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.)

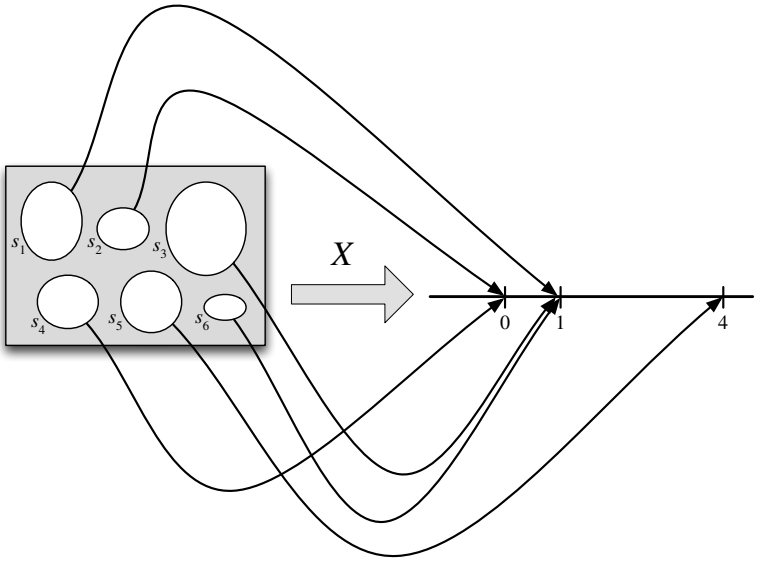


FIGURE 3.1
A random variable maps the sample space into the real line. The r.v. X depicted here is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The randomness comes from choosing a random pebble according to the probability function P for the sample space.

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. □

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.

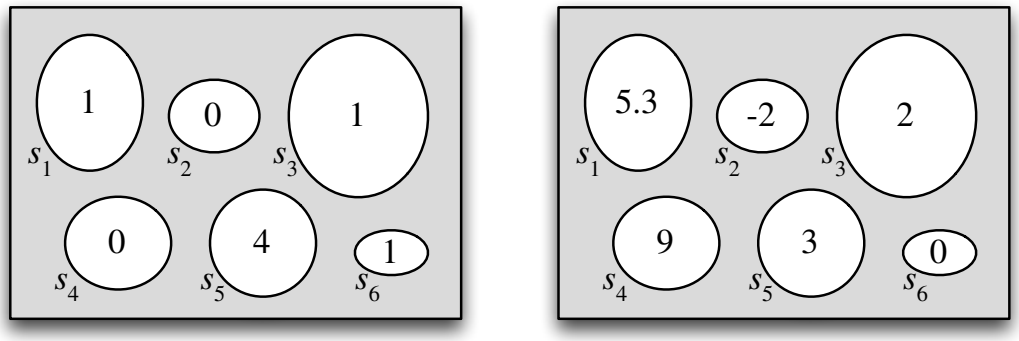


FIGURE 3.2
Two random variables defined on the same sample space.

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.

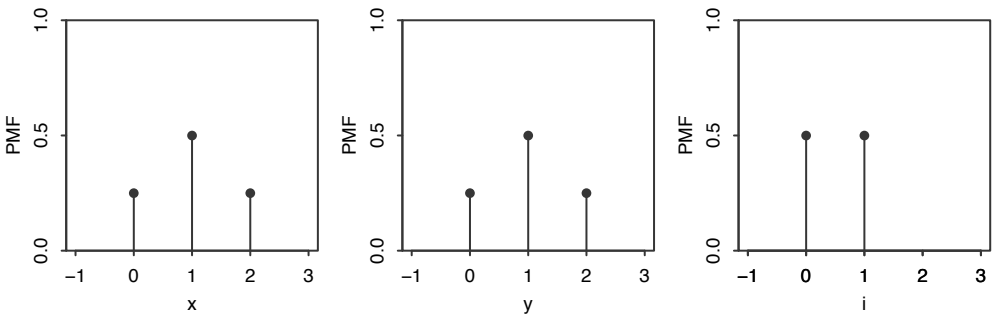


FIGURE 3.3
Left to right: PMFs of X, Y, and I, with X the number of Heads in two fair coin tosses, Y the number of Tails, and I the indicator of Heads on the first toss.

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. □

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 $T = 14 - t$.

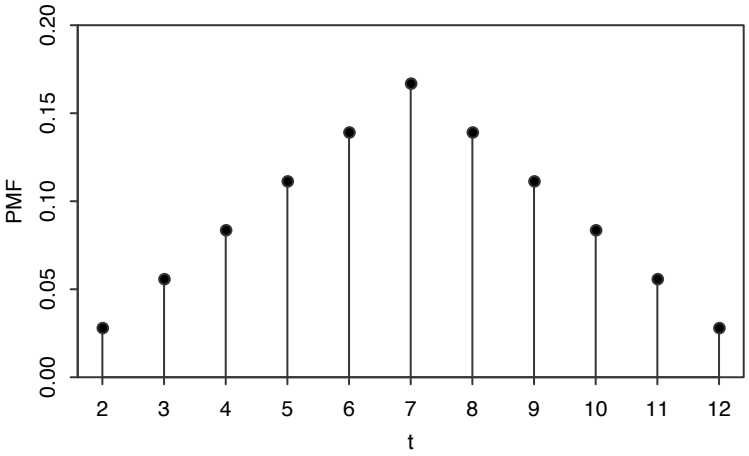


FIGURE 3.4
PMF of the sum of two die rolls.

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: $\sum_{j=1}^{\infty} p_X(x_j) = 1$.

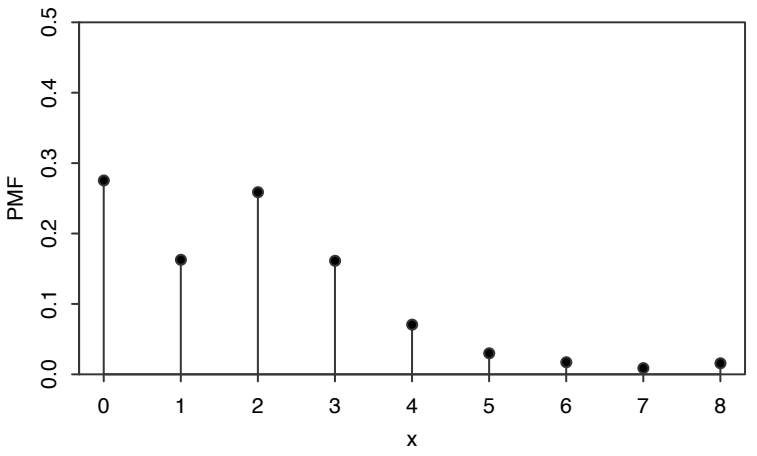


FIGURE 3.5
PMF of the number of children in a randomly selected U.S. household.

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.*

3.3 Bernoulli and Binomial

Some distributions are so ubiquitous in probability and statistics that they have their own names. We will introduce these *named distributions* throughout the book, starting with a very simple but useful case: an r.v. that can take on only two possible values, 0 and 1.

Definition 3.3.1 (Bernoulli distribution)**.** An r.v. X is said to have the *Bernoulli distribution* with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$, where $0 < p < 1$. We write this as $X \sim \text{Bern}(p)$. The symbol $\sim$ is read "is distributed as".

Any r.v. whose possible values are 0 and 1 has a Bern(p) distribution, with p the probability of the r.v. equaling 1. This number p in Bern(p) is called the *parameter* of the distribution; it determines which specific Bernoulli distribution we have. Thus there is not just one Bernoulli distribution, but rather a *family* of Bernoulli distributions, indexed by p. For example, if $X \sim \text{Bern}(1/3)$, it would be correct but incomplete to say "X is Bernoulli"; to fully specify the distribution of X, we should both say its name (Bernoulli) and its parameter value (1/3), which is the point of the notation $X \sim \text{Bern}(1/3)$.

Any event has a Bernoulli r.v. that is naturally associated with it, equal to 1 if the event happens and 0 otherwise. This is called the *indicator random variable* of the event; we will see that such r.v.s are extremely useful.

Definition 3.3.2 (Indicator random variable)**.** The *indicator random variable* of an event A is the r.v. which equals 1 if A occurs and 0 otherwise. We will denote the indicator r.v. of A by I_A or $I(A)$. Note that $I_A \sim \text{Bern}(p)$ with $p = P(A)$.

We often imagine Bernoulli r.v.s using coin tosses, but this is just convenient language for discussing the following general story.

Story 3.3.3 (Bernoulli trial)**.** An experiment that can result in either a "success" or a "failure" (but not both) is called a *Bernoulli trial*. A Bernoulli random variable can be thought of as the *indicator of success* in a Bernoulli trial: it equals 1 if success occurs and 0 if failure occurs in the trial. □

Because of this story, the parameter p is often called the *success probability* of the Bern(p) distribution. Once we start thinking about Bernoulli trials, it's hard not to start thinking about what happens when we have more than one trial.

Story 3.3.4 (Binomial distribution)**.** Suppose that n *independent* Bernoulli trials are performed, each with the same success probability p. Let X be the number of successes. The distribution of X is called the *Binomial distribution* with parameters n and p. We write $X \sim \text{Bin}(n, p)$ to mean that X has the Binomial distribution with parameters n and p, where n is a positive integer and $0 < p < 1$. □

Notice that we define the Binomial distribution not by its PMF, but by a *story*

about the type of experiment that could give rise to a random variable with a Binomial distribution. The most famous distributions in statistics all have stories which explain why they are so often used as models for data, or as the building blocks for more complicated distributions.

Thinking about the named distributions first and foremost in terms of their stories has many benefits. It facilitates pattern recognition, allowing us to see when two problems are essentially identical in structure; it often leads to cleaner solutions that avoid PMF calculations altogether; and it helps us understand how the named distributions are connected to one another. Here it is clear that $\text{Bern}(p)$ is the same distribution as $\text{Bin}(1, p)$: the Bernoulli is a special case of the Binomial.

Using the story definition of the Binomial, let's find its PMF.

Theorem 3.3.5 (Binomial PMF)**.** If $X \sim \text{Bin}(n, p)$, then the PMF of X is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$ (and $P(X = k) = 0$ otherwise).

☣ **3.3.6.** To save writing, it is often left implicit that a PMF is zero wherever it is not specified to be nonzero, but in any case it is important to understand what the support of a random variable is, and good practice to check that PMFs are valid. If two discrete r.v.s have the same PMF, then they also must have the same support. So we sometimes refer to the support of a discrete *distribution*; this is the support of any r.v. with that distribution.

Proof. An experiment consisting of n independent Bernoulli trials produces a sequence of successes and failures. The probability of any specific sequence of k successes and $n - k$ failures is $p^k(1-p)^{n-k}$. There are $\binom{n}{k}$ such sequences, since we just need to select where the successes are. Therefore, letting X be the number of successes,

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

for $k = 0, 1, \ldots, n$, and $P(X = k) = 0$ otherwise. This is a valid PMF because it is nonnegative and it sums to 1 by the binomial theorem. ■

Figure 3.6 shows plots of the Binomial PMF for various values of n and p. Note that the PMF of the $\text{Bin}(10, 1/2)$ distribution is symmetric about 5, but when the success probability is not $1/2$, the PMF is *skewed*. For a fixed number of trials n, X tends to be larger when the success probability is high and lower when the success probability is low, as we would expect from the story of the Binomial distribution. Also recall that in any PMF plot, the sum of the heights of the vertical bars must be 1.

We've used Story 3.3.4 to find the $\text{Bin}(n, p)$ PMF. The story also gives us a straightforward proof of the fact that if X is Binomial, then $n - X$ is also Binomial.

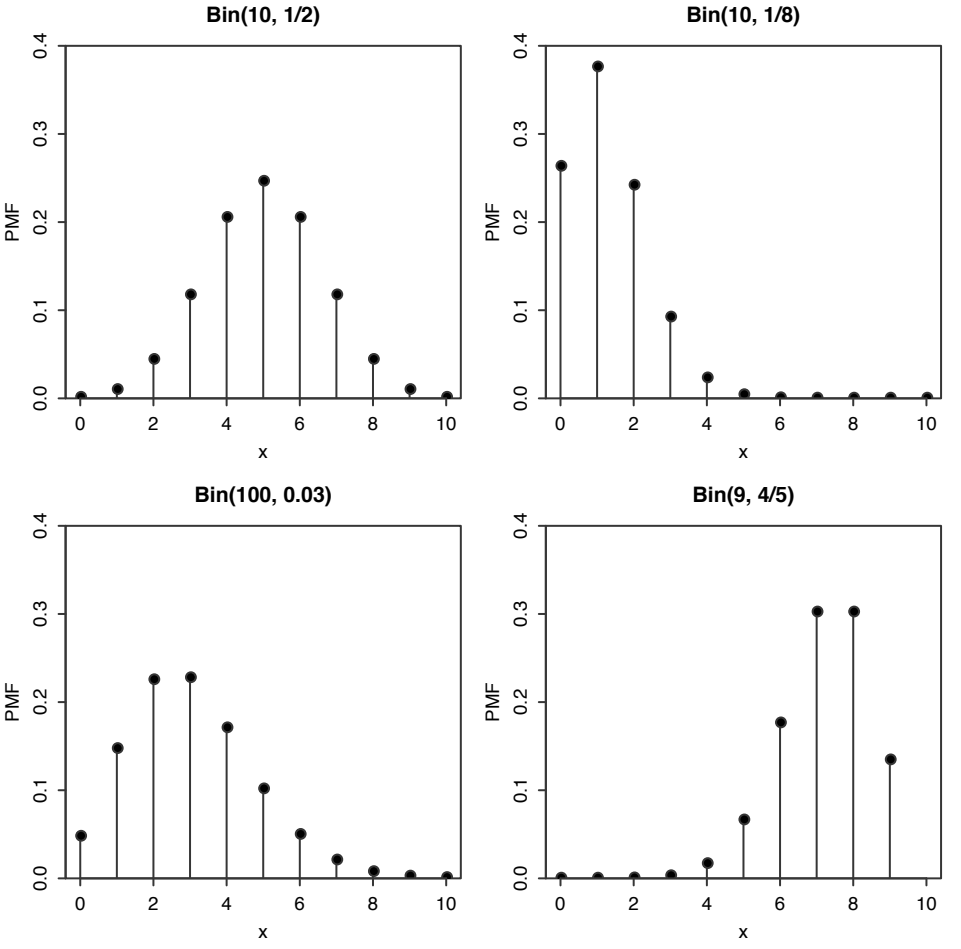


FIGURE 3.6
Some Binomial PMFs. In the lower left, we plot the Bin$(100, 0.03)$ PMF between 0 and 10 only, as the probability of more than 10 successes is close to 0.

Theorem 3.3.7. Let $X \sim \text{Bin}(n, p)$, and $q = 1 - p$ (we often use q to denote the failure probability of a Bernoulli trial). Then $n - X \sim \text{Bin}(n, q)$.

Proof. Using the story of the Binomial, interpret X as the number of successes in n independent Bernoulli trials. Then $n - X$ is the number of failures in those trials. Interchanging the roles of success and failure, we have $n - X \sim \text{Bin}(n, q)$. Alternatively, we can check that $n - X$ has the $\text{Bin}(n, q)$ PMF. Let $Y = n - X$. The PMF of Y is

$$P(Y = k) = P(X = n - k) = \binom{n}{n-k} p^{n-k} q^k = \binom{n}{k} q^k p^{n-k},$$

for $k = 0, 1, \ldots, n$. ■

Corollary 3.3.8. Let $X \sim \text{Bin}(n, p)$ with $p = 1/2$ and n even. Then the distribution of X is symmetric about $n/2$, in the sense that $P(X = n/2 + j) = P(X = n/2 - j)$ for all nonnegative integers j.

Proof. By Theorem 3.3.7, $n - X$ is also $\text{Bin}(n, 1/2)$, so

$$P(X = k) = P(n - X = k) = P(X = n - k)$$

for all nonnegative integers k. Letting $k = n/2 + j$, the desired result follows. This explains why the $\text{Bin}(10, 1/2)$ PMF is symmetric about 5 in Figure 3.6. ■

Example 3.3.9 (Coin tosses continued)**.** Going back to Example 3.1.2, we now know that $X \sim \text{Bin}(2, 1/2)$, $Y \sim \text{Bin}(2, 1/2)$, and $I \sim \text{Bern}(1/2)$. Consistent with Theorem 3.3.7, X and $Y = 2 - X$ have the same distribution, and consistent with Corollary 3.3.8, the distribution of X (and of Y) is symmetric about 1. □

3.4 Hypergeometric

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn *with replacement* yields a $\text{Bin}(n, w/(w + b))$ distribution for the number of white balls obtained in n trials, since the draws are independent Bernoulli trials, each with probability $w/(w+b)$ of success. If we instead sample *without replacement*, as illustrated in Figure 3.7, then the number of white balls follows a *Hypergeometric distribution*.

Story 3.4.1 (Hypergeometric distribution)**.** Consider an urn with w white balls and b black balls. We draw n balls out of the urn at random without replacement, such that all $\binom{w+b}{n}$ samples are equally likely. Let X be the number of white balls in the sample. Then X is said to have the *Hypergeometric distribution* with parameters w, b, and n; we denote this by $X \sim \text{HGeom}(w, b, n)$. □

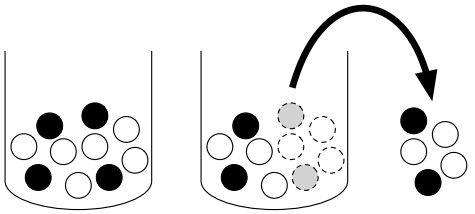


FIGURE 3.7
Hypergeometric story. An urn contains $w = 6$ white balls and $b = 4$ black balls. We sample $n = 5$ without replacement. The number X of white balls in the sample is Hypergeometric; here we observe $X = 3$.

As with the Binomial distribution, we can obtain the PMF of the Hypergeometric distribution from the story.

Theorem 3.4.2 (Hypergeometric PMF)**.** If $X \sim \text{HGeom}(w, b, n)$, then the PMF of X is

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}},$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n-k \leq b$, and $P(X = k) = 0$ otherwise.

Proof. To get $P(X = k)$, we first count the number of possible ways to draw exactly k white balls and $n - k$ black balls from the urn (without distinguishing between different orderings for getting the same set of balls). If $k > w$ or $n - k > b$, then the draw is impossible. Otherwise, there are $\binom{w}{k}\binom{b}{n-k}$ ways to draw k white and $n-k$ black balls by the multiplication rule, and there are $\binom{w+b}{n}$ total ways to draw n balls. Since all samples are equally likely, the naive definition of probability gives

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

for integers k satisfying $0 \leq k \leq w$ and $0 \leq n-k \leq b$. This PMF is valid because the numerator, summed over all k, equals $\binom{w+b}{n}$ by Vandermonde's identity (Example 1.5.3), so the PMF sums to 1. ■

The Hypergeometric distribution comes up in many scenarios which, on the surface, have little in common with white and black balls in an urn. The essential structure of the Hypergeometric story is that items in a population are classified using two sets of *tags*: in the urn story, each ball is either white or black (this is the first set of tags), and each ball is either sampled or not sampled (this is the second set of tags). Furthermore, at least one of these sets of tags is assigned completely at random (in the urn story, the balls are sampled randomly, with all sets of the correct size equally likely). Then $X \sim \text{HGeom}(w, b, n)$ represents the number of twice-tagged items: in the urn story, balls that are *both* white and sampled.

The next two examples show seemingly dissimilar scenarios that are nonetheless isomorphic to the urn story.

Example 3.4.3 (Elk capture-recapture)**.** A forest has N elk. Today, m of the elk are captured, tagged, and released into the wild. At a later date, n elk are recaptured at random. Assume that the recaptured elk are equally likely to be any set of n of the elk, e.g., an elk that has been captured does not learn how to avoid being captured again.

By the story of the Hypergeometric, the number of tagged elk in the recaptured sample is $\text{HGeom}(m, N-m, n)$. The m tagged elk in this story correspond to the white balls and the $N-m$ untagged elk correspond to the black balls. Instead of sampling n balls from the urn, we recapture n elk from the forest. □

Example 3.4.4 (Aces in a poker hand)**.** In a five-card hand drawn at random from a well-shuffled standard deck, the number of aces in the hand has the $\text{HGeom}(4, 48, 5)$ distribution, which can be seen by thinking of the aces as white balls and the non-aces as black balls. Using the Hypergeometric PMF, the probability that the hand has exactly three aces is

$$\frac{\binom{4}{3}\binom{48}{2}}{\binom{52}{5}} \approx 0.0017.$$ □

The following table summarizes how the above examples can be thought of in terms of two sets of tags. In each example, the r.v. of interest is the number of items falling into both the second and the fourth columns: white and sampled, tagged and recaptured, ace and in one's hand.

Story	**First set of tags**		**Second set of tags**	
urn	white	black	sampled	not sampled
elk	tagged	untagged	recaptured	not recaptured
cards	ace	not ace	in hand	not in hand

The next theorem describes a symmetry between two Hypergeometric distributions with different parameters; the proof follows from *swapping* the two sets of tags in the Hypergeometric story.

Theorem 3.4.5. The $\text{HGeom}(w, b, n)$ and $\text{HGeom}(n, w+b-n, w)$ distributions are identical. That is, if $X \sim \text{HGeom}(w, b, n)$ and $Y \sim \text{HGeom}(n, w+b-n, w)$, then X and Y have the same distribution.

Proof. Using the story of the Hypergeometric, imagine an urn with w white balls, b black balls, and a sample of size n made without replacement. Let $X \sim \text{HGeom}(w, b, n)$ be the number of white balls in the sample, thinking of white/black as the first set of tags and sampled/not sampled as the second set of tags. Let $Y \sim \text{HGeom}(n, w+b-n, w)$ be the number of sampled balls among the white balls, thinking of sampled/not sampled as the first set of tags and white/black as

the second set of tags. Both X and Y count the number of white sampled balls, so they have the same distribution.

Alternatively, we can check algebraically that X and Y have the same PMF:

$$P(X=k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} = \frac{w!b!n!(w+b-n)!}{k!(w+b)!(w-k)!(n-k)!(b-n+k)!},$$

$$P(Y=k) = \frac{\binom{n}{k}\binom{w+b-n}{w-k}}{\binom{w+b}{w}} = \frac{w!b!n!(w+b-n)!}{k!(w+b)!(w-k)!(n-k)!(b-n+k)!}.$$

We prefer the story proof because it is less tedious and more memorable. ■

☣ **3.4.6** (Binomial vs. Hypergeometric)**.** The Binomial and Hypergeometric distributions are often confused. Both are discrete distributions taking on integer values between 0 and n for some n, and both can be interpreted as the number of successes in n Bernoulli trials (for the Hypergeometric, each tagged elk in the recaptured sample can be considered a success and each untagged elk a failure). However, a crucial part of the Binomial story is that the Bernoulli trials involved are *independent*. The Bernoulli trials in the Hypergeometric story are *dependent*, since the sampling is done without replacement: knowing that one elk in our sample is tagged decreases the probability that the second elk will also be tagged.

3.5 Discrete Uniform

A very simple story, closely connected to the naive definition of probability, describes picking a random number from some finite set of possibilities.

Story 3.5.1 (Discrete Uniform distribution)**.** Let C be a finite, nonempty set of numbers. Choose one of these numbers uniformly at random (i.e., all values in C are equally likely). Call the chosen number X. Then X is said to have the *Discrete Uniform distribution* with parameter C; we denote this by $X \sim \text{DUnif}(C)$. □

The PMF of $X \sim \text{DUnif}(C)$ is

$$P(X=x) = \frac{1}{|C|}$$

for $x \in C$ (and 0 otherwise), since a PMF must sum to 1. As with questions based on the naive definition of probability, questions based on a Discrete Uniform distribution reduce to counting problems. Specifically, for $X \sim \text{DUnif}(C)$ and any $A \subseteq C$, we have

$$P(X \in A) = \frac{|A|}{|C|}.$$

Example 3.5.2 (Random slips of paper)**.** There are 100 slips of paper in a hat, each of which has one of the numbers $1, 2, \ldots, 100$ written on it, with no number appearing more than once. Five of the slips are drawn, one at a time.

First consider random sampling with *replacement (with equal probabilities).*

(a) What is the distribution of how many of the drawn slips have a value of at least 80 written on them?

(b) What is the distribution of the value of the jth draw (for $1 \leq j \leq 5$)?

(c) What is the probability that the number 100 is drawn at least once?

Now consider random sampling without *replacement (with all sets of five slips equally likely to be chosen).*

(d) What is the distribution of how many of the drawn slips have a value of at least 80 written on them?

(e) What is the distribution of the value of the jth draw (for $1 \leq j \leq 5$)?

(f) What is the probability that the number 100 is drawn in the sample?

Solution:

(a) By the story of the Binomial, the distribution is $\text{Bin}(5, 0.21)$.

(b) Let X_j be the value of the jth draw. By symmetry, $X_j \sim \text{DUnif}(1, 2, \ldots, 100)$. There aren't certain slips that love being chosen on the jth draw and others that avoid being chosen then; all are equally likely.

(c) Taking complements,

$$P(X_j = 100 \text{ for at least one } j) = 1 - P(X_1 \neq 100, \ldots, X_5 \neq 100).$$

By the naive definition of probability, this is

$$1 - (99/100)^5 \approx 0.049.$$

This solution just uses new notation for concepts from Chapter 1. It is useful to have this new notation since it is compact and flexible. In the above calculation, it is important to see why

$$P(X_1 \neq 100, \ldots, X_5 \neq 100) = P(X_1 \neq 100) \ldots P(X_5 \neq 100).$$

This follows from the naive definition in this case, but a more general way to think about such statements is through *independence* of r.v.s, a concept discussed in detail in Section 3.8.

(d) By the story of the Hypergeometric, the distribution is $\text{HGeom}(21, 79, 5)$.

(e) Let Y_j be the value of the jth draw. By symmetry, $Y_j \sim \text{DUnif}(1, 2, \ldots, 100)$.

Learning any Y_i gives information about the other values (so $Y_1, \ldots, Y_5$ are *not* independent, as defined in Section 3.8), but symmetry still holds since, unconditionally, the jth slip drawn is equally likely to be any of the slips. This is the *unconditional* distribution of Y_j: we are working from a vantage point before drawing any of the slips.

For further insight into why each of $Y_1, \ldots, Y_5$ is Discrete Uniform and how to think about Y_j unconditionally, imagine that instead of one person drawing five slips, one at a time, there are five people who draw one slip each, all reaching into the hat *simultaneously*, with all possibilities equally likely for who gets which slip. This formulation does not change the problem in any important way, and it helps avoid getting distracted by irrelevant chronological details. Label the five people $1, 2, \ldots, 5$ in some way, e.g., from youngest to oldest, and let Z_j be the value drawn by person j. By symmetry, $Z_j \sim \text{DUnif}(1, 2, \ldots, 100)$ for each j; the Z_j's are dependent but, looked at individually, each person is drawing a uniformly random slip.

(f) The events $Y_1 = 100, \ldots, Y_5 = 100$ are disjoint since we are now sampling without replacement, so

$$P(Y_j = 100 \text{ for some } j) = P(Y_1 = 100) + \cdots + P(Y_5 = 100) = 0.05.$$

Sanity check: This answer makes sense intuitively since we can just as well think of first choosing five random slips out of 100 blank slips and then randomly writing the numbers from 1 to 100 on the slips, which gives a $5/100$ chance that the number 100 is on one of the five chosen slips.

It would be bizarre if the answer to (c) were greater than or equal to the answer to (f), since sampling without replacement makes it easier to find the number 100. (For the same reason, when searching for a lost possession it makes more sense to sample locations without replacement than with replacement.) But it makes sense that the answer to (c) is only slightly less than the answer to (f), since it is unlikely in (c) that the same slip will be sampled more than once (though by the birthday problem it's less unlikely than many people would guess).

More generally, if k slips are drawn without replacement, where $0 \leq k \leq 100$, then the same reasoning gives that the probability of drawing the number 100 is $k/100$. Note that this makes sense in the extreme case $k = 100$, since in that case we draw *all* of the slips. □

3.6 Cumulative distribution functions

Another function that describes the distribution of an r.v. is the *cumulative distribution function* (CDF). Unlike the PMF, which only discrete r.v.s possess, the CDF is defined for *all* r.v.s.

Definition 3.6.1. The *cumulative distribution function* (CDF) of an r.v. X is the function F_X given by $F_X(x) = P(X \leq x)$. When there is no risk of ambiguity, we sometimes drop the subscript and just write F (or some other letter) for a CDF.

The next example demonstrates that for discrete r.v.s, we can freely convert between CDF and PMF.

Example 3.6.2. Let $X \sim \text{Bin}(4, 1/2)$. Figure 3.8 shows the PMF and CDF of X.

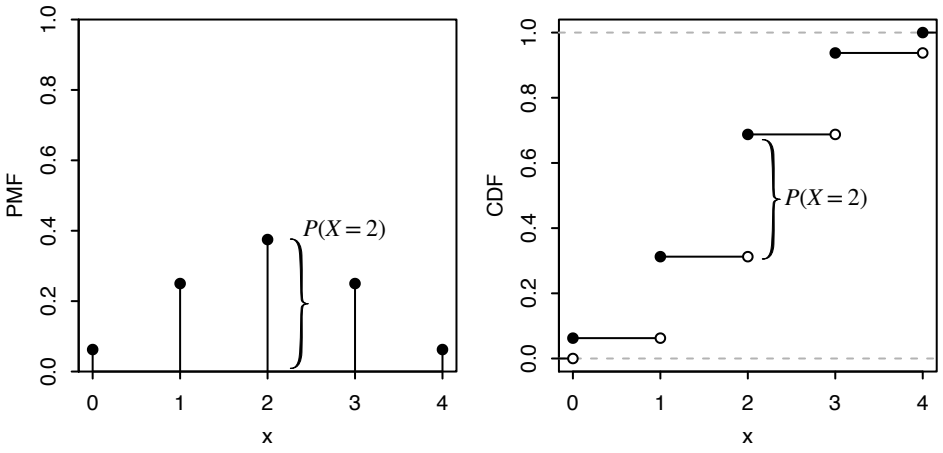


FIGURE 3.8
Bin$(4, 1/2)$ PMF and CDF. The height of the vertical bar $P(X = 2)$ in the PMF is also the height of the jump in the CDF at 2.

- *From PMF to CDF*: To find $P(X \leq 1.5)$, which is the CDF evaluated at 1.5, we sum the PMF over all values of the support that are less than or equal to 1.5:

$$P(X \leq 1.5) = P(X = 0) + P(X = 1) = \left(\frac{1}{2}\right)^4 + 4\left(\frac{1}{2}\right)^4 = \frac{5}{16}.$$

  Similarly, the value of the CDF at an arbitrary point x is the sum of the heights of the vertical bars of the PMF at values less than or equal to x.

- *From CDF to PMF*: The CDF of a discrete r.v. consists of jumps and flat regions. The height of a jump in the CDF at x is equal to the value of the PMF at x. For example, in Figure 3.8, the height of the jump in the CDF at 2 is the same as the height of the corresponding vertical bar in the PMF; this is indicated in the figure with curly braces. The flat regions of the CDF correspond to values outside the support of X, so the PMF is equal to 0 in those regions. □

Valid CDFs satisfy the following criteria.

Theorem 3.6.3 (Valid CDFs)**.** Any CDF F has the following properties.

- Increasing: If $x_1 \leq x_2$, then $F(x_1) \leq F(x_2)$.

- Right-continuous: As in Figure 3.8, the CDF is continuous except possibly for having some jumps. Wherever there is a jump, the CDF is continuous from the right. That is, for any a, we have

$$F(a) = \lim_{x \to a^+} F(x).$$

- Convergence to 0 and 1 in the limits:

$$\lim_{x \to -\infty} F(x) = 0 \text{ and } \lim_{x \to \infty} F(x) = 1.$$

Proof. The above criteria are true for *all* CDFs, but for simplicity we will only prove it for the case where F is the CDF of a discrete r.v. X whose possible values are $0, 1, 2, \ldots$. As an example of how to visualize the criteria, consider Figure 3.8: the CDF shown there is increasing (with some flat regions), continuous from the right (it is continuous except at jumps, and each jump has an open dot at the bottom and a closed dot at the top), and it converges to 0 as $x \to -\infty$ and to 1 as $x \to \infty$ (in this example, it reaches 0 and 1; in some examples, one or both of these values may be approached but never reached).

The first criterion is true since the event $\{X \leq x_1\}$ is a subset of the event $\{X \leq x_2\}$, so $P(X \leq x_1) \leq P(X \leq x_2)$.

For the second criterion, note that

$$P(X \leq x) = P(X \leq \lfloor x \rfloor),$$

where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. For example, $P(X \leq 4.9) = P(X \leq 4)$ since X is integer-valued. So $F(a+b) = F(a)$ for any $b > 0$ that is small enough so that $a + b < \lfloor a \rfloor + 1$, e.g., for $a = 4.9$, this holds for $0 < b < 0.1$. This implies $F(a) = \lim_{x \to a^+} F(x)$ (in fact, it's much stronger since it says $F(x)$ *equals* $F(a)$ when x is close enough to a and on the right).

For the third criterion, we have $F(x) = 0$ for $x < 0$, and

$$\lim_{x \to \infty} F(x) = \lim_{x \to \infty} P(X \leq \lfloor x \rfloor) = \lim_{x \to \infty} \sum_{n=0}^{\lfloor x \rfloor} P(X = n) = \sum_{n=0}^{\infty} P(X = n) = 1. \quad \blacksquare$$

The converse is true too: we will show in Chapter 5 that given any function F meeting these criteria, we can construct a random variable whose CDF is F.

To recap, we have now seen three equivalent ways of expressing the distribution of a random variable. Two of these are the PMF and the CDF: we know these two functions contain the same information, since we can always figure out the CDF from the PMF and vice versa. Generally the PMF is easier to work with for discrete r.v.s, since evaluating the CDF requires a summation.

A third way to describe a distribution is with a story that explains (in a precise way) how the distribution can arise. We used the stories of the Binomial and Hypergeometric distributions to derive the corresponding PMFs. Thus the story and the PMF also contain the same information, though we can often achieve more intuitive proofs with the story than with PMF calculations.

3.7 Functions of random variables

In this section we will discuss what it means to take a function of a random variable, and we will build understanding for why *a function of a random variable is a random variable*. That is, if X is a random variable, then X^2, e^X, and $\sin(X)$ are also random variables, as is $g(X)$ for any function $g : \mathbb{R} \to \mathbb{R}$.

For example, imagine that two basketball teams (A and B) are playing a seven-game match, and let X be the number of wins for team A (so $X \sim \text{Bin}(7, 1/2)$ if the teams are evenly matched and the games are independent). Let $g(x) = 7 - x$, and let $h(x) = 1$ if $x \geq 4$ and $h(x) = 0$ if $x < 4$. Then $g(X) = 7 - X$ is the number of wins for team B, and $h(X)$ is the indicator of team A winning the majority of the games. Since X is an r.v., both $g(X)$ and $h(X)$ are also r.v.s.

To see how to define functions of an r.v. formally, let's rewind a bit. At the beginning of this chapter, we considered a random variable X defined on a sample space with 6 elements. Figure 3.1 used arrows to illustrate how X maps each pebble in the sample space to a real number, and the left half of Figure 3.2 showed how we can equivalently imagine X writing a real number inside each pebble.

Now we can, if we want, apply the same function g to all the numbers inside the pebbles. Instead of the numbers $X(s_1)$ through $X(s_6)$, we now have the numbers $g(X(s_1))$ through $g(X(s_6))$, which gives a new mapping from sample outcomes to real numbers—we've created a new random variable, $g(X)$.

Definition 3.7.1 (Function of an r.v.)**.** For an experiment with sample space S, an r.v. X, and a function $g : \mathbb{R} \to \mathbb{R}$, $g(X)$ is the r.v. that maps s to $g(X(s))$ for all $s \in S$.

Taking $g(x) = \sqrt{x}$ for concreteness, Figure 3.9 shows that $g(X)$ is the *composition* of the functions X and g, saying "first apply X, then apply g". Figure 3.10 represents $g(X)$ more succinctly by directly labeling the sample outcomes. Both figures show us that $g(X)$ is an r.v.; if X crystallizes to 4, then $g(X)$ crystallizes to 2.

Given a discrete r.v. X with a known PMF, how can we find the PMF of $Y = g(X)$? In the case where g is a one-to-one function, the answer is straightforward: the support of Y is the set of all $g(x)$ with x in the support of X, and

$$P(Y = g(x)) = P(g(X) = g(x)) = P(X = x).$$

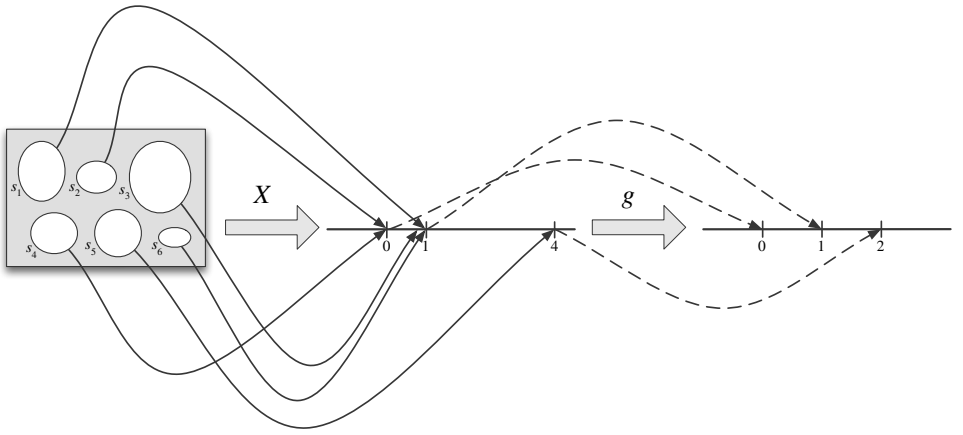


FIGURE 3.9
The r.v. X is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The function g is the square root function. Composing X and g gives the random variable $g(X) = \sqrt{X}$, which has possible values 0, 1, and 2.

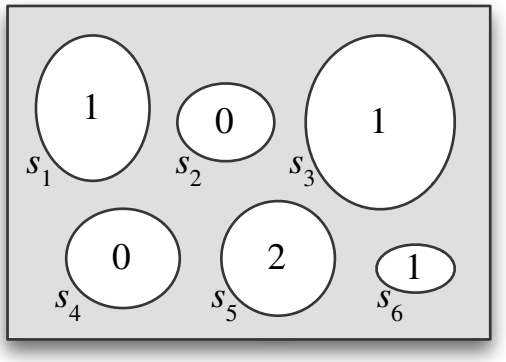


FIGURE 3.10
Since $g(X) = \sqrt{X}$ labels each pebble with a number, it is an r.v.

The case where $Y = g(X)$ with g one-to-one is illustrated in the following tables; the idea is that if the distinct possible values of X are $x_1, x_2, \ldots$ with probabilities $p_1, p_2, \ldots$ (respectively), then the distinct possible values of Y are $g(x_1), g(x_2), \ldots$, with the *same* list $p_1, p_2, \ldots$ of probabilities.

x	$P(X = x)$
x_1	p_1
x_2	p_2
x_3	p_3
$\vdots$	$\vdots$

PMF of X, in table form

y	$P(Y = y)$
$g(x_1)$	p_1
$g(x_2)$	p_2
$g(x_3)$	p_3
$\vdots$	$\vdots$

PMF of Y, in table form

This suggests a strategy for finding the PMF of an r.v. with an unfamiliar distribution: try to express the r.v. as a one-to-one function of an r.v. with a known distribution. The next example illustrates this method.

Example 3.7.2 (Random walk)**.** A particle moves n steps on a number line. The particle starts at 0, and at each step it moves 1 unit to the right or to the left, with equal probabilities. Assume all steps are independent. Let Y be the particle's position after n steps. Find the PMF of Y.

Solution:

Consider each step to be a Bernoulli trial, where right is considered a success and left is considered a failure. Then the number of steps the particle takes to the right is a $\text{Bin}(n, 1/2)$ random variable, which we can name X. If $X = j$, then the particle has taken j steps to the right and $n - j$ steps to the left, giving a final position of $j - (n - j) = 2j - n$. So we can express Y as a one-to-one function of X, namely, $Y = 2X - n$. Since X takes values in $\{0, 1, 2, \ldots, n\}$, Y takes values in $\{-n, 2 - n, 4 - n, \ldots, n\}$.

The PMF of Y can then be found from the PMF of X:

$$P(Y = k) = P(2X - n = k) = P(X = (n + k)/2) = \binom{n}{\frac{n+k}{2}} \left(\frac{1}{2}\right)^n,$$

if k is an integer between $-n$ and n (inclusive) such that $n+k$ is an even number. □

If g is not one-to-one, then for a given y, there may be multiple values of x such that $g(x) = y$. To compute $P(g(X) = y)$, we need to sum up the probabilities of X taking on any of these candidate values of x.

Theorem 3.7.3 (PMF of $g(X)$)**.** Let X be a discrete r.v. and $g : \mathbb{R} \to \mathbb{R}$. Then the support of $g(X)$ is the set of all y such that $g(x) = y$ for at least one x in the support of X, and the PMF of $g(X)$ is

$$P(g(X) = y) = \sum_{x:g(x)=y} P(X = x),$$

for all y in the support of $g(X)$.

Example 3.7.4. Continuing as in the previous example, let D be the particle's distance from the origin after n steps. Assume that n is even. Find the PMF of D.

Solution:

We can write $D = |Y|$; this is a function of Y, but it isn't one-to-one. The event $D = 0$ is the same as the event $Y = 0$. For $k = 2, 4, \ldots, n$, the event $D = k$ is the same as the event $\{Y = k\} \cup \{Y = -k\}$. So the PMF of D is

$$P(D = 0) = \binom{n}{\frac{n}{2}} \left(\frac{1}{2}\right)^n,$$

$$P(D = k) = P(Y = k) + P(Y = -k) = 2\binom{n}{\frac{n+k}{2}} \left(\frac{1}{2}\right)^n,$$

for $k = 2, 4, \ldots, n$. In the final step we used symmetry (imagine a new random walk that moves left each time our random walk moves right, and vice versa) to see that $P(Y = k) = P(Y = -k)$. □

The same reasoning we have used to handle functions of one random variable can be extended to deal with functions of multiple random variables. We have already seen an example of this with the addition function (which maps two numbers x, y to their sum $x + y$): in Example 3.2.5, we saw how to view $T = X + Y$ as an r.v. in its own right, where X and Y are obtained by rolling dice.

Definition 3.7.5 (Function of two r.v.s)**.** Given an experiment with sample space S, if X and Y are r.v.s that map $s \in S$ to $X(s)$ and $Y(s)$ respectively, then $g(X, Y)$ is the r.v. that maps s to $g(X(s), Y(s))$.

Note that we are assuming that X and Y are defined on the same sample space S. Usually we assume that S is chosen to be rich enough to encompass whatever r.v.s we wish to work with. For example, if X is based on a coin flip and Y is based on a die roll, and we initially were using the sample space $S_1 = \{H, T\}$ for X and the sample space $S_2 = \{1, 2, 3, 4, 5, 6\}$ for Y, we can easily redefine X and Y so that both are defined on the richer space $S = S_1 \times S_2 = \{(s_1, s_2) : s_1 \in S_1, s_2 \in S_2\}$.

One way to understand the mapping from S to $\mathbb{R}$ represented by the r.v. $g(X, Y)$ is with a table displaying the values of X, Y, and $g(X, Y)$ under various possible outcomes. Interpreting $X + Y$ as an r.v. is intuitive: if we observe $X = x$ and $Y = y$, then $X + Y$ crystallizes to $x + y$. For a less familiar example like $\max(X, Y)$, students often are unsure how to interpret it as an r.v. But the idea is the same: if we observe $X = x$ and $Y = y$, then $\max(X, Y)$ crystallizes to $\max(x, y)$.

