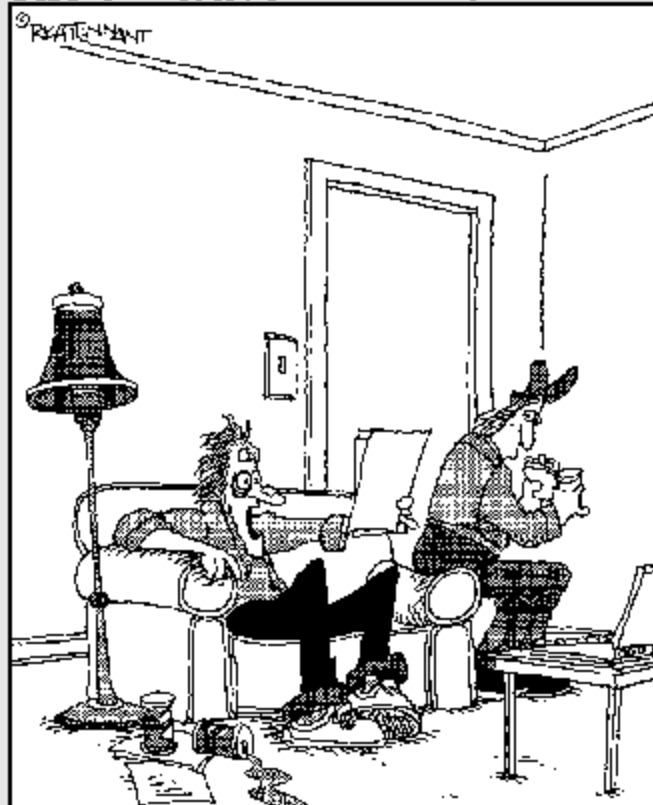


Part IV

Taking It Up a Notch: Advanced Probability Models

The 5th Wave

By Rich Tennant



"Theoretically, dating her twin sister should have increased my probability of marrying the right girl."

In this part . . .

part IV builds on your repertoire of probability models to include the Poisson, negative binomial, geometric, and hypergeometric distributions. These distributions allow you to model probabilities for situations beyond the binomial, such as the number of customers who enter a bank in one hour, the number of tries it takes to make a free throw or win the lottery (or both!), or the chance of being selected at random to participate on a committee. (We all love committees, right? If not, you can also find the probability of *not* being selected. . . .)

With the information you pocket here, you'll be able to perform tasks from modeling the number of times you need to roll a die before ten 6s come up to estimating the number of phone calls coming in to a psychic hotline from folks looking for the winning lottery numbers.

Chapter 13

Working with the Poisson (a Nonpoisonous) Distribution

In This Chapter

- ▶ Examining the Poisson probability model
- ▶ Figuring less-than, greater-than, and other probabilities for the Poisson
- ▶ Calculating the expected value and variance of the Poisson distribution and Poisson process
- ▶ Making an approximation for the Poisson with a normal distribution

The Poisson distribution helps you answer questions involving the number of arrivals, or events that occur, in a fixed period of time — such as the number of airplanes that come in to an airport in two hours or the number of accidents that occur at a certain intersection in a month. The Poisson distribution is often used to model birth and death rates, because these rates represent events that occur over a given period of time. And it can also be used to model occurrences of an event in a fixed amount of space, such as the number of defects in a 10'-x-10' piece of carpet.

In this chapter, you discover how to recognize the Poisson distribution probability model, find probabilities, and calculate the expected value and variance of the Poisson distribution. You also find out how to approximate the Poisson with a normal distribution (see Chapter 9) in cases where the numbers get out of hand (this is similar to approximating the binomial with the normal distribution, which I describe in Chapter 10).



Because Simeon Denis Poisson (a 19th-century ecologist) invented the Poisson model, you capitalize its name; you leave other probability models, such as the binomial, in lowercase form. And even though it looks like the word “poison,” you don’t pronounce Poisson that way; you pronounce it “Pwas-on,” with the accent on the “on.”

Counting On Arrivals with the Poisson Model

The *Poisson distribution* allows you to model and find probabilities for the number of events that occur in a specified period of time or within a specified area of space. With the Poisson, you can find the chance that more than two phone calls will come in to a helpline within 15 minutes; the chance of getting a chocolate-chip cookie with no chocolate chips in it (you wouldn't want that, would you?); or the chance that a typist makes more than two errors on a single page, just to name a few examples.

Meeting conditions for the Poisson model

As you build your repertoire of probability models, it becomes more and more important to know the conditions you need to check to identify which model is which. The first step in solving a Poisson probability problem is making sure Poisson is the proper model to use.



A random variable X has a Poisson distribution if the following conditions hold:

- ✓ X counts the number of events or occurrences within a specified time or space.
- ✓ The events occur independently of each other.
- ✓ No two events can happen at exactly the same time.

Because you count the number of events or occurrences in a fixed time or space, the Poisson random variable can take on any positive integer from zero to infinity (0, 1, 2, 3, and so on); therefore, a Poisson random variable is a discrete random variable with a countably infinite number of possible values (see Chapter 7).

“Now hold on,” you say. “It isn’t really possible to obtain an infinite number of typos on a page or an infinite number of phone calls in 15 minutes.” Good point. However, because it’s impossible to determine exactly where the cutoff is for such situations, you can go ahead and let X take on any integer from zero to infinity. You know that the probabilities for extremely high numbers of occurrences will be small, as determined by the Poisson probability mass function (pmf), which I discuss in the section “Determining Probabilities for the Poisson.”

Pitting Poisson versus binomial



In order to successfully identify the probability model you need to be working with, you need to recognize the difference between the Poisson and the binomial probability models (for information on the binomial probability model, see Chapter 8). Both the Poisson and binomial are discrete, which means you use them for counting outcomes that occur. However, the binomial model counts the number of “successes” (outcomes that have the desired characteristic of interest) in n fixed trials, so X can take on integer values from only zero to n = the number of trials. The Poisson, on the other hand, doesn’t have any “trials”; it merely observes a situation over a fixed period of time or space and counts how many occurrences happen in that time or space. And because the Poisson features no fixed cutoff, X can take on any integer from zero to infinity.

A problem can create scenarios for the binomial and Poisson that sound very similar, requiring you to be very clear about their differences. For example, suppose you’re sitting at an intersection watching traffic for your driver’s education class. Your job is to select 50 cars at random and keep a count of how many come to complete stops. Which model should you use, the binomial or Poisson? Because you have a fixed number of trials (50), and each trial is success or failure with equal probability (because you take a random sample), the binomial model holds (see Chapter 8). The important step to key in on is the fixed number of trials.

Now suppose your job is to sit at the intersection and count the number of times you see a car roll through the stop sign in a two-hour period. The job calls on you to fix the time period but not the number of trials (because you have no way of knowing how many cars will come through the intersection). Now you should use the Poisson model.

Determining Probabilities for the Poisson

When you know that X falls under a Poisson probability model (see the previous section to be sure), you can use previously established Poisson formulas to find Poisson probabilities for X . You can do this because all Poisson probability models have the same formula for calculating probabilities; the only differences are in the rates of the occurrences that you expect to see in the situations.

You have two possible ways to calculate probabilities for a Poisson: the probability mass function and the cumulative distribution function.

The pmf of the Poisson

The *pmf* [probability mass function, denoted $P(x)$; see Chapter 7] gives you the formula for calculating the probability that X equals a certain number.



The formula for the pmf for the Poisson distribution is $\frac{e^{-\lambda} \lambda^x}{x!}$, for $x = 0, 1, 2, 3, \dots, \infty$, where λ is the average (or mean) rate of occurrence of the events over the fixed time or space. (The problem has to give you the mean.) When people in the know run across data that fits this model, they say that X has a Poisson distribution with mean λ .



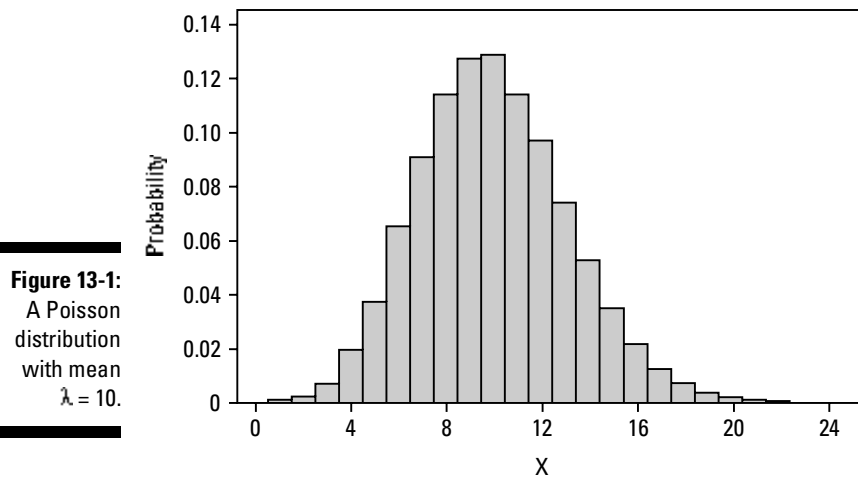
Recall that the letter “e” in math terms means the inverse natural logarithm (inverse of \ln) and that e^1 is approximately equal to 2.72. Most scientific or graphing calculators include a button for e^x , and you can use it to raise “e” to any power you want.

For example, suppose you’re working at a helpdesk and, from past experience, you know that you typically get ten inquiries per hour on average. Assuming you can’t have two inquiries at exactly the same time, the probability model for X — the actual number of inquiries per hour — would have a Poisson distribution with λ equal to ten per hour. The pmf of X is $P(x) = \frac{e^{-10} 10^x}{x!}$, for $x = 0, 1, 2, 3, \dots, \infty$. A graph of the pmf of the Poisson distribution with mean $\lambda = 10$ is shown in Figure 13-1.

Here’s how to interpret the graph of the pmf in Figure 13-1. The probabilities start out low at $X = 0$, because if you expect ten inquiries per hour, the chance of getting only a few inquiries in an hour is small. The probabilities increase as X increases. The graph “peaks out” around ten (the expected number of customers per hour); the values then get smaller and smaller as X gets larger and larger. After X passes 22, almost no probability is left; that’s because if you expect ten inquiries per hour, getting more than 22 in an hour should happen very rarely.

You can also use the pmf to calculate the probability of getting a certain number of inquiries in an hour. Suppose that you want the probability of getting 15 inquiries in the next hour: $P(X = 15)$, or $P(15)$. Putting 15 in for X in the pmf, you get $P(15) = \frac{e^{-10} 10^{15}}{15!}$, which equals 0.035. If you want the probability of getting 10 inquiries in an hour, you put 10 in for X to get $P(10) = \frac{e^{-10} 10^{10}}{10!} = 0.125$. As you may expect, the probability of a number that’s significantly far from 10 should be small, because if the mean is 10 inquiries, you shouldn’t expect a very large or very small number of inquiries to occur with high probability.

For example, the chance of getting 20 inquiries is $P(20) = \frac{e^{-10} 10^{20}}{20!}$, which equals 0.002, and the chance of getting no inquiries in an hour is $P(0) = \frac{e^{-10} 10^0}{0!}$, which equals 0.00005.



You may be wondering why the probability that $X = 10$ isn't higher than 0.125, because the average number of inquiries is expected to be 10. Although this value of X is the one that has the highest probability (see Figure 13-1), the result will vary from hour to hour. Because the total of all the probabilities has to be equal to one (see Chapter 7), the probability will be spread out among different values of X , with the most probability occurring at λ and less and less probability occurring as you move away from it on either side. The Poisson must start at zero, so it isn't necessarily symmetric; this may cause the probability distribution to have a skewed look because it doesn't have time to develop a tail in some instances on the left side.



You need to keep track of your units with Poisson probability models. In the previous example, the mean of ten applies to the expected number of inquiries per hour, and you have to reflect that "per hour" in your description of X . (In the section "Changing Units Over Time or Space: The Poisson Process" later in this chapter, you find out how to calculate probabilities for a Poisson when the units are changed.)

