3

Random variables and their distributions

In this chapter, we introduce *random variables*, an incredibly useful concept that simplifies notation and expands our ability to quantify uncertainty and summarize the results of experiments. Random variables are essential throughout the rest of this book, and throughout statistics, so it is crucial to think through what they mean, both intuitively and mathematically.

3.1 Random variables

To see why our current notation can quickly become unwieldy, consider again the gambler's ruin problem from Chapter 2. In this problem, we may be very interested in how much wealth each gambler has at any particular time. So we could make up notation like letting A_{jk} be the event that gambler A has exactly j dollars after k rounds, and similarly defining an event B_{jk} for gambler B, for all j and k.

This is already too complicated. Furthermore, we may also be interested in other quantities, such as the difference in their wealths (gambler A's minus gambler B's) after k rounds, or the duration of the game (the number of rounds until one player is bankrupt). Expressing the event "the duration of the game is r rounds" in terms of the A_{jk} and B_{jk} would involve a long, awkward string of unions and intersections. And then what if we want to express gambler A's wealth as the equivalent amount in euros rather than dollars? We can multiply a *number* in dollars by a currency exchange rate, but we can't multiply an *event* by an exchange rate.

Instead of having convoluted notation that obscures how the quantities of interest are related, wouldn't it be nice if we could say something like the following?

> Let X_k be the wealth of gambler A after k rounds. Then $Y_k = N - X_k$ is the wealth of gambler B after k rounds (where N is the fixed total wealth); $X_k - Y_k = 2X_k - N$ is the difference in wealths after k rounds; $c_k X_k$ is the wealth of gambler A in euros after k rounds, where c_k is the euros per dollar exchange rate after k rounds; and the duration is $R = \min\{n : X_n = 0 \text{ or } Y_n = 0\}$.

The notion of a random variable will allow us to do exactly this! It needs to be introduced carefully though, to make it both conceptually and technically correct. Sometimes a definition of "random variable" is given that is a barely paraphrased version of "a random variable is a variable that takes on random values", but such a feeble attempt at a definition fails to say where the randomness come from. Nor does it help us to derive properties of random variables: we're familiar with working with algebraic equations like $x^2 + y^2 = 1$, but what are the valid mathematical operations if x and y are *random* variables? To make the notion of random variable precise, we define it as a *function* mapping the sample space to the real line. (See the math appendix for review of some concepts about functions.)

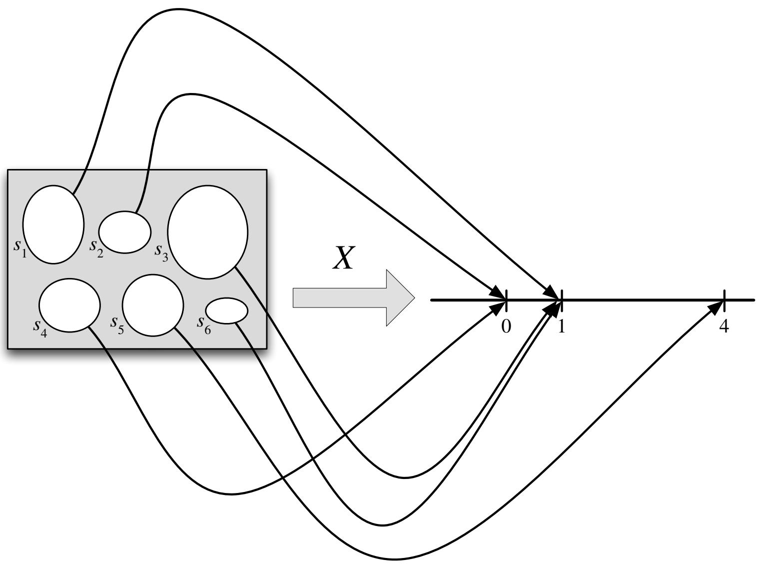

FIGURE 3.1
A random variable maps the sample space into the real line. The r.v. X depicted here is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The randomness comes from choosing a random pebble according to the probability function P for the sample space.