Example 3.7.6 (Maximum of two die rolls)**.** We roll two fair 6-sided dice. Let X be the number on the first die and Y the number on the second die. The following table gives the values of X, Y, and $\max(X, Y)$ under 7 of the 36 outcomes in the sample space, analogously to the table in Example 3.2.5.

s	X	Y	$\max(X,Y)$
$(1,2)$	1	2	2
$(1,6)$	1	6	6
$(2,5)$	2	5	5
$(3,1)$	3	1	3
$(4,3)$	4	3	4
$(5,4)$	5	4	5
$(6,6)$	6	6	6

So $\max(X,Y)$ assigns a numerical value to each sample outcome. The PMF is

$$\begin{aligned}
P(\max(X,Y)=1) &= 1/36,\\
P(\max(X,Y)=2) &= 3/36,\\
P(\max(X,Y)=3) &= 5/36,\\
P(\max(X,Y)=4) &= 7/36,\\
P(\max(X,Y)=5) &= 9/36,\\
P(\max(X,Y)=6) &= 11/36.
\end{aligned}$$

These probabilities can be obtained by tabulating the values of $\max(x,y)$ in a 6×6 grid and counting how many times each value appears in the grid, or with calculations such as

$$\begin{aligned}
P(\max(X,Y)=5) &= P(X=5,Y\leq 4)+P(X\leq 4,Y=5)+P(X=5,Y=5)\\
&= 2P(X=5,Y\leq 4)+1/36\\
&= 2(4/36)+1/36 = 9/36.
\end{aligned}$$

□

☣ **3.7.7** (Category errors and sympathetic magic)**.** Many common mistakes in probability can be traced to confusing two of the following fundamental objects with each other: distributions, random variables, events, and numbers. Such mistakes are examples of *category errors*. In general, a category error is a mistake that doesn't just happen to be wrong, but in fact is necessarily wrong since it is based on the wrong category of object. For example, answering the question "How many people live in Boston?" with "-42" or "π" or "pink elephants" would be a category error—we may not know the population size of a city, but we do know that it is a nonnegative integer at any point in time. To help avoid being categorically wrong, always think about what category an answer should have.

An especially common category error is to confuse a random variable with its distribution. We call this error *sympathetic magic*; this term comes from anthropology, where it is used for the belief that one can influence an object by manipulating a representation of that object. The following saying sheds light on the distinction between a random variable and its distribution:

> The word is not the thing; the map is not the territory. – Alfred Korzybski

We can think of the distribution of a random variable as a map or *blueprint* describing the r.v. Just as different houses can share the same blueprint, different r.v.s can have the same distribution, even if the *experiments* they summarize, and the *sample spaces* they map from, are not the same.

Here are two examples of sympathetic magic:

- Given an r.v. X, trying to get the PMF of $2X$ by multiplying the PMF of X by 2. It does not make sense to multiply a PMF by 2, since the probabilities would no longer sum to 1. As we saw above, if X takes on values x_j with probabilities p_j, then $2X$ takes on values $2x_j$ with probabilities p_j. Therefore the PMF of $2X$ is a horizontal stretch of the PMF of X; it is *not* a vertical stretch, as would result from multiplying the PMF by 2. Figure 3.11 shows the PMF of a discrete r.v. X with support $\{0, 1, 2, 3, 4\}$, along with the PMF of $2X$, which has support $\{0, 2, 4, 6, 8\}$. Note that X can take on odd values, but $2X$ is necessarily even.

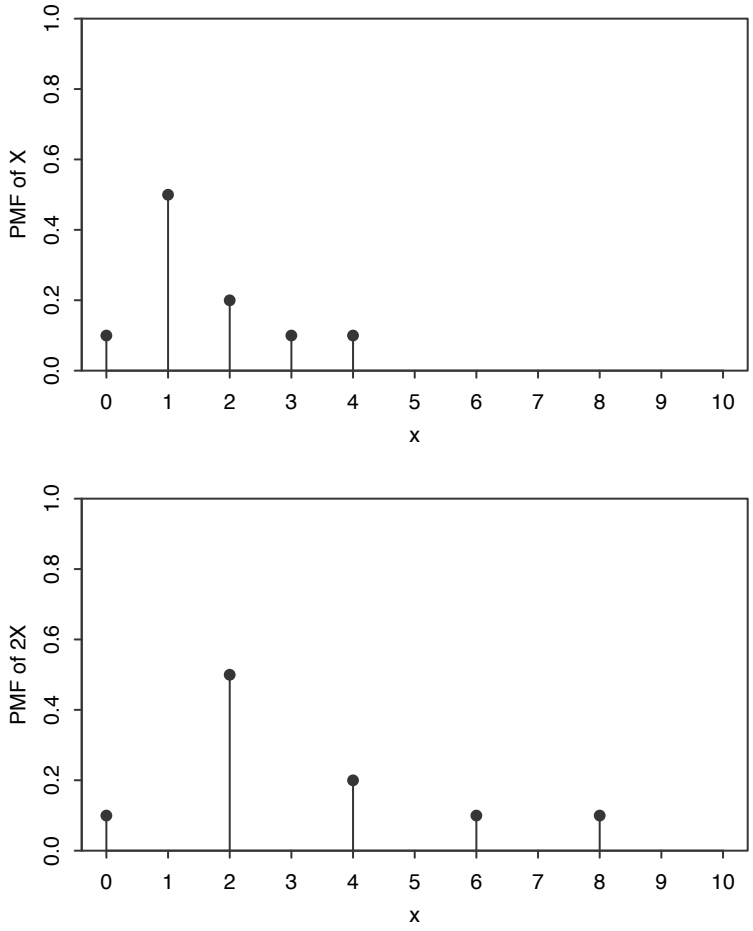


FIGURE 3.11
PMF of X (above) and PMF of $2X$ (below).

- Claiming that because X and Y have the same distribution, X must always equal Y, i.e., $P(X = Y) = 1$. Just because two r.v.s have the same distribution does not mean they are always equal, or *ever* equal. We saw this in Example 3.2.5. As another example, consider flipping a fair coin once. Let X be the indicator of Heads and $Y = 1 - X$ be the indicator of Tails. Both X and Y have the Bern(1/2) distribution, but the event $X = Y$ is impossible. The PMFs of X and Y are the same function, but X and Y are different mappings from the sample space to the real numbers.

  If Z is the indicator of Heads in a second flip (independent of the first flip), then Z is also Bern(1/2), but Z is not the same r.v. as X. Here

$$P(Z = X) = P(HH \text{ or } TT) = 1/2.$$

3.8 Independence of r.v.s

Just as we had the notion of independence of events, we can define independence of random variables. Intuitively, if two r.v.s X and Y are independent, then knowing the value of X gives no information about the value of Y, and vice versa. The definition formalizes this idea.

Definition 3.8.1 (Independence of two r.v.s)**.** Random variables X and Y are said to be *independent* if

$$P(X \leq x, Y \leq y) = P(X \leq x)P(Y \leq y),$$

for all $x, y \in \mathbb{R}$.

In the discrete case, this is equivalent to the condition

$$P(X = x, Y = y) = P(X = x)P(Y = y),$$

for all x, y with x in the support of X and y in the support of Y.

The definition for more than two r.v.s is analogous.

Definition 3.8.2 (Independence of many r.v.s)**.** Random variables $X_1, \ldots, X_n$ are *independent* if

$$P(X_1 \leq x_1, \ldots, X_n \leq x_n) = P(X_1 \leq x_1) \ldots P(X_n \leq x_n),$$

for all $x_1, \ldots, x_n \in \mathbb{R}$. For infinitely many r.v.s, we say that they are independent if every finite subset of the r.v.s is independent.

Comparing this to the criteria for independence of n events, it may seem strange that the independence of $X_1, \ldots, X_n$ requires just one equality, whereas for events we

needed to verify pairwise independence for all $\binom{n}{2}$ pairs, three-way independence for all $\binom{n}{3}$ triplets, and so on. However, upon closer examination of the definition, we see that independence of r.v.s requires the equality to hold for *all* possible $x_1, \ldots, x_n$—infinitely many conditions! If we can find even a single list of values $x_1, \ldots, x_n$ for which the above equality fails to hold, then $X_1, \ldots, X_n$ are not independent.

☣ **3.8.3.** If $X_1, \ldots, X_n$ are independent, then they are pairwise independent, i.e., X_i is independent of X_j for $i \neq j$. The idea behind proving that X_i and X_j are independent is to let all the x_k other than x_i, x_j go to ∞ in the definition of independence, since we already know $X_k < \infty$ is true (though it takes some work to give a complete justification for the limit). But pairwise independence does *not* imply independence in general, as we saw in Chapter 2 for events.

Example 3.8.4. In a roll of two fair dice, if X is the number on the first die and Y is the number on the second die, then $X + Y$ is not independent of $X - Y$ since

$$0 = P(X + Y = 12, X - Y = 1) \neq P(X + Y = 12)P(X - Y = 1) = \frac{1}{36} \cdot \frac{5}{36}.$$

Knowing the total is 12 tells us the difference must be 0, so the r.v.s provide information about each other. □

If X and Y are independent then it is also true, e.g., that X^2 is independent of Y^4, since if X^2 provided information about Y^4, then X would give information about Y (using X^2 and Y^4 as intermediaries: X determines X^2, which would give information about Y^4, which in turn would give information about Y). More generally, we have the following result (for which we omit a formal proof).

Theorem 3.8.5 (Functions of independent r.v.s)**.** If X and Y are independent r.v.s, then any function of X is independent of any function of Y.

Definition 3.8.6 (i.i.d.)**.** We will often work with random variables that are independent and have the same distribution. We call such r.v.s *independent and identically distributed*, or *i.i.d.* for short.

☣ **3.8.7** (i. vs. i.d.)**.** "Independent" and "identically distributed" are two often-confused but completely different concepts. Random variables are independent if they provide no information about each other; they are identically distributed if they have the same PMF (or equivalently, the same CDF). Whether two r.v.s are independent has nothing to do with whether they have the same distribution. We can have r.v.s that are:

- independent and identically distributed. Let X be the result of a die roll, and let Y be the result of a second, independent die roll. Then X and Y are i.i.d.
- independent and not identically distributed. Let X be the result of a die roll, and let Y be the closing price of the Dow Jones (a stock market index) a month from now. Then X and Y provide no information about each other (one would fervently hope), and X and Y do not have the same distribution.

- dependent and identically distributed. Let X be the number of Heads in n independent fair coin tosses, and let Y be the number of Tails in those same n tosses. Then X and Y are both distributed $\text{Bin}(n, 1/2)$, but they are highly dependent: if we know X, then we know Y perfectly.
- dependent and not identically distributed. Let X be the indicator of whether the majority party retains control of the House of Representatives in the U.S. after the next election, and let Y be the average favorability rating of the majority party in polls taken within a month of the election. Then X and Y are dependent, and X and Y do not have the same distribution.

By taking a sum of i.i.d. Bernoulli r.v.s, we can write down the story of the Binomial distribution in an algebraic form.

Theorem 3.8.8. If $X \sim \text{Bin}(n, p)$, viewed as the number of successes in n independent Bernoulli trials with success probability p, then we can write $X = X_1 + \cdots + X_n$ where the X_i are i.i.d. $\text{Bern}(p)$.

Proof. Let $X_i = 1$ if the ith trial was a success, and 0 if the ith trial was a failure. It's as though we have a person assigned to each trial, and we ask each person to raise their hand if their trial was a success. If we count the number of raised hands (which is the same as adding up the X_i), we get the total number of successes. ■

An important fact about the Binomial distribution is that the sum of independent Binomial r.v.s with the same success probability is also Binomial.

Theorem 3.8.9. If $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, and X is independent of Y, then $X + Y \sim \text{Bin}(n + m, p)$.

Proof. We present three proofs, since each illustrates a useful technique.

1. LOTP: We can directly find the PMF of $X + Y$ by conditioning on X (or Y, whichever we prefer) and using the law of total probability:

$$\begin{aligned} P(X+Y=k) &= \sum_{j=0}^{k} P(X+Y=k|X=j)P(X=j) \\ &= \sum_{j=0}^{k} P(Y=k-j)P(X=j) \\ &= \sum_{j=0}^{k} \binom{m}{k-j} p^{k-j} q^{m-k+j} \binom{n}{j} p^j q^{n-j} \\ &= p^k q^{n+m-k} \sum_{j=0}^{k} \binom{m}{k-j} \binom{n}{j} \\ &= \binom{n+m}{k} p^k q^{n+m-k}. \end{aligned}$$

In the second line, we used the independence of X and Y to justify dropping the conditioning in

$$P(X+Y=k|X=j) = P(Y=k-j|X=j) = P(Y=k-j),$$

and in the last line, we used the fact that

$$\sum_{j=0}^{k} \binom{m}{k-j}\binom{n}{j} = \binom{n+m}{k}$$

by Vandermonde's identity. The resulting expression is the $\text{Bin}(n+m,p)$ PMF, so $X+Y \sim \text{Bin}(n+m,p)$.

2. Representation: A much simpler proof is to represent both X and Y as the sum of i.i.d. $\text{Bern}(p)$ r.v.s: $X = X_1 + \cdots + X_n$ and $Y = Y_1 + \cdots + Y_m$, where the X_i and Y_j are all i.i.d. $\text{Bern}(p)$. Then $X+Y$ is the sum of $n+m$ i.i.d. $\text{Bern}(p)$ r.v.s, so its distribution, by the previous theorem, is $\text{Bin}(n+m,p)$.

3. Story: By the Binomial story, X is the number of successes in n independent trials and Y is the number of successes in m additional independent trials, all with the same success probability, so $X+Y$ is the total number of successes in the $n+m$ trials, which is the story of the $\text{Bin}(n+m,p)$ distribution. ■

Of course, if we have a definition for independence of r.v.s, we should have an analogous definition for conditional independence of r.v.s.

Definition 3.8.10 (Conditional independence of r.v.s)**.** Random variables X and Y are *conditionally independent* given an r.v. Z if for all $x, y \in \mathbb{R}$ and all z in the support of Z,

$$P(X \le x, Y \le y|Z=z) = P(X \le x|Z=z)P(Y \le y|Z=z).$$

For discrete r.v.s, an equivalent definition is to require

$$P(X=x, Y=y|Z=z) = P(X=x|Z=z)P(Y=y|Z=z).$$

As we might expect from the name, this is the definition of independence, except that we condition on $Z=z$ everywhere, and require the equality to hold for all z in the support of Z.

Definition 3.8.11 (Conditional PMF)**.** For any discrete r.v.s X and Z, the function $P(X=x|Z=z)$, when considered as a function of x for fixed z, is called the *conditional PMF of X given $Z=z$*.

Independence of r.v.s does not imply conditional independence, nor vice versa. First let us show why independence does not imply conditional independence.

Example 3.8.12 (Matching pennies)**.** Consider the simple game called *matching pennies*. Each of two players, A and B, has a fair penny. They flip their pennies independently. If the pennies match, A wins; otherwise, B wins. Let X be 1 if A's penny lands Heads and -1 otherwise, and define Y similarly for B (the r.v.s X and Y are called *random signs*).

Let $Z = XY$, which is 1 if A wins and -1 if B wins. Then X and Y are unconditionally independent, but given $Z = 1$, we know that $X = Y$ (the pennies match). So X and Y are conditionally dependent given Z. □

Example 3.8.13 (Two friends)**.** Consider again the "I have only two friends who ever call me" scenario from Example 2.5.11, except now with r.v. notation. Let X be the indicator of Alice calling me next Friday, Y be the indicator of Bob calling me next Friday, and Z be the indicator of exactly one of them calling me next Friday. Then X and Y are independent (by assumption). But given $Z = 1$, we have that X and Y are completely dependent: given that $Z = 1$, we have $Y = 1 - X$. □

Next let's see why conditional independence does not imply independence.

Example 3.8.14 (Mystery opponent)**.** Suppose that you are going to play two games of tennis against one of two identical twins. Against one of the twins, you are evenly matched, and against the other you have a 3/4 chance of winning. Suppose that you can't tell which twin you are playing against until after the two games. Let Z be the indicator of playing against the twin with whom you're evenly matched, and let X and Y be the indicators of victory in the first and second games, respectively.

Conditional on $Z = 1$, X and Y are i.i.d. Bern(1/2), and conditional on $Z = 0$, X and Y are i.i.d. Bern(3/4). So X and Y are conditionally independent given Z. Unconditionally, X and Y are dependent because observing $X = 1$ makes it more likely that we are playing the twin who is worse. That is,

$$P(Y = 1|X = 1) > P(Y = 1).$$

Past games give us information which helps us infer who our opponent is, which in turn helps us predict future games! Note that this example is isomorphic to the "random coin" scenario from Example 2.3.7. □

3.9 Connections between Binomial and Hypergeometric

The Binomial and Hypergeometric distributions are connected in two important ways. As we will see in this section, we can get from the Binomial to the Hypergeometric by *conditioning*, and we can get from the Hypergeometric to the Binomial by *taking a limit*. We'll start with a motivating example.

Example 3.9.1 (Fisher exact test)**.** A scientist wishes to study whether women or

men are more likely to have a certain disease, or whether they are equally likely. A random sample of n women and m men is gathered, and each person is tested for the disease (assume for this problem that the test is completely accurate). The numbers of women and men in the sample who have the disease are X and Y respectively, with $X \sim \text{Bin}(n, p_1)$ and $Y \sim \text{Bin}(m, p_2)$, independently. Here p_1 and p_2 are unknown, and we are interested in testing whether $p_1 = p_2$ (this is known as a *null hypothesis* in statistics).

Consider a 2×2 table with rows corresponding to disease status and columns corresponding to gender. Each entry is the count of how many people have that disease status and gender, so $n + m$ is the sum of all 4 entries. Suppose that it is observed that $X + Y = r$.

The *Fisher exact test* is based on conditioning on both the row and column sums, so n, m, r are all treated as fixed, and then seeing if the observed value of X is "extreme" compared to this conditional distribution. Assuming the null hypothesis, find the conditional PMF of X given $X + Y = r$.

Solution:

First we'll build the 2×2 table, treating n, m, and r as fixed.

	Women	**Men**	**Total**
Disease	x	$r - x$	r
No disease	$n - x$	$m - r + x$	$n + m - r$
Total	n	m	$n + m$

Next, let's compute the conditional PMF $P(X = x|X + Y = r)$. By Bayes' rule,

$$\begin{aligned} P(X = x|X + Y = r) &= \frac{P(X + Y = r|X = x)P(X = x)}{P(X + Y = r)} \\ &= \frac{P(Y = r - x)P(X = x)}{P(X + Y = r)}. \end{aligned}$$

The step $P(X+Y = r|X = x) = P(Y = r-x)$ is justified by the independence of X and Y. Assuming the null hypothesis and letting $p = p_1 = p_2$, we have $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, independently, so $X + Y \sim \text{Bin}(n + m, p)$. Thus,

$$\begin{aligned} P(X = x|X + Y = r) &= \frac{\binom{m}{r-x} p^{r-x}(1-p)^{m-r+x} \binom{n}{x} p^x (1-p)^{n-x}}{\binom{n+m}{r} p^r (1-p)^{n+m-r}} \\ &= \frac{\binom{n}{x}\binom{m}{r-x}}{\binom{n+m}{r}}. \end{aligned}$$

So the conditional distribution of X is Hypergeometric with parameters n, m, r.

To understand why the Hypergeometric appeared, seemingly out of nowhere, let's connect this problem to the elk story for the Hypergeometric. In the elk story, we are

interested in the distribution of the number of tagged elk in the recaptured sample. By analogy, think of women as tagged elk and men as untagged elk. Instead of recapturing r elk at random from the forest, we infect $X + Y = r$ people with the disease; under the null hypothesis, the set of diseased people is equally likely to be any set of r people. Thus, conditional on $X + Y = r$, X represents the number of women among the r diseased individuals. This is exactly analogous to the number of tagged elk in the recaptured sample, which is distributed $\text{HGeom}(n, m, r)$.

An interesting fact, which turns out to be useful in statistics, is that the conditional distribution of X does not depend on p: unconditionally, $X \sim \text{Bin}(n, p)$, but p disappears from the parameters of the conditional distribution! This makes sense upon reflection, since once we know $X + Y = r$, we can work directly with the fact that we have a population with r diseased and $n + m - r$ healthy people, without worrying about the value of p that originally generated the population. □

This motivating example serves as a proof of the following theorem.

Theorem 3.9.2. If $X \sim \text{Bin}(n, p)$, $Y \sim \text{Bin}(m, p)$, and X is independent of Y, then the conditional distribution of X given $X + Y = r$ is $\text{HGeom}(n, m, r)$.

In the other direction, the Binomial is a limiting case of the Hypergeometric.

Theorem 3.9.3. If $X \sim \text{HGeom}(w, b, n)$ and $N = w + b \to \infty$ such that $p = w/(w + b)$ remains fixed, then the PMF of X converges to the $\text{Bin}(n, p)$ PMF.

Proof. We take the stated limit of the $\text{HGeom}(w, b, n)$ PMF:

$$\begin{aligned}
P(X = k) &= \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}} \\
&= \binom{n}{k}\frac{\binom{w+b-n}{w-k}}{\binom{w+b}{w}} \quad \text{by Theorem 3.4.5} \\
&= \binom{n}{k}\frac{w!}{(w-k)!}\frac{b!}{(b-n+k)!}\frac{(w+b-n)!}{(w+b)!} \\
&= \binom{n}{k}\frac{w(w-1)\dots(w-k+1)b(b-1)\dots(b-n+k+1)}{(w+b)(w+b-1)\dots(w+b-n+1)} \\
&= \binom{n}{k}\frac{p\left(p-\frac{1}{N}\right)\dots\left(p-\frac{k-1}{N}\right)q\left(q-\frac{1}{N}\right)\dots\left(q-\frac{n-k-1}{N}\right)}{\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\dots\left(1-\frac{n-1}{N}\right)}.
\end{aligned}$$

As $N \to \infty$, the denominator goes to 1, and the numerator goes to $p^k q^{n-k}$. Thus

$$P(X = k) \to \binom{n}{k} p^k q^{n-k},$$

which is the $\text{Bin}(n, p)$ PMF. ■

The stories of the Binomial and Hypergeometric provide intuition for this result: given an urn with w white balls and b black balls, the Binomial distribution arises

from sampling n balls from the urn with replacement, while the Hypergeometric arises from sampling without replacement. As the number of balls in the urn grows very large relative to the number of balls that are drawn, sampling with replacement and sampling without replacement become essentially equivalent. In practical terms, this theorem tells us that if $N = w + b$ is large relative to n, we can approximate the HGeom(w, b, n) PMF by the Bin$(n, w/(w + b))$ PMF.

The birthday problem implies that it is surprisingly likely that some ball will be sampled more than once if sampling with replacement; for example, if 1,200 out of 1,000,000 balls are drawn randomly with replacement, then there is about a 51% chance that some ball will be drawn more than once! But this becomes less and less likely as N grows, and even if it is likely that there will be a few coincidences, the approximation can still be reasonable if it is very likely that the vast majority of balls in the sample are sampled only once each.

3.10 Recap

A random variable (r.v.) is a function assigning a real number to every possible outcome of an experiment. The distribution of an r.v. X is a full specification of the probabilities for the events associated with X, such as $\{X = 3\}$ and $\{1 \leq X \leq 5\}$. The distribution of a discrete r.v. can be defined using a PMF, a CDF, or a story. The PMF of X is the function $P(X = x)$ for $x \in \mathbb{R}$. The CDF of X is the function $P(X \leq x)$ for $x \in \mathbb{R}$. A story for X describes an experiment that could give rise to a random variable with the same distribution as X.

For a PMF to be valid, it must be nonnegative and sum to 1. For a CDF to be valid, it must be increasing, right-continuous, converge to 0 as $x \to -\infty$, and converge to 1 as $x \to \infty$.

It is important to distinguish between a random variable and its distribution: the distribution is a blueprint for building the r.v., but different r.v.s can have the same distribution, just as different houses can be built from the same blueprint.

Four named discrete distributions are the Bernoulli, Binomial, Hypergeometric, and Discrete Uniform. Each of these is actually a *family* of distributions, indexed by parameters; to fully specify one of these distributions, we need to give both the name and the parameter values.

- A Bern(p) r.v. is the indicator of success in a Bernoulli trial with probability of success p.
- A Bin(n, p) r.v. is the number of successes in n independent Bernoulli trials, all with the same probability p of success.

- A HGeom(w, b, n) r.v. is the number of white balls obtained in a sample of size n drawn without replacement from an urn of w white and b black balls.
- A DUnif(C) r.v. is obtained by randomly choosing an element of the finite set C, with equal probabilities for each element.

A function of a random variable is still a random variable. If we know the PMF of X, we can find $P(g(X) = k)$, the PMF of $g(X)$, by translating the event $\{g(X) = k\}$ into an equivalent event involving X, then using the PMF of X.

Two random variables are independent if knowing the value of one r.v. gives no information about the value of the other. This is unrelated to whether the two r.v.s are identically distributed. In Chapter 7, we will learn how to deal with dependent random variables by considering them jointly rather than separately.

We have now seen four fundamental types of objects in probability: distributions, random variables, events, and numbers. Figure 3.12 shows connections between these four fundamental objects. A CDF can be used as a blueprint for generating an r.v., and then there are various events describing the behavior of the r.v., such as the events $X \leq x$ for all x. Knowing the probabilities of these events determines the CDF, taking us full circle. For a discrete r.v. we can also use the PMF as a blueprint, and go from distribution to r.v. to events and back again.

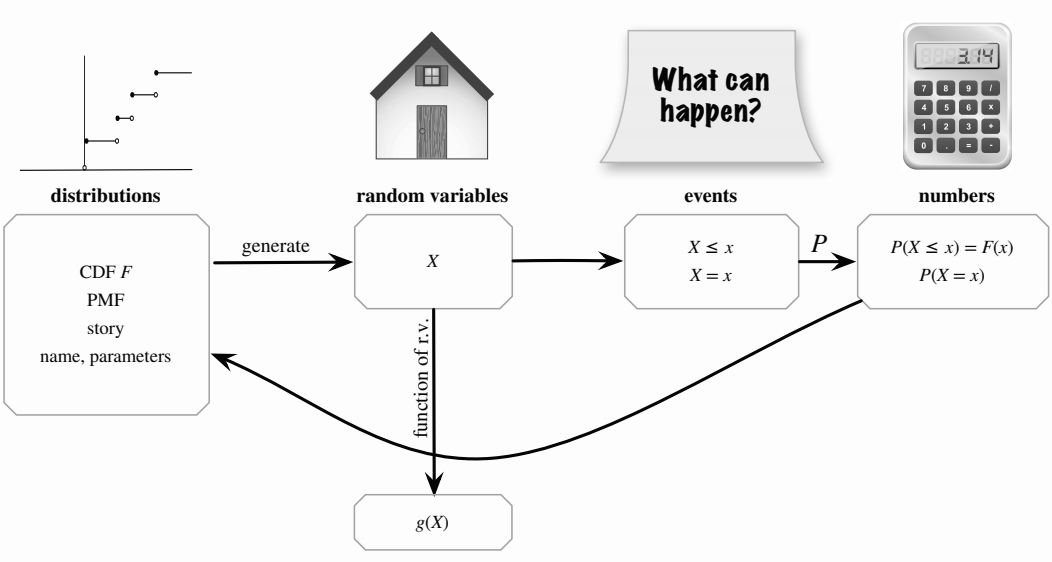


FIGURE 3.12
Four fundamental objects in probability: distributions (blueprints), random variables, events, and numbers. From a CDF F we can generate an r.v. X. From X, we can generate many other r.v.s by taking functions of X. There are various events describing the behavior of X. Most notably, for any constant x the events $X \leq x$ and $X = x$ are of interest. Knowing the probabilities of these events for all x gives us the CDF and (in the discrete case) the PMF, taking us full circle.

3.11 R

Distributions in R

All of the named distributions that we'll encounter in this book have been implemented in R. In this section we'll explain how to work with the Binomial and Hypergeometric distributions in R. We will also explain in general how to generate r.v.s from any discrete distribution with a finite support. Typing `help(distributions)` gives a handy list of built-in distributions; many others are available through R packages that can be loaded.

In general, for many named discrete distributions, three functions starting with `d`, `p`, and `r` will give the PMF, CDF, and random generation, respectively. Note that the function starting with `p` is not the PMF, but rather is the CDF.

Binomial distribution

The Binomial distribution is associated with the following three R functions: `dbinom`, `pbinom`, and `rbinom`. For the Bernoulli distribution we can just use the Binomial functions with $n = 1$.

- `dbinom` is the Binomial PMF. It takes three inputs: the first is the value of x at which to evaluate the PMF, and the second and third are the parameters n and p. For example, `dbinom(3,5,0.2)` returns the probability $P(X = 3)$ where $X \sim \text{Bin}(5, 0.2)$. In other words,

$$\texttt{dbinom(3,5,0.2)} = \binom{5}{3}(0.2)^3(0.8)^2 = 0.0512.$$

- `pbinom` is the Binomial CDF. It takes three inputs: the first is the value of x at which to evaluate the CDF, and the second and third are the parameters. `pbinom(3,5,0.2)` is the probability $P(X \leq 3)$ where $X \sim \text{Bin}(5, 0.2)$. So

$$\texttt{pbinom(3,5,0.2)} = \sum_{k=0}^{3} \binom{5}{k}(0.2)^k(0.8)^{5-k} = 0.9933.$$

- `rbinom` is a function for generating Binomial random variables. For `rbinom`, the first input is *how many* r.v.s we want to generate, and the second and third inputs are still the parameters. Thus the command `rbinom(7,5,0.2)` produces realizations of seven i.i.d. $\text{Bin}(5, 0.2)$ r.v.s. When we ran this command, we got

```
2 1 0 0 1 0 0
```

  but you'll probably get something different when you try it!

We can also evaluate PMFs and CDFs at an entire vector of values. For example, recall that `0:n` is a quick way to list the integers from 0 to n. The command `dbinom(0:5,5,0.2)` returns 6 numbers, $P(X = 0), P(X = 1), \ldots, P(X = 5)$, where $X \sim \text{Bin}(5, 0.2)$.

Hypergeometric distribution

The Hypergeometric distribution also has three functions: `dhyper`, `phyper`, and `rhyper`. As one might expect, `dhyper` is the Hypergeometric PMF, `phyper` is the Hypergeometric CDF, and `rhyper` generates Hypergeometric r.v.s. Since the Hypergeometric distribution has three parameters, each of these functions takes *four* inputs. For `dhyper` and `phyper`, the first input is the value at which we wish to evaluate the PMF or CDF, and the remaining inputs are the parameters of the distribution.

Thus `dhyper(k,w,b,n)` returns $P(X = k)$ where $X \sim \text{HGeom}(w, b, n)$, and `phyper(k,w,b,n)` returns $P(X \leq k)$. For `rhyper`, the first input is the number of r.v.s we want to generate, and the remaining inputs are the parameters; `rhyper(100,w,b,n)` generates 100 i.i.d. $\text{HGeom}(w, b, n)$ r.v.s.

Discrete distributions with finite support

We can generate r.v.s from *any* discrete distribution with finite support using the `sample` command. When we first introduced the `sample` command, we said that it can be used in the form `sample(n,k)` or `sample(n,k,replace=TRUE)` to sample k times from the integers 1 through n, either without or with replacement. For example, to generate 5 independent $\text{DUnif}(1, 2, \ldots, 100)$ r.v.s, we can use the command `sample(100,5,replace=TRUE)`.

It turns out that `sample` is far more versatile. If we want to sample from the values $x_1, \ldots, x_n$ with probabilities $p_1, \ldots, p_n$, we simply create a vector `x` containing all the x_i and a vector `p` containing all the p_i, then feed them into `sample`. Suppose we want realizations of i.i.d. r.v.s $X_1, \ldots, X_{100}$ whose PMF is

$$\begin{aligned} P(X_j = 0) &= 0.25, \\ P(X_j = 1) &= 0.5, \\ P(X_j = 5) &= 0.1, \\ P(X_j = 10) &= 0.15, \end{aligned}$$

and $P(X_j = x) = 0$ for all other values of x. First, we use the `c` function to create vectors with the support of the distribution and the corresponding probabilities.

```
x <- c(0,1,5,10)
p <- c(0.25,0.5,0.1,0.15)
```

Next, we use `sample`. Here's how to get 100 draws from the PMF above:

```
sample(x,100,prob=p,replace=TRUE)
```

The inputs are the vector `x` to sample from, the sample size (100 in this case), the probabilities `p` to use when sampling from `x` (if this is omitted, the probabilities are assumed equal), and whether to sample with replacement.

3.12 Exercises

Exercises marked with Ⓢ have detailed solutions at `http://stat110.net`.

PMFs and CDFs

1. People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, i.e., before person X arrives no two people have the same birthday, but when person X arrives there is a match. Find the PMF of X.

2. (a) Independent Bernoulli trials are performed, with probability 1/2 of success, until there has been at least one success. Find the PMF of the number of trials performed.

   (b) Independent Bernoulli trials are performed, with probability 1/2 of success, until there has been at least one success and at least one failure. Find the PMF of the number of trials performed.

3. Let X be an r.v. with CDF F, and $Y = \mu + \sigma X$, where μ and σ are real numbers with $\sigma > 0$. (Then Y is called a *location-scale transformation* of X; we will encounter this concept many times in Chapter 5 and beyond.) Find the CDF of Y, in terms of F.

4. Let n be a positive integer and
$$F(x) = \frac{\lfloor x \rfloor}{n}$$
for $0 \leq x \leq n$, $F(x) = 0$ for $x < 0$, and $F(x) = 1$ for $x > n$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x. Show that F is a CDF, and find the PMF that it corresponds to.

5. (a) Show that $p(n) = \left(\frac{1}{2}\right)^{n+1}$ for $n = 0, 1, 2, \ldots$ is a valid PMF for a discrete r.v.

   (b) Find the CDF of a random variable with the PMF from (a).

6. Ⓢ *Benford's law* states that in a very large variety of real-life data sets, the first digit approximately follows a particular distribution with about a 30% chance of a 1, an 18% chance of a 2, and in general
$$P(D = j) = \log_{10}\left(\frac{j+1}{j}\right), \text{ for } j \in \{1, 2, 3, \ldots, 9\},$$
where D is the first digit of a randomly chosen element. Check that this is a valid PMF (using properties of logs, not with a calculator).

7. Bob is playing a video game that has 7 levels. He starts at level 1, and has probability p_1 of reaching level 2. In general, given that he reaches level j, he has probability p_j of reaching level $j+1$, for $1 \leq j \leq 6$. Let X be the highest level that he reaches. Find the PMF of X (in terms of $p_1, \ldots, p_6$).

8. There are 100 prizes, with one worth \$1, one worth \$2, ..., and one worth \$100. There are 100 boxes, each of which contains one of the prizes. You get 5 prizes by picking random boxes one at a time, *without replacement*. Find the PMF of how much your most valuable prize is worth (as a simple expression in terms of binomial coefficients).

9. Let F_1 and F_2 be CDFs, $0 < p < 1$, and $F(x) = pF_1(x) + (1-p)F_2(x)$ for all x.

   (a) Show directly that F has the properties of a valid CDF (see Theorem 3.6.3). The distribution defined by F is called a *mixture* of the distributions defined by F_1 and F_2.

   (b) Consider creating an r.v. in the following way. Flip a coin with probability p of Heads. If the coin lands Heads, generate an r.v. according to F_1; if the coin lands Tails, generate an r.v. according to F_2. Show that the r.v. obtained in this way has CDF F.

10. (a) Is there a discrete distribution with support $1, 2, 3, \ldots$, such that the value of the PMF at n is proportional to $1/n$?

    Hint: See the math appendix for a review of some facts about series.

    (b) Is there a discrete distribution with support $1, 2, 3, \ldots$, such that the value of the PMF at n is proportional to $1/n^2$?

11. Ⓢ Let X be an r.v. whose possible values are $0, 1, 2, \ldots$, with CDF F. In some countries, rather than using a CDF, the convention is to use the function G defined by $G(x) = P(X < x)$ to specify a distribution. Find a way to convert from F to G, i.e., if F is a known function, show how to obtain $G(x)$ for all real x.

12. (a) Give an example of r.v.s X and Y such that $F_X(x) \leq F_Y(x)$ for all x, where the inequality is strict for some x. Here F_X is the CDF of X and F_Y is the CDF of Y. For the example you gave, sketch the CDFs of both X and Y on the same axes. Then sketch their PMFs on a second set of axes.

    (b) In Part (a), you found an example of two different CDFs where the first is less than or equal to the second everywhere. Is it possible to find two different PMFs where the first is less than or equal to the second everywhere? In other words, find discrete r.v.s X and Y such that $P(X = x) \leq P(Y = x)$ for all x, where the inequality is strict for some x, or show that it is impossible to find such r.v.s.

13. Let X, Y, Z be discrete r.v.s such that X and Y have the same conditional distribution given Z, i.e., for all a and z we have

$$P(X = a|Z = z) = P(Y = a|Z = z).$$

    Show that X and Y have the same distribution (unconditionally, not just when given Z).

14. Let X be the number of purchases that Fred will make on the online site for a certain company (in some specified time period). Suppose that the PMF of X is $P(X = k) = e^{-\lambda}\lambda^k/k!$ for $k = 0, 1, 2, \ldots$. This distribution is called the *Poisson distribution* with parameter λ, and it will be studied extensively in later chapters.

    (a) Find $P(X \geq 1)$ and $P(X \geq 2)$ without summing infinite series.

    (b) Suppose that the company only knows about people who have made at least one purchase on their site (a user sets up an account to make a purchase, but someone who has never made a purchase there doesn't appear in the customer database). If the company computes the number of purchases for everyone in their database, then these data are draws from the *conditional* distribution of the number of purchases, given that at least one purchase is made. Find the conditional PMF of X given $X \geq 1$. (This conditional distribution is called a *truncated Poisson distribution*.)

Named distributions

15. Find the CDF of an r.v. $X \sim \text{DUnif}(1, 2, \ldots, n)$.

16. Let $X \sim \text{DUnif}(C)$, and B be a nonempty subset of C. Find the conditional distribution of X, given that X is in B.

17. An airline overbooks a flight, selling more tickets for the flight than there are seats on the plane (figuring that it's likely that some people won't show up). The plane has 100 seats, and 110 people have booked the flight. Each person will show up for the flight with probability 0.9, independently. Find the probability that there will be enough seats for everyone who shows up for the flight.

18. Ⓢ (a) In the World Series of baseball, two teams (call them A and B) play a sequence of games against each other, and the first team to win four games wins the series. Let p be the probability that A wins an individual game, and assume that the games are independent. What is the probability that team A wins the series?

    (b) Give a clear intuitive explanation of whether the answer to (a) depends on whether the teams always play 7 games (and whoever wins the majority wins the series), or the teams stop playing more games as soon as one team has won 4 games (as is actually the case in practice: once the match is decided, the two teams do not keep playing more games).

19. In a chess tournament, n games are being played, independently. Each game ends in a win for one player with probability 0.4 and ends in a draw (tie) with probability 0.6. Find the PMFs of the number of games ending in a draw, and of the number of players whose games end in draws.

20. Suppose that a lottery ticket has probability p of being a winning ticket, independently of other tickets. A gambler buys 3 tickets, hoping this will triple the chance of having at least one winning ticket.

    (a) What is the distribution of how many of the 3 tickets are winning tickets?

    (b) Show that the probability that at least 1 of the 3 tickets is winning is $3p - 3p^2 + p^3$, in two different ways: by using inclusion-exclusion, and by taking the complement of the desired event and then using the PMF of a certain named distribution.

    (c) Show that the gambler's chances of having at least one winning ticket do not quite triple (compared with buying only one ticket), but that they do *approximately* triple if p is small.

21. Ⓢ Let $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$, independent of X. Show that $X - Y$ is *not* Binomial.

22. There are two coins, one with probability p_1 of Heads and the other with probability p_2 of Heads. One of the coins is randomly chosen (with equal probabilities for the two coins). It is then flipped $n \geq 2$ times. Let X be the number of times it lands Heads.

    (a) Find the PMF of X.

    (b) What is the distribution of X if $p_1 = p_2$?

    (c) Give an intuitive explanation of why X is *not* Binomial for $p_1 \neq p_2$ (its distribution is called a *mixture* of two Binomials). You can assume that n is large for your explanation, so that the frequentist interpretation of probability can be applied.

23. There are n people eligible to vote in a certain election. Voting requires registration. Decisions are made independently. Each of the n people will register with probability p_1. Given that a person registers, they will vote with probability p_2. Given that a person votes, they will vote for Kodos (who is one of the candidates) with probability p_3. What is the distribution of the number of votes for Kodos (give the PMF, fully simplified, or the name of the distribution, including its parameters)?

24. Let X be the number of Heads in 10 fair coin tosses.

    (a) Find the conditional PMF of X, given that the first two tosses both land Heads.

    (b) Find the conditional PMF of X, given that at least two tosses land Heads.

25. Ⓢ Alice flips a fair coin n times and Bob flips another fair coin $n+1$ times, resulting in independent $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n+1, \frac{1}{2})$.

    (a) Show that $P(X < Y) = P(n - X < n + 1 - Y)$.

    (b) Compute $P(X < Y)$.

    Hint: Use (a) and the fact that X and Y are integer-valued.

26. If $X \sim \text{HGeom}(w, b, n)$, what is the distribution of $n - X$? Give a short proof.

27. Recall de Montmort's matching problem from Chapter 1: in a deck of n cards labeled 1 through n, a match occurs when the number on the card matches the card's position in the deck. Let X be the number of matching cards. Is X Binomial? Is X Hypergeometric?

28. Ⓢ There are n eggs, each of which hatches a chick with probability p (independently). Each of these chicks survives with probability r, independently. What is the distribution of the number of chicks that hatch? What is the distribution of the number of chicks that survive? (Give the PMFs; also give the names of the distributions and their parameters, if applicable.)

29. Ⓢ A sequence of n independent experiments is performed. Each experiment is a success with probability p and a failure with probability $q = 1 - p$. Show that conditional on the number of successes, all valid possibilities for the list of outcomes of the experiment are equally likely.

30. A certain company has $n + m$ employees, consisting of n women and m men. The company is deciding which employees to promote.

    (a) Suppose for this part that the company decides to promote t employees, where $1 \leq t \leq n + m$, by choosing t random employees (with equal probabilities for each set of t employees). What is the distribution of the number of women who get promoted?

    (b) Now suppose that instead of having a predetermined number of promotions to give, the company decides independently for each employee, promoting the employee with probability p. Find the distributions of the number of women who are promoted, the number of women who are not promoted, and the number of employees who are promoted.

    (c) In the set-up from (b), find the conditional distribution of the number of women who are promoted, given that exactly t employees are promoted.

31. Once upon a time, a famous statistician offered tea to a lady. The lady claimed that she could tell whether milk had been added to the cup before or after the tea. The statistician decided to run some experiments to test her claim.

    (a) The lady is given 6 cups of tea, where it is known in advance that 3 will be milk-first and 3 will be tea-first, in a completely random order. The lady gets to taste each and then guess which 3 were milk-first. Assume for this part that she has no ability whatsoever to distinguish milk-first from tea-first cups of tea. Find the probability that at least 2 of her 3 guesses are correct.

    (b) Now the lady is given one cup of tea, with probability 1/2 of it being milk-first. She needs to say whether she thinks it was milk-first. Let p_1 be the lady's probability of being correct given that it was milk-first, and p_2 be her probability of being correct given that it was tea-first. She claims that the cup was milk-first. Find the *posterior odds* that the cup is milk-first, given this information.

32. In Evan's history class, 10 out of 100 key terms will be randomly selected to appear on the final exam; Evan must then choose 7 of those 10 to define. Since he knows the format of the exam in advance, Evan is trying to decide how many key terms he should study.

    (a) Suppose that Evan decides to study s key terms, where s is an integer between 0 and 100. Let X be the number of key terms appearing on the exam that he has studied. What is the distribution of X? Give the name and parameters, in terms of s.

    (b) Using R or other software, calculate the probability that Evan knows at least 7 of the 10 key terms that appear on the exam, assuming that he studies $s = 75$ key terms.

33. A book has n typos. Two proofreaders, Prue and Frida, independently read the book. Prue catches each typo with probability p_1 and misses it with probability $q_1 = 1 - p_1$, independently, and likewise for Frida, who has probabilities p_2 of catching and $q_2 = 1-p_2$ of missing each typo. Let X_1 be the number of typos caught by Prue, X_2 be the number caught by Frida, and X be the number caught by at least one of the two proofreaders.

    (a) Find the distribution of X.

    (b) For this part only, assume that $p_1 = p_2$. Find the conditional distribution of X_1 given that $X_1 + X_2 = t$.

34. There are n students at a certain school, of whom $X \sim \text{Bin}(n, p)$ are Statistics majors. A simple random sample of size m is drawn ("simple random sample" means sampling without replacement, with all subsets of the given size equally likely).

    (a) Find the PMF of the number of Statistics majors in the sample, using the law of total probability (don't forget to say what the support is). You can leave your answer as a sum (though with some algebra it can be simplified, by writing the binomial coefficients in terms of factorials and using the binomial theorem).

    (b) Give a story proof derivation of the distribution of the number of Statistics majors in the sample; simplify fully.

    Hint: Does it matter whether the students declare their majors before or after the random sample is drawn?