The cdf of the Poisson



The cdf [cumulative distribution function, denoted $F(x)$; see Chapter 7] gives you a formula for the accumulated probability from zero to X for any particular value of X . You use the cdf to find the probability that X is less than, greater than, or between two numbers.

The cdf shows all the probability accumulated up to any point x on the distribution. The cdf for the Poisson distribution has the formula

$\sum_{x \leq x} \frac{e^{-\lambda} \lambda^x}{x!}$ for $x = 0, 1, 2, 3, \dots, \infty$. Notice that the formula sums up the pmf (see the previous section) over all values up through the one you're interested in, x . For example, the cdf of a Poisson distribution with mean ten is shown in Figure 13-2. Notice it starts at zero for $X = 0$ and approaches one as X heads to infinity.

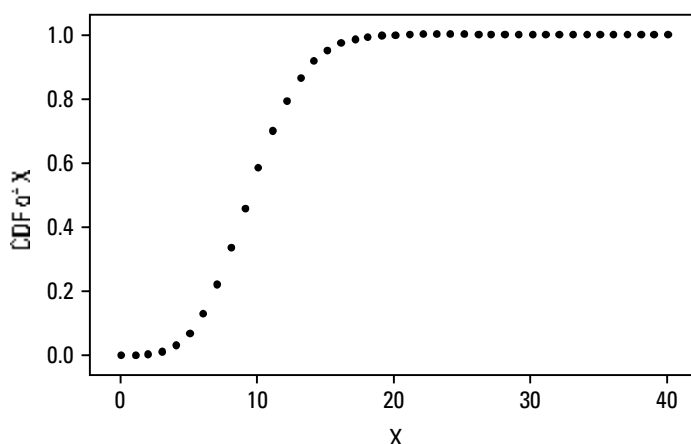


Figure 13-2:
The cdf of
a Poisson
distribution
with mean
 $\lambda = 10$.



Ready for some good news? You don't have to use the formula for the cdf to find values of $F(x)$. You can use a table that has the calculations ready to go for you. The trusty table is Table A-3 in the Appendix.

To use the Poisson table to find the value of $F(x)$ for a Poisson distribution, you need to know only two things:

- ✓ The value of λ .
- ✓ The value of X that you want the probability accumulated up to

Intersect the row for the value of X with the column for the value of λ to find the value of the cdf at that point.

Figuring less-than or equal-to (or strictly less-than) probabilities

Calculating a less-than or equal-to probability means adding up all the probabilities for the values of X that are less than or equal to the number you want. For example, if you want the probability that X is less than or equal to four, you want $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$. This is the same as finding the value of the cdf at four, written as $F(4)$. So you can avoid having to do all the tedious adding calculations, a table has already been created to find less-than or equal-to probabilities for a Poisson distribution with several different values of λ . In this section, you find out how to use this table (Table A-3 in the Appendix).

The Poisson table is split into two parts; the first part is for values of λ between zero and one, and the second part is for values of λ from one to twenty.

For example, a carpet manufacturer wants to do a quality-control study on its carpet, and the company knows from past experience that it can expect about 1 blemish per 100 square feet of carpet. In this example, λ equals 1 blemish per 100 square feet of carpet. Suppose that you want the probability of having less than or equal to 2 blemishes in 100 square feet of carpet: $P(X \leq 2)$, or $F(2)$, for a Poisson with λ equal to 1. Look at the top part of the Poisson table and find the column where λ equals 1 and the row where X equals 2. If you intersect that row and column, you find the number 0.920. That means the chance that a 100-square-foot piece of carpet has 2 or fewer blemishes is 0.920, or 92 percent.

You can also use the Poisson table to find specific less-than probabilities. However, a strictly less-than probability requires a little finesse before you can use the table because it doesn't include the "equal to" part of the probability. If you want to find the probability that X is less than four, for example, you want to find the probability of all values up to but not including four: $P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$. In terms of the cdf, the probability that X is less than four is the same as $F(3)$.

For example, to find the probability of having less than 2 blemishes in 100 square feet of carpet, you need $P(X < 2)$. Because the Poisson table gives probabilities for only less than or equal to, you have to rewrite $P(X < 2)$ as $P(X \leq 1)$. Now look at the top part of the table in the column for $\lambda = 1$ and the row for $X = 1$ to find the number 0.736. Now you have the probability of having fewer than 2 blemishes in a 100-square-foot piece of carpet.

Figuring greater-than probabilities

You can use the Poisson table (see Table A-3 in the Appendix) to find greater-than probabilities. In order to use the table to find your probability, however, you need to rewrite your greater-than probability as a less-than or equal-to probability, using complements. For example, the probability that X is greater



than six is the same as one minus the probability that X is less than or equal to five. The probability that X is less than or equal to five is something you can find on the Poisson table; just remember to take the “one minus” to get your final answer.

Sticking with the carpet example from the previous section, suppose that you want to find the probability of having more than 2 blemishes in a 100-square-foot piece of carpet: $P(X > 2)$. Using complements to utilize the Poisson table, you can rewrite the problem as $1 - P(X \leq 2)$, which is $1 - 0.920 = 0.08$. The probability of the carpet having at least 2 blemishes is $P(X \geq 2)$, which is $1 - P(X \leq 1)$. That gives you $1 - 0.736 = 0.264$, or 26.4 percent.

Figuring probabilities between two values

The probability of X being between two values is the probability for certain values of X only; it takes on various forms depending on whether you want to include the specific values on the borders of the inequality. For example, $P(3 < X < 6)$ means the probability of all values of X between 3 and 6, not including 3 or 6 — in other words, the probability that X equals 4 or 5. But $P(3 \leq X \leq 6)$ means the probability of all values of X between 3 and 6, including 3 and 6 — in other words, the probability that X equals 3, 4, 5, or 6. Of course, you can have combinations of these inequalities, like $P(3 \leq X < 6)$, which includes 3 but not 6, and $P(3 < X \leq 6)$, which includes 6 but not 3.



To find the probability of being between two values, you have to rewrite the probability statement so that it involves only less-than or equal-to signs (due to the fact that the Poisson table [see Table A-3 in the Appendix] shows only these probabilities), and then you find the values of the cdf for each number and subtract them (largest minus smallest).

Sticking with the carpet example from the previous two sections (where $\lambda = 1$), suppose that you want to find the probability of getting between 3 and 5 blemishes (including 3 and 5) in a 100-square-foot piece of carpet: $P(3 \leq X \leq 5)$. You find this probability by looking at the probability of less than or equal to 5 in the Poisson table, which is denoted $F(5)$, and subtracting the probability of less than or equal to 2, which is $F(2)$. This gives you the probability of all values of X between 3 and 5. The value of the cdf at $X = 5$ is 0.999. The value of the cdf at $X = 2$ is 0.920. Subtract these two values to get $F(5) - F(2) = 0.999 - 0.920 = 0.079$, or 7.9 percent. The probability of having 3, 4, or 5 blemishes in a 100-square-foot piece of carpet is 7.9 percent.

Identifying the Expected Value and Variance of the Poisson

The expected value and variance of the Poisson distribution show how many arrivals or occurrences you can expect in a fixed time or space, as well as the amount of variability in these results from one experiment to the next. What's interesting about the Poisson distribution is that you see a very close relationship between the expected value and the variance.

The *expected value* of any distribution is the overall average value of the distribution. The notation you use for the expected value is μ_x , or $E(X)$. In terms of the Poisson distribution, $\mu_x = E(X) = \lambda$.



The problem has to give you λ for the Poisson distribution. It may not identify it as the expected value, so you have to remember this fact. Or the problem may not tell you specifically what λ is, but it tells you the mean is “such and such.” And that “such and such” represents the value of λ .

The *variance* of any distribution is the overall average squared distance from the mean, denoted by σ_x^2 . The standard deviation is the square root of the variance, denoted by σ_x . For the Poisson distribution, the variance is $\sigma_x^2 = \lambda$. Because this changes the units to square units, which may not make sense (especially if you have to deal with something like phone calls per minute), you normally look at the standard deviation as your measure of variability in the results. The standard deviation of a Poisson distribution is $\sigma_x = \sqrt{\lambda}$.

For example, suppose that customers come into a bank with $\lambda = 20$ per hour, according to a Poisson distribution. In this case, the expected value is $\mu_x = \lambda = 20$ customers per hour, and the standard deviation is $\sigma_x = \sqrt{\lambda} = \sqrt{20} = 4.47$ customers per hour.



Always be aware of and include the units that come with the mean for a Poisson distribution. The mean is really a rate per unit, and that unit is important — especially when the units change. (See the next section for more on the topic of units.)

Changing Units Over Time or Space: The Poisson Process

The *Poisson process* is a probability model that takes a Poisson distribution and changes its units of time or space. It uses a function that allows you to easily change the units of a Poisson distribution and find probabilities, expected values, and variances instantly. Suppose, for example, that X is a random variable with a mean of λ per unit. A new random variable $Y = \alpha X$ is a Poisson process with mean $\alpha\lambda$, where α is the number that you multiply λ by to change from the old units to the new units. In other words, a change in units for X is reflected in the formula for the expected value (and variance).

For example, suppose that X counts the number of occurrences of an accident in one year, with mean $\lambda = 1$ per year. Following this, $2X$ counts the number of occurrences of an accident in two years, with mean $2\lambda = 2 * 1 = 2$ per year. So, α equals 2 in this case.



Sometimes you can easily identify what multiplier you need, and sometimes the task may be difficult, but you can always set up a proportion problem to figure out what the multiplier should be. (Remember those from algebra? And you thought you'd never use them . . .) For example, suppose that X is a Poisson with mean 16 occurrences per hour, and you want to look at 15-minute intervals only. What's the multiplier you need to move from rate per hour to rate per 15 minutes? You know that 60 minutes make up 1 hour, so you can write $\frac{16}{1 \text{ hour}} = \frac{16}{60 \text{ min}} = \frac{\alpha}{15 \text{ min}} \rightarrow 60\alpha = 240 \rightarrow \alpha = 4$. It also makes sense that if you expect 16 occurrences per hour on average, you can expect about 4 occurrences in 15 minutes (which is $\frac{1}{4}$ of an hour) on average. (Notice that I had to convert from hours to minutes in the middle part of the equation to avoid errors.)

You can apply the Poisson process to the carpet example from the section "The cdf of the Poisson." Assume that you have 200 feet of carpet. The number of blemishes, Y , in 200 square feet has a Poisson distribution with mean $2 * 1 = 2$. You can now call this λ and put it into a Poisson distribution. The expected value of Y is $\mu_y = 2$; the variance is equal to $\sigma_y^2 = \lambda = 2$ blemishes per 200 square feet (squared); and the standard deviation is equal to $\sigma_y = \sqrt{\lambda} = \sqrt{2}$, which equals 1.41 blemishes per 200 square feet. (See the previous section to find out how to calculate these figures.)



Notice that the standard deviation for Y , the number of blemishes in 200 square feet of carpet (1.41), isn't the same as two times the standard deviation of X , the number of blemishes in 100 square feet (one). The difference occurs because with 200 square feet of carpet, the blemishes have more space to occur in, so you see more variability in when and where they'll happen. Just because you have two blemishes every 200 square feet on average doesn't mean each one will end up in a separate 100-square-foot chunk, so, although it works for expected values, you can't just take the results from the original Poisson distribution and double them in every case to get answers for a Poisson process with a multiplier of two.

Now suppose you cut down the square footage to 50 square feet of carpet. If the original random variable X is the number of blemishes per 100 square feet, and Y is the number of blemishes per 50 square feet, using the Poisson process, you have $Y = \frac{1}{2}X$, so the multiplier, λ , equals $\frac{1}{2}$ or 0.50. You can say that Y has a Poisson distribution with mean equal to $\frac{1}{2}$ times the mean of the original Poisson distribution, X , so you have $\frac{1}{2} * 1 = \frac{1}{2}$ blemish per 50 square feet. The expected value of Y is a half of a blemish per 50 square feet; the variance is $\frac{1}{2}$ blemish per 50 square feet (squared); and the standard deviation is $\sqrt{\frac{1}{2}} = 0.71$ blemishes per 50 square feet. The probability of getting at most two blemishes per 50 square feet is $P(X \leq 2) = 0.986$ by looking at Table A-3 in the Appendix (intersect the row for $X = 2$ with the column for $\lambda = 0.50$).

