chapter 8

The Binomial and Geometric Distributions

ACTIVITY 8 A Gaggle of Girls

The Ferrells have 3 children: Jennifer, Jessica, and Jaclyn. If we assume that a couple is equally likely to have a girl or a boy, then how unusual is it for a family like the Ferrells to have 3 children who are all girls? We have encountered problems like this in an earlier chapter. But this time we're going to use the method of simulation. If success = girl, and failure = boy, then p(success) = 0.5. We will define the random variable X as the number of girls. Then we want to simulate families with 3 children. Our goal is to determine the long-term relative frequency of a family with 3 girls, that is, $P(X = 3)$.

1. Using a random number table, let even digits represent "girl" and odd digits represent "boy." Select a row, and beginning at that row, read off numbers 3 digits at a time. Each 3 digits will constitute one trial. Use tally marks in a table like this one to record the results:

3 girls	
Not 3 girls	

Do at least 40 trials. Then combine your results with those of other students in the class to obtain at least 200 trials. Calculate the relative frequency of the event {3 girls}.

2. For variety, do the same thing as before, but this time using the calculator. Using the codes 1 = girl and 0 = boy, enter the command `randInt(0,1,3)`. This command instructs the calculator to randomly pick a whole number from the set {0, 1} and to do this 3 times. The outcome {0, 0, 1}, using our codes, means {*boy, boy, girl*}, in that order. Continue to press ENTER and count until you have 40 trials. Use a tally mark to record each time you observe a {1, 1, 1} result. Calculate the relative frequency for the event {3 girls}.

3. *Extra for programming experts:* Write a calculator program to carry out the process described above. Allow the user to specify the number of trials, and have the calculator report the relative frequency of {3 girls} as a decimal number.

4. Determine the total number of outcomes for this experiment. List the outcomes in the sample space. Then complete the probability distribution table for the random variable X = number of girls.

X	0	1	2	3
$P(X)$				

Do the results of your simulations come close to the theoretical value for $P(X = 3)$?

INTRODUCTION

In practice, we frequently encounter experimental situations where there are two outcomes of interest. Some examples are:

- We use a coin toss to see which of the two football teams gets the choice of kicking off or receiving to begin the game.
- A basketball player shoots a free throw; the outcomes of interest are {she makes the shot; she misses}.
- A young couple prepares for their first child; the possible outcomes are {boy; girl}.
- A quality control inspector selects a widget coming off the assembly line; he is interested in whether or not the widget meets production requirements.

In this chapter we will explore two important classes of distributions—the binomial distributions and the geometric distributions—and learn some of their properties. We will use what we have learned about probability and random variables from previous chapters, with the view toward completing the necessary foundation to study inference.

8.1 THE BINOMIAL DISTRIBUTIONS

In Activity 8, we simulated families with 3 children to discover how often the children would be all girls. Flipping a fair coin 3 times and letting heads represent having a girl and tails represent having a boy would produce exactly the same results. The characterizing features of this experiment are as follows: A *trial* consists of flipping the coin once. There are two outcomes: heads = girl (success), and tails = boy (failure). We will flip the coin 3 times. The coin flips are independent in the sense that the outcome of one coin flip has no influence on the outcome of the next flip. And last, the probability of success (girl) is the same for each coin flip (trial). A situation where these four conditions are satisfied is said to be a ***binomial setting.***

binomial setting

THE BINOMIAL SETTING

1. Each observation falls into one of just two categories, which for convenience we call "success" or "failure."
2. There is a fixed number n of observations.
3. The n observations are all **independent**. That is, knowing the result of one observation tells you nothing about the other observations.
4. The probability of success, call it p, is the same for each observation.

If you are presented with an experimental setting, it is important to be able to recognize it as a binomial setting or a geometric setting (covered in the next section) or neither. If you can verify that each of these four conditions is satisfied, you will be able to make use of known properties of binomial situations to gain more insights.

binomial random variable

If data are produced in a binomial setting, then the random variable X = number of successes is called a ***binomial random variable***, and the probability distribution of X is called a *binomial distribution.*

BINOMIAL DISTRIBUTION

The distribution of the count X of successes in the binomial setting is the **binomial distribution** with parameters n and p. The parameter n is the number of observations, and p is the probability of a success on any one observation. The possible values of X are the whole numbers from 0 to n. As an abbreviation, we say that X is $B(n, p)$.

The binomial distributions are an important class of discrete probability distributions. Pay attention to the binomial setting because not all counts have binomial distributions.

EXAMPLE 8.1 BLOOD TYPES

Blood type is inherited. If both parents carry genes for the O and A blood types, each child has probability 0.25 of getting two O genes and so of having blood type O. Different children inherit independently of each other. The number of O blood types among 5 children of these parents is the count X of successes in 5 independent observations with probability 0.25 of a success on each observation. So X has the binomial distribution with $n = 5$ and $p = 0.25$. We say that X is $B(5, 0.25)$.

EXAMPLE 8.2 DEALING CARDS

Deal 10 cards from a shuffled deck and count the number X of red cards. There are 10 observations, and each gives either a red or a black card. A "success" is a red card. But the observations are *not* independent. If the first card is black, the second is more likely to be red because there are more red cards than black cards left in the deck. The count X does *not* have a binomial distribution.

EXAMPLE 8.3 INSPECTING SWITCHES

An engineer chooses an SRS of 10 switches from a shipment of 10,000 switches. Suppose that (unknown to the engineer) 10% of the switches in the shipment are bad. The engineer counts the number X of bad switches in the sample.

This is not quite a binomial setting. Just as removing one card in Example 8.2 changed the makeup of the deck, removing one switch changes the proportion of bad

switches remaining in the shipment. So the state of the second switch chosen is not independent of the first. But removing one switch from a shipment of 10,000 changes the makeup of the remaining 9999 switches very little. In practice, the distribution of X is very close to the binomial distribution with $n = 10$ and $p = 0.1$.

Example 8.3 shows how we can use the binomial distributions in the statistical setting of selecting an SRS. When the population is much larger than the sample, a count of successes in an SRS of size n has approximately the binomial distribution with n equal to the sample size and p equal to the proportion of successes in the population.

EXAMPLE 8.4 AIRCRAFT ENGINE RELIABILITY

Engineers define reliability as the probability that an item will perform its function under specific conditions for a specific period of time. If an aircraft engine turbine has probability 0.999 of performing properly for an hour of flight, the number of turbines in a fleet of 350 engines that fly for an hour without failure has the $B(350, 0.999)$ distribution. This binomial distribution is obtained by assuming, as seems reasonable, that the turbines fail independently of each other. A common cause of failure, such as sabotage, would destroy the independence and make the binomial model inappropriate.

EXERCISES

8.1 BINOMIAL SETTING? In each situation below, is it reasonable to use a binomial distribution for the random variable X? Give reasons for your answer in each case.

(a) An auto manufacturer chooses one car from each hour's production for a detailed quality inspection. One variable recorded is the count X of finish defects (dimples, ripples, etc.) in the car's paint.

(b) The pool of potential jurors for a murder case contains 100 persons chosen at random from the adult residents of a large city. Each person in the pool is asked whether he or she opposes the death penalty; X is the number who say "Yes."

(c) Joe buys a ticket in his state's "Pick 3" lottery game every week; X is the number of times in a year that he wins a prize.

8.2 BINOMIAL SETTING? In each of the following cases, decide whether or not a binomial distribution is an appropriate model, and give your reasons.

(a) Fifty students are taught about binomial distributions by a television program. After completing their study, all students take the same examination. The number of students who pass is counted.

(b) A student studies binomial distributions using computer-assisted instruction. After the initial instruction is completed, the computer presents 10 problems. The student solves each problem and enters the answer; the computer gives additional instruction between problems if the student's answer is wrong. The number of problems that the student solves correctly is counted.

(c) A chemist repeats a solubility test 10 times on the same substance. Each test is conducted at a temperature 10° higher than the previous test. She counts the number of times that the substance dissolves completely.

Finding binomial probabilities

We will give a formula later for the probability that a binomial random variable takes any of its values. In practice, you will rarely have to use this formula for calculations. The TI-83/89 and most statistical software packages calculate binomial probabilities.

EXAMPLE 8.5 INSPECTING SWITCHES

A quality engineer selects an SRS of 10 switches from a large shipment for detailed inspection. Unknown to the engineer, 10% of the switches in the shipment fail to meet the specifications. What is the probability that no more than 1 of the 10 switches in the sample fail inspection?

The count X of bad switches in the sample has approximately the $B(10, 0.1)$ distribution. Figure 8.1 is a probability histogram for this distribution.

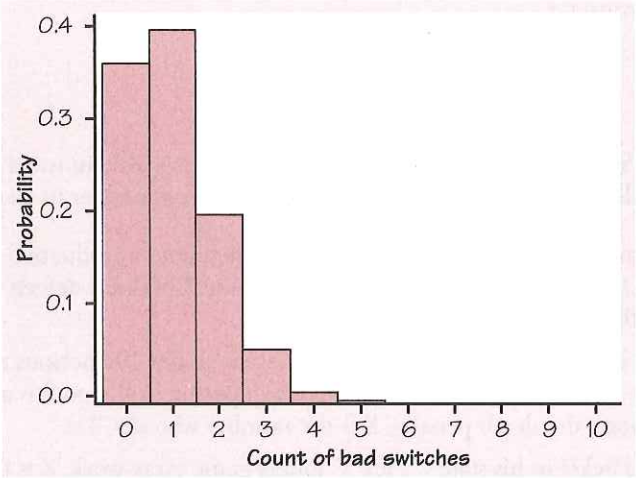


FIGURE 8.1 Probability histogram for the binomial distribution with $n = 10$ and $p = 0.1$.

The distribution is strongly skewed. Although X can take any whole-number value from 0 to 10, the probabilities of values larger than 5 are so small that they do not appear in the histogram. We want to calculate

$$P(X \le 1) = P(X = 0) + P(X = 1)$$

when X is $B(10, 0.1)$. The TI-83 command `binompdf(n,p,X)` and the TI-89 command `tistat.binomPdf(n,p,X)` calculate the binomial probability of the value X. The suffix `pdf` stands for "probability distribution function." We met the probability distribution in Chapter 7.

pdf

Given a discrete random variable X, the **probability distribution function** assigns a probability to each value of X. The probabilities must satisfy the rules for probabilities given in Chapter 6.

The `binompdf` command is found under 2nd (DISTR) / `0:binompdf` on the TI-83. On the TI-89, it's in the CATALOG under Flash Apps. The TI-83 command `binompdf(10,.1,0)` and the TI-89 command `tistat.binompdf(10,.1,0)` calculate the binomial probability that X = 0 to be 0.3486784401. The command `binompdf(10,.1,1)` returns probability 0.387420489. Thus,

$$\begin{aligned} P(X \le 1) &= P(X = 0) + P(X = 1) \\ &= 0.3487 + 0.3874 = 0.7361 \end{aligned}$$

About 74% of all samples will contain no more than 1 bad switch. A sample of size 10 cannot be trusted to alert the engineer to the presence of unacceptable items in the shipment.

EXAMPLE 8.6 CORINNE'S FREE THROWS

Corinne is a basketball player who makes 75% of her free throws over the course of a season. In a key game, Corinne shoots 12 free throws and makes only 7 of them. The fans think that she failed because she was nervous. Is it unusual for Corinne to perform this poorly? To answer this question, assume that free throws are independent with probability 0.75 of a success on each shot. (Studies of long sequences of free throws have found no evidence that they are dependent, so this is a reasonable assumption.) The number X of baskets (successes) in 12 attempts has the $B(12, 0.75)$ distribution.

We want the probability of making a basket on at most 7 free throws. This is

$$\begin{aligned} P(X \le 7) &= P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = 7) \\ &= 0.0000 + 0.0000 + 0.0000 + 0.0004 + 0.0024 + 0.0115 + 0.0401 \\ &\quad + 0.1032 \\ &= 0.1576 \end{aligned}$$

Corinne will make at most 7 of her 12 free throws about 16% of the time, or roughly in one of every six games. While below her average level, this performance is well within the range of the usual chance variation in her shooting.