35. Ⓢ Players A and B take turns in answering trivia questions, starting with player A answering the first question. Each time A answers a question, she has probability p_1 of getting it right. Each time B plays, he has probability p_2 of getting it right.

    (a) If A answers m questions, what is the PMF of the number of questions she gets right?

    (b) If A answers m times and B answers n times, what is the PMF of the total number of questions they get right (you can leave your answer as a sum)? Describe exactly when/whether this is a Binomial distribution.

    (c) Suppose that the first player to answer correctly wins the game (with no predetermined maximum number of questions that can be asked). Find the probability that A wins the game.

36. There are n voters in an upcoming election in a certain country, where n is a large, even number. There are two candidates: Candidate A (from the Unite Party) and Candidate B (from the Untie Party). Let X be the number of people who vote for Candidate A. Suppose that each voter chooses randomly whom to vote for, independently and with equal probabilities.

    (a) Find an exact expression for the probability of a tie in the election (so the candidates end up with the same number of votes).

(b) Use Stirling's approximation, which approximates the factorial function as

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

to find a simple approximation to the probability of a tie. Your answer should be of the form $1/\sqrt{cn}$, with c a constant (which you should specify).

37. Ⓢ A message is sent over a noisy channel. The message is a sequence $x_1, x_2, \ldots, x_n$ of n bits ($x_i \in \{0,1\}$). Since the channel is noisy, there is a chance that any bit might be corrupted, resulting in an error (a 0 becomes a 1 or vice versa). Assume that the error events are independent. Let p be the probability that an individual bit has an error ($0 < p < 1/2$). Let $y_1, y_2, \ldots, y_n$ be the received message (so $y_i = x_i$ if there is no error in that bit, but $y_i = 1 - x_i$ if there is an error there).

To help detect errors, the nth bit is reserved for a parity check: x_n is defined to be 0 if $x_1 + x_2 + \cdots + x_{n-1}$ is even, and 1 if $x_1 + x_2 + \cdots + x_{n-1}$ is odd. When the message is received, the recipient checks whether y_n has the same parity as $y_1 + y_2 + \cdots + y_{n-1}$. If the parity is wrong, the recipient knows that at least one error occurred; otherwise, the recipient assumes that there were no errors.

(a) For $n = 5, p = 0.1$, what is the probability that the received message has errors which go undetected?

(b) For general n and p, write down an expression (as a sum) for the probability that the received message has errors which go undetected.

(c) Give a simplified expression, not involving a sum of a large number of terms, for the probability that the received message has errors which go undetected.

Hint for (c): Letting

$$a = \sum_{k \text{ even},\, k \geq 0} \binom{n}{k} p^k (1-p)^{n-k} \text{ and } b = \sum_{k \text{ odd},\, k \geq 1} \binom{n}{k} p^k (1-p)^{n-k},$$

the binomial theorem makes it possible to find simple expressions for $a + b$ and $a - b$, which then makes it possible to obtain a and b.

Independence of r.v.s

38. (a) Give an example of dependent r.v.s X and Y such that $P(X < Y) = 1$.

(b) Give an example of independent r.v.s X and Y such that $P(X < Y) = 1$.

39. Give an example of two discrete random variables X and Y on the same sample space such that X and Y have the same distribution, with support $\{1, 2, \ldots, 10\}$, but the event $X = Y$ *never* occurs. If X and Y are independent, is it still possible to construct such an example?

40. Suppose X and Y are discrete r.v.s such that $P(X = Y) = 1$. This means that X and Y always take on the same value.

(a) Do X and Y have the same PMF?

(b) Is it possible for X and Y to be independent?

41. If X, Y, Z are r.v.s such that X and Y are independent and Y and Z are independent, does it follow that X and Z are independent?

Hint: Think about simple and extreme examples.

42. Ⓢ Let X be a random day of the week, coded so that Monday is 1, Tuesday is 2, etc. (so X takes values $1, 2, \ldots, 7$, with equal probabilities). Let Y be the next day after X (again represented as an integer between 1 and 7). Do X and Y have the same distribution? What is $P(X < Y)$?

43. (a) Is it possible to have two r.v.s X and Y such that X and Y have the same distribution but $P(X < Y) \geq p$, where:
    - $p = 0.9$?
    - $p = 0.99$?
    - $p = 0.9999999999999$?
    - $p = 1$?

    For each, give an example showing it is possible, or prove it is impossible.
    Hint: Do the previous question first.

    (b) Consider the same question as in Part (a), but now assume that X and Y are independent. Do your answers change?

44. For x and y binary digits (0 or 1), let $x \oplus y$ be 0 if $x = y$ and 1 if $x \neq y$ (this operation is called *exclusive or* (often abbreviated to XOR), or *addition mod 2*).

    (a) Let $X \sim \text{Bern}(p)$ and $Y \sim \text{Bern}(1/2)$, independently. What is the distribution of $X \oplus Y$?

    (b) With notation as in (a), is $X \oplus Y$ independent of X? Is $X \oplus Y$ independent of Y? Be sure to consider both the case $p = 1/2$ and the case $p \neq 1/2$.

    (c) Let $X_1, \ldots, X_n$ be i.i.d. Bern(1/2). For each nonempty subset J of $\{1, 2, \ldots, n\}$, let

    $$Y_J = \bigoplus_{j \in J} X_j,$$

    where the notation means to "add" in the $\oplus$ sense all the elements of J; the order in which this is done doesn't matter since $x \oplus y = y \oplus x$ and $(x \oplus y) \oplus z = x \oplus (y \oplus z)$. Show that $Y_J \sim \text{Bern}(1/2)$ and that these $2^n - 1$ r.v.s are pairwise independent, but not independent. For example, we can use this to simulate 1023 pairwise independent fair coin tosses using only 10 independent fair coin tosses.

    Hint: Apply the previous parts with $p = 1/2$. Show that if J and K are two different nonempty subsets of $\{1, 2, \ldots, n\}$, then we can write $Y_J = A \oplus B, Y_K = A \oplus C$, where A consists of the X_i with $i \in J \cap K$, B consists of the X_i with $i \in J \cap K^c$, and C consists of the X_i with $i \in J^c \cap K$. Then A, B, C are independent since they are based on disjoint sets of X_i. Also, at most one of these sets of X_i can be empty. If $J \cap K = \emptyset$, then $Y_J = B, Y_K = C$. Otherwise, compute $P(Y_J = y, Y_K = z)$ by conditioning on whether $A = 1$.

Mixed practice

45. Ⓢ A new treatment for a disease is being tested, to see whether it is better than the standard treatment. The existing treatment is effective on 50% of patients. It is believed initially that there is a 2/3 chance that the new treatment is effective on 60% of patients, and a 1/3 chance that the new treatment is effective on 50% of patients. In a pilot study, the new treatment is given to 20 random patients, and is effective for 15 of them.

    (a) Given this information, what is the probability that the new treatment is better than the standard treatment?

    (b) A second study is done later, giving the new treatment to 20 new random patients. Given the results of the first study, what is the PMF for how many of the new patients the new treatment is effective on? (Letting p be the answer to (a), your answer can be left in terms of p.)

46. Independent Bernoulli trials are performed, with success probability 1/2 for each trial. An important question that often comes up in such settings is how many trials to perform. Many controversies have arisen in statistics over the issue of how to analyze data coming from an experiment where the number of trials can depend on the data collected so far.

    For example, if we can follow the rule "keep performing trials until there are more than twice as many failures as successes, and then stop", then naively looking at the ratio of failures to successes (if and when the process stops) will give more than 2:1 rather than the true theoretical 1:1 ratio; this could be a very misleading result! However, it might *never* happen that there are more than twice as many failures as successes; in this problem, you will find the probability of that happening.

    (a) Two gamblers, A and B, make a series of bets, where each has probability 1/2 of winning a bet, but A gets \$2 for each win and loses \$1 for each loss (a very favorable game for A!). Assume that the gamblers are allowed to borrow money, so they can and do gamble forever. Let p_k be the probability that A, starting with \$$k$, will ever reach \$0, for each $k \geq 0$. Explain how this story relates to the original problem, and how the original problem can be solved if we can find p_k.

    (b) Find p_k.

    Hint: As in the gambler's ruin, set up and solve a difference equation for p_k. We have $p_k \to 0$ as $k \to \infty$ (you don't need to prove this, but it should make sense since the game is so favorable to A, which will result in A's fortune going to ∞; a formal proof, not required here, could be done using the *law of large numbers*, an important theorem from Chapter 10). The solution can be written neatly in terms of the golden ratio.

    (c) Find the probability of ever having more than twice as many failures as successes with independent Bern(1/2) trials, as originally desired.

47. A copy machine is used to make n pages of copies per day. The machine has two trays in which paper gets loaded, and each page used is taken randomly and independently from one of the trays. At the beginning of the day, the trays are refilled so that they each have m pages.

    (a) Let pbinom(x, n, p) be the CDF of the Bin(n, p) distribution, evaluated at x. In terms of pbinom, find a simple expression for the probability that both trays have enough paper on any particular day, when this probability is strictly between 0 and 1 (also specify the values of m for which the probability is 0 and the values for which it is 1).

    Hint: Be careful about whether inequalities are strict, since the Binomial is discrete.

    (b) Using a computer, find the smallest value of m for which there is at least a 95% chance that both trays have enough paper on a particular day, for $n = 10, n = 100, n = 1000$, and $n = 10000$.

    Hint: If you use R, you may find the following commands useful:
    `g <- function(m,n)` *[your answer from (a)]* defines a function g such that $g(m, n)$ is your answer from (a), `g(1:100,100)` gives the vector $(g(1, 100), \ldots, g(100, 100))$, `which(v>0.95)` gives the indices of the components of vector **v** that exceed 0.95, and `min(w)` gives the minimum of a vector **w**.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

4

Expectation

4.1 Definition of expectation

In the previous chapter, we introduced the *distribution* of a random variable, which gives us full information about the probability that the r.v. will fall into any particular set. For example, we can say how likely it is that the r.v. will exceed 1000, that it will equal 5, or that it will be in the interval $[0, 7]$. It can be unwieldy to manage so many probabilities though, so often we want just one number summarizing the "average" value of the r.v.

There are several senses in which the word "average" is used, but by far the most commonly used is the *mean* of an r.v., also known as its *expected value*. In addition, much of statistics is about understanding *variability* in the world, so it is often important to know how "spread out" the distribution is; we will formalize this with the concepts of *variance* and *standard deviation*. As we'll see, variance and standard deviation are defined in terms of expected values, so the uses of expected values go far beyond just computing averages.

Given a list of numbers $x_1, x_2, \ldots, x_n$, the familiar way to average them is to add them up and divide by n. This is called the *arithmetic mean*, and is defined by

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

More generally, we can define a *weighted mean* of $x_1, \ldots, x_n$ as

$$\text{weighted-mean}(x) = \sum_{j=1}^{n} x_j p_j,$$

where the weights $p_1, \ldots, p_n$ are pre-specified nonnegative numbers that add up to 1 (so the unweighted mean $\bar{x}$ is obtained when $p_j = 1/n$ for all j).

The definition of expectation for a discrete r.v. is inspired by the weighted mean of a list of numbers, with weights given by probabilities.

Definition 4.1.1 (Expectation of a discrete r.v.)**.** The *expected value* (also called the *expectation* or *mean*) of a discrete r.v. X whose distinct possible values are

$x_1, x_2, \ldots$ is defined by

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j).$$

If the support is finite, then this is replaced by a finite sum. We can also write

$$E(X) = \sum_x \underbrace{x}_{\text{value}} \underbrace{P(X = x)}_{\text{PMF at } x},$$

where the sum is over the support of X (in any case, $xP(X = x)$ is 0 for any x not in the support). The expectation is undefined if $\sum_{j=1}^{\infty} |x_j| P(X = x_j)$ diverges, since then the series for $E(X)$ diverges or its value depends on the order in which the x_j are listed.

In words, the expected value of X is a weighted average of the possible values that X can take on, weighted by their probabilities. Let's check that the definition makes sense in a few simple examples:

1. Let X be the result of rolling a fair 6-sided die, so X takes on the values $1, 2, 3, 4, 5, 6$, with equal probabilities. Intuitively, we should be able to get the average by adding up these values and dividing by 6. Using the definition, the expected value is

   $$E(X) = \frac{1}{6}(1 + 2 + \cdots + 6) = 3.5,$$

   as we expected. Note that X *never* equals its mean in this example. This is similar to the fact that the average number of children per household in some country could be 1.8, but that doesn't mean that a typical household has 1.8 children!

2. Let $X \sim \text{Bern}(p)$ and $q = 1 - p$. Then

   $$E(X) = 1p + 0q = p,$$

   which makes sense intuitively since it is between the two possible values of X, compromising between 0 and 1 based on how likely each is. This is illustrated in Figure 4.1 for a case with $p < 1/2$: two pebbles are being balanced on a seesaw. For the seesaw to balance, the fulcrum (shown as a triangle) must be at p, which in physics terms is the *center of mass*.

   The frequentist interpretation would be to consider a large number of independent Bernoulli trials, each with probability p of success. Writing 1 for "success" and 0 for "failure", in the long run we would expect to have data consisting of a list of numbers where the proportion of 1's is very close to p. The average of a list of 0's and 1's *is* the proportion of 1's.

3. Let X have 3 distinct possible values, a_1, a_2, a_3, with probabilities p_1, p_2, p_3, respectively. Imagine running a simulation where n independent draws

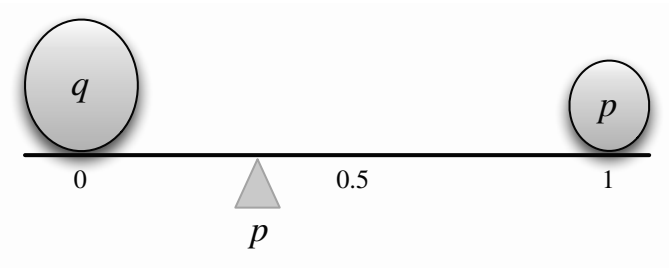


FIGURE 4.1
Center of mass of two pebbles, depicting that $E(X) = p$ for $X \sim \text{Bern}(p)$. Here q and p denote the masses of the two pebbles.

from the distribution of X are generated. For n large, we would expect to have about $p_1 n$ a_1's, $p_2 n$ a_2's, and $p_3 n$ a_3's. (We will look at a more mathematical version of this example when we study the law of large numbers in Chapter 10.) If the simulation results are close to these expected results, then the arithmetic mean of the simulation results is approximately

$$\frac{p_1 n \cdot a_1 + p_2 n \cdot a_2 + p_3 n \cdot a_3}{n} = p_1 a_1 + p_2 a_2 + p_3 a_3 = E(X).$$

Note that $E(X)$ depends only on the *distribution* of X. This follows directly from the definition, but is worth recording since it is fundamental.

Proposition 4.1.2. If X and Y are discrete r.v.s with the same distribution, then $E(X) = E(Y)$ (if either side exists).

Proof. In the definition of $E(X)$, we only need to know the PMF of X. ■

The converse of the above proposition is false since the expected value is just a one-number summary, not nearly enough to specify the entire distribution; it's a measure of where the "center" is but does not determine, for example, how spread out the distribution is or how likely the r.v. is to be positive. Figure 4.2 shows an example of two different PMFs with the same expected value (balancing point).

☣ **4.1.3** (Replacing an r.v. by its expectation)**.** For any discrete r.v. X, the expected value $E(X)$ is a *number* (if it exists). A common mistake is to replace an r.v. by its expectation without justification, which is wrong both mathematically (X is a function, $E(X)$ is a constant) and statistically (it ignores the variability of X), except in the degenerate case where X is a constant.

Notation 4.1.4. We often abbreviate $E(X)$ to EX. Similarly, we often abbreviate $E(X^2)$ to EX^2, and $E(X^n)$ to EX^n.

☣ **4.1.5.** Paying attention to the order of operations is crucial when working with expectation. As stated above, EX^2 is the expectation of the random variable X^2, *not* the square of the number EX. Unless the parentheses explicitly indicate otherwise,

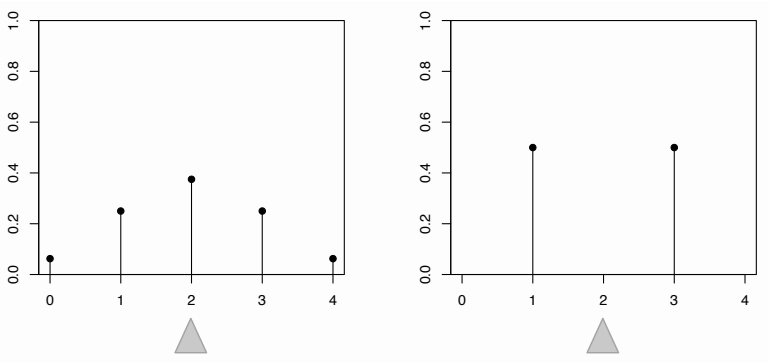


FIGURE 4.2
The expected value does not determine the distribution: different PMFs can have the same balancing point.

for the expectation of an r.v. raised to a power, first we take the power and then we take the expectation. For example, $E(X-1)^4$ is $E\left((X-1)^4\right)$, not $(E(X-1))^4$.

4.2 Linearity of expectation

The most important property of expectation is *linearity*: the expected value of a sum of r.v.s is the sum of the individual expected values.

Theorem 4.2.1 (Linearity of expectation)**.** For any r.v.s X, Y and any constant c,

$$E(X+Y) = E(X) + E(Y),$$
$$E(cX) = cE(X).$$

The second equation says that we can take out constant factors from an expectation; this is both intuitively reasonable and easily verified from the definition. The first equation, $E(X+Y) = E(X) + E(Y)$, also seems reasonable when X and Y are independent. What may be surprising is that it holds even if X and Y are dependent! To build intuition for this, consider the extreme case where X always equals Y. Then $X + Y = 2X$, and both sides of $E(X+Y) = E(X) + E(Y)$ are equal to $2E(X)$, so linearity still holds even in the most extreme case of dependence.

Linearity is true for all r.v.s, not just discrete r.v.s, but in this chapter we will prove it only for discrete r.v.s. Before proving linearity, it is worthwhile to recall some basic facts about averages. If we have a list of numbers, say $(1, 1, 1, 1, 1, 3, 3, 5)$, we can calculate their mean by adding all the values and dividing by the length of the list, so that each element of the list gets a weight of $\frac{1}{8}$:

$$\frac{1}{8}(1+1+1+1+1+3+3+5) = 2.$$

But another way to calculate the mean is to group together all the 1's, all the 3's, and all the 5's, and then take a weighted average, giving appropriate weights to 1's, 3's, and 5's:

$$\frac{5}{8} \cdot 1 + \frac{2}{8} \cdot 3 + \frac{1}{8} \cdot 5 = 2.$$

This insight—that averages can be calculated in two ways, *ungrouped* or *grouped*—is all that is needed to prove linearity! Recall that X is a function which assigns a real number to every outcome s in the sample space. The r.v. X may assign the same value to multiple sample outcomes. When this happens, our definition of expectation groups all these outcomes together into a *super-pebble* whose weight, $P(X = x)$, is the total weight of the constituent pebbles. This grouping process is illustrated in Figure 4.3 for a hypothetical r.v. taking values in $\{0, 1, 2\}$. So our definition of expectation corresponds to the grouped way of taking averages.

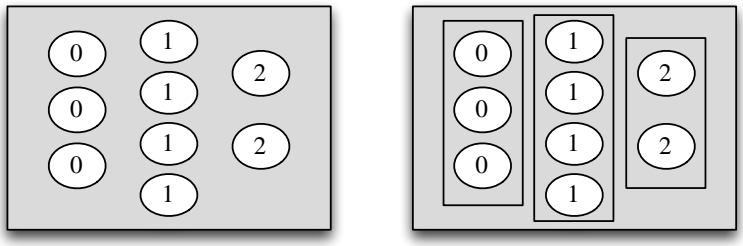


FIGURE 4.3
Left: X assigns a number to each pebble in the sample space. Right: Grouping the pebbles by the value that X assigns to them, the 9 pebbles become 3 super-pebbles. The weight of a super-pebble is the sum of the weights of the constituent pebbles.

The advantage of this definition is that it allows us to work with the distribution of X directly, without returning to the sample space. The disadvantage comes when we have to prove theorems like this one, for if we have another r.v. Y on the same sample space, the super-pebbles created by Y are different from those created from X, with different weights $P(Y = y)$; this makes it difficult to combine $\sum_x xP(X = x)$ and $\sum_y yP(Y = y)$.

Fortunately, we know there's another equally valid way to calculate an average: we can take a weighted average of the values of individual pebbles. In other words, if $X(s)$ is the value that X assigns to pebble s, we can take the weighted average

$$E(X) = \sum_s X(s)P(\{s\}),$$

where $P(\{s\})$ is the weight of pebble s. This corresponds to the ungrouped way of taking averages. The advantage of this definition is that it breaks down the sample space into the smallest possible units, so we are now using the *same* weights $P(\{s\})$ for every random variable defined on this sample space. If Y is another random

variable, then

$$E(Y) = \sum_s Y(s)P(\{s\}).$$

We *can* combine $\sum_s X(s)P(\{s\})$ and $\sum_s Y(s)P(\{s\})$, which gives

$$E(X)+E(Y) = \sum_s X(s)P(\{s\})+\sum_s Y(s)P(\{s\}) = \sum_s (X+Y)(s)P(\{s\}) = E(X+Y).$$

Another intuition for linearity of expectation is via the concept of *simulation*. If we simulate many, many times from the distribution of X, the histogram of the simulated values will look very much like the true PMF of X. In particular, the *arithmetic mean* of the simulated values will be very close to the true value of $E(X)$ (the precise nature of this convergence is described by the law of large numbers, an important theorem that we will discuss in detail in Chapter 10).

Let X and Y be r.v.s summarizing a certain experiment. Suppose we perform the experiment n times, where n is a very large number, and we write down the values realized by X and Y each time. For each repetition of the experiment, we obtain an X value, a Y value, and (by adding them) an $X + Y$ value. In Figure 4.4, each row represents a repetition of the experiment. The left column contains the draws of X, the middle column contains the draws of Y, and the right column contains the draws of $X + Y$.

There are two ways to calculate the sum of all the numbers in the last column. The straightforward way is just to add all the numbers in that column. But an equally valid way is to add all the numbers in the first column, add all the numbers in the second column, and then add the two column sums.

Dividing by n everywhere, what we've argued is that the following procedures are equivalent:

- Taking the arithmetic mean of all the numbers in the last column. By the law of large numbers, this is very close to $E(X + Y)$.
- Taking the arithmetic mean of the first column and the arithmetic mean of the second column, then adding the two column means. By the law of large numbers, this is very close to $E(X) + E(Y)$.

Linearity of expectation thus emerges as a simple fact about arithmetic (we're just adding numbers in two different orders)! Notice that nowhere in our argument did we rely on whether X and Y were independent. In fact, in Figure 4.4, X and Y appear to be dependent: Y tends to be large when X is large, and Y tends to be small when X is small (in the language of Chapter 7, we say that X and Y are *positively correlated*). But this dependence is irrelevant: shuffling the draws of Y could completely alter the pattern of dependence between X and Y, but would have no effect on the column sums.

X	Y	$X+Y$
3	4	7
2	2	4
6	8	14
10	23	33
1	−3	−2
1	0	1
5	9	14
4	1	5
⋮	⋮	⋮

$$\frac{1}{n}\sum_{i=1}^{n} x_i \quad + \quad \frac{1}{n}\sum_{i=1}^{n} y_i \quad = \quad \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i)$$

$$E(X) \quad + \quad E(Y) \quad = \quad E(X+Y)$$

FIGURE 4.4
Intuitive view of linearity of expectation. Each row represents a repetition of the experiment; the three columns are the realized values of X, Y, and $X+Y$, respectively. Adding all the numbers in the last column is equivalent to summing the first column and the second column separately, then adding the two column sums. So the mean of the last column is the sum of the first and second column means; this is linearity of expectation.

Linearity is an extremely handy tool for calculating expected values, often allowing us to bypass the definition of expected value altogether. Let's use linearity to find the expectations of the Binomial and Hypergeometric distributions.

Example 4.2.2 (Binomial expectation)**.** For $X \sim \text{Bin}(n,p)$, let's find $E(X)$ in two ways. By definition of expectation,

$$E(X) = \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}.$$

From Example 1.5.2, we know $k\binom{n}{k} = n\binom{n-1}{k-1}$, so

$$\begin{aligned}\sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} &= n\sum_{k=0}^{n}\binom{n-1}{k-1}p^k q^{n-k} \\ &= np\sum_{k=1}^{n}\binom{n-1}{k-1}p^{k-1}q^{n-k} \\ &= np\sum_{j=0}^{n-1}\binom{n-1}{j}p^j q^{n-1-j} \\ &= np.\end{aligned}$$

The sum in the penultimate line equals 1 because it is the sum of the $\text{Bin}(n-1,p)$ PMF (or by the binomial theorem). Therefore, $E(X) = np$.

This proof required us to remember combinatorial identities and manipulate binomial coefficients. Using linearity of expectation, we obtain a *much* shorter path to the same result. Let's write X as the sum of n independent $\text{Bern}(p)$ r.v.s:

$$X = I_1 + \cdots + I_n,$$

where each I_j has expectation $E(I_j) = 1p + 0q = p$. By linearity,

$$E(X) = E(I_1) + \cdots + E(I_n) = np. \qquad \square$$

Example 4.2.3 (Hypergeometric expectation)**.** Let $X \sim \text{HGeom}(w,b,n)$, interpreted as the number of white balls in a sample of size n drawn without replacement from an urn with w white and b black balls. As in the Binomial case, we can write X as a sum of Bernoulli random variables,

$$X = I_1 + \cdots + I_n,$$

where I_j equals 1 if the jth ball in the sample is white and 0 otherwise. By symmetry, $I_j \sim \text{Bern}(p)$ with $p = w/(w+b)$, since unconditionally the jth ball drawn is equally likely to be any of the balls.

Unlike in the Binomial case, the I_j are not independent, since the sampling is without replacement: given that a ball in the sample is white, there is a lower

chance that another ball in the sample is white. However, linearity still holds for dependent random variables! Thus,

$$E(X) = nw/(w+b).$$ □

As another example of the power of linearity, we can give a quick proof of the intuitive idea that "bigger r.v.s have bigger expectations".

Proposition 4.2.4 (Monotonicity of expectation)**.** Let X and Y be r.v.s such that $X \geq Y$ with probability 1. Then $E(X) \geq E(Y)$, with equality holding if and only if $X = Y$ with probability 1.

Proof. This result holds for all r.v.s, but we will prove it only for discrete r.v.s since this chapter focuses on discrete r.v.s. The r.v. $Z = X - Y$ is nonnegative (with probability 1), so $E(Z) \geq 0$ since $E(Z)$ is defined as a sum of nonnegative terms. By linearity,

$$E(X) - E(Y) = E(X - Y) \geq 0,$$

as desired. If $E(X) = E(Y)$, then by linearity we also have $E(Z) = 0$, which implies that $P(X = Y) = P(Z = 0) = 1$ since if even one term in the sum defining $E(Z)$ is positive, then the whole sum is positive. ■

4.3 Geometric and Negative Binomial

We now introduce two more famous discrete distributions, the Geometric and Negative Binomial, and calculate their expected values.

Story 4.3.1 (Geometric distribution)**.** Consider a sequence of independent Bernoulli trials, each with the same success probability $p \in (0, 1)$, with trials performed until a success occurs. Let X be the number of *failures* before the first successful trial. Then X has the *Geometric distribution* with parameter p; we denote this by $X \sim \text{Geom}(p)$. □

For example, if we flip a fair coin until it lands Heads for the first time, then the number of Tails before the first occurrence of Heads is distributed as Geom(1/2).

To get the Geometric PMF from the story, imagine the Bernoulli trials as a string of 0's (failures) ending in a single 1 (success). Each 0 has probability $q = 1 - p$ and the final 1 has probability p, so a string of k failures followed by one success has probability $q^k p$.

Theorem 4.3.2 (Geometric PMF)**.** If $X \sim \text{Geom}(p)$, then the PMF of X is

$$P(X = k) = q^k p$$

for $k = 0, 1, 2, \ldots$, where $q = 1 - p$.

This is a valid PMF because, summing a geometric series (see the math appendix for a review of geometric series), we have

$$\sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = p \cdot \frac{1}{1-q} = 1.$$

Just as the binomial theorem shows that the Binomial PMF is valid, a geometric series shows that the Geometric PMF is valid! A geometric series can also be used to obtain the Geometric CDF.

Theorem 4.3.3 (Geometric CDF)**.** If $X \sim \text{Geom}(p)$, then the CDF of X is

$$F(x) = \begin{cases} 1 - q^{\lfloor x \rfloor + 1}, & \text{if } x \geq 0; \\ 0, & \text{if } x < 0, \end{cases}$$

where $q = 1 - p$ and $\lfloor x \rfloor$ is the greatest integer less than or equal to x.

Proof. Let F be the CDF of X. We will find $F(x)$ first for the case $x < 0$, then for the case that x is a nonnegative integer, and lastly for the case that x is a nonnegative real number. For $x < 0$, $F(x) = 0$ since X can't be negative. For n a nonnegative integer,

$$F(n) = \sum_{k=0}^{n} P(X = k) = p \sum_{k=0}^{n} q^k = p \cdot \frac{1 - q^{n+1}}{1 - q} = 1 - q^{n+1}.$$

We can also get the same result from the fact that the event $X \geq n + 1$ means that the first $n + 1$ trials were failures:

$$F(n) = 1 - P(X > n) = 1 - P(X \geq n + 1) = 1 - q^{n+1}.$$

For real $x \geq 0$,

$$F(x) = P(X \leq x) = P(X \leq \lfloor x \rfloor),$$

since X always takes on integer values. For example,

$$P(X \leq 3.7) = P(X \leq 3) + P(3 < X \leq 3.7) = P(X \leq 3).$$

Therefore, F is as claimed. ■

Figure 4.3 displays the Geom(0.5) PMF and CDF from 0 to 6. All Geometric PMFs have a similar shape; the greater the success probability p, the more quickly the PMF decays to 0.

☣ **4.3.4** (Conventions for the Geometric)**.** There are differing conventions for the definition of the Geometric distribution; some sources define the Geometric as the total number of *trials*, including the success. In this book, the Geometric distribution excludes the success, and the *First Success* distribution includes the success.

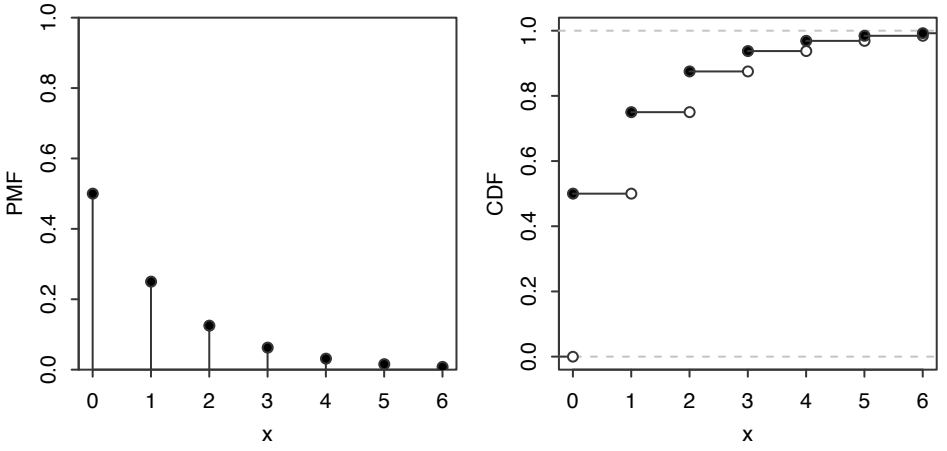


FIGURE 4.5
Geom(0.5) PMF and CDF.

Definition 4.3.5 (First Success distribution)**.** In a sequence of independent Bernoulli trials with success probability p, let Y be the number of *trials* until the first successful trial, including the success. Then Y has the *First Success distribution* with parameter p; we denote this by $Y \sim \text{FS}(p)$.

It is easy to convert back and forth between the two but important to be careful about which convention is being used. If $Y \sim \text{FS}(p)$ then $Y - 1 \sim \text{Geom}(p)$, and we can convert between the PMFs of Y and $Y - 1$ by writing

$$P(Y = k) = P(Y - 1 = k - 1).$$

Conversely, if $X \sim \text{Geom}(p)$, then $X + 1 \sim \text{FS}(p)$.

Example 4.3.6 (Geometric expectation)**.** Let $X \sim \text{Geom}(p)$. By definition,

$$E(X) = \sum_{k=0}^{\infty} kq^k p,$$

where $q = 1 - p$. This sum looks unpleasant; it's not a geometric series because of the extra k multiplying each term. But we notice that each term looks similar to kq^{k-1}, the derivative of q^k (with respect to q), so let's start there:

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}.$$

This geometric series converges since $0 < q < 1$. Differentiating both sides with respect to q, we get

$$\sum_{k=0}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}.$$

Finally, if we multiply both sides by pq, we recover the original sum we wanted to find:

$$E(X) = \sum_{k=0}^{\infty} kq^k p = pq \sum_{k=0}^{\infty} kq^{k-1} = pq \frac{1}{(1-q)^2} = \frac{q}{p}.$$

In Example 9.1.8, we will give a story proof of the same result, based on first-step analysis: condition on the result of the first trial in the story interpretation of X. If the first trial is a success, we know $X = 0$ and if it's a failure, we have one wasted trial and then are back where we started. □

Example 4.3.7 (First Success expectation)**.** Since we can write $Y \sim \text{FS}(p)$ as $Y = X + 1$ where $X \sim \text{Geom}(p)$, we have

$$E(Y) = E(X + 1) = \frac{q}{p} + 1 = \frac{1}{p}.$$ □

The Negative Binomial distribution generalizes the Geometric distribution: instead of waiting for just one success, we can wait for any predetermined number r of successes.

Story 4.3.8 (Negative Binomial distribution)**.** In a sequence of independent Bernoulli trials with success probability p, if X is the number of *failures* before the rth success, then X is said to have the *Negative Binomial distribution* with parameters r and p, denoted $X \sim \text{NBin}(r, p)$. □

Both the Binomial and the Negative Binomial distributions are based on independent Bernoulli trials; they differ in the *stopping rule* and in what they are counting. The Binomial counts the number of successes in a fixed number of *trials*; the Negative Binomial counts the number of failures until a fixed number of *successes*.

In light of these similarities, it comes as no surprise that the derivation of the Negative Binomial PMF bears a resemblance to the corresponding derivation for the Binomial.

Theorem 4.3.9 (Negative Binomial PMF)**.** If $X \sim \text{NBin}(r, p)$, then the PMF of X is

$$P(X = n) = \binom{n + r - 1}{r - 1} p^r q^n$$

for $n = 0, 1, 2 \ldots$, where $q = 1 - p$.

Proof. Imagine a string of 0's and 1's, with 1's representing successes. The probability of any *specific* string of n 0's and r 1's is $p^r q^n$. How many such strings are there? Because we stop as soon as we hit the rth success, the string must terminate in a 1. Among the other $n + r - 1$ positions, we choose $r - 1$ places for the remaining 1's to go. So the overall probability of exactly n failures before the rth success is

$$P(X = n) = \binom{n + r - 1}{r - 1} p^r q^n, \quad n = 0, 1, 2, \ldots.$$ ■

Just as a Binomial r.v. can be represented as a sum of i.i.d. Bernoullis, a Negative Binomial r.v. can be represented as a sum of i.i.d. Geometrics.

Theorem 4.3.10. Let $X \sim \text{NBin}(r, p)$, viewed as the number of failures before the rth success in a sequence of independent Bernoulli trials with success probability p. Then we can write $X = X_1 + \cdots + X_r$ where the X_i are i.i.d. Geom(p).

Proof. Let X_1 be the number of failures until the first success, X_2 be the number of failures between the first success and the second success, and in general, X_i be the number of failures between the $(i-1)$st success and the ith success.

Then $X_1 \sim \text{Geom}(p)$ by the story of the Geometric distribution. After the first success, the number of additional failures until the next success is still Geometric! So $X_2 \sim \text{Geom}(p)$, and similarly for all the X_i. Furthermore, the X_i are independent because the trials are all independent of each other. Adding the X_i, we get the total number of failures before the rth success, which is X. ■

Using linearity, the expectation of the Negative Binomial now follows without any additional calculations.

Example 4.3.11 (Negative Binomial expectation)**.** Let $X \sim \text{NBin}(r, p)$. By the previous theorem, we can write $X = X_1 + \cdots + X_r$, where the X_i are i.i.d. Geom(p). By linearity,

$$E(X) = E(X_1) + \cdots + E(X_r) = r \cdot \frac{q}{p}.$$ □

The next example is a famous problem in probability and an instructive application of the Geometric and First Success distributions. It is usually stated as a problem about collecting coupons, hence its name, but we'll use toys instead of coupons.

Example 4.3.12 (Coupon collector)**.** Suppose there are n types of toys, which you are collecting one by one, with the goal of getting a complete set. When collecting toys, the toy types are random (as is sometimes the case, for example, with toys included in cereal boxes or included with kids' meals from a fast food restaurant). Assume that each time you collect a toy, it is equally likely to be any of the n types. What is the expected number of toys needed until you have a complete set?

Solution:

Let N be the number of toys needed; we want to find $E(N)$. Our strategy will be to break up N into a sum of simpler r.v.s so that we can apply linearity. So write

$$N = N_1 + N_2 + \cdots + N_n,$$

where N_1 is the number of toys until the first toy type you haven't seen before (which is always 1, as the first toy is always a new type), N_2 is the additional number of toys until the second toy type you haven't seen before, and so forth. Figure 4.6 illustrates these definitions with $n = 3$ toy types.

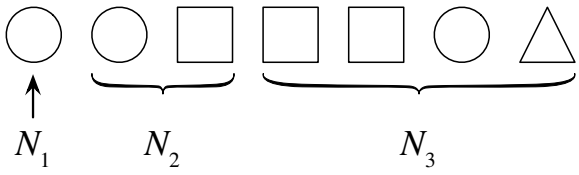


FIGURE 4.6
Coupon collector, $n = 3$. Here N_1 is the time (number of toys collected) until the first new toy type, N_2 is the additional time until the second new type, and N_3 is the additional time until the third new type. The total number of toys for a complete set is $N_1 + N_2 + N_3$.

By the story of the FS distribution, $N_2 \sim \text{FS}((n-1)/n)$: after collecting the first toy type, there's a $1/n$ chance of getting the same toy you already had (failure) and an $(n-1)/n$ chance you'll get something new (success). Similarly, N_3, the additional number of toys until the third new toy type, is distributed $\text{FS}((n-2)/n)$. In general,

$$N_j \sim \text{FS}((n-j+1)/n).$$

By linearity,

$$\begin{aligned} E(N) &= E(N_1) + E(N_2) + E(N_3) + \cdots + E(N_n) \\ &= 1 + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + n \\ &= n\sum_{j=1}^{n} \frac{1}{j}. \end{aligned}$$

For large n, this is very close to $n(\log n + 0.577)$.

Before we leave this example, let's take a moment to connect it to our proof of Theorem 4.3.10, the representation of the Negative Binomial as a sum of i.i.d. Geometrics. In both problems, we are waiting for a specified number of successes, and we approach the problem by considering the intervals between successes. There are two major differences:

- In Theorem 4.3.10, we exclude the successes themselves, so the number of failures between two successes is Geometric. In the coupon collector problem, we include the successes because we want to count the total number of toys, so we have First Success r.v.s instead.
- In Theorem 4.3.10, the probability of success in each trial never changes, so the total number of failures is a sum of *i.i.d.* Geometrics. In the coupon collector problem, the probability of success decreases after each success, since it becomes harder and harder to find a new toy type you haven't seen before; so the N_j are not identically distributed, though they are independent. □

☣ **4.3.13** (Expectation of a nonlinear function of an r.v.)**.** Expectation is linear,

but in general we do *not* have $E(g(X)) = g(E(X))$ for arbitrary functions g. We must be careful not to move the E around when g is not linear. The next example shows a situation in which $E(g(X))$ is *very* different from $g(E(X))$.

Example 4.3.14 (St. Petersburg paradox)**.** Suppose a wealthy stranger offers to play the following game with you. You will flip a fair coin until it lands Heads for the first time, and you will receive \$2 if the game lasts for 1 round, \$4 if the game lasts for 2 rounds, \$8 if the game lasts for 3 rounds, and in general, $\$2^n$ if the game lasts for n rounds. What is the fair value of this game (the expected payoff)? How much would you be willing to pay to play this game once?

Solution:

Let X be your winnings from playing the game. By definition, $X = 2^N$ where N is the number of rounds that the game lasts. Then X is 2 with probability 1/2, 4 with probability 1/4, 8 with probability 1/8, and so on, so

$$E(X) = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \cdots = \infty.$$

The expected winnings are infinite! On the other hand, the number of rounds N that the game lasts is the number of tosses until the first Heads, so $N \sim \text{FS}(1/2)$ and $E(N) = 2$. Thus $E(2^N) = \infty$ while $2^{E(N)} = 4$. Infinity certainly does not equal 4, illustrating the danger of confusing $E(g(X))$ with $g(E(X))$ when g is not linear.

This problem is often considered a paradox because although the game's expected payoff is infinite, most people would not be willing to pay very much to play the game (even if they could afford to lose the money). One explanation is to note that *the amount of money in the real world is finite*. Suppose that if the game lasts longer than 40 rounds, the wealthy stranger flees the country and you get nothing. Since $2^{40} \approx 1.1 \times 10^{12}$, this still gives you the potential to earn over a trillion dollars, and anyway it's incredibly unlikely that the game will last longer than 40 rounds. But in this setting, your expected value is

$$E(X) = \sum_{n=1}^{40} \frac{1}{2^n} \cdot 2^n + \sum_{n=41}^{\infty} \frac{1}{2^n} \cdot 0 = 40.$$

Is this drastic reduction because the wealthy stranger may flee the country? Let's suppose instead that the wealthy stranger caps your winnings at 2^{40}, so if the game lasts more than 40 rounds you will get this amount rather than walking away empty-handed. Now your expected value is

$$E(X) = \sum_{n=1}^{40} \frac{1}{2^n} \cdot 2^n + \sum_{n=41}^{\infty} \frac{1}{2^n} \cdot 2^{40} = 40 + 1 = 41,$$

an increase of only \$1 from the previous scenario. The ∞ in the St. Petersburg paradox is driven by an infinite "tail" of extremely rare events where you get extremely large payoffs. Cutting off this tail at some point, which makes sense in the real world, dramatically reduces the expected value of the game. □

4.4 Indicator r.v.s and the fundamental bridge

This section is devoted to *indicator random variables*, which we already encountered in the previous chapter but will treat in much greater detail here. In particular, we will show that indicator r.v.s are an extremely useful tool for calculating expected values.

Recall from the previous chapter that the indicator r.v. I_A (or $I(A)$) for an event A is defined to be 1 if A occurs and 0 otherwise. So I_A is a Bernoulli random variable, where success is defined as "A occurs" and failure is defined as "A does not occur". Some useful properties of indicator r.v.s are summarized below.

Theorem 4.4.1 (Indicator r.v. properties)**.** Let A and B be events. Then the following properties hold.

1. $(I_A)^k = I_A$ for any positive integer k.
2. $I_{A^c} = 1 - I_A$.
3. $I_{A \cap B} = I_A I_B$.
4. $I_{A \cup B} = I_A + I_B - I_A I_B$.

Proof. Property 1 holds since $0^k = 0$ and $1^k = 1$ for any positive integer k. Property 2 holds since $1 - I_A$ is 1 if A does not occur and 0 if A occurs. Property 3 holds since $I_A I_B$ is 1 if both I_A and I_B are 1, and 0 otherwise. Property 4 holds since

$$I_{A \cup B} = 1 - I_{A^c \cap B^c} = 1 - I_{A^c} I_{B^c} = 1 - (1 - I_A)(1 - I_B) = I_A + I_B - I_A I_B. \quad \blacksquare$$

Indicator r.v.s provide a link between probability and expectation; we call this fact the *fundamental bridge*.

Theorem 4.4.2 (Fundamental bridge between probability and expectation)**.** There is a one-to-one correspondence between events and indicator r.v.s, and the probability of an event A is the expected value of its indicator r.v. I_A:

$$P(A) = E(I_A).$$

Proof. For any event A, we have an indicator r.v. I_A. This is a one-to-one correspondence since A uniquely determines I_A and vice versa (to get from I_A back to A, we can use the fact that $A = \{s \in S : I_A(s) = 1\}$). Since $I_A \sim \text{Bern}(p)$ with $p = P(A)$, we have $E(I_A) = P(A)$. $\blacksquare$

The fundamental bridge connects events to their indicator r.v.s, and allows us to express *any* probability as an expectation. As an example, we give a short proof of inclusion-exclusion and a related inequality known as *Boole's inequality* or *Bonferroni's inequality* using indicator r.v.s.

