Chapter 68
The binomial, hypergeometric, and negative binomial random variables

Questions answered in this chapter:

- What is a binomial random variable?

- How do I use the BINOM.DIST and BINOM.DIST.RANGE functions to compute binomial probabilities?

- If equal numbers of people prefer Coke to Pepsi and Pepsi to Coke, and I ask 100 people whether they prefer Coke to Pepsi, what is the probability that exactly 60 people prefer Coke to Pepsi and the probability that between 40 and 60 people prefer Coke to Pepsi?

- Of all the elevator rails my company produces, 3 percent are considered defective. We are about to ship a batch of 10,000 elevator rails to a customer. To determine whether the batch is acceptable, the customer will randomly choose a sample of 100 rails and check whether each sampled rail is defective. If two or fewer sampled rails are defective, the customer will accept the batch. How can I determine the probability that the batch will be accepted?

- Airlines do not like flights with empty seats. Suppose that, on average, 95 percent of all ticket purchasers show up for a flight. If the airline sells 105 tickets for a 100-seat flight, what is the probability that the flight will be overbooked?

- The local Village Deli knows that 1,000 customers come for lunch each day. On average, 20 percent order the specialty vegetarian sandwich. These sandwiches are made in advance. How many should the deli make if they want to have a 5 percent chance of running out of vegetarian sandwiches?

- What is the hypergeometric random variable?

- What is the negative binomial random variable?

Answers to this chapter's questions

What is a binomial random variable?

A *binomial random variable* is a discrete random variable used to calculate probabilities in a situation in which all three of the following apply:

- *n* independent trials occur.

- Each trial results in one of two outcomes: success or failure.

- In each trial, the probability of success (*p*) remains constant.

In such a situation, the binomial random variable can be used to calculate probabilities related to the number of successes in a given number of trials. I let *x* be the random variable denoting the number of successes occurring in *n* independent trials, when the probability of success on each trial is *p*. Here are some examples in which the binomial random variable is relevant.

Coke or Pepsi   Assume that equal numbers of people prefer Coke to Pepsi and Pepsi to Coke. You ask 100 people whether they prefer Coke to Pepsi. You're interested in the probability that exactly 60 people prefer Coke to Pepsi and the probability that from 40 through 60 people prefer Coke to Pepsi. In this situation, you have a binomial random variable defined by the following:

- Trial: Survey individuals

- Success: Prefer Coke

- *p* equals 0.50

- *n* equals 100

Let *x* equal the number of people sampled who prefer Coke. You want to determine the probability that *x*=60 and also the probability that $40 \le x \le 60$.

Elevator rails   Of all the elevator rails you produce, 3 percent are considered defective. You are about to ship a batch of 10,000 elevator rails to a customer. To determine whether the batch is acceptable, the customer will randomly choose a sample of 100 rails and check whether each sampled rail is defective. If two or fewer sampled rails are defective, the customer will accept the batch. You want to determine the probability that the batch will be accepted.

You have a binomial random variable defined by the following:

- Trial: Look at a sampled rail

- Success: Rail is defective

- *p* equals 0.03

- *n* equals 100

Let *x* equal the number of defective rails in the sample. You want to find the probability that $x \le 2$.

Airline overbooking   Airlines don't like flights with empty seats. Suppose that, on average, 95 percent of all ticket purchasers show up for a flight. If the airline sells 105 tickets for a 100-seat flight, what is the probability that the flight will be overbooked?

You have a binomial random variable defined by the following:

- Trial: Individual ticket holders

- Success: Ticket holder shows up

- *p* equals 0.95

- *n* equals 105

Let *x* equal the number of ticket holders who show up. Then you want to find the probability that *x* ≥ 101.

How do I use the BINOM.DIST and BINOM.DIST.RANGE functions to compute binomial probabilities?

