of the mass of these pennies.

2. Use the pennies data set for this problem. As always, be sure to read the description of the data set if you have not already done so.

   (a) Create a histogram of all the penny masses in this data set, with a reasonable bin width. Don’t forget to label the axes and include a title for the histogram.

   (b) Describe the shape of this histogram.

   (c) Explain the shape of this histogram by looking also at the other variable in the data set, the year of the penny.

   (d) If you have done Problem 1, indicate why the numerical summaries of penny masses computed in that problem are not particularly useful.

3. Use the introductory statistics surveys data set for this problem. As always, be sure to read the description of the data set if you have not already done so.

   (a) Make a histogram of the shoe sizes of female students in the data set. Also make a histogram of the shoe sizes of male students in the data set. Include all relevant axis labels, and make the plots have the same axis limits to help in comparing the two histograms.

   (b) Make side-by-side boxplots of female and male shoe sizes in the data set, including all relevant axis labels.
(c) Interpret what you see in these plots, remembering that the data came from self-reporting in response to the question “What is your shoe size?”. (Hint: the context of the data is quite important here. The obvious interpretation needs some additional thought to arrive at a correct interpretation.)

(d) Describe the distribution of the number of states visited by students in this data set, making any relevant plot(s).

4. Use the quizzes and finals data set to answer the question: do the distributions of quiz averages appear similar for the three sections in this data set? More specifically, compare and contrast these distributions by computing the relevant numerical and graphical summaries (preferably several of both) that we have studied so far.