Example 4.4.3 (Boole, Bonferroni, and inclusion-exclusion)**.** Let $A_1, A_2, \ldots, A_n$ be events. Note that

$$I(A_1 \cup \cdots \cup A_n) \leq I(A_1) + \cdots + I(A_n),$$

since if the left-hand side is 0 this is immediate, and if the left-hand side is 1 then at least one term on the right-hand side must be 1. Taking the expectation of both sides and using linearity and the fundamental bridge, we have

$$P(A_1 \cup \cdots \cup A_n) \leq P(A_1) + \cdots + P(A_n),$$

which is called *Boole's inequality* or *Bonferroni's inequality*. To prove inclusion-exclusion for $n = 2$, we can take the expectation of both sides in Property 4 of Theorem 4.4.1. For general n, we can use properties of indicator r.v.s as follows:

$$\begin{aligned} 1 - I(A_1 \cup \cdots \cup A_n) &= I(A_1^c \cap \cdots \cap A_n^c) \\ &= (1 - I(A_1)) \cdots (1 - I(A_n)) \\ &= 1 - \sum_i I(A_i) + \sum_{i<j} I(A_i) I(A_j) - \cdots + (-1)^n I(A_1) \cdots I(A_n). \end{aligned}$$

Taking the expectation of both sides, by the fundamental bridge we have proven the inclusion-exclusion theorem. □

Conversely, the fundamental bridge is also extremely useful in many expected value problems. We can often express a complicated discrete r.v. whose distribution we don't know as a sum of indicator r.v.s, which are extremely simple. The fundamental bridge lets us find the expectation of the indicators; then, using linearity, we obtain the expectation of our original r.v. This strategy is extremely useful and versatile—in fact, we already used it when deriving the expectations of the Binomial and Hypergeometric distributions earlier in this chapter!

Recognizing problems that are amenable to this strategy and then defining the indicator r.v.s takes practice, so it is important to study a lot of examples and solve a lot of problems. In applying the strategy to a random variable that counts the number of [noun]s, we should have an indicator for each potential [noun]. This [noun] could be a person, place, or thing; we will see examples of all three types.

We'll start by revisiting two problems from Chapter 1, de Montmort's matching problem and the birthday problem.

Example 4.4.4 (Matching continued)**.** We have a well-shuffled deck of n cards, labeled 1 through n. A card is a *match* if the card's position in the deck matches the card's label. Let X be the number of matches; find $E(X)$.

Solution:

First let's check whether X could have any of the named distributions we have studied. The Binomial and Hypergeometric are the only two candidates since the value of X must be an integer between 0 and n. But neither of these distributions

has the right support because X can't take on the value $n-1$: if $n-1$ cards are matches, then the nth card must be a match as well. So X does not follow a named distribution we have studied, but we can readily find its mean using indicator r.v.s: let's write $X = I_1 + I_2 + \cdots + I_n$, where

$$I_j = \begin{cases} 1 & \text{if the } j\text{th card in the deck is a match,} \\ 0 & \text{otherwise.} \end{cases}$$

In other words, I_j is the indicator for A_j, the event that the jth card in the deck is a match. We can imagine that each I_j "raises its hand" to be counted if its card is a match; adding up the raised hands, we get the total number of matches, X.

By the fundamental bridge,

$$E(I_j) = P(A_j) = \frac{1}{n}$$

for all j. So by linearity,

$$E(X) = E(I_1) + \cdots + E(I_n) = n \cdot \frac{1}{n} = 1.$$

The expected number of matched cards is 1, regardless of n. Even though the I_j are dependent in a complicated way that makes the distribution of X neither Binomial nor Hypergeometric, linearity still holds. □

Example 4.4.5 (Distinct birthdays, birthday matches)**.** In a group of n people, under the usual assumptions about birthdays, what is the expected number of distinct birthdays among the n people, i.e., the expected number of days on which at least one of the people was born? What is the expected number of birthday matches, i.e., pairs of people with the same birthday?

Solution:

Let X be the number of distinct birthdays, and write $X = I_1 + \cdots + I_{365}$, where

$$I_j = \begin{cases} 1 & \text{if the } j\text{th day is represented,} \\ 0 & \text{otherwise.} \end{cases}$$

We create an indicator for each *day* of the year because X counts the number of *days* of the year that are represented. By the fundamental bridge,

$$E(I_j) = P(j\text{th day is represented}) = 1 - P(\text{no one born on day } j) = 1 - \left(\frac{364}{365}\right)^n$$

for all j. Then by linearity,

$$E(X) = 365\left(1 - \left(\frac{364}{365}\right)^n\right).$$

Now let Y be the number of birthday matches. Label the people as $1, 2, \ldots, n$, and order the $\binom{n}{2}$ pairs of people in some definite way. Then we can write

$$Y = J_1 + \cdots + J_{\binom{n}{2}},$$

where J_i is the indicator of the ith pair of people having the same birthday. We create an indicator for each *pair of people* since Y counts the number of *pairs of people* with the same birthday. The probability of any two people having the same birthday is $1/365$, so again by the fundamental bridge and linearity,

$$E(Y) = \frac{\binom{n}{2}}{365}. \qquad \square$$

In addition to the fundamental bridge and linearity, the last two examples used a basic form of symmetry to simplify the calculations greatly: within each sum of indicator r.v.s, each indicator had the same expected value. For example, in the matching problem the probability of the jth card being a match does not depend on j, so we can just take n times the expected value of the first indicator r.v.

Other forms of symmetry can also be extremely helpful when available. The next two examples showcase a form of symmetry that stems from having equally likely permutations. Note how symmetry, linearity, and the fundamental bridge are used in tandem to make seemingly very hard problems manageable.

Example 4.4.6 (Putnam problem)**.** A permutation $a_1, a_2, \ldots, a_n$ of $1, 2, \ldots, n$ has a *local maximum* at j if $a_j > a_{j-1}$ and $a_j > a_{j+1}$ (for $2 \leq j \leq n-1$; for $j = 1$, a local maximum at j means $a_1 > a_2$ while for $j = n$, it means $a_n > a_{n-1}$). For example, $4, 2, 5, 3, 6, 1$ has 3 local maxima, at positions 1, 3, and 5. The Putnam exam (a famous, hard math competition, on which the median score is often a 0) from 2006 posed the following question: for $n \geq 2$, what is the average number of local maxima of a random permutation of $1, 2, \ldots, n$, with all $n!$ permutations equally likely?

Solution:

This problem can be solved quickly using indicator r.v.s, symmetry, and the fundamental bridge. Let $I_1, \ldots, I_n$ be indicator r.v.s, where I_j is 1 if there is a local maximum at position j, and 0 otherwise. We are interested in the expected value of $\sum_{j=1}^{n} I_j$. For $1 < j < n$, $EI_j = 1/3$ since having a local maximum at j is equivalent to a_j being the largest of a_{j-1}, a_j, a_{j+1}, which has probability $1/3$ since all orders are equally likely. For $j = 1$ or $j = n$, we have $EI_j = 1/2$ since then there is only one neighbor. Thus, by linearity,

$$E\left(\sum_{j=1}^{n} I_j\right) = 2 \cdot \frac{1}{2} + (n-2) \cdot \frac{1}{3} = \frac{n+1}{3}. \qquad \square$$

The next example introduces the *Negative Hypergeometric* distribution, which completes the following table. The table shows the distributions for four sampling schemes: the sampling can be done with or without replacement, and the stopping rule can require a fixed number of draws or a fixed number of successes.

	With replacement	**Without replacement**
Fixed number of trials	Binomial	Hypergeometric
Fixed number of successes	Negative Binomial	Negative Hypergeometric

Example 4.4.7 (Negative Hypergeometric)**.** An urn contains w white balls and b black balls, which are randomly drawn one by one *without replacement*, until r white balls have been obtained. The number of black balls drawn before drawing the rth white ball has a *Negative Hypergeometric* distribution with parameters w, b, r. We denote this distribution by $\text{NHGeom}(w, b, r)$. Of course, we assume that $r \leq w$. For example, if we shuffle a deck of cards and deal them one at a time, the number of cards dealt before uncovering the first ace is $\text{NHGeom}(4, 48, 1)$.

As another example, suppose a college offers g good courses and b bad courses (for some definition of "good" and "bad"), and a student wants to find 4 good courses to take. Not having any idea which of the courses are good, the student randomly tries out courses one at a time, stopping when they have obtained 4 good courses. Then the number of bad courses the student tries out is $\text{NHGeom}(g, b, 4)$.

We can obtain the PMF of $X \sim \text{NHGeom}(w, b, r)$ by noting that, in the urn context, $X = k$ means that the $(r + k)$th ball chosen is white and exactly $r - 1$ of the first $r + k - 1$ balls chosen are white. This gives

$$P(X = k) = \frac{\binom{w}{r-1}\binom{b}{k}}{\binom{w+b}{r+k-1}} \cdot \frac{w - r + 1}{w + b - r - k + 1}$$

for $k = 0, 1, \ldots, b$ (and 0 otherwise).

Alternatively, we can imagine that we continue drawing balls until the urn has been emptied out; this is valid since whether or not we continue to draw balls after obtaining the rth white ball has no effect on X. Think of the $w + b$ balls as lined up in a random order, the order in which they will be drawn.

Then $X = k$ means that among the first $r + k - 1$ balls there are exactly $r - 1$ white balls, then there is a white ball, and then among the last $w + b - r - k$ balls there are exactly $w - r$ white balls. All $\binom{w+b}{w}$ possibilities for the locations of the white balls in the line are equally likely. So by the naive definition of probability, we have the following slightly simpler expression for the PMF:

$$P(X = k) = \frac{\binom{r+k-1}{r-1}\binom{w+b-r-k}{w-r}}{\binom{w+b}{w}},$$

for $k = 0, 1, \ldots, b$ (and 0 otherwise).

Finding the expected value of a Negative Hypergeometric r.v. directly from the definition of expectation results in complicated sums. But the answer is very simple: for $X \sim \text{NHGeom}(w, b, r)$, we have $E(X) = rb/(w + 1)$.

Let's prove this using indicator r.v.s. As explained above, we can assume that we

continue drawing balls until the urn is empty. First consider the case $r = 1$. Label the black balls as $1, 2, \ldots, b$, and let I_j be the indicator of black ball j being drawn before any white balls have been drawn. Then $P(I_j = 1) = 1/(w+1)$ since, listing out the order in which black ball j and the white balls are drawn (ignoring the other balls), all orders are equally likely by symmetry, and $I_j = 1$ is equivalent to black ball j being first in this list. So by linearity,

$$E\left(\sum_{j=1}^{b} I_j\right) = \sum_{j=1}^{b} E(I_j) = b/(w+1).$$

Sanity check: This answer makes sense since it is increasing in b, decreasing in w, and correct in the extreme cases $b = 0$ (when no black balls will be drawn) and $w = 0$ (when all the black balls will be exhausted before drawing a nonexistent white ball). Moreover, note that $b/(w+1)$ looks similar to, but is strictly smaller than, b/w, which is the expected value of a $\text{Geom}(w/(w+b))$ r.v. It makes sense that sampling without replacement should give a smaller expected waiting time than sampling with replacement. Similarly, if you are searching for something you lost, it makes more sense to choose locations to check without replacement, rather than wasting time looking over and over again in locations you already ruled out.

For general r, write $X = X_1 + X_2 + \cdots + X_r$, where X_1 is the number of black balls before the first white ball, X_2 is the number of black balls after the first white ball but before the second white ball, etc. By essentially the same argument we used to handle the $r = 1$ case, we have $E(X_j) = b/(w+1)$ for each j. So by linearity,

$$E(X) = rb/(w+1). \qquad \square$$

Closely related to indicator r.v.s is an alternative expression for the expectation of a nonnegative integer-valued r.v. X. Rather than summing up values of X times values of the PMF of X, we can sum up probabilities of the form $P(X > n)$ (known as *tail probabilities*), over nonnegative integers n.

Theorem 4.4.8 (Expectation via survival function)**.** Let X be a nonnegative integer-valued r.v. Let F be the CDF of X, and $G(x) = 1 - F(x) = P(X > x)$. The function G is called the *survival function* of X. Then

$$E(X) = \sum_{n=0}^{\infty} G(n).$$

That is, we can obtain the expectation of X by summing up the survival function (or, stated otherwise, summing up *tail probabilities* of the distribution).

Proof. For simplicity, we will prove the result only for the case that X is *bounded*, i.e., there is a nonnegative integer b such that X is always at most b. We can represent X as a sum of indicator r.v.s: $X = I_1 + I_2 + \cdots + I_b$, where $I_n = I(X \geq n)$. For

example, if $X = 7$ occurs, then I_1 through I_7 equal 1 while the other indicators equal 0.

By linearity and the fundamental bridge, and the fact that $\{X \geq k\}$ is the same event as $\{X > k-1\}$,

$$E(X) = \sum_{k=1}^{b} E(I_k) = \sum_{k=1}^{b} P(X \geq k) = \sum_{n=0}^{b-1} P(X > n) = \sum_{n=0}^{\infty} G(n). \quad \blacksquare$$

As a quick example, we use the above result to give another derivation of the mean of a Geometric r.v.

Example 4.4.9 (Geometric expectation redux)**.** Let $X \sim \text{Geom}(p)$, and $q = 1-p$. Using the Geometric story, $\{X > n\}$ is the event that the first $n+1$ trials are all failures. So by Theorem 4.4.8,

$$E(X) = \sum_{n=0}^{\infty} P(X > n) = \sum_{n=0}^{\infty} q^{n+1} = \frac{q}{1-q} = \frac{q}{p},$$

confirming what we already knew about the mean of a Geometric. $\square$

4.5 Law of the unconscious statistician (LOTUS)

As we saw in the St. Petersburg paradox, $E(g(X))$ does *not* equal $g(E(X))$ in general if g is not linear. So how do we correctly calculate $E(g(X))$? Since $g(X)$ is an r.v., one way is to first find the distribution of $g(X)$ and then use the definition of expectation. Perhaps surprisingly, it turns out that it is possible to find $E(g(X))$ directly using the distribution of X, without first having to find the distribution of $g(X)$. This is done using the *law of the unconscious statistician* (LOTUS).

Theorem 4.5.1 (LOTUS)**.** If X is a discrete r.v. and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then

$$E(g(X)) = \sum_{x} g(x) P(X = x),$$

where the sum is taken over all possible values of X.

This means that we can get the expected value of $g(X)$ knowing only $P(X = x)$, the PMF of X; we don't need to know the PMF of $g(X)$. The name comes from the fact that in going from $E(X)$ to $E(g(X))$ it is tempting just to change x to $g(x)$ in the definition, which can be done very easily and mechanically, perhaps in a state of unconsciousness. On second thought, it may sound too good to be true that finding the distribution of $g(X)$ is not needed for this calculation, but LOTUS says it *is* true.

Before proving LOTUS in general, let's see why it is true in some special cases. Let X have support $0, 1, 2, \ldots$ with probabilities $p_0, p_1, p_2, \ldots$, so the PMF is $P(X = n) = p_n$. Then X^3 has support $0^3, 1^3, 2^3, \ldots$ with probabilities $p_0, p_1, p_2, \ldots$, so

$$E(X) = \sum_{n=0}^{\infty} np_n,$$

$$E(X^3) = \sum_{n=0}^{\infty} n^3 p_n.$$

As claimed by LOTUS, to edit the expression for $E(X)$ into an expression for $E(X^3)$, we can just change the n in front of the p_n to an n^3. This was an easy example since the function $g(x) = x^3$ is one-to-one. But LOTUS holds much more generally. The key insight needed for the proof of LOTUS for general g is the same as the one we used for the proof of linearity: the expectation of $g(X)$ can be written in ungrouped form as

$$E(g(X)) = \sum_s g(X(s))P(\{s\}),$$

where the sum is over all the pebbles in the sample space, but we can also group the pebbles into super-pebbles according to the value that X assigns to them. Within the super-pebble $X = x$, $g(X)$ always takes on the value $g(x)$. Therefore,

$$\begin{aligned} E(g(X)) &= \sum_s g(X(s))P(\{s\}) \\ &= \sum_x \sum_{s:X(s)=x} g(X(s))P(\{s\}) \\ &= \sum_x g(x) \sum_{s:X(s)=x} P(\{s\}) \\ &= \sum_x g(x)P(X = x). \end{aligned}$$

In the last step, we used the fact that $\sum_{s:X(s)=x} P(\{s\})$ is the weight of the super-pebble $X = x$.

4.6 Variance

One important application of LOTUS is for finding the *variance* of a random variable. Like expected value, variance is a single-number summary of the distribution of a random variable. While the expected value tells us the center of mass of a distribution, the variance tells us how spread out the distribution is.

Definition 4.6.1 (Variance and standard deviation)**.** The *variance* of an r.v. X is

$$\text{Var}(X) = E(X - EX)^2.$$

The square root of the variance is called the *standard deviation* (SD):

$$\text{SD}(X) = \sqrt{\text{Var}(X)}.$$

Recall that when we write $E(X - EX)^2$, we mean the expectation of the random variable $(X - EX)^2$, *not* $(E(X - EX))^2$ (which is 0 by linearity).

The variance of X measures how far X is from its mean on average, but instead of simply taking the average difference between X and its mean EX, we take the average *squared* difference. To see why, note that the average deviation from the mean, $E(X - EX)$, always equals 0 by linearity; positive and negative deviations cancel each other out. By squaring the deviations, we ensure that both positive and negative deviations contribute to the overall variability. However, because variance is an average squared distance, it has the wrong units: if X is in dollars, $\text{Var}(X)$ is in squared dollars. To get back to our original units, we take the square root; this gives us the standard deviation.

One might wonder why variance isn't defined as $E|X-EX|$, which would achieve the goal of counting both positive and negative deviations while maintaining the same units as X. This measure of variability isn't nearly as popular as $E(X - EX)^2$, for a variety of reasons. Most notably, the absolute value function isn't differentiable at 0, whereas the squaring function is differentiable everywhere and is central in various fundamental mathematical results such as the Pythagorean theorem.

An equivalent expression for variance is $\text{Var}(X) = E(X^2) - (EX)^2$. This formula is often easier to work with when doing actual calculations. Since this is the variance formula we will use over and over again, we state it as its own theorem.

Theorem 4.6.2. For any r.v. X,

$$\text{Var}(X) = E(X^2) - (EX)^2.$$

Proof. Let $\mu = EX$. Expanding $(X - \mu)^2$ and using linearity, the variance of X is

$$E(X - \mu)^2 = E(X^2 - 2\mu X + \mu^2) = E(X^2) - 2\mu EX + \mu^2 = E(X^2) - \mu^2. \quad \blacksquare$$

Variance has the following properties. The first two are easily verified from the definition, the third will be addressed in a later chapter, and the last one is proven just after stating it.

- $\text{Var}(X + c) = \text{Var}(X)$ for any constant c. Intuitively, if we shift a distribution to the left or right, that should affect the center of mass of the distribution but not its spread.
- $\text{Var}(cX) = c^2\text{Var}(X)$ for any constant c.
- If X and Y are independent, then $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. We prove this and discuss it more in Chapter 7. This is not true in general if X and Y are dependent. For example, in the extreme case where X always equals Y, we have

$$\text{Var}(X + Y) = \text{Var}(2X) = 4\text{Var}(X) > 2\text{Var}(X) = \text{Var}(X) + \text{Var}(Y)$$

if $\text{Var}(X) > 0$ (which will be true unless X is a constant, as the next property shows).

- $\text{Var}(X) \geq 0$, with equality if and only if $P(X = a) = 1$ for some constant a. In other words, the only random variables that have zero variance are constants (which can be thought of as degenerate r.v.s); all other r.v.s have positive variance.

To prove the last property, note that $\text{Var}(X)$ is the expectation of the *nonnegative* r.v. $(X - EX)^2$, so $\text{Var}(X) \geq 0$. If $P(X = a) = 1$ for some constant a, then $E(X) = a$ and $E(X^2) = a^2$, so $\text{Var}(X) = 0$. Conversely, suppose that $\text{Var}(X) = 0$. Then $E(X - EX)^2 = 0$, which shows that $(X - EX)^2 = 0$ has probability 1, which in turn shows that X equals its mean with probability 1.

☣ **4.6.3** (Variance is not linear)**.** Unlike expectation, variance is *not* linear. The constant comes out *squared* in $\text{Var}(cX) = c^2\text{Var}(X)$, and the variance of the sum of r.v.s may not be the sum of their variances if they are dependent.

Example 4.6.4 (Geometric and Negative Binomial variance)**.** In this example we'll use LOTUS to compute the variance of the Geometric distribution.

Let $X \sim \text{Geom}(p)$. We already know $E(X) = q/p$. By LOTUS,

$$E(X^2) = \sum_{k=0}^{\infty} k^2 P(X = k) = \sum_{k=0}^{\infty} k^2 pq^k = \sum_{k=1}^{\infty} k^2 pq^k.$$

We'll find this using a tactic similar to how we found the expectation, starting from the geometric series

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$

and taking derivatives. After differentiating once with respect to q, we have

$$\sum_{k=1}^{\infty} kq^{k-1} = \frac{1}{(1-q)^2}.$$

We start the sum from $k = 1$ since the $k = 0$ term is 0 anyway. If we differentiate again, we'll get $k(k-1)$ instead of k^2 as we want, so let's replenish our supply of q's by multiplying both sides by q. This gives

$$\sum_{k=1}^{\infty} kq^k = \frac{q}{(1-q)^2}.$$

Now we are ready to take another derivative:

$$\sum_{k=1}^{\infty} k^2 q^{k-1} = \frac{1+q}{(1-q)^3},$$

so

$$E(X^2) = \sum_{k=1}^{\infty} k^2 pq^k = pq\frac{1+q}{(1-q)^3} = \frac{q(1+q)}{p^2}.$$

Finally,

$$\text{Var}(X) = E(X^2) - (EX)^2 = \frac{q(1+q)}{p^2} - \left(\frac{q}{p}\right)^2 = \frac{q}{p^2}.$$

This is also the variance of the First Success distribution, since shifting by a constant does not affect the variance.

Since an NBin(r,p) r.v. can be represented as a sum of r i.i.d. Geom(p) r.v.s by Theorem 4.3.10, and since variance is additive for independent random variables, it follows that the variance of the NBin(r,p) distribution is $r \cdot \frac{q}{p^2}$. □

LOTUS is an all-purpose tool for computing $E(g(X))$ for any g, but as it usually leads to complicated sums, it should be used as a last resort. For variance calculations, our trusty indicator r.v.s can sometimes be used in place of LOTUS, as in the next example.

Example 4.6.5 (Binomial variance)**.** Let's find the variance of $X \sim \text{Bin}(n,p)$ using indicator r.v.s to avoid tedious sums. Represent $X = I_1 + I_2 + \cdots + I_n$, where I_j is the indicator of the jth trial being a success. Each I_j has variance

$$\text{Var}(I_j) = E(I_j^2) - (E(I_j))^2 = p - p^2 = p(1-p).$$

(Recall that $I_j^2 = I_j$, so $E(I_j^2) = E(I_j) = p$.)

Since the I_j are independent, we can add their variances to get the variance of their sum:

$$\text{Var}(X) = \text{Var}(I_1) + \cdots + \text{Var}(I_n) = np(1-p).$$

Alternatively, we can find $E(X^2)$ by first finding $E\binom{X}{2}$. The latter sounds more complicated, but actually it is simpler since $\binom{X}{2}$ is the number of *pairs* of successful trials. Creating an indicator r.v. for each pair of trials, we have

$$E\binom{X}{2} = \binom{n}{2}p^2.$$

Thus,

$$n(n-1)p^2 = E(X(X-1)) = E(X^2) - E(X) = E(X^2) - np,$$

which again gives

$$\text{Var}(X) = E(X^2) - (EX)^2 = (n(n-1)p^2 + np) - (np)^2 = np(1-p).$$

Exercise 48 uses this strategy to find the variance of the Hypergeometric. □

4.7 Poisson

The last discrete distribution that we'll introduce in this chapter is the Poisson, which is an extremely popular distribution for modeling discrete data. We'll introduce its PMF, mean, and variance, and then discuss its story in more detail.

Definition 4.7.1 (Poisson distribution)**.** An r.v. X has the *Poisson distribution* with parameter λ, where $\lambda > 0$, if the PMF of X is

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, 2, \ldots.$$

We write this as $X \sim \text{Pois}(\lambda)$.

This is a valid PMF because of the Taylor series $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$.

Example 4.7.2 (Poisson expectation and variance)**.** Let $X \sim \text{Pois}(\lambda)$. We will show that the mean and variance are both equal to λ. For the mean, we have

$$\begin{aligned} E(X) &= e^{-\lambda} \sum_{k=0}^{\infty} k \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} \\ &= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} e^{\lambda} = \lambda. \end{aligned}$$

First we dropped the $k = 0$ term because it was 0. Then we took a λ out of the sum so that what was left inside was just the Taylor series for e^{λ}.

To get the variance, we first find $E(X^2)$. By LOTUS,

$$E(X^2) = \sum_{k=0}^{\infty} k^2 P(X = k) = e^{-\lambda} \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!}.$$

From here, the derivation is very similar to that of the variance of the Geometric. Differentiate the familiar series

$$\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$$

with respect to λ and replenish:

$$\begin{aligned} \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{k!} &= e^{\lambda}, \\ \sum_{k=1}^{\infty} k \frac{\lambda^k}{k!} &= \lambda e^{\lambda}. \end{aligned}$$

Rinse and repeat:

$$\begin{aligned} \sum_{k=1}^{\infty} k^2 \frac{\lambda^{k-1}}{k!} &= e^{\lambda} + \lambda e^{\lambda} = e^{\lambda}(1 + \lambda), \\ \sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} &= e^{\lambda}\lambda(1 + \lambda). \end{aligned}$$

Finally,

$$E(X^2) = e^{-\lambda}\sum_{k=0}^{\infty} k^2\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda}\lambda(1+\lambda) = \lambda(1+\lambda),$$

so

$$\text{Var}(X) = E(X^2) - (EX)^2 = \lambda(1+\lambda) - \lambda^2 = \lambda.$$

Thus, the mean and variance of a Pois(λ) r.v. are both equal to λ. □

Figure 4.7 shows the PMF and CDF of the Pois(2) and Pois(5) distributions from $k = 0$ to $k = 10$. It appears that the mean of the Pois(2) is around 2 and the mean of the Pois(5) is around 5, consistent with our findings above. The PMF of the Pois(2) is highly skewed, but as λ grows larger, the skewness is reduced and the PMF becomes more bell-shaped.

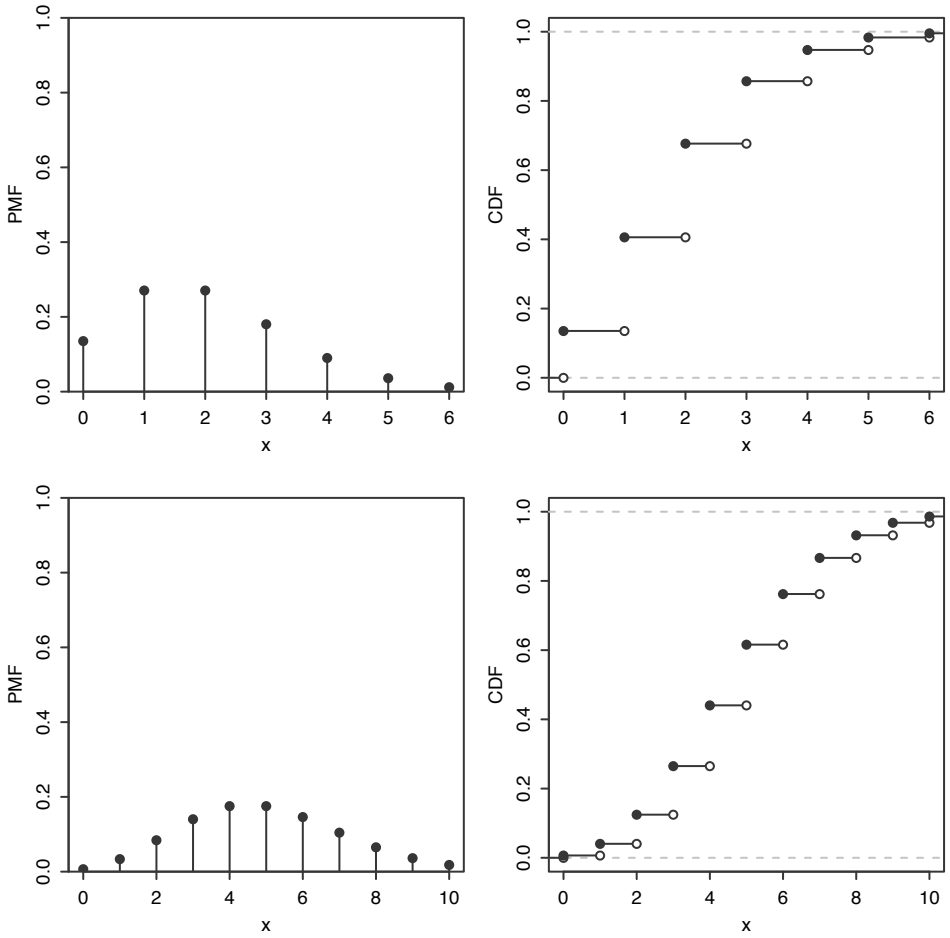


FIGURE 4.7
Top: Pois(2) PMF and CDF. Bottom: Pois(5) PMF and CDF.

The Poisson distribution is often used in situations where we are counting the

number of successes in a particular region or interval of time, and there are a large number of trials, each with a small probability of success. For example, the following random variables could follow a distribution that is approximately Poisson.

- The number of emails you receive in an hour. There are a lot of people who could potentially email you in that hour, but it is unlikely that any specific person will actually email you in that hour. Alternatively, imagine subdividing the hour into milliseconds. There are 3.6×10^6 seconds in an hour, but in any specific millisecond it is unlikely that you will get an email.
- The number of chips in a chocolate chip cookie. Imagine subdividing the cookie into small cubes; the probability of getting a chocolate chip in a single cube is small, but the number of cubes is large.
- The number of earthquakes in a year in some region of the world. At any given time and location, the probability of an earthquake is small, but there are a large number of possible times and locations for earthquakes to occur over the course of the year.

The parameter λ is interpreted as the *rate* of occurrence of these rare events; in the examples above, λ could be 20 (emails per hour), 10 (chips per cookie), and 2 (earthquakes per year). The *Poisson paradigm* says that in applications similar to the ones above, we can approximate the distribution of the number of events that occur by a Poisson distribution.

Approximation 4.7.3 (Poisson paradigm)**.** Let $A_1, \ldots, A_n$ be events with $p_j = P(A_j)$, where n is large, the p_j are small, and the A_j are independent or weakly dependent. Let

$$X = \sum_{j=1}^{n} I(A_j)$$

count how many of the A_j occur. Then X is approximately distributed as Pois(λ), with $\lambda = \sum_{j=1}^{n} p_j$.

Proving that the above approximation is good is difficult, and would require first giving precise definitions of weak dependence (there are various ways to measure dependence of r.v.s) and of good approximations (there are various ways to measure how good an approximation is). A remarkable theorem is that if the A_j are independent, $N \sim \text{Pois}(\lambda)$, and B is any set of nonnegative integers, then

$$|P(X \in B) - P(N \in B)| \leq \min\left(1, \frac{1}{\lambda}\right) \sum_{j=1}^{n} p_j^2.$$

This gives an upper bound on how much error is incurred from using a Poisson approximation. It also makes more precise how small the p_j should be: we want $\sum_{j=1}^{n} p_j^2$ to be very small, or at least very small compared to λ. The result can be shown using an advanced technique known as the *Stein-Chen method*.

The Poisson paradigm is also called the *law of rare events*. The interpretation of "rare" is that the p_j are small, not that λ is small. For example, in the email example, the low probability of getting an email from a specific person in a particular hour is offset by the large number of people who could send you an email in that hour.

In the examples we gave above, the number of events that occur isn't *exactly* Poisson because a Poisson random variable has no upper bound, whereas how many of $A_1, \ldots, A_n$ occur is at most n, and there is a limit to how many chocolate chips can be crammed into a cookie. But the Poisson distribution often gives good *approximations*. Note that the conditions for the Poisson paradigm to hold are fairly flexible: the n trials can have different success probabilities, and the trials don't have to be independent, though they should not be very dependent. So there are a wide variety of situations that can be cast in terms of the Poisson paradigm. This makes the Poisson a popular model, or at least a starting point, for data whose values are nonnegative integers (called *count data* in statistics).

Example 4.7.4 (Balls in boxes)**.** There are k distinguishable balls and n distinguishable boxes. The balls are randomly placed in the boxes, with all n^k possibilities equally likely. Problems in this setting are called *occupancy problems*, and are at the core of many widely used algorithms in computer science.

(a) Find the expected number of empty boxes (fully simplified, *not* as a sum).

(b) Find the probability that at least one box is empty. Express your answer as a sum of at most n terms.

(c) Now let $n = 1000$, $k = 5806$. The expected number of empty boxes is then approximately 3. Find a good approximation as a decimal for the probability that at least one box is empty. The handy fact $e^3 \approx 20$ may help.

Solution:

(a) Let I_j be the indicator r.v. for the jth box being empty. Then

$$E(I_j) = P(I_j = 1) = \left(1 - \frac{1}{n}\right)^k.$$

By linearity,

$$E\left(\sum_{j=1}^{n} I_j\right) = \sum_{j=1}^{n} E(I_j) = n\left(1 - \frac{1}{n}\right)^k.$$

(b) The probability is 1 for $k < n$. In general, let A_j be the event that box j is empty. By inclusion-exclusion,

$$\begin{aligned} P(A_1 \cup A_2 \cup \cdots \cup A_n) &= \sum_{j=1}^{n} (-1)^{j+1} \binom{n}{j} P(A_1 \cap A_2 \cap \cdots \cap A_j) \\ &= \sum_{j=1}^{n-1} (-1)^{j+1} \binom{n}{j} \left(1 - \frac{j}{n}\right)^k. \end{aligned}$$

(c) The number X of empty boxes is approximately Pois(3), since there are a lot of boxes but each is very unlikely to be empty; the probability that a specific box is empty is $(1-\frac{1}{n})^k = \frac{1}{n} \cdot E(X) \approx 0.003$. So

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-3} \approx 1 - \frac{1}{20} = 0.95. \qquad \square$$

Poisson approximation greatly simplifies obtaining a good approximate solution to the birthday problem discussed in Chapter 1, and makes it possible to obtain good approximations to various variations which would be hard to solve exactly.

Example 4.7.5 (Birthday problem continued)**.** If we have m people and make the usual assumptions about birthdays, then each pair of people has probability $p = 1/365$ of having the same birthday, and there are $\binom{m}{2}$ pairs. By the Poisson paradigm the distribution of the number X of birthday matches is approximately Pois(λ), where $\lambda = \binom{m}{2}\frac{1}{365}$. Then the probability of at least one match is

$$P(X \geq 1) = 1 - P(X = 0) \approx 1 - e^{-\lambda}.$$

For $m = 23$, $\lambda = 253/365$ and $1-e^{-\lambda} \approx 0.500002$, which agrees with our finding from Chapter 1 that we need 23 people to have a 50-50 chance of a matching birthday.

Note that even though $m = 23$ is fairly small, the relevant quantity in this problem is actually $\binom{m}{2}$, which is the total number of "trials" for a successful birthday match, so the Poisson approximation still performs well. $\square$

Example 4.7.6 (Near-birthday problem)**.** What if we want to find the number of people required in order to have a 50-50 chance that two people would have birthdays within one day of each other (i.e., on the same day or one day apart)? Unlike the original birthday problem, this is difficult to obtain an exact answer for, but the Poisson paradigm still applies. The probability that any two people have birthdays within one day of each other is 3/365 (choose a birthday for the first person, and then the second person needs to be born on that day, the day before, or the day after). Again there are $\binom{m}{2}$ possible pairs, so the number of within-one-day matches is approximately Pois(λ) where $\lambda = \binom{m}{2}\frac{3}{365}$. Then a calculation similar to the one above tells us that we need $m = 14$ or more. This was a quick approximation, but it turns out that $m = 14$ is the exact answer! $\square$

Example 4.7.7 (Birth-minute and birth-hour)**.** There are 1600 sophomores at a certain college. Throughout this example, make the usual assumptions as in the birthday problem.

(a) Find a Poisson approximation for the probability that there are two sophomores who were born not only on the same day of the year, but also at the same hour *and* the same minute (e.g., both sophomores were born at 8:20 pm on March 31, not necessarily in the same year).

(b) With assumptions as in (a), what is the probability that there are *four* sophomores who were born not only on the same day, but also at the same hour (e.g., all were born between 2 pm and 3 pm on March 31, not necessarily in the same year)?

Give two different Poisson approximations for this value, one based on creating an indicator r.v. for each quadruplet of sophomores, and the other based on creating an indicator r.v. for each possible day-hour. Which do you think is more accurate?

Solution:

(a) This is the birthday problem, with $c = 365 \cdot 24 \cdot 60 = 525600$ categories rather than 365 categories.[1] Let $n = 1600$. Creating an indicator r.v. for each pair of sophomores, by linearity the expected number of pairs born on the same day-hour-minute is

$$\lambda_1 = \binom{n}{2}\frac{1}{c}.$$

By Poisson approximation, the probability of at least one match is approximately

$$1 - \exp(-\lambda_1) \approx 0.9122.$$

This approximation is very accurate: typing `pbirthday(1600, classes=365*24*60)` in R yields 0.9125.

(b) Now there are $b = 365 \cdot 24 = 8760$ categories. Let's explore two different methods of Poisson approximation.

Method 1: Create an indicator for each set of 4 sophomores. By linearity, the expected number of sets of 4 sophomores born on the same day-hour is

$$\lambda_2 = \binom{n}{4}\frac{1}{b^3}.$$

Poisson approximation gives that the desired probability is approximately

$$1 - \exp(-\lambda_2) \approx 0.333.$$

Method 2: Create an indicator for each possible day-hour. Let I_j be the indicator for at least 4 people having been born on the jth day-hour of the year (ordered chronologically), for $1 \leq j \leq b$. Let $p = 1/b$ and $q = 1 - p$. Then

$$\begin{aligned} E(I_j) &= P(I_j = 1) \\ &= 1 - P(\text{at most 3 people born on the } j\text{th day-hour}) \\ &= 1 - q^n - npq^{n-1} - \binom{n}{2}p^2q^{n-2} - \binom{n}{3}p^3q^{n-3}. \end{aligned}$$

The expected number of day-hours on which at least 4 sophomores were born is

$$\lambda_3 = b \cdot E(I_1),$$

with $E(I_1)$ as above. We then have the Poisson approximation

$$1 - \exp(-\lambda_3) \approx 0.295.$$

[1] The song "Seasons of Love" from *Rent* gives a musical interpretation of this fact.

The command `pbirthday(1600, classes = 8760, coincident=4)` in R gives that the correct answer is 0.296. So Method 2 is more accurate than Method 1.

An intuitive explanation for why Method 1 is less accurate is that there is a more substantial dependence in the indicators in that method. For example, being given that sophomores $1, 2, 3, 4$ share the same birth day-hour greatly increases the chance that sophomores $1, 2, 3, 5$ share the same birth day-hour. In contrast, knowing that at least 4 sophomores were born on a specific day-hour provides very little information about whether at least 4 were born on a different specific day-hour. □

4.8 Connections between Poisson and Binomial

The Poisson and Binomial distributions are closely connected, and their relationship is exactly parallel to the relationship between the Binomial and Hypergeometric distributions that we examined in the previous chapter: we can get from the Poisson to the Binomial by *conditioning*, and we can get from the Binomial to the Poisson by *taking a limit*.

Our results will rely on the fact that the sum of independent Poissons is Poisson, just as the sum of independent Binomials is Binomial. We'll prove this result using the law of total probability for now; in Chapter 6 we'll learn a faster method that uses a tool called the moment generating function. Chapter 13 gives further insight into these results.

Theorem 4.8.1 (Sum of independent Poissons)**.** If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

Proof. To get the PMF of $X+Y$, condition on X and use the law of total probability:

$$\begin{aligned} P(X+Y=k) &= \sum_{j=0}^{k} P(X+Y=k|X=j)P(X=j) \\ &= \sum_{j=0}^{k} P(Y=k-j)P(X=j) \\ &= \sum_{j=0}^{k} \frac{e^{-\lambda_2}\lambda_2^{k-j}}{(k-j)!}\frac{e^{-\lambda_1}\lambda_1^{j}}{j!} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}}{k!}\sum_{j=0}^{k}\binom{k}{j}\lambda_1^{j}\lambda_2^{k-j} \\ &= \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^{k}}{k!}. \end{aligned}$$

The last step used the binomial theorem. Since we've arrived at the Pois$(\lambda_1 + \lambda_2)$ PMF, we have $X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$.

The story of the Poisson distribution provides intuition for this result. If there are two different types of events occurring at rates λ_1 and λ_2, independently, then the overall event rate is $\lambda_1 + \lambda_2$. ■

Theorem 4.8.2 (Poisson given a sum of Poissons)**.** If $X \sim \text{Pois}(\lambda_1)$, $Y \sim \text{Pois}(\lambda_2)$, and X is independent of Y, then the conditional distribution of X given $X + Y = n$ is Bin $(n, \lambda_1/(\lambda_1 + \lambda_2))$.

Proof. Exactly as in the corresponding proof for the Binomial and Hypergeometric, we use Bayes' rule to compute the conditional PMF $P(X = k|X + Y = n)$:

$$\begin{aligned} P(X = k|X + Y = n) &= \frac{P(X + Y = n|X = k)P(X = k)}{P(X + Y = n)} \\ &= \frac{P(Y = n - k)P(X = k)}{P(X + Y = n)}. \end{aligned}$$

Now we plug in the PMFs of X, Y, and $X + Y$; the last of these is distributed Pois$(\lambda_1 + \lambda_2)$ by the previous theorem. This gives

$$\begin{aligned} P(X = k|X + Y = n) &= \frac{\left(\frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!}\right)\left(\frac{e^{-\lambda_1}\lambda_1^{k}}{k!}\right)}{\frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}} \\ &= \binom{n}{k}\frac{\lambda_1^k\lambda_2^{n-k}}{(\lambda_1+\lambda_2)^n} \\ &= \binom{n}{k}\left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^k\left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{n-k}, \end{aligned}$$

which is the Bin$(n, \lambda_1/(\lambda_1 + \lambda_2))$ PMF, as desired. ■

Conversely, if we take the limit of the Bin(n, p) distribution as $n \to \infty$ and $p \to 0$ with np fixed, we arrive at a Poisson distribution. This provides the basis for the *Poisson approximation to the Binomial distribution*.

Theorem 4.8.3 (Poisson approximation to Binomial)**.** If $X \sim \text{Bin}(n, p)$ and we let $n \to \infty$ and $p \to 0$ such that $\lambda = np$ remains fixed, then the PMF of X converges to the Pois(λ) PMF. More generally, the same conclusion holds if $n \to \infty$ and $p \to 0$ in such a way that np converges to a constant λ.

This is a special case of the Poisson paradigm, where the A_j are independent with the same probabilities, so that $\sum_{j=1}^{n} I(A_j)$ has a Binomial distribution. In this special case, we can prove that the Poisson approximation makes sense just by taking a limit of the Binomial PMF.

Proof. We will prove this for the case that $\lambda = np$ is fixed while $n \to \infty$ and $p \to 0$, by showing that the Bin(n,p) PMF converges to the Pois(λ) PMF. For $0 \leq k \leq n$,

$$\begin{aligned} P(X=k) &= \binom{n}{k} p^k (1-p)^{n-k} \\ &= \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k} \\ &= \frac{\lambda^k}{k!} \frac{n(n-1)\dots(n-k+1)}{n^k} \left(1-\frac{\lambda}{n}\right)^n \left(1-\frac{\lambda}{n}\right)^{-k}. \end{aligned}$$

Letting $n \to \infty$ with k fixed,

$$\begin{aligned} \frac{n(n-1)\dots(n-k+1)}{n^k} &\to 1, \\ \left(1-\frac{\lambda}{n}\right)^n &\to e^{-\lambda}, \\ \left(1-\frac{\lambda}{n}\right)^{-k} &\to 1, \end{aligned}$$

where the $e^{-\lambda}$ comes from the compound interest formula from Section A.2.5 of the math appendix. So

$$P(X=k) \to \frac{e^{-\lambda}\lambda^k}{k!},$$

which is the Pois(λ) PMF. ■

This theorem implies that if n is large, p is small, and np is moderate, we can approximate the Bin(n,p) PMF by the Pois(np) PMF. The main thing that matters here is that p should be small; in fact, the result mentioned after the statement of the Poisson paradigm says in this case that the error in approximating $P(X \in B) \approx P(N \in B)$ for $X \sim \text{Bin}(n,p), N \sim \text{Pois}(np)$ is at most $\min(p, np^2)$.

