Chapter 1

Basic Concepts

> *If it can go wrong, it will.* —Murphy
>
> *The phenomenon of toast falling from a table to land butter-side down on the floor is popularly held to be empirical proof of the existence of Murphy's Law.* —Matthews, 1995, p. 172

In this chapter, we present a few basic concepts about statistical modeling. These concepts, which form a cornerstone for the rest of the book, are most easily discussed in the context of a simple example. We use an example motivated by Murphy's Law. As noted by Matthews in the above quote, the probability toast falls butter-side down is an important quantity in testing Murphy's law. In order to estimate this probability, toast can be flipped; Figure 1.1 provides an example of a toast flipper. The toast is first buttered and then placed butter-side up at Point A of the device. Then the experimenter pushes on the lever (Point B) and observes whether the toast hits the floor butter-side up or butter-side down. We will call the whole act *flipping toast* as it is analogous to flipping coins.

1.1 Random Variables

1.1.1 Outcomes and Events

The basic ideas of modeling are explained with reference to two simple experiments. In the first, Experiment 1, a piece of toast is flipped once. In

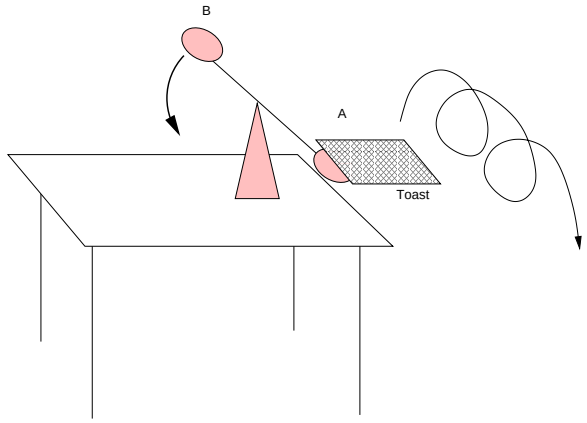

Figure 1.1: Flipping toast. A piece of toast is placed butter-side at Point A. Then, force is applied to Point B, launching the toast off the table.

the second, Experiment 2, a piece of toast is flipped twice. The first step in analysis is describing the outcomes and sample space.

Definition 1 (Outcome) *An outcome is a possible result of an experiment.*

There are two possible outcomes for Experiment 1: (1) the piece of toast falls butter-side down or (2) it falls butter-side up. We denote the former outcome by D and the latter by U (for "down" and "up," respectively). For Experiment 2, the outcomes are denoted by ordered pairs as follows:

- (D, D) : The first and second pieces fall butter-side down.
- (D, U) : The first piece falls butter-side down and the second falls butter-side up.
- (U, D) : The first piece falls butter-side up and the second falls butter-side down.
- (U, U) : The first and second pieces fall butter-side up.

Definition 2 (Sample Space) *The sample space is the set of all outcomes.*

The sample space for Experiment 1 is $\{U, D\}$. The sample space for Experiment 2 is $\{(D, D), (U, D), (D, U), (U, U)\}$.

Although outcomes describe the results of experiments, they are not sufficient for most analyses. To see this insufficiency, consider Murphy's Law in Experiment 2. We are interested in whether one or more of the flips is butter-side up. There is no outcome that uniquely represents this event. Therefore, it is common to consider *events*:

Definition 3 (Events) *Events are sets of outcomes. They are, equivalently, subsets of the sample space.*

There are four events associated with Experiment 1 and sixteen associated with Experiment 2. For Experiment 1, the four events are:

$$\{U\},\ \{D\},\ \{U, D\},\ \emptyset.$$

The event $\{U, D\}$ refers to the case that either the toast falls butter-side down or it falls butter-side up. Barring the miracle that the toast lands on its side, it will always land either butter-side up or butter-side down. Even though this event seems uninformative, it is still a legitimate event and is included. The null set ($\emptyset$) is an empty set; it has no elements.

The 16 events for Experiment 2 are:

$\{(D, D)\}$,	$\{(D, U)\}$,	$\{(U, D)\}$,
$\{(U, U)\}$,	$\{(D, D), (D, U)\}$,	$\{(D, D), (U, D)\}$,
$\{(D, D), (U, U)\}$,	$\{(D, U), (U, D)\}$,	$\{(D, U), (U, U)\}$,
$\{(U, D), (U, U)\}$,	$\{(D, U), (U, D), (U, U)\}$,	$\{(D, D), (U, D), (U, U)\}$,
$\{(D, D), (D, U), (U, U)\}$,	$\{(D, D), (D, U), (U, D)\}$,	$\{(D, D), (D, U), (U, D), (U, U)\}$

$\emptyset$.

1.1.2 Probability

Definition 4 (Probability) *Probabilities are numbers assigned to events. The number reflects our degree of belief in the plausibility of the event. This number ranges from zero (the event will never occur in the experiment) to one (the event will always occur). The probability of event A is denoted $Pr(A)$.*

To explain how probability works, it is helpful at first to assume the probabilities of at least some of the events are known. For now, let's assume the probability that toast falling butter-side down. is .7; e.g., $Pr(\{D\}) = .7$, It is desirable to place probabilities on the other events as well. The following concepts are useful in doing so:

Definition 5 (Union) *The union of sets A and B, denoted $A \cup B$, is the set of all elements that are either in A or in B. For example, if $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$, then $A \cup B = \{1, 2, 3, 4, 5\}$.*

Definition 6 (Intersection) *The intersection of sets A and B, denoted $A \cap B$, is the set of all elements that are both in A and in B. For example, if $A = \{1, 2, 3\}$ and $B = \{3, 4, 5\}$, then $A \cap B = \{3\}$.*

Probabilities are placed on events by applying the following three rules of probability:

Event	Pr
$\{D\}$	.7
$\{U\}$	.3
$\{U, D\}$	1
$\emptyset$	0

Table 1.1: Probabilities of events in Experiment 1.

Definition 7 (The Three Rules of Probability) *The three rules of probability are:*

1. $Pr(A) \geq 0$, *where* A *is any event,*
2. $Pr(\textit{sample space}) = 1$, *and*
3. *if* $A \cap B = \emptyset$, *then* $Pr(A \cup B) = Pr(A) + Pr(B)$.

The three rules are known as Kolmogorov Axioms of probability (Kolmogorov, 1950). It is relatively easy to apply the rules to the probability of events. Table 1.1 shows the probabilities of events for Experiment 1.