EXAMPLE 8.7 THREE GIRLS

In Activity 8 we wanted to determine the probability that all 3 children in a family are girls. In this case, the random variable of interest, X = the number of girls, has the $B(3, 0.5)$ distribution. We want to find the probability that the number of

girls is 3, that is, $P(X = 3)$. The TI-83 command `binompdf(3,.5,3)` and the TI-89 command `tistat.binomPdf(3,.5,3)` return the probability 0.125.

In applications we frequently want to find the probability that a random variable takes a range of values. The *cumulative* binomial probability is useful in these cases.

cdf

Given a random variable X, the **cumulative distribution function** (cdf) of X calculates the sum of the probabilities for 0, 1, 2, ..., up to the value X. That is, it calculates the probability of obtaining at most X successes in n trials.

For the count X of defective switches in Example 8.5, the command `binomcdf(10,.1,1)` and the TI-89 command `tistat.binomCdf(10,.1,1)` output 0.736098903 for the cumulative probability $P(X \le 1)$.

EXAMPLE 8.8 IS CORINNE IN A SLUMP?

In Example 8.6. Corinne shoots $n = 12$ free throws and makes only 7 of them. Since she is a 75% free-throw shooter ($p = 0.75$), we wanted to know if it was unusual for Corinne to perform this poorly. If X = number of baskets made on free throws, then X has the $B(12, 0.75)$ distribution, and we need to find the probability that she makes at most 7 of her free throws, that is, $P(X \le 7)$. The TI-83 command `binomcdf(12,.75,7)` and the TI-89 command `tistat.binomCdf(12,.75,7)` calculates the cumulative probability $P(X \le 7)$ to be 0.1576436761. We round the answer to four decimal places and report that the probability that Corinne makes *at most* 7 of her 12 free throws is 0.1576.

The pdf table for Corinne's shots looks like this:

X	0	1	2	3	4	5	6
P(X)	0.000	0.000	0.000	0.000	0.002	0.011	0.040
X	7	8	9	10	11	12	
P(X)	0.103	0.194	0.258	0.232	0.127	0.032	

If we denote the cumulative distribution function by $F(X)$, we can record the cumulative sum of the probabilities in a third row of the table:

X	0	1	2	3	4	5	6
P(X)	0.000	0.000	0.000	0.000	0.002	0.011	0.040
F(X)	0.000	0.000	0.000	0.000	0.003	0.014	0.054

X	7	8	9	10	11	12
$P(X)$	0.103	0.194	0.258	0.232	0.127	0.032
$F(X)$	0.158	0.351	0.609	0.842	0.968	1

Notice that terms sometimes don't appear to add up as they should. The cumulative function $F(4)$, for example, should equal $P(0) + P(1) + P(2) + P(3) + P(4)$. Of course, the culprit is roundoff error. With your calculator, enter the integers 0 to 12 into L_1/list1, the corresponding binomial probabilities into L_2/list2, and use the command `binomcdf(12,.75,`L_1`)`$\rightarrow L_3$ (`tistat.binomCdf(12,.75,list1)`$\rightarrow$`list3` on the TI-89) to enter the cumulative probabilities into L_3/list3.

In addition to being helpful in answering questions involving wording such as "find the probability that it takes at most 6 trials," the cdf is also particularly useful for calculating the probability that it takes *more* than a certain number of trials to see the first success. This calculation uses the complement rule:

$$P(X > n) = 1 - P(X \leq n) \quad n = 2, 3, 4, \ldots$$

EXERCISES

Use your calculator's binomial pdf or cdf commands to find the following probabilities:

8.3 INHERITING BLOOD TYPE Each child born to a particular set of parents has probability 0.25 of having blood type O. Suppose these parents have 5 children. Let X = number of children who have type O blood. Then X is $B(5, 0.25)$.

(a) What is the probability that exactly 2 of the children have type O blood?

(b) Make a table for the pdf of the random variable X. Then use the calculator to find the probabilities of all possible values of X, and complete the table.

(c) Verify that the sum of the probabilities is 1.

(d) Construct a histogram of the pdf.

(e) Use the calculator to find the cumulative probabilities, and add these values to your pdf table. Then construct a cumulative distribution histogram. How is this histogram different from the histogram for Corinne's free throws?

8.4 GUESSING ON A TRUE-FALSE QUIZ Suppose that James guesses on each question of a 50-item true-false quiz. Find the probability that James passes if

(a) a score of 25 or more correct is needed to pass.

(b) a score of 30 or more correct is needed to pass.

(c) a score of 32 or more correct is needed to pass.

8.5 GUESSING ON A MULTIPLE-CHOICE QUIZ Suppose that Erin guesses on each question of a multiple-choice quiz.

(a) If each question has four different choices, find the probability that Erin gets one or more correct answers on a 10-item quiz.

(b) If the quiz consists of three questions, question 1 has 3 possible answers, question two has 4 possible answers, and question 3 has 5 possible answers, find the probability that Erin gets one or more correct answers.

8.6 DAD'S IN THE POKEY According to a 2000 study by the Bureau of Justice Statistics, approximately 2% of the nation's 72 million children had a parent behind bars—nearly 1.5 million minors. Let X be the number of children who had an incarcerated parent. Suppose that 100 children are randomly selected.

(a) Does X satisfy the requirements for a binomial setting? Explain. If $X = B(n, p)$, what are n and p?

(b) Describe $P(X = 0)$ in words. Then find $P(X = 0)$ and $P(X = 1)$.

(c) What is the probability that 2 or more of the 100 children have a parent behind bars?

8.7 DO OUR ATHLETES GRADUATE? A university claims that 80% of its basketball players get degrees. An investigation examines the fate of all 20 players who entered the program over a period of several years that ended six years ago. Of these players, 11 graduated and the remaining 9 are no longer in school. If the university's claim is true, the number of players among the 20 who graduate should have the binomial distribution with $n = 20$ and $p = 0.8$. What is the probability that exactly 11 out of 20 players graduate?

8.8 MARITAL STATUS Among employed women, 25% have never been married. Select 10 employed women at random.

(a) The number in your sample who have never been married has a binomial distribution. What are n and p?

(b) What is the probability that exactly 2 of the 10 women in your sample have never been married?

(c) What is the probability that 2 or fewer have never been married?

Binomial formulas

We can find a formula for the probability that a binomial random variable takes any value by adding probabilities for the different ways of getting exactly that many successes in n observations. Here is the example we will use to show the idea.

EXAMPLE 8.9 INHERITING BLOOD TYPE

Each child born to a particular set of parents has probability 0.25 of having blood type O. If these parents have 5 children, what is the probability that exactly 2 of them have type O blood?

The count of children with type O blood is a binomial random variable X with $n = 5$ tries and probability $p = 0.25$ of a success on each try. We want $P(X = 2)$.

Because the method doesn't depend on the specific example, let's use "S" for success and "F" for failure for short. Do the work in two steps.

Step 1. Find the probability that a specific 2 of the 5 tries give successes, say the first and the third. This is the outcome SFSFF. Here's how to find the probability of this outcome:

- The probability that the first try is a success is 0.25. That is, in many repetitions, we succeed on the first try 25% of the time.
- Out of all the repetitions with a success on the first try, 75% have a failure on the second try. So the proportion of repetitions on which the first two tries are SF is (0.25)(0.75). We can multiply here because the tries are *independent.* That is, the first try has no influence on the second.
- Keep going: Of these repetitions, the proportion 0.25 have S on the third try. So the probability of SFS is (0.25)(0.75)(0.25). After two more tries, the probability of SFSFF is the product of the try-by-try probabilities:

$$(0.25)(0.75)(0.25)(0.75)(0.75) = (0.25)^2(0.75)^3$$

Step 2. Observe that the probability of *any one* arrangement of 2 S's and 3 F's has this same probability. That's true because we multiply together 0.25 twice and 0.75 three times whenever we have 2 S's and 3 F's. The probability that $X = 2$ is the probability of getting 2 S's and 3 F's in any arrangement whatsoever. Here are all the possible arrangements:

SSFFF	SFSFF	SFFSF	SFFFS	FSSFF
FSFSF	FSFFS	FFSSF	FFSFS	FFFSS

There are 10 of them, all with the same probability. The overall probability of 2 successes is therefore

$$P(X = 2) = 10(0.25)^2(0.75)^3 = 0.2637$$

The pattern of this calculation works for any binomial probability. To use it, we need to be able to count the number of arrangements of k successes in n observations without actually listing them. We use the following fact to do the counting:

BINOMIAL COEFFICIENT

The number of ways of arranging k successes among n observations is given by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

for $k = 0, 1, 2, \ldots, n$.

factorial

The formula for binomial coefficients uses the ***factorial*** notation. For any positive whole number n, its factorial $n!$ is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

Also, $0! = 1$.

Notice that the larger of the two factorials in the denominator of a binomial coefficient will cancel much of the $n!$ in the numerator. For example, the binomial coefficient we need for Example 8.9 is

$$\begin{aligned}\binom{5}{2} &= \frac{5!}{2!\,3!} \\ &= \frac{(5)(4)(3)(2)(1)}{(2)(1)\times(3)(2)(1)} \\ &= \frac{(5)(4)}{(2)(1)} = \frac{20}{2} = 10\end{aligned}$$

The notation $\binom{n}{k}$ is *not* related to the fraction $\frac{n}{k}$. A helpful way to remember its meaning is to read it as "binomial coefficient n choose k." Binomial coefficients have many uses in mathematics, but we are interested in them only as an aid to finding binomial probabilities. The binomial coefficient $\binom{n}{k}$ counts the number of ways in which k successes can be distributed among n observations. The binomial probability $P(X = k)$ is this count multiplied by the probability of any specific arrangement of the k successes. Here is the formula we seek:

BINOMIAL PROBABILITY

If X has the binomial distribution with n observations and probability p of success on each observation, the possible values of X are $0, 1, 2, \ldots, n$. If k is any one of these values,

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

EXAMPLE 8.10 DEFECTIVE SWITCHES

The number X of switches that fail inspection in Example 8.3 has approximately the binomial distribution with $n = 10$ and $p = 0.1$. The probability that no more than 1 switch fails is

$$
\begin{aligned}
P(X \le 1) &= P(X=1) + P(X=0) \\
&= \binom{10}{1}(0.1)^1(0.9)^9 + \binom{10}{0}(0.1)^0(0.9)^{10} \\
&= \frac{10!}{1!9!}(0.1)(0.3874) + \frac{10!}{0!10!}(1)(0.3487) \\
&= (10)(0.1)(0.3874) + (1)(1)(0.3487) \\
&= 0.3874 + 0.3487 \\
&= 0.7361
\end{aligned}
$$

Notice that the calculation uses the facts that $0! = 1$ and that $a^0 = 1$ for any number a other than 0.

EXERCISES

In each of the following exercises, you are to use the binomial probability formula to answer the question. You may not use the binomial pdf command on your calculator. Begin with the formula, and show substitution into the formula.

8.9 BLOOD TYPES The count X of children with type O blood among 5 children whose parents carry genes for both the O and the A blood types is $B(5, 0.25)$. See Example 8.1 on page 440. Use the binomial probability formula to find $P(X = 3)$.

8.10 BROCCOLI PLANTS Suppose you purchase a bundle of 10 bare-root broccoli plants. The sales clerk tells you that on average you can expect 5% of the plants to die before producing any broccoli. Assume that the bundle is a random sample of plants. Use the binomial formula to find the probability that you will lose at most one of the broccoli plants.

8.11 MORE ON BLOOD TYPES Use the binomial probability formula to find the probability that at least one of the children in Exercise 8.9 has blood type O. (*Hint:* Do not calculate more than one binomial formula.)

8.12 GRADUATION RATE FOR ATHLETES See Exercise 8.7 on page 446. The number of athletes who graduate is $B(20, 0.8)$. Use the binomial probability formula to find the probability that all 20 graduate. What's the probability that not all of the 20 graduate?

8.13 HISPANIC REPRESENTATION A factory employs several thousand workers, of whom 30% are Hispanic. If the 15 members of the union executive committee were chosen from the workers at random, the number of Hispanics on the committee would have the binomial distribution with $n = 15$ and $p = 0.3$.

(a) What is the probability that exactly 3 members of the committee are Hispanic?

(b) What is the probability that none of the committee members are Hispanic?