Example 4.8.4 (Visitors to a website)**.** The owner of a certain website is studying the distribution of the number of visitors to the site. Every day, a million people independently decide whether to visit the site, with probability $p = 2 \times 10^{-6}$ of visiting. Give a good approximation for the probability of getting *at least three* visitors on a particular day.

Solution:

Let $X \sim \text{Bin}(n,p)$ be the number of visitors, where $n = 10^6$. It is easy to run into computational difficulties or numerical errors in exact calculations with this distribution since n is so large and p is so small. But since n is large, p is small, and $np = 2$ is moderate, Pois(2) is a good approximation. This gives

$$P(X \geq 3) = 1 - P(X < 3) \approx 1 - e^{-2} - e^{-2} \cdot 2 - e^{-2} \cdot \frac{2^2}{2!} = 1 - 5e^{-2} \approx 0.3233,$$

which turns out to be extremely accurate. □

4.9 *Using probability and expectation to prove existence

An amazing and beautiful fact is that we can use probability and expectation to prove the *existence* of objects with properties we care about. This technique is called the *probabilistic method*, and it is based on two simple but surprisingly powerful ideas. Suppose I want to show that there exists an object in a collection with a certain property. This desire seems at first to have nothing to do with probability; I could simply examine each object in the collection one by one until finding an object with the desired property.

The probabilistic method rejects such painstaking inspection in favor of random selection: our strategy is to pick an object *at random* from the collection and show that there is a positive probability of the random object having the desired property. Note that we are not required to compute the exact probability, but merely to show it is greater than 0. If we can show that the probability of the property holding is positive, then we know that there must exist an object with the property—even if we don't know how to explicitly construct such an object.

Similarly, suppose each object has a score, and I want to show that there exists an object with a "good" score—that is, a score exceeding a particular threshold. Again, we proceed by choosing a *random object* and considering its score, X. We know there is an object in the collection whose score is at least $E(X)$—it's impossible for every object to be below average! If $E(X)$ is already a good score, then there must also be an object in the collection with a good score. Thus we can show the existence of an object with a good score by showing that the average score is already good.

Let's state the two key ideas formally.

- The possibility principle: Let A be the event that a randomly chosen object in a collection has a certain property. If $P(A) > 0$, then there exists an object with the property.
- The good score principle: Let X be the score of a randomly chosen object. If $E(X) \geq c$, then there is an object with a score of at least c.

To see why the possibility principle is true, consider its contrapositive: if there is no object with the desired property, then the probability of a randomly chosen object having the property is 0. Similarly, the contrapositive of the good score principle is "if all objects have a score below c, then the average score is below c", which is true since a weighted average of numbers less than c is a number less than c.

The probabilistic method doesn't tell us *how* to find an object with the desired property; it only assures us that one exists.

Example 4.9.1. A group of 100 people are assigned to 15 committees of size 20, such that each person serves on 3 committees. Show that there exist 2 committees that have at least 3 people in common.

Solution:

A direct approach is inadvisable here: one would have to list all possible committee assignments and compute, for each one, the number of people in common in every pair of committees. The probabilistic method lets us bypass brute-force calculations. To prove the existence of two committees with an overlap of at least three people, we'll calculate the *average* overlap of two *randomly chosen* committees in an arbitrary committee assignment. So choose two committees at random, and let X be the number of people on both committees. We can represent $X = I_1 + I_2 + \cdots + I_{100}$, where $I_j = 1$ if the jth person is on both committees and 0 otherwise. By symmetry, all of the indicators have the same expected value, so $E(X) = 100E(I_1)$, and we just need to find $E(I_1)$.

By the fundamental bridge, $E(I_1)$ is the probability that person 1 (whom we'll name Bob) is on both committees (which we'll call A and B). There are a variety of ways to calculate this probability; one way is to think of Bob's committees as 3 tagged elk in a population of 15. Then A and B are a sample of 2 elk, made without replacement. Using the HGeom$(3, 12, 2)$ PMF, the probability that both of these elk are tagged (i.e., the probability that both committees contain Bob) is $\binom{3}{2}\binom{12}{0}/\binom{15}{2} = 1/35$. Therefore,

$$E(X) = 100/35 = 20/7,$$

which is just shy of the desired "good score" of 3. But hope is not lost! The good score principle says there exist two committees with an overlap of at least $20/7$, but since the overlap between two committees must be an integer, an overlap of at least $20/7$ implies an overlap of at least 3. Thus, there exist two committees with at least 3 people in common. □

4.9.1 *Communicating over a noisy channel

Another major application of the probabilistic method is in *information theory*, the subject which studies (among other things) how to achieve reliable communication across a noisy channel. Consider the problem of trying to send a message when there is noise. This problem is encountered by millions of people every day, such as when talking on the phone (you may be misheard). Suppose that the message you want to send is represented as a binary vector $x \in \{0, 1\}^k$, and that you want to use a *code* to improve the chance that your message will get through successfully.

Definition 4.9.2 (Codes and rates)**.** Given positive integers k and n, a *code* is a function c that assigns to each input message $x \in \{0, 1\}^k$ a *codeword* $c(x) \in \{0, 1\}^n$. The *rate* of this code is k/n (the number of input bits per output bit). After $c(x)$ is sent, a *decoder* takes the received message, which may be a corrupted version of $c(x)$, and attempts to recover the correct x.

For example, an obvious code would be to repeat yourself a bunch of times, sending x a bunch of times in a row, say m (with m odd); this is called a *repetition code*. The

receiver could then *decode* by going with the majority, e.g., decoding the first bit of x as a 1 if that bit was received more times as a 1 than as a 0. But this code may be very inefficient; to get the probability of failure very small, you may need to repeat yourself many times, resulting in a very low rate $1/m$ of communication.

Claude Shannon, the founder of information theory, showed something amazing: even in a very noisy channel, there is a code allowing for very reliable communication at a rate that does not go to 0 as we require the probability of failure to be lower and lower. His proof was even more amazing: he studied the performance of a completely *random* code. Richard Hamming, who worked with Shannon at Bell Labs, described Shannon's approach as follows.

> Courage is another attribute of those who do great things. Shannon is a good example. For some time he would come to work at about 10:00 am, play chess until about 2:00 pm and go home.
>
> The important point is how he played chess. When attacked he seldom, if ever, defended his position, rather he attacked back. Such a method of playing soon produces a very interrelated board. He would then pause a bit, think and advance his queen saying, "I ain't [scared] of nothin'." It took me a while to realize that of course that is why he was able to prove the existence of good coding methods. Who but Shannon would think to average over all random codes and expect to find that the average was close to ideal? I learned from him to say the same to myself when stuck, and on some occasions his approach enabled me to get significant results. [15]

We will prove a version of Shannon's result, for the case of a channel where each transmitted bit gets flipped (from 0 to 1 or from 1 to 0) with probability p, independently. First we need two definitions. A natural measure of distance between binary vectors, named after Hamming, is as follows.

Definition 4.9.3 (Hamming distance)**.** For two binary vectors v and w of the same length, the *Hamming distance* $d(v, w)$ is the number of positions in which they differ. We can write this as

$$d(v, w) = \sum_i |v_i - w_i|.$$

The following function arises very frequently in information theory.

Definition 4.9.4 (Binary entropy function)**.** For $0 < p < 1$, the *binary entropy function* H is given by

$$H(p) = -p \log_2 p - (1 - p) \log_2 (1 - p).$$

We also define $H(0) = H(1) = 0$.

The interpretation of $H(p)$ in information theory is that it is a measure of how much information we get from observing a Bern(p) r.v.; $H(1/2) = 1$ says that a fair coin flip provides 1 bit of information, while $H(1) = 0$ says that with a coin that

always lands Heads, there's no information gained from being told the result of the flip, since we already know the result.

Now consider a channel where each transmitted bit gets flipped with probability p, independently. Intuitively, it may seem that smaller p is always better, but note that $p = 1/2$ is actually the worst-case scenario. In that case, technically known as a *useless channel*, it is impossible to send information over the channel: the output will be independent of the input! Analogously, in deciding whether to watch a movie, would you rather hear a review from someone you always disagree with or someone you agree with half the time? We now prove that for $0 < p < 1/2$, it is possible to communicate very reliably with rate very close to $1 - H(p)$.

Theorem 4.9.5 (Shannon)**.** Consider a channel where each transmitted bit gets flipped with probability p, independently. Let $0 < p < 1/2$ and $\epsilon > 0$. There exists a code with rate at least $1 - H(p) - \epsilon$ that can be decoded with probability of error less than ϵ.

Proof. We can assume that $1 - H(p) - \epsilon > 0$, since otherwise there is no constraint on the rate. Let n be a large positive integer (chosen according to conditions given below), and

$$k = \lceil n(1 - H(p) - \epsilon) \rceil + 1.$$

The ceiling function is there since k must be an integer. Choose $p' \in (p, 1/2)$ such that $|H(p') - H(p)| < \epsilon/2$ (this can be done since H is continuous). We will now study the performance of a *random* code C. To generate a random code C, we need to generate a random encoded message $C(x)$ for all possible input messages x.

For each $x \in \{0,1\}^k$, choose $C(x)$ to be a uniformly random vector in $\{0,1\}^n$ (making these choices independently). So we can think of $C(x)$ as a vector consisting of n i.i.d. Bern(1/2) r.v.s. The rate k/n exceeds $1 - H(p) - \epsilon$ by definition, but let's see how well we can decode the received message!

Let $x \in \{0,1\}^k$ be the input message, $C(x)$ be the encoded message, and $Y \in \{0,1\}^n$ be the received message. For now, treat x as deterministic. But $C(x)$ is random since the codewords are chosen randomly, and Y is random since $C(x)$ is random and due to the random noise in the channel. Intuitively, we hope that $C(x)$ will be close to Y (in Hamming distance) and $C(z)$ will be far from Y for all $z \neq x$, in which case it will be clear how to decode Y and the decoding will succeed. To make this precise, decode Y as follows:

If there exists a unique $z \in \{0,1\}^k$ such that $d(C(z), Y) \leq np'$, decode Y to that z; otherwise, declare decoder failure.

We will show that for n large enough, the probability of the decoder failing to recover the correct x is less than ϵ. There are two things that could go wrong:

(a) $d(C(x), Y) > np'$, or

(b) There could be some impostor $z \neq x$ with $d(C(z), Y) \leq np'$.

Note that $d(C(x), Y)$ is an r.v., so $d(C(x), Y) > np'$ is an event. To handle (a), represent

$$d(C(x), Y) = B_1 + \cdots + B_n \sim \text{Bin}(n, p),$$

where B_i is the indicator of the ith bit being flipped. The law of large numbers (see Chapter 10) says that as n grows, the r.v. $d(C(x), Y)/n$ will get very close to p (its expected value), and so will be very unlikely to exceed p':

$$P(d(C(x), Y) > np') = P\left(\frac{B_1 + \cdots + B_n}{n} > p'\right) \to 0 \text{ as } n \to \infty.$$

So by choosing n large enough, we can make

$$P(d(C(x), Y) > np') < \epsilon/4.$$

To handle (b), note that $d(C(z), Y) \sim \text{Bin}(n, 1/2)$ for $z \neq x$, since the n bits in $C(z)$ are i.i.d. Bern(1/2), independent of Y (to show this in more detail, condition on Y using LOTP). Let $B \sim \text{Bin}(n, 1/2)$. By Boole's inequality,

$$P(d(C(z), Y) \leq np' \text{ for some } z \neq x) \leq (2^k - 1)P(B \leq np').$$

To simplify notation, suppose that np' is an integer. A crude way to upper bound a sum of m terms is to use m times the largest term, and a crude way to upper bound a binomial coefficient $\binom{n}{j}$ is to use $r^{-j}(1-r)^{-(n-j)}$ for any $r \in (0, 1)$. Combining these two crudities,

$$P(B \leq np') = \frac{1}{2^n}\sum_{j=0}^{np'}\binom{n}{j} \leq \frac{np'+1}{2^n}\binom{n}{np'} \leq (np'+1)2^{nH(p')-n},$$

using the fact that $(p')^{-np'}(q')^{-nq'} = 2^{nH(p')}$ for $q' = 1 - p'$. Thus,

$$2^k P(B \leq np') \leq (np'+1)2^{n(1-H(p)-\epsilon)+2+n(H(p)+\epsilon/2)-n} = 4(np'+1)2^{-n\epsilon/2} \to 0,$$

so we can choose n to make $P(d(C(z), Y) \leq np' \text{ for some } z \neq x) < \epsilon/4$.

Assume that k and n have been chosen in accordance with the above, and let $F(c, x)$ be the event of failure when code c is used with input message x. Putting together the above results, we have shown that for a *random* C and any *fixed* x,

$$P(F(C, x)) < \epsilon/2.$$

It follows that for each x, there is a code c with $P(F(c, x)) < \epsilon/2$, but this is not good enough: we want *one* code that works well for *all* x! Let X be a uniformly random input message in $\{0, 1\}^k$, independent of C. By LOTP, we have

$$P(F(C, X)) = \sum_x P(F(C, x))P(X = x) < \epsilon/2.$$

Again using LOTP, but this time conditioning on C, we have

$$\sum_c P(F(c,X))P(C=c) = P(F(C,X)) < \epsilon/2.$$

Therefore, there exists a code c such that $P(F(c,X)) < \epsilon/2$, i.e., a code c such that the probability of failure for a random input message X is less than $\epsilon/2$. Lastly, we will improve c, obtaining a code that works well for *all* x, not just a random x. We do this by *expurgating* the worst 50% of the x's. That is, remove as legal input messages the 2^{k-1} values of x with the highest failure probabilities for code c. For all remaining x, we have $P(F(c,x)) < \epsilon$, since otherwise more than half of the $x \in \{0,1\}^k$ would have more than double the average failure probability (see Markov's inequality in Chapter 10 for more about this kind of argument). By relabeling the remaining x using vectors in $\{0,1\}^{k-1}$, we obtain a code $c' : \{0,1\}^{k-1} \to \{0,1\}^n$ with rate $(k-1)/n \geq 1 - H(p) - \epsilon$ and probability less than ϵ of failure for all input messages in $\{0,1\}^{k-1}$. ■

There is also a converse to the above theorem, showing that if we require the rate to be at least $1 - H(p) + \epsilon$, it is impossible to find codes that make the probability of error arbitrarily small. This is why $1 - H(p)$ is called the *capacity* of the channel. Shannon also obtained analogous results for much more general channels. These results give theoretical bounds on what can be achieved, without saying explicitly which codes to use. Decades of subsequent work have been devoted to developing specific codes that work well in practice, by coming close to the Shannon bound and allowing for efficient encoding and decoding.

4.10 Recap

The expectation of a discrete r.v. X is

$$E(X) = \sum_x xP(X=x).$$

An equivalent "ungrouped" way of calculating expectation is

$$E(X) = \sum_s X(s)P(\{s\}),$$

where the sum is taken over pebbles in the sample space. Expectation is a single number summarizing the center of mass of a distribution. A single-number summary of the spread of a distribution is the variance, defined by

$$\text{Var}(X) = E(X - EX)^2 = E(X^2) - (EX)^2.$$

The square root of the variance is called the standard deviation.

Expectation is linear:

$$E(cX) = cE(X) \text{ and } E(X+Y) = E(X) + E(Y),$$

regardless of whether X and Y are independent or not. Variance is *not* linear:

$$\text{Var}(cX) = c^2\text{Var}(X),$$

and

$$\text{Var}(X+Y) \neq \text{Var}(X) + \text{Var}(Y)$$

in general (an important exception is when X and Y are independent).

A very important strategy for calculating the expectation of a discrete r.v. X is to express it as a sum of *indicator r.v.s*, and then apply linearity and the fundamental bridge. This technique is especially powerful because the indicator r.v.s need not be independent; linearity holds even for dependent r.v.s. The strategy can be summarized in the following three steps.

1. Represent the r.v. X as a sum of indicator r.v.s. To decide how to define the indicators, think about what X is counting. For example, if X is the number of local maxima, as in the Putnam problem, then we should create an indicator for each local maximum that could occur.
2. Use the fundamental bridge to calculate the expected value of each indicator. When applicable, symmetry may be very helpful at this stage.
3. By linearity of expectation, $E(X)$ can be obtained by adding up the expectations of the indicators.

Another tool for computing expectations is LOTUS, which says we can calculate the expectation of $g(X)$ using only the PMF of X, via

$$E(g(X)) = \sum_x g(x)P(X = x).$$

If g is non-linear, it is a grave mistake to attempt to calculate $E(g(X))$ by swapping the E and the g.

Four new discrete distributions to add to our list are the Geometric, Negative Binomial, Negative Hypergeometric, and Poisson distributions. A Geom(p) r.v. is the number of failures before the first success in a sequence of independent Bernoulli trials with probability p of success, and an NBin(r, p) r.v. is the number of failures before r successes. The Negative Hypergeometric is similar to the Negative Binomial except, in terms of drawing balls from an urn, the Negative Hypergeometric samples *without* replacement and the Negative Binomial samples *with* replacement. (We also introduced the First Success distribution, which is just a Geometric shifted so that the success is included.)

A Poisson r.v. is often used as an approximation for the number of successes that

occur when there are many independent or weakly dependent trials, where each trial has a small probability of success. In the Binomial story, all the trials have the same probability p of success, but in the Poisson approximation, different trials can have different (but small) probabilities p_j of success.

The Poisson, Binomial, and Hypergeometric distributions are mutually connected via the operations of conditioning and taking limits, as illustrated in Figure 4.8. In the rest of this book, we'll continue to introduce new named distributions and add them to this family tree, until everything is connected!

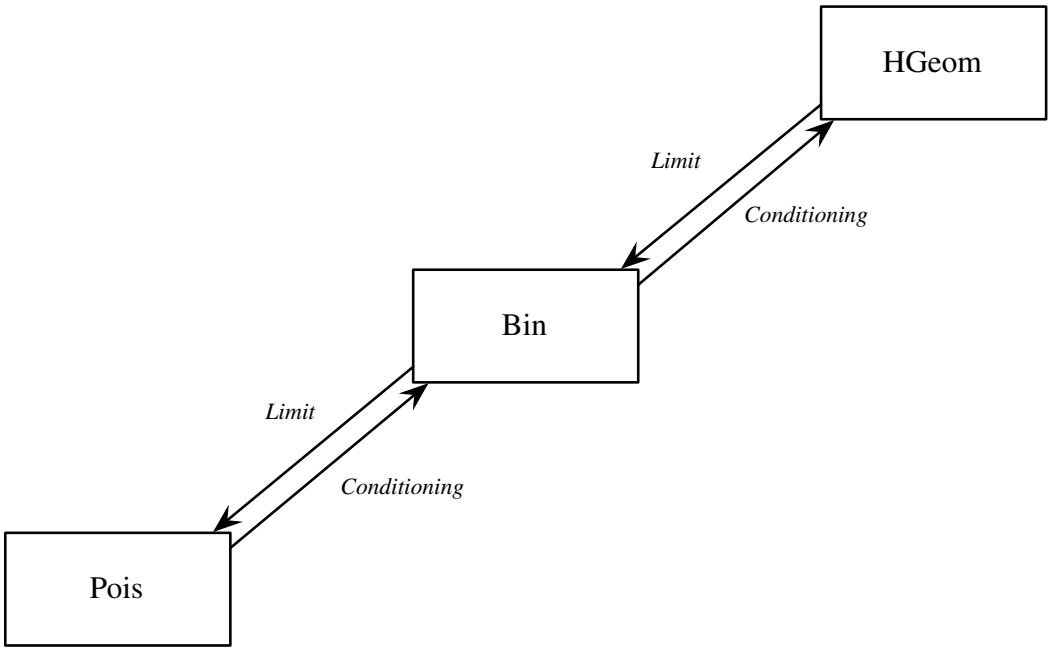


FIGURE 4.8
Relationships between the Poisson, Binomial, and Hypergeometric.

Figure 4.9 expands upon the corresponding figure from the previous chapter, further exploring the connections between the four fundamental objects we have considered: distributions, random variables, events, and numbers.

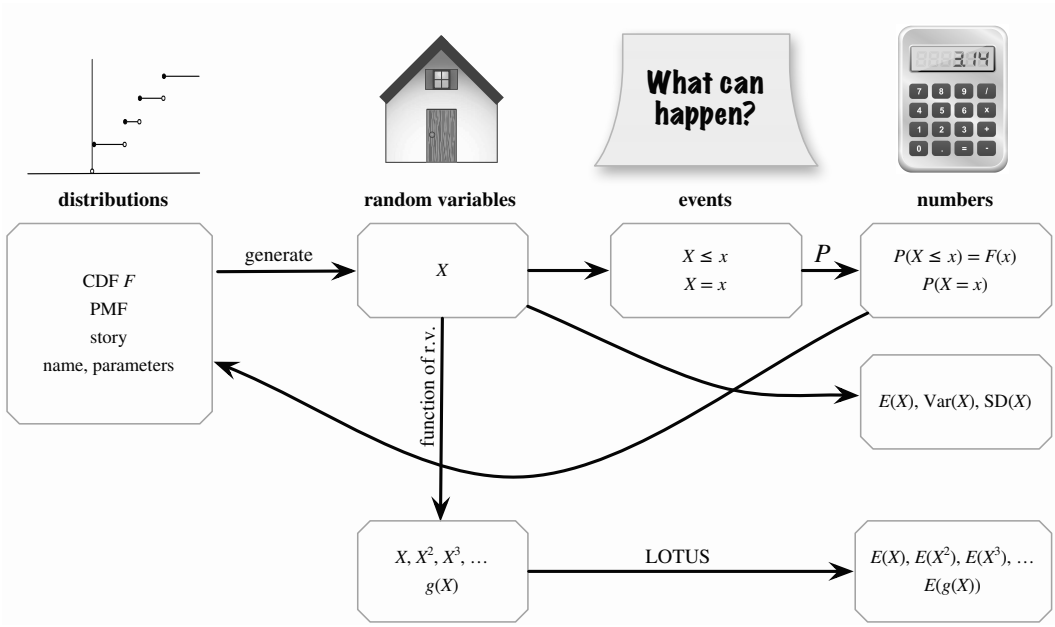


FIGURE 4.9
Four fundamental objects in probability: distributions, random variables, events, and numbers. From an r.v. X, we can generate many other r.v.s by taking functions of X, and we can use LOTUS to find their expected values. The mean, variance, and standard deviation of X express the average and spread of the distribution of X (in particular, they only depend on F, not directly on X itself).

4.11 R

Geometric, Negative Binomial, and Poisson

The three functions for the Geometric distribution in R are `dgeom`, `pgeom`, and `rgeom`, corresponding to the PMF, CDF, and random generation. For `dgeom` and `pgeom`, we need to supply the following as inputs: (1) the value at which to evaluate the PMF or CDF, and (2) the parameter p. For `rgeom`, we need to input (1) the number of random variables to generate and (2) the parameter p.

For example, to calculate $P(X = 3)$ and $P(X \leq 3)$ where $X \sim \text{Geom}(0.5)$, we use `dgeom(3,0.5)` and `pgeom(3,0.5)`, respectively. To generate 100 i.i.d. Geom(0.8) r.v.s, we use `rgeom(100,0.8)`. If instead we want 100 i.i.d. FS(0.8) r.v.s, we just need to add 1 to include the success: `rgeom(100,0.8)+1`.

For the Negative Binomial distribution, we have `dnbinom`, `pnbinom`, and `rnbinom`.

These take three inputs. For example, to calculate the NBin(5, 0.5) PMF at 3, we type `dnbinom(3,5,0.5)`.

Finally, for the Poisson distribution, the three functions are `dpois`, `ppois`, and `rpois`. These take two inputs. For example, to find the Pois(10) CDF at 2, we type `ppois(2,10)`.

Matching simulation

Continuing with Example 4.4.4, let's use simulation to calculate the expected number of matches in a deck of cards. As in Chapter 1, we let `n` be the number of cards in the deck and perform the experiment 10^4 times using `replicate`.

```
n <- 100
r <- replicate(10^4,sum(sample(n)==(1:n)))
```

Now `r` contains the number of matches from each of the 10^4 simulations. But instead of looking at the probability of at least one match, as in Chapter 1, we now want to find the expected number of matches. We can approximate this by the average of all the simulation results, that is, the arithmetic mean of the elements of `r`. This is accomplished with the `mean` function:

```
mean(r)
```

The command `mean(r)` is equivalent to `sum(r)/length(r)`. The result we get is very close to 1, confirming the calculation we did in Example 4.4.4 using indicator r.v.s. You can verify that no matter what value of `n` you choose, `mean(r)` will be very close to 1.

Distinct birthdays simulation

Let's calculate the expected number of distinct birthdays in a group of k people by simulation. We'll let $k = 20$, but you can choose whatever value of k you like.

```
k <- 20
r <- replicate(10^4,{bdays <- sample(365,k,replace=TRUE);
                     length(unique(bdays))})
```

In the second line, `replicate` repeats the expression in the curly braces 10^4 times, so we just need to understand what is inside the curly braces. First, we sample k times with replacement from the numbers 1 through 365 and call these the birthdays of the k people, `bdays`. Then, `unique(bdays)` removes duplicates in the vector `bdays`, and `length(unique(bdays))` calculates the length of the vector after duplicates have been removed. The two commands need to be separated by a semicolon.

Now `r` contains the number of distinct birthdays that we observed in each of the 10^4 simulations. The average number of distinct birthdays across the 10^4 simulations

is `mean(r)`. We can compare the simulated value to the theoretical value that we found in Example 4.4.5 using indicator r.v.s:

```
mean(r)
365*(1-(364/365)^k)
```

When we ran the code, both the simulated and theoretical values gave us approximately 19.5.

4.12 Exercises

Exercises marked with Ⓢ have detailed solutions at `http://stat110.net`.

Expectations and variances

1. Bobo, the amoeba from Chapter 2, currently lives alone in a pond. After one minute Bobo will either die, split into two amoebas, or stay the same, with equal probability. Find the expectation and variance for the number of amoebas in the pond after one minute.
2. In the Gregorian calendar, each year has either 365 days (a normal year) or 366 days (a leap year). A year is randomly chosen, with probability 3/4 of being a normal year and 1/4 of being a leap year. Find the mean and variance of the number of days in the chosen year.
3. (a) A fair die is rolled. Find the expected value of the roll.

   (b) Four fair dice are rolled. Find the expected total of the rolls.
4. A fair die is rolled some number of times. You can choose whether to stop after $1, 2$, or 3 rolls, and your decision can be based on the values that have appeared so far. You receive the value shown on the last roll of the die, in dollars. What is your optimal strategy (to maximize your expected winnings)? Find the expected winnings for this strategy.

   Hint: Start by considering a simpler version of this problem, where there are at most 2 rolls. For what values of the first roll should you continue for a second roll?
5. Find the mean and variance of a Discrete Uniform r.v. on $1, 2, \ldots, n$.

   Hint: See the math appendix for some useful facts about sums.
6. Two teams are going to play a best-of-7 match (the match will end as soon as either team has won 4 games). Each game ends in a win for one team and a loss for the other team. Assume that each team is equally likely to win each game, and that the games played are independent. Find the mean and variance of the number of games played.
7. A certain small town, whose population consists of 100 families, has 30 families with 1 child, 50 families with 2 children, and 20 families with 3 children. The *birth rank* of one of these children is 1 if the child is the firstborn, 2 if the child is the secondborn, and 3 if the child is the thirdborn.

   (a) A random family is chosen (with equal probabilities), and then a random child within that family is chosen (with equal probabilities). Find the PMF, mean, and variance of the child's birth rank.

(b) A random child is chosen in the town (with equal probabilities). Find the PMF, mean, and variance of the child's birth rank.

8. A certain country has four regions: North, East, South, and West. The populations of these regions are 3 million, 4 million, 5 million, and 8 million, respectively. There are 4 cities in the North, 3 in the East, 2 in the South, and there is only 1 city in the West. Each person in the country lives in exactly one of these cities.

   (a) What is the average size of a city in the country? (This is the arithmetic mean of the populations of the cities, and is also the expected value of the population of a city chosen uniformly at random.)

   Hint: Give the cities *names* (labels).

   (b) Show that without further information it is impossible to find the variance of the population of a city chosen uniformly at random. That is, the variance depends on how the people within each region are allocated between the cities in that region.

   (c) A region of the country is chosen uniformly at random, and then a city within that region is chosen uniformly at random. What is the expected population size of this randomly chosen city?

   Hint: First find the selection probability for each city.

   (d) Explain intuitively why the answer to (c) is larger than the answer to (a).

9. Consider the following simplified scenario based on *Who Wants to Be a Millionaire?*, a game show in which the contestant answers multiple-choice questions that have 4 choices per question. The contestant (Fred) has answered 9 questions correctly already, and is now being shown the 10th question. He has no idea what the right answers are to the 10th or 11th questions are. He has one "lifeline" available, which he can apply on any question, and which narrows the number of choices from 4 down to 2. Fred has the following options available.
   (a) Walk away with \$16,000.
   (b) Apply his lifeline to the 10th question, and then answer it. If he gets it wrong, he will leave with \$1,000. If he gets it right, he moves on to the 11th question. He then leaves with \$32,000 if he gets the 11th question wrong, and \$64,000 if he gets the 11th question right.
   (c) Same as the previous option, except not using his lifeline on the 10th question, and instead applying it to the 11th question (if he gets the 10th question right).

   Find the expected value of each of these options. Which option has the highest expected value? Which option has the lowest variance?

10. Consider the St. Petersburg paradox (Example 4.3.14), except that you receive $\$n$ rather than $\$2^n$ if the game lasts for n rounds. What is the fair value of this game? What if the payoff is $\$n^2$?

11. Martin has just heard about the following exciting gambling strategy: bet \$1 that a fair coin will land Heads. If it does, stop. If it lands Tails, double the bet for the next toss, now betting \$2 on Heads. If it does, stop. Otherwise, double the bet for the next toss to \$4. Continue in this way, doubling the bet each time and then stopping right after winning a bet. Assume that each individual bet is fair, i.e., has an expected net winnings of 0. The idea is that

$$1 + 2 + 2^2 + 2^3 + \cdots + 2^n = 2^{n+1} - 1,$$

   so the gambler will be \$1 ahead after winning a bet, and then can walk away with a profit.

   Martin decides to try out this strategy. However, he only has \$31, so he may end up walking away bankrupt rather than continuing to double his bet. On average, how much money will Martin win?

12. Let X be a discrete r.v. with support $-n, -n+1, \ldots, 0, \ldots, n-1, n$ for some positive integer n. Suppose that the PMF of X satisfies the symmetry property $P(X = -k) = P(X = k)$ for all integers k. Find $E(X)$.

13. Ⓢ Are there discrete random variables X and Y such that $E(X) > 100E(Y)$ but Y is greater than X with probability at least 0.99?

14. Let X have PMF

$$P(X = k) = cp^k/k \text{ for } k = 1, 2, \ldots,$$

where p is a parameter with $0 < p < 1$ and c is a normalizing constant. We have $c = -1/\log(1-p)$, as seen from the Taylor series

$$-\log(1-p) = p + \frac{p^2}{2} + \frac{p^3}{3} + \ldots.$$

This distribution is called the *Logarithmic* distribution (because of the log in the above Taylor series), and has often been used in ecology. Find the mean and variance of X.

15. Player A chooses a random integer between 1 and 100, with probability p_j of choosing j (for $j = 1, 2, \ldots, 100$). Player B guesses the number that player A picked, and receives from player A that amount in dollars if the guess is correct (and 0 otherwise).

(a) Suppose for this part that player B knows the values of p_j. What is player B's optimal strategy (to maximize expected earnings)?

(b) Show that if both players choose their numbers so that the probability of picking j is proportional to $1/j$, then neither player has an incentive to change strategies, assuming the opponent's strategy is fixed. (In game theory terminology, this says that we have found a *Nash equilibrium*.)

(c) Find the expected earnings of player B when following the strategy from (b). Express your answer both as a sum of simple terms and as a numerical approximation. Does the value depend on what strategy player A uses?

16. The dean of Blotchville University boasts that the average class size there is 20. But the reality experienced by the majority of students there is quite different: they find themselves in huge courses, held in huge lecture halls, with hardly enough seats or Haribo gummi bears for everyone. The purpose of this problem is to shed light on the situation. For simplicity, suppose that every student at Blotchville University takes only one course per semester.

(a) Suppose that there are 16 seminar courses, which have 10 students each, and 2 large lecture courses, which have 100 students each. Find the dean's-eye-view average class size (the simple average of the class sizes) and the student's-eye-view average class size (the average class size experienced by students, as it would be reflected by surveying students and asking them how big their classes are). Explain the discrepancy intuitively.

(b) Give a short proof that for *any* set of class sizes (not just those given above), the dean's-eye-view average class size will be strictly less than the student's-eye-view average class size, unless all classes have exactly the same size.

Hint: Relate this to the fact that variances are nonnegative.

17. The sociologist Elizabeth Wrigley-Field posed the following puzzle [29]:

> *American fertility fluctuated dramatically in the decades surrounding the Second World War. Parents created the smallest families during the Great Depression, and the largest families during the postwar Baby Boom. Yet children born during the Great Depression came from larger families than those born during the Baby Boom. How can this be?*

(a) For a particular era, let n_k be the number of American families with exactly k children, for each $k \geq 0$. (Assume for simplicity that American history has cleanly been separated into eras, where each era has a well-defined set of families, and each family has a well-defined set of children; we are ignoring the fact that a particular family's size may change over time, that children grow up, etc.) For each $j \geq 0$, let

$$m_j = \sum_{k=0}^{\infty} k^j \, n_k.$$

For a *family* selected randomly in that era (with all families equally likely), find the expected number of children in the family. Express your answer only in terms of the m_j's.

(b) For a *child* selected randomly in that era (with all children equally likely), find the expected number of children in the child's family, only in terms of the m_j's.

(c) Give an intuitive explanation in words for which of the answers to (a) and (b) is larger, or whether they are equal. Explain how this relates to the Wrigley-Field puzzle.

Named distributions

18. Ⓢ A fair coin is tossed repeatedly, until it has landed Heads at least once and has landed Tails at least once. Find the expected number of tosses.

19. Ⓢ A coin is tossed repeatedly until it lands Heads for the first time. Let X be the number of tosses that are required (including the toss that landed Heads), and let p be the probability of Heads, so that $X \sim \text{FS}(p)$. Find the CDF of X, and for $p = 1/2$ sketch its graph.

20. Let $X \sim \text{Bin}(100, 0.9)$. For each of the following parts, construct an example showing that it is possible, or explain clearly why it is impossible. In this problem, Y is a random variable on the same probability space as X; note that X and Y are not necessarily independent.

    (a) Is it possible to have $Y \sim \text{Pois}(0.01)$ with $P(X \geq Y) = 1$?

    (b) Is it possible to have $Y \sim \text{Bin}(100, 0.5)$ with $P(X \geq Y) = 1$?

    (c) Is it possible to have $Y \sim \text{Bin}(100, 0.5)$ with $P(X \leq Y) = 1$?

21. Ⓢ Let $X \sim \text{Bin}(n, \frac{1}{2})$ and $Y \sim \text{Bin}(n+1, \frac{1}{2})$, independently.

    (a) Let $V = \min(X, Y)$ be the smaller of X and Y, and let $W = \max(X, Y)$ be the larger of X and Y. So if X crystallizes to x and Y crystallizes to y, then V crystallizes to $\min(x, y)$ and W crystallizes to $\max(x, y)$. Find $E(V) + E(W)$.

    (b) Show that $E|X - Y| = E(W) - E(V)$, with notation as in (a).

    (c) Compute $\text{Var}(n - X)$ in two different ways.

22. Ⓢ Raindrops are falling at an average rate of 20 drops per square inch per minute. What would be a reasonable distribution to use for the number of raindrops hitting a particular region measuring 5 inches^2 in t minutes? Why? Using your chosen distribution, compute the probability that the region has no rain drops in a given 3-second time interval.

23. Ⓢ Alice and Bob have just met, and wonder whether they have a mutual friend. Each has 50 friends, out of 1000 other people who live in their town. They think that it's unlikely that they have a friend in common, saying "each of us is only friends with 5% of the people here, so it would be very unlikely that our two 5%'s overlap."

    Assume that Alice's 50 friends are a random sample of the 1000 people (equally likely

to be any 50 of the 1000), and similarly for Bob. Also assume that knowing who Alice's friends are gives no information about who Bob's friends are.

(a) Compute the expected number of mutual friends Alice and Bob have.

(b) Let X be the number of mutual friends they have. Find the PMF of X.

(c) Is the distribution of X one of the important distributions we have looked at? If so, which?

24. Let $X \sim \text{Bin}(n,p)$ and $Y \sim \text{NBin}(r,p)$. Using a story about a sequence of Bernoulli trials, prove that $P(X < r) = P(Y > n - r)$.

25. Ⓢ Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match. Find the expected number of games played.

Hint: Consider the first two games as a pair, then the next two as a pair, etc.

26. Nick and Penny are independently performing independent Bernoulli trials. For concreteness, assume that Nick is flipping a nickel with probability p_1 of Heads and Penny is flipping a penny with probability p_2 of Heads. Let $X_1, X_2, \ldots$ be Nick's results and $Y_1, Y_2, \ldots$ be Penny's results, with $X_i \sim \text{Bern}(p_1)$ and $Y_j \sim \text{Bern}(p_2)$.

(a) Find the distribution and expected value of the first time at which they are simultaneously successful, i.e., the smallest n such that $X_n = Y_n = 1$.

Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.

(b) Find the expected time until at least one has a success (including the success).

Hint: Define a new sequence of Bernoulli trials and use the story of the Geometric.

(c) For $p_1 = p_2$, find the probability that their first successes are simultaneous, and use this to find the probability that Nick's first success precedes Penny's.

27. Ⓢ Let X and Y be $\text{Pois}(\lambda)$ r.v.s, and $T = X + Y$. Suppose that X and Y are *not* independent, and in fact $X = Y$. Prove or disprove the claim that $T \sim \text{Pois}(2\lambda)$ in this scenario.

28. William is on a treasure hunt. There are t pieces of treasure, each of which is hidden in one of n locations. William searches these locations one by one, without replacement, until he has found all the treasure. (Assume that no location contains more than one piece of treasure, and that William *will* find the treasure piece when he searches a location that does have treasure.) Let X be the number of locations that William searches during his treasure hunt. Find the distribution of X, and find $E(X)$.

29. Let $X \sim \text{Geom}(p)$, and define the function f by $f(x) = P(X = x)$, for all real x. Find $E(f(X))$. (The notation $f(X)$ means first evaluate $f(x)$ in terms of p and x, and then plug in X for x; it is *not* correct to say "$f(X) = P(X = X) = 1$".)

30. (a) Use LOTUS to show that for $X \sim \text{Pois}(\lambda)$ and any function g,

$$E(Xg(X)) = \lambda E(g(X+1)),$$

assuming that both sides exist. This is called the *Stein-Chen identity* for the Poisson.

(b) Find the third moment $E(X^3)$ for $X \sim \text{Pois}(\lambda)$ by using the identity from (a) and a bit of algebra to reduce the calculation to the fact that X has mean λ and variance λ.

31. In many problems about modeling count data, it is found that values of zero in the data are far more common than can be explained well using a Poisson model (we can make $P(X = 0)$ large for $X \sim \text{Pois}(\lambda)$ by making λ small, but that also constrains the mean and variance of X to be small since both are λ). The *Zero-Inflated Poisson* distribution

is a modification of the Poisson to address this issue, making it easier to handle frequent zero values gracefully.

A Zero-Inflated Poisson r.v. X with parameters p and λ can be generated as follows. First flip a coin with probability of p of Heads. Given that the coin lands Heads, $X = 0$. Given that the coin lands Tails, X is distributed Pois(λ). Note that if $X = 0$ occurs, there are two possible explanations: the coin could have landed Heads (in which case the zero is called a *structural zero*), or the coin could have landed Tails but the Poisson r.v. turned out to be zero anyway.

For example, if X is the number of chicken sandwiches consumed by a random person in a week, then $X = 0$ for vegetarians (this is a structural zero), but a chicken-eater could still have $X = 0$ occur by chance (since they might happen not to eat any chicken sandwiches that week).

(a) Find the PMF of a Zero-Inflated Poisson r.v. X.

(b) Explain why X has the same distribution as $(1 - I)Y$, where $I \sim \text{Bern}(p)$ is independent of $Y \sim \text{Pois}(\lambda)$.

(c) Find the mean of X in two different ways: directly using the PMF of X, and using the representation from (b). For the latter, you can use the fact (which we prove in Chapter 7) that if r.v.s Z and W are independent, then $E(ZW) = E(Z)E(W)$.

(d) Find the variance of X.

32. Ⓢ A discrete distribution has the *memoryless property* if for X a random variable with that distribution, $P(X \geq j + k | X \geq j) = P(X \geq k)$ for all nonnegative integers j, k.

(a) If X has a memoryless distribution with CDF F and PMF $p_i = P(X = i)$, find an expression for $P(X \geq j + k)$ in terms of $F(j), F(k), p_j, p_k$.

(b) Name a discrete distribution which has the memoryless property. Justify your answer with a clear interpretation in words or with a computation.

33. Find values of w, b, r such that the Negative Hypergeometric distribution with parameters w, b, r reduces to a Discrete Uniform on $\{0, 1, \ldots, n\}$. Justify your answer both in terms of the story of the Negative Hypergeometric and in terms of its PMF.

Indicator r.v.s

34. Ⓢ Randomly, k distinguishable balls are placed into n distinguishable boxes, with all possibilities equally likely. Find the expected number of empty boxes.

35. Ⓢ A group of 50 people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.). Find the expected number of pairs of people with the same birthday, and the expected number of days in the year on which at least two of these people were born.

36. Ⓢ A group of $n \geq 4$ people are comparing their birthdays (as usual, assume their birthdays are independent, are not February 29, etc.). Let I_{ij} be the indicator r.v. of i and j having the same birthday (for $i < j$). Is I_{12} independent of I_{34}? Is I_{12} independent of I_{13}? Are the I_{ij} independent?

37. Ⓢ A total of 20 bags of Haribo gummi bears are randomly distributed to 20 students. Each bag is obtained by a random student, and the outcomes of who gets which bag are independent. Find the average number of bags of gummi bears that the first three students get in total, and find the average number of students who get at least one bag.

38. Each of $n \geq 2$ people puts their name on a slip of paper (no two have the same name). The slips of paper are shuffled in a hat, and then each person draws one (uniformly at random at each stage, without replacement). Find the average number of people who draw their own names.

39. Two researchers independently select simple random samples from a population of size N, with sample sizes m and n (for each researcher, the sampling is done without replacement, with all samples of the prescribed size equally likely). Find the expected size of the overlap of the two samples.

40. In a sequence of n independent fair coin tosses, what is the expected number of occurrences of the pattern HTH (consecutively)? Note that overlap is allowed, e.g., $HTHTH$ contains two overlapping occurrences of the pattern.

41. You have a well-shuffled 52-card deck. On average, how many pairs of adjacent cards are there such that both cards are red?

42. Suppose there are n types of toys, which you are collecting one by one. Each time you collect a toy, it is equally likely to be any of the n types. What is the expected number of distinct toy types that you have after you have collected t toys? (Assume that you will definitely collect t toys, whether or not you obtain a complete set before then.)

43. A building has n floors, labeled $1, 2, \ldots, n$. At the first floor, k people enter the elevator, which is going up and is empty before they enter. Independently, each decides which of floors $2, 3, \ldots, n$ to go to and presses that button (unless someone has already pressed it).

    (a) Assume for this part only that the probabilities for floors $2, 3, \ldots, n$ are equal. Find the expected number of stops the elevator makes on floors $2, 3, \ldots, n$.

    (b) Generalize (a) to the case that floors $2, 3, \ldots, n$ have probabilities $p_2, \ldots, p_n$ (respectively); you can leave your answer as a finite sum.

44. Ⓢ There are 100 shoelaces in a box. At each stage, you pick two random ends and tie them together. Either this results in a longer shoelace (if the two ends came from different pieces), or it results in a loop (if the two ends came from the same piece). What are the expected number of steps until everything is in loops, and the expected number of loops after everything is in loops? (This is a famous interview problem; leave the latter answer as a sum.)

    Hint: For each step, create an indicator r.v. for whether a loop was created then, and note that the number of free ends goes down by 2 after each step.