For Experiment 2, let's assume the following probabilities on the events corresponding to single outcomes: $Pr(\{(D, D)\}) = .49$, $Pr(\{(D, U)\}) = .21$, $Pr(\{(U, D)\}) = .21$, and $Pr(\{(U, U)\}) = .09$. Using the above rules, we can compute the probabilities on the events. These probabilities are shown in Table 1.2.

1.1.3 Random Variables

Suppose we are interested in the number of butter-side-down flips. According to Murphy's Law, this number should be large. For each experiment, this number varies according the probability of events. The concept of *random variable* captures this dependence:

Event	Probability
$\{(D,D)\}$	.49
$\{(D,U)\}$	.21
$\{(U,D)\}$	.21
$\{(U,U)\}$	.09
$\{(D,D),(D,U)\}$	.70
$\{(D,D),(U,D)\}$	.70
$\{(D,D),(U,U)\}$	.58
$\{(D,U),(U,D)\}$	.42
$\{(D,U),(U,U)\}$	.30
$\{(U,D),(U,U)\}$	.30
$\{(D,U),(U,D),(U,U)\}$	.51
$\{(D,D),(U,D),(U,U)\}$	.79
$\{(D,D),(D,U),(U,U)\}$	.79
$\{(D,D),(D,U),(U,D)\}$	.91
$\{(D,D),(D,U),(U,D),(U,U)\}$	1
$\emptyset$	0

Table 1.2: Probabilities of events in Experiment 2.

Outcome	Value of X
D	1
U	0

Table 1.3: Definition of a random variable X for Experiment 1.

Outcome	Value of X
(D, D)	2
(D, U)	1
(U, D)	1
(U, U)	0

Table 1.4: Definition of a random variable X for Experiment 2.

Definition 8 (Random Variable) *A random variable (RV) is a function that maps events into sets of real numbers. Probabilities on events become probabilities on the corresponding sets of real numbers.*

Let random variable X denote the number of butter-side down flips. X is defined for Experiments 1 and 2 in Tables 1.3 and 1.4, respectively.

Random variables map experimental results into numbers. This mapping also applies to probability: probabilities on events in the natural word transfer to numbers. Tables 1.5 and 1.6 show this mapping.

Value of X	Corresponding Event	Probability
1	$\{D\}$	.7,
0	$\{U\}$	.3,

Table 1.5: Probabilities associated with random variable X for Experiment 1.

Value of X	Corresponding Event	Probability
2	$\{(D, D)\}$	.49
1	$\{(D, U), (U, D)\}$	.42
0	$\{(U, U)\}$	.09

Table 1.6: Probabilities associated with random variable X for Experiment 2.

Random variables are typically typeset in upper-case; e.g., X. Random variables take on values and these values are typeset in lower-case; e.g., x. The expression $Pr(X = x)$ refers to the probability that an event corresponding to x will occur. The mappings in Tables 1.5 and 1.6 are typically expressed as the function $f(x) = Pr(X = x)$, which is called the *probability mass function.*

Definition 9 (Probability Mass Function) *Probability mass function, f(x), provides the probability for a particular value of a random variable.*

For Experiment 1, the probability mass function of X is

$$f(x) = \begin{cases} .3 & x = 0 \\ .7 & x = 1 \\ 0 & \text{Otherwise} \end{cases} \tag{1.1}$$

For Experiment 2, it is

$$f(x) = \begin{cases} .09 & x = 0 \\ .42 & x = 1 \\ .49 & x = 2 \\ 0 & \text{Otherwise} \end{cases} \tag{1.2}$$

Probability mass functions always sum to 1, i.e., $\sum_x f(x) = 1$ as a consequence of the Kolomogorov Axioms in Definition 7. As can be seen, the above probability mass functions sum to 1.

As discussed in the preface, we use the computer package **R** to aid with statistical analysis. **R** can be used to plot these probability mass functions. The following code is for plotting the probability mass function for Experiment 2, shown in Eq. (1.2). The first step in plotting is to assign values to x and f. These assignments are implemented with the statements `x=c(0,1,2)` and `f=c(.09,.42,.49)`. The symbol `c()` stands for "concatenate" and is used to join several numbers into a vector. The values of a variable may seen by simply typing the variable name at the **R** prompt; e.g.,

```
> x=c(0,1,2)
> f=c(.09,.42,.49)
> x
[1] 0 1 2
> f
[1] 0.09 0.42 0.49
```

The function `plot()` can be used to plot one variable as a function of another. Try `plot(x,f,type='h')`. The resulting graph should look like Figure 1.2. There are several types of plots in **R** including scatter plots, bar plots, and line plots. The type of plot is specified with the `type` option. The option `type='h'` specifies thin vertical lines, which is appropriate for plotting probability mass functions. Help on any **R** command is available by typing help with the command name in parentheses, e.g., `help(plot)`. The points on top of the lines were added with the command `points(x,f)`.

The random variable X, which denotes the number of butter-side-down flips, is known as a *discrete random variable*. The reason is that probability is assigned to discrete points; for X, it is the discrete points of 0, 1, and 2. There is another type of random variable, a *continuous random variable*, in which probability is assigned to intervals rather than points. The differences between discrete and continuous random variables will be discussed in Chapter 4.

1.1.4 Parameters

Up to now, we have assumed probabilities on some events (those corresponding to outcomes) and used the laws of probability to assign probabilities to the other events. In experiments, though, we do not assume probabilities; instead, we estimate them from data. We introduce the concept of a parameter

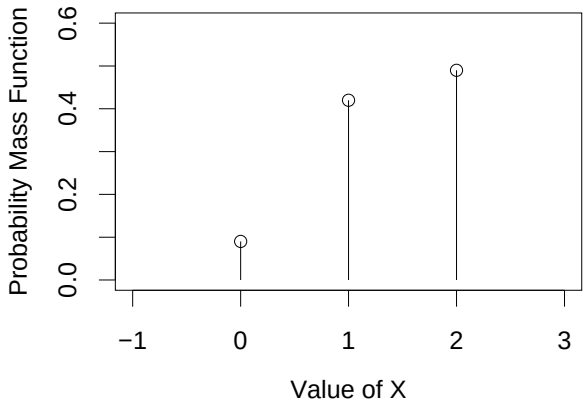

Figure 1.2: Probability mass functions for random variable X, the number of butter-side down flips, in Experiment 2.

to avoid assuming probabilities.