8.14 CORINNE'S FREE THROWS. Use the binomial probability formula to show that the probability that Corinne makes exactly 7 of her 12 free throws is 0.1032 (see Example 8.6 on page 443).

Binomial mean and standard deviation

If a count X has the binomial distribution based on n observations with probability p of success, what is its mean μ? We can guess the answer. If a basketball player makes 75% of her free throws, the mean number made in 12 tries should be 75% of 12, or 9. In general, the mean of a binomial distribution should be $\mu = np$. To derive the expressions for the mean and standard deviation in the general case, let X represent the number of successes in a single trial. Then X takes two values, 1 (for success) and 0 (for failure). We'll let p be the probability of success on a single trial, and introduce $q = 1 - p$ as the probability of failure. This is common notation. Then the probability distribution is simply

X	0	1
$P(X)$	q	p

The expected value for this one trial is

$$E(X) = \mu_X = 0\ (q) + 1\ (p) = p$$

The variance is

$$\sigma_X^2 = (0-p)^2 q + (1-p)^2 p = p^2 q + pq^2 = pq(p+q) = pq$$

Now define a new random variable Y to be the number of successes in n independent trials. Then $Y = X_1 + X_2 + \cdots + X_n$. Using the rules for means and variances of linear combinations of *independent* random variables, we can say that

$$\begin{aligned}\mu_Y &= \mu_{(X_1+X_2+\cdots+X_n)} = \mu_{X_1} + \mu_{X_2} + \cdots + \mu_{X_n} \\ &= p + p + \cdots + p \\ &= np\end{aligned}$$

and

$$\begin{aligned}\sigma_Y^2 &= \sigma^2_{X_1+X_2+\cdots+X_n} = \sigma^2_{X_1} + \sigma^2_{X_2} + \cdots + \sigma^2_{X_n} \\ &= pq + pq + \cdots + pq \\ &= npq = np(1-p)\end{aligned}$$

and the standard deviation of Y is $\sqrt{np(1-p)}$. Here is what we have shown:

MEAN AND STANDARD DEVIATION OF A BINOMIAL RANDOM VARIABLE

If a count X has the binomial distribution with number of observations n and probability of success p, the mean and standard deviation of X are

$$\mu = np$$

$$\sigma = \sqrt{np(1-p)}$$

Important note: These short formulas are good only for binomial distributions. They can't be used for other discrete random variables.

EXAMPLE 8.11 BAD SWITCHES

Continuing Example 8.10, the count X of bad switches is binomial with $n = 10$ and $p = 0.1$. This is the sampling distribution the engineer would see if she drew all possible SRSs of 10 switches from the shipment and recorded the value of X for each sample.

The mean and standard deviation of the binomial distribution are

$$\mu = np$$
$$= (10)(0.1) = 1$$
$$\sigma = \sqrt{np(1-p)}$$
$$= \sqrt{(10)(0.1)(0.9)} = \sqrt{0.9} = 0.9487$$

The mean is marked on the probability histogram in Figure 8.2.

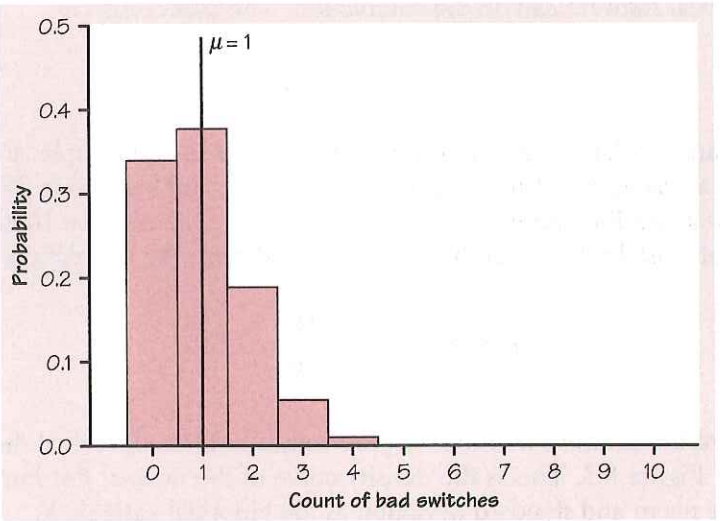


FIGURE 8.2 Probability histogram for the binomial distribution with $n = 10$ and $p = 0.1$.

The normal approximation to binomial distributions

The formula for binomial probabilities becomes awkward as the number of trials n increases. You can use software or a statistical calculator to handle some problems for which the formula is not practical. Here is another alternative: *as the number of trials n gets larger, the binomial distribution gets close to a normal distribution*. When n is large, we can use normal probability calculations to approximate hard-to-calculate binomial probabilities.

EXAMPLE 8.12 ATTITUDES TOWARD SHOPPING

Are attitudes toward shopping changing? Sample surveys show that fewer people enjoy shopping than in the past. A recent survey asked a nationwide random sample of 2500 adults if they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and time-consuming."[1] The population that the poll wants to draw conclusions about is all U.S. residents aged 18 and over. Suppose that in fact 60% of all adult U.S. residents would say "Agree" if asked the same question. What is the probability that 1520 or more of the sample agree?

Because there are more than 195 million adults, we can take the responses of 2500 randomly chosen adults to be independent. So the number in our sample who agree that shopping is frustrating is a random variable X having the binomial distribution with $n = 2500$ and $p = 0.6$. To find the probability that at least 1520 of the people in the sample find shopping frustrating, we must add the binomial probabilities of all outcomes from $X = 1520$ to $X = 2500$. This isn't practical. Here are three ways to do this problem.

1. Statistical software can do the calculation. The exact result is

$$P(X \geq 1520) = 0.2131$$

2. We can simulate a large number of repetitions of the sample. Figure 8.3 displays a histogram of the counts X from 1000 samples of size 2500 when the truth about the population is $p = 0.6$. Because 221 of these 1000 samples have X at least 1520, the probability estimated from the simulation is

$$P(X \geq 1520) = \frac{221}{1000} = 0.221$$

3. Both of the previous methods require software. Instead, look at the normal curve in Figure 8.3. This is the density curve of the normal distribution with the same mean and standard deviation as the binomial variable X:

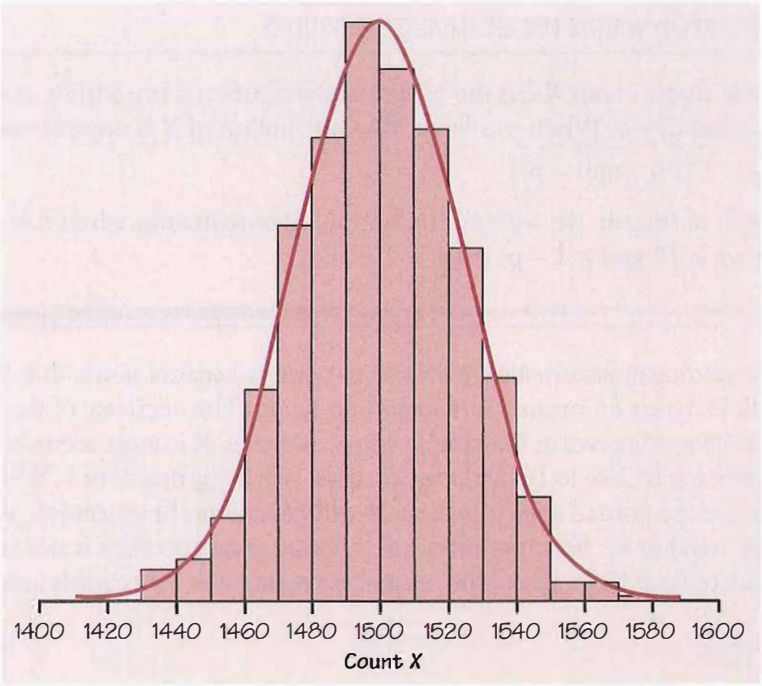


FIGURE 8.3 Histogram of 1000 binomial counts ($n = 2500$, $p = 0.6$) and the normal density curve that approximates this binomial distribution.

$$\mu = np = (2500)(0.6) = 1500$$
$$\sigma = \sqrt{np(1-p)} = \sqrt{(2500)(0.6)(0.4)} = 24.49$$

As the figure shows, this normal distribution approximates the binomial distribution quite well. So we can do a normal calculation.

EXAMPLE 8.13 NORMAL CALCULATION OF A BINOMIAL PROBABILITY

If we act as though the count X has the $N(1500, 24.49)$ distribution, here is the probability we want, using Table A:

$$P(X \geq 1520) = P\left(\frac{X - 1500}{24.49} \geq \frac{1520 - 1500}{24.49}\right)$$
$$= P(Z \geq 0.82)$$
$$= 1 - 0.7939 = 0.2061$$

The normal approximation 0.2061 differs from the software result 0.2131 by only 0.007.

NORMAL APPROXIMATION FOR BINOMIAL DISTRIBUTIONS

Suppose that a count X has the binomial distribution with n trials and success probability p. When n is large, the distribution of X is approximately normal, $N\left(np, \sqrt{np(1-p)}\right)$.

As a rule of thumb, we will use the normal approximation when n and p satisfy $np \geq 10$ and $n(1-p) \geq 10$.

The normal approximation is easy to remember because it says that X is normal with its binomial mean and standard deviation. The accuracy of the normal approximation improves as the sample size n increases. It is most accurate for any fixed n when p is close to 1/2 and least accurate when p is near 0 or 1. Whether or not you use the normal approximation should depend on how accurate your calculations need to be. For most statistical purposes great accuracy is not required. Our "rule of thumb" for use of the normal approximation reflects this judgment.

EXERCISES

8.15 ATTITUDES ON SHOPPING Refer to Example 8.12 on attitudes toward shopping.

(a) Verify that the rule of thumb conditions are satisfied for using the normal approximation to the binomial distribution.

(b) Use your calculator and the cumulative binomial function to verify the exact answer for the probability that at least 1520 people in the sample find shopping frustrating is 0.2131. What is the probability correct to 6 decimal places?

(c) What is the probability that at most 1468 people in the sample would agree with the statement that shopping is frustrating?

8.16 HISPANIC COMMITTEE MEMBERS

(a) What is the mean number of Hispanics on randomly chosen committees of 15 workers in Exercise 8.13 (page 449)?

(b) What is the standard deviation σ of the count X of Hispanic members?

(c) Suppose that 10% of the factory workers were Hispanic. Then $p = 0.1$. What is σ in this case? What is σ if $p = 0.01$? What does your work show about the behavior of the standard deviation of a binomial distribution as the probability of a success gets closer to 0?

8.17 DO OUR ATHLETES GRADUATE?

(a) Find the mean number of graduates out of 20 players in the setting of Exercise 8.12 (page 449).

(b) Find the standard deviation σ of the count X.

(c) Suppose that the 20 players came from a population of which $p = 0.9$ graduated. What is the standard deviation σ of the count of graduates? If $p = 0.99$, what is σ? What does your work show about the behavior of the standard deviation of a binomial distribution as the probability p of success gets closer to 1?

8.18 MARITAL STATUS OF EMPLOYED WOMEN You choose 10 employed women at random, as in Exercise 8.8 (page 446). What is the mean number of women in such a sample who have never been married? What is the standard deviation?

8.19 POLLING Many local polls of public opinion use samples of size 400 to 800. Consider a poll of 400 adults in Richmond that asks the question "Do you approve of President George W. Bush's response to the World Trade Center terrorists attacks in September 2001?" Suppose we know that President Bush's approval rating on this issue nationally is 92% a week after the incident.

(a) What is the random variable X? Is X binomial? Explain.

(b) Calculate the binomial probability that at most 358 of the 400 adults in the Richmond poll answer "Yes" to this question.

(c) Find the expected number of people in the sample who indicate approval. Find the standard deviation of X.

(d) Perform a normal approximation to answer the question in (b), and compare the results of the binomial calculation and the normal approximation. Is the normal approximation satisfactory?

8.20 A MARKET RESEARCH SURVEY You operate a restaurant. You read that a sample survey by the National Restaurant Association shows that 40% of adults are committed to eating nutritious food when eating away from home. To help plan your menu, you decide to conduct a sample survey in your own area. You will use random digit dialing to contact an SRS of 200 households by telephone.