Approximating a Poisson with a Normal

You often come across times when calculations for probabilities for a Poisson distribution get out of control, making your job of finding the values of the pmf and/or the cumulative distribution function (cdf) difficult. For example, suppose you have a Poisson distribution with a mean of 50, and you want the probability that $X = 45$: $\frac{e^{-50} 50^{45}}{45!}$. If you try to use your calculator to do any of the parts of this calculation, you discover that it won't budge! So, you can imagine that finding the cdf at 45, which would sum all the probabilities for X from 0 to 45, is a no-can-do situation as well.

Because the calculations for a Poisson distribution for large values of λ are hard to get, you often must use an approximation. And just like the approximation you do for the binomial distribution when it gets unwieldy (see Chapter 10), you can use the normal distribution to approximate the Poisson. All you have to do is figure out what to put in for the values of μ , the mean of the approximating normal distribution, and σ , the standard deviation. In this section, I discuss the normal approximation to the Poisson distribution.

Satisfying conditions for using the normal approximation

The Poisson distribution must meet certain conditions to merit the use of the normal distribution for approximation. If you have a Poisson distribution with mean $\lambda = 20$ or more, you can use a normal distribution to approximate the Poisson. The larger the value of λ , the better job the normal approximation does at approximating probabilities for the Poisson distribution. Using the normal approximation makes it much easier to find probabilities for the Poisson compared to using the pmf formula.



For λ beyond 20, the Poisson table (see Table A-3 in the Appendix) doesn't provide values for the cdf (see the section "Determining Probabilities for the Poisson" earlier in this chapter); experts had to cut it off somewhere, and most Poisson cdf tables cut it off at $\lambda = 20$.

To understand why a normal approximation works well for a Poisson in certain situations, it helps to look at the graph of the probability mass function of a Poisson (for more on the pmf, see the section "Determining Probabilities for the Poisson" earlier in this chapter).

Figure 13-3 shows a histogram of the Poisson distribution with a mean of $\lambda = 2$. Notice that the left side of the graph appears to be cut off, and a tail shows up on the right side. The graph sets up this way because the mean is two, which is very close to zero, and the left side of the graph doesn't have enough room to include much probability (because X can't be less than zero in a Poisson distribution). This makes the graph of the pmf skewed to the right (see Chapter 7 for more on shapes of distributions). The normal approximation doesn't work well here, because the value of λ is too small.

Figure 13-3:
A Poisson
distribution
where the
mean, $\lambda = 2$,
is too small
to use the
normal.

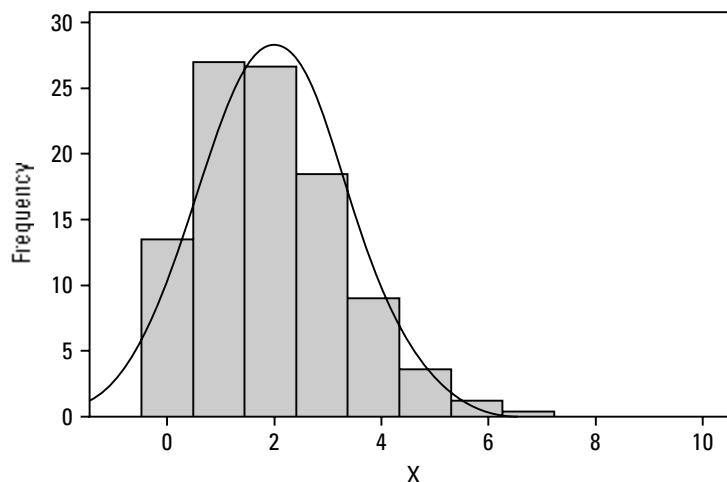


Figure 13-4 shows a Poisson distribution with a mean of 5. You can see that the graph is starting to move over away from zero, and a tail is beginning to develop on the left side. However, the graph is still skewed right because the mean is still quite close to zero. A normal approximation won't work well in this case because the value of λ is too small.

Figure 13-5 shows a Poisson distribution with a mean of 20. Now you can see a bell-shaped curve really forming because the left side is fully available; you can attribute this to the mean being far enough from zero for the entire left tail to appear. At this point, you can safely use the normal distribution to approximate the Poisson distribution.

Figure 13-4:
A Poisson distribution where the mean, $\lambda = 5$, is too small to use the normal.

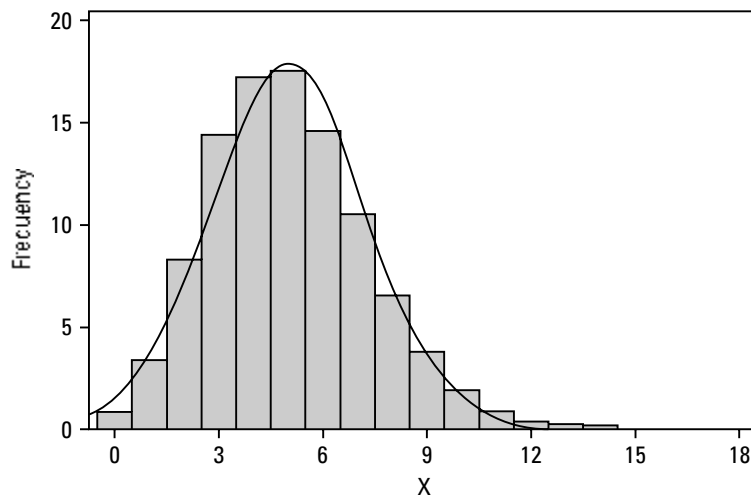
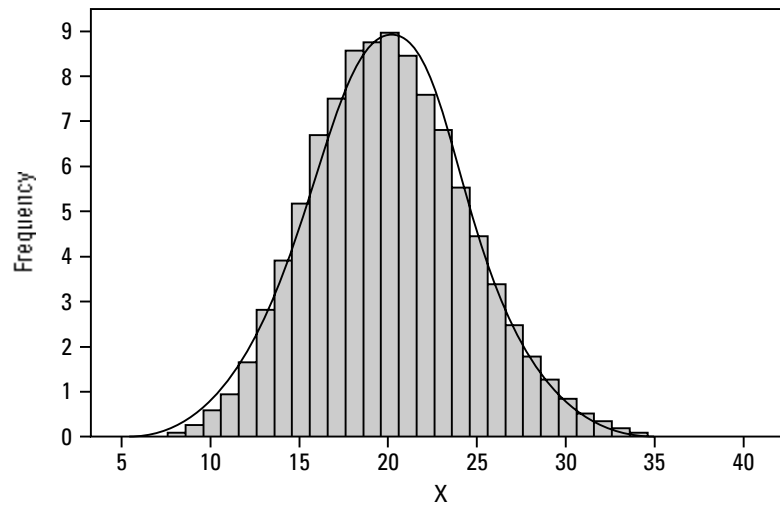


Figure 13-5:
A Poisson distribution where the mean, $\lambda = 20$, is large enough to use a normal approximation.



Completing steps to approximate the Poisson with a normal

If you know that λ , the mean of the Poisson distribution, is greater than 20, and you want to use the normal approximation to get probabilities for the Poisson, you use the steps from Chapter 10 for approximating a probability by using a normal distribution. You need to know what to put in for the mean and standard deviation of the normal distribution; just put in the mean and standard deviation of the Poisson, which are $\mu_x = \lambda$ and $\sigma_x = \sqrt{\lambda}$, respectively.

Here are the steps for approximating Poisson probabilities by using a normal distribution:

- 1. Make sure your value of λ is large enough (the larger the better, but at least 20).**
- 2. Translate the problem, using probability notation, into one of the following: $P(X < a)$, $P(X > b)$, or $(a < X < b)$.**
- 3. Transform a (or b) into a z value, using the Z-formula: $Z = \frac{X - \mu_x}{\sigma_x}$.**
In the case of the Poisson, use $\mu_x = \lambda$ for the mean and $\sigma_x = \sqrt{\lambda}$ for the standard deviation.
- 4. Look up the value on the Z table (Table A-2 in the Appendix), and find the probability of being less than or equal to X.**
- 5. If you have a less-than problem, you're done. If you have a greater-than problem, take one minus the result from Step 4. If you have a between-values problem, do Steps 2–4 for b (the larger of the two values) and then for a (the smaller of the two values) and subtract the results.**
- 6. Answer the original question in the context of the problem (in other words, in the language of X, not Z). The answer remains the same.**

For example, suppose customers enter a bank at an average rate of five customers per ten minutes. You want the probability that more than 35 customers will come in a given hour (note the change in units). This example is a Poisson process with multiplier six (because ten minutes times six equals one hour, the new unit of time; see the section “Changing Units Over Time or Space: The Poisson Process”). The number of customers in one hour — call it X — has a Poisson distribution with mean $\mu = 6 * 5 = 30$. (You also know that the standard deviation is $\sqrt{\lambda} = \sqrt{30}$, which is 5.48 customers per hour.) You want the probability that X is more than 35, so you want $P(X > 35)$.



The Poisson distribution at work

The Poisson distribution has plenty of real-world applications. For instance, you can use it to model people entering and exiting a queue (a line that forms where people wait to be served and then exit) at any time. Life insurance companies use it to model births and deaths so they can figure out what premiums to charge and when those premiums are likely to come due in the long run. Highway departments use it to assess roads and highways by counting and modeling the number of accidents that occur

and which intersections are more dangerous than others.

Sound too boring? Well, the Poisson distribution also helps make sure you have almost no chance of biting into a chocolate-chip cookie without eating any chocolate chips. The cookie manufacturers make sure the mean of the Poisson distribution that counts the number of chips per cookie is so large that the chance of the actual number of chips being zero is almost zero. What a relief!

Following the steps to find probabilities with a normal distribution, you get the following:

1. $\lambda = 30$, is large enough to use the normal approximation.
2. You want $P(X > 35)$.
3. Using the Z-formula with mean $\mu_x = \lambda = 30$ and standard deviation $\sigma_x = \sqrt{\lambda} = \sqrt{30} = 5.48$ (the mean and standard deviation of the Poisson, respectively), you have $Z = \frac{X - \mu_x}{\sigma_x} = \frac{35 - 30}{\sqrt{30}} = \frac{5}{5.48} = 0.91$.
So, $P(X > 35)$ is approximately equal to $P(Z > 0.91)$.
4. Looking up 0.91 on the Z table (Table A-2 in the Appendix), you get 0.8186.
Because the Z table gives you the probability below the Z value you look up, you know that $P(Z < 0.91)$ equals 0.8186.
5. You have a greater-than problem, so to find $P(Z > 0.91)$, you take the complement: $1 - P(Z < 0.91) = 1 - 0.8186 = 0.1814$.
6. In the original context of the problem, you state that the probability of having more than 35 customers enter the bank during a one-hour period is approximately 0.1814, or 18.14 percent.



Don't forget that your final answer is an approximation when you use the normal distribution to approximate the Poisson. So, in your final answer, make sure you state it that way. You can also say, "I feel confident that this approximation is close, because λ is greater than 20."



When the values of λ are beyond 20, the normal distribution gives pretty close answers to the actual Poisson probabilities, but you still see a little bit of difference until the values of λ get past 40 or so. To address this issue, you can use a continuity correction so that your answers with the normal distribution are closer to the exact answers you'd get if you use the Poisson distribution. All you have to do is subtract 0.50 from your value of X if you're doing a greater-than or equal-to probability and add 0.50 to your value of X if you're doing a less-than or equal-to probability before you use the Z -formula. (To get the full details on how to incorporate the continuity correction, see Chapter 10.) In the example I introduce in this section, you want $P(X > 35) = P(X \geq 36)$. Using the continuity correction, you really find $P(X \geq 35.5)$ with the Z -formula, which comes out to 0.1587 rather than 0.1814.