Definition 3.1.1 (Random variable)**.** Given an experiment with sample space S, a *random variable* (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$. It is common, but not required, to denote random variables by capital letters.

Thus, a random variable X assigns a numerical value $X(s)$ to each possible outcome s of the experiment. The randomness comes from the fact that we have a random experiment (with probabilities described by the probability function P); the mapping itself is deterministic, as illustrated in Figure 3.1. The same r.v. is shown in a simpler way in the left panel of Figure 3.2, in which we inscribe the values inside the pebbles.

This definition is abstract but fundamental; one of the most important skills to develop when studying probability and statistics is the ability to go back and forth between abstract ideas and concrete examples. Relatedly, it is important to work on recognizing the essential pattern or structure of a problem and how it connects to problems you have studied previously. We will often discuss stories that involve tossing coins or drawing balls from urns because they are simple, convenient scenarios to work with, but many other problems are *isomorphic*: they have the same essential structure, but in a different guise.

To start, let's consider a coin-tossing example. The structure of the problem is that we have a sequence of trials where there are two possible outcomes for each trial. Here we think of the possible outcomes as H (Heads) and T (Tails), but we could just as well think of them as "success" and "failure" or as 1 and 0, for example.

Example 3.1.2 (Coin tosses)**.** Consider an experiment where we toss a fair coin twice. The sample space consists of four possible outcomes: $S = \{HH, HT, TH, TT\}$. Here are some random variables on this space (for practice, you can think up some of your own). Each r.v. is a numerical summary of some aspect of the experiment.

- Let X be the number of Heads. This is a random variable with possible values 0, 1, and 2. Viewed as a function, X assigns the value 2 to the outcome HH, 1 to the outcomes HT and TH, and 0 to the outcome TT. That is,

$$X(HH) = 2, X(HT) = X(TH) = 1, X(TT) = 0.$$

- Let Y be the number of Tails. In terms of X, we have $Y = 2 - X$. In other words, Y and $2 - X$ are the same r.v.: $Y(s) = 2 - X(s)$ for all s.
- Let I be 1 if the first toss lands Heads and 0 otherwise. Then I assigns the value 1 to the outcomes HH and HT and 0 to the outcomes TH and TT. This r.v. is an example of what is called an *indicator random variable* since it indicates whether the first toss lands Heads, using 1 to mean "yes" and 0 to mean "no".

We can also encode the sample space as $\{(1,1), (1,0), (0,1), (0,0)\}$, where 1 is the code for Heads and 0 is the code for Tails. Then we can give explicit formulas for X, Y, I:

$$X(s_1, s_2) = s_1 + s_2,\ Y(s_1, s_2) = 2 - s_1 - s_2,\ I(s_1, s_2) = s_1,$$

where for simplicity we write $X(s_1, s_2)$ to mean $X((s_1, s_2))$, etc.

For most r.v.s we will consider, it is tedious or infeasible to write down an explicit formula in this way. Fortunately, it is usually unnecessary to do so, since (as we saw in this example) there are other ways to define an r.v., and (as we will see throughout the rest of this book) there are many ways to study the properties of an r.v. other than by doing computations with an explicit formula for what it maps each outcome s to. $\square$

As in the previous chapters, for a sample space with a finite number of outcomes we can visualize the outcomes as pebbles, with the mass of a pebble corresponding to its probability, such that the total mass of the pebbles is 1. A random variable simply labels each pebble with a number. Figure 3.2 shows two random variables defined on the same sample space: the pebbles or outcomes are the same, but the real numbers assigned to the outcomes are different.