(a) If the national result holds in your area, it is reasonable to use the binomial distribution with $n = 200$ and $p = 0.4$ to describe the count X of respondents who seek nutritious food when eating out. Explain why.

(b) What is the mean number of nutrition-conscious people in your sample if $p = 0.4$ is true? What is the standard deviation?

(c) What is the probability that X lies between 75 and 85? Make sure that the rule of thumb conditions are satisfied, and then use a normal approximation to answer the question.

Binomial distribution with the calculator

The following Technology Toolbox summarizes some important calculator techniques when working in a binomial setting:

TECHNOLOGY TOOLBOX *Exploring binomial distributions*

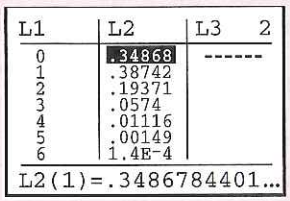
For illustration purposes, we will use the sample of $n = 10$ switches with probability $p = 0.10$ of a defective switch from Example 8.3 (page 440). The random variable X is the number of defective switches (success) and $X = B(10, 0.1)$. To have the calculator make the probability distribution table and plot a histogram for the distribution of defective switches in a sample of 10 switches, proceed as follows:

TECHNOLOGY TOOLBOX *Exploring binomial distributions (continued)*

TI-83

1. Enter the values of X into list L_1, either through the STAT/EDIT mode or by entering the command `seq(X,X,0,10,1)→L1`. (The `seq` command is under 2nd/LIST/OPS/5:seq. The syntax is: the first X is the function, the second X is the counting variable, the next two numbers define the starting and ending values, and the last number is the increment.)

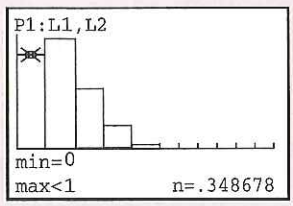
2. Enter the binomial probabilities into list L_2. Highlight L_2 and press 2nd VARS (DISTR)/ `0:binompdf(`. Then complete the command: `binompdf(10,0.1,L1)`. Note that the largest probability listed is about 0.3874. This will help us define our viewing window.



3. Deselect or delete any active defined functions in the Y = window.

4. Define Plot1 to be a histogram with Xlist:L_1 and Freq:L_2.

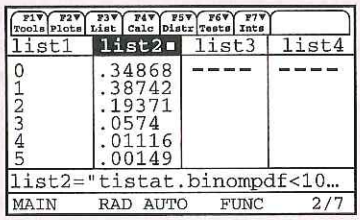
5. Set the viewing window to be $X[0,11]_1$ and $Y[-.15,.5]_1$. Press TRACE and use the left and right cursor keys to inspect heights of various bars in the histogram.



TI-89

1. Enter the values of X into list1, either through the Statistics/List Editor or by entering the command `seq(X,X,0,10,1)→list1`. The `seq` command is in the CATALOG. The syntax is: the first X is the function, the second X is the counting variable, the next two numbers define the starting and ending values, and the last number is the increment.)

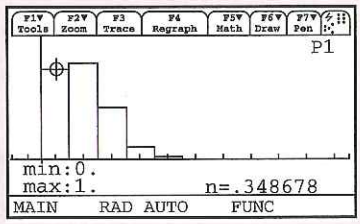
2. Highlight list2 and press CATALOG, type F3 (Flash Apps), choose `binompdf(`. Press ENTER and complete the command: `tistat.binomPdf(10,0.1,list1)` and then ENTER.



3. Deselect or delete any active defined functions in the Y = window.

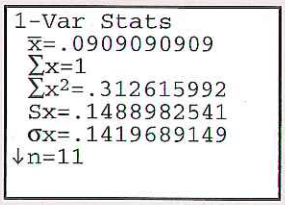
4. Define Plot1 to be a histogram using list1 for *x* and list2 for frequency.

5. Define the viewing window: ♦ F2 (WINDOW). Specify $X[0,11]_1$ and $Y[-.15,.5]_1$. Press ♦ F3. Here is the pdf.

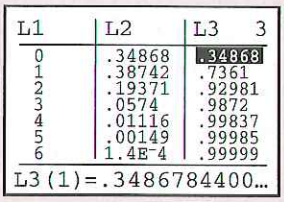


Outcomes larger than 6 do not have probability exactly 0, but their probabilities are so small that the rounded values are 0.0000. Verify that the sum of the probabilities is 1.

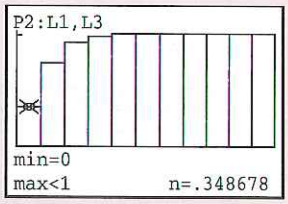
TECHNOLOGY TOOLBOX *Exploring binomial distributions (continued)*

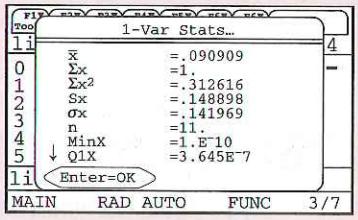


6. To calculate the cumulative probabilities, highlight list L_3. Press 2nd DISTR, and select `A:binomcdf(`. Complete the command: `binomcdf(10,.1,L1)`. Press ENTER. The cumulative probabilities are in L_3.

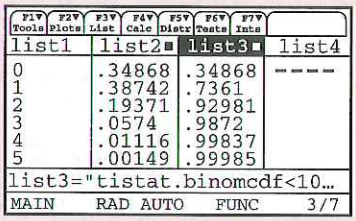


7. Turn off Plot1 and turn on Plot2. Define Plot2 to be a histogram with Xlist: L_1 and Ylist: L_3. In the viewing window, set Ymin = -.3 and Ymax = 1.2. Here is the histogram for the cdf:

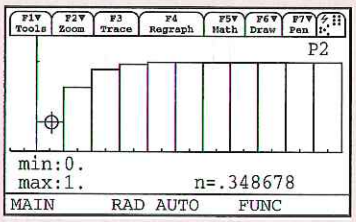




6. To calculate the cumulative distribution values, highlight list3 and press CATALOG F3 (Flash Apps), and choose `binomcdf(`. Press ENTER and complete the command: `tistat.binomCdf(10,.1,list1)` and then ENTER.



7. Turn off Plot1 (F2 (Plot Setup)), highlight Plot1, then F4 (✓) to deselect Plot1. Define Plot2 to be a histogram, except this time specify list3 for the frequency. In the viewing window, set Ymin = -.3 and Ymax = 1.2. Here is the histogram for the cdf:



Simulating binomial experiments

In order to simulate a binomial experiment, you need to know how the random variable X and "success" are defined, the probability of success, and the number of trials. But if you know these things, you can apply the rules learned in this section to calculate the probabilities of events exactly. So perhaps simulation methods are not as important in a binomial setting as they are in other settings. On

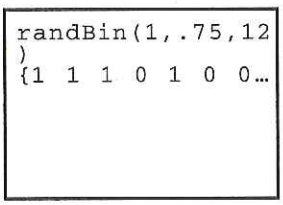
the other hand, being able to simulate a binomial experiment can give credence to results obtained by applying formulas and rules when the results may be less than convincing to someone who knows no statistics.

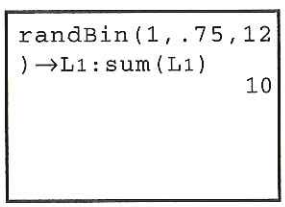
EXAMPLE 8.14 CORINNE'S FREE THROWS

Recall that Corinne's free throw percentage was 75% (see Example 8.6). In a particular game, she had 12 attempts and she made only 7. The question was, "How unusual was it for Corinne to make at most 7 shots out of 12 attempts?" In Example 8.6, we calculated this binomial probability to be $P(X \leq 7) = 0.1576$. Now we will use the calculator to simulate 12 attempted shots and we will count the number of hits (baskets). Let X = number of hits in 12 free throw attempts. Note that the probability of "success" is 0.75. To set up the simulation, we will assign the digit 0 to a miss and a 1 to a hit. The command `randBin(1,.75,12)` simulates 12 free throw attempts. In the long run, this random function will select the number 1 75% of the time and the number 0 25% of the time.

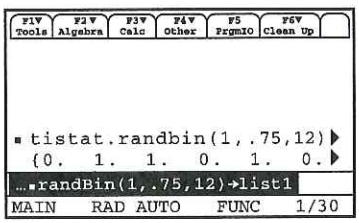
Here are the results of one simulated game on the TI-83: the first three shots were hits, the fourth was a miss, the fifth was a hit, and the next two were misses, and so forth. If we repeated this many times and counted the proportion of times Corinne had 7 or fewer hits, that would give an estimate of the probability $P(X \leq 7)$ that Corinne made at most 7 of her 12 attempts. One way to automate this more is to assign these results to list L_1/list1 and then sum the entries in the list. Enter the TI-83 command `randBin(1,.75,12)→L1:sum(L1)` (or for the TI-89, press CATALOG F3 (Flash Apps) and select `randbin(` and then complete the command: `tistat.randbin(1,.75,12)`).

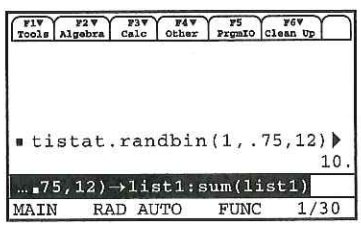
TI-83





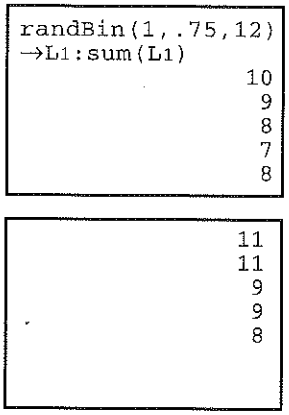
TI-89

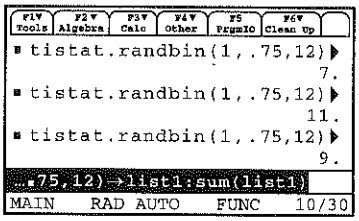




Continue pressing the ENTER key until you have 10 numbers. Record these numbers.

This makes 10 repetitions (i.e., simulates 10 games) for both calculators. (The TI-89 results were 10, 10, 10, 9, 10, 12, 9, 7, 11, 9.) So far Corinne has made 7 or fewer shots in 1 out of 10 games, for a relative frequency of 0.10. Compare this with the binomial probability of 0.1576 for this event. Continue to press ENTER to simulate 10 more





games. Calculate the relative frequency for 20 games and so on. According to the law of large numbers, these relative frequencies should get closer to 0.1576 as the number of simulated games increases. Continue in this fashion until you have simulated 50 games. Are your cumulative results close to 0.1576?

EXERCISES

8.21 CORINNE'S FREE THROWS Use lists L_1/list1 and L_2/list2 on your calculator to construct a pdf for Corinne's free throw probabilities. (Refer to Examples 8.6 and 8.8 on page 443 and 444.) Use the random variable X = number of baskets made on free throws. Then on both calculators, execute the command cumSum(L_2)→L_3. (cumSum is found under 2nd / LIST / OPS / 6 : cumSum on the TI-83; cumSum is found in the CATALOG on the TI-89). What do you think this command does? Then use the binomcdf command to enter the cumulative probabilities into list L_4/list4. Compare L_3/list3 and L_4/list4. Are they the same?

8.22 SIMULATING DEFECTIVE SWITCHES

(a) Use the calculator's randBin function to simulate the random selection of 10 switches from the $B(10, 0.1)$ distribution, and assign these 10 results to L_1/list1. Then use the 1-Var Stats function to find the mean number of defective switches among the 10. Compare this result with the known mean $\mu = 1$. Repeat these steps to find the mean of 25 randomly selected switches and then the mean of 50 randomly selected switches. What effect, if any, does the number of switches sampled have on the mean number of defective switches?

(b) Do the same as in (a) for the distribution $B(12, 0.75)$ of Corinne's free throws made in 12 attempts (see Examples 8.6 and 8.8). How do your results for samples of size 10, 25, and 50 compare with the true mean number of successes?