Microsoft Excel 2016 includes the BINOM.DIST and BINOM.DIST.RANGE functions, which you can use to compute binomial probabilities. If you want to compute the probability of *x* or fewer successes, for a binomial random variable having *n* trials with probability of success *p*, simply enter BINOM.DIST(x,n, p,1). If you want to compute the probability of exactly *x* successes for a binomial random variable having *n* trials with probability of success of *p*, enter BINOM.DIST(*x,n,p*,0). Entering 1 as the last argument of BINOM.DIST yields a "cumulative" probability; entering 0 yields the "probability mass function" for any particular value. (Note that a last argument of True can be used instead of a 1, and a last argument of False can be used instead of a 0.)

The function BINOM.DIST.RANGE(n,p,s1,s2) gives the probability of obtaining between *s1* and *s2* successes (inclusive) in *n* independent trials with probability of success *p* on each trial.

Here are a few examples of using the BINOM.DIST function to calculate probabilities of interest. You'll find the data and analysis in the file Binomialexamples.xlsx, which is shown in Figure 68-1.

	B	C	D	E	F
2					
3	Coke vs. Pepsi	Binom.dist			
4	Probability exactly 60 people prefer Coke to Pepsi	0.010843867	=BINOM.DIST(60,100,0.5,0)	BINOM.DIST.RANGE	
5	Probability from 40 through 60 people prefer Coke to Pepsi	0.9647998	=BINOM.DIST(60,100,0.5,1)-BINOM.DIST(39,100,0.5,1)		0.9647998 =BINOM.DIST.RANGE(100,0.5,40,60)
6					
7	Elevator Rails				
8	Probability 2 or fewer of 100 rails are defective	0.419775083	=BINOM.DIST(2,100,0.03,1)		0.419775083 =BINOM.DIST.RANGE(100,0.03,0,2)
9	Airline Overbooking				
10	Probability flight is overbooked if 105 tickets are sold	0.392433721	=1-BINOM.DIST(100,105,0.95,1)		0.392433721 =BINOM.DIST.RANGE(105,0.95,101,105) F68xx01

FIGURE 68-1 Using the binomial random variable.

If equal numbers of people prefer Coke to Pepsi and Pepsi to Coke, and I ask 100 people whether they prefer Coke to Pepsi, what is the probability that exactly 60 people prefer Coke to Pepsi and the

     3/3/2017

probability that between 40 and 60 people prefer Coke to Pepsi?

You have n=100 and p=0.5. You seek the probability that x=60 and the probability that $40 \leq x \leq 60$, where x equals the number of people who prefer Coke to Pepsi. First, you find the probability that x=60 by entering in cell C4 the formula =BINOM.DIST(60,100,0.5,0). Excel returns the value 0.011.

To use the BINOM.DIST function to compute the probability that $40 \leq x \leq 60$, you can note that the probability that $40 \leq x \leq 60$ equals the probability that $x \leq 60$ minus the probability that $x \leq 39$. Thus, you can obtain the probability that from 40 through 60 people prefer Coke by entering in cell C5 the formula =BINOM.DIST(60,100,0.5,1)–BINOM.DIST(39,100,0.5,1). Excel returns the value 0.9648. So, if Coke and Pepsi are equally preferred, it is very unlikely that in a sample of 100 people, Coke or Pepsi would be more than 10 percent ahead. If a sample of 100 people shows Coke or Pepsi to be more than 10 percent ahead, you should probably doubt that Coke and Pepsi are equally preferred. Alternatively, in cell E5 I used the formula =BINOM.DIST.RANGE(100,0.5,40,60) to compute the same probability.

Of all the elevator rails my company produces, 3 percent are considered defective. We are about to ship a batch of 10,000 elevator rails to a customer. To determine whether the batch is acceptable, the customer will randomly choose a sample of 100 rails and check whether each sampled rail is defective. If two or fewer sampled rails are defective, the customer will accept the batch. How can I determine the probability that the batch will be accepted?

If you let x equal the number of defective rails in a batch, you have a binomial random variable with n=100 and p=0.03. You seek the probability that $x \leq 2$. Simply enter in cell C8 the formula =BINOM.DIST(2,100,0.03,1). Excel returns the value 0.42 Thus, the batch will be accepted 42 percent of the time. Alternatively, in cell E7 I used the formula =BINOM.DIST.RANGE(100,0.03,0,2) to compute the same probability.