Definition 10 (Parameter) *Parameters are mathematical variables on which a random variable may depend. Probabilities on events are functions of parameters.*

For example, let the outcomes of Experiment 1 depend on parameter p as follows: $Pr(D) = p$. Because the probability of all outcomes must sum 1.0, the probability of event U must be $1-p$. The resulting probability mass function for X, the number of butter-side down flips, is

$$f(x;p) = \begin{cases} 1-p & x = 0 \\ p & x = 1 \\ 0 & \text{Otherwise} \end{cases} \tag{1.3}$$

The use of the semicolon in $f(x;p)$ indicates that the function is of one variable, x, for a given value of the parameter p.

Let's consider the probabilities in Experiment 2 to be parameters defined as $p_1 = Pr(D,D)$, $p_2 = Pr(D,U)$, $p_3 = Pr(U,D)$. By the laws of probability,

$Pr(U,U) = 1 - p_1 - p_2 - p_3$. The resulting probability mass function on X is

$$f(x; p_1, p_2, p_3) = \begin{cases} 1 - p_1 - p_2 - p_3 & x = 0 \\ p_2 + p_3 & x = 1 \\ p_1 & x = 2 \\ 0 & \text{Otherwise} \end{cases} \tag{1.4}$$

This function is still of one variable, x, for given values of p_1, p_2, and p_3.

Although the account of probability and random variables presented here is incomplete, it is sufficient for the development that follows in the book. A more complete basic treatment can be found in mathematical statistics textbooks such as Hogg & Craig (1978) and Rice (1995).

Problem 1.1.1 (Your Turn)
You and a friend are playing a game with a four-sided die. Each of the sides, labeled A, B, C, and D, has equal probability of landing up on any given throw.

1. There are two throws left in the game; list all of the possible outcomes for the last two throws. Hint: these outcomes may be expressed as ordered pairs.

2. In order to win, you need to throw an A or B on each of the final two throws. List all the outcomes that are elements in the event that you win.

3. In this game there are only two possible outcomes: winning and losing. Given the information in Problem 2, what is the probability mass function for the random variable that maps the event that you win to 0 and the event that you lose to 1? Plot this function in **R**.

1.2 Binomial Distribution

Experiment 1, in which a piece of toast is flipped once, plays an important role in developing more sophisticated models. Experiment 1 is an example

of a Bernoulli trial, defined below.

Definition 11 (Bernoulli Trial) *Bernoulli trials are experiments with two mutually exclusive (or* dichotomous*) outcomes. Examples include a flip of toast; the sex of a baby (assuming only male or female outcomes); or, for our purposes, whether a participant produces a correct response (or not) on a trial in a psychology experiment. By convention, one of the outcomes is called a success, the other a failure. A random variable is distributed as a Bernoulli if it has two possible values: 0 (for failure) and 1 (for success).*

As a matter of notation, we will consider the butter-side-down result in toast flipping as a success and the butter-side-up result as a failure. Random variables from Bernoulli trials have a single parameter p and probability mass function given by Eq. (1.3). In general, the value of p is not known a priori and must be estimated from the data.

Experiment 2 is a sequence of two Bernoulli trials. We let X_1 and X_2 denote the outcomes (either success of failure) of these two trials. Let p_1 and p_2 be the probability of success parameter for X_1 and X_2, respectively. If $p_1 = p_2$, then random variables X_1 and X_2 are called *identical*. Note that identical does not mean that the results of the flips are the same, e.g., both flips are successes or both are failures. Instead, it means that the probabilities of a success are the same. The concept of identical random variables can be extended: whenever two random variables have the same probability mass function, they are identical. If the result of a Bernoulli trial is not affected by the result of the others, then the RVs are called *independent*. If two random variables are both independent and identically distributed, then they are called *iid*.

In order to understand the concepts of independent and identically distributed, it may help to consider a concrete example. Suppose a basketball player is shooting free throws. If one throw influences the next, for instance, if a player gets discouraged because he or she misses a throw, this is a violation of independence, but not necessarily of identical distribution. Although one throw may effect the next, if *on average* they are the same, identical

distribution is not violated. If a player gets tired and does worse over time, regardless of the outcome of his or her throws, then identical distribution is violated. It is possible to violate one and not the other.

Definition 12 (Bernoulli Process) *A sequence of independent and identically distributed Bernoulli trials is called a Bernoulli Process.*

In a Bernoulli process, each X_i is a function of the same parameter p. Furthermore, because each trial is independent, the order of outcomes is unimportant. Therefore, it makes sense to define a new random variable, Y, which is the total number of successes, i.e., $Y = \sum_{i=1}^{N} X_i$.

Definition 13 (Binomial Random Variable) *The random variable which denotes the number of successes in a Bernoulli process is called a binomial. The binomial random variable has a single parameter, p (probability of success on a single trial). It is also a function of a constant, N, the number of trials. It can take any integer value from 0 to N.*

Definition 14 (Random Variable Notation) *It is common to use the character "$\sim$" to indicate the distribution of a random variable. A binomial random variable is indicated as $Y \sim Binomial(p, N)$. Here, Y is the number of successes in N trials where the probability of success on each trial is p.*

The variable p is considered a parameter but the variable N is not. The value of N is known exactly and supplied by the experimenter. The true value of p is generally unknown and must be estimated. The probability

mass function of a binomial random variable describes the probability of observing y successes in N trials:

$$Pr(Y = y) = f(y; p) = \begin{cases} \begin{pmatrix} N \\ y \end{pmatrix} p^y (1-p)^{N-y} & y = 0, .., N \\ 0 & \text{Otherwise} \end{cases} \tag{1.5}$$

The term $\begin{pmatrix} N \\ y \end{pmatrix}$ refers to the "choose function," given by

$$\begin{pmatrix} N \\ y \end{pmatrix} = \frac{N!}{y!(N-y)!}.$$

Let's look at the probability mass function in **R**. Try the following for $N = 20$ and $p = .7$:

```
y=0:20    #assigns x to 0,1,2,..,20
f=dbinom(y,20,.7)  #probability mass function
plot(y,f,type='h')
points(y,f)
```

In the above code, `dbinom()` is an **R** function that returns the probability mass function of a binomial. Variable `y` is a vector taking on 21 values (0,1,..,20). Because the first argument of `dbinom()` is a vector the output is also a vector with one element for each element of `y`. Type `f` to see the 21 values. The first value of `f` corresponds to the first value of `y`; the second value of `f` to the second value of `y`; and so on. The resulting plot is shown in Figure 1.3.