8.23 SIMULATING COMMITTEE SELECTION Refer to Exercise 8.13 (page 449). Construct a simulation to estimate the probability that in a committee of 15 members, 3 or fewer members are Hispanic. Describe the design of your experiment, including the correspondence between digits and outcomes in the experiment, and report the relative frequency for 30 repetitions.

8.24 STUDENT INDEBTEDNESS According to the General Accounting Office and the student loan agency Nellie Mae, the average college student credit-card debt in 2000 was $2,748, and a third of students have four or more credit cards. Assume that a randomly selected student has probability 0.33 of having four or more credit cards. Use simulation methods

to determine the probability that more than 12 students in a sample of 30 have four or more credit cards.

8.25 SIMULATING MARRIAGE Refer to Exercise 8.8 (page 446). Construct a simulation to estimate the probability that 2 or fewer of a random sample of 10 employed women have never been married. Describe the design of your experiment, including the correspondence between digits and outcomes in the experiment and the number of repetitions you carried out. Report your results.

8.26 DRAWING POKER CHIPS There are 50 poker chips in a container, 25 of which are red, 15 white, and 10 blue. You draw a chip without looking 25 times, each time returning the chip to the container.

(a) What is the expected number of white chips you will draw in 25 draws?

(b) What is the standard deviation of the number of blue chips that you will draw?

(c) Simulate 25 draws by hand or by calculator. Repeat the process as many times as you think necessary.

(d) Based on your answers to parts (a) to (c), is it likely or unlikely that you will draw 9 or fewer blue chips?

(e) Is it likely or unlikely that you will draw 15 or fewer blue chips?

SUMMARY

A count X of successes has a binomial distribution in the **binomial setting**: there are n observations; the observations are **independent** of each other; each observation results in a success or a failure; and each observation has the same probability p of a success.

If X has the binomial distribution with parameters n and p, the possible values of X are the whole numbers 0, 1, 2, ..., n. The **binomial probability** that X takes any value is

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

The **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

counts the number of ways k successes can be arranged among n observations. Here the **factorial** $n!$ is

$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

for positive whole numbers n, and $0! = 1$.

Given a random variable X, the **probability distribution function** (pdf) assigns a probability to each value of X. For each value of X, the **cumulative distribution function** (cdf) assigns the sum of the probabilities for values less than or equal to X.

The **mean** and **standard deviation** of a binomial count X are

$$\mu = np$$
$$\sigma = \sqrt{np(1-p)}$$

The **normal approximation** to the binomial distribution says that if X is a count having the binomial distribution with parameters n and p, then when n is large, X is approximately $N\left(np, \sqrt{np(1-p)}\right)$. We will use this approximation when $np \geq 10$ and $n(1-p) \geq 10$.

SECTION 8.1 EXERCISES

8.27 RANDOM DIGITS Each entry in a table of random digits like Table B has probability 0.1 of being a 0, and digits are independent of each other.

(a) What is the probability that a group of five digits from the table will contain at least one 0?

(b) What is the mean number of 0s in lines 40 digits long?

8.28 TESTING ESP In a test for ESP (extrasensory perception), a subject is told that cards the experimenter can see but the subject cannot contain either a star, a circle, a wave, or a square. As the experimenter looks at each of 20 cards in turn, the subject names the shape on the card. A subject who is just guessing has probability 0.25 of guessing correctly on each card.

(a) The count of correct guesses in 20 cards has a binomial distribution. What are n and p?

(b) What is the mean number of correct guesses in many repetitions?

(c) What is the probability of exactly 5 correct guesses?

8.29 RANDOM STOCK PRICES A believer in the "random walk" theory of stock markets thinks that an index of stock prices has probability 0.65 of increasing in any year. Moreover, the change in the index in any given year is not influenced by whether it rose or fell in earlier years. Let X be the number of years among the next 5 years in which the index rises.

(a) X has a binomial distribution. What are n and p?

(b) What are the possible values that X can take?

(c) Find the probability of each value X. Draw a probability histogram for the distribution of X.

(d) What are the mean and standard deviation of this distribution? Mark the location of the mean on the histogram.

8.30 LIE DETECTORS A federal report finds that lie detector tests given to truthful persons have probability about 0.2 of suggesting that the person is deceptive.[2]

(a) A company asks 12 job applicants about thefts from previous employers, using a lie detector to assess their truthfulness. Suppose that all 12 answer truthfully. What is the probability that the lie detector says all 12 are truthful? What is the probability that the lie detector says at least 1 is deceptive?

(b) What is the mean number among 12 truthful persons who will be classified as deceptive? What is the standard deviation of this number?

(c) What is the probability that the number classified as deceptive is less than the mean?

8.31 A MARKET RESEARCH SURVEY Return to the restaurant sample described in Exercise 8.20 (page 455). You find 100 of your 200 respondents concerned about nutrition. Is this reason to believe that the percent in your area is higher than the national 40%? To answer this question, find the probability that X is 100 or larger if $p = 0.4$ is true. If this probability is very small, that is reason to think that p is actually greater than 0.4.

8.32 PLANNING A SURVEY You are planning a sample survey of small businesses in your area. You will choose an SRS of businesses listed in the telephone book's Yellow Pages. Experience shows that only about half the businesses you contact will respond.

(a) If you contact 150 businesses, it is reasonable to use the binomial distribution with $n = 150$ and $p = 0.5$ for the number X who respond. Explain why.

(b) What is the expected number (the mean) who will respond?

(c) What is the probability that 70 or fewer will respond? (Use the normal approximation.)

(d) How large a sample must you take to increase the mean number of respondents to 100?

8.33 ARE WE SHIPPING ON TIME? Your mail-order company advertises that it ships 90% of its orders within three working days. You select an SRS of 100 of the 5000 orders received in the past week for an audit. The audit reveals that 86 of these orders were shipped on time.

(a) If the company really ships 90% of its orders on time, what is the probability that 86 or fewer in an SRS of 100 orders are shipped on time?

(b) A critic says, "Aha! You claim 90%, but in your sample the on-time percentage is only 86%. So the 90% claim is wrong." Explain in simple language why your probability calculation in (a) shows that the result of the sample does not refute the 90% claim.

8.34 AIDS TEST A test for the presence of antibodies to the AIDS virus in blood has probability 0.99 of detecting the antibodies when they are present. Suppose that during a year 20 units of blood with AIDS antibodies pass through a blood bank.

(a) Take X to be the number of these 20 units that the test detects. What is the distribution of X?

(b) What is the probability that the test detects all 20 contaminated units? What is the probability that at least 1 unit is not detected?

(c) What is the mean number of units among the 20 that will be detected? What is the standard deviation of the number detected?

8.35 SIMULATING GRADUATION Refer to Exercise 8.7 (page 446). Construct a simulation to estimate the probability that at most 11 of the 20 basketball players graduated. Describe the design of your experiment, including the correspondence between digits and outcomes in the experiment and the number of repetitions you carried out. Report your results.

8.36 ALLERGY RELIEF Clinical trials of the popular allergy medicine Allegra-D (fexofenadine HCl 60 mg/pseudoephedrine HCl 120 mg Extended Release Tablets) found that 13% of the 215 subjects reported headache as an adverse reaction to the drug.[3] Assume that in fact 13% of all users of this medicine experience headaches after taking this medicine. Suppose that 8 allergy sufferers are selected at random. This exercise uses the statistical software Minitab to answer several questions. To calculate binomial probabilities with Minitab, begin by entering the integers 0 to 8 in column 1 and naming this column VALUES. Then select **Calc > Probability Distributions > Binomial.** Then select **Probability** to indicate that you want individual probabilities. Specify the **Number of trials** and the **Probability of success.** Select **input column** and specify VALUES in column C1 to tell Minitab to calculate binomial probabilities for each of the values in that column. Then click **OK.** The following results are produced:

```
MTB >  PDF 'VALUES';
SUBC>    Binomial 8 .13.
          K          P( X = K)
       0.00             0.3282
       1.00             0.3923
       2.00             0.2052
       3.00             0.0613
       4.00             0.0115
       5.00             0.0014
       6.00             0.0001
       7.00             0.0000
       8.00             0.0000
MTB>
```

To calculate the cumulative distribution, make the same menu choices, except this time select **Cumulative probability** instead of **Probability.** The following output is produced:

```
MTB >  CDF 'VALUES';
SUBC>    Binomial 8 .13.
          K          P( X LESS OR = K)
       0.00             0.3282
       1.00             0.7206
       2.00             0.9257
       3.00             0.9871
       4.00             0.9985
       5.00             0.9999
       6.00             1.0000
       7.00             1.0000
       8.00             1.0000
```

From the printouts, calculate the probability that out of the 8 randomly selected subjects, the number experiencing headaches is

(a) exactly 3

(b) at most 2

(c) less than 2

(d) at least 3 but no more than 5

(e) either less than 2 or more than 5

8.2 THE GEOMETRIC DISTRIBUTIONS

In the case of a binomial random variable, the number of trials is fixed beforehand, and the binomial variable X counts the number of successes in that fixed number of trials. If there are n trials then the possible values of X are 0, 1, 2, ..., n. By way of comparison, there are situations in which the goal is to obtain a fixed number of successes. In particular, if the goal is to obtain one success, a random variable X can be defined that counts the number of trials needed to obtain that first success. A random variable that satisfies the above description is called *geometric*, and the distribution produced by this random variable is called a ***geometric distribution***. The possible values of a geometric random variable are 1, 2, 3, . . . , that is, an infinite set, because it is theoretically possible to proceed indefinitely without ever obtaining a success. Consider the following situations:

geometric distribution

- Flip a coin until you get a *head*.
- Roll a die until you get a 3.
- In basketball, attempt a three-point shot until you make a *basket*.

Notice that all of these situations involve counting the number of trials until an event of interest happens. We are now ready to characterize the geometric setting.

A random variable X is geometric provided that the following conditions are met:

THE GEOMETRIC SETTING

1. Each observation falls into one of just two categories, which for convenience we call "success" or "failure."
2. The probability of a success, call it p, is the same for each observation.
3. The observations are all **independent**.
4. The variable of interest is the number of trials required to obtain the first success.

EXAMPLE 8.15 ROLL A DIE

An experiment consists of rolling a single die. The event of interest is rolling a 3; this event is called a success. The random variable is defined as X = the number of trials until a 3 occurs. To verify that this is a geometric setting, note that rolling a 3 will represent a success, and rolling any other number will represent a failure. The probability of rolling a 3 on each roll is the same: 1/6. The observations are independent. A trial consists of rolling the die once. We roll the die until a 3 appears. Since all of the requirements are satisfied, this experiment describes a geometric setting.

EXAMPLE 8.16 DRAW AN ACE

Suppose you repeatedly draw cards without replacement from a deck of 52 cards until you draw an ace. There are two categories of interest: ace = success; not ace = failure. But is the probability of success the same for each trial? No. The probability of an ace on the first card is 4/52. If you don't draw an ace on the first card, then the probability of an ace on the second card is 4/51. Since the result of the first draw affects probabilities on the second draw (and on all successive draws required), the trials are not independent. So this is not a geometric setting.

Using the setting of Example 8.15, let's calculate some probabilities.

$X = 1$: $P(X = 1) = P(\text{success on first roll}) = 1/6$

$X = 2$: $P(X = 2) = P(\text{success on second roll})$

$= P(\text{failure on first roll and success on second roll})$

$= P(\text{failure on first roll}) \times P(\text{success on second roll})$

$= (5/6) \times (1/6)$

(since trials are independent).

$X = 3$: $P(X = 3) = P(\text{failure on first roll}) \times P(\text{failure on second roll})$

$\times P(\text{success on third roll})$

$= (5/6) \times (5/6) \times (1/6)$

Continue the process. The pattern suggests that a general formula for the variable X is

$$P(X = n) = (5/6)^{n-1}(1/6)$$

Now we can state the following principle:

RULE FOR CALCULATING GEOMETRIC PROBABILITIES

If X has a geometric distribution with probability p of success and $(1 - p)$ of failure on each observation, the possible values of X are 1, 2, 3, ... If n is any

RULE FOR CALCULATING GEOMETRIC PROBABILITIES (*continued*)

one of these values, the probability that the first success occurs on the nth trial is

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

Although the setting for the geometric distribution is very similar to the binomial setting, there are some striking differences. In rolling a die, for example, it is possible that you will have to roll the die many times before you roll a 3. In fact, it is theoretically possible to roll the die forever without rolling a 3 (although the probability gets closer and closer to 0 the longer you roll the die without getting a 3). The probability of observing the first 3 on the fiftieth roll of the die is $P(X = 50) = 0.0000$.