45. Show that for any events $A_1, \ldots, A_n$,

$$P(A_1 \cap A_2 \cdots \cap A_n) \geq \sum_{j=1}^{n} P(A_j) - n + 1.$$

    Hint: First prove a similar-looking statement about indicator r.v.s, by interpreting what the events $I(A_1 \cap A_2 \cdots \cap A_n) = 1$ and $I(A_1 \cap A_2 \cdots \cap A_n) = 0$ mean.

46. You have a well-shuffled 52-card deck. You turn the cards face up one by one, without replacement. What is the expected number of non-aces that appear before the first ace? What is the expected number between the first ace and the second ace?

47. You are being tested for psychic powers. Suppose that you do not have psychic powers. A standard deck of cards is shuffled, and the cards are dealt face down one by one. Just after each card is dealt, you name any card (as your prediction). Let X be the number of cards you predict correctly. (See Diaconis [5] for much more about the statistics of testing for psychic powers.)

    (a) Suppose that you get no feedback about your predictions. Show that no matter what strategy you follow, the expected value of X stays the same; find this value. (On the other hand, the *variance* may be very different for different strategies. For example, saying "Ace of Spades" every time gives variance 0.)

    Hint: Indicator r.v.s.

(b) Now suppose that you get partial feedback: after each prediction, you are told immediately whether or not it is right (but without the card being revealed). Suppose you use the following strategy: keep saying a specific card's name (e.g., "Ace of Spades") until you hear that you are correct. Then keep saying a different card's name (e.g., "Two of Spades") until you hear that you are correct (if ever). Continue in this way, naming the same card over and over again until you are correct and then switching to a new card, until the deck runs out. Find the expected value of X, and show that it is very close to $e - 1$.

Hint: Indicator r.v.s.

(c) Now suppose that you get complete feedback: just after each prediction, the card is revealed. Call a strategy "stupid" if it allows, e.g., saying "Ace of Spades" as a guess after the Ace of Spades has already been revealed. Show that any non-stupid strategy gives the same expected value for X; find this value.

Hint: Indicator r.v.s.

48. Ⓢ Let X be Hypergeometric with parameters w, b, n.

(a) Find $E\binom{X}{2}$ *by thinking*, without any complicated calculations.

(b) Use (a) to find the variance of X. You should get

$$\text{Var}(X) = \frac{N-n}{N-1}npq,$$

where $N = w + b, p = w/N, q = 1 - p$.

49. There are n prizes, with values \$1, \$2, ..., \$$n$. You get to choose k random prizes, without replacement. What is the expected total value of the prizes you get?

Hint: Express the total value in the form $a_1 I_1 + \cdots + a_n I_n$, where the a_j are constants and the I_j are indicator r.v.s. Or find the expected value of the jth prize received directly.

50. Ten random chords of a circle are chosen, independently. To generate each of these chords, two independent uniformly random points are chosen on the circle (intuitively, "uniformly" means that the choice is completely random, with no favoritism toward certain angles; formally, it means that the probability of any arc is proportional to the length of that arc). On average, how many pairs of chords intersect?

Hint: Consider two random chords. An equivalent way to generate them is to pick four independent uniformly random points on the circle, and then pair them up randomly.

51. Ⓢ A hash table is being used to store the phone numbers of k people, storing each person's phone number in a uniformly random location, represented by an integer between 1 and n (see Exercise 27 from Chapter 1 for a description of hash tables). Find the expected number of locations with no phone numbers stored, the expected number with exactly one phone number, and the expected number with more than one phone number (should these quantities add up to n?).

52. A coin with probability p of Heads is flipped n times. The sequence of outcomes can be divided into *runs* (blocks of H's or blocks of T's), e.g., $HHHTTHTTTH$ becomes $\boxed{HHH}\boxed{TT}\boxed{H}\boxed{TTT}\boxed{H}$, which has 5 runs. Find the expected number of runs.

Hint: Start by finding the expected number of tosses (other than the first) where the outcome is different from the previous one.

53. A coin with probability p of Heads is flipped 4 times. Let X be the number of occurrences of HH (for example, $THHT$ has 1 occurrence and $HHHH$ has 3 occurrences). Find $E(X)$ and $\text{Var}(X)$.

54. A population has N people, with ID numbers from 1 to N. Let y_j be the value of some numerical variable for person j, and

$$\bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j$$

be the population average of the quantity. For example, if y_j is the height of person j then $\bar{y}$ is the average height in the population, and if y_j is 1 if person j holds a certain belief and 0 otherwise, then $\bar{y}$ is the proportion of people in the population who hold that belief. In this problem, $y_1, y_2, \ldots, y_n$ are thought of as constants rather than random variables.

A researcher is interested in learning about $\bar{y}$, but it is not feasible to measure y_j for all j. Instead, the researcher gathers a random sample of size n, by choosing people one at a time, with equal probabilities at each stage and without replacement. Let W_j be the value of the numerical variable (e.g., height) for the jth person in the sample. Even though $y_1, \ldots, y_n$ are constants, W_j is a random variable because of the random sampling. A natural way to estimate the unknown quantity $\bar{y}$ is using

$$\bar{W} = \frac{1}{n}\sum_{j=1}^{n} W_j.$$

Show that $E(\bar{W}) = \bar{y}$ in two different ways:

(a) by directly evaluating $E(W_j)$ using symmetry;

(b) by showing that $\bar{W}$ can be expressed as a sum over the population by writing

$$\bar{W} = \frac{1}{n}\sum_{j=1}^{N} I_j y_j,$$

where I_j is the indicator of person j being included in the sample, and then using linearity and the fundamental bridge.

55. Ⓢ Consider the following algorithm, known as *bubble sort*, for sorting a list of n distinct numbers into increasing order. Initially they are in a random order, with all orders equally likely. The algorithm compares the numbers in positions 1 and 2, and swaps them if needed, then it compares the new numbers in positions 2 and 3, and swaps them if needed, etc., until it has gone through the whole list. Call this one "sweep" through the list. After the first sweep, the largest number is at the end, so the second sweep (if needed) only needs to work with the first $n-1$ positions. Similarly, the third sweep (if needed) only needs to work with the first $n-2$ positions, etc. Sweeps are performed until $n-1$ sweeps have been completed or there is a swapless sweep.

For example, if the initial list is 53241 (omitting commas), then the following 4 sweeps are performed to sort the list, with a total of 10 comparisons:

$$53241 \to 35241 \to 32541 \to 32451 \to 32415.$$

$$32415 \to 23415 \to 23415 \to 23145.$$

$$23145 \to 23145 \to 21345.$$

$$21345 \to 12345.$$

(a) An *inversion* is a pair of numbers that are out of order (e.g., 12345 has no inversions, while 53241 has 8 inversions). Find the expected number of inversions in the original list.

(b) Show that the expected number of comparisons is between $\frac{1}{2}\binom{n}{2}$ and $\binom{n}{2}$.

Hint: For one bound, think about how many comparisons are made if $n-1$ sweeps are done; for the other bound, use Part (a).

56. A certain basketball player practices shooting free throws over and over again. The shots are independent, with probability p of success.

    (a) In n shots, what is the expected number of streaks of 7 consecutive successful shots? (Note that, for example, 9 in a row counts as 3 streaks.)

    (b) Now suppose that the player keeps shooting until making 7 shots in a row for the first time. Let X be the number of shots taken. Show that $E(X) \leq 7/p^7$.

    Hint: Consider the first 7 trials as a block, then the next 7 as a block, etc.

57. Ⓢ An urn contains red, green, and blue balls. Balls are chosen randomly with replacement (each time, the color is noted and then the ball is put back). Let r, g, b be the probabilities of drawing a red, green, blue ball, respectively ($r + g + b = 1$).

    (a) Find the expected number of balls chosen before obtaining the first red ball, not including the red ball itself.

    (b) Find the expected number of different *colors* of balls obtained before getting the first red ball.

    (c) Find the probability that at least 2 of n balls drawn are red, given that at least 1 is red.

58. Ⓢ Job candidates $C_1, C_2, \ldots$ are interviewed one by one, and the interviewer compares them and keeps an updated list of rankings (if n candidates have been interviewed so far, this is a list of the n candidates, from best to worst). Assume that there is no limit on the number of candidates available, that for any n the candidates $C_1, C_2, \ldots, C_n$ are equally likely to arrive in any order, and that there are no ties in the rankings given by the interview.

    Let X be the index of the first candidate to come along who ranks as better than the very first candidate C_1 (so C_X is better than C_1, but the candidates after 1 but prior to X (if any) are worse than C_1. For example, if C_2 and C_3 are worse than C_1 but C_4 is better than C_1, then $X = 4$. All 4! orderings of the first 4 candidates are equally likely, so it could have happened that the first candidate was the best out of the first 4 candidates, in which case $X > 4$.

    What is $E(X)$ (which is a measure of how long, on average, the interviewer needs to wait to find someone better than the very first candidate)?

    Hint: Find $P(X > n)$ by interpreting what $X > n$ says about how C_1 compares with other candidates, and then apply the result of Theorem 4.4.8.

59. People are arriving at a party one at a time. While waiting for more people to arrive they entertain themselves by comparing their birthdays. Let X be the number of people needed to obtain a birthday match, i.e., before person X arrives there are no two people with the same birthday, but when person X arrives there is a match.

    Assume for this problem that there are 365 days in a year, all equally likely. By the result of the birthday problem from Chapter 1, for 23 people there is a 50.7% chance of a birthday match (and for 22 people there is a less than 50% chance). But this has to do with the *median* of X (defined below); we also want to know the *mean* of X, and in this problem we will find it, and see how it compares with 23.

    (a) A *median* of a random variable Y is a value m for which $P(Y \leq m) \geq 1/2$ and $P(Y \geq m) \geq 1/2$ (this is also called a median of the *distribution* of Y; note that the notion is completely determined by the CDF of Y). Every distribution has a median, but for some distributions it is not unique. Show that 23 is the *unique* median of X.

(b) Show that $X = I_1 + I_2 + \cdots + I_{366}$, where I_j is the indicator r.v. for the event $X \geq j$. Then find $E(X)$ in terms of p_j's defined by $p_1 = p_2 = 1$ and for $3 \leq j \leq 366$,

$$p_j = \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right)\cdots\left(1 - \frac{j-2}{365}\right).$$

(c) Compute $E(X)$ numerically. In R, the pithy command `cumprod(1-(0:364)/365)` produces the vector $(p_2, \ldots, p_{366})$.

(d) Find the variance of X, both in terms of the p_j's and numerically.

Hint: What is I_i^2, and what is $I_i I_j$ for $i < j$? Use this to simplify the expansion

$$X^2 = I_1^2 + \cdots + I_{366}^2 + 2\sum_{j=2}^{366}\sum_{i=1}^{j-1} I_i I_j.$$

Note: In addition to being an entertaining game for parties, the birthday problem has many applications in computer science, such as in a method called the *birthday attack* in cryptography. It can be shown that if there are n days in a year and n is large, then

$$E(X) \approx \sqrt{\frac{\pi n}{2}} + \frac{2}{3}.$$

60. Elk dwell in a certain forest. There are N elk, of which a simple random sample of size n is captured and tagged (so all $\binom{N}{n}$ sets of n elk are equally likely). The captured elk are returned to the population, and then a new sample is drawn. This is an important method that is widely used in ecology, known as *capture-recapture*. If the new sample is also a simple random sample, with some fixed size, then the number of tagged elk in the new sample is Hypergeometric.

    For this problem, assume that instead of having a fixed sample size, elk are sampled one by one without replacement until m tagged elk have been recaptured, where m is specified in advance (of course, assume that $1 \leq m \leq n \leq N$). An advantage of this sampling method is that it can be used to avoid ending up with a very small number of tagged elk (maybe even zero), which would be problematic in many applications of capture-recapture. A disadvantage is not knowing how large the sample will be.

    (a) Find the PMFs of the number of untagged elk in the new sample (call this X) and of the total number of elk in the new sample (call this Y).

    (b) Find the expected sample size EY using symmetry, linearity, and indicator r.v.s.

    (c) Suppose that m, n, N are such that EY is an integer. If the sampling is done with a fixed sample size equal to EY rather than sampling until exactly m tagged elk are obtained, find the expected number of tagged elk in the sample. Is it less than m, equal to m, or greater than m (for $n < N$)?

LOTUS

61. Ⓢ For $X \sim \text{Pois}(\lambda)$, find $E(X!)$ (the average factorial of X), if it is finite.

62. For $X \sim \text{Pois}(\lambda)$, find $E(2^X)$, if it is finite.

63. For $X \sim \text{Geom}(p)$, find $E(2^X)$ (if it is finite) and $E(2^{-X})$ (if it is finite). For each, make sure to clearly state what the values of p are for which it is finite.

64. Ⓢ Let $X \sim \text{Geom}(p)$ and let t be a constant. Find $E(e^{tX})$, as a function of t (this is known as the *moment generating function*; we will see in Chapter 6 how this function is useful).

65. Ⓢ The number of fish in a certain lake is a Pois(λ) random variable. Worried that there might be no fish at all, a statistician adds one fish to the lake. Let Y be the resulting number of fish (so Y is 1 plus a Pois(λ) random variable).

    (a) Find $E(Y^2)$.

    (b) Find $E(1/Y)$.

66. Ⓢ Let X be a Pois(λ) random variable, where λ is fixed but unknown. Let $\theta = e^{-3\lambda}$, and suppose that we are interested in estimating θ based on the data. Since X is what we observe, our estimator is a function of X, call it $g(X)$. The *bias* of the estimator $g(X)$ is defined to be $E(g(X)) - \theta$, i.e., how far off the estimate is on average; the estimator is *unbiased* if its bias is 0.

    (a) For estimating λ, the r.v. X itself is an unbiased estimator. Compute the bias of the estimator $T = e^{-3X}$. Is it unbiased for estimating θ?

    (b) Show that $g(X) = (-2)^X$ is an unbiased estimator for θ. (In fact, it turns out to be the only unbiased estimator for θ.)

    (c) Explain intuitively why $g(X)$ is a silly choice for estimating θ, despite (b), and show how to improve it by finding an estimator $h(X)$ for θ that is always at least as good as $g(X)$ and sometimes strictly better than $g(X)$. That is,

    $$|h(X) - \theta| \leq |g(X) - \theta|,$$

    with the inequality sometimes strict.

Poisson approximation

67. Ⓢ Law school courses often have assigned seating to facilitate the Socratic method. Suppose that there are 100 first-year law students, and each takes the same two courses: Torts and Contracts. Both are held in the same lecture hall (which has 100 seats), and the seating is uniformly random and independent for the two courses.

    (a) Find the probability that no one has the same seat for both courses (exactly; you should leave your answer as a sum).

    (b) Find a simple but accurate approximation to the probability that no one has the same seat for both courses.

    (c) Find a simple but accurate approximation to the probability that at least two students have the same seat for both courses.

68. Ⓢ A group of n people play "Secret Santa" as follows: each puts their name on a slip of paper in a hat, picks a name randomly from the hat (without replacement), and then buys a gift for that person. Unfortunately, they overlook the possibility of drawing one's own name, so some may have to buy gifts for themselves (on the bright side, some may like self-selected gifts better). Assume $n \geq 2$.

    (a) Find the expected value of the number X of people who pick their own names.

    (b) Find the expected number of pairs of people, A and B, such that A picks B's name and B picks A's name (where $A \neq B$ and order doesn't matter).

    (c) What is the *approximate* distribution of X if n is large (specify the parameter value or values)? What does $P(X = 0)$ converge to as $n \to \infty$?

69. A survey is being conducted in a city with a million (10^6) people. A sample of size 1000 is collected by choosing people in the city at random, *with* replacement and with equal probabilities for everyone in the city. Find a simple, accurate approximation to the probability that at least one person will get chosen more than once (in contrast, Exercise 26 from Chapter 1 asks for an exact answer).

    Hint: Indicator r.v.s are useful here, but creating 1 indicator for each of the million people is *not* recommended since it leads to a messy calculation. Feel free to use the fact that $999 \approx 1000$.

70. Ⓢ Ten million people enter a certain lottery. For each person, the chance of winning is one in ten million, independently.

    (a) Find a simple, good approximation for the PMF of the number of people who win the lottery.

    (b) Congratulations! You won the lottery. However, there may be other winners. Assume now that the number of winners other than you is $W \sim \text{Pois}(1)$, and that if there is more than one winner, then the prize is awarded to one randomly chosen winner. Given this information, find the probability that you win the prize (simplify).

71. In a group of 90 people, find a simple, good approximation for the probability that there is at least one pair of people such that they share a birthday *and* their biological mothers share a birthday. Assume that no one among the 90 people is the biological mother of another one of the 90 people, nor do two of the 90 people have the same biological mother. Express your answer as a fully simplified fraction in the form a/b, where a and b are positive integers and $b \leq 100$.

    Make the usual assumptions as in the birthday problem. To simplify the calculation, you can use the approximations $365 \approx 360$ and $89 \approx 90$, and the fact that $e^x \approx 1 + x$ for $x \approx 0$.

72. Use Poisson approximations to investigate the following types of coincidences. The usual assumptions of the birthday problem apply.

    (a) How many people are needed to have a 50% chance that at least one of them has the same birthday as *you*?

    (b) How many people are needed to have a 50% chance that there is at least one pair of people who not only were born on the same day of the year, but also were born at the same *hour* (e.g., two people born between 2 pm and 3 pm are considered to have been born at the same hour)?

    (c) Considering that only 1/24 of pairs of people born on the same day were born at the same hour, why isn't the answer to (b) approximately $24 \cdot 23$?

    (d) With 100 people, there is a 64% chance that there is at least one set of 3 people with the same birthday (according to R, using `pbirthday(100,classes=365,coincident=3)` to compute it). Provide two different Poisson approximations for this value, one based on creating an indicator r.v. for each triplet of people, and the other based on creating an indicator r.v. for each day of the year. Which is more accurate?

73. A chess tournament has 100 players. In the first round, they are randomly paired to determine who plays whom (so 50 games are played). In the second round, they are again randomly paired, independently of the first round. In both rounds, all possible pairings are equally likely. Let X be the number of people who play against the same opponent twice.

    (a) Find the expected value of X.

    (b) Explain why X is *not* approximately Poisson.

    (c) Find good approximations to $P(X = 0)$ and $P(X = 2)$, by thinking about games in the second round such that the same pair played each other in the first round.

*Existence

74. Ⓢ Each of 111 people names their 5 favorite movies out of a list of 11 movies.

(a) Alice and Bob are 2 of the 111 people. Assume *for this part only* that Alice's 5 favorite movies out of the 11 are random, with all sets of 5 equally likely, and likewise for Bob, independently. Find the expected number of movies in common to Alice's and Bob's lists of favorite movies.

(b) Show that there are 2 movies such that at least 21 of the people name both of these movies as favorites.

75. Ⓢ The circumference of a circle is colored with red and blue ink such that 2/3 of the circumference is red and 1/3 is blue. Prove that no matter how complicated the coloring scheme is, there is a way to inscribe a square in the circle such that at least three of the four corners of the square touch red ink.

76. Ⓢ A hundred students have taken an exam consisting of 8 problems, and for each problem at least 65 of the students got the right answer. Show that there exist two students who collectively got everything right, in the sense that for each problem, at least one of the two got it right.

77. Ⓢ Ten points in the plane are designated. You have ten circular coins (of the same radius). Show that you can position the coins in the plane (without stacking them) so that all ten points are covered.

Hint: Consider a *honeycomb tiling* of the plane (this is a way to divide the plane into hexagons). You can use the fact from geometry that if a circle is inscribed in a hexagon then the ratio of the area of the circle to the area of the hexagon is $\frac{\pi}{2\sqrt{3}} > 0.9$.

78. Ⓢ Let S be a set of binary strings $a_1 \dots a_n$ of length n (where juxtaposition means concatenation). We call S *k-complete* if for any indices $1 \leq i_1 < \cdots < i_k \leq n$ and any binary string $b_1 \dots b_k$ of length k, there is a string $s_1 \dots s_n$ in S such that $s_{i_1} s_{i_2} \dots s_{i_k} = b_1 b_2 \dots b_k$. For example, for $n = 3$, the set $S = \{001, 010, 011, 100, 101, 110\}$ is 2-complete since all 4 patterns of 0's and 1's of length 2 can be found in any 2 positions. Show that if $\binom{n}{k} 2^k (1 - 2^{-k})^m < 1$, then there exists a k-complete set of size at most m.

Mixed practice

79. A hacker is trying to break into a password-protected website by randomly trying to guess the password. Let m be the number of possible passwords.

(a) Suppose for this part that the hacker makes random guesses (with equal probability), *with replacement*. Find the average number of guesses it will take until the hacker guesses the correct password (including the successful guess).

(b) Now suppose that the hacker guesses randomly, *without replacement*. Find the average number of guesses it will take until the hacker guesses the correct password (including the successful guess).

Hint: Use symmetry.

(c) Show that the answer to (a) is greater than the answer to (b) (except in the degenerate case $m = 1$), and explain why this makes sense intuitively.

(d) Now suppose that the website locks out any user after n incorrect password attempts, so the hacker can guess at most n times. Find the PMF of the number of guesses that the hacker makes, both for the case of sampling with replacement and for the case of sampling without replacement.

80. A fair 20-sided die is rolled repeatedly, until a gambler decides to stop. The gambler receives the amount shown on the die when the gambler stops. The gambler decides in advance to roll the die until a value of m or greater is obtained, and then stop (where m is a fixed integer with $1 \leq m \leq 20$).

    (a) What is the expected number of rolls (simplify)?

    (b) What is the expected square root of the number of rolls (as a sum)?

81. Ⓢ A group of 360 people is going to be split into 120 teams of 3 (where the order of teams and the order within a team don't matter).

    (a) How many ways are there to do this?

    (b) The group consists of 180 married couples. A random split into teams of 3 is chosen, with all possible splits equally likely. Find the expected number of teams containing married couples.

82. Ⓢ The gambler de Méré asked Pascal whether it is more likely to get at least one six in 4 rolls of a die, or to get at least one double-six in 24 rolls of a pair of dice. Continuing this pattern, suppose that a group of n fair dice is rolled $4 \cdot 6^{n-1}$ times.

    (a) Find the expected number of times that "all sixes" is achieved (i.e., how often among the $4 \cdot 6^{n-1}$ rolls it happens that all n dice land 6 simultaneously).

    (b) Give a simple but accurate approximation of the probability of having at least one occurrence of "all sixes", for n large (in terms of e but not n).

    (c) de Méré finds it tedious to re-roll so many dice. So after one normal roll of the n dice, in going from one roll to the next, with probability 6/7 he leaves the dice in the same configuration and with probability 1/7 he re-rolls. For example, if $n = 3$ and the 7th roll is $(3, 1, 4)$, then 6/7 of the time the 8th roll remains $(3, 1, 4)$ and 1/7 of the time the 8th roll is a new random outcome. Does the expected number of times that "all sixes" is achieved stay the same, increase, or decrease (compared with (a))? Give a short but clear explanation.

83. Ⓢ Five people have just won a \$100 prize, and are deciding how to divide the \$100 up between them. Assume that whole dollars are used, not cents. Also, for example, giving \$50 to the first person and \$10 to the second is different from vice versa.

    (a) How many ways are there to divide up the \$100, such that each gets at least \$10?

    (b) Assume that the \$100 is randomly divided up, with all of the possible allocations counted in (a) equally likely. Find the expected amount of money that the first person receives.

    (c) Let A_j be the event that the jth person receives more than the first person (for $2 \leq j \leq 5$), when the \$100 is randomly allocated as in (b). Are A_2 and A_3 independent?

84. Ⓢ Joe's iPod has 500 different songs, consisting of 50 albums of 10 songs each. He listens to 11 random songs on his iPod, with all songs equally likely and chosen independently (so repetitions may occur).

    (a) What is the PMF of how many of the 11 songs are from his favorite album?

    (b) What is the probability that there are 2 (or more) songs from the same album among the 11 songs he listens to?

    (c) A pair of songs is a *match* if they are from the same album. If, say, the 1st, 3rd, and 7th songs are all from the same album, this counts as 3 matches. Among the 11 songs he listens to, how many matches are there on average?

85. Ⓢ Each day that the Mass Cash lottery is run in Massachusetts, 5 of the integers from 1 to 35 are chosen (randomly and without replacement).

(a) When playing this lottery, find the probability of guessing exactly 3 numbers right, given that you guess at least 1 of the numbers right.

(b) Find an exact expression for the expected number of days needed so that all of the $\binom{35}{5}$ possible lottery outcomes will have occurred.

(c) Approximate the probability that after 50 days of the lottery, every number from 1 to 35 has been picked at least once.

86. A certain country has three political parties, denoted by A, B, and C. Each adult in the country is a member of exactly one of the three parties. There are n adults in the country, consisting of n_A members of party A, n_B members of party B, and n_C members of party C, where n_A, n_B, n_C are positive integers with $n_A + n_B + n_C = n$.

A simple random sample of size m is chosen from the adults in the country (the sampling is done *without* replacement, and all possible samples of size m are equally likely). Let X be the number of members of party A in the sample, Y be the number of members of party B in the sample, and Z be the number of members of party C in the sample.

(a) Find $P(X = x, Y = y, Z = z)$, for x, y, z nonnegative integers with $x + y + z = m$.

(b) Find $E(X)$.

(c) Find $\text{Var}(X)$, and briefly explain why your answer makes sense in the extreme cases $m = 1$ and $m = n$.

87. The U.S. Senate consists of 100 senators, with 2 from each of the 50 states. There are d Democrats in the Senate. A committee of size c is formed, by picking a random set of senators such that all sets of size c are equally likely.

(a) Find the expected number of Democrats on the committee.

(b) Find the expected number of states represented on the committee (by at least one senator).

(c) Find the expected number of states such that both of the state's senators are on the committee.

(d) Each state has a *junior senator* and a *senior senator* (based on which of them has served longer). A committee of size 20 is formed randomly, with all sets of 20 senators equally likely. Find the distribution of the number of junior senators on the committee, and the expected number of junior senators on the committee.

(e) For the committee from (d), find the expected number of states such that both senators from that state are on the committee.

88. A certain college has g good courses and b bad courses, where g and b are positive integers. Alice, who is hoping to find a good course, randomly shops courses one at a time (without replacement) until she finds a good course.

(a) Find the expected number of bad courses that Alice shops before finding a good course (as a simple expression in terms of g and b).

(b) Should the answer to (a) be less than, equal to, or greater than b/g? Explain this using properties of the Geometric distribution.

89. A DNA sequence can be represented as a sequence of letters, where the *alphabet* has 4 letters: A,C,T,G. Suppose such a sequence is generated randomly, where the letters are independent and the probabilities of A,C,T,G are p_A, p_C, p_T, p_G, respectively.

    (a) In a DNA sequence of length 115, what is the variance of the number of occurrences of the letter C?

    (b) In a DNA sequence of length 115, what is the expected number of occurrences of the expression CATCAT? Note that, for example, the expression CATCATCAT counts as 2 occurrences.

    (c) In a DNA sequence of length 6, what is the probability that the expression CAT occurs at least once?

90. Alice is conducting a survey in a town with population size 1000. She selects a simple random sample of size 100 (i.e., sampling without replacement, such that all samples of size 100 are equally likely). Bob is also conducting a survey in this town. Bob selects a simple random sample of size 20, independent of Alice's sample. Let A be the set of people in Alice's sample and B be the set of people in Bob's sample.

    (a) Find the expected number of people in $A \cap B$.

    (b) Find the expected number of people in $A \cup B$.

    (c) The 1000 people consist of 500 married couples. Find the expected number of couples such that both members of the couple are in Bob's sample.

91. The *Wilcoxon rank sum test* is a widely used procedure for assessing whether two groups of observations come from the same distribution. Let group 1 consist of i.i.d. $X_1, \ldots, X_m$ with CDF F and group 2 consist of i.i.d. $Y_1, \ldots, Y_n$ with CDF G, with all of these r.v.s independent. Assume that the probability of 2 of the observations being equal is 0 (this will be true if the distributions are continuous).

    After the $m+n$ observations are obtained, they are listed in increasing order, and each is assigned a *rank* between 1 and $m+n$: the smallest has rank 1, the second smallest has rank 2, etc. Let R_j be the rank of X_j among all the observations for $1 \leq j \leq m$, and let $R = \sum_{j=1}^{m} R_j$ be the sum of the ranks for group 1.

    Intuitively, the Wilcoxon rank sum test is based on the idea that a very large value of R is evidence that observations from group 1 are usually larger than observations from group 2 (and vice versa if R is very small). But how large is "very large" and how small is "very small"? Answering this precisely requires studying the distribution of the *test statistic* R.

    (a) The *null hypothesis* in this setting is that $F = G$. Show that if the null hypothesis is true, then $E(R) = m(m+n+1)/2$.

    (b) The *power* of a test is an important measure of how good the test is about saying to reject the null hypothesis if the null hypothesis is false. To study the power of the Wilcoxon rank sum test, we need to study the distribution of R in general. So for this part, we do *not* assume $F = G$. Let $p = P(X_1 > Y_1)$. Find $E(R)$ in terms of m, n, p.

    Hint: Write R_j in terms of indicator r.v.s for X_j being greater than various other r.v.s.

92. The legendary Caltech physicist Richard Feynman and two editors of *The Feynman Lectures on Physics* (Michael Gottlieb and Ralph Leighton) posed the following problem about how to decide what to order at a restaurant. You plan to eat m meals at a certain restaurant, where you have never eaten before. Each time, you will order one dish.

    The restaurant has n dishes on the menu, with $n \geq m$. Assume that if you had tried all the dishes, you would have a definite ranking of them from 1 (your least favorite) to n (your favorite). If you knew which your favorite was, you would be happy to order it always (you never get tired of it).

Before you've eaten at the restaurant, this ranking is completely unknown to you. After you've tried some dishes, you can rank those dishes amongst themselves, but don't know how they compare with the dishes you haven't yet tried. There is thus an *exploration-exploitation tradeoff*: should you try new dishes, or should you order your favorite among the dishes you have tried before?

A natural strategy is to have two phases in your series of visits to the restaurant: an *exploration phase*, where you try different dishes each time, and an *exploitation phase*, where you always order the best dish you obtained in the exploration phase. Let k be the length of the exploration phase (so $m-k$ is the length of the exploitation phase).

Your goal is to maximize the expected sum of the ranks of the dishes you eat there (the rank of a dish is the "true" rank from 1 to n that you would give that dish if you could try all the dishes). Show that the optimal choice is

$$k = \sqrt{2(m+1)} - 1,$$

or this rounded up or down to an integer if needed. Do this in the following steps:

(a) Let X be the rank of the best dish that you find in the exploration phase. Find the expected sum of the ranks of all the dishes you eat (including both phases), in terms of k, n, and $E(X)$.

(b) Find the PMF of X, as a simple expression in terms of binomial coefficients.

(c) Show that

$$E(X) = \frac{k(n+1)}{k+1}.$$

Hint: Use Example 1.5.2 (about the team captain) and Exercise 20 from Chapter 1 (about the hockey stick identity).

(d) Use calculus to find the optimal value of k.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

5

Continuous random variables

So far we have been working with discrete random variables, whose possible values can be written down as a list. In this chapter we will discuss *continuous* r.v.s, which can take on any real value in an interval (possibly of infinite length, such as $(0, \infty)$ or the entire real line). First we'll look at properties of continuous r.v.s in general. Then we'll introduce three famous continuous distributions—the Uniform, Normal, and Exponential—which, in addition to having important stories in their own right, serve as building blocks for many other useful continuous distributions.

5.1 Probability density functions

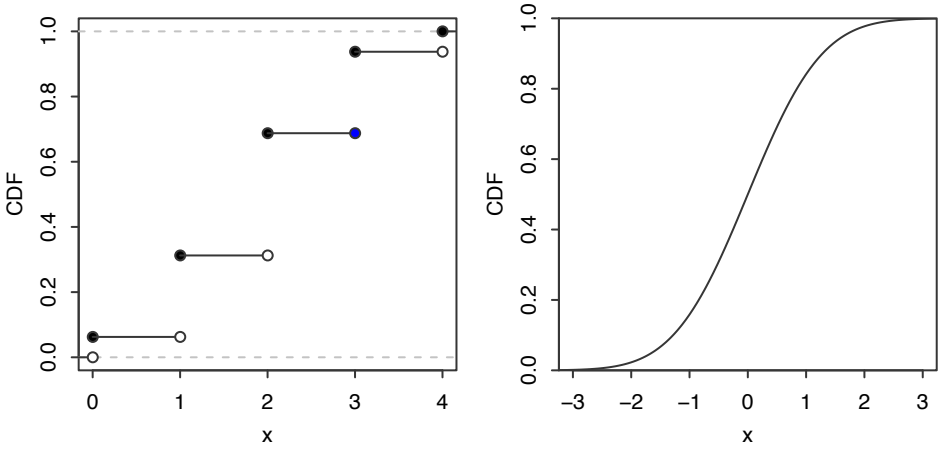


FIGURE 5.1
Discrete vs. continuous r.v.s. Left: The CDF of a discrete r.v. has jumps at each point in the support. Right: The CDF of a continuous r.v. increases smoothly.

Recall that for a discrete r.v., the CDF jumps at every point in the support, and is flat everywhere else. In contrast, for a continuous r.v. the CDF increases smoothly; see Figure 5.1 for a comparison of discrete vs. continuous CDFs.

Definition 5.1.1 (Continuous r.v.)**.** An r.v. has a *continuous distribution* if its

CDF is differentiable. We also allow there to be endpoints (or finitely many points) where the CDF is continuous but not differentiable, as long as the CDF is differentiable everywhere else. A *continuous random variable* is a random variable with a continuous distribution.

For discrete r.v.s, the CDF is awkward to work with because of its jumpiness, and its derivative is almost useless since it's undefined at the jumps and 0 everywhere else. But for continuous r.v.s, the CDF is often convenient to work with, and its derivative is a very useful function, called the *probability density function*.

Definition 5.1.2 (Probability density function)**.** For a continuous r.v. X with CDF F, the *probability density function* (PDF) of X is the derivative f of the CDF, given by $f(x) = F'(x)$. The *support* of X, and of its distribution, is the set of all x where $f(x) > 0$.

An important way in which continuous r.v.s differ from discrete r.v.s is that for a continuous r.v. X, $P(X = x) = 0$ for all x. This is because $P(X = x)$ is the height of a jump in the CDF at x, but the CDF of X has no jumps! Since the PMF of a continuous r.v. would just be 0 everywhere, we work with a PDF instead.

The PDF is analogous to the PMF in many ways, but there is a key difference: for a PDF f, the quantity $f(x)$ is *not* a probability, and in fact it is possible to have $f(x) > 1$ for some values of x. To obtain a probability, we need to *integrate* the PDF. The fundamental theorem of calculus tells us how to get from the PDF back to the CDF.

Proposition 5.1.3 (PDF to CDF)**.** Let X be a continuous r.v. with PDF f. Then the CDF of X is given by

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Proof. By definition of PDF, F is an antiderivative of f. So by the fundamental theorem of calculus,

$$\int_{-\infty}^{x} f(t)dt = F(x) - F(-\infty) = F(x). \qquad \blacksquare$$

The above result is analogous to how we obtained the value of a discrete CDF at x by summing the PMF over all values less than or equal to x; here we *integrate* the PDF over all values up to x, so the CDF is the *accumulated area* under the PDF. Since we can freely convert between the PDF and the CDF using the inverse operations of integration and differentiation, both the PDF and CDF carry complete information about the distribution of a continuous r.v.

Since the PDF determines the distribution, we should be able to use it to find the probability of X falling into an interval (a, b). A handy fact is that we can include or exclude the endpoints as we wish without altering the probability, since the endpoints have probability 0:

$$P(a < X < b) = P(a < X \leq b) = P(a \leq X < b) = P(a \leq X \leq b).$$

☣ **5.1.4** (Including or excluding endpoints)**.** We can be carefree about including or excluding endpoints as above for continuous r.v.s, but we must not be careless about this for discrete r.v.s.

By definition of CDF and the fundamental theorem of calculus,

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx.$$

Therefore, to find the probability of X falling in the interval $(a, b]$ (or (a, b), $[a, b)$, or $[a, b]$) using the PDF, we simply integrate the PDF from a to b. In general, for an arbitrary region $A \subseteq \mathbb{R}$,

$$P(X \in A) = \int_A f(x)dx.$$

In summary:

To get a desired probability, integrate the PDF over the appropriate range.

Just as a valid PMF must be nonnegative and sum to 1, a valid PDF must be nonnegative and integrate to 1.

Theorem 5.1.5 (Valid PDFs)**.** The PDF f of a continuous r.v. must satisfy the following two criteria:

- Nonnegative: $f(x) \geq 0$;
- Integrates to 1: $\int_{-\infty}^{\infty} f(x)dx = 1$.

Proof. The first criterion is true because probability is nonnegative; if $f(x_0)$ were negative, then we could integrate over a tiny region around x_0 and get a negative probability. Alternatively, note that the PDF at x_0 is the slope of the CDF at x_0, so $f(x_0) < 0$ would imply that the CDF is *decreasing* at x_0, which is not allowed. The second criterion is true since $\int_{-\infty}^{\infty} f(x)dx$ is the probability of X falling somewhere on the real line, which is 1. ■

Conversely, any such function f *is* the PDF of some r.v. This is because if f satisfies these properties, we can integrate it as in Proposition 5.1.3 to get a function F satisfying the properties of a CDF. Then a version of Universality of the Uniform, the main concept in Section 5.3, can be used to create an r.v. with CDF F.

Now let's look at some specific examples of PDFs. The two distributions in the following examples are named the Logistic and Rayleigh distributions, but we won't discuss their stories here; their appearance is intended mainly as a way of getting comfortable with PDFs.

Example 5.1.6 (Logistic)**.** The Logistic distribution has CDF

$$F(x) = \frac{e^x}{1 + e^x}, \quad x \in \mathbb{R}.$$

To get the PDF, we differentiate the CDF, which gives

$$f(x) = \frac{e^x}{(1+e^x)^2}, \quad x \in \mathbb{R}.$$

Let $X \sim$ Logistic. To find $P(-2 < X < 2)$, integrate the PDF from -2 to 2:

$$P(-2 < X < 2) = \int_{-2}^{2} \frac{e^x}{(1+e^x)^2} dx = F(2) - F(-2) \approx 0.76.$$

The integral was easy to evaluate since we already knew that F was an antiderivative for f, and we had a nice expression for F. Otherwise, we could have made the substitution $u = 1 + e^x$, so $du = e^x dx$, giving

$$\int_{-2}^{2} \frac{e^x}{(1+e^x)^2} dx = \int_{1+e^{-2}}^{1+e^2} \frac{1}{u^2} du = \left(-\frac{1}{u}\right)\Big|_{1+e^{-2}}^{1+e^2} \approx 0.76.$$

Figure 5.2 shows the Logistic PDF (left) and CDF (right). On the PDF, the probability $P(-2 < X < 2)$ is represented by the shaded area; on the CDF, it is represented by the height of the curly brace. You can check that the properties of a valid PDF and CDF are satisfied. □

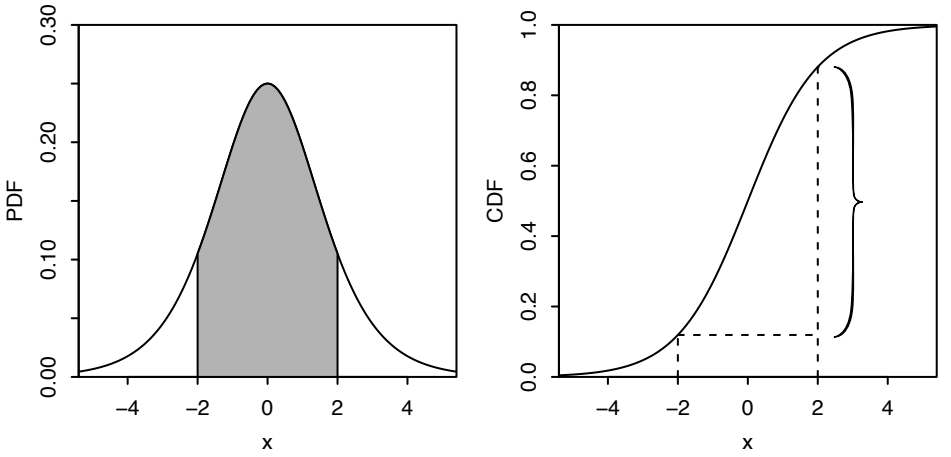


FIGURE 5.2
Logistic PDF and CDF. The probability $P(-2 < X < 2)$ is indicated by the shaded area under the PDF and the height of the curly brace on the CDF.

Example 5.1.7 (Rayleigh)**.** The Rayleigh distribution has CDF

$$F(x) = 1 - e^{-x^2/2}, \quad x > 0.$$

To get the PDF, we differentiate the CDF, which gives

$$f(x) = xe^{-x^2/2}, \quad x > 0.$$

For $x \leq 0$, both the CDF and the PDF are equal to 0.

Let $X \sim$ Rayleigh. To find $P(X > 2)$, we need to integrate the PDF from 2 to ∞. We can do that by making the substitution $u = -x^2/2$, but since we already have the CDF in a nice form we know the integral is $F(\infty) - F(2) = 1 - F(2)$:

$$P(X > 2) = \int_2^\infty xe^{-x^2/2}dx = 1 - F(2) \approx 0.14.$$

The Rayleigh PDF and CDF are plotted in Figure 5.3. Again, probability is represented by a shaded area on the PDF and a vertical height on the CDF. □

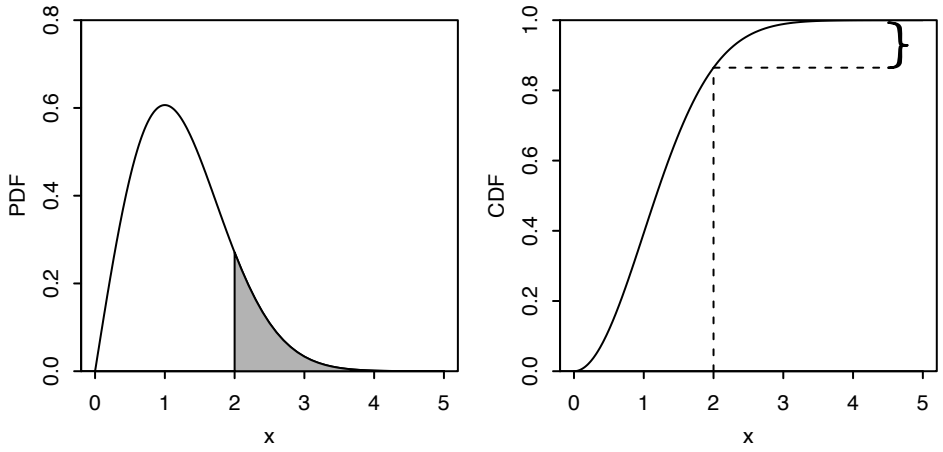


FIGURE 5.3
Rayleigh PDF and CDF. The probability $P(X > 2)$ is indicated by the shaded area under the PDF and the height of the curly brace on the CDF.

Although the height of a PDF at x does not represent a probability, it is closely related to the probability of falling into a tiny interval around x, as the following intuition explains.

Intuition 5.1.8 (Units)**.** Let F be the CDF and f be the PDF of a continuous r.v. X. As mentioned earlier, $f(x)$ is *not* a probability; for example, we could have $f(3) > 1$, and we know $P(X = 3) = 0$. But thinking about the probability of X being *very close* to 3 gives us a way to interpret $f(3)$. Specifically, the probability of X being in a tiny interval of length ϵ, centered at 3, will essentially be $f(3)\epsilon$. This is because

$$P(3 - \epsilon/2 < X < 3 + \epsilon/2) = \int_{3-\epsilon/2}^{3+\epsilon/2} f(x)dx \approx f(3)\epsilon,$$

if the interval is so tiny that f is approximately the constant $f(3)$ on that interval. In general, we can think of $f(x)dx$ as the probability of X being in an infinitesimally small interval containing x, of length dx.

In practice, X often has *units* in some system of measurement, such as units of

distance, time, area, or mass. Thinking about the units is not only important in applied problems, but also it often helps in checking that answers make sense.

Suppose for concreteness that X is a length, measured in centimeters (cm). Then $f(x) = dF(x)/dx$ is the probability per cm at x, which explains why $f(x)$ is a probability *density*. Probability is a dimensionless quantity (a number without physical units), so the units of $f(x)$ are cm^{-1}. Therefore, to be able to get a probability again, we need to multiply $f(x)$ by a length. When we do an integral such as $\int_0^5 f(x)dx$, this is achieved by the often-forgotten dx. □

Intuition 5.1.9 (Simulation)**.** For another intuitive way to think about PDFs, consider the following graphical way to *simulate* a draw X from a continuous distribution, based on looking at the graph of the PDF. To generate X, choose a uniformly random point under the PDF curve; this means that the probability of any region under the curve is the *area* of that region. Then let X be the x-coordinate of the random point. This is illustrated in Figure 5.4.