The goal of Experiment 1 and 2 is to learn about p. These experiments, however, are too small to learn much. Instead, we need a larger experiment with more flips; for generality, consider the case in which we had N flips. One common-sense approach is to take the number of successes in a Bernoulli process, Y and divide by N, the number of trials. A function of random variables that estimates a parameter is called an *estimator*. The common-sense estimator of p is $\hat{p} = Y/N$. It is conventional to place the caret over an estimator as in $\hat{p}$. This distinguishes it from the true, but unknown parameter p. Note that because $\hat{p}$ is a function of a random variable, it is also a random variable. Because estimators are random variables themselves, studying them requires more background about random variables.

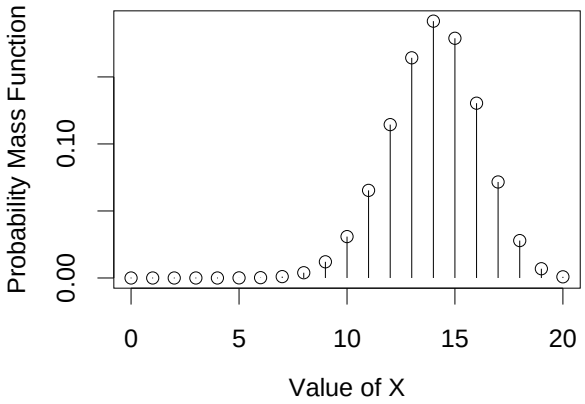

Figure 1.3: Probability Mass function for a binomial random variable with $N = 20$ and $p = .7$

1.3 Expected Values of Random Variables

1.3.1 Expected Value

The expected value is the center or theoretical average of a random variable. For example, the center of the distribution in Figure 1.3 is at 14. Expected value is defined as:

Definition 15 (Expected Value) *The expected value of a discrete random variable X is given as:*

$$E(X) = \sum_x x f(x; p). \tag{1.6}$$

The expected value of a random variable is closely related to the concept of an average, or mean. Typically, to compute a mean of a set of values, one adds all the values together and divides by the total number of values. In the case of an expected value, however, each possible value is weighted by the probability that it will occur before summing. The expected value of a random variable is also called its first moment, population mean, or simply its mean. It is important to differentiate between the expected value of a random variable and the sample mean, which will be discussed subsequently.

Consider the following example. Let X be a random variable denoting the outcome of a Bernoulli trial with parameter $p = .7$. Then, the expected value is given by $\mathrm{E}(X) = \sum_x xf(x;p) = (0 \times .3) + (1 \times .7) = .7$. More generally, the expected value of a Bernoulli trial with parameter p is $E(X) = p$.

1.3.2 Variance

Whereas the expected value measures the center of a random variable, variance measures its spread.

Definition 16 (Variance) *The variance of a discrete random variable is given as:*

$$V(X) = \sum_x [x - E(X)]^2 f(x;p) \tag{1.7}$$

Just as the expected value is a weighted sum, so is the variance of a random variable. The variance is the sum of all possible squared deviations from the expected value, weighted by their probability of occuring. An equivalent equation for variance is given as $\mathrm{V}(X) = \mathrm{E}[(X - \mathrm{E}(X))^2]$; that is, variance is the expected squared deviation of a random variable from its mean. The variance of a random variable is different from the variance of a sample, which will be discussed subsequently.

Another common measure of the spread of a random variable is the *standard deviation*. The standard deviation of a random variable is square-root of variance, e.g. $\mathrm{SD}(X) = \sqrt{\mathrm{V}(X)}$. Standard deviation is often used as a measure of spread rather than variance because it is in the same units as the random variable. Variance, in contrast, is in squared units, which are more difficult to interpret. The standard deviation of an estimator has its own name: *standard error*.

Definition 17 (Standard Error) *The standard deviation of a parameter estimator is called the standard error of the estimator.*

Problem 1.3.1 (Your Turn)

1. It is common in psychology to ask people their opinion of statements, e.g., "I am content with my life." Responses are often collected on a Likert scale; e.g., *1=strongly disagree, 2=disagree, 3=neutral, 4=agree, 5=strongly agree*. The answer may be considered a random variable. Suppose the probability mass function for the above question is given as $f(x) = (.05, .15, .25, .35, .2)$ for $x = 1, \ldots, 5$, respectively. Plot this probability mass function. Compute the expected value. Does the expected value appear to be at center of the distribution? Compute the variance.

2. Let Y be a binomial RV with $N = 3$ and parameter p. Show $\mathrm{E}(Y) = 3p$.

1.3.3 Expected Value of Functions of Random Variables

It is often necessary to consider functions of random variables. For example, the common sense estimator of p, $\hat{p} = Y/N$, is a function of random variable Y. The following two rules are convenient in computing the expected value of functions of random variables.

Definition 18 (Two rules of expected values) *Let X, Y, and Z all denote random variables, and let $Z = g(X)$. The following rules apply to the expected values:*

1. *The expected value of the sum is the sum of the expected values:*
$$E(X + Y) = E(X) + E(Y)$$
, and

2. *the expected value of a function of a random variable is*
$$E(Z) = \sum_x g(x)f(x; p),$$
where f is the probability mass function of X.

The first rule can be used to find the expected value of a binomial random variable. By definition, binomial RV Y is defined as $Y = \sum_{1=1}^{N} X_i$, where the X_i are iid Bernoulli trials. Hence, by Rule 1,

$$E(Y) = E(\sum_i X_i) = \sum_i E(X_i) = \sum_i p = Np$$

The second rule can be used to find the expected value of $\hat{p}$. The random variable $\hat{p} = g(Y)$ is $g(Y) = Y/N$. The expected value of $\hat{p}$ is given by:

$$\begin{aligned} \mathrm{E}(\hat{p}) &= \mathrm{E}(g(Y)) \\ &= \sum_x (x/N)f(x; p) \\ &= (1/N)\sum_x xf(x; p) \\ &= (1/N)\mathrm{E}(Y) \\ &= (1/N)(Np) \\ &= p. \end{aligned}$$

While $\hat{p}$ may vary from experiment to experiment, its average will be p.

