and $E(X_3) = \frac{6}{4}$. Continuing in this fashion, we can find the remaining $E(X_i)$'s. It follows that a customer will have to make *14.7* trips to the store, on the average, to collect a complete set of six letters:

$$\begin{aligned} E(X) &= \sum_{i=1}^{6} E(X_i) \\ &= 1 + \frac{6}{5} + \frac{6}{4} + \frac{6}{3} + \frac{6}{2} + \frac{6}{1} \\ &= 14.7 \end{aligned}$$

■

Questions

4.4.1. Because of her past convictions for mail fraud and forgery, Jody has a 30% chance each year of having her tax returns audited. What is the probability that she will escape detection for at least three years? Assume that she exaggerates, distorts, misrepresents, lies, and cheats every year.

4.4.2. A teenager is trying to get a driver's license. Write out the formula for the pdf $p_x(k)$, where the random variable X is the number of tries that he needs to pass the road test. Assume that his probability of passing the exam on any given attempt is 0.10. On the average, how many attempts is he likely to require before he gets his license?

4.4.3. Is the following set of data likely to have come from the geometric pdf $p_X(k) = \left(\frac{3}{4}\right)^{k-1} \cdot \left(\frac{1}{4}\right)$, $k = 1, 2, \ldots$? Explain.

2	8	1	2	2	5	1	2	8	3
5	4	2	4	7	2	2	8	4	7
2	6	2	3	5	1	3	3	2	5
4	2	2	3	6	3	6	4	9	3
3	7	5	1	3	4	3	4	6	2

4.4.4. Recently married, a young couple plans to continue having children until they have their first girl. Suppose the probability that a child is a girl is $\frac{1}{2}$, the outcome of each birth is an independent event, and the birth at which the first girl appears has a geometric distribution. What is the couple's expected family size? Is the geometric pdf a reasonable model here? Discuss.

4.4.5. Show that the cdf for a geometric random variable is given by $F_X(t) = P(X \le t) = 1 - (1 - p)^{[t]}$, where $[t]$ denotes the greatest integer in t, $t \ge 0$.

4.4.6. Suppose three fair dice are tossed repeatedly. Let the random variable X denote the roll on which a sum of 4 appears for the first time. Use the expression for $F_x(t)$ given in Question 4.4.5 to evaluate $P(65 \le X \le 75)$.

4.4.7. Let Y be an exponential random variable, where $f_Y(y) = \lambda e^{-\lambda y}$, $0 \le y$. For any positive integer n, show that $P(n \le Y \le n + 1) = e^{-\lambda n}(1 - e^{-\lambda})$. Note that if $p = 1 - e^{-\lambda}$, the "discrete" version of the exponential pdf is the geometric pdf.

4.4.8. Sometimes the geometric random variable is defined to be the number of trials, *X, preceding* the first success. Write down the corresponding pdf and derive the moment-generating function for X two ways—(1) by evaluating $E(e^{tX})$ directly and (2) by using Theorem 3.12.3.

4.4.9. Differentiate the moment-generating function for a geometric random variable and verify the expressions given for $E(X)$ and $\text{Var}(X)$ in Theorem 4.4.1.

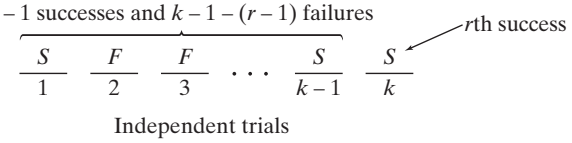
4.4.10. Suppose that the random variables X_1 and X_2 have mgfs $M_{X_1}(t) = \frac{\frac{1}{2}e^t}{1-\left(1-\frac{1}{2}\right)e^t}$ and $M_{X_2}(t) = \frac{\frac{1}{4}e^t}{1-\left(1-\frac{1}{4}t\right)e^t}$, respectively. Let $X = X_1 + X_2$. Does X have a geometric distribution? Assume that X_1 and X_2 are independent.

4.4.11. The *factorial moment-generating function* for any random variable W is the expected value of t^w. Moreover $\frac{d^r}{dt^r}E(t^W)\,|_{t=1} = E[W(W-1)\cdots(W-r+1)]$. Find the factorial moment-generating function for a geometric random variable and use it to verify the expected value and variance formulas given in Theorem 4.4.1.

4.5 The Negative Binomial Distribution

The geometric distribution introduced in Section 4.4 can be generalized in a very straightforward fashion. Imagine waiting for the rth (instead of the first) success in a series of independent trials, where each trial has a probability of p of ending in success (see Figure 4.5.1).