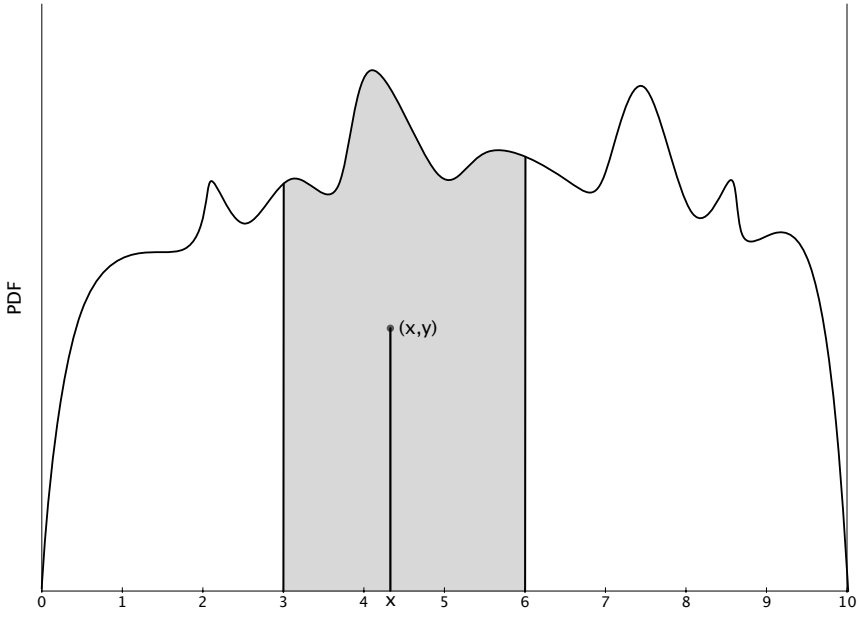


FIGURE 5.4
A complicated PDF (no numbers are shown on the vertical axis since the scale is whatever it needs to be to make the area under the curve 1). To generate an r.v. X with this PDF, choose a uniformly random point (x, y) under the curve and let $X = x$. This method works since, for example, X will be in the interval $[3, 6]$ if and only if the randomly chosen point (x, y) is in the shaded region.

Then X has the desired distribution since, by construction, $P(a \leq X \leq b)$ is the area under the PDF curve between the lines $x = a$ and $x = b$. Thinking about this method helps build intuition for PDFs, by giving us a feel for random variables sampled according to a particular PDF curve. □

The definition of expectation for continuous r.v.s is analogous to the definition for discrete r.v.s: replace the sum with an integral and the PMF with the PDF.

Definition 5.1.10 (Expectation of a continuous r.v.)**.** The *expected value* (also called the *expectation* or *mean*) of a continuous r.v. X with PDF f is

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

As in the discrete case, the expectation of a continuous r.v. may or may not exist. When discussing expectations, it would be very tedious to have to add "(if it exists)" after every mention of an expectation not yet shown to exist, so we will often leave this implicit.

The integral is taken over the entire real line, but if the support of X is not the entire real line we can just integrate over the support. The units in this definition make sense: if X is measured in centimeters, then so is $E(X)$, since $xf(x)dx$ has units of $\text{cm} \cdot \text{cm}^{-1} \cdot \text{cm} = \text{cm}$.

With this definition, the expected value retains its interpretation as a center of mass. As shown in Figure 5.5, using the Rayleigh PDF for illustrative purposes, the expected value is the balancing point of the PDF, just as it was the balancing point of the PMF in the discrete case.

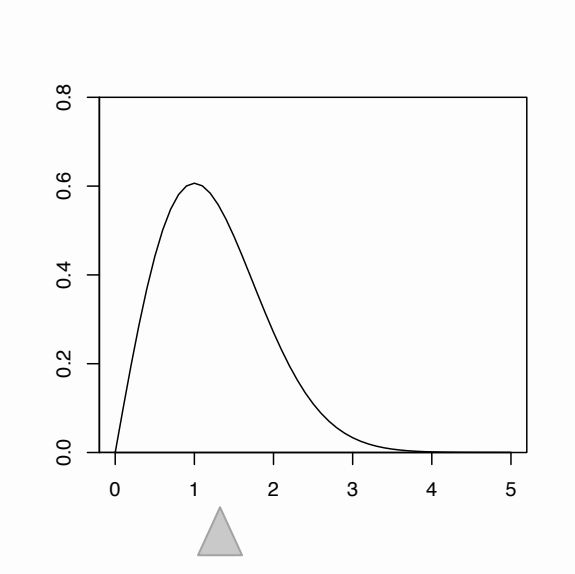


FIGURE 5.5
The expected value of a continuous r.v. is the balancing point of the PDF.

Linearity of expectation holds for continuous r.v.s, as it did for discrete r.v.s (we will show this later in Example 7.2.4). LOTUS also holds for continuous r.v.s, replacing the sum with an integral and the PMF with the PDF:

Theorem 5.1.11 (LOTUS, continuous)**.** If X is a continuous r.v. with PDF f and g is a function from $\mathbb{R}$ to $\mathbb{R}$, then

$$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

We now have all the tools we need to tackle the named distributions of this chapter, starting with the Uniform distribution.

5.2 Uniform

Intuitively, a Uniform r.v. on the interval (a, b) is a completely random number between a and b. We formalize the notion of "completely random" on an interval by specifying that the PDF should be *constant* over the interval.

Definition 5.2.1 (Uniform distribution)**.** A continuous r.v. U is said to have the *Uniform distribution* on the interval (a, b) if its PDF is

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a < x < b, \\ 0 & \text{otherwise.} \end{cases}$$

We denote this by $U \sim \text{Unif}(a, b)$.

This is a valid PDF because the area under the curve is just the area of a rectangle with width $b - a$ and height $1/(b - a)$. The CDF is the accumulated area under the PDF:

$$F(x) = \begin{cases} 0 & \text{if } x \leq a, \\ \frac{x-a}{b-a} & \text{if } a < x < b, \\ 1 & \text{if } x \geq b. \end{cases}$$

The Uniform distribution that we will most frequently use is the Unif$(0, 1)$ distribution, also called the standard Uniform. The Unif$(0, 1)$ PDF and CDF are particularly simple: $f(x) = 1$ and $F(x) = x$ for $0 < x < 1$. Figure 5.6 shows the Unif$(0, 1)$ PDF and CDF side by side.

For a general Unif(a, b) distribution, the PDF is constant on (a, b), and the CDF is ramp-shaped, increasing linearly from 0 to 1 as x ranges from a to b.

For Uniform distributions, *probability is proportional to length.*

Proposition 5.2.2. Let $U \sim \text{Unif}(a, b)$, and let (c, d) be a subinterval of (a, b), of length l (so $l = d - c$). Then the probability of U being in (c, d) is proportional to l. For example, a subinterval that is twice as long has twice the probability of containing U, and a subinterval of the same length has the same probability.

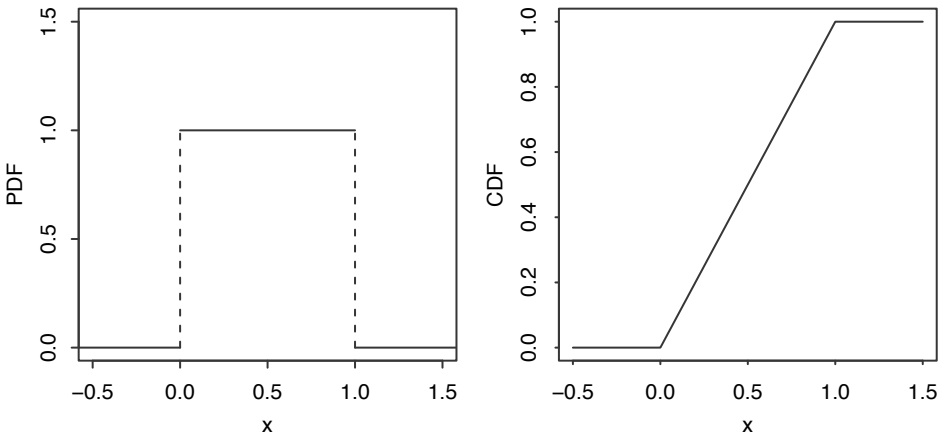


FIGURE 5.6
Unif$(0, 1)$ PDF and CDF.

Proof. Since the PDF of U is the constant $\frac{1}{b-a}$ on (a, b), the area under the PDF from c to d is $\frac{l}{b-a}$, which is a constant times l. ■

The above proposition is a very special property of the Uniform; for any other distribution, there are intervals of the same length that have different probabilities. Even after conditioning on a Uniform r.v. being in a certain subinterval, we *still* have a Uniform distribution and thus still have probability proportional to length (within that subinterval); we show this below.

Proposition 5.2.3. Let $U \sim \text{Unif}(a, b)$, and let (c, d) be a subinterval of (a, b). Then the conditional distribution of U given $U \in (c, d)$ is Unif(c, d).

Proof. For u in (c, d), the conditional CDF at u is

$$P\left(U \leq u | U \in (c, d)\right) = \frac{P(U \leq u, c < U < d)}{P(U \in (c, d))} = \frac{P(U \in (c, u])}{P(U \in (c, d))} = \frac{u - c}{d - c}.$$

The conditional CDF is 0 for $u \leq c$ and 1 for $u \geq d$. So the conditional distribution of U is as claimed. ■

Example 5.2.4. Let's illustrate the above propositions for $U \sim \text{Unif}(0, 1)$. In this special case, the support has length 1, so probability *is* length: the probability of U falling into the interval $(0, 0.3)$ is 0.3, as is the probability of falling into $(0.3, 0.6)$, $(0.4, 0.7)$, or any other interval of length 0.3 within $(0, 1)$.

Now suppose that we learn that $U \in (0.4, 0.7)$. Given this information, the conditional distribution of U is Unif$(0.4, 0.7)$. Then the conditional probability of $U \in (0.4, 0.6)$ is $2/3$, since $(0.4, 0.6)$ provides $2/3$ of the length of $(0.4, 0.7)$. The conditional probability of $U \in (0, 0.6)$ is also $2/3$, since we discard the points to the left of 0.4 when conditioning on $U \in (0.4, 0.7)$. □

Next, let's derive the mean and variance of $U \sim \text{Unif}(a, b)$. The expectation is extremely intuitive: the PDF is constant, so its balancing point should be the midpoint of (a, b). This is exactly what we find by using the definition of expectation for continuous r.v.s:

$$E(U) = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{1}{b-a}\left(\frac{b^2}{2} - \frac{a^2}{2}\right) = \frac{a+b}{2}.$$

For the variance, we first find $E(U^2)$ using the continuous version of LOTUS:

$$E(U^2) = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{3} \cdot \frac{b^3 - a^3}{b-a}.$$

Then

$$\text{Var}(U) = E(U^2) - (EU)^2 = \frac{1}{3} \cdot \frac{b^3 - a^3}{b-a} - \left(\frac{a+b}{2}\right)^2,$$

which reduces, after factoring $b^3 - a^3 = (b-a)(a^2+ab+b^2)$ and simplifying, to

$$\text{Var}(U) = \frac{(b-a)^2}{12}.$$

The above derivation isn't terribly painful, but there is an easier path, using a technique that is often useful for continuous distributions. The technique is called *location-scale transformation*, and it relies on the observation that shifting and scaling a Uniform r.v. produces another Uniform r.v. Shifting is considered a change of *location* and scaling is a change of *scale*, hence the term location-scale. For example, if X is Uniform on the interval $(1, 2)$, then $X + 5$ is Uniform on the interval $(6, 7)$, $2X$ is Uniform on the interval $(2, 4)$, and $2X + 5$ is Uniform on $(7, 9)$.

Definition 5.2.5 (Location-scale transformation)**.** Let X be a random variable and $Y = \sigma X + \mu$, where σ and μ are constants with $\sigma > 0$. Then we say that Y has been obtained as a *location-scale transformation* of X. Here μ controls how the location is changed and σ controls how the scale is changed.

☣ **5.2.6.** In a location-scale transformation, starting with $X \sim \text{Unif}(a, b)$ and transforming it to $Y = cX + d$ where c and d are constants with $c > 0$, Y is a *linear* function of X and Uniformity is preserved: $Y \sim \text{Unif}(ca + d, cb + d)$. But if Y is defined as a *nonlinear* transformation of X, then Y will *not* be Uniform in general. For example, for $X \sim \text{Unif}(a, b)$ with $0 \leq a < b$, the transformed r.v. $Y = X^2$ has support (a^2, b^2) but is *not* Uniform on that interval. Chapter 8 explores transformations of r.v.s in detail.

In studying Uniform distributions, a useful strategy is to start with an r.v. that has the simplest Uniform distribution, figure things out in the friendly simple case, and then use a location-scale transformation to handle the general case.

Let's see how this works for finding the expectation and variance of the $\text{Unif}(a, b)$

distribution. The location-scale strategy says to start with $U \sim \text{Unif}(0,1)$. Since the PDF of U is just 1 on the interval $(0,1)$, it is easy to see that

$$E(U) = \int_0^1 x dx = \frac{1}{2},$$
$$E(U^2) = \int_0^1 x^2 dx = \frac{1}{3},$$
$$\text{Var}(U) = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Now that we know the answers for U, transforming U into a general $\text{Unif}(a,b)$ r.v. takes just two steps. First we change the support from an interval of length 1 to an interval of length $b-a$, so we multiply U by the scaling factor $b-a$ to obtain a $\text{Unif}(0, b-a)$ r.v. Then we shift everything until the left endpoint of the support is at a. Thus, if $U \sim \text{Unif}(0,1)$, the random variable

$$\tilde{U} = a + (b-a)U$$

is distributed $\text{Unif}(a,b)$. Now the mean and variance of $\tilde{U}$ follow directly from properties of expectation and variance. By linearity of expectation,

$$E(\tilde{U}) = E(a + (b-a)U) = a + (b-a)E(U) = a + \frac{b-a}{2} = \frac{a+b}{2}.$$

By the fact that additive constants don't affect the variance while multiplicative constants come out squared,

$$\text{Var}(\tilde{U}) = \text{Var}(a + (b-a)U) = \text{Var}((b-a)U) = (b-a)^2\text{Var}(U) = \frac{(b-a)^2}{12}.$$

These agree with our previous answers.

The technique of location-scale transformation will work for any family of distributions such that shifting and scaling an r.v. whose distribution in the family produces another r.v. whose distribution is in the family. This technique does not apply to families of discrete distributions (with a fixed support) since, for example, shifting or scaling $X \sim \text{Bin}(n,p)$ changes the support and produces an r.v. that is no longer Binomial. A Binomial r.v. must be able to take on all integer values between 0 and some upper bound, but $X+4$ can't take on any value in $\{0,1,2,3\}$ and $2X$ can only take even values, so neither of these r.v.s has a Binomial distribution.

☣ **5.2.7** (Beware of sympathetic magic)**.** When using location-scale transformations, the shifting and scaling should be applied to the *random variables* themselves, not to their PDFs. To confuse these two would be an instance of sympathetic magic (see ☣ 3.7.7), and would result in invalid PDFs. For example, let $U \sim \text{Unif}(0,1)$, so the PDF f has $f(x) = 1$ on $(0,1)$ (and $f(x) = 0$ elsewhere). Then $3U + 1 \sim \text{Unif}(1,4)$, but $3f + 1$ is the function that equals 4 on $(0,1)$ and 1 elsewhere, which is not a valid PDF since it does not integrate to 1.

5.3 Universality of the Uniform

In this section, we will discuss a remarkable property of the Uniform distribution: given a Unif(0, 1) r.v., we can construct an r.v. with *any continuous distribution we want.* Conversely, given an r.v. with an arbitrary continuous distribution, we can create a Unif(0, 1) r.v. We call this the *universality of the Uniform*, because it tells us the Uniform is a universal starting point for building r.v.s with other distributions. Universality of the Uniform also goes by many other names, such as the *probability integral transform*, *inverse transform sampling*, the *quantile transformation*, and even the *fundamental theorem of simulation.*

To keep the proofs simple, we will state universality of the Uniform for a case where we know the inverse of the desired CDF exists. Similar ideas can be used to simulate a random draw from *any* desired CDF as a function of a Unif(0, 1) r.v.

Theorem 5.3.1 (Universality of the Uniform)**.** Let F be a CDF which is a continuous function and strictly increasing on the support of the distribution. This ensures that the inverse function F^{-1} exists, as a function from $(0, 1)$ to $\mathbb{R}$. We then have the following results.

1. Let $U \sim \text{Unif}(0, 1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.
2. Let X be an r.v. with CDF F. Then $F(X) \sim \text{Unif}(0, 1)$.

Let's make sure we understand what each part of the theorem is saying. The first part says that if we start with $U \sim \text{Unif}(0, 1)$ and a CDF F, then we can create an r.v. whose CDF is F by plugging U into the inverse CDF F^{-1}. Since F^{-1} is a function (known as the *quantile function*), U is an r.v., and a function of an r.v. is an r.v., $F^{-1}(U)$ is an r.v.; universality of the Uniform says its CDF is F.

The second part of the theorem goes in the reverse direction, starting from an r.v. X whose CDF is F and then creating a Unif(0, 1) r.v. Again, F is a function, X is an r.v., and a function of an r.v. is an r.v., so $F(X)$ is an r.v. Since any CDF is between 0 and 1 everywhere, $F(X)$ must take values between 0 and 1. Universality of the Uniform says that the distribution of $F(X)$ is Uniform on $(0, 1)$.

☣ **5.3.2.** The second part of universality of the Uniform involves plugging a random variable X *into its own CDF* F. This may seem strangely self-referential, but it makes sense because F is just a function (that satisfies the properties of a valid CDF), and a function of an r.v. is an r.v. There is a potential notational confusion, however: $F(x) = P(X \leq x)$ by definition, but it would be incorrect to say "$F(X) = P(X \leq X) = 1$". Rather, we should first find an expression for the CDF as a function of x, then replace x with X to obtain an r.v. For example, if the CDF of X is $F(x) = 1 - e^{-x}$ for $x > 0$, then $F(X) = 1 - e^{-X}$.

Understanding the statement of the theorem is the difficult part; the proof is just a couple of lines for each direction.

Proof.

1. Let $U \sim \text{Unif}(0,1)$ and $X = F^{-1}(U)$. For all real x,

$$P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x),$$

so the CDF of X is F, as claimed. For the last equality, we used the fact that $P(U \leq u) = u$ for $u \in (0,1)$.

2. Let X have CDF F, and find the CDF of $Y = F(X)$. Since Y takes values in $(0,1)$, $P(Y \leq y)$ equals 0 for $y \leq 0$ and equals 1 for $y \geq 1$. For $y \in (0,1)$,

$$P(Y \leq y) = P(F(X) \leq y) = P(X \leq F^{-1}(y)) = F(F^{-1}(y)) = y.$$

Thus Y has the $\text{Unif}(0,1)$ CDF. ■

To gain more insight into what the quantile function F^{-1} and universality of the Uniform mean, let's consider an example that is familiar to millions of students: percentiles on an exam.

Example 5.3.3 (Percentiles)**.** A large number of students take a certain exam, graded on a scale from 0 to 100. Let X be the score of a random student. Continuous distributions are easier to deal with here, so let's approximate the discrete distribution of scores using a continuous distribution. Suppose that X is continuous, with a CDF F that is strictly increasing on $(0, 100)$. In reality, there are only finitely many students and only finitely many possible scores, but a continuous distribution may be a good approximation.

Suppose that the median score on the exam is 60, i.e., half of the students score above 60 and the other half score below 60 (a convenient aspect of assuming a continuous distribution is that we don't need to worry about how many students had scores *equal* to 60). That is, $F(60) = 1/2$, or, equivalently, $F^{-1}(1/2) = 60$.

If Fred scores a 72 on the exam, then his *percentile* is the fraction of students who score below a 72. This is $F(72)$, which is some number in $(1/2, 1)$ since 72 is above the median. In general, a student with score x has percentile $F(x)$. Going the other way, if we start with a percentile, say 0.95, then $F^{-1}(0.95)$ is the score that has that percentile. A percentile is also called a *quantile*, which is why F^{-1} is called the quantile function. The function F converts scores to quantiles, and the function F^{-1} converts quantiles to scores.

The strange operation of plugging X into its own CDF now has a natural interpretation: $F(X)$ is the percentile attained by a random student. It often happens that the distribution of scores on an exam looks very non-Uniform. For example, there is no reason to think that 10% of the scores are between 70 and 80, even though $(70, 80)$ covers 10% of the range of possible scores.

On the other hand, the distribution of *percentiles* of the students *is* Uniform: the universality property says that $F(X) \sim \text{Unif}(0,1)$. For example, 50% of the students have a percentile of at least 0.5. Universality of the Uniform is expressing the fact that 10% of the students have a percentile between 0 and 0.1, 10% have a percentile between 0.1 and 0.2, 10% have a percentile between 0.2 and 0.3, and so on—a fact that is clear from the definition of percentile. □

To illustrate universality of the Uniform, we will apply it to the two distributions we encountered in the previous section, the Logistic and Rayleigh.

Example 5.3.4 (Universality with Logistic)**.** The Logistic CDF is

$$F(x) = \frac{e^x}{1+e^x}, \quad x \in \mathbb{R}.$$

Suppose we have $U \sim \text{Unif}(0,1)$ and wish to generate a Logistic r.v. Part 1 of the universality property says that $F^{-1}(U) \sim$ Logistic, so we first invert the CDF to get F^{-1}:

$$F^{-1}(u) = \log\left(\frac{u}{1-u}\right).$$

Then we plug in U for u:

$$F^{-1}(U) = \log\left(\frac{U}{1-U}\right).$$

Therefore $\log\left(\frac{U}{1-U}\right) \sim$ Logistic.

We can verify directly that $\log\left(\frac{U}{1-U}\right)$ has the required CDF: start from the definition of CDF, do some algebra to isolate U on one side of the inequality, and then use the CDF of the Uniform distribution. Let's work through these calculations once for practice:

$$\begin{aligned} P\left(\log\left(\frac{U}{1-U}\right) \leq x\right) &= P\left(\frac{U}{1-U} \leq e^x\right) \\ &= P\left(U \leq e^x(1-U)\right) \\ &= P\left(U \leq \frac{e^x}{1+e^x}\right) \\ &= \frac{e^x}{1+e^x}, \end{aligned}$$

which is indeed the Logistic CDF.

We can also use simulation to visualize how universality of the Uniform works. To this end, we generated 1 million Unif(0, 1) random variables. We then transformed each of these values u into $\log\left(\frac{u}{1-u}\right)$; if the universality of the Uniform is correct, the transformed numbers should follow a Logistic distribution.

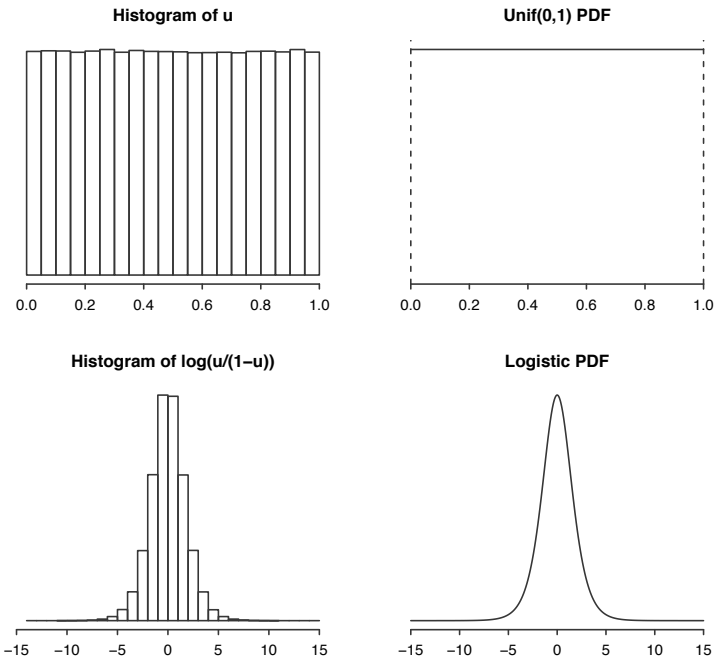


FIGURE 5.7
Top: Histogram of 10^6 draws of $U \sim \text{Unif}(0,1)$, with Unif$(0,1)$ PDF for comparison. Bottom: Histogram of 10^6 draws of $\log\left(\frac{U}{1-U}\right)$, with Logistic PDF for comparison.

Figure 5.7 displays a histogram of the realizations of U alongside the Unif$(0,1)$ PDF; below that, we have a histogram of the realizations of $\log\left(\frac{U}{1-U}\right)$ next to the Logistic PDF. As we can see, the second histogram looks very much like the Logistic PDF. Thus, by applying F^{-1}, we were able to transform our Uniform draws into Logistic draws, exactly as claimed by the universality of the Uniform.

Conversely, Part 2 of the universality property states that if $X \sim$ Logistic, then

$$F(X) = \frac{e^X}{1+e^X} \sim \text{Unif}(0,1). \qquad \square$$

Example 5.3.5 (Universality with Rayleigh)**.** The Rayleigh CDF is

$$F(x) = 1 - e^{-x^2/2}, \quad x > 0.$$

The quantile function (the inverse of the CDF) is

$$F^{-1}(u) = \sqrt{-2\log(1-u)},$$

so if $U \sim \text{Unif}(0,1)$, then $F^{-1}(U) = \sqrt{-2\log(1-U)} \sim$ Rayleigh.

We again generated 1 million realizations of $U \sim \text{Unif}(0,1)$ and transformed them

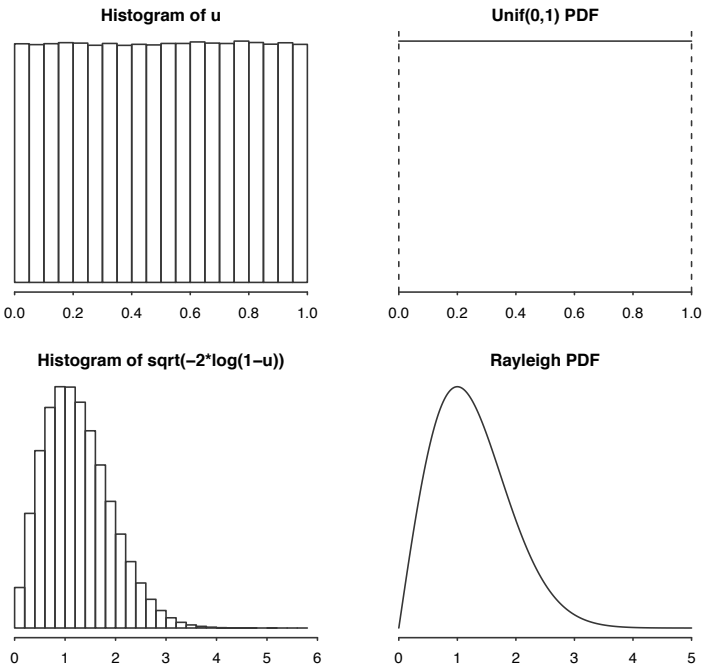


FIGURE 5.8
Top: Histogram of 1 million draws from $U \sim \text{Unif}(0,1)$, with Unif(0, 1) PDF for comparison. Bottom: Histogram of 1 million draws from $\sqrt{-2\log(1-U)}$, with Rayleigh PDF for comparison.

to produce 1 million realizations of $\sqrt{-2\log(1-U)}$. As Figure 5.8 shows, the realizations of $\sqrt{-2\log(1-U)}$ look very similar to the Rayleigh PDF, as predicted by the universality of the Uniform.

Conversely, if $X \sim$ Rayleigh, then $F(X) = 1 - e^{-X^2/2} \sim \text{Unif}(0,1)$. □

Next, let us consider the extent to which universality of the Uniform holds for discrete random variables. The CDF F of a discrete r.v. has jumps and flat regions, so F^{-1} does not exist (in the usual sense). But Part 1 still holds in the sense that given a Uniform random variable, we can construct an r.v. with *any discrete distribution we want*. The difference is that instead of working with the CDF, which is not invertible, it is more straightforward to work with the PMF.

Suppose we want to use $U \sim \text{Unif}(0,1)$ to construct a discrete r.v. X with PMF $p_j = P(X = j)$ for $j = 0, 1, 2, \ldots, n$. As illustrated in Figure 5.9, we can chop up the interval $(0,1)$ into pieces of lengths $p_0, p_1, \ldots, p_n$. By the properties of a valid PMF, the sum of the p_j's is 1, so this perfectly divides up the interval, without overshooting or undershooting.

Now define X to be the r.v. which equals 0 if U falls into the p_0 interval, 1 if U falls into the p_1 interval, 2 if U falls into the p_2 interval, and so on. Then X is a discrete

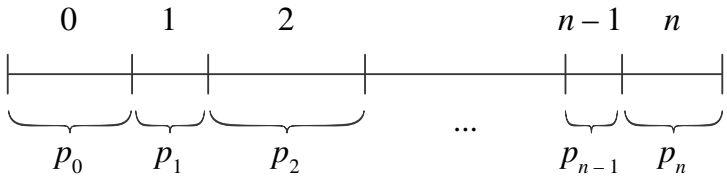


FIGURE 5.9
Given a PMF, chop up the interval $(0, 1)$ into pieces, with lengths given by the PMF values.

r.v. taking on values 0 through n. The probability that $X = j$ is the probability that U falls into the interval of length p_j. But for a Unif$(0, 1)$ r.v., probability *is* length, so $P(X = j)$ is precisely p_j, as desired!

The same trick will work for a discrete r.v. that can take on infinitely many values, such as a Poisson; we'll need to chop $(0, 1)$ into infinitely many pieces, but the total length of the pieces is still 1.

We now know how to take an arbitrary PMF and create an r.v. with that PMF. This fulfills our promise from Chapter 3 that any function with the properties given in Theorem 3.2.7 is the PMF of some r.v.

☣ **5.3.6.** Part 2 of universality of the Uniform, on the other hand, fails for discrete r.v.s. A function of a discrete r.v. is still discrete, so if X is discrete, then $F(X)$ is still discrete. So $F(X)$ doesn't have a Uniform distribution. For example, if $X \sim \text{Bern}(p)$, then $F(X)$ has only two possible values: $F(0) = 1 - p$ and $F(1) = 1$.

The upshot of universality is that we can use a Uniform r.v. U to generate r.v.s from both continuous and discrete distributions: in the continuous case, we can plug U into the inverse CDF, and in the discrete case, we can chop up the unit interval according to the desired PMF. Part 1 of universality of the Uniform is often useful in practice when running simulations (since the software being used may know how to generate Uniform r.v.s but not know how to generate r.v.s with the distribution of interest), though the extent to which it is useful depends on how tractable it is to compute the inverse CDF. Part 2 is important for certain widely used techniques in statistical inference, by providing a transformation that converts an r.v. with an *unknown* distribution to an r.v. with a *known*, simple distribution: the Uniform.

Using our analogy of distributions as blueprints and r.v.s as houses, the beauty of the universality property is that the Uniform distribution is a very simple blueprint, and it's easy to create a house from that blueprint; universality of the Uniform then gives us a simple rule for remodeling the Uniform house into a house with any other blueprint, no matter how complicated!

To conclude this section, we give an elegant identity that is often useful for finding the expectation of a nonnegative r.v. The identity also has a neat visual interpreta-

tion related to universality of the Uniform, LOTUS, and the relationship between CDFs and quantile functions.

Definition 5.3.7. The *survival function* of an r.v. X with CDF F is the function G given by $G(x) = 1 - F(x) = P(X > x)$.

Theorem 5.3.8 (Expectation by integrating the survival function)**.** Let X be a nonnegative r.v. Its expectation can be found by integrating its survival function:

$$E(X) = \int_0^\infty P(X > x)dx.$$

This result is the continuous analog of Theorem 4.4.8 (note though that it holds for *any* nonnegative r.v., not just for continuous nonnegative r.v.s). Actuaries sometimes call it the *Darth Vader rule*, for obscure reasons; statisticians are more likely to refer to it as finding the expectation by *integrating the survival function*.

Proof. For any number $x \geq 0$, we can write

$$x = \int_0^x dt = \int_0^\infty I(x > t)dt,$$

where $I(x > t)$ is 1 if $x \geq t$ and 0 otherwise. So

$$X(s) = \int_0^\infty I(X(s) > t)dt,$$

for each s in the sample space. We can write this more compactly as

$$X = \int_0^\infty I(X > t)dt.$$

Taking the expectation of both sides and swapping the E with the integral (which can be justified using results from real analysis), we have

$$E(X) = E\left(\int_0^\infty I(X > t)dt\right) = \int_0^\infty E(I(X > t))dt = \int_0^\infty P(X > t)dt. \quad \blacksquare$$

For a visual explanation of this identity, we can graph a CDF and interpret a certain area in two different ways: as the integral of the survival function, and as the integral of the quantile function.

A prototypical CDF of a nonnegative, continuous r.v. with CDF F is shown in Figure 5.10, with the area between the CDF curve and the horizontal line $p = 1$ shaded. This area can be found by integrating $1 - F(x)$, the difference between the line and the curve, from 0 to ∞.

But another way to find this area is to turn your head sideways and integrate with

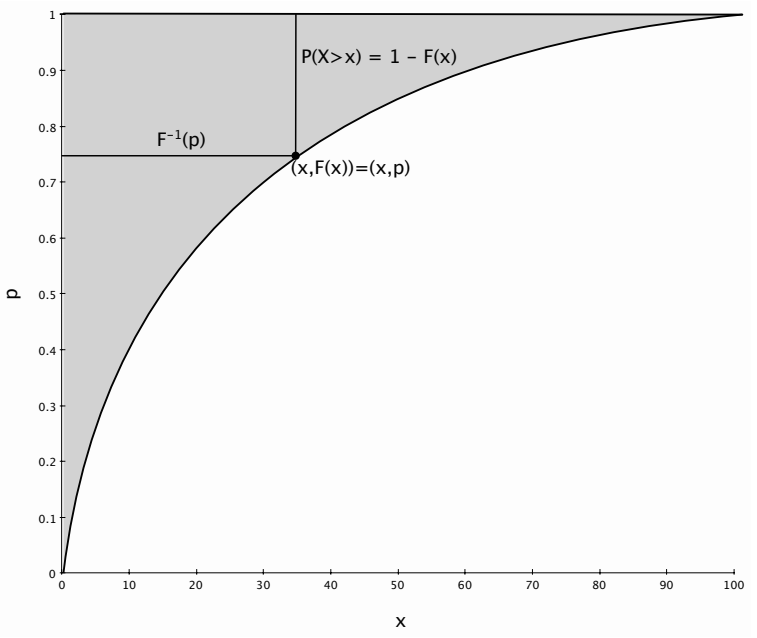


FIGURE 5.10
The area above a certain CDF and below the line $p = 1$ is shaded. This area can be interpreted in two ways: as the integral of the survival function, or as the integral of the quantile function.

respect to the vertical axis variable p rather than the horizontal axis variable x. This gives the integral of the quantile function, which, letting $U \sim \text{Unif}(0, 1)$, is

$$\int_0^1 F^{-1}(p)dp = E(F^{-1}(U)) = E(X),$$

by LOTUS and universality of the Uniform. So again we have

$$\int_0^\infty (1 - F(x))dx = \int_0^1 F^{-1}(p)dp = E(X).$$

5.4 Normal

The Normal distribution is a famous continuous distribution with a bell-shaped PDF. It is extremely widely used in statistics because of a theorem, the *central limit theorem*, which says that under very weak assumptions, the sum of a large number of i.i.d. random variables has an approximately Normal distribution, *regardless* of the distribution of the individual r.v.s. This means we can start with independent r.v.s from almost any distribution, discrete or continuous, but once we add up a bunch of them, the distribution of the resulting r.v. looks like a Normal distribution.

The central limit theorem is a topic for Chapter 10, but in the meantime, we'll introduce the properties of the Normal PDF and CDF and derive the expectation and variance of the Normal distribution. To do this, we will again use the strategy of location-scale transformation by starting with the simplest Normal distribution, the *standard Normal*, which is centered at 0 and has variance 1. After deriving the properties of the standard Normal, we'll be able to get to any Normal distribution we want by shifting and scaling.

Definition 5.4.1 (Standard Normal distribution)**.** A continuous r.v. Z is said to have the *standard Normal distribution* if its PDF φ is given by

$$\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

We write this as $Z \sim \mathcal{N}(0,1)$ since, as we will show, Z has mean 0 and variance 1.

The constant $\frac{1}{\sqrt{2\pi}}$ in front of the PDF may look surprising (why is something with π needed in front of something with e, when there are no circles in sight?), but it's exactly what is needed to make the PDF integrate to 1. Such constants are called *normalizing constants* because they normalize the total area under the PDF to 1. We'll verify soon that this is a valid PDF.

The standard Normal CDF Φ is the accumulated area under the PDF:

$$\Phi(z) = \int_{-\infty}^{z} \varphi(t)dt = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

Some people, upon seeing the function Φ for the first time, express dismay that it is left in terms of an integral. Unfortunately, we have little choice in the matter: it turns out to be mathematically impossible to find a closed-form expression for the antiderivative of φ, meaning that we cannot express Φ as a finite sum of more familiar functions like polynomials or exponentials. But closed-form or no, it's still a well-defined function: if we give Φ an input z, it returns the accumulated area under the PDF from $-\infty$ up to z.

Notation 5.4.2. We can tell the Normal distribution must be special because the standard Normal PDF and CDF get their own Greek letters. By convention, we use φ for the standard Normal PDF and Φ for the CDF. We will often use Z to denote a standard Normal random variable.

The standard Normal PDF and CDF are plotted in Figure 5.11. The PDF is bell-shaped and symmetric about 0, and the CDF is S-shaped. These have the same general shape as the Logistic PDF and CDF that we saw Example 5.1.6, but the Normal PDF decays to 0 much more quickly.

There are several important symmetry properties that can be deduced from the standard Normal PDF and CDF.

1. *Symmetry of PDF*: φ satisfies $\varphi(z) = \varphi(-z)$, i.e., φ is an even function.

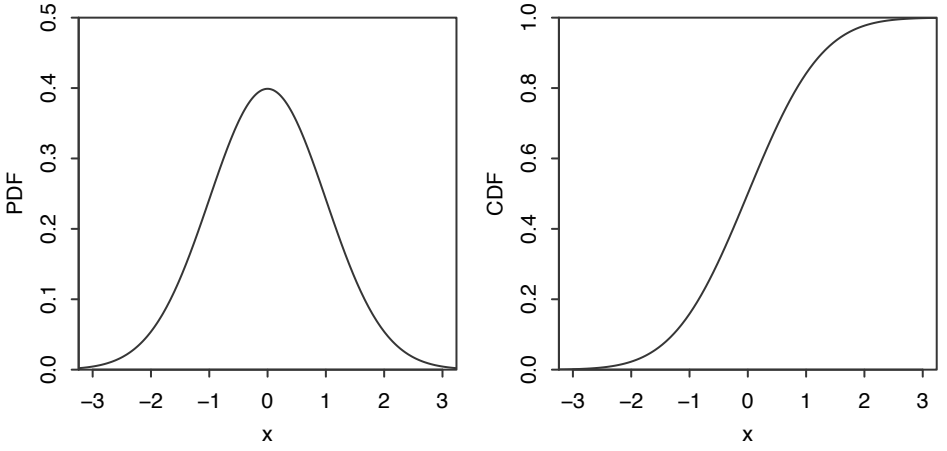


FIGURE 5.11
Standard Normal PDF φ (left) and CDF Φ (right).

2. *Symmetry of tail areas*: The area under the PDF curve to the left of -2, which is $P(Z \leq -2) = \Phi(-2)$ by definition, equals the area to the right of 2, which is $P(Z \geq 2) = 1 - \Phi(2)$. In general, we have

$$\Phi(z) = 1 - \Phi(-z)$$

for all z. This can be seen visually by looking at the PDF curve, and mathematically by substituting $u = -t$ below and using the fact that PDFs integrate to 1:

$$\Phi(-z) = \int_{-\infty}^{-z} \varphi(t)dt = \int_{z}^{\infty} \varphi(u)du = 1 - \int_{-\infty}^{z} \varphi(u)du = 1 - \Phi(z).$$

3. *Symmetry of Z and $-Z$*: If $Z \sim \mathcal{N}(0,1)$, then $-Z \sim \mathcal{N}(0,1)$ as well. To see this, note that the CDF of $-Z$ is

$$P(-Z \leq z) = P(Z \geq -z) = 1 - \Phi(-z),$$

but that is $\Phi(z)$, according to what we just argued. So $-Z$ has CDF Φ.

We need to prove three key facts about the standard Normal, and then we'll be ready to handle general Normal distributions: we need to show that φ is a valid PDF, that $E(Z) = 0$, and that $\text{Var}(Z) = 1$.

To verify the validity of φ, we'll show that the total area under $e^{-z^2/2}$ is $\sqrt{2\pi}$. However, we can't find the antiderivative of $e^{-z^2/2}$ directly, again because of the annoying fact that the antiderivative isn't expressible in closed form. But this doesn't mean we can't do *definite* integrals, with some ingenuity.

An amazing trick saves the day here: write down the integral *twice*. Usually, writing

down the same problem repeatedly is more a sign of frustration than a problem-solving strategy. But in this case, it allows a neat conversion to polar coordinates:

$$\begin{aligned}\left(\int_{-\infty}^{\infty} e^{-z^2/2}dz\right)\left(\int_{-\infty}^{\infty} e^{-z^2/2}dz\right) &= \left(\int_{-\infty}^{\infty} e^{-x^2/2}dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}dy\right) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}}dxdy \\ &= \int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}rdrd\theta.\end{aligned}$$

In the first step, we used the fact that z is just a dummy variable in each integral, so we are allowed to give it a different name (or two different names, one for each integral). The extra r that appears in the final step comes from the Jacobian of the transformation to polar coordinates, as explained in Section A.7.2 of the math appendix. That r is also what saves us from the impossibility of the original integral, since we can now use the substitution $u = r^2/2$, $du = rdr$. This gives

$$\begin{aligned}\int_0^{2\pi}\int_0^{\infty} e^{-r^2/2}rdrd\theta &= \int_0^{2\pi}\left(\int_0^{\infty} e^{-u}du\right)d\theta \\ &= \int_0^{2\pi} 1d\theta = 2\pi.\end{aligned}$$

Therefore,

$$\int_{-\infty}^{\infty} e^{-z^2/2}dz = \sqrt{2\pi},$$

as we wanted to show.

The expectation of the standard Normal has to be 0, by the symmetry of the PDF; no other balancing point would make sense. We can also see this symmetry by looking at the definition of $E(Z)$:

$$E(Z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ze^{-z^2/2}dz,$$

and since $g(z) = ze^{-z^2/2}$ is an odd function (see Section A.2.3 of the math appendix for more on even and odd functions), the area under g from $-\infty$ to 0 cancels the area under g from 0 to ∞. Therefore $E(Z) = 0$. In fact, the same argument shows that $E(Z^n) = 0$ for any odd positive integer n.[1]

Getting the mean was easy (one might even say it was EZ), but the variance

[1] A subtlety is that $\infty - \infty$ is undefined, so we also want to check that the area under the curve $z^n e^{-z^2/2}$ from 0 to ∞ is *finite*. But this is true since $e^{-z^2/2}$ goes to 0 extremely quickly (faster than exponential decay), more than offsetting the growth of the polynomial z^n.

calculation is a bit more involved. By LOTUS,

$$\begin{aligned}\text{Var}(Z) &= E(Z^2) - (EZ)^2 = E(Z^2) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz \\ &= \frac{2}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2} dz\end{aligned}$$

The last step uses the fact that $z^2 e^{-z^2/2}$ is an even function. Now we use integration by parts with $u = z$ and $dv = ze^{-z^2/2}dz$, so $du = dz$ and $v = -e^{-z^2/2}$:

$$\begin{aligned}\text{Var}(Z) &= \frac{2}{\sqrt{2\pi}} \left(-ze^{-z^2/2} \Big|_0^{\infty} + \int_0^{\infty} e^{-z^2/2} dz \right) \\ &= \frac{2}{\sqrt{2\pi}} \left(0 + \frac{\sqrt{2\pi}}{2} \right) \\ &= 1.\end{aligned}$$

The first term of the integration by parts equals 0 because $e^{-z^2/2}$ decays much faster than z grows, and the second term is $\sqrt{2\pi}/2$ because it's half of the total area under $e^{-z^2/2}$, which we've already proved is $\sqrt{2\pi}$. So indeed, the standard Normal distribution has mean 0 and variance 1.