1.4 Sequences of Random Variables

1.4.1 Realizations

Consider Experiment 1, the single flip of toast and the random variable, X, the number of butter-side-down flips. Before the experiment, there are two possible values that X could take with nonzero probability, 0 and 1. Afterward, there is one result. The result is called the realization of X.

Definition 19 (Realization) *The realization of a RV is the value it attains in an experiment.*

Consider Experiment 2 in which two pieces of toast are flipped. Before the experiment is conducted, the possible values of X are 0, 1, and 2. Afterward, the realization of X can be only one of these values. The same is true of estimators. Consider the random variables $Y \sim \text{Binomial}(p, N)$ and common-sense estimate $\hat{p} = Y/N$. After an experiment, these will have realizations denoted y and y/N, respectively. The realization of an estimator is called an estimate.

It is easy to generate realizations in **R**. For binomial random variables, the appropriate function is `rbinom()`: type `rbinom(1, 20, .7)`. The first argument is the number of realizations, which is 1. The second is N, the number of trials. The third is p, the probability of success on a trial. Try the command a few times. The output of each command is one realization of an experiment with 20 trials.

1.4.2 Law of Large Numbers

Consider `rbinom(5, 20, .7)`. This should yield five replicates; the five realizations from five separate experiments. There are two interpretations of the five realizations. The first, sometimes prominent in undergraduate introductory texts, is that these five numbers are samples from a common distribution. The second, which is more common in advanced treatments of probability, is that the realizations are from different, though independent and identically distributed, random variables. Replicate experiments can be

represented as a sequence of random variables, and in this case, we write:

$$Y_i \overset{iid}{\sim} \text{Binomial}(p = .7, N = 20) \;\; i = 1, \ldots, 5.$$

Each Y_i is a different random variable, but all Y_i are independent and distributed as identical binomials. Each i could repeasent a different trial, a different person, or a different experimental condition.

Of course, we are not limited to 5 replicates; for example `y=rbinom(200, 20, .7)` produces 200 replicates and stores them in vector `y`. To see a histogram, type `hist(y, breaks=seq(-.5, 20.5, 1), freq=T)`. We prefer a different type of histogram for looking at realizations of discrete random variables—one in which the y-axis is not the raw counts but the proportion, or relative frequency, of counts. These histograms are called *relative frequency histograms.*

Definition 20 (Relative Frequency Histogram) *Let $Y_i \overset{iid}{\sim} Y$ be a sequence of M independent and identically distributed discrete random variables and let $y_1, .., y_M$ be a sequence of corresponding realizations. Let $h_M(j)$ be the proportion of realizations with value j. The relative frequency histogram is a plot of $h_M(j)$ against j.*

Relative frequency histograms may be drawn in **R** with the following code:

```
freqs=table(y) #frequencies of realization values
props=freqs/200 # proportions of realization values
plot(props, xlab='Value', ylab='Relative Frequency')
```

The code draws the histogram as a series of lines. The relative histogram plot looks like a probability mass function. Figure 1.4A shows that this is no coincidence. The lines are the relative frequency histogram; the points are the probability mass function for a binomial with $N = 20$ and $p = .7$ (The points were produced with the `points()` function. The specific form is `points(0:21,dbinom(0:21,20,.7),pch=21)`).

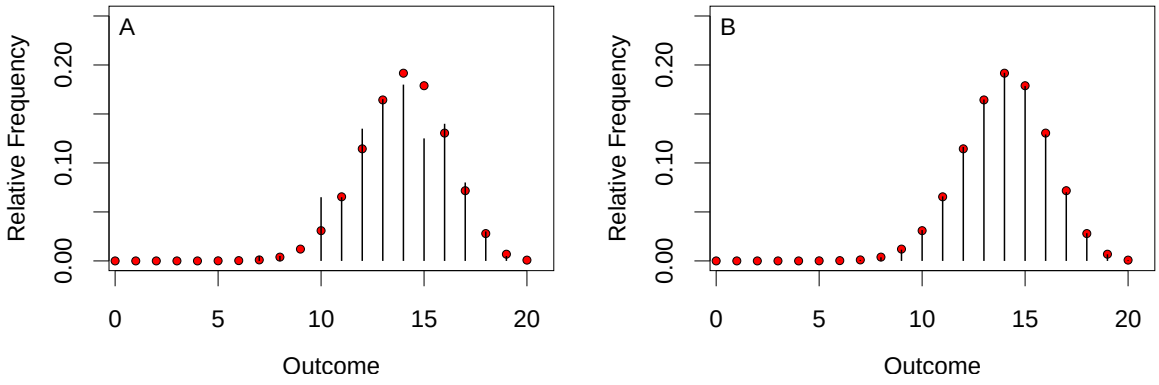

Figure 1.4: A. Relative Frequency histogram and probability mass function roughly match with 200 realizations. B. The match is near perfect with 100,000 realizations.

The match between the relative histogram and the pmf is not exact. The problem is that there are only 200 realizations. Figure 1.4B shows the match between probability mass function and the relative frequency histogram when there are 10,000 realizations. Here, the match is nearly perfect. This match indicates that as the number of realizations grows, the relative frequency histogram converges to the probability mass function. The convergence is a consequence of the Law of Large Numbers. The Law of Large says, informally, that the proportion of realizations attaining a particular value will converge to the true probability of that realization. More formally,

$$\lim_{M \to \infty} h_M(j) = f(j; p),$$

where f is the probability mass function of Y.

The fact that the relative frequency histogram of samples converges to the probability mass function is immensely helpful in understanding random variables. Often, it is difficult to write down the probability mass function of a random variable but easy to generate samples of realizations. By generating a sequence of realizations from independent and identically distributed random variables, it is possible to see how the probability mass functions behaves. This approach is called the *simulation approach* and we use it liberally as a teaching tool.