Chapter 14

Covering All the Angles of the Geometric Distribution

In This Chapter

- Sizing up the geometric probability model
- Utilizing the pmf to find probabilities for a geometric distribution
- Using simple formulas to measure the expected value and variance

The *geometric distribution* is named for the way its *probability mass function* (pmf) looks. If you've taken advanced algebra or calculus, you might remember the geometric series, which is a sum of the values of a fraction that's between 0 and 1, taken to higher and higher powers (starting with the fraction to the zero power, which is one). For example, one geometric series sums the values of $\frac{1}{2}$ taken to higher and higher powers: $(\frac{1}{2})^0 + (\frac{1}{2})^1 + (\frac{1}{2})^2 + (\frac{1}{2})^3$, and so on. You might remember that mathematicians call the fraction r , and the sum of the geometric series for r is $1 \div (1 - r)$. In our example, the sum would be $1 \div (1 - \frac{1}{2})$, which comes out to $1 \div \frac{1}{2} = 1 * 2$, which is 2.

Now, if you skip the first term in this series, which is r to the zero power, or one, the sum of the geometric series in the previous example is $2 - 1 = 1$, and one is a great number because it represents the total of all probabilities for a random variable X . If you let the probabilities for X be $\frac{1}{2}$, $\frac{1}{2}$ squared, $\frac{1}{2}$ cubed, and so on, you then have a geometric series which sums to one, and that can be a probability distribution for some random variable X . This particular distribution is called the *geometric distribution*.

And the coolest thing about this is that taking a fraction to higher and higher powers happens all the time in probability; namely when you're repeating an experiment over and over again, more and more times, until the desired outcome happens. And that's what you work with in this chapter. (By the way, if the math explanation I used here made your eyes glaze over, don't worry. Probability, in my opinion, is a lot easier to understand than math anyway, so you'll do just fine when you get to the actual problems.)

In this chapter, you figure out how to recognize the geometric probability model, find probabilities under the geometric distribution, and calculate the expected value and variance of the geometric distribution.

Shaping Up the Geometric Distribution

The geometric distribution allows you to model and find probabilities for the number of trials needed until the first success occurs. For example, you can use the geometric distribution to find the chance that you'll need to buy more than ten instant-win lottery tickets to win a prize; the chance of ten days passing before the first accident occurs at a certain intersection; or the chance that a typist can go more than ten pages before making an error.

Here you find out what conditions a problem must meet to warrant the use of a geometric distribution, and you look at the differences and similarities between the geometric distribution and the binomial distribution (see Chapter 8).

Meeting the conditions for a geometric distribution

A probability problem presented to you must meet the following conditions in order for a random variable, X , within the problem to have a geometric distribution:

- ✓ The problem presents a sequence of independent trials of some random process.
- ✓ You can classify the outcomes of each trial into two groups: success or failure.
- ✓ The probability of success must be the same for each trial; let p be the probability of success (which means one minus p is the probability of failure).
- ✓ X counts the total number of trials up to and including the first success (which means the number of failures prior to the first success is X minus one).

Choosing the geometric distribution over the binomial and Poisson

In order to properly apply the geometric distribution to a given experiment, you need to recognize the difference between the geometric, binomial, and Poisson probability models (for information on the binomial and Poisson probability models, see Chapters 8 and 13). All three of these models are discrete, which means they take on either a finite or countably infinite number of possible values, and they each note success or failure on each trial. The binomial model counts the number of successes in n number of fixed trials, so X can take on integer values from only zero to n = the number of trials. The geometric distribution, on the other hand, doesn't require a fixed set of trials. It keeps observing trials until the first success occurs, and then it stops. So, the random variable, X , is the number of trials required to get that first success. And because you don't have to deal with a fixed cutoff, X can take on any integer value from zero to infinity (much like the Poisson model; see Chapter 13). Finally, the Poisson model fixes only the time or space. It has no number of trials, and it doesn't fix the number of successes. It merely counts the occurrences within that fixed time/space.

For example, if you flip a coin four times, you can have many combinations of heads and tails (indicated by H and T) — 16, in fact: HHHH, HHHT, HHTH, HTHH, THHH, HHTT, HTHT, HTTH, THHT, THTH, TTHH, HTTT, THTT, TTHT, TTTH, and TTTT. With the binomial model, any of these combinations are possible, and you would be counting the number of heads on the four tosses. The random variable, X , would go from zero through four.

With the geometric distribution, on the other hand, you don't have a fixed number of flips; you have a fixed number of successes needed (one success), and you flip until you get the first success. The random variable counts the total number of flips needed to get there. So, X can go from one to infinity. If you flip until you get a head, the outcomes for the geometric distribution look like this: H (you get the head right away — $X = 1$); TH (you get the first head on the second toss — $X = 2$); TTH (you get the first head on the third toss — $X = 3$); TTTH ($X = 4$); and so on.

With a Poisson, you wouldn't even be counting outcomes of flips. If you have a room full of students flipping coins, you may count something like the number of times a coin lands on the floor in a 10 minute period.



Some problems can create scenarios for the binomial and geometric and even Poisson distributions that sound very similar, requiring you to be very clear about their differences. For example, suppose you're sitting at a corner on campus watching students go by. Suppose your job is to select 50 students at random and count how many of them are wearing a backpack. Which model are you using, binomial or geometric? Because you have a fixed number of trials ($n = 50$) and each trial is success or failure with equal probability (because you took a random sample), the binomial model holds (see Chapter 8). The key to recognizing the proper model is the fixed number of trials.

Now suppose your job is to sit at this corner and count the number of students that go past until you see the first one with a backpack. Here, you're fixing the number of successes (one backpack wearer) but not the number of trials needed to get there (because you have no way of knowing how many students you'll have to count). This problem presents a geometric distribution.

Finally, suppose your job is to sit at this corner for two hours and count the number of students that go by wearing backpacks. In this situation, the time is fixed, not the number of trials or the number of successes. So, you know this problem needs a Poisson distribution.

The bottom line? Don't confuse the binomial, geometric, and Poisson distributions because you will be putting the wrong numbers in the wrong places in the formulas, and this will result in wrong answers. If the number of trials is fixed ahead of time and you want the number of successes, you have a binomial. If the number of successes is fixed and you want the number of trials it takes to get there, you have a geometric. If the time period is fixed but not the number of successes or trials, you use Poisson.

Finding Probabilities for the Geometric by Using the pmf

Finding probabilities for the geometric distribution amounts to multiplying the probabilities of success and failure. Because X is the number of trials up to and including the first success, you have only one success, and it has probability p . It also means you have $x - 1$ failures leading up to that success, each with probability $1 - p$. And, the best part, you don't have to worry about

the number of ways to rearrange those successes and failures because they all have to happen in a certain order: all the failures first, followed by that first success. So, finding probabilities for the geometric distribution is intuitive as well as fairly straightforward. (Don't ask questions, just go with it!)

Building the pmf for the geometric

The probability mass function of any discrete random variable X lists the possible values of X , along with their probabilities (Chapter 7). The formula for the pmf of the geometric distribution is $P(x) = p(1-p)^{x-1}$, for $x = 1, 2, 3, \dots, \infty$, where the number of trials up to and including the first success is x , the number of failures before the first success is $x - 1$, the probability of failure is $1 - p$, and the probability of success is p . Because each trial is independent, you use the multiplication rule to take $(1 - p) * (1 - p) * (1 - p) * \dots * (1 - p)$ for the $x - 1$ failures leading up to the success and multiply by p for that first success. And then you stop. The formula just rewrites this by using powers of $(1 - p)$ multiplied in the end by p .

For example, if you want the probability of rolling a die three times before getting your first 2, use the geometric pmf with $x = 3$ and $p = \frac{1}{6}$. The probability is $\frac{1}{6}$ squared, times $\frac{1}{6}$, which is 0.116. If you increase that number of rolls to 5, the probability is $\frac{1}{6}$ to the fourth power, times $\frac{1}{6}$, which is 0.0804. The probability is reduced because it shouldn't take that long to get your first 2.



The possible values of X start at one in the pmf because you have to have at least one trial to get the first success.



The *cumulative distribution function* (cdf) gives you the probability that X is less than or equal to any number x . Probability courses typically don't use the cumulative distribution function for the geometric distribution; nor do you have a table (like Table A-3 for the Poisson; see Chapter 13) for either the pmf (probability mass function) or the cdf of the geometric. Therefore, I don't include either in this book. The reason is that with geometric problems, you're typically interested only in the probability that it takes a certain number of trials to come up with the first success. And the formula is fairly straightforward to calculate. Also, if you need to find the value of the cdf — the accumulated probability up to a certain point — you'll find that the probabilities drop off to tiny values after a certain point, so you don't need to calculate many of them most of the time.

Applying geometric probabilities

The pmf of the geometric distribution looks similar to the pmf of the binomial distribution (see Chapter 8). However, the binomial pmf counts the number of ways to get x successes in n trials times the probability of an outcome with x successes in n trials. And the number-of-ways part involves counting rules, such as combinations and permutations (see Chapter 5). But because the geometric distribution counts the number of trials that pass until the first success occurs, you know that a sequence of failures must take place up to that first success, and you have only one way to show that: Put all your failures first and your success at the very end. The only question is how many failures will be involved.

For example, suppose a basketball player's free-throw average is 0.70. What's the chance that the player's first basket hits the net on the third shot? In other words, you want $P(X = 3)$. The problem meets the conditions of the geometric distribution because you have independent trials, each being success or failure, and you count the trials until the first success (for more on the conditions, see the section "Meeting the conditions for a geometric distribution"). You can also reasonably assume that the probability of success, p , will stay the same (0.70) throughout the experiment.

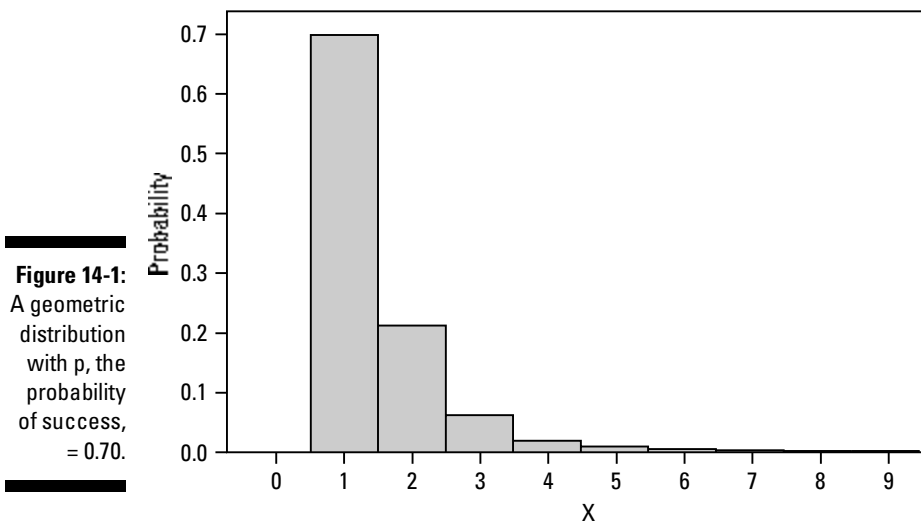
Using the pmf of the geometric, you find that $P(X = 3)$ equals $(1 - p)^2 p$, because the player has two failures first and then a success. Because $p = 0.70$, you have $P(X = 3) = (1 - 0.70)^2 (0.70) = 0.063$. Would the probability of getting the first basket in two tries be higher or lower than the probability of getting the first basket in three tries? Find out by finding $P(X = 2) = (1 - 0.70)^1 (0.70) = 0.21$. The probability for two tries is much higher because the probability of getting a basket on any trial is high, 0.70, so it shouldn't take too long to get that first basket!