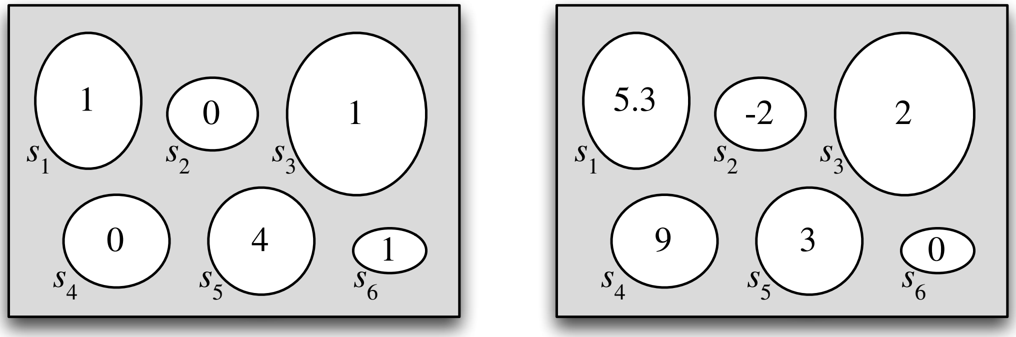

FIGURE 3.2
Two random variables defined on the same sample space.

As we've mentioned earlier, the source of the randomness in a random variable is the experiment itself, in which a sample outcome $s \in S$ is chosen according to a probability function P. Before we perform the experiment, the outcome s has not yet been realized, so we don't know the value of X, though we could calculate the probability that X will take on a given value or range of values. After we perform the experiment and the outcome s has been realized, the random variable crystallizes into the numerical value $X(s)$.

Random variables provide *numerical* summaries of the experiment in question. This is very handy because the sample space of an experiment is often incredibly complicated or high-dimensional, and the outcomes $s \in S$ may be non-numeric. For example, the experiment may be to collect a random sample of people in a certain city and ask them various questions, which may have numeric (e.g., age or height) or non-numeric (e.g., political party or favorite movie) answers. The fact that r.v.s take on numerical values is a very convenient simplification compared to having to work with the full complexity of S at all times.

3.2 Distributions and probability mass functions

There are two main types of random variables used in practice: *discrete* r.v.s and *continuous* r.v.s. In this chapter and the next, our focus is on discrete r.v.s. Continuous r.v.s are introduced in Chapter 5.

Definition 3.2.1 (Discrete random variable). A random variable X is said to be *discrete* if there is a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite list of values $a_1, a_2, \ldots$ such that $P(X = a_j \text{ for some } j) = 1$. If X is a discrete r.v., then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the *support* of X.

Most commonly in applications, the support of a discrete r.v. is a set of integers. In contrast, a *continuous* r.v. can take on any real value in an interval (possibly even the entire real line); such r.v.s are defined more precisely in Chapter 5. It is also possible to have an r.v. that is a hybrid of discrete and continuous, such as by flipping a coin and then generating a discrete r.v. if the coin lands Heads and generating a continuous r.v. if the coin lands Tails. But the starting point for understanding such r.v.s is to understand discrete and continuous r.v.s.

Given a random variable, we would like to be able to describe its behavior using the language of probability. For example, we might want to answer questions about the probability that the r.v. will fall into a given range: if L is the lifetime earnings of a randomly chosen U.S. college graduate, what is the probability that L exceeds a million dollars? If M is the number of major earthquakes in California in the next five years, what is the probability that M equals 0?

The *distribution* of a random variable provides the answers to these questions; it specifies the probabilities of all events associated with the r.v., such as the probability of it equaling 3 and the probability of it being at least 110. We will see that there are several equivalent ways to express the distribution of an r.v. For a discrete r.v., the most natural way to do so is with a *probability mass function*, which we now define.

Definition 3.2.2 (Probability mass function)**.** The *probability mass function* (PMF) of a discrete r.v. X is the function p_X given by $p_X(x) = P(X = x)$. Note that this is positive if x is in the support of X, and 0 otherwise.