We can use the simulation approach to approximate the probability mass function for the common-sense estimator $\hat{p} = Y/N$ with the following **R** code:

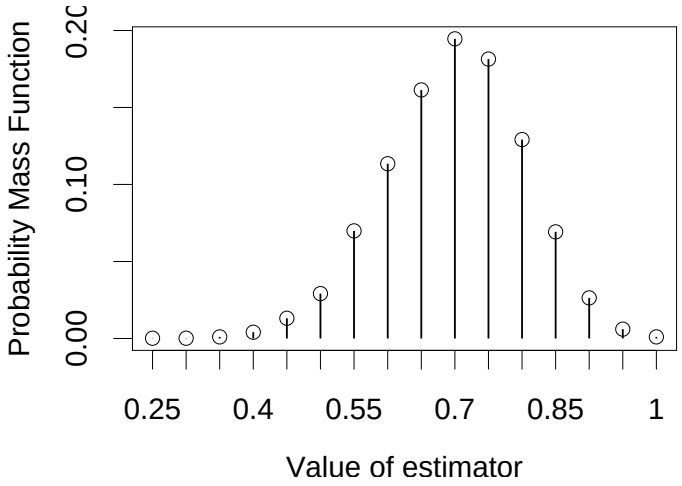

Figure 1.5: Simulated probability mass function for a common-sense estimator of p for a binomial with $N = 20$ and $p = .7$.

```
y=rbinom(10000,20,.7)
p.hat=y/20  #10,000 iid replicates of p-hat
freq=table(p.hat)
plot(freq/10000,type='h')
```

The resulting plot is shown in Figure 1.5. The plot shows the approximate probability mass function for the $\hat{p}$ estimator. The distribution of an estimator is so often of interest that it has a special name: a *sampling distribution.*

Definition 21 (Sampling Distribution) *A sampling distribution is the probability mass function of an estimator.*

1.5 Estimators

Estimators are random variables that are used to estimate parameters from data. We have seen one estimator, the common-sense estimator of p in a

binomial: $\hat{p} = Y/N$. Two others are the sample mean and sample variance defined below, which are used as estimators for the expected value and variance of an RV, respectively.

Definition 22 (Sample Mean and Sample Variance) *Let $Y_1, Y_2, ..., Y_M$ be a collection of M random variables. The sample mean and sample variance are defined as*

$$\bar{Y} = \frac{\sum_i Y_i}{M} and \tag{1.8}$$

$$s_Y^2 = \frac{\sum_i (Y_i - \bar{Y})^2}{M-1}, \tag{1.9}$$

respectively.

How good are these estimators? To answer this question, we first discuss properties of estimators.

1.5.1 Properties of estimators

To evaluate the usefulness of estimators, statisticians usually discuss three basic properties: *bias*, *efficiency*, and *consistency*. Bias and efficiency are illustrated in Table 1.7. The data are the results of weighing a hypothetical person of 170 lbs on two hypothetical scales four separate times. *Bias* refers to the mean of repeated estimates. Scale A is unbiased because the mean of the estimates equals the true value of 170 lbs. Scale B is biased. The mean is 172 lbs which is 2 lbs. greater than true value of 170 lbs. Examining the values for scale B, however, reveals that scale B has a smaller degree of error than scale A. Scale B is called more *efficient* than Scale A. High efficiency means that expected error is low. Bias and efficiency have the same meaning for estimators as they do for scales. Bias refers to the difference between the average value of an estimator and a true value. Efficiency refers to the amount of spread in an estimator around the true value.

The bias and efficiency of any estimator depends on the sample size. For example, the common-sense estimator $\hat{p}$ is $\hat{p} = (\sum_{i=1}^{N} Y_i/N)$ provides a better

Table 1.7: Two Hypothetical Scales

	Scale A	Scale B
	180	174
	160	170
	175	173
	165	171
Mean	170	172
Bias	0	2.0
RMSE	7.91	2.55

estimate with increasing N. Let $\hat{\theta}_N$ denote an estimator which estimates parameter θ, for a sample size of N.

Definition 23 (Bias) *The bias of an estimator is given by B_N:*

$$B_N = E(\hat{\theta_N}) - \theta$$

Bias refers to the expected value of an estimator. We have already proven that estimator $\hat{p}$ is unbiased (Section 1.3.3). Both sample mean and sample variance are also unbiased. Other common estimators, however, are biased. One example is the sample correlation. Fortunately, this bias reduces toward zero with increasing N. Unbiasedness is certainly desirable, but not critical. Many of the estimators discussed in this book will have some degree of bias.

Problem 1.5.1 (Your Turn)
Let $Y_i; i = 1..N$ be a sequence of N independent and identically distributed random variables. Show that the sample mean is unbiased for all N (hint: use the rules of expected value in Definition 18).

Definition 24 (Efficiency) *Efficiency refers to the expected degree of error in estimation. We use root-mean-squared error (RMSE) as a measure of efficiency:*

$$RMSE = \sqrt{E[(\hat{\theta}_N - \theta)^2]} \tag{1.10}$$

More efficient efficient estimators have less error, on average, than less efficient estimators. Sample mean and sample variance are the most efficient unbiased estimators of expected value and variance, respectively. One of the main issues is estimation is the trade-off between bias and efficiency. Often, the most efficient estimator of a parameter is biased, and this facet is explored in the following section.

The final property of estimators is consistency.

Definition 25 (Consistency) *An estimator is consistent if*

$$\lim_{N \to \infty} RMSE(\hat{\theta}_N) = 0$$

Consistency means that as the sample sizes gets larger and larger, the estimator converges to the true value of the parameter. If an estimator is consistent, then one can estimate the parameter to arbitrary accuracy. To get more accurate estimates, one simply increases the sample size. Conversely, if an estimator is inconsistent, then there is a limit to how accurately the parameter can be estimated, even with infinitely large samples. Most common estimators in psychology, including the sample mean, sample variance, and sample correlation, are consistent.

Because sample means and sample variances converge to expected value and variances, respectively, they can be used to estimate these properties. For example, let's approximate the expected value, variance, and standard error of $\hat{p}$ with the sample statistics in **R**. We first generate a sequence of realizations $y_1, .., y_M$ for binomial random variables $Y_i \overset{iid}{\sim} Y$ $i = 1, .., M$. For each realization, we compute an estimate $p_i = y_i/N$. The sample mean, sample variance, and sample standard deviation approximate the expected value, variance, and standard error. To see this, run the following **R** code:

```
y=rbinom(10000,20,.7)
p.hat=y/20
mean(p.hat)   #sample mean
var(p.hat)    #sample variance (N-1 in denominator)
sd(p.hat)     #sample std. deviation (N-1 in denominator)
```

Problem 1.5.2 (Your Turn)
How does the standard error of $\hat{p}$ depend on the number of trials N?