If you take the previous example further, you can see how a pattern develops in the pmf of the geometric. Table 14-1 shows you the first few values in the pmf. Notice how each is equal to 0.30 times the value before it. That happens because the only difference between the values is another power put on the $(1 - 0.70)$ factor. The probabilities technically never reach zero, but after you reach the highest possible probability, the probabilities get closer and closer to zero as X gets larger and larger. Table 14-1 shows the probabilities to four decimal places only (the probabilities for all values past $X = 9$ are smaller than 1 in 10,000).

Table 14-1 Geometric pmf for $p = 0.70$	
X	$P(x)$
1	$P(X=1) = (1-p)^{1-1}p = (1-0.70)^0(0.70) = 0.7000$
2	$(1-p)^{2-1}p = (1-0.70)^1(0.70) = 0.2100$
3	$(1-p)^{3-1}p = (1-0.70)^2(0.70) = 0.0630$
4	$(1-p)^{4-1}p = (1-0.70)^3(0.70) = 0.0189$
5	$(1-p)^{5-1}p = (1-0.70)^4(0.70) = 0.0057$
6	$(1-p)^{6-1}p = (1-0.70)^5(0.70) = 0.0017$
7	$(1-p)^{7-1}p = (1-0.70)^6(0.70) = 0.0005$
8	$(1-p)^{8-1}p = (1-0.70)^7(0.70) = 0.0002$
9	$(1-p)^{9-1}p = (1-0.70)^8(0.70) = 0.0000$

The graph of the geometric distribution for $p = 0.70$ is shown in Figure 14-1. Notice that the graph is skewed to the right. Right skewness is a characteristic of geometric distributions. Because their values can technically go on to infinity, and each probability is equal to $(1-p)$ times the last one, the probabilities get smaller and smaller after they peak out; so, they tend to have long right tails.



Uncovering the Expected Value and Variance of the Geometric

The expected value of a random variable is the overall average outcome over the long term, and the variance is the average amount of variability in your results over repeated experiments (see Chapter 7). Although you can calculate the expected value $E(X)$ and the variance $V(X)$ of a geometric distribution from scratch by using the formulas I present in Chapter 7, the math goes a little beyond the scope of this book. It involves taking the sums of a modified version of the geometric series. Lucky for you, people have already worked out the probabilities and found that the results come out very nicely, and these nice people want to give you the benefit of using these results without having to do all the extra calculation work. In this section, I present the easy-to-use formulas for the expected value and variance of the geometric distribution.

The expected value of the geometric

The expected value of a geometric distribution has a surprisingly great connection to your intuition. Suppose you roll a die until a 1 comes up. The chance of getting a 1 on any individual roll is $\frac{1}{6}$. So how long do you expect, on average, to have to roll the die before the 1 comes up? You guessed it — 6 rolls. This just happens to be one over the probability of getting a 6 (not a coincidence). The formulas for the variance and standard deviation are also related to the probability of success.

The formula for the expected value of the geometric states that if p is the probability of achieving success in a geometric distribution, the expected number of trials until you get a success is $E(X) = \frac{1}{p}$.

Suppose you have a 10 percent chance of winning with an instant-lottery ticket. How many tickets should you expect to have to buy before you win? You may guess ten tickets, and you would be right. Now, this doesn't guarantee that if you buy ten tickets, you'll win; the probability just presents an overall expected average number of trials until that first success. In this example, $p = 0.10$ is the probability of achieving success and the expected number of trials until you get a success is $E(X) = \frac{1}{p} = 1 \div 0.10 = 10$.



Figuring out “how long” with the geometric distribution

People use the geometric distribution whenever they want to model the number of trials that occur until the first success happens. For example, players in a casino can use the distribution to determine the number of times they should expect to have to play slot machines until they win big. Or, from the other point of view, the game manufacturers can use the distribution to

determine how long players will keep playing casino games without winning that first time; with this info, they know how to set the probabilities for the machines. Outside the casino, manufacturers use the geometric distribution to determine how long they can expect production lines to last without breaking down and the amount of variability they can expect.

Now imagine that you want to win the mega lottery, which has a 100-million-dollar jackpot with a probability of 1 in 143 million. How many tickets should you expect to have to buy before you win? That’s right — 143 million tickets! Of course, people win with far fewer trials than that because variability exists in the outcomes of any geometric distribution, but this ‘winning right away’ situation applies to only a tiny number of people compared to the total number of people who play and lose over and over again.

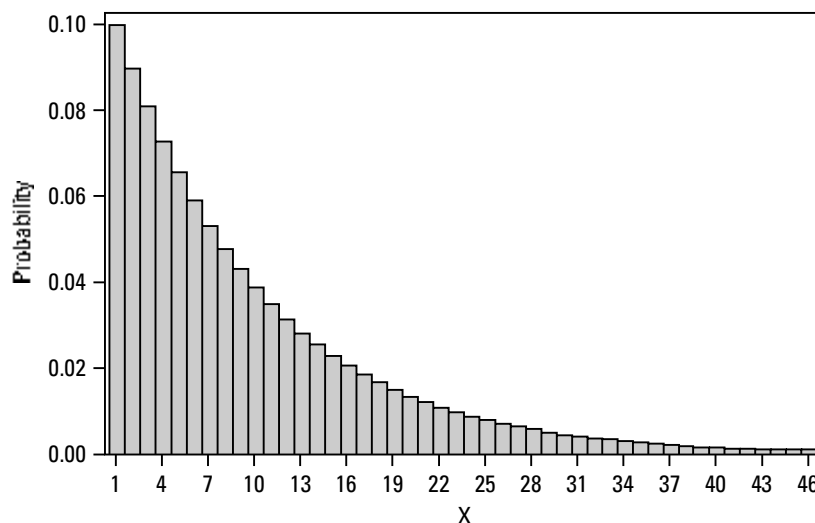
The variance and standard deviation of the geometric

You measure the variance in the results of a geometric distribution with the formula $V(X) = \frac{1-p}{p^2}$. If the probability of failure ($1 - p$) is small (this appears in the numerator), the probability of success (p) is large (this number squared goes in the denominator), so the overall variance (the numerator divided by the denominator) is small. If the probability of failure is large, the probability of success is small, so the overall variance is large.

The standard deviation of the geometric distribution is the square root of the variance, which gives you $\sqrt{\frac{1-p}{p^2}}$.

For example, suppose your chance of winning with an instant-lottery ticket is $p = 0.10$, or 10 percent. The mean number of tickets you need to purchase to win the first time is $1 \div 0.10 = 10$. What's the standard deviation? Because $p = 0.10$, and you have a geometric distribution, you use the formula for the standard deviation for a geometric to get $\sqrt{\frac{1-p}{p^2}} = \sqrt{\frac{1-0.10}{0.10^2}} = \sqrt{\frac{0.9}{0.01}}$, which equals 9.49. The average deviation in the results is 9.49 tickets. You calculate a rather large value for a standard deviation, but it does match well with the real world — with gambling, you just never know what's going to happen. Figure 14-2 shows a picture of the geometric distribution with $p = 0.10$. It has a large amount of spread compared to Figure 14-1, where $p = 0.70$. Spread is measured by standard deviation; when p is smaller, you have more spread in the values than when p is larger.

Figure 14-2:
A geometric
distribution
with p , the
probability
of success,
 $= 0.10$.



Chapter 15

Making a Positive out of the Negative Binomial Distribution

In This Chapter

- ▶ Grasping and distinguishing the negative binomial probability model
- ▶ Figuring probabilities for a negative binomial distribution
- ▶ Determining the expected value and variance for a negative binomial

The negative binomial probability model is a more general version of the geometric probability model (see Chapter 14). The geometric model counts the total number of trials that take place until the first success occurs, such as the number of times you need to flip a coin until a head comes up. The negative binomial model, on the other hand, counts the total number of trials until the k th success occurs (where k can be any integer starting at one and going to infinity) — for example, the number of times you have to flip a coin until three heads come up or the number of people you have to survey before you find three Democrats who support a certain issue. Many of the same ideas from the geometric distribution apply in this chapter, but those concepts are expanded to allow you to build on more than one success.

In this chapter, you discover how to recognize the negative binomial probability model, find probabilities under this model, and calculate the expected value and variance of a negative binomial distribution.

Recognizing the Negative Binomial Model

The *negative binomial distribution* allows you to model and find probabilities for the number of trials needed to run an experiment until the k th success occurs. For example, the distribution could model the number of parts you have to examine before you find five defective ones (the longer the better,

right?); the number of names you have to call in a “must-be-present-to-win” drawing before you can give away your three prizes; or the number of people you have to randomly sample before you find five people who’ve read the whole series of *Harry Potter* books (that shouldn’t take long!).



Textbooks may use different definitions of what the random variable, X , counts, because probability offers two common ways to introduce X . If the books don’t use the same definition I do, they let X count the total number of failures before the k th success. In that case, the formulas look a little different, but the basic ideas are the same.

As with any other probability model, a situation must meet certain conditions in order to merit a negative binomial distribution. In this section, you run through these conditions and look at the differences and similarities between the negative binomial model, the geometric model (see Chapter 14), and the binomial model (see Chapter 8).

Checking off the conditions for a negative binomial model

Any problem or event you encounter must meet the following conditions in order for a random variable, X , to have a negative binomial distribution:

- ✓ You observe a sequence of independent trials from some random process.
- ✓ You can classify the outcomes of each trial into two groups: success or failure.
- ✓ The probability of success is the same for each trial; let p be the probability of success (which means $1 - p$ is the probability of failure). Some textbooks use the letter q to denote the probability of failure rather than $1 - p$.
- ✓ X counts the total number of trials up to and including the k th success. (The process features k successes in x number of trials, so it has to also feature $x - k$ total failures.)

Comparing and contrasting the negative binomial, geometric, and binomial models

The negative binomial, geometric (see Chapter 14), and binomial (see Chapter 8) models have similarities and differences that you must note in order to tell which type of problem you’re working with. If you’re taking a probability course, your instructor is likely to give you problems on exams without telling you which distribution the problems call for — you have to

determine that information yourself. In this section, I outline some tips for determining which distribution you have in probability situations.

How the models are the same

The binomial, geometric, and negative binomial models have some important similarities to note. Here's a list to keep in mind:

- ✓ All three distributions are discrete, which means they have a finite or countably infinite number of possible values (see Chapter 7).
- ✓ All record the outcome of each trial as either a success or failure.
- ✓ All contain p , the probability of success on a given trial, and $1 - p$ as the probability of failure.
- ✓ All the trials are independent for each of the three distributions.
- ✓ The negative binomial and geometric distributions are directly related.

The geometric is equal to a negative binomial random variable, X , with $k = 1$, where k is the number of successes, because you stop when you reach the first success in a geometric. The negative binomial extends beyond the geometric, allowing you to have any number of successes. (See Chapter 14 for full details on the geometric distribution.)

How the models differ

The binomial, geometric, and negative binomial models have some major differences. When you know which items to key in on, you can quickly tell the difference between them. Following is a list of the differences:

- ✓ **The binomial model counts the number of successes in n fixed trials, so the random variable, X , can take on integer values from only zero to n = the number of trials.** The negative binomial and geometric models, on the other hand, don't have a fixed set of trials. What's fixed in the negative binomial and geometric distributions is the number of successes you need; because you don't deal with fixed cutoffs, X can take on an infinite number of possible values. (See Chapter 8 for full details on the binomial distribution.)
- ✓ **The negative binomial model gets its name because it does the opposite of what the binomial does.** The negative binomial fixes the number of successes and counts the number of trials needed to get those successes; a binomial, on the other hand, fixes the number of trials and counts the number of successes within those trials.
- ✓ **The geometric model counts the number of trials until the first success occurs.** The geometric is a special case of the negative binomial distribution, where $k = 1$. And unlike the binomial distribution, you don't work with fixed numbers of trials for the geometric — it counts the number of trials needed to get a fixed number of successes. (See Chapter 14 for more on the geometric model.)