Really, your chance of success is not exactly 3 percent on each trial. For example, if the first 10 rails are defective, the chance the next rail is defective has dropped to 290/9,990; if the first 10 rails are not defective, the chance the next rail is defective is 300/9,990. Therefore, the probability of success on the eleventh trial is not independent of the probability of success on one of the first 10 trials. Despite this fact, the binomial random variable can be used as an approximation when a sample is drawn and the sample size is less than 10 percent of the total population. Here, the population equals 10,000, and the sample size is 100. Exact probabilities involving sampling from a finite population can be calculated with the *hypergeometric random variable*, which I'll discuss later in this chapter.

Airlines do not like flights with empty seats. Suppose that, on average, 95 percent of all ticket purchasers show up for a flight. If the airline sells 105 tickets for a 100-seat flight, what is the probability that the flight will be overbooked?

Let x equal the number of ticket holders who show up for the flight. You have n=105 and p=0.95. You seek the probability that $x \geq 101$. Note that the probability that $x \leq 101$ equals 1 minus the probability that $x \leq 100$. So, to compute the probability that the flight is overbooked, you enter in cell C10 the

formula =1–BINOM.DIST(100,105,0.95,1). Excel yields 0.392, which means there is a 39.2 percent chance that the flight will be overbooked. Alternatively, in cell E10 I computed the same probability by using the following formula:

`=BINOM.DIST.RANGE(105,0.95,101,105)`

The local Village Deli knows that 1,000 customers come for lunch each day. On average, 20 percent order the specialty vegetarian sandwich. These sandwiches are made in advance. How many should the deli make if they want to have a 5 percent chance of running out of vegetarian sandwiches?

In Excel 2016, the function BINOM.INV, with the syntax BINOM.INV(trials, probability of success, *alpha*), determines the smallest number x, for which the probability of less than or equal to x successes is at least *alpha*. In earlier versions of Excel, the function CRITBINOM(trials, probability of success, *alpha*) yielded the same results as BINOM.INV. In this example, *trials* equals 1,000, *probability of success* equals 0.2, and *alpha* equals 0.95. As shown in Figure 68-2, if the deli orders 221 sandwiches, the probability that demand will be less than or equal to 221 is at least 0.95. Also, the probability that 220 or fewer sandwiches will be demanded is less than 0.95.

	G	H
4	Customers	1000
5	fraction wanting veggie	0.2
6	prob of filling demand	0.95
7		
8		221 =BINOM.INV(H4,H5,H6)

F68xx02

FIGURE 68-2 An example of the BINOM.INV function.

F68xx02: This figure shows how the BINOM.INV function can be used to determine how to meet 95 percent of the demand.

What is the hypergeometric random variable?

The *hypergeometric random variable* governs a situation such as the following:

- A bowl contains N balls.

- Each ball is one of two types (called *success* or *failure*).

- There are s successes in the bowl.

- A sample of size n is drawn from the bowl.

Let's look at an example in the file Hypergeom.dist.xlsx, which is shown in Figure 68-3. The Excel 2016 formula HYPERGEOM.DIST(x,n,s,N,0) gives the probability of x successes if n balls are drawn from a bowl containing N balls, of which s are marked as *success*. The Excel 2016 formula HYPERGEOM.DIST(x,n,s,N,1) gives the probability of less than or equal to x successes if n balls are

drawn from a bowl containing *N* balls, of which *s* are marked as *success*. (As with the BINOM.DIST function, True can be used to replace 1 and False to replace 0.)

For example, suppose that 40 of the Fortune 500 companies have a woman CEO. The 500 CEOs are analogous to the balls in the bowl (*N*=500) and the 40 women are representative of the *s* successes in the bowl. Then, copying from D8 to D9:D18 the formula =HYPERGEOM.DIST(C8,Sample_size,Population_women,Population_size,FALSE) gives the probability that a sample of 10 Fortune 500 companies will have 0, 1, 2,…, 10 women CEOs. Here *Sample_size* equals 10, *Population_women* equals 40, and *Population_size* equals 500. You can substitute FALSE for *0* in the formula.