Let's use the simulation method to further study the common-sense estimator of the expected value of the binomial, the sample mean. Suppose in an experiment, we had ten binomial RVs, each the result of 20 toast flips. Here is a formal definition of the problem:

$$\begin{aligned} Y_i &\overset{iid.}{\sim} \text{Binomial}(p, 20), \quad i = 1...10, \\ \bar{Y} &= \frac{\sum_i Y_i}{10}. \end{aligned}$$

The following code generates 10 replicates from a binomial, each of 20 flips. Here we have defined a custom function called `bsms()` (bsms stands for "binomial sample mean sampler"). Try it a few times. This is analogous to having 10 people each flip 20 coins, then returning the mean number of heads across people.

```
#define function
bsms=function(m,n,p)
{
z=rbinom(m,n,p)
mean(z)
}

#call function
bsms(10,20,.7)
```

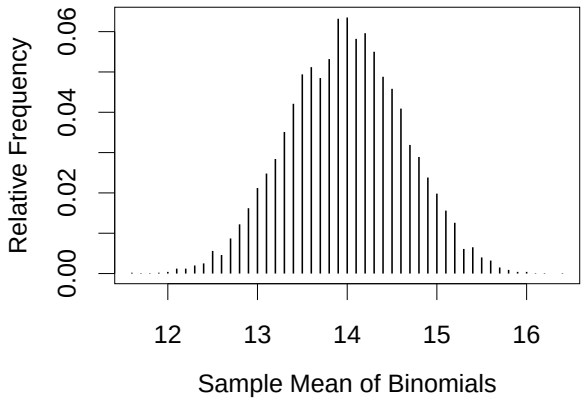

Figure 1.6: Relative Frequency plot of 10,000 calls to the function `bsms()`. for this plot, `bsms()` computed the mean of 10 realizations from binomials with $N = 20$ and $p = .7$

The above code returns a single number as output: the sample mean of 10 binomials. Since the sample mean is an estimator, it has a sampling distribution. The `bsms()` function returns one realization of the sample mean. If we are interested in the sampling distribution of the sample mean, we need to sample it many times and plot the results in a relative frequency histogram. This can be done by repeatedly calling `bsms()`. Here is the code for 10,000 replicates of `bsms()`:

```
M=10000
bsms.realization=1:M  #define the vector ppes.realization
for(m in 1:M) bsms.realization[m]=bsms(10,20,.7)
bsms.props=table(bsms.realization)/M
plot(ppes.props, xlab="Estimate of Expected Value (Sample Mean)",
ylab="Relative Frequency", type='h')
```

The resulting histogram is shown in Figure 1.6. The new programming element is the `for` loop. Within it, function `bsms()` is called `M` times, each result being stored to a different element of `bsms.realization`. However, we cannot reference elements in a vector without first reserving space. The line `bsms.realization=1:M` defines the vector, and in the process, reserves space for it.

Problem 1.5.3 (Your Turn)

1. What is the expected value of the sample mean of ten binomial random variables with $N = 20$ and $p = .5$? What is the approximate value from the above simulation? Are the values close? What is the simulation approximation for the standard error?
2. Manipulate the number of trials, N, in each binomial RV through a few levels: 5 trials, 20 trials, 80 trials. What is the effect on the sampling distribution of $\bar{Y}$?
3. Manipulate the number of random variables in the sample mean though a few levels: e.g., a mean of 4, 10, or 50 binomials. What is the effect on the sampling distribution $\bar{Y}$?
4. What is the effect of raising or lower the number of replicates `M`?

1.6 Three Binomial Probability Estimators

Consider the three following estimators for p: $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$.

$$\hat{p}_0 = \frac{Y}{N}, \tag{1.11}$$

$$\hat{p}_1 = \frac{Y + .5}{N + 1}, \tag{1.12}$$

$$\hat{p}_2 = \frac{Y + 1}{N + 2}. \tag{1.13}$$

Let's use **R** to examine the properties of these three estimators for 10 flips with $p = .7$. The following code uses the simulation method. It draws 10,000 replicates from a binomial distribution and computes the value for each estimator for each replicate.

```
p=.7
N=10
z=rbinom(10000,N,p)
est.p0=z/N
est.p1=(z+.5)/(N+1)
est.p2=(z+1)/(N+2)
bias.p0=mean(est.p0)-p
rmse.p0=sqrt(mean((est.p0-p)^2))
bias.p1=mean(est.p1)-p
rmse.p1=sqrt(mean((est.p1-p)^2))
bias.p2=mean(est.p2)-p
rmse.p2=sqrt(mean((est.p2-p)^2))
```

Figure 1.7 shows the sampling distributions for the three estimators. These sampling distributions tend to be roughly centered around the true value of the parameter, $p = .7$. Estimator $\hat{p}_2$ is the least spread out, followed by $\hat{p_1}$ and $\hat{p_0}$. Bias and efficiency of the estimators are indicated. Although estimator $\hat{p_0}$ is unbiased, it is also the least efficient! Figure 1.8 shows bias and efficiency for all three estimators for the full range of p. The conventional estimator $\hat{p}_0$ is unbiased for all true values of p, but the other two estimators are biased for extreme probabilities. None of the estimators are always more efficient than the others. For intermediate probabilities, estimator $\hat{p}_2$ is most efficient; for extreme probabilities, estimator $\hat{p}_0$ is most efficient. Typically, researchers have some idea of what types of probabilities of success to expect in their experiments. This knowledge can therefore be used to help pick the best estimator for a particular situation. We recommend $\hat{p}_1$ as a versatile alternative to $\hat{p}_0$ for many applications even though it is not the common-sense estimator.