Formulating Probabilities for the Negative Binomial

Probabilities for the negative binomial distribution are based on the fact that you continue with an experiment until you achieve k successes, and then you stop. So, the last trial is the k th success — it can't be a failure. The trials before that point have to contain the rest of the successes ($k - 1$) and a certain number of failures. The random part of this process is where those $k - 1$ successes come in and how many failures happen along the way. In this section, I develop the probability mass function (pmf) for finding probabilities for the negative binomial distribution. You may notice that it looks much like the pmf for the binomial distribution (see Chapter 8), with the added restriction that the last trial has to result in the k th success.

Developing the negative binomial probability formula

The formula for finding probability for a negative binomial distribution is given by the following: $P(X = x) = \binom{x-1}{k-1} p^k (1-p)^{x-k}$, for $x = k, k+1, \dots, \infty$,

where p is the probability of success, $1 - p$ is the probability of failure, k is the number of successes you want to get, and X is the number of trials it takes to achieve the k successes (it counts the total number of trials up to and including the k th success). The possible values of X start at k because you need at least k trials to get k successes, and you can go through infinity because you can't tell how many trials you'll need to get k successes, especially if the probability of success, p , is very small.



When talking about the random variable itself, I use X (capitalized). When talking about a particular value of the random variable X , I use x (lowercase).

The probability formula for any discrete random variable (one that takes on a finite or countably infinite number of values; see Chapter 7) is also known as the probability mass function for X (or pmf). A pmf is a formula that gives you the possible values of X and their probabilities.

For example, suppose you want to roll a fair die until you get a total of ten 1s (and then stop). In this case, X is a negative binomial distribution with $p = \frac{1}{6}$ (because the die is fair), $1 - p = \frac{5}{6}$, and $k = 10$ successes (1s) needed. X is the total number of trials needed to get ten 1s. The pmf for X in this case is

$$P(X = x) = \binom{x-1}{9} \left(\frac{1}{6}\right)^{10} \left(1 - \frac{1}{6}\right)^{x-10}, \text{ for } x = 10, 11, 12, \dots, \infty. \text{ Notice that it has}$$

two parts — a combination part (based on k and x) and a probability part (based on p and $1 - p$). The probability part focuses on the probability of a single outcome, where all the failures occur first and then all the successes. Then you have to multiply that by the combination part, which takes care of the number of ways to rearrange all the successes except the last one (which has to occur in the last trial).

Suppose you want to find the probability of rolling the die 12 times before you get the 10 successes (1s). In other words, you want $P(X = 12)$. Do you think the probability is high or low? It should be low, because the chance of needing only 12 tries to get ten 1s is very small. In the pmf, you have $k = 10$, $p = \frac{1}{6}$ (probability of getting a 1), and $1 - p = \frac{5}{6}$. The value of X in this case is 12 (the total number of rolls).

Putting these values into the formula for the pmf, you get

$$P(X = 12) = \binom{12-1}{10-1} \left(\frac{1}{6}\right)^{10} \left(1 - \frac{1}{6}\right)^{12-10} = \binom{11}{9} \left(\frac{1}{6}\right)^{10} \left(\frac{5}{6}\right)^2 =$$

$(55)(0.000000017)(0.6944)$, which gives you 0.000000649. The probability is very small.

The chance of having to roll the die 20 times before getting ten 1s should be larger than the chance when you roll the die 12 times. In this situation, you have $X = 20$, and everything else in the problem remains the same. Putting the values into the pmf, you get

$$P(X = 20) = \binom{20-1}{10-1} \left(\frac{1}{6}\right)^{10} \left(1 - \frac{1}{6}\right)^{20-10} = \binom{19}{9} \left(\frac{1}{6}\right)^{10} \left(\frac{5}{6}\right)^{10} =$$

$(92,378) * (0.000000017) * (0.1615)$, which equals 0.000247. This probability is small, but it's larger than the one for $X = 12$.



Notice also that the pmf for the negative binomial is just an expanded version of the pmf for the geometric (Chapter 14), allowing for k successes to occur before stopping, rather than just one success like the geometric does.

Applying the negative binomial pmf

Applying the negative binomial distribution to a problem means first identifying the problem as a negative binomial by noticing that you need to find the number of trials needed to get a fixed number of successes, where you stop when you get the desired number of successes. The number of trials is the random part, not the number of successes. Next, you need to identify k , the desired number of successes, p , the probability of success, and x , the number of trials you want to find the probability for. Then you put the numbers into the negative binomial probability mass function (pmf), and you're on your way. In this section, you practice doing just that.

Suppose, for example, that a basketball player's free-throw average is 0.40, or 40 percent. The player wants to shoot baskets until he makes four free throws total (not necessarily four in a row; if he isn't good, he may be there all day!). What's the chance that he'll need ten tries to accomplish his goal? Remember, X counts the total number of trials until he makes the fourth free throw, so you're looking for the probability that X equals 10 [$P(X = 10)$]. And notice that if the player takes ten tries to make four free throws, he must have $10 - 4 = 6$ failures.

Stepping through the probability

To find the probability for X when X has a negative binomial distribution, follow these steps:

1. **Check to make sure that the conditions of the negative binomial model are met (see the section "Checking off the conditions for a negative binomial model" earlier in this chapter).**

For the free-throw example I present in the introduction to this section, the problem meets the conditions of the negative binomial model because you have independent trials, each being a success or failure, and you're counting the trials until the k th success. You can also assume that the probability of success, p , will stay the same (0.40) throughout the experiment.

2. **Identify x , k , p , and $1 - p$.**

Using the pmf of the negative binomial for the free-throw example, you have $x = 10$ (trials), $k = 4$ (successes), and the number of failures is $x - k = 10 - 4 = 6$. You also have $p = 0.40$ (the probability of success) and $1 - p = 0.60$ (the probability of failure). This gives you

$$\binom{x-1}{k-1} p^k (1-p)^{x-k} = \binom{10-1}{4-1} 0.40^4 0.60^{10-4}. \text{ The problem simplifies to } \binom{9}{3} 0.40^4 0.60^6 = \frac{9!}{3!6!} 0.40^4 0.60^6, \text{ which equals 0.100.}$$

So, even though the player has a 40-percent chance of making the free throws, and he should expect to make four shots in ten attempts, the chance of him making them so that his fourth success comes on the tenth try is only 0.100, or 10 percent. The chance is low because the player has many ways to experience successes and failures.



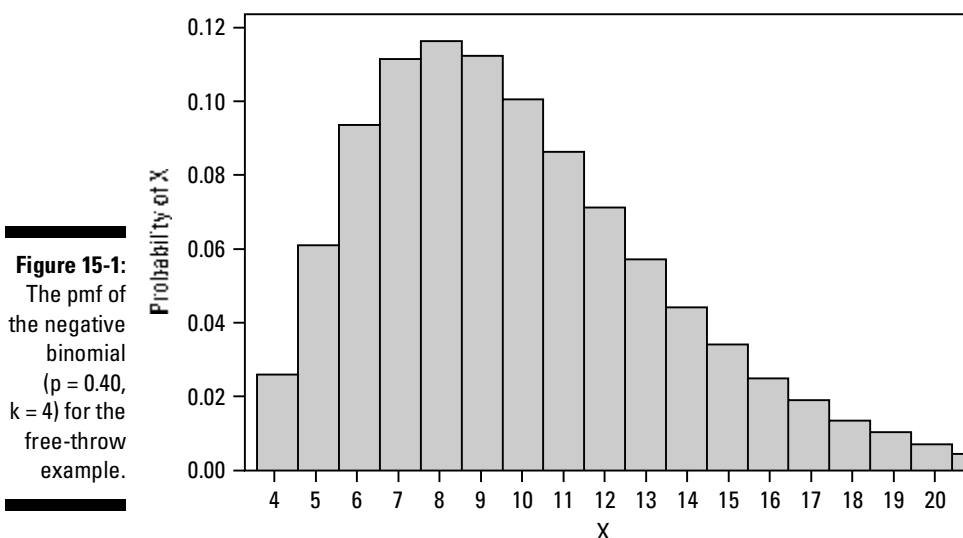
The coefficient in the free-throw problem is the combination 9 choose 3. You want the probability that the fourth success comes on the tenth try, so you force the tenth shot to be a success when calculating the probability. That means of the remaining nine shots, three of them have to be successes, and you're counting the number of ways to arrange those. (After you arrange the successes, the failures automatically take their places.)

Taking calculations for the negative binomial a bit further

Is it easier or harder to achieve, say, a fourth success when you have more tries? What about fewer tries? You can take calculations for negative binomial probability further to see how a pattern develops in the pmf of the negative binomial as the total number of trials changes. Table 15-1 shows you the first few values in the pmf for the free-throw example I present in the introduction to this section.

Table 15-1 First 20 Values of the Negative Binomial pmf ($p = 0.40$, $k = 4$) for the Free-Throw Example	
<i>X = Total # of Trials to Make 4 Baskets</i>	<i>P(x)</i>
4	0.026
5	0.061
6	0.092
7	0.111
8	0.116
9	0.111
10	0.100
11	0.086
12	0.071
13	0.057
14	0.044
15	0.034
16	0.025
17	0.019
18	0.014
19	0.010
20	0.007

Figure 15-1 shows the graph of the pmf that corresponds to Table 15-1.



The figure shows that the pmf of the distribution is skewed to the right, peaking at around eight (the mode of the distribution). The probabilities increase to a point and then start decreasing and trailing off into infinity. Although the point where the peak occurs differs for each negative binomial situation, all the situations have this overall skewed shape. You can also see that the median (the point that splits the probability in half on the graph) appears to be around eight, too. Because this distribution is skewed to the right, you can expect the mean to be driven upward (see Chapter 7 for all the basics on the shape, center, and spread of probability distributions).

Because the basketball player has a probability of success of 0.40, it's pretty tough for him to get all the successes right away ($X = 4$), but the task becomes easier as the number of trials increases (to a point). The easiest situation is when the player has around eight tries to make it happen, because the probability of success is 0.40. If the probability were smaller, the player would have a much harder time getting four successes in eight tries; the peak would come much later for a negative binomial with $p < 0.40$.



You may think that the probabilities should keep increasing as the number of trials increases. Why, then, do the probabilities start to decrease when you pass $X = 8$? Because it becomes harder and harder to get *only* 4 successes in that number of trials. You would expect, for example, that with 20 tries and a 40 percent chance of success, you would see more than 4 baskets made. The probabilities then continue to decrease forever as X gets larger and larger; there's no end to the probability distribution for a negative binomial. You must account for the tiny chance that the player could be at the gym forever waiting for that fourth basket to drop. The free-throw experiment is an example of a discrete random variable that has a countably infinite number of possible values (see Chapter 7 for more on this topic).



Checking on product quality with the negative binomial

You can use the negative binomial distribution whenever you want to model the number of trials that occur until the k th success occurs. For example, people in quality control often use the negative binomial to determine if the manufacturing process is going according to specifications. They can keep sampling items until they reach the fourth defective one and count how many items they sample until it happens. If it takes a long time to reach the occurrence, that's good news — they would conclude that the process is "in control." If it happens right away, that's

bad news — the managers would conclude that the process is "out of control."

As with any probability situation, however, people can make mistakes in their conclusions. For example, if quality control managers expect the probability of defectives to be small (close to zero), they expect the variability in the results to be larger, too (although after a "long time" passes, exactly how long it takes may not matter so much). But that's the magic of probability. You never know *exactly* what's going to happen.

Exploring the Expected Value and Variance of the Negative Binomial

In the following pages, I present easy-to-use formulas for the expected value, variance, and standard deviation of the negative binomial distribution. The probability of success helps determine the mean of the negative binomial, and as you may expect, it also influences the spread in the distribution. I also explain how the expected value and variance of the negative binomial relate to those of the geometric distribution (see Chapter 14).