A probability distribution table for the geometric random variable is strange indeed because it never ends; that is, the number of table entries is infinite. The rule for calculating geometric probabilities shown above can be used to construct the table:

X	1	2	3	4	5	6	7...
$P(X)$	p	$(1-p)p$	$(1-p)^2p$	$(1-p)^3p$	$(1-p)^4p$	$(1-p)^5p$	$(1-p)^6p$...

The probabilities (i.e., the entries in the second row) are the terms of a *geometric sequence* (hence the name for this random variable). You may recall from your study of algebra that the general form for a geometric sequence is

$$a, ar, ar^2, ar^3, \ldots, ar^{n-1}, \ldots$$

where a is the first term, r is the ratio of one term in the sequence to the next, and the nth term is ar^{n-1}. You may also recall that even though the sequence continues forever, and even though you could never finish adding the terms, the sequence does have a sum (one of the implausible truths of the infinite!). This sum is

$$\frac{a}{1-r}$$

In order for the geometric random variable to have a valid pdf, the probabilities in the second row of the table must add to 1. Using the formula for the sum of a geometric sequence, we have

$$\sum_{i=1}^{\infty} P(x_i) = p + (1-p)p + (1-p)^2p + \cdots$$

$$= \frac{p}{1-(1-p)} = \frac{p}{p} = 1$$

EXAMPLE 8.17 ROLL A DIE

The rule for calculating geometric probabilities can be used to construct a probability distribution table for X = number of rolls of a die until a 3 occurs:

X	1	2	3	4	5	6	7	...
$P(X)$	0.1667	0.1389	0.1157	0.0965	0.0804	0.0670	0.0558	...

Here's one way to find these probabilities with your calculator:

1. Enter the probability of success, 1/6. Press ENTER.
2. Enter *(5/6) and press ENTER.
3. Continue to press ENTER repeatedly.

```
1/6
        .1666666667
Ans*(5/6)
        .1388888889
        .1157407407
        .0964506173
        .0803755144
```

Verify that the entries in the second row are as shown:

X	1	2	3	4
$P(X)$	1/6	5/36	25/216	125/1296

Figure 8.4 is a graph of the distribution of X. As you might expect, the probability distribution histogram is strongly skewed to the right with a peak at the leftmost value, 1. It is easy to see why this must be so, since the height of each bar after the first is the height of the previous bar times the probability of failure $1 - p$. Since you're multiplying the height of each bar by a number less than 1, each new bar will be shorter than the previous bar, and hence the histogram will be right-skewed. Always.

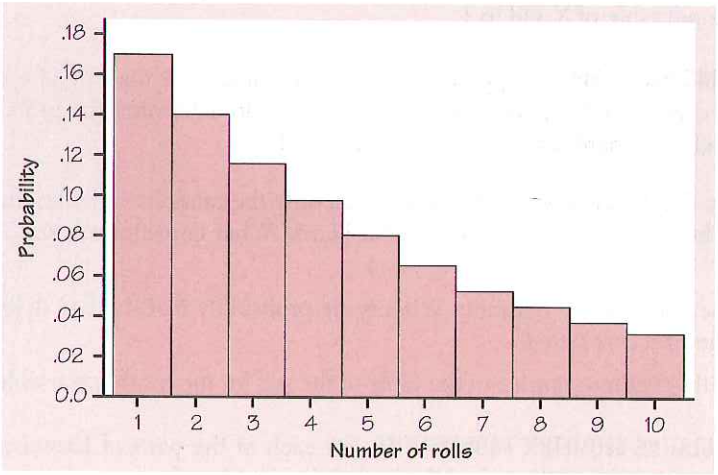


FIGURE 8.4 Probability histogram for the geometric distribution.

EXERCISES

8.37 GEOMETRIC SETTING? For each of the following, determine if the experiment describes a geometric distribution. If it does, describe the two events of interest (success and failure), what constitutes a trial, and the probability of success on one trial. If the random variable is not geometric, identify a condition of the geometric setting that is not satisfied.

(a) Flip a coin until you observe a tail.

(b) Record the number of times a player makes both shots in a one-and-one foul-shooting situation. (In this situation, you get to attempt a second shot only if you make your first shot.)

(c) Draw a card from a deck, observe the card, and replace the card within the deck. Count the number of times you draw a card in this manner until you observe a jack.

(d) Buy a "Match 6" lottery ticket every day until you win the lottery. (In a "Match 6" lottery, a player chooses 6 different numbers from the set {1, 2, 3, ..., 44}. A lottery representative draws 6 different numbers from this set. To win, the player must match all 6 numbers, in any order.)

(e) There are 10 red marbles and 5 blue marbles in a jar. You reach in and, without looking, select a marble. You want to know how many marbles you will have to draw (without replacement), on average, in order to be sure that you have 3 red marbles.

8.38 ROLL A PRIME An experiment consists of rolling a die until a prime number (2, 3, or 5) is observed. Let X = number of rolls required to get the first prime number.

(a) Verify that X has a geometric distribution.

(b) Construct a probability distribution table to include at least 5 entries for the probabilities of X. Record probabilities to four decimal places.

(c) Construct a graph of the pdf of X.

(d) Compute the cdf of X and plot its histogram.

(e) Use the formula for the sum of a geometric sequence to show that the probabilities in the pdf table of X add to 1.

8.39 TESTING HARD DRIVES Suppose we have data that suggest that 3% of a company's hard disk drives are defective. You have been asked to determine the probability that the first defective hard drive is the fifth unit tested.

(a) Verify that this is a geometric setting. Identify the random variable; that is, write X = number of _____ and fill in the blank. What constitutes a success in this situation?

(b) Answer the original question: What is the probability that the first defective hard drive is the fifth unit tested?

(c) Find the first four entries in the table of the pdf for the random variable X.

8.40 CALCULATING GEOMETRIC PROBABILITIES For each of the parts of Exercise 8.37 that describes a geometric setting, find the probability that X = 4.

The expected value and other properties of the geometric random variable

If you're flipping a fair coin, how many times would you expect to have to flip the coin in order to observe the first head? If you're rolling a die, how many times would you expect to have to roll the die in order to observe the first 3? If you said 2 coin tosses and 6 rolls of the die, then your intuition is serving you well. To derive an expression for the mean (expected value) of a geometric random variable, we begin with the probability distribution table. The notation will be simplified if we let p = probability of success and let q = probability of failure. Then $q = 1 - p$ and the probability distribution table looks like this:

X	1	2	3	4	...
P(X)	p	pq	pq^2	pq^3	...

The mean (expected value) of X is calculated as follows:

$$\begin{aligned}\mu_X &= 1(p) + 2(pq) + 3(pq^2) + 4(pq^3) + \cdots \\ &= p(1 + 2q + 3q^2 + 4q^3 + \cdots)\end{aligned}$$

Multiplying both sides by q, we have

$$q\mu_X = p(q + 2q^2 + 3q^3 + 4q^4 + \cdots)$$

Now subtract this equation from the previous equation, and group like terms on the right.

$$\mu_X - q\mu_X = p(1 + q + q^2 + q^3 + \cdots)$$

$$\mu_X(1-q) = p\left(\frac{1}{1-q}\right)$$

$$\mu_X = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

Deriving the variance and standard deviation of the geometric random variable X is considerably more work and would take us too far afield.

Here are the facts:

THE MEAN AND STANDARD DEVIATION OF A GEOMETRIC RANDOM VARIABLE

If X is a geometric random variable with probability of success p on each trial, then the **mean**, or **expected value**, of the random variable, that is, the expected number of trials required to get the first success, is $\mu = 1/p$. The variance of X is $(1 - p)/p^2$.

EXAMPLE 8.18 ARCADE GAME

Glenn likes the game at the state fair where you toss a coin into a saucer. You win if the coin comes to rest in the saucer without sliding off. Glenn has played this game many times and has determined that on average he wins 1 out of every 12 times he plays. He believes that his chances of winning are the same for each toss. He has no reason to think that his tosses are not independent. Let X be the number of tosses until a win. Glenn believes that this describes a geometric setting.

Since $E(X) = 12 = 1/p$, the probability of success on any given trial is

$$p = 1/12 = 0.0833$$

The variance of X is

$$\sigma_X^2 = \frac{1-p}{p^2} = \frac{11/12}{1/144} = 132$$

And the standard deviation is $\sigma_X \simeq 11.5$

There is another interesting result that relates to the probability that it takes *more* than a certain number of trials to achieve success. Here are the steps:

$$\begin{aligned} P(X > n) &= 1 - P(X \le n) \\ &= 1 - (p + qp + q^2p + \cdots + q^{n-1}p) \\ &= 1 - p(1 + q + q^2 + \cdots + q^{n-1}) \\ &= 1 - p\left(\frac{1-q^n}{1-q}\right) \\ &= 1 - p\left(\frac{1-q^n}{p}\right) \\ &= 1 - (1 - q^n) \\ &= q^n = (1-p)^n \end{aligned}$$

We summarize as follows:

$P(X > n)$

The probability that it takes *more* than n trials to see the first success is

$$P(X > n) = (1 - p)^n$$

EXAMPLE 8.19 APPLYING THE FORMULA

Roll a die until a 3 is observed. The probability that it takes more than 6 rolls to observe a 3 is

$$P(X > 6) = (1 - p)^n = (5/6)^6 \cong 0.335$$

Let Y be the number of Glenn's coin tosses until a coin stays in the saucer (see Example 8.18). The expected number is 12. The probability that it takes more than 12 tosses to win a stuffed animal is

$$P(X > 12) = (11/12)^{12} \cong 0.352$$

The probability that it takes more than 24 tosses to win a stuffed animal is

$$P(X > 24) = (11/12)^{24} \cong 0.124$$

The following Technology Toolbox summarizes some calculator techniques when working in a geometric setting:

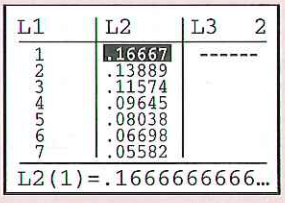
TECHNOLOGY TOOLBOX *Exploring geometric distributions*

For illustration purposes, we will use the roll of a die with $n = 6$ equally likely outcomes and probability $p = 1/6$ of rolling a 3, from Example 8.15 (page 465). The random variable X is the number of rolls until a 3 is observed.

To have the calculator calculate the probability distribution table and plot a histogram for the distribution, proceed as follows:

TI-83

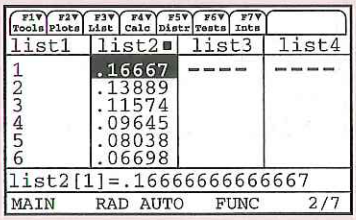
1. Enter the numbers 1 to 10 in list L_1. Next, enter the probabilities into L_2 by first highlighting L_2. Then press 2nd VARS (DISTR). Scroll down and select `D:geometpdf(`. Complete the command: `geometpdf(1/6,L1)`, and press ENTER. Here are the results:



2. Specify the dimensions of an appropriate viewing window. Scanning the list of values gives you insight into reasonable dimensions for the window. Specify $X[0,11]_1$ and $Y[-.05,.2]_{.1}$.

TI-89

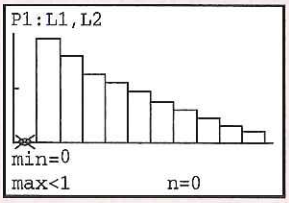
1. Enter the numbers 1 to 10 in list1. Next, enter the probabilities into list2 by first highlighting list2. Then press CATALOG F3 (Flash Apps) and scroll down to select `geomPdf(`. Complete the command: `TIStat.geomPdf(1/6,list1)`. Here are the results:



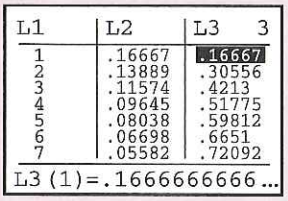
2. Specify these dimensions for the viewing window: $X[0,11]_1$ and $Y[-.05,.2]_{.1}$.