Figure 4.5.1



Let the random variable X denote the trial at which the rth success occurs. Then

$$\begin{aligned} p_X(k) = P(X = k) &= P(r\text{th success occurs on } k\text{th trial}) \\ &= P(r-1 \text{ successes occur in first } k-1 \text{ trials and} \\ &\quad \text{success occurs on } k\text{th trial}) \\ &= P(r-1 \text{ successes occur in first } k-1 \text{ trials}) \\ &\quad \cdot P(\text{Success occurs on } k\text{th trial}) \\ &= \binom{k-1}{r-1} p^{r-1}(1-p^{k-1-(r-1)}) \cdot p \\ &= \binom{k-1}{r-1} p^r (1-p)^{k-r}, k = r, r+1, \ldots \end{aligned} \tag{4.5.1}$$

Any random variable whose pdf has the form given in Equation 4.5.1 is said to have a *negative binomial distribution* (with parameter p).

Comment Two equivalent formulations of the negative binomial structure are widely used. Sometimes X is defined to be the number of trials *preceding* the rth success; other times, X is taken to be the number of trials in *excess of* r that are necessary to achieve the rth success. The underlying probability structure is the same, however X is defined. We will primarily use Equation 4.5.1; properties of the other two definitions for X will be covered in the exercises.

Theorem 4.5.1 *Let X have a negative binomial distribution with $p_X(k) = \binom{k-1}{r-1} p^r.(1-p)^{k-r}$, $k = r, r+1, \ldots$. Then*

1. $M_X(t) = \left[\frac{pe^t}{1-(1-p)e^t}\right]^r$
2. $E(X) = \frac{r}{p}$
3. $\text{Var}(X) = \frac{r(1-p)}{p^2}$

Proof All of these results follow immediately from the fact that X can be written as the sum of r independent geometric random variables, $X_1, X_2, \ldots, X_r$, each with parameter p. That is,

$$\begin{aligned} X &= \text{total number of trials to achieve } r\text{th success} \\ &= \text{number of trials to achieve 1st success} \\ &\quad + \text{number of additional trials to achieve 2nd success} + \cdots \\ &\quad + \text{number of additional trials to achieve } r\text{th success} \\ &= X_1 + X_2 + \cdots + X_r \end{aligned}$$

where

$$p_{X_i}(k) = (1-p)^{k-1} p, \quad k = 1, 2, \ldots, \quad i = 1, 2, \ldots, r$$

Therefore,

$$M_X(t) = M_{X_1}(t)M_{X_2}(t)\dots M_{X_r}(t)$$
$$= \left[\frac{pe^t}{1-(1-p)e^t}\right]^r$$

Also, from Theorem 4.4.1,

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_r)$$
$$= \frac{1}{p} + \frac{1}{p} + \cdots + \frac{1}{p}$$
$$= \frac{r}{p}$$

and

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_r)$$
$$= \frac{1-p}{p^2} + \frac{1-p}{p^2} + \cdots + \frac{1-p}{p^2}$$
$$= \frac{r(1-p)}{p^2}$$

□

Example 4.5.1 The California Mellows are a semipro baseball team. Eschewing all forms of violence, the laid-back Mellow batters never swing at a pitch, and should they be fortunate enough to reach base on a walk, they never try to steal. On the average, how many runs will the Mellows score in a nine-inning road game, assuming the opposing pitcher has a 50% probability of throwing a strike on any given pitch (83)?

The solution to this problem illustrates very nicely the interplay between the physical constraints imposed by a question (in this case, the rules of baseball) and the mathematical characteristics of the underlying probability model. The negative binomial distribution appears *twice* in this analysis, along with several of the properties associated with expected values and linear combinations.

To begin, we calculate the probability of a Mellow batter striking out. Let the random variable X denote the number of pitches necessary for that to happen. Clearly, $X = 3, 4, 5$, or 6 (why can X not be larger than 6?), and

$$p_X(k) = P(X = k) = P(\text{2 strikes are called in the first } k-1 \text{ pitches and the } k\text{th pitch is the 3rd strike})$$
$$= \binom{k-1}{2}\left(\frac{1}{2}\right)^3\left(\frac{1}{2}\right)^{k-3}, \quad k = 3, 4, 5, 6$$

Therefore,

$$P(\text{Batter strikes out}) = \sum_{k=3}^{6} p_X(k) = \left(\frac{1}{2}\right)^3 + \binom{3}{2}\left(\frac{1}{2}\right)^4 + \binom{4}{2}\left(\frac{1}{2}\right)^5 + \binom{5}{2}\left(\frac{1}{2}\right)^6$$
$$= \frac{21}{32}$$