The expected value of the negative binomial

For the expected value of the negative binomial, the expected number of trials needed to obtain k successes is $E(X) = \frac{k}{p}$. The probability of success, p , is the determining factor here (because k is fixed). A high chance for success means you should expect to achieve your goal quickly — notice that a large value of p in the denominator results in a small value of $E(X)$. A low chance for success means you shouldn't expect to achieve your goal quickly — notice that a small value of p in the denominator results in a large value of $E(X)$.



It helps to recognize that the formula for $E(X)$ for the negative binomial is just k times the formula for $E(X)$ for the geometric distribution (see Chapter 14). That makes sense because the geometric distribution concentrates on the first success (so $k = 1$). In the negative binomial formula for $E(X)$, you would then have $1 * (1 \div p) = 1 \div p$.

In the free-throw example I present in the previous section, p is 0.40, $1 - p = 0.60$, and $k = 4$. The expected number of trials needed to achieve four successes is $E(X) = 4 \div 0.40 = 10$ tries. This value makes sense when you look at Figure 15-1. The graph peaks at $X = 8$, but the mean is higher than that because the distribution is skewed to the right. All those unlikely but possible situations where the basketball player needs a ton of tries to make four baskets have to be taken into account, which drives up the mean.



The formula for the expected value doesn't guarantee that the basketball player will achieve his fourth success on the tenth shot. You're just finding an overall expected average number of trials to obtain the k th success, given the particular value of p . To figure out how much variability you can expect in your results, you need the standard deviation.

The variance and standard deviation of the negative binomial

The formula for the variance of the negative binomial distribution is $V(X) = \frac{k(1-p)}{p^2}$. The standard deviation of the negative binomial is the square root of the variance, which is $\sqrt{\frac{k(1-p)}{p^2}}$.

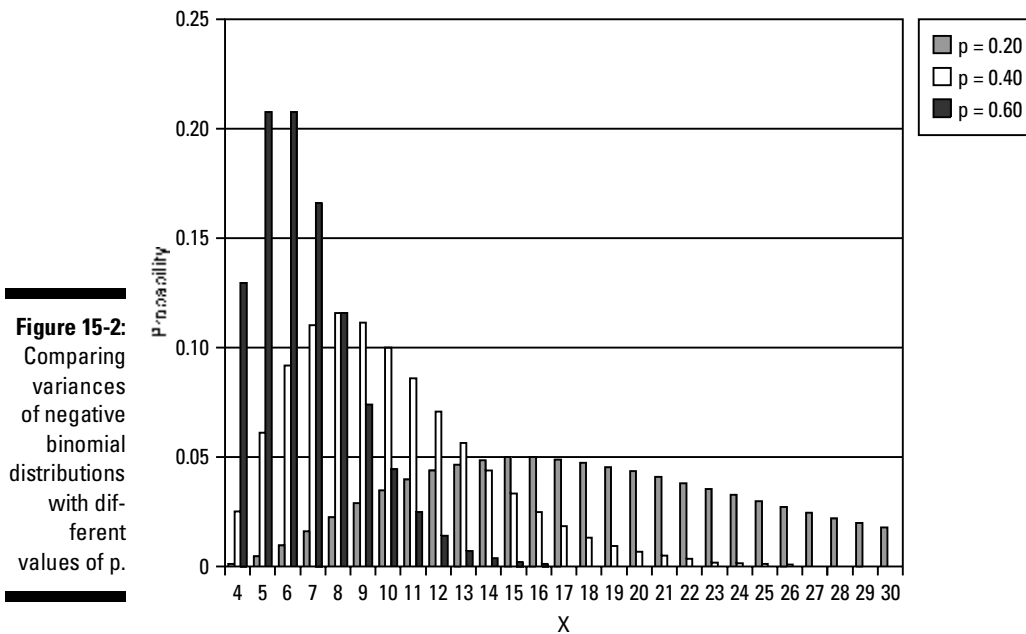
Looking at the free-throw example from the previous sections, you have $p = 0.40$, $1 - p = 0.60$, and $k = 4$. The variance in the number of tries needed to make four baskets is $V(X) = \frac{k(1-p)}{p^2} = \frac{4(1-0.40)}{0.40^2}$, which equals 15.

The standard deviation is the square root of 15, which is 3.87. That means the average amount of deviation you can expect in the number of tries needed to make four baskets is 3.87.

Notice how the trade-off between the probability of success (in the denominator) and the probability of failure (in the numerator) helps you figure out how much variability to expect in the results when you do the experiment over and over again. If your value of p is close to one, you should expect plenty of successes right away, and they'll happen close together, which means the distribution will have less variation. Take it as good news if p is close to one (because your expected number of trials is low and is expected to stay consistent). In the formula, a larger value of p in the denominator makes $V(X)$ smaller and forces $1 - p$ in the numerator to be smaller, which decreases $V(X)$ even more.

However, a value of p closer to zero has the opposite effect; it increases $V(X)$ because the expected successes are likely to happen much later, and you don't know exactly when. Take it as somewhat good news if p is close to zero (because your expected number of trials is high, but you expect your results to change a lot from experiment to experiment; keep hope alive!).

Figure 15-2 shows the pmf for different negative binomial distributions; notice how different their spreads are. If p is close to one, the variability is low; if p is close to zero, the amount of spread (measured by the variance) is larger.



Applying the expected value and variance formulas

Suppose, for example, that you, Bob, and Jaimie play basketball in your driveway, and you keep track of your overall percentages when making baskets from your free-throw line. Suppose your percentage is $p = 0.90$ (after all, it is your driveway), Bob's percentage is $p = 0.50$, and Jaimie's percentage is a pretty sad $p = 0.10$. Which of you will have the most variability in making versus missing shots? To find the answer, you set up a game where you keep shooting until you make 20 baskets. The baskets don't have to fall in a row, but when you make the 20th basket, you stop. The person who takes the least amount of shots wins. Who do you expect to take the least amount of shots to get to 20 baskets? And how much variability can you expect in these results if you play the game over and over?

It makes sense that you should win because you have the highest percentage. How long should it take each of you to get there? Using the formula for expected value of the negative binomial (with $k = 20$), you take $20 \div p$ for each person to get the expected number of trials until 20 baskets are made. For you, that value is $20 \div 0.90 = 22.22$ (remember that this is a long-term average, so you'll shoot between 22 and 23 baskets on average). Bob's expected value is $20 \div 0.50 = 40$ tries, and Jaimie, well, he's in it for the long haul. His expected number of tries is $20 \div 0.10$, which is a whopping 200 tries before the 20th basket falls. (Maybe you should shorten the game to just 10 baskets . . .)

Now, if you repeat this game over and over, assuming all the percentages remain steady, the amount of variability in the results should be the lowest for you and the highest for Jaimie. Using the formula for the variance of the negative binomial, your variance is $20 * (1 - 0.90) \div 0.90^2$, which is 2.47. Your standard deviation is the square root of this number, which is 1.57. So, on average, the number of tries you need to make 20 baskets will deviate by only 1.57. For Bob, the variance is $20 * (1 - 0.50) \div 0.50^2$, which is 40, and his standard deviation is the square root, which is 6.32. He's got a higher amount of variability than you have from game to game, which means he's less consistent. What about Jaimie? His variance is $20 * (1 - 0.10) \div 0.10^2$, which is 1,800, with a standard deviation of 42.43. Jaimie's standard deviation is the highest, which means he's the least consistent, but there's an upside to this. With that much variability, he can do a lot worse than expected or a lot better than expected from game to game. That's where lottery winners are born!

Chapter 16

Remaining Calm about the Hypergeometric Distribution

In This Chapter

- ▶ Covering the basics of the hypergeometric probability model
 - ▶ Pitting the hypergeometric model against other major probability models
 - ▶ Following the hypergeometric formula and the associated boundary conditions
 - ▶ Figuring probabilities for a hypergeometric distribution with the pmf
 - ▶ Determining the expected value and variance
-

You're called upon to use the scary-sounding but harmless hypergeometric distribution in situations where you take a sample without replacement and you want to find the probability that a specific number of items from the sample have a certain desired characteristic of interest. It's very similar to the binomial distribution (see Chapter 8), except that you sample without replacement (which means that after you draw an item or individual from the population, you don't put the item or individual back into the population; hence, the number of items or individuals remaining decreases by one each time). To find probabilities for the hypergeometric, you in essence divide the population into two groups: the group that has the desired characteristic of interest, and the group that doesn't. And to get the sample you need, you take a certain number from each group.

In this chapter, you discover how to recognize when the hypergeometric probability model is in play, find probabilities under the hypergeometric model, and calculate the expected value and variance of a hypergeometric distribution.

Zooming In on the Conditions for the Hypergeometric Model


As with other probability models I discuss in this book, situations must meet certain conditions before you can apply the hypergeometric model. In this section, you review these conditions for the hypergeometric and look at how the hypergeometric differs from the binomial distribution.

A probability situation must meet the following conditions in order for a random variable, X , to have a hypergeometric distribution:

- ✓ You sample without replacement from a population of N total individuals (see the introduction to this chapter for a definition of without replacement).
- ✓ Every individual in the population has an equal chance of being sampled.
- ✓ You can classify the total population into two groups, or subpopulations:
 - **The marked subpopulation:** Individuals who have the characteristic that you're interested in; for example, owning a cell phone, living on a farm, or having a particular disease.
 - **The unmarked subpopulation:** Represented by everyone else in the overall population who isn't considered to be among the marked subpopulation.
- ✓ The total number of marked individuals in the population and the total population size are fixed and known.
- ✓ X counts the total number of marked individuals in the sample.

For example, suppose you have a group of 20 people — 12 men and 8 women. You randomly select five people to be on a committee, and it turns out that all five people you select are men. You wonder, "What's the chance of that happening?" This is a job for the hypergeometric distribution. Why?

- ✓ You're sampling 5 people from the population of 20 individuals without replacement.
- ✓ Because you conduct a random sample, you make sure every individual has an equal chance of being selected.
- ✓ You can classify the total (fixed) population of 20 people into two groups: 12 men and 8 women — and both of these numbers are fixed.

 You define X as the total number of men in the sample because that's the group you're interested in — you want to know the probability of selecting five men. For this example, the 12 men are considered to be the marked population. (Because you want the probability that you select five men, you want $P[X = 5]$.) The eight women in the population are the unmarked population. (In this problem, in order to choose five people, all men, the number of women selected has to be zero.)



The hypergeometric distribution is similar to the binomial in that you are taking a sample and determining how many individuals in the sample have a certain characteristic of interest. However, with the binomial you are sampling with replacement, so your probability of “success” (being in the desired category) is the same for each individual, and in the hypergeometric you are sampling without replacement, so the probability of “success” is not the same for each individual. After you determine if the sampling is done with or without replacement, you'll know which distribution to use.

Finding Probabilities for the Hypergeometric Model

The hypergeometric distribution uses combinations (see Chapter 5) to find probabilities. A *combination* is a formula used to figure out the number of ways to select a certain number of items from a group without replacement, where the order of the items doesn't matter. In this section, I present the hypergeometric probability mass function (pmf) used to find probabilities, its applications, and an overall discussion of how the formula comes about.

Setting up the hypergeometric pmf

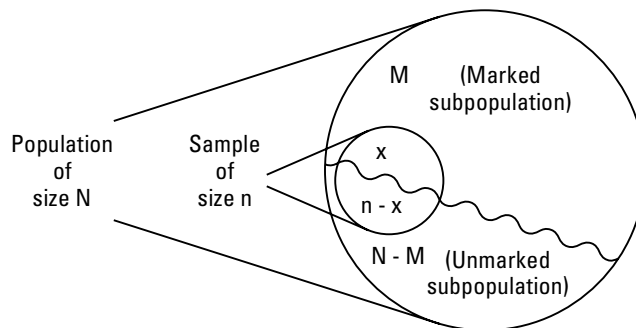
The probability mass function (pmf) of the hypergeometric distribution is

$$\frac{\binom{M}{x} \binom{N-M}{n-x}}{\binom{N}{n}}, \text{ where } \max(0, M+n-N) \leq x \leq \min(M, n); n \text{ is the sample size,}$$

x is the number of marked individuals in the sample, N is the total population size, and M is the total number of marked individuals in the population.