☣ **3.2.3.** In writing $P(X = x)$, we are using $X = x$ to denote an *event*, consisting of all outcomes s to which X assigns the number x. This event is also written as $\{X = x\}$; formally, $\{X = x\}$ is defined as $\{s \in S : X(s) = x\}$, but writing $\{X = x\}$ is shorter and more intuitive. Going back to Example 3.1.2, if X is the number of Heads in two fair coin tosses, then $\{X = 1\}$ consists of the sample outcomes HT and TH, which are the two outcomes to which X assigns the number 1. Since $\{HT, TH\}$ is a subset of the sample space, it is an event. So it makes sense to talk about $P(X = 1)$, or more generally, $P(X = x)$. If $\{X = x\}$ were anything other than an event, it would make no sense to calculate its probability! It does not make sense to write "$P(X)$"; we can only take the probability of an event, not of an r.v.

Let's look at a few examples of PMFs.

Example 3.2.4 (Coin tosses continued)**.** In this example we'll find the PMFs of all the random variables in Example 3.1.2, the example with two fair coin tosses. Here are the r.v.s we defined, along with their PMFs:

- X, the number of Heads. Since X equals 0 if TT occurs, 1 if HT or TH occurs, and 2 if HH occurs, the PMF of X is the function p_X given by

$$p_X(0) = P(X = 0) = 1/4,$$
$$p_X(1) = P(X = 1) = 1/2,$$
$$p_X(2) = P(X = 2) = 1/4,$$

and $p_X(x) = 0$ for all other values of x.

- $Y = 2 - X$, the number of Tails. Reasoning as above or using the fact that

$$P(Y = y) = P(2 - X = y) = P(X = 2 - y) = p_X(2 - y),$$

the PMF of Y is

$$p_Y(0) = P(Y = 0) = 1/4,$$
$$p_Y(1) = P(Y = 1) = 1/2,$$
$$p_Y(2) = P(Y = 2) = 1/4,$$

and $p_Y(y) = 0$ for all other values of y.

Note that X and Y have the same PMF (that is, p_X and p_Y are the same function) even though X and Y are not the same r.v. (that is, X and Y are two different functions from $\{HH, HT, TH, TT\}$ to the real line).

- I, the indicator of the first toss landing Heads. Since I equals 0 if TH or TT occurs and 1 if HH or HT occurs, the PMF of I is

$$p_I(0) = P(I = 0) = 1/2,$$
$$p_I(1) = P(I = 1) = 1/2,$$

and $p_I(i) = 0$ for all other values of i.

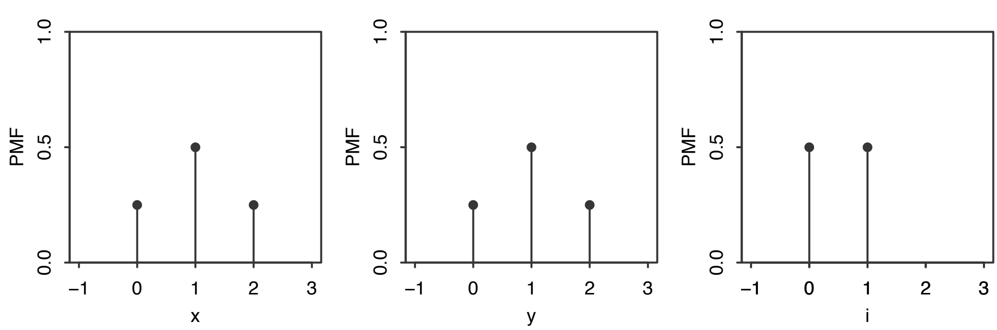

FIGURE 3.3
Left to right: PMFs of X, Y, and I, with X the number of Heads in two fair coin tosses, Y the number of Tails, and I the indicator of Heads on the first toss.