TECHNOLOGY TOOLBOX *Exploring geometric distributions (continued)*

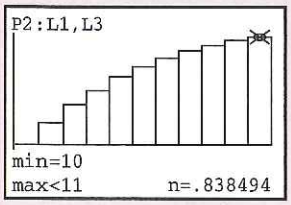
3. When you define a histogram for Plot1, specify Xlist: L_1 and Freq: L_2. The resulting plot shows that the distribution is strongly right-skewed.



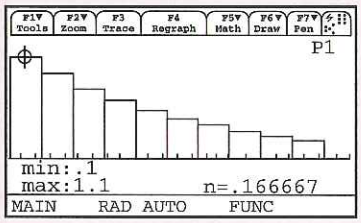
4. Next install the cdf as list L_3. In the STAT/ Edit window, place the cursor on list L_3. Enter the formula `geometcdf(1/6,L1)` and press ENTER.



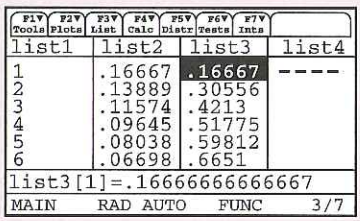
5. To plot the cumulative distribution histogram, first specify the viewing window: $X[0,11]_1$ and $Y[-.3,1]_1$. The deselect Plot1 and define Plot2 to be a histogram with Xlist: L_1 and Freq: L_3. Press GRAPH. Here is the cdf histogram:



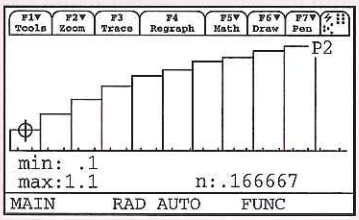
3. From the Statistics/List Editor, press F2 (Plots). Select 1:Plot Setup. Define Plot 1 to be a histogram using list1 for the X-values and list2 for the frequency. To plot the histogram, press ◆ GRAPH. Here is the pdf histogram:



4. To calculate the cumulative distribution values, highlight list3 and press CATALOG F3 (Flash Apps), then scroll down and select `geomCdf(`. Complete the command `tistat.geomCdf(1/6,list1)`. Here is the geometric cdf:



5. Deselect Plot 1. Then define Plot 2 to be a histogram using list3 for the frequency. Here is the histogram for the cdf:



Simulating geometric experiments

Geometric simulations are frequently called "waiting time" simulations because you continue to conduct trials and wait until a "success" is observed.

Conducting a geometric simulation by hand is generally pretty easy but tedious. Conducting a geometric simulation by calculator or computer usually takes more effort initially, but the payoff is that you can quickly run a rather large number of repetitions and frequently get results that are quite respectable. Here is an example:

EXAMPLE 8.20 SHOW ME THE MONEY!

In 1986–1987, Cheerios™ cereal boxes displayed a dollar bill on the front of the box and a cartoon character who said, "Free $1 bill in every 20th box." Here is a simulation to determine the number of boxes of Cheerios you would expect to buy in order to get one of the "free" dollar bills.

Let a two-digit number, 00 to 99, represent a box of Cheerios, and let the digits 01 to 05 represent a box of Cheerios with a $1 bill in it. The digits 00 and 06 to 99 represent boxes without the $1 bill. Starting at the third block of digits in line 127 of Table B, we select digits in pairs:

23	33	06	43	59	40	08	61	69	25
85	11	73	60	71	15	68	91	42	27
06	56	51	43	74	13	35	24	93	67
81	98	28	72	09	36	75	95	89	84
68	28	82	29	13	18	63	84	43	03

So in our first run of this simulation, we had to buy 50 boxes of Cheerios until we found one with a $1 bill in it! If you don't usually buy Cheerios, would this promotion induce you to buy a box in hopes of getting one with a dollar in it?

Why did it take so many boxes (50) to achieve success? Since the probability of success on a single trial is $p = 1/20 = 0.05$, we know that the mean (expected value) is $E(X) = 1/p = 20$, so a value of 50 in our simulation seems high. But the variance is

$$\sigma^2(X) = \frac{1-p}{p^2} = \frac{0.95}{0.0025} = 380$$

and

$$\sigma(X) = \sqrt{380} \cong 19.49$$

Our simulated result of 50 is about 1.5 standard deviations to the right of the mean, 20. So perhaps we should not be too surprised with our 50. (Keep in mind that the standard deviation is not an appropriate measure of spread for strongly skewed distributions, and this geometric distribution *is* strongly skewed right.)

EXERCISES

8.41 FLIP A COIN Consider the following experiment: flip a coin until a head appears.

(a) Identify the random variable X.

(b) Construct the pdf table for X. Then plot the probability histogram.

(c) Compute the cdf and plot its histogram.

8.42 ARCADE GAME Refer to Example 8.19 on page 471.

(a) Use the formula for calculating $P(X > n)$ on page 470 to find the probability that it takes more than 10 tosses until Glenn wins a stuffed animal.

(b) Find the answer to (a) by calculating the probability of the complementary event: $1 - P(X \leq 10)$. Your results should agree, of course.

(*Note:* The formula for $P(X > n)$ is not practically important since there are other ways to answer the question. But it's a nice little result, and it's quite easy to derive.)

8.43 ROLL A DIE

(a) Plot the cumulative distribution histogram for the die-rolling experiment described in Example 8.15 with the pdf table in Example 8.17.

(b) Find the probability that it takes more than 10 rolls to observe a 3.

(c) Find the smallest positive integer k for which $P(X \leq k) > 0.99$.

8.44 LANGUAGE SKILLS The State Department is trying to identify an individual who speaks Farsi to fill a foreign embassy position. They have determined that 4% of the applicant pool are fluent in Farsi.

(a) If applicants are contacted randomly, how many individuals can they expect to interview in order to find one who is fluent in Farsi?

(b) What is the probability that they will have to interview more than 25 until they find one who speaks Farsi? More than 40?

8.45 SHOOTING FREE THROWS A basketball player makes 80% of her free throws. We put her on the free-throw line and ask her to shoot free throws until she misses one. Let X = the number of free throws the player takes until she misses.

(a) What assumption do you need to make in order for the geometric model to apply? With this assumption, verify that X has a geometric distribution. What action constitutes "success" in this context?

(b) What is the probability that the player will make 5 shots before she misses?

(c) What is the probability that she will make at most 5 shots before she misses?

8.46 GAME OF CHANCE Three friends each toss a coin. The odd man wins; that is, if one coin comes up different from the other two, that person wins that round. If the coins all match, then no one wins and they toss again. We're interested in the number of times the players will have to toss the coins until someone wins.

(a) What is the probability that no one will win on a given coin toss?

(b) Define a success as "someone wins on a given coin toss." What is the probability of a success?

(c) Define the random variable of interest: X = number of _____. Is X binomial? Geometric? Justify your answer.

(d) Construct a probability distribution table for X. Then extend your table by the addition of cumulative probabilities in a third row.

(e) What is the probability that it takes no more than 2 rounds for someone to win?

(f) What is the probability that it takes more than 4 rounds for someone to win?

(g) What is the expected number of tosses needed for someone to win?

(h) Use the `randInt` function on your calculator to simulate 25 rounds of play. Then calculate the relative frequencies for X = 1, 2, 3, ... Compare the results of your simulation with the theoretical probabilities you calculated in (d).

SUMMARY

A count X of successes has a **geometric distribution** in the geometric setting if the following are satisfied: each observation results in a success or a failure; each observation has the same probability p of success; observations are independent; and X counts the number of trials required to obtain the first success. A geometric random variable differs from a binomial variable because in the geometric setting the number of trials varies and the desired number of defined successes (1) is fixed in advance.

If X has the geometric distribution with probability of success p, the possible values of X are the positive integers 1, 2, 3, ... The **geometric probability** that X takes any value is

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

The **mean** (expected value) of a geometric count X is $1/p$.

The standard deviation is

$$\sqrt{\frac{(1-p)}{p^2}}$$

The probability that it takes *more* than n trials to see the first success is

$$P(X > n) = (1 - p)^n$$

SECTION 8.2 EXERCISES

8.47 DRAWING MARBLES There are 20 red marbles, 10 blue marbles, and 5 white marbles in a jar. An experiment consists of selecting a marble without looking, noting the color, and then replacing the marble in the jar. We're interested in the number of marbles you would have to draw in order to be sure you have a red marble.

(a) Is this a binomial or a geometric setting? Explain your choice, and write a description of the random variable X.

(b) Calculate the probability of drawing a red marble on the second draw. Calculate the probability of drawing a red marble by the second draw. Calculate the probability that it would take more than 2 draws to get a red marble.

(c) What single calculator command will install the first 20 values of X into L_1/list1? What single command will install the corresponding probabilities into L_2/list2? What single command will install the cumulative probabilities into L_3/list3? Enter these commands in the Home screen. Copy this information from your calculator onto your paper to make an expanded probability distribution table (with the cdf as the third row).

(d) Construct a probability distribution histogram as STAT PLOT1, and then construct a cumulative distribution histogram as STAT PLOT2.

8.48 DRAWING MARBLES II This is a continuation of Exercise 8.47. Given the jar containing red, white, and blue marbles, Joey thinks a more interesting problem would be to find the number of marbles you would have to draw, without replacing them in the jar, to be sure that you have 2 red marbles.

(a) Does this experiment describe a geometric setting? Why or why not?

(b) Would your answer to (a) change if the marble was replaced after each draw? Explain.

(c) Design and carry out a simulation to determine the number of marbles you would have to draw, with replacement, until you get 2 red marbles. Compare the results from your simulation with the results from the previous exercise.

8.49 MULTIPLE-CHOICE Carla makes random guesses on a multiple-choice test that has five choices for each question. We want to know how many questions Carla answers until she gets one correct.

(a) Define a success in this context, and define the random variable X of interest. What is the probability of success?

(b) What is the probability that Carla's first correct answer occurs on problem 5?

(c) What is the probability that it takes more than 4 questions before Carla answers one correctly?

(d) Construct a probability distribution table for X.

(e) If Carla took a test like this test many times and randomly guessed at each question, what would be the average number of questions she would have to answer before she answered one correctly?

8.50 IT'S A BOY! In some cultures, it is considered very important to have a son to carry on the family name. Suppose that a couple in one of these cultures plans to have children until they have exactly one son.

(a) Find the average number of children per family in such a culture.

(b) What is the expected number of girls in this family?

(c) Describe a simulation that could be used to find approximate answers to the questions in (a) and (b).

8.51 FAMILY PLANNING, I Example 5.24 (page 313) used simulation techniques to explore the following situation: A couple plans to have children until they have a girl or until they have four children, whichever comes first.

(a) List the outcomes in the sample space for this "experiment." What event represents a success?

(b) Let X = the number of boys in this family. What values can X take? Use an appropriate probability rule to calculate the probability for each value of X, and make a probability distribution table for X. Then show that the sum of the probabilities is 1.

(c) Let Y = the number of children produced in this family until a girl is produced. Show that Y starts out as a geometric distribution but then is stopped abruptly. Make a probability distribution table for Y.

(d) What is the expected number of children for this couple?

(e) What is the probability that this couple will have more than the expected number of children?

(f) At the end of Example 5.24, it states that the probability of having a girl in this situation is 0.938. How can you prove this?

8.52 FAMILY PLANNING, II This is a continuation of Exercise 8.51. A couple plans to have children until they have a girl or until they have four children, whichever comes first. Use the random number table (Table B), beginning on line 130, to simulate 25 repetitions of this childbearing strategy. As in Example 5.24, since a girl and boy are equally likely, let the digits 0 to 4 represent a girl, and let digits 5 to 9 represent a boy. Write the digits in a string until you observe a girl, write B or G under each digit, and write the number of children noted at the bottom. The first two repetitions would be recorded as

6	9	0		5	1
B	B	G		B	G
	3				2

Then find the mean of the 25 repetitions. How do your results compare with the theoretical expected value of 1.8 children?