Finding a woman CEO is a success. In the sample of 10, for example, there's a probability of 0.431 that no women CEOs will be in the sample. By the way, you could have approximated this probability with the formula BINOMDIST(0,10,0.08,0), yielding 0.434, which is very close to the true probability of 0.431. In cell F10, I computed the probability that at most 2 of the 10 people in the sample would be women with the formula =HYPGEOM.DIST(2,Sample_Size,Population_women,Population_size,TRUE). Thus there is a 96.2 percent chance that at most two people in the sample will be women. Of course, I also could have obtained this answer by adding together the highlighted cells (D8:D10).
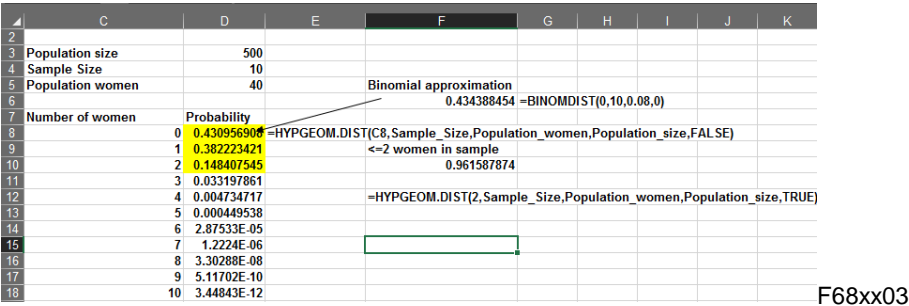
	C	D	E	F	G	H	I	J	K
2									
3	Population size	500							
4	Sample Size	10							
5	Population women	40		Binomial approximation					
6				0.434388454 =BINOMDIST(0,10,0.08,0)					
7	Number of women	Probability							
8		0	0.430956908 =HYPGEOM.DIST(C8,Sample_Size,Population_women,Population_size,FALSE)						
9		1	0.382223421		<=2 women in sample				
10		2	0.148407545		0.961587874				
11		3	0.033197861						
12		4	0.004734717		=HYPGEOM.DIST(2,Sample_Size,Population_women,Population_size,TRUE)				
13		5	0.000449538						
14		6	2.87533E-05						
15		7	1.2224E-06						
16		8	3.30288E-08						
17		9	5.11702E-10						
18		10	3.44843E-12						

F68xx03

FIGURE 68-3 Using the hypergeometric random variable.

F68xx03: This figure illustrates the use of the hypergeometric random variable to compute probabilities when sampling without replacement.

What is the negative binomial random variable?

The *negative binomial random variable* applies to the same situation as the binomial random variable, but the negative binomial random variable gives the probability of *f* failures occurring before the *s*th success. Thus =NEGBINOM.DIST(f,s,p,0) gives the probability that exactly *f* failures will occur before the *s*th success when the probability of success is *p* for each trial, and =NEGBINOM.DIST(f,s,p,1) gives the probability that at most *f* failures will occur before the *s*th success when the probability of success is *p* for each trial. For example, consider a baseball team that wins 40 percent of their games (see the file Negbinom.dist.xlsx and Figure 68-4). Copying from E9 to E34 the formula =NEGBINOM.DIST(D9,2,.4,False) gives the probability of 0, 1, 2,…, 25 losses occurring before the

second win. Note here that success equals a game won. For example, there is a 19.2 percent chance the team will lose exactly one game before winning two games. In cell G10, I used the formula =NEGBINOM.DIST(3,2,0.4,TRUE) to compute the chance that at most three losses will occur before the team wins two games (0.663). Of course, I could have obtained this answer by simply adding up the shaded cells (E9:E12).