The PMFs of X, Y, and I are plotted in Figure 3.3. Vertical bars are drawn to make it easier to compare the heights of different points. $\square$

Example 3.2.5 (Sum of die rolls). We roll two fair 6-sided dice. Let $T = X + Y$ be the total of the two rolls, where X and Y are the individual rolls. The sample space of this experiment has 36 equally likely outcomes:

$$S = \{(1, 1), (1, 2), \ldots, (6, 5), (6, 6)\}.$$

For example, 7 of the 36 outcomes s are shown in the table below, along with the corresponding values of X, Y, and T. After the experiment is performed, we observe values for X and Y, and then the observed value of T is the sum of those values.

s	X	Y	$X + Y$
$(1, 2)$	1	2	3
$(1, 6)$	1	6	7
$(2, 5)$	2	5	7
$(3, 1)$	3	1	4
$(4, 3)$	4	3	7
$(5, 4)$	5	4	9
$(6, 6)$	6	6	12

Since the dice are fair, the PMF of X is

$$P(X = j) = 1/6,$$

for $j = 1, 2, \ldots, 6$ (and $P(X = j) = 0$ otherwise); we say that X has a *Discrete Uniform* distribution on $1, 2, \ldots, 6$. Similarly, Y is also Discrete Uniform on $1, 2, \ldots, 6$.

Note that Y has the same *distribution* as X but is not the same *random variable* as X. In fact, we have

$$P(X = Y) = 6/36 = 1/6.$$

Two more r.v.s in this experiment with the same distribution as X are $7 - X$ and $7 - Y$. To see this, we can use the fact that for a standard die, $7 - X$ is the value on the bottom if X is the value on the top. If the top value is equally likely to be any of the numbers $1, 2, \ldots, 6$, then so is the bottom value. Note that even though $7 - X$ has the same distribution as X, it is *never* equal to X in a run of the experiment!

Let's now find the PMF of T. By the naive definition of probability,

$$\begin{aligned}
P(T = 2) &= P(T = 12) = 1/36,\\
P(T = 3) &= P(T = 11) = 2/36,\\
P(T = 4) &= P(T = 10) = 3/36,\\
P(T = 5) &= P(T = 9) = 4/36,\\
P(T = 6) &= P(T = 8) = 5/36,\\
P(T = 7) &= 6/36.
\end{aligned}$$

For all other values of t, $P(T = t) = 0$. We can see directly that the support of T

is $\{2, 3, \ldots, 12\}$ just by looking at the possible totals for two dice, but as a check, note that

$$P(T = 2) + P(T = 3) + \cdots + P(T = 12) = 1,$$

which shows that all possibilities have been accounted for. The symmetry property of T that appears above, $P(T = t) = P(T = 14-t)$, makes sense since each outcome $\{X = x, Y = y\}$ which makes $T = t$ has a corresponding outcome $\{X = 7 - x, Y = 7 - y\}$ of the same probability which makes $T = 14 - t$.

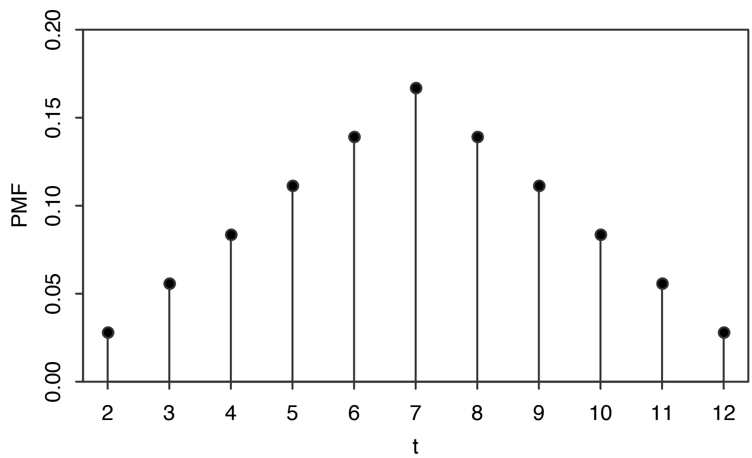

FIGURE 3.4
PMF of the sum of two die rolls.