The general Normal distribution has two parameters, denoted μ and σ^2, which correspond to the mean and variance (so the standard Normal is the special case where $\mu = 0$ and $\sigma^2 = 1$). Starting with a standard Normal r.v. $Z \sim \mathcal{N}(0, 1)$, we can get a Normal r.v. with any mean and variance by a location-scale transformation (shifting and scaling).

Definition 5.4.3 (Normal distribution)**.** If $Z \sim \mathcal{N}(0, 1)$, then

$$X = \mu + \sigma Z$$

is said to have the *Normal distribution* with mean μ and variance σ^2, for any real μ and σ^2 with $\sigma > 0$. We denote this by $X \sim \mathcal{N}(\mu, \sigma^2)$.

It's clear by properties of expectation and variance that X does in fact have mean μ and variance σ^2:

$$\begin{aligned}E(\mu + \sigma Z) &= E(\mu) + \sigma E(Z) = \mu, \\ \text{Var}(\mu + \sigma Z) &= \text{Var}(\sigma Z) = \sigma^2 \text{Var}(Z) = \sigma^2.\end{aligned}$$

Note that we multiply Z by the standard deviation σ, not σ^2; else the units would be wrong and X would have variance σ^4.

Of course, if we can get from Z to X, then we can get from X back to Z. The process of getting a standard Normal from a non-standard Normal is called, appropriately enough, *standardization*. For $X \sim \mathcal{N}(\mu, \sigma^2)$, the *standardized version* of X is

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1).$$

We can use standardization to find the CDF and PDF of X in terms of the standard Normal CDF and PDF.

Theorem 5.4.4 (Normal CDF and PDF)**.** Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then the CDF of X is

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right),$$

and the PDF of X is

$$f(x) = \varphi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}.$$

Proof. For the CDF, we start from the definition $F(x) = P(X \leq x)$, standardize, and use the CDF of the standard Normal:

$$F(x) = P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

Then we differentiate to get the PDF, remembering to apply the chain rule:

$$\begin{aligned} f(x) &= \frac{d}{dx}\Phi\left(\frac{x-\mu}{\sigma}\right) \\ &= \varphi\left(\frac{x-\mu}{\sigma}\right)\frac{1}{\sigma}. \end{aligned}$$

We can also write out the PDF as

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

■

Finally, three important benchmarks for the Normal distribution are the probabilities of falling within one, two, and three standard deviations of the mean. The 68-95-99.7% rule tells us that these probabilities are what the name suggests.

Theorem 5.4.5 (68-95-99.7% rule)**.** If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\begin{aligned} P(|X-\mu| < \sigma) &\approx 0.68, \\ P(|X-\mu| < 2\sigma) &\approx 0.95, \\ P(|X-\mu| < 3\sigma) &\approx 0.997. \end{aligned}$$

We can use this rule to get quick approximations of Normal probabilities.[2] Often it is easier to apply the rule after standardizing, in which case we have

$$\begin{aligned} P(|Z| < 1) &\approx 0.68, \\ P(|Z| < 2) &\approx 0.95, \\ P(|Z| < 3) &\approx 0.997. \end{aligned}$$

[2]The 68-95-99.7% rule says that 95% of the time, a Normal random variable will fall within ± 2 standard deviations of its mean. An even more accurate approximation says that 95% of the time, a Normal r.v. is within ± 1.96 SDs of its mean. This explains why the number 1.96 comes up very often in statistics in the context of 95% confidence intervals, which are often created by taking an estimate and putting a buffer zone of 1.96 SDs on either side.

Example 5.4.6 (Practice with the standard Normal CDF)**.** Let $X \sim \mathcal{N}(-1, 4)$. What is $P(|X| < 3)$, exactly (in terms of Φ) and approximately?

Solution:

The event $|X| < 3$ is the same as the event $-3 < X < 3$. We use standardization to express this event in terms of the standard Normal r.v. $Z = (X - (-1))/2$, then apply the 68-95-99.7% rule to get an approximation. The exact answer is

$$P(-3 < X < 3) = P\left(\frac{-3-(-1)}{2} < \frac{X-(-1)}{2} < \frac{3-(-1)}{2}\right) = P(-1 < Z < 2),$$

which is $\Phi(2) - \Phi(-1)$. The 68-95-99.7% rule tells us that $P(-1 < Z < 1) \approx 0.68$ and $P(-2 < Z < 2) \approx 0.95$. In other words, going from ± 1 standard deviation to ± 2 standard deviations adds approximately $0.95 - 0.68 = 0.27$ to the area under the curve. By symmetry, this is evenly divided between the areas $P(-2 < Z < -1)$ and $P(1 < Z < 2)$. Therefore,

$$P(-1 < Z < 2) = P(-1 < Z < 1) + P(1 < Z < 2) \approx 0.68 + \frac{0.27}{2} = 0.815.$$

This is close to the correct value, $\Phi(2) - \Phi(-1) \approx 0.8186$. □

As we will see later in the book, several important distributions can be obtained through transforming Normal r.v.s in natural ways, e.g., squaring or exponentiating. Chapter 8 delves into transformations in depth, but meanwhile there is a lot that we can do just using LOTUS and properties of CDFs.

Example 5.4.7 (Folded Normal)**.** Let $Y = |Z|$ with $Z \sim \mathcal{N}(0, 1)$. The distribution of Y is called a *Folded Normal* with parameters $\mu = 0$ and $\sigma^2 = 1$. In this example, we will derive the mean, variance, and distribution of Y. At first sight, Y may seem tricky to deal with since the absolute value function is not differentiable at 0 (due to its sharp corner), but Y has a perfectly valid continuous distribution.

(a) Find $E(Y)$.

(b) Find $\text{Var}(Y)$.

(c) Find the CDF and PDF of Y.

Solution:

(a) We will derive the PDF of Y later in this example, but to find $E(Y)$, LOTUS says we can work directly with the PDF of Z:

$$E(Y) = E|Z| = \int_{-\infty}^{\infty} |z| \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 2\int_{0}^{\infty} z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \sqrt{\frac{2}{\pi}}.$$

(b) Note that $Y^2 = Z^2$, so we do *not* need to do another integral! We have

$$E(Y^2) = E(Z^2) = \text{Var}(Z) = 1,$$

so

$$\text{Var}(Y) = E(Y^2) - (E(Y))^2 = 1 - \frac{2}{\pi}.$$

(c) For $y \leq 0$, the CDF of Y is $F_Y(y) = P(Y \leq y) = 0$. For $y > 0$, the CDF is

$$F_Y(y) = P(Y \leq y) = P(|Z| \leq y) = P(-y \leq Z \leq y) = \Phi(y) - \Phi(-y) = 2\Phi(y) - 1.$$

So the PDF of Y is $2\varphi(y)$ for $y \geq 0$, and 0 otherwise, where φ is the $\mathcal{N}(0,1)$ PDF.

Sanity check: Note that $2\Phi(y) - 1 \to 2 - 1 = 1$ as $y \to \infty$, as it must, and that the CDF of Y is a continuous function since Φ is continuous and at 0 there is not a jump: $2\Phi(0) - 1 = 0$. Also, the PDF of Y makes sense since taking the absolute value of Z "folds" the probability mass of a negative range of values of Z over to the positive side, e.g., the probability for Z values between -2 and -1 contributes to the probability for Y values between 1 and 2. This results in zero density for negative values and double the density for positive values. □

5.5 Exponential

The Exponential distribution is the continuous counterpart to the Geometric distribution. Recall that a Geometric random variable counts the number of failures before the first success in a sequence of Bernoulli trials. The story of the Exponential distribution is analogous, but we are now waiting for a success in *continuous* time, where successes arrive at a rate of λ successes per unit of time. The average number of successes in a time interval of length t is λt, though the actual number of successes varies randomly. An Exponential random variable represents the waiting time until the first arrival of a success.

Definition 5.5.1 (Exponential distribution)**.** A continuous r.v. X is said to have the *Exponential distribution* with parameter λ, where $\lambda > 0$, if its PDF is

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

We denote this by $X \sim \text{Expo}(\lambda)$.

The corresponding CDF is

$$F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

The Expo(1) PDF and CDF are plotted in Figure 5.12. Note the resemblance to the Geometric PMF and CDF pictured in Chapter 4. Exercise 43 explores the sense in which the Geometric converges to the Exponential, in the limit where the Bernoulli trials are performed faster and faster but with smaller and smaller success probabilities.

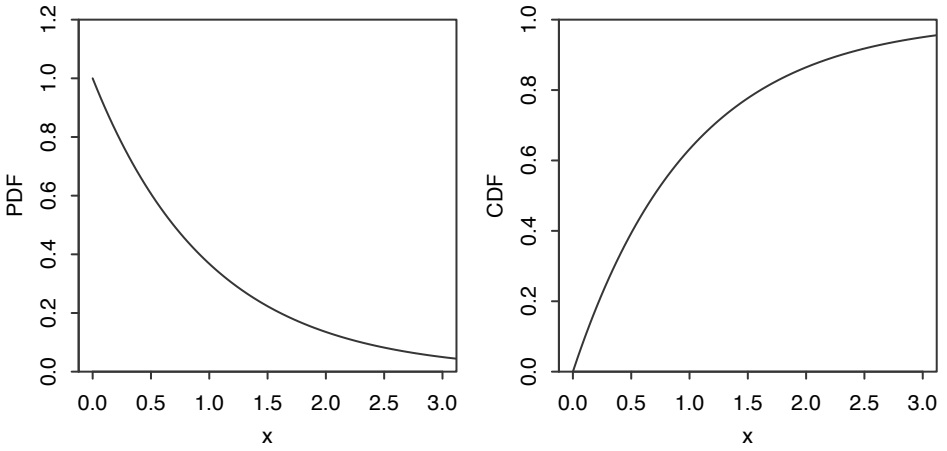


FIGURE 5.12
Expo(1) PDF and CDF.

We've seen how all Uniform and Normal distributions are related to one another via location-scale transformations, and we might wonder whether the Exponential distribution allows this too. Exponential r.v.s are defined to have support $(0, \infty)$, and shifting would change the left endpoint. But scale transformations work nicely, and we can use scaling to get from the simple Expo(1) to the general Expo(λ): if $X \sim \text{Expo}(1)$, then

$$Y = \frac{X}{\lambda} \sim \text{Expo}(\lambda),$$

since

$$P(Y \leq y) = P\left(\frac{X}{\lambda} \leq y\right) = P(X \leq \lambda y) = 1 - e^{-\lambda y}, \quad y > 0.$$

Conversely, if $Y \sim \text{Expo}(\lambda)$, then $\lambda Y \sim \text{Expo}(1)$.

This means that just as we did for the Uniform and the Normal, we can get the mean and variance of the Exponential distribution by starting with $X \sim \text{Expo}(1)$. Both $E(X)$ and $\text{Var}(X)$ are obtained using standard integration by parts calculations. This gives

$$E(X) = \int_0^\infty x e^{-x} dx = 1,$$
$$E(X^2) = \int_0^\infty x^2 e^{-x} dx = 2,$$
$$\text{Var}(X) = E(X^2) - (EX)^2 = 1.$$

In the next chapter we'll introduce a new tool called the moment generating function, which will let us get these results without integration.

For $Y = X/\lambda \sim \text{Expo}(\lambda)$ we then have

$$E(Y) = \frac{1}{\lambda}E(X) = \frac{1}{\lambda},$$
$$\text{Var}(Y) = \frac{1}{\lambda^2}\text{Var}(X) = \frac{1}{\lambda^2},$$

so the mean and variance of the $\text{Expo}(\lambda)$ distribution are $1/\lambda$ and $1/\lambda^2$, respectively. As we'd expect intuitively, the faster the rate of arrivals λ, the shorter the average waiting time.

The Exponential distribution has a very special property called the *memoryless property*, which says that even if you've waited for hours or days without success, the success isn't any more likely to arrive soon. In fact, you might as well have just started waiting 10 seconds ago. The definition formalizes this idea.

Definition 5.5.2 (Memoryless property)**.** A continuous distribution is said to have the *memoryless property* if a random variable X from that distribution satisfies

$$P(X \geq s + t | X \geq s) = P(X \geq t)$$

for all $s, t \geq 0$.

Here s represents the time you've already spent waiting; the definition says that after you've waited s minutes, the probability you'll have to wait another t minutes is exactly the same as the probability of having to wait t minutes with no previous waiting time under your belt. Another way to state the memoryless property is that conditional on $X \geq s$, the additional waiting time $X - s$ is still distributed $\text{Expo}(\lambda)$. In particular, this implies

$$E(X | X \geq s) = s + E(X) = s + \frac{1}{\lambda}.$$

(Conditional expectation is explained in detail in Chapter 9, but the meaning should already be clear: for any r.v. X and event A, $E(X|A)$ is the expected value of X given A; this can be defined by replacing the unconditional PMF or PDF of X in the definition of $E(X)$ by the conditional PMF or PDF of X given A.)

Using the definition of conditional probability, we can directly verify that the Exponential distribution has the memoryless property. Let $X \sim \text{Expo}(\lambda)$. Then

$$P(X \geq s + t | X \geq s) = \frac{P(X \geq s + t)}{P(X \geq s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = e^{-\lambda t} = P(X \geq t).$$

What are the implications of the memoryless property? If you're waiting at a bus stop and the time until the bus arrives has an Exponential distribution, then conditional on your having waited 30 minutes, the bus isn't due to arrive soon. The distribution simply *forgets* that you've been waiting for half an hour, and your remaining wait time is the same as if you had just shown up to the bus stop. If the

lifetime of a machine has an Exponential distribution, then no matter how long the machine has been functional, conditional on having lived that long, the machine is as good as new: there is no wear-and-tear effect that makes the machine more likely to break down soon. If human lifetimes were Exponential, then conditional on having survived to the age of 80, your remaining lifetime would have the same distribution as that of a newborn baby!

Clearly, the memoryless property is not an appropriate description for human or machine lifetimes. Why then do we care about the Exponential distribution?

1. Some physical phenomena, such as radioactive decay, truly do exhibit the memoryless property, so the Exponential is an important model in its own right.
2. The Exponential distribution is well-connected to other named distributions. In the next section, we'll see how the Exponential and Poisson distributions can be united by a shared story, and we'll discover many more connections in later chapters.
3. The Exponential serves as a building block for more flexible distributions, such as the *Weibull distribution* (introduced in Chapter 6), that allow for a wear-and-tear effect (where older units are due to break down) or a survival-of-the-fittest effect (where the longer you've lived, the stronger you get). To understand these distributions, we first have to understand the Exponential.

The memoryless property is a very special property of the Exponential distribution: no other continuous distribution on $(0, \infty)$ is memoryless! Let's prove this.

Theorem 5.5.3. If X is a positive continuous random variable with the memoryless property, then X has an Exponential distribution.

Proof. Suppose X is a positive continuous r.v. with the memoryless property. Let F be the CDF of X and G be the survival function of X, given by $G(x) = 1 - F(x)$. We will show that $G(x) = e^{-\lambda x}$ for some λ, by first showing that $G(xt) = G(t)^x$ for all real $x > 0$. The memoryless property says that

$$G(s + t) = G(s)G(t)$$

for all $s, t \geq 0$. Putting $s = t$, we have

$$G(2t) = G(t)^2,$$

so

$$G(3t) = G(2t + t) = G(2t)G(t) = G(t)^3, G(4t) = G(t)^4, \ldots.$$

We now have

$$G(mt) = G(t)^m$$

for m a positive integer. Let's extend this in stages. Replacing t by $t/2$ in $G(2t) = G(t)^2$, we have $G(t/2) = G(t)^{1/2}$. Similarly,

$$G\left(\frac{t}{n}\right) = G(t)^{1/n}$$

for any positive integer n. It follows that

$$G\left(\frac{m}{n}t\right) = \left(G\left(\frac{t}{n}\right)\right)^m = G(t)^{m/n}$$

for any positive integers m, n, so

$$G(xt) = G(t)^x$$

for all positive *rational* numbers x. Any positive real number can be written as a limit of positive rational numbers so, using the fact that G is a continuous function, the above equation holds for all positive *real* numbers x. Taking $t = 1$, we have

$$G(x) = G(1)^x = e^{-\lambda x},$$

where $\lambda = -\log(G(1)) > 0$. This is exactly the form we wanted for G, so X has an Exponential distribution. ■

The memoryless property is defined analogously for *discrete* distributions: a discrete distribution is memoryless if for X an r.v. with that distribution,

$$P(X \geq j + k | X \geq j) = P(X \geq k)$$

for all nonnegative integers j, k. In view of the analogy between the Geometric and Exponential stories (or if you have solved Exercise 32 from Chapter 4), you might guess that the Geometric distribution is memoryless. If so, you would be correct! If we're waiting for the first Heads in a sequence of fair coin tosses, and in a streak of bad luck we happen to get ten Tails in a row, this has no impact on how many additional tosses we'll need: the coin isn't due for a Heads, nor conspiring against us to perpetually land Tails. The coin is memoryless. The Geometric is the only memoryless discrete distribution taking values in $\{0, 1, 2, \dots\}$, and the Exponential is the only memoryless continuous distribution taking values in $(0, \infty)$.

As practice with the memoryless property, the following example chronicles the adventures of Fred, who experiences firsthand the frustrations of the memoryless property after moving to a town with a memoryless public transportation system.

Example 5.5.4 (Blissville and Blotchville)**.** Fred lives in Blissville, where buses always arrive exactly on time, with the time between successive buses fixed at 10 minutes. Having lost his watch, he arrives at the bus stop at a uniformly random time on a certain day (assume that buses run 24 hours a day, every day, and that the time that Fred arrives is independent of the bus arrival process).

(a) What is the distribution of how long Fred has to wait for the next bus? What is the average time that Fred has to wait?

(b) Given that the bus has not yet arrived after 6 minutes, what is the probability that Fred will have to wait at least 3 more minutes?

(c) Fred moves to Blotchville, a city with inferior urban planning and where buses are much more erratic. Now, when any bus arrives, the time until the next bus arrives is an Exponential random variable with mean 10 minutes. Fred arrives at the bus stop at a random time (assume that Blotchville has followed and will follow this system for all of eternity, and that the time that Fred arrives is independent of the bus arrival process). What is the distribution of Fred's waiting time for the next bus? What is the average time that Fred has to wait?

(d) When Fred complains to a friend how much worse transportation is in Blotchville, the friend says: "Stop whining so much! You arrive at a uniform instant between the previous bus arrival and the next bus arrival. The average length of that interval between buses is 10 minutes, but since you are equally likely to arrive at any time in that interval, your average waiting time is only 5 minutes."

Fred disagrees, both from experience and from solving Part (c) while waiting for the bus. Explain what is wrong with the friend's reasoning.

Solution:

(a) The distribution is Uniform on $(0, 10)$, so the mean is 5 minutes.

(b) Let T be the waiting time. Then

$$P(T \geq 6+3|T>6) = \frac{P(T \geq 9, T>6)}{P(T>6)} = \frac{P(T \geq 9)}{P(T>6)} = \frac{1/10}{4/10} = \frac{1}{4}.$$

In particular, Fred's waiting time in Blissville is not memoryless; conditional on having waited 6 minutes already, there's only a $1/4$ chance that he'll have to wait at least another 3 minutes, whereas if he had just showed up, there would be a $P(T \geq 3) = 7/10$ chance of having to wait at least 3 minutes.

(c) By the memoryless property, the distribution is Exponential with parameter $1/10$ (and mean 10 minutes) regardless of when Fred arrives; how much longer the next bus will take to arrive is independent of how long ago the previous bus arrived. The average time that Fred has to wait is 10 minutes.

(d) Fred's friend is making the mistake, explained in ☣ 4.1.3, of replacing a random variable (the time between buses) by its expectation (10 minutes), thereby ignoring the variability in interarrival times. The *average* length of a time interval between two buses is 10 minutes, but Fred is not equally likely to arrive at any of these intervals: Fred is more likely to arrive during a long interval between buses than to arrive during a short interval between buses. For example, if one interval between buses is 50 minutes and another interval is 5 minutes, then Fred is 10 times more likely to arrive during the 50-minute interval.

This phenomenon is known as *length-biased sampling*, and it comes up in many real-life situations. For example, asking randomly chosen mothers how many children they have yields a different distribution from asking randomly chosen people how many siblings they have, including themselves. Asking students the sizes of their classes and averaging those results may give a much higher value than taking a list of classes and averaging the sizes of each; this is called the *class size paradox*. See exercises 16 and 17 from Chapter 4 for more about the class size paradox and length-biased sampling.

Fred's adventures in Blissville and Blotchville continue in the exercises (see also MacKay [17] for more of Fred's adventures). The bus arrivals in Blotchville follow a *Poisson process*, which is the topic of the next section. □

5.6 Poisson processes

The Exponential distribution is closely connected to the Poisson distribution, as suggested by our use of λ for the parameters of both distributions. In this section we will see that the Exponential and Poisson are linked by a common story, which is the story of the *Poisson process*. A Poisson process is a sequence of arrivals occurring at different points on a timeline, such that the number of arrivals in a particular interval of time has a Poisson distribution. Poisson processes are discussed in much greater detail in Chapter 13, but we already have the tools to understand the definition and basic properties.

Definition 5.6.1 (Poisson process)**.** A process of arrivals in continuous time is called a *Poisson process* with rate λ if the following two conditions hold:

1. The number of arrivals that occur in an interval of length t is a $\text{Pois}(\lambda t)$ random variable.

2. The numbers of arrivals that occur in disjoint intervals are independent of each other. For example, the numbers of arrivals in the intervals $(0, 10), [10, 12)$, and $[15, \infty)$ are independent.

In this section, we will focus on Poisson processes on $(0, \infty)$, but we can also define Poisson processes on $(-\infty, \infty)$ or other intervals, and in Chapter 13 we will introduce Poisson processes in more than one dimension. A sketch of a Poisson process on $(0, \infty)$ is pictured in Figure 5.13. Each X marks the spot of an arrival.

For concreteness, suppose that the arrivals are emails landing in an inbox according to a Poisson process with rate λ. There are several things we might want to know about this process. One question we could ask is: in one hour, *how many* emails will arrive? The answer comes directly from the definition, which tells us that the

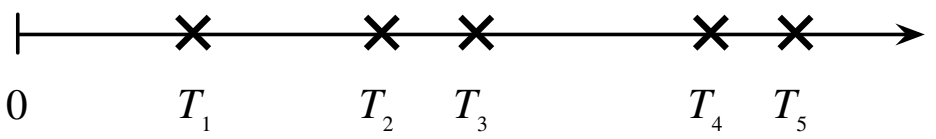


FIGURE 5.13
A Poisson process on $(0, \infty)$. Each X corresponds to an arrival.

number of emails in an hour follows a Pois(λ) distribution. Notice that the number of emails is a nonnegative integer, so a discrete distribution is appropriate.

But we could also flip the question around and ask: *how long* does it take until the first email arrives (measured relative to some fixed starting point)? The waiting time for the first email is a positive real number, so a continuous distribution on $(0, \infty)$ is appropriate. Let T_1 be the time until the first email arrives. To find the distribution of T_1, we just need to understand one crucial fact: saying that the waiting time for the first email is greater than t is the same as saying that *no emails* have arrived between 0 and t. In other words, if N_t is the number of emails that arrive at or before time t, then

$$T_1 > t \text{ is the same event as } N_t = 0.$$

We call this the *count-time duality* because it connects a discrete r.v., N_t, which *counts* the number of arrivals, with a continuous r.v., T_1, which marks the *time* of the first arrival. More generally, the count-time duality says that

$$T_n > t \text{ is the same event as } N_t < n.$$

Saying that the nth arrival has not happened yet as of time t is equivalent to saying that, up until time t, there have been fewer than n arrivals.

If two events are the same, they have the same probability. Since $N_t \sim \text{Pois}(\lambda t)$ by the definition of Poisson process,

$$P(T_1 > t) = P(N_t = 0) = \frac{e^{-\lambda t}(\lambda t)^0}{0!} = e^{-\lambda t}.$$

Therefore $P(T_1 \leq t) = 1 - e^{-\lambda t}$, so $T_1 \sim \text{Expo}(\lambda)$! The time until the first arrival in a Poisson process of rate λ has an Exponential distribution with parameter λ.

What about $T_2 - T_1$, the time between the first and second arrivals? Since disjoint intervals in a Poisson process are independent by definition, the past is irrelevant once the first arrival occurs. Thus $T_2 - T_1$ is independent of the time until the first arrival, and by the same argument as before, $T_2 - T_1$ also has an Exponential distribution with rate λ.

Similarly, $T_3 - T_2 \sim \text{Expo}(\lambda)$ independently of T_1 and $T_2 - T_1$. Continuing in this way, we deduce that all the interarrival times are i.i.d. Expo(λ) random variables. Thus, Poisson processes tie together two important distributions, one discrete and one

continuous, and the use of a common symbol λ for both the Poisson and Exponential parameters is felicitous notation, for λ is the arrival rate in the process that unites the two distributions.

☣ **5.6.2.** The total time until the second arrival, T_2, is the sum of two independent Expo(λ) r.v.s, T_1 and $T_2 - T_1$. This does *not* have an Exponential distribution, but rather a Gamma distribution, which is introduced in Chapter 8.

The story of the Poisson process provides intuition for the fact, shown below, that the minimum of independent Exponential r.v.s is another Exponential r.v.

Example 5.6.3 (Minimum of independent Expos)**.** Let $X_1, \ldots, X_n$ be independent, with $X_j \sim \text{Expo}(\lambda_j)$. Let $L = \min(X_1, \ldots, X_n)$. Show that $L \sim \text{Expo}(\lambda_1 + \cdots + \lambda_n)$, and interpret this intuitively.

Solution:

We can find the distribution of L by considering its *survival function* $P(L > t)$, since the survival function is 1 minus the CDF.

$$\begin{aligned} P(L > t) &= P(\min(X_1, \ldots, X_n) > t) = P(X_1 > t, \ldots, X_n > t) \\ &= P(X_1 > t) \cdots P(X_n > t) = e^{-\lambda_1 t} \cdots e^{-\lambda_n t} = e^{-(\lambda_1 + \cdots + \lambda_n)t}. \end{aligned}$$

The second equality holds since saying that the minimum of the X_j is greater than t is the same as saying that all of the X_j are greater than t. The third equality holds by independence of the X_j. Thus, L has the survival function (and the CDF) of an Exponential distribution with parameter $\lambda_1 + \cdots + \lambda_n$.

Intuitively, we can interpret the λ_j as the rates of n independent Poisson processes. We can imagine, for example, X_1 as the waiting time for a green car to pass by, X_2 as the waiting time for a blue car to pass by, and so on, assigning a color to each X_j. Then L is the waiting time for a car with any of these colors to pass by, so it makes sense that L has a combined rate of $\lambda_1 + \cdots + \lambda_n$. □

☣ **5.6.4.** The minimum of independent Exponentials is Exponential, but the maximum of independent Exponentials is *not* Exponential. However, the result about such a minimum turns out to be useful in studying such a maximum, as illustrated in the next two examples.

Example 5.6.5 (Maximum of 3 independent Exponentials)**.** Three students are working independently on their probability homework. All 3 start at 1 pm on a certain day, and each takes an Exponential time with mean 6 hours to complete the homework. What is the earliest time at which all 3 students will have completed the homework, on average?

Solution: Label the students as $1, 2, 3$, and let X_j be how long it takes student j to finish the homework. Let $\lambda = 1/6$, and let T be the time when all 3 students will have completed the homework, so $T = \max(X_1, X_2, X_3)$ with $X_i \sim \text{Expo}(\lambda)$. The CDF of T is

$$P(T \leq t) = P(X_1 \leq t, X_2 \leq t, X_3 \leq t) = (1 - e^{-\lambda t})^3.$$

So the PDF of T is

$$f_T(t) = 3\lambda e^{-\lambda t}(1 - e^{-\lambda t})^2.$$

In particular, T is *not* Exponential.

Finding $E(T)$ by integrating $tf_T(t)$ is possible but not especially pleasant. A neater approach is to use the memoryless property and the fact that the minimum of independent Exponentials *is* Exponential. We can decompose

$$T = T_1 + T_2 + T_3,$$

where $T_1 = \min(X_1, X_2, X_3)$ is how long it takes for one student to complete the homework, T_2 is the additional time it takes for a second student to complete the homework, and T_3 is the additional time until all 3 have completed the homework. Then $T_1 \sim \text{Expo}(3\lambda)$, by the result of Example 5.6.3.

By the memoryless property, at the first time when a student completes the homework the other two students are starting from fresh, so $T_2 \sim \text{Expo}(2\lambda)$. Again by the memoryless property, $T_3 \sim \text{Expo}(\lambda)$. The memoryless property also implies that T_1, T_2, T_3 are independent (which would be very useful if we were finding $\text{Var}(T)$). By linearity,

$$E(T) = \frac{1}{3\lambda} + \frac{1}{2\lambda} + \frac{1}{\lambda} = 2 + 3 + 6 = 11,$$

which shows that on average, the 3 students will have all completed the homework at midnight, 11 hours after they started. □

Example 5.6.6 (Machine repair)**.** A certain machine often breaks down and needs to be fixed. At time 0, the machine is working. It works for an $\text{Expo}(\lambda)$ period of time (measured in days), and then breaks down. It then takes an $\text{Expo}(\lambda)$ amount of time to get it fixed, after which it will work for an $\text{Expo}(\lambda)$ time until it breaks down again, after which it will take an $\text{Expo}(\lambda)$ time to get it fixed, etc. Assume that these $\text{Expo}(\lambda)$ r.v.s are i.i.d.

(a) A *transition* occurs when the machine switches from working to being broken, or switches from being broken to working. Find the distribution of the number of transitions that occur in the time interval $(0, t)$.

(b) Hoping to reduce the frequency of breakdowns, the machine is redesigned so that it can continue to function even if one component has failed. The redesigned machine has 5 components, each of which works for an $\text{Expo}(\lambda)$ amount of time and then fails, independently. The machine works properly if and only if at most one component has failed. Currently, all 5 components are working (none have failed). Find the expected time until the machine breaks down.

Solution:

(a) The times between transitions are i.i.d. $\text{Expo}(\lambda)$, so the times at which transitions occur follow a Poisson process of rate λ. So the desired distribution is $\text{Pois}(\lambda t)$.

(b) The time until a component fails is $\text{Expo}(5\lambda)$. Then by the memoryless property,

the additional time until another component fails is Expo(4λ). So the expected time until the machine breaks down is

$$\frac{1}{5\lambda} + \frac{1}{4\lambda} = \frac{9}{20\lambda}.$$

□

5.7 Symmetry of i.i.d. continuous r.v.s

Continuous r.v.s that are independent and identically distributed have an important symmetry property: all possible orderings are equally likely. Intuitively, this is because if all we are told is that $X_1, \ldots, X_n$ are i.i.d., then they are interchangeable, in the sense that we have been given no information that distinguishes one X_i from another X_j. This is reminiscent of the fact that it is common for someone to say "Do you want the good news first or the bad news first?" but rare for someone to say "I have two pieces of news. Which do you want to hear first?", since in the latter case no distinguishing information has been provided for the two pieces of news.

Proposition 5.7.1. Let $X_1, \ldots, X_n$ be i.i.d. from a continuous distribution. Then

$$P(X_{a_1} < X_{a_2} < \cdots < X_{a_n}) = \frac{1}{n!}$$

for any permutation $a_1, a_2, \ldots, a_n$ of $1, 2, \ldots, n$.

Proof. Let F be the CDF of X_j. By symmetry, all orderings of $X_1, \ldots, X_n$ are equally likely. For example, $P(X_3 < X_2 < X_1) = P(X_1 < X_2 < X_3)$ since both sides have exactly the same structure: they are both of the form $P(A < B < C)$ where A, B, C are i.i.d. draws from F. For any i and j with $i \neq j$, the probability of the *tie* $X_i = X_j$ is 0 since X_i and X_j are independent continuous r.v.s. So the probability of there being at least one tie among $X_1, \ldots, X_n$ is also 0, since

$$P\left(\bigcup_{i \neq j} \{X_i = X_j\}\right) \leq \sum_{i \neq j} P(X_i = X_j) = 0.$$

Thus, $X_1, \ldots, X_n$ are distinct with probability 1, and the probability of any particular ordering is $1/n!$. ■

☣ **5.7.2.** This proposition may fail if the r.v.s are dependent. Let $n = 2$, and consider the extreme case where X_1 and X_2 are so dependent that they are always equal, i.e., $X_1 = X_2$ with probability 1. Then $P(X_1 < X_2) = P(X_2 < X_1) = 0$. For dependent X_1, X_2 we can also make $P(X_1 < X_2) \neq P(X_2 < X_1)$. For an example, see Exercise 42 from Chapter 3.

☣ **5.7.3.** If X and Y are i.i.d. *continuous* r.v.s, then

$$P(X < Y) = P(Y < X) = \frac{1}{2},$$

by symmetry and since the probability of a tie is 0. In contrast, if X and Y are i.i.d. *discrete* r.v.s, it is still true that $P(X < Y) = P(Y < X)$ by symmetry, but this number is less than $1/2$ because of the possibility of a tie. For example, if X and Y are i.i.d. nonnegative integer-valued r.v.s with $P(X = j) = c_j$, then

$$1 = P(X < Y) + P(X = Y) + P(Y < X) = 2P(X < Y) + P(X = Y),$$

so

$$P(X < Y) = \frac{1}{2} \cdot (1 - P(X = Y)) = \frac{1}{2} \cdot \left(1 - \sum_{j=0}^{\infty} c_j^2\right) < \frac{1}{2}.$$

The *ranks* of a list of distinct numbers are defined by giving the smallest number a rank of 1, the second smallest a rank of 2, and so on. For example, the ranks for $3.14, 2.72, 1.41, 1.62$ are $4, 3, 1, 2$. Proposition 5.7.1 says that the ranks of i.i.d. continuous $X_1, \ldots, X_n$ are a uniformly random permutation of the numbers $1, \ldots, n$. The next example shows how we can use this symmetry property in conjunction with indicator r.v.s in problems involving *records*, such as the record level of rainfall or the record performance on a high jump.

Example 5.7.4 (Records)**.** Athletes compete one at a time at the high jump. Let X_j be how high the jth jumper jumped, with $X_1, X_2, \ldots$ i.i.d. with a continuous distribution. We say that the jth jumper sets a *record* if X_j is greater than all of $X_{j-1}, \ldots, X_1$.

(a) Is the event "the 110th jumper sets a record" independent of the event "the 111th jumper sets a record"?

(b) Find the mean number of records among the first n jumpers. What happens to the mean as $n \to \infty$?

(c) A *double record* occurs at time j if *both* the jth and $(j-1)$st jumpers set records. Find the mean number of double records among the first n jumpers. What happens to the mean as $n \to \infty$?

Solution:

(a) Let I_j be the indicator r.v. for the jth jumper setting a record. By symmetry, $P(I_j = 1) = 1/j$ (as any of the first j jumps is equally likely to be the highest of those jumps). Also,

$$P(I_{110} = 1, I_{111} = 1) = \frac{109!}{111!} = \frac{1}{110 \cdot 111},$$

since in order for both the 110th and 111th jumps to be records, we need the highest of the first 111 jumps to be in position 111 and the second highest to be in position 110, and the remaining 109 can be in any order. So

$$P(I_{110} = 1, I_{111} = 1) = P(I_{110} = 1)P(I_{111} = 1),$$

which shows that the 110th jumper setting a record is independent of the 111th jumper setting a record. Intuitively, this makes sense since learning that the 111th jumper sets a record gives us no information about the "internal" matter of how the first 110 jumps are arranged amongst themselves.

(b) By linearity, the expected number of records among the first n jumpers is $\sum_{j=1}^{n} \frac{1}{j}$, which goes to ∞ as $n \to \infty$ since the harmonic series diverges.

(c) Let J_j be the indicator r.v. for a double record occurring at time j, for $2 \leq j \leq n$. Then $P(J_j = 1) = \frac{1}{j(j-1)}$, following the logic of Part (a). So the expected number of double records is

$$\sum_{j=2}^{n} \frac{1}{j(j-1)} = \sum_{j=2}^{n} \left(\frac{1}{j-1} - \frac{1}{j} \right) = 1 - \frac{1}{n},$$

since all the other terms cancel out. Thus, the expected number of records goes to ∞ as $n \to \infty$, but the expected number of double records goes to 1. □

5.8 Recap

A continuous r.v. can take on any value in an interval, although the probability that it equals any particular value is 0. The CDF of a continuous r.v. is differentiable, and the derivative is called the probability density function (PDF). Probability is given by area under the PDF curve, *not* by the value of the PDF at a point. We must integrate the PDF to get a probability. The table below summarizes and compares some important concepts in the discrete case and the continuous case.

	Discrete r.v.	**Continuous r.v.**
CDF	$F(x) = P(X \leq x)$	$F(x) = P(X \leq x)$
PMF/PDF	$P(X = x)$	$f(x) = F'(x)$
	• PMF is height of jump of F at x.	• PDF is derivative of F.
	• PMF is nonnegative.	• PDF is nonnegative.
	• PMF sums to 1.	• PDF integrates to 1.
	• $P(X \in A) = \sum_{x \in A} P(X = x)$.	• $P(X \in A) = \int_A f(x)dx$.
Expectation	$E(X) = \sum_x xP(X = x)$	$E(X) = \int_{-\infty}^{\infty} xf(x)dx$
LOTUS	$E(g(X)) = \sum_x g(x)P(X = x)$	$E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)dx$

Three important continuous distributions are the Uniform, Normal, and Exponential. A Unif(a, b) r.v. is a "completely random" number in the interval (a, b), and it has the property that probability is proportional to length. The universality of the Uniform tells us how we can use a Unif$(0, 1)$ r.v. to construct r.v.s from other distributions we may be interested in; it also says that if we plug a continuous r.v. into its own CDF, the resulting r.v. has a Unif$(0, 1)$ distribution.

A $\mathcal{N}(\mu, \sigma^2)$ r.v. has a symmetric bell-shaped PDF centered at μ, with σ controlling how spread out the curve is. The mean is μ and standard deviation is σ. The 68-95-99.7% rule gives important benchmarks for the probability of a Normal r.v. falling within 1, 2, and 3 standard deviations of its mean.

An Expo(λ) r.v. represents the waiting time for the first success in continuous time, analogous to how a Geometric r.v. represents the number of failures before the first success in discrete time; the parameter λ can be interpreted as the rate at which successes arrive. The Exponential distribution has the memoryless property, which says that conditional on our having waited a certain amount of time without success, the distribution of the remaining wait time is exactly the same as if we hadn't waited at all. In fact, the Exponential is the *only* positive continuous distribution with the memoryless property.

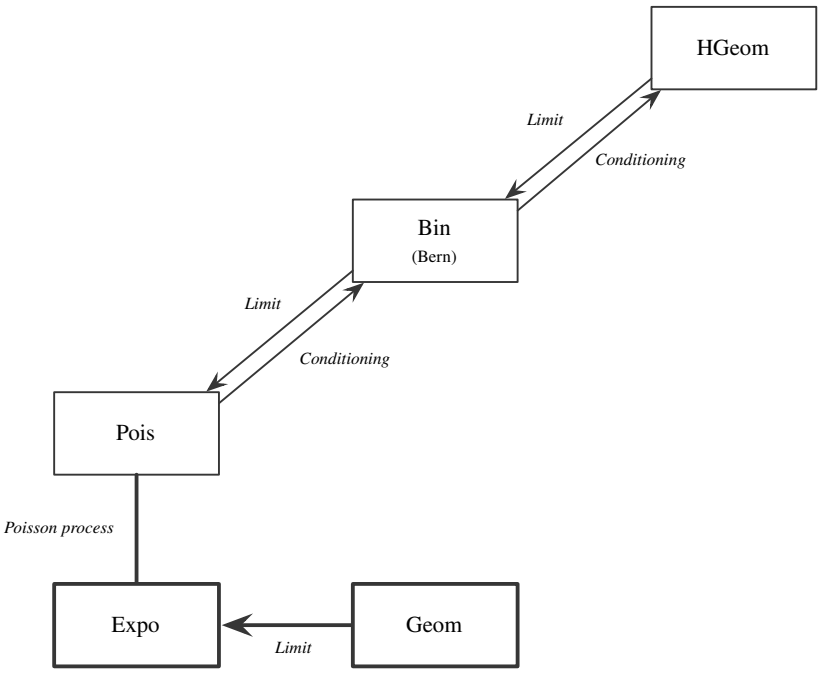
A Poisson process is a sequence of arrivals in continuous time such that the number of arrivals in an interval is Poisson (with mean proportional to the length of the interval) and disjoint intervals have independent numbers of arrivals. The interarrival times in a Poisson process of rate λ are i.i.d. Expo(λ) r.v.s.

A useful symmetry property of i.i.d. r.v.s $X_1, X_2, \ldots, X_n$ is that all orderings are equally likely. For example, $P(X_1 < X_2 < X_3) = P(X_3 < X_2 < X_1)$. If the X_j are continuous in addition to being i.i.d., then we can also conclude, e.g., that $P(X_1 < X_2 < X_3) = 1/6$, whereas in the discrete case we also have to account for the possibility of ties.

A new strategy that we learned for continuous distributions is location-scale transformation, which says that if shifting and scaling will not take us outside the family of distributions we're studying, then we can start with the simplest member of the family, find the answer for the simple case, then use shifting and scaling to arrive at the general case. For the three main distributions of this chapter, this approach works as follows.

- Uniform: If $U \sim \text{Unif}(0, 1)$, then $\tilde{U} = a + (b - a)U \sim \text{Unif}(a, b)$.
- Normal: If $Z \sim \mathcal{N}(0, 1)$, then $X = \mu + \sigma Z \sim \mathcal{N}(\mu, \sigma^2)$.
- Exponential: If $X \sim \text{Expo}(1)$, then $Y = X/\lambda \sim \text{Expo}(\lambda)$. We do not consider shifts here since a nonzero shift would prevent the support from being $(0, \infty)$.

We can now add the Exponential and Geometric distributions to our diagram of connections between distributions: the Exponential is a continuous limit of the Geometric, and the Poisson and Exponential are connected by the Poisson process.



And in our map of the four fundamental objects in probability, we add the PDF as another blueprint for continuous random variables.

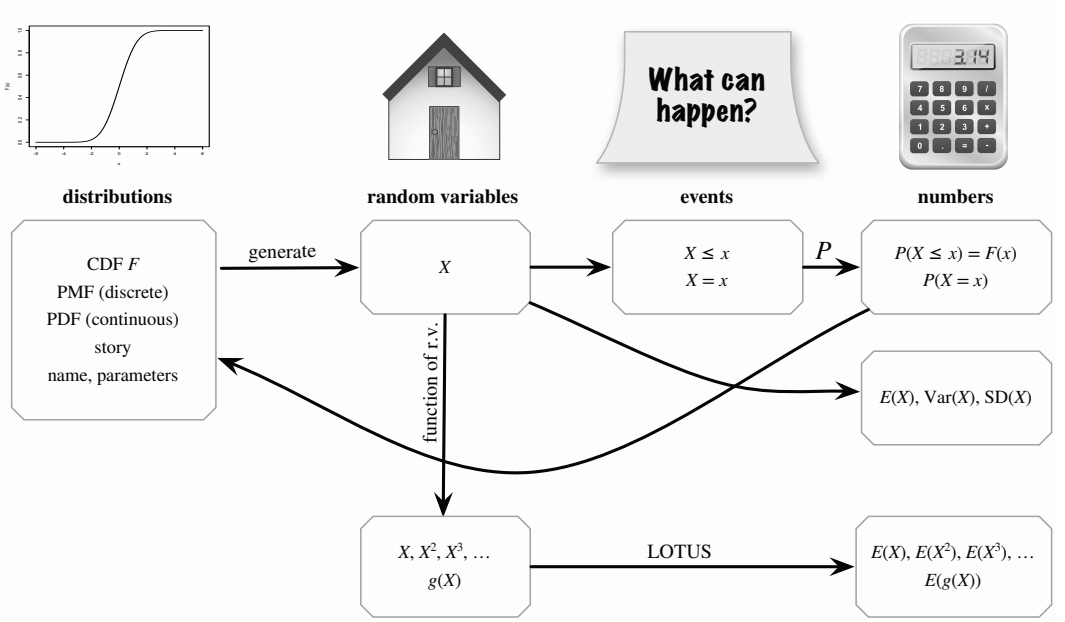


FIGURE 5.14
Four fundamental objects in probability: distributions, random variables, events, and numbers. For a continuous r.v. X, we have $P(X = x) = 0$, so we use the PDF as a blueprint in place of the PMF.