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.

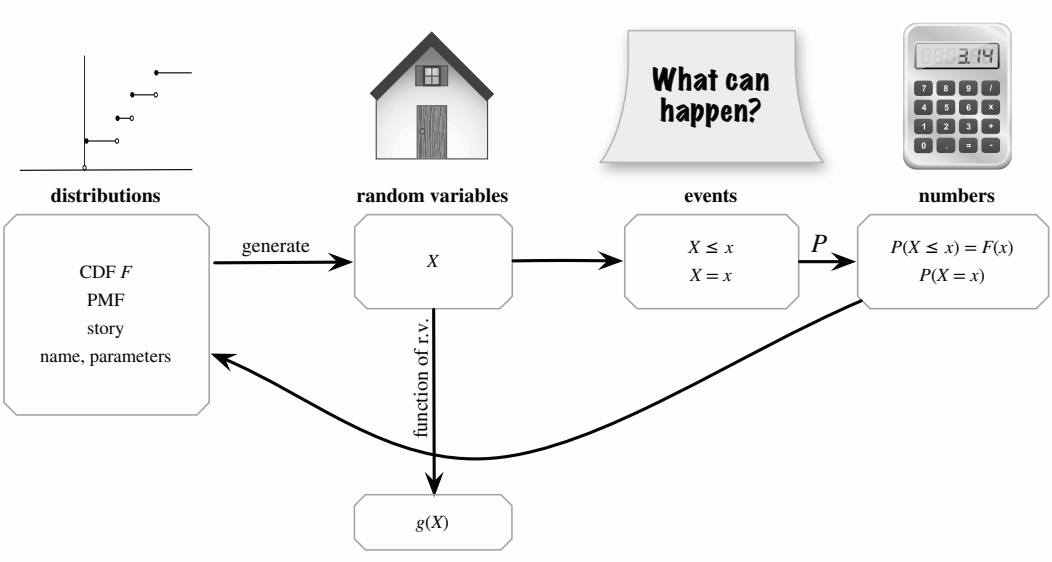

FIGURE 3.12
Four fundamental objects in probability: distributions (blueprints), random variables, events, and numbers. From a CDF F we can generate an r.v. X. From X, we can generate many other r.v.s by taking functions of X. There are various events describing the behavior of X. Most notably, for any 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 $p_1, \ldots, p_6$).