Now, let the random variable W denote the number of walks the Mellows get in a given inning. In order for W to take on the value w, exactly two of the first $w+2$ batters must strike out, as must the $(w+3)$rd (see Figure 4.5.2). The pdf for W, then, is a negative binomial with $p = P(\text{Batter strikes out}) = \frac{21}{32}$:

$$p_W(w) = P(W = w) = \binom{w+2}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^w, \quad w = 0, 1, 2, \ldots$$

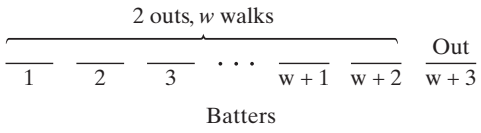


Figure 4.5.2

In order for a run to score, the pitcher must walk a Mellows batter with the bases loaded. Let the random variable R denote the total number of runs walked in during a given inning. Then

$$R = \begin{cases} 0 & \text{if } w \le 3 \\ w - 3 & \text{if } w > 3 \end{cases}$$

and

$$\begin{aligned} E(R) = \quad & = \sum_{w=4}^{\infty}(w-3)\binom{w+2}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^w \\ & = \sum_{w=0}^{\infty}(w-3)\cdot P(W=w) - \sum_{w=0}^{3}(w-3)\cdot P(W=w) \\ & = E(W) - 3 + \sum_{w=0}^{3}(3-w)\cdot\binom{w+2}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^w \end{aligned} \tag{4.5.2}$$

To evaluate $E(W)$ using the statement of Theorem 4.5.1 requires a linear transformation to rescale W to the format of Equation 4.5.1. Let

$$T = W + 3 = \text{total number of Mellow batters appearing in a given inning}$$

Then

$$p_T(t) = p_W(t-3) = \binom{t-1}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^{t-3}, \quad t = 3, 4, \ldots$$

which we recognize as a negative binomial pdf with $r = 3$ and $p = \frac{21}{32}$. Therefore,

$$E(T) = \frac{3}{21/32} = \frac{32}{7}$$

which makes $E(W) = E(T) - 3 = \frac{32}{7} - 3 = \frac{11}{7}$.

From Equation 4.5.2, then, the expected number of runs scored by the Mellows in a given inning is *0.202:*

$$E(R) = \frac{11}{7} - 3 + 3 \cdot \binom{2}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^0 + 2 \cdot \binom{3}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^1 + 1 \cdot \binom{4}{2}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^2 = 0.202$$

Each of the nine innings, of course, would have the same value for $E(R)$, so the expected number of runs in a *game* is the sum $0.202 + 0.202 + \cdots + 0.202 = 9(0.202)$, or *1.82*. ■

Case Study 4.5.1

Natural phenomena that are particularly complicated for whatever reasons may be impossible to describe with any single, easy-to-work-with probability model. An effective Plan B in those situations is to break the phenomenon down into simpler components and simulate the contributions of each of those components by using randomly generated observations. These are called Monte Carlo analyses, an example of which is described in detail in Section 4.7.

The fundamental requirement of any simulation technique is the ability to generate random observations from specified pdfs. In practice, this is done using computers because the number of observations needed is huge. In principle, though, the same, simple procedure can be used, by hand, to generate random observations from any discrete pdf.

Recall Example 4.5.1 and the random variable W, where W is the number of walks the Mellow batters are issued in a given inning. It was shown that $p_W(w)$ is the particular negative binomial pdf,

$$p_W(w) = P(W = w) = \binom{w+2}{w}\left(\frac{21}{32}\right)^3\left(\frac{11}{32}\right)^w, \quad w = 0, 1, 2, \ldots$$

Suppose a record is kept of the numbers of walks the Mellow batters receive in each of the next one hundred innings the team plays. What might that record look like?

The answer is, the record will look like a random sample of size 100 drawn from $p_W(w)$. Table 4.5.1 illustrates a procedure for generating such a sample. The first two columns show $p_W(w)$ for the nine values of w likely to occur (0 through 8). The third column parcels out the one hundred digits 00 through 99 into nine intervals whose lengths correspond to the values of $p_W(w)$.

There are twenty-nine two-digit numbers, for example, in the interval 28 to 56, with each of those numbers having the same probability of 0.01. Any random two-digit number that falls anywhere in that interval will then be mapped into the value $w = 1$ (which will happen, in the long run, 29% of the time).

Tables of random digits are typically presented in blocks of twenty-five (see Figure 4.5.3).

(Continued on next page)