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.

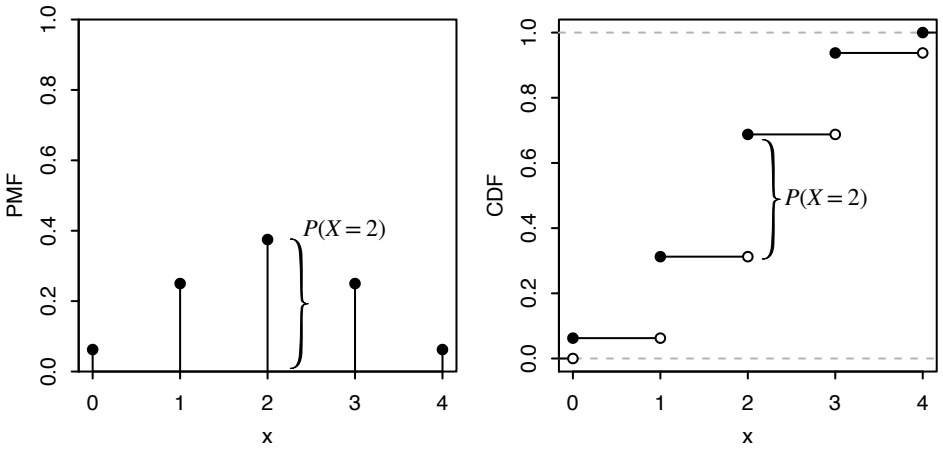

FIGURE 3.8
Bin$(4, 1/2)$ PMF and CDF. The height of the vertical bar $P(X = 2)$ in the PMF is also the height of the jump in the CDF at 2.

- *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

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

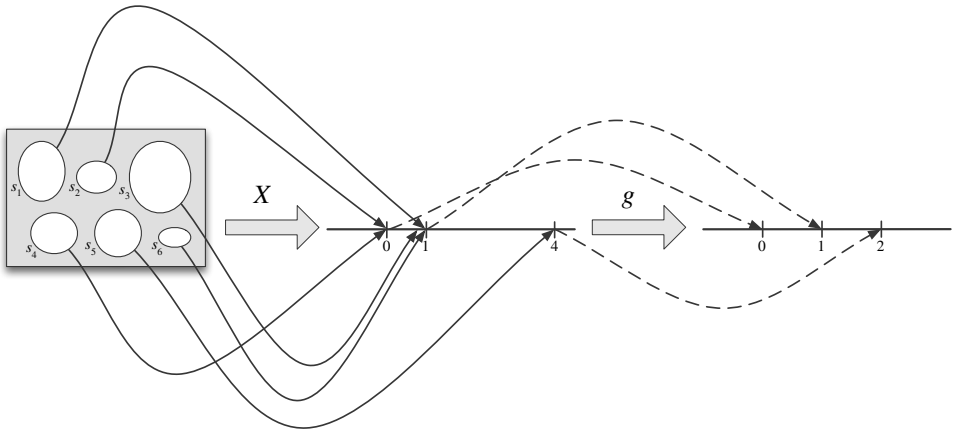

FIGURE 3.9
The r.v. X is defined on a sample space with 6 elements, and has possible values 0, 1, and 4. The function g is the square root function. Composing X and g gives the random variable $g(X) = \sqrt{X}$, which has possible values 0, 1, and 2.

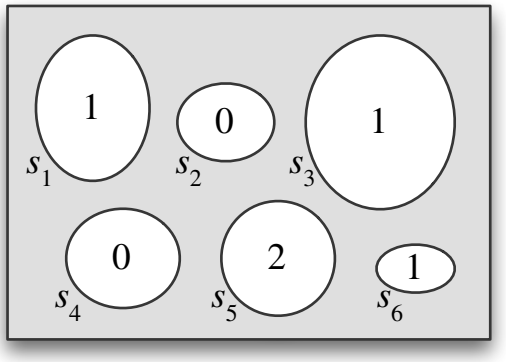

FIGURE 3.10
Since $g(X) = \sqrt{X}$ labels each pebble with a number, it is an r.v.

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.

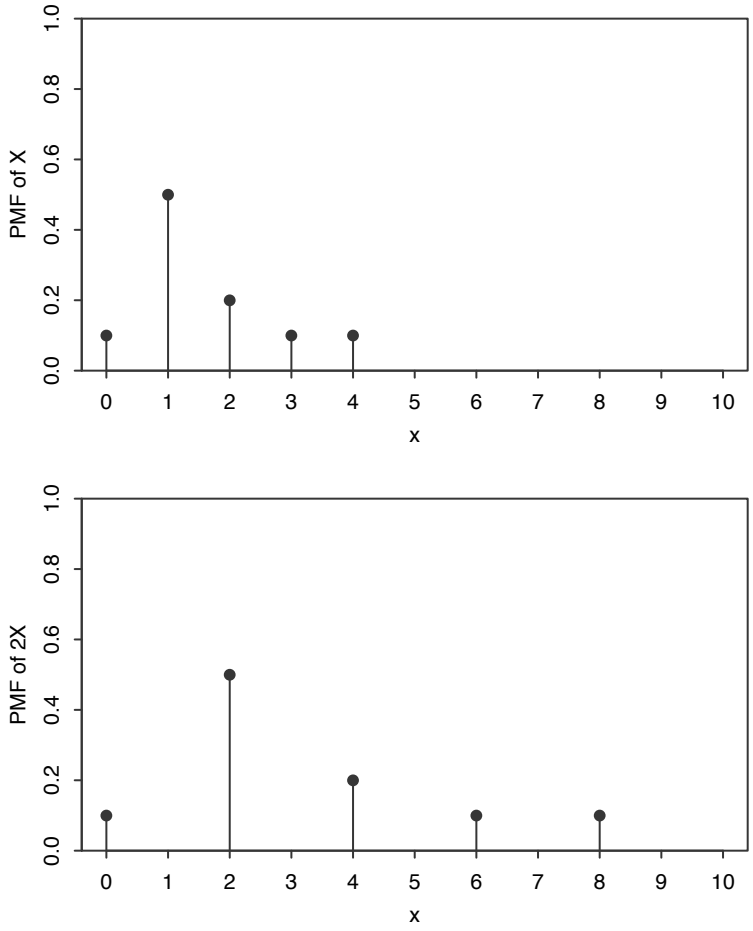

FIGURE 3.11
PMF of X (above) and PMF of $2X$ (below).

- 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