8.53 FAMILY PLANNING, III This is a continuation of Exercises 8.51 and 8.52. Devise a simulation procedure for the calculator to approximate the expected number of children. List the steps and commands you use as well as the number of repetitions and the results. Alternatively, incorporate these steps into a calculator program similar to the programs SPIN123 (page 329) or FLIP50 (page 92).

8.54 MAKING THE CONNECTION This exercise provides visual reinforcement of the relationship between the probability of success and the mean (expected value) of a geometric random variable.

(a) Begin by completing the table below, where X = probability of success and Y = expected value.

X	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
Y									

(b) Make a scatterplot of the points (X, Y).

(c) Enter the data into your calculator, and transform the data assuming a power function model.

(d) Remember that the purpose of transforming data is to make the data points linear so that the method of least squares can be employed. Sketch the plot of the transformed data.

(e) What is the correlation, r, for the transformed data?

(f) Write the equation of the power function. Draw the power function curve on your scatterplot.

(g) Briefly explain the connection between this curve and what you have learned about the expected value of a geometric random variable.

CHAPTER REVIEW

The previous chapter introduced discrete and continuous random variables and described methods for finding means and variances, as well as rules for means and variances. This chapter focused on two important classes of discrete random variables, each of which involves two outcomes or events of interest. Both require independent trials and the same probability of success on each trial. The **binomial** random variable requires a fixed number of trials; the **geometric** random variable has the property that the number of trials varies. Both the binomial and the geometric settings occur sufficiently often in applications that they deserve special attention. Here is a checklist of the major skills you should have acquired by studying this chapter:

A. BINOMIAL

1. Identify a random variable as binomial by verifying four conditions: two outcomes (success and failure); fixed number of trials; independent trials; and the same probability of success for each trial.

2. Use a TI-83/89 or the formula to determine binomial probabilities and to construct probability distribution tables and histograms.

3. Calculate cumulative distribution functions for binomial random variables, and construct cumulative distribution tables and histograms.

4. Calculate means (expected values) and standard deviations of binomial random variables.

5. Use a normal approximation to the binomial distribution to compute probabilities.

B. GEOMETRIC

1. Identify a random variable as geometric by verifying four conditions: two outcomes (success and failure); the same probability of success for each trial; independent trials; and the count of interest is the number of trials required to get the first success.

2. Use formulas or a TI-83/89 to determine geometric probabilities and to construct probability distribution tables and histograms.

3. Calculate cumulative distribution functions for geometric random variables, and construct cumulative distribution tables and histograms.

4. Calculate expected values and standard deviations of geometric random variables.

CHAPTER 8 REVIEW EXERCISES

8.55 BINOMIAL SETTING? In each of the following cases, decide whether or not a binomial distribution is an appropriate model, and give your reasons.

(a) You want to know what percent of married people believe that mothers of young children should not be employed outside the home. You plan to interview 50 people, and for the sake of convenience you decide to interview both the husband and wife in 25 married couples. The random variable X is the number among the 50 persons interviewed who think mothers should not be employed.

(b) You are interested in attitudes toward drinking among the 75 members of a fraternity. You choose 25 members at random to interview. One question is "Have you had five or more drinks at one time during the last week?" Suppose that in fact 20% of the 75 members would say "Yes." Explain why you cannot safely use the $B(25, 0.2)$ distribution for the count X in your sample who say "Yes."

8.56 VIRGINIA ROAD FATALITIES In 2001 there were 930 road fatalities in Virginia, according to the Virginia Department of Motor Vehicles. Of these, 355 were alcohol-related. A DMV analyst wants to randomly select several groups of 25 road fatalities for further study. Find the mean and standard deviation for the number of alcohol-related road fatalities in such groups of 25. What is the probability that such a group will have no more than 5 alcohol-related road fatalities?

8.57 SEVEN BROTHERS! This exercise is an extension of Activity 8. There's a movie classic entitled *Seven Brides for Seven Brothers*. Even if these brothers had a few sisters, this many brothers is unusual. We will assume that there are no sisters.

(a) Let X = number of boys in a family of 7 children. Assume that sons and daughters are equally likely outcomes. Do you think the distribution of X will be skewed left, symmetric, or skewed right? The answer to this question depends on what fact?

(b) Use the `binompdf` command to construct a pdf table for X. Then construct a probability distribution histogram and a cumulative distribution histogram for X. Keep a written record of your numerical results as they are produced by your calculator, as well as sketches of the histograms.

(c) What is the probability that all of the 7 children are boys?

8.58 GET A HEAD Suppose we toss a penny repeatedly until we get a head. We want to determine the probability that the first head comes up in an *odd* number of tosses (1, 3, 5, and so on).

(a) Toss a penny until the first head occurs, and repeat the procedure 50 times. Keep a record of the results of the first toss and of the number of tosses needed to get a head on each of your 50 repetitions.

(b) Based on the result of your first toss in the 50 repetitions, estimate the probability of getting a head on the first toss.

(c) Use your 50 repetitions to estimate the probability that the first head appears on an odd-numbered toss.

8.59 ARMED AND DANGEROUS According to a 1997 Centers for Disease Control study of risky behavior, roughly one in five teenagers carries a weapon. Hispanics were most likely to arm themselves, with 23% carrying a gun, knife, or club, compared with 22% of blacks and 17% of whites. Suppose that 4 teenagers are selected at random and subjected to a search. Suppose that success is defined as "the teen is carrying a gun, knife, or club."

(a) What is the probability, p, of success?

(b) Make a list of all of the possible results of the search of the 4 teenagers. Use S to represent success, and F for failure. For each of these responses, write a product of four factors for that combination of successes and failures. For example, SSFS is one such response, and the probability of that outcome is $(0.2)(0.2)(0.8)(0.2) = 0.0064$. Display the probabilities to four decimal places.

(c) Draw a tree diagram to show the possible outcomes.

(d) List the outcomes in which exactly 2 of the 4 students are found to carry a gun, knife, or club.

(e) What are the probabilities of the outcomes in part (d)? Briefly explain why all of these probabilities are the same.

8.60 TOOTH DECAY AND GUM DISEASE Dentists are increasingly concerned about the growing trend of local school districts to grant soft drink companies exclusive rights to install soda pop machines in schools in return for money—usually millions—that goes directly into school coffers. According to a recent study by the National Soft Drink Association, 62% of schools nationally already have such contracts. This comes at a time when dentists are seeing an alarming increase in horribly decayed teeth and eroded enamel in the mouths of teenagers and young adults. With ready access to soft drinks, children tend to drink them all day. That, combined with no opportunity to brush, leads to disaster, dentists say. Suppose that 20 schools around the country are randomly selected and asked if they have a soft drink contract. Find the probability that the number of "Yes" answers is

(a) exactly 8

(b) at most 8

(c) at least 4

(d) between 4 and 12, inclusive

(e) Identify the random variable of interest, X. Then write the probability distribution table for X.

(f) Draw a probability histogram for X.

8.61 FAITH AND HEALING A higher percentage of southerners believe in God and prayer, according to a 1998 study by the University of North Carolina's Institute for Research in Social Science. The survey was conducted by means of telephone interviews with 844 adults in 12 southern states and 413 adults in other states. One of the findings was that 46% of southerners believe they have been healed by prayer, compared with 28% of others. Assume that the results of the UNC survey are true for the region. Suppose that 20 southerners are selected at random and asked if they believe they have been healed by prayer. Find the probability that the number who answer "Yes" to this question is

(a) exactly 10

(b) between 10 and 15

(c) over 75% of the 20

(d) less than 8

8.62 CONTAMINATED SUPERMARKET MEAT In an October 2001 study by the University of Maryland and the Food and Drug Administration, 200 samples of ground beef, ground chicken, ground turkey, and ground pork were collected from supermarkets around Washington, D.C. Forty-one samples, or about 20%, were contaminated by salmonella. Salmonella is a microorganism that can produce flu-like symptoms of fever, diarrhea, and vomiting within 12 to 36 hours after eating improperly cooked food contaminated by it. Assume that 20% of all supermarket ground meat and poultry is contaminated by salmonella. Suppose 7 ground meat and poultry samples are selected from supermarkets at random. Let X denote the number of those chosen that are contaminated by salmonella. The Minitab printout below provides the probability distribution for the random variable X.

```
MTB > PDF 'Values';
SUBC > Binomial 7 .2.
      K          P(X = K)
   0.00           0.2097
   1.00           0.3670
   2.00           0.2753
   3.00           0.1147
   4.00           0.0287
   5.00           0.0043
   6.00           0.0004
   7.00           0.0000
MTB >
```

From the printout, determine the probability that of the 7 meat and poultry samples chosen, the number contaminated by salmonella is

(a) exactly 2

(b) at least 2

(c) less than 2

(d) between 2 and 5, inclusive

8.63 FIRST HIT OF THE SEASON Suppose that Roberto, a well-known major league baseball player, finished last season with a .325 batting average. He wants to calculate the probability that he will get his first hit of this new season in his first at-bat. You define a success as getting a hit and define the random variable X = number of at-bats until Roberto gets his first hit.

(a) What is the probability that Roberto will get a hit on his first at-bat (i.e., that $X = 1$)?

(b) What is the probability that it will take him at most 3 at-bats to get his first hit?

(c) What is the probability that it will take him more than 4 at-bats to get his first hit?

(d) Roberto wants to know the expected number of at-bats until he gets a hit. What would you tell him?

(e) Enter the first 10 values of X into L_1/list1, the corresponding geometric probabilities into L_2/list2, and the cumulative probabilities into L_3/list3.

(f) Construct a probability distribution histogram as STAT PLOT1, and then construct a cumulative distribution histogram as STAT PLOT2.

You show this analysis to Roberto, and he is so impressed he gives you two free tickets to his first game.

8.64 QUALITY CONTROL Many manufacturing companies use statistical techniques to ensure that the products they make meet standards. One common way to do this is to take a random sample of products at regular intervals throughout the production shift. Assuming that the process is working properly, the mean measurements from these random samples will vary normally around the target mean μ, with a standard deviation of σ.

(a) If the process is working properly, what is the probability that 4 out of 5 consecutive sample means fall within the interval $(\mu - \sigma, \mu + \sigma)$?

(b) If the process is working properly, what is the probability that the first sample mean that is greater than $\mu + 2\sigma$ is the one from the fourth sample taken?

8.65 A MINI-THEOREM Suppose that $X = B(n, p)$. Show that $P(X \geq 1) = 1 - (1 - p)^n$.

NOTES AND DATA SOURCES

1. The survey question is reported in Trish Hall, "Shop? Many say 'Only if I must,'" *New York Times*, November 28, 1990. In fact, 66% (1650 of 2500) in the sample said "Agree."

2. Office of Technology Assessment, *Scientific Validity of Polygraph Testing: A Research Review and Evaluation*, Government Printing Office, Washington, D.C., 1983.
3. Prescribing information, including results of clinical trials, as of November 2000, appear frequently in newspaper and magazine advertisements and can be found at the Web site www.allegra.com.

Courtesy Dr. David Blackwell, UCA, Berkeley

DAVID BLACKWELL

Mathematics in the Service of Statistics

Statistical practice rests in part on statistical theory. Statistics has been advanced not only by people concerned with practical problems, from Florence Nightingale to R. A. Fisher and John Tukey, but also by people whose first love is mathematics for its own sake. *David Blackwell (1919–)* is one of the major contemporary contributors to the mathematical study of statistics.

Blackwell grew up in Illinois, earned a doctorate in mathematics at the age of 22, and in 1944 joined the faculty of Howard University in Washington, D.C. "It was the ambition of every black scholar in those days to get a job at Howard University," he says. "That was the best job you could hope for." Society changed, and in 1954 Blackwell became professor of statistics at the University of California at Berkeley.

Washington, D.C., had an active statistical community, and the young mathematician Blackwell soon began to work on mathematical aspects of statistics. He explored the behavior of statistical procedures that, rather than working with a fixed sample, keep taking observations until there is enough information to reach a firm conclusion. He found insights into statistical inference by thinking of inference as a game in which nature plays against the statistician. Blackwell's work uses probability theory, the mathematics that describes chance behavior. This chapter presents the probabilistic ideas needed to understand the reasoning of statistical inference.

Statistics has been advanced not only by people concerned with practical problems but also by people whose first love is mathematics for its own sake.