	D	E	F	G	H	I
4	40% chance of success					
5						
6						
7						
8	Losses before 2nd win	prob		Probability <=3 failures		
9	0	0.16	=NEGBINOM.DIST(D9,2,0.4,FALSE)	before 2nd win		
10	1	0.192	=NEGBINOM.DIST(D10,2,0.4,FALSE)	0.66304		
11	2	0.1728	=NEGBINOM.DIST(D11,2,0.4,FALSE)			
12	3	0.13824	=NEGBINOM.DIST(D12,2,0.4,FALSE)			
13	4	0.10368				
14	5	0.07465				
15	6	0.05225				
16	7	0.03583				
17	8	0.02419				
18	9	0.01612				
19	10	0.01064				
20	11	0.00697				

F68xx04

FIGURE 68-4 Using the negative binomial random variable.

F68xx04: This figure shows calculations involving the negative binomial random variable.

Problems

1. Suppose that, on average, 4 percent of all CD drives received by a computer company are defective. The company has adopted the following policy: Sample 50 CD drives in each shipment, and accept the shipment if none are defective. Using this information, determine the following:

   - What fraction of shipments will be accepted?

   - If the policy changes so that a shipment is accepted if only one CD drive in the sample is defective, what fraction of shipments will be accepted?

   - What is the probability that a sample size of 50 will contain at least 10 defective CD drives?

2. Use the airline overbooking data to do the following:

   - Determine how the probability of overbooking varies as the number of tickets sold varies from 100 through 115. Hint: Use a one-way data table.

   - Show how the probability of overbooking varies as the number of tickets sold varies from 100 through 115 and the probability that a ticket holder shows up varies from 80 percent

through 95 percent. Hint: Use a two-way data table.

3. Suppose that during each year, a given mutual fund has a 50 percent chance of beating the Standard & Poor's 500 Stock Index (S&P Index). In a group of 100 mutual funds, what is the probability that at least 10 funds will beat the S&P Index during at least 8 out of 10 years?

4. Professional basketball player Steve Nash is a 90 percent foul shooter. Answer the following questions:

   - If he shoots 100 free throws, what is the probability that he will miss more than 15 shots?

   - How good a foul shooter would Steve Nash be if he had only a 5 percent chance of making fewer than 90 free throws out of 100 attempts? Hint: Use Goal Seek in the What-If Analysis options.

5. When tested for extra sensory perception (ESP), participants are asked to identify the shape of a card from a 25-card deck. The deck consists of five cards, of each of five shapes. If a person identifies 12 cards correctly, what would you conclude?

6. Suppose that in a group of 100 people, 20 have the flu and 80 do not. If you randomly select 30 people, what is the chance that at least 10 people have the flu?

7. A student is selling magazines for a school fundraiser. There is a 20 percent chance that a given house will buy a magazine. He needs to sell five magazines. Determine the probability that he will need to visit 5, 6, 7,…, 100 houses to sell five magazines.

8. Use the BINOM.DIST function to verify that the BINOM.INV function yields the correct answer to my Village Deli example.

9. In the seventeenth century, French mathematicians Pierre de Fermat and Blaise Pascal were inspired to formulate the modern probability theory after trying to solve the *problem of points*. Here is a simple example: Pascal and Fermat take turns tossing coins. Fermat wins a point if a head is tossed, and Pascal wins a point if a tail is tossed. If Pascal is ahead 8-7 and the first player with 10 points wins, what is the chance that Pascal will win?

10. Suppose that each time the Houston Astros play the New York Yankees, the Astros have a 35 percent chance of winning the game. If the Astros play the Yankees 10 times, what is the chance the Astros win at least half the games?

11. I make 55 percent of my free throws. If I shoot 400 free throws, what is the chance that I make between 200 and 300 free throws inclusive?

12. Suppose 10 percent of all staplers are defective. If 60 staplers are sold, what is the chance that at least 10 staplers are defective?

13. A factory has five assembly lines. Each assembly line is down (not working) p percent of the time. What must p equal so that there is a 95 percent chance that at least one assembly line is

working?