The PMF of T is plotted in Figure 3.4; it has a triangular shape, and the symmetry noted above is very visible. □

Example 3.2.6 (Children in a U.S. household)**.** Suppose we choose a household in the United States at random. Let X be the number of children in the chosen household. Since X can only take on integer values, it is a discrete r.v. The probability that X takes on the value x is proportional to the number of households in the United States with x children.

Using data from the 2010 General Social Survey [23], we can approximate the proportion of households with 0 children, 1 child, 2 children, etc., and hence approximate the PMF of X, which is plotted in Figure 3.5. □

We will now state the properties of a valid PMF.

Theorem 3.2.7 (Valid PMFs)**.** Let X be a discrete r.v. with support $x_1, x_2, \ldots$ (assume these values are distinct and, for notational simplicity, that the support is countably infinite; the analogous results hold if the support is finite). The PMF p_X of X must satisfy the following two criteria:

- Nonnegative: $p_X(x) > 0$ if $x = x_j$ for some j, and $p_X(x) = 0$ otherwise;
- Sums to 1: $\sum_{j=1}^{\infty} p_X(x_j) = 1$.

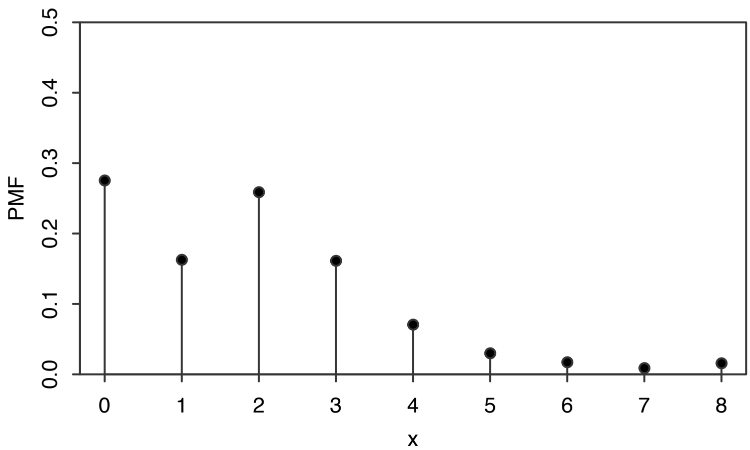

FIGURE 3.5
PMF of the number of children in a randomly selected U.S. household.

Proof. The first criterion is true since probability is nonnegative. The second is true since X must take on *some* value, and the events $\{X = x_j\}$ are disjoint, so

$$\sum_{j=1}^{\infty} P(X = x_j) = P\left(\bigcup_{j=1}^{\infty}\{X = x_j\}\right) = P(X = x_1 \text{ or } X = x_2 \text{ or } \ldots) = 1. \quad \blacksquare$$

Conversely, if distinct values $x_1, x_2, \ldots$ are specified and we have a function satisfying the two criteria above, then this function *is* the PMF of some r.v.; we will show how to construct such an r.v. in Chapter 5.

We claimed earlier that the PMF is one way of expressing the distribution of a discrete r.v. This is because once we know the PMF of X, we can calculate the probability that X will fall into a given subset of the real numbers by summing over the appropriate values of x, as the next example shows.

Example 3.2.8. Returning to Example 3.2.5, let T be the sum of two fair die rolls. We have already calculated the PMF of T. Now suppose we're interested in the probability that T is in the interval $[1, 4]$. There are only three values in the interval $[1, 4]$ that T can take on, namely, 2, 3, and 4. We know the probability of each of these values from the PMF of T, so

$$P(1 \leq T \leq 4) = P(T = 2) + P(T = 3) + P(T = 4) = 6/36. \quad \square$$

In general, given a discrete r.v. X and a set B of real numbers, if we know the PMF of X we can find $P(X \in B)$, the probability that X is in B, by summing up the heights of the vertical bars at points in B in the plot of the PMF of X. *Knowing the PMF of a discrete r.v. determines its distribution.*