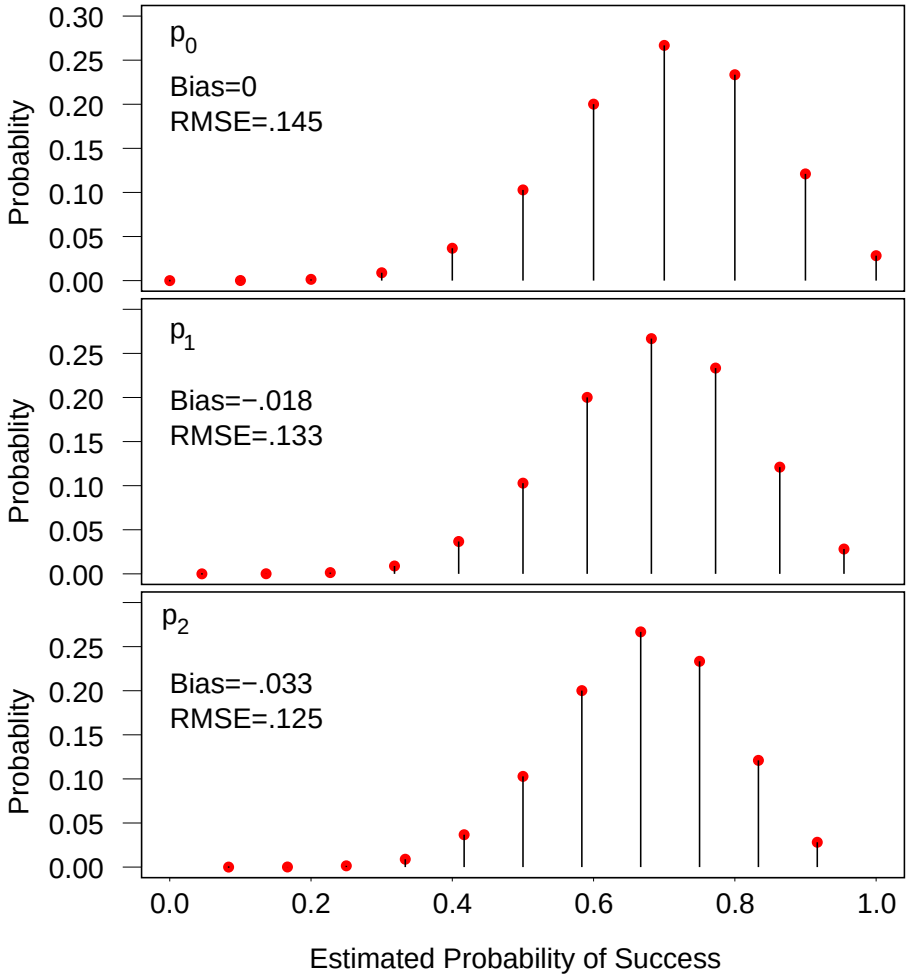

Figure 1.7: Sampling distribution of $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$. Bias and root-mean-squared-error (RMSE) are included. This figure depicts the case that there are $N = 10$ trials with a $p = .7$.

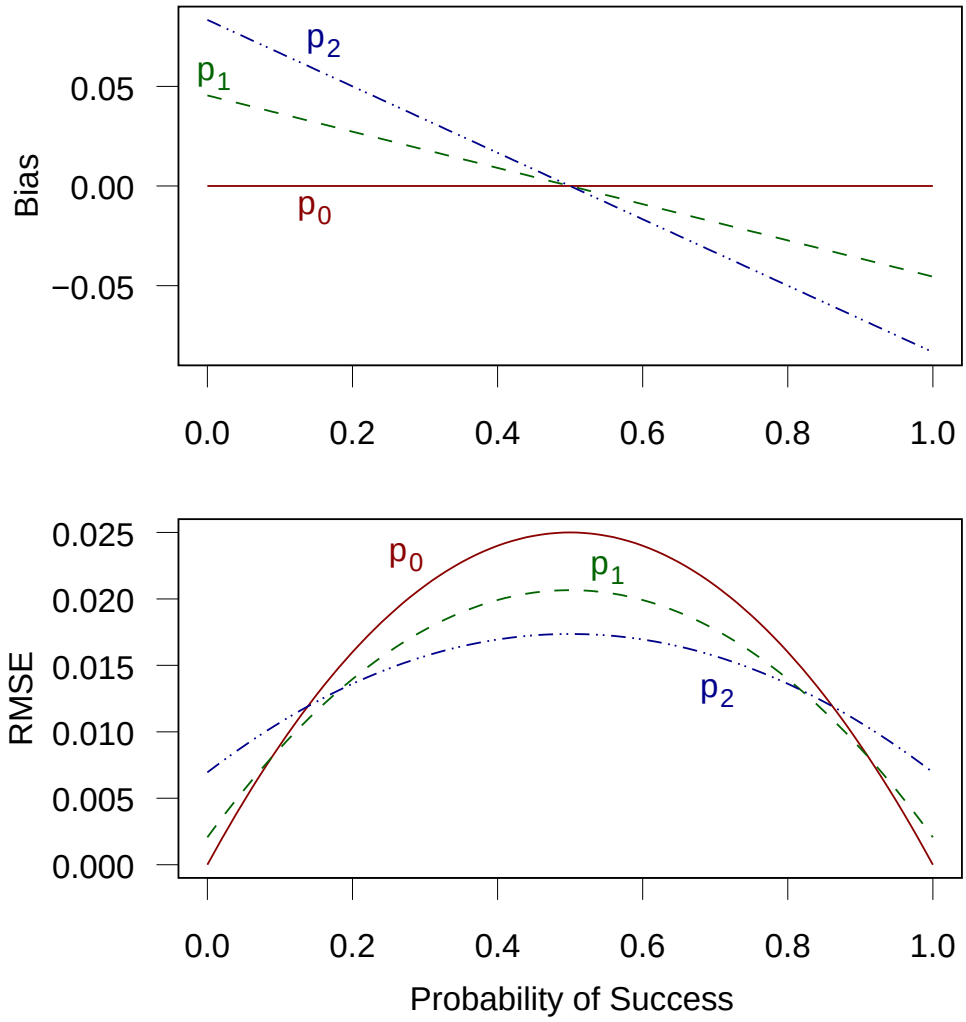

Figure 1.8: Bias and root-mean-squared-error (RMSE) for the three estimators as a function of true probability. Solid, dashed, and dashed-dotted lines denote the characteristics of $\hat{p}_0$, $\hat{p}_1$, and $\hat{p}_2$, respectively.

Problem 1.6.1 (Your Turn)
The estimators of the binomial probability parameter discussed above all have the form $(Y + a)/(N + 2a)$. We have advocated using the estimator $\hat{p}_1 = (Y + .5)/(N + 1)$, but there are many other possible estimators besides $a = .5$.. Examine what happens to the efficiency of an estimator as a gets large. Why would we choose $a = .5$ over, say, $a = 20$?