Here's what's going on in the sampling process to get the hypergeometric. You have a population of size N , and you classify it into two separate subpopulations: M individuals in the marked subpopulation and $N - M$ individuals in the unmarked subpopulation (see Figure 16-1). The marked population has the desired characteristic of interest, and the unmarked population doesn't.

Figure 16-1:
Sampling
from
marked and
unmarked
individuals
in the hyper-
geometric.



Now you reach into the population and take a random sample of size n . Some members of that sample are members of the marked subpopulation, and other members of the sample are members of the unmarked population (you don't know in advance which ones you're picking, so you don't know how many you'll get of each type). Then x represents the number in the sample that are from the marked population and $n - x$ members of the sample are from the unmarked population.

The pmf of the hypergeometric lays out all the possible values of X and their probabilities under the hypergeometric model. You need two combinations in the numerator, because you want to select x items from the marked population (of $N - M$ total individuals) and $n - x$ items from the unmarked population (of M total individuals). The denominator contains one combination, which represents the number of ways to choose a sample of size n from the entire population of N individuals.

Say, for example, that you form a committee of 5 from 20 people, with 12 men and 8 women. You have $n = 5$ as the sample size; X is the total number of men on the committee; $N = 20$ as the total amount of people in the population; and $M = 12$ as the total number of men in the population. The pmf for this example

$$\text{is } P(X = x) = \frac{\binom{12}{x} \binom{20-12}{5-x}}{\binom{20}{5}}.$$



Notice the part in the pmf that tells you the possible values of X — namely that $\max(0, M + n - N) \leq x \leq \min(M, n)$. For this particular probability model, defining the possible values for X is more complex than usual. Some seemingly strange but important conditions apply to X as seen in the

previous formula. This double inequality breaks down into two boundary conditions for X that have to happen at the same time (see the next section “Breaking down the boundary conditions for X ” for more on the boundary conditions).

For the committee example, you have $\max(0, 12 + 5 - 20) \leq x \leq \min(12, 5)$. This simplifies to $0 \leq x \leq 5$, which makes sense because x represents the number of men in the sample, and the sample size is 5, so you can have 0, 1, 2, 3, 4, or 5 men in the sample. You know you can have up to 5 men in the sample because 12 men appear in the population (so you won’t run out, so to speak).



Using the language of combinations (see Chapter 5) helps you quickly identify how to set up the probabilities for a hypergeometric. In fact, hypergeometric problems are nothing but combinations themselves. The numerator is “ M choose x ” times “ $N - M$ choose $n - x$,” and the denominator is “ N choose n .”

In symbols, you have $\binom{M}{x} \binom{N-M}{n-x}$ for the numerator, and $\binom{N}{n}$ for the denominator. Now you have the pmf for the hypergeometric distribution.



Because you divide the entire population into two groups and classify everyone in the sample into two groups, the two “top numbers” in the numerator combinations have to sum to N , and the two “bottom numbers” in the numerator combinations have to sum to n . This is a good way to check to be sure you’ve set up your hypergeometric distribution correctly.

Breaking down the boundary conditions for X

X has two boundary conditions that have to happen at the same time:

- ✓ X must be less than or equal to the minimum of two values: M and n .
- ✓ X must be greater than or equal to the maximum of two values: zero and $M + n - N$.

Depending on whether M turns out to be small, relative to n , or large, relative to n , your results for $\min(M, n)$ and $\max(0, M + n - N)$ in the boundary conditions will be different. In this section, you look at two examples that help you understand how to figure out and understand boundary conditions when M is small or large, relative to n .

Boundary conditions when M is small

Here’s an example that you can use to practice setting up boundary conditions for small values of M . Suppose Sue has ten guests who choose (without replacement) envelopes from a bag that contains 22 envelopes: 2 envelopes

contain \$20 each, and 20 envelopes contain \$1 each. The population is the total group being selected from, which is the 22 envelopes, so $N = 22$. The marked population is the number of envelopes with the characteristic of interest (the \$20), so M is 2. The sample size is $n = 10$ because ten guests attend the party and get to choose an envelope (without replacement). You want to count X , the number of guests who receive a \$20 prize. What are the possible values of X ? According to the first boundary condition, X has to be less than or equal to the minimum of $M = 2$ and $n = 10$. So, X has to be less than or equal to 2.

Now look at Sue's party problem with the second boundary in mind. $N = 22$ envelopes, $M = 2$ envelopes that are marked (contain \$20), and $n = 10$ guests who each choose an envelope (without replacement). Here, X represents the number of guests who select a \$20 prize. You know that the maximum possible value of X is 2; what's the minimum value of X ? According to the second boundary condition, X must be greater than or equal to the maximum of 0 and $(M + n - N)$, which is $2 + 10 - 22 = -10$. In this case, the maximum of 0 and -10 is 0, so X has to be greater than or equal to 0. (Zero is always in this boundary condition to prevent X from being negative.)

Putting the two boundary conditions together for Sue's party, X , the number of people winning a \$20 prize, has to be greater than or equal to 0 and less than or equal to 2; in other words, X can take on three possible values: 0, 1, or 2.

Boundary conditions when M is large

You can stick with the example of Sue's party I use in the previous section to look at larger possible values of M . The second boundary condition in the previous section turned out to be zero, the easier situation. But now suppose Sue has 20 \$20 prizes rather than 2, and the other 2 prizes are for \$1 each. In other words, suppose Sue has 22 prizes, and 20 of them are \$20 envelopes. She has 10 guests who each choose an envelope without replacement; let X be the number of \$20 prizes chosen. What are the boundary conditions on X ? They both change compared to Sue's previous situation.

According to the first boundary condition applied to Sue's new scenario, X has to be less than or equal to the minimum of M (which is 20) and n (which is 10). That means X has to be less than or equal to 10. This makes sense because she has only 10 people choosing, and you can't choose more \$20 prizes than there are guests.

Now, according to the second boundary condition, X has to be greater than or equal to the maximum of 0 and $M + n - N$. In this case $M = 20$, $n = 10$, and $N = 22$, so $M + n - N$ equals $20 + 10 - 22$, which is 8. So, X has to be greater than or equal to the maximum of 0 and 8 (which is 8). Why is this? Observing the lowest number of \$20 envelopes chosen happens when the highest number of \$1 envelopes are chosen. Because Sue provides only two of the \$1 envelopes, when ten people select from the bag, two can pick the \$1 envelopes, but the

rest of them have to get \$20 envelopes. This is represented mathematically by saying $n - X$ (the number of \$1 prizes chosen) has to be less than or equal to $N - M$ (the number of \$1 prizes in the bag). Rewritten, this says that $n - X \leq N - M$, or that $X > M + n - N$. So in this second example, the number of \$20 prizes, X , has to be greater than or equal to 8 and less than or equal to 10, so X must be 8, 9, or 10.



I believe it's always better and easier to understand the formulas and conditions as much as possible instead of trying to memorize them. You can rely on your understanding much better than your memory in exam situations. Try to understand the formulas and conditions whenever you can to increase your probability of being successful.

Finding and using the pmf to calculate probabilities

The biggest application of the hypergeometric distribution is to find probabilities in situations where you sample without replacement and you want to find the probability of getting a certain number of individuals who have a desired characteristic of interest. That means, given the question, you can determine for that problem what constitutes the marked subpopulation — the one with the individuals being counted by X . In this section, you work through the set up and calculation of hypergeometric probabilities.

To calculate a probability for X by using the hypergeometric probability mass function (pmf; see the section “Setting up the hypergeometric pmf”), perform the following steps:

- 1. Determine the values of N (total population size); M (number of marked individuals in the population); n (the sample size); and X (what you're interested in counting). (Notice that the definition of marked individuals depends on what X is counting.)**

Take a look at the committee example I present in the section “Zooming In on the Conditions for the Hypergeometric Model.” The total population size is $N = 20$, the total number of men in the population is $M = 12$, the sample size is $n = 5$, and X is counting the number of men in the sample. You want to know the chance of choosing all men for the committee of five.

- 2. If you need to find the entire pmf, determine the boundary conditions for X — the minimum and maximum values of X . (See the section “Breaking down the boundary conditions for X .”) If you need only a single probability for X , proceed to Step 3.**

You don't want the entire pmf of X — just one probability for X — so you skip Step 2 and go on to Step 3.

3. Write down the probability you want to find in terms of X.

Because you want the probability that men make up the whole sample, you want $P(X = 5)$.

4. Substitute the values of M, N, n, and X into the pmf for the hypergeometric and simplify.

Putting the values of $N = 20$, $M = 12$, $N - M = 8$, $n = 5$, and $X = 5$ into the pmf of the hypergeometric, you find that $P(X = 5)$ is equal to

$$P(X = 5) = \frac{\binom{12}{5} \binom{20-12}{5-5}}{\binom{20}{5}} = \frac{\binom{12}{5} \binom{8}{0}}{\binom{20}{5}} = \frac{\left(\frac{12!}{5!7!}\right) * 1}{\frac{20!}{5!15!}} = \frac{792}{15,504}, \text{ which is } 0.05.$$

In other words, if you have 12 men and 8 women in a group, and you randomly choose 5 people to be on a committee, you'll choose all men for your committee about 5 percent of the time. Not very often!



Make sure you represent only one of the two populations with X throughout all your calculations. Switching the role of X midstream isn't allowed. In the committee example, if you define X as the number of men in the sample, and you want the probability of choosing five men to be on the committee, you want $P(X = 5)$. If you want the probability of choosing five women to be on the committee, you want $P(X = 0)$ because five women making up the committee means that you choose zero men to be on the committee, and X represents the men, not the women. Note also that in most cases, the probability of choosing five men isn't equal to the probability of choosing five women (when the total populations of men and women aren't the same size).

Table 16-1 shows the entire pmf of X for the committee example — it shows the probability of choosing X men to be on the committee for all possible values of X: zero through five.

Table 16-1 The pmf of X = Number of Men on a Committee of 5 (M = 12, N = 20)

<i>X = Number of Men on the Committee</i>	<i>P(X = x)</i>
0	0.004
1	0.054
2	0.238
3	0.397
4	0.255
5	0.051

Measuring the Expected Value and Variance of the Hypergeometric

In this section, you find and work with formulas for the expected value, variance, and standard deviation of the hypergeometric distribution.

The expected value of the hypergeometric

The *expected value*, $E(X)$, of the hypergeometric distribution is the number of marked individuals you expect to find in the sample. The total number of marked individuals in the population is M , and the total population size is N , so the total proportion of marked individuals in the population is $M \div N$. Now you just need to multiply by the sample size to get the expected number of individuals in the sample that belong to the marked population. This gives you the formula for the expected value of X : $n * \frac{M}{N}$.



The formula for $E(X)$ for the hypergeometric distribution is similar in nature to the expected value for the binomial distribution (see Chapter 8); that is, the proportion of “successes” (or marked individuals) times the sample size gives you the expected number of “successes” (or marked individuals) in your sample.

Using the committee example from the section “Zooming In on the Conditions for the Hypergeometric Model,” assume that you have a total population, N , of 20 members, with the number of marked individuals, M , equal to 12 men (with 8 women making up the rest of the population). You want to choose at random a sample size, n , of five people to be on a committee. What’s the expected number of men on the committee? Using the formula for $E(X)$ for a hypergeometric, you have $n * \frac{M}{N} = 5 * \frac{12}{20}$, which equals 3.

The variance and standard deviation of the hypergeometric

The variance of a distribution, $V(X)$, is the average squared deviation (or distance) from the mean in the long term, and the standard deviation, $S(X)$, is the average expected deviation from the mean in the long term (see Chapter 7). In other words, how much do you expect your results to vary if you were to take your sample again? The variance depends on the sample size, n , the population size, N , and the total number of individuals in the marked population (who have the desired characteristic of interest), M .

The formula for the variance of the hypergeometric distribution is

$\frac{nM}{N^2(N-1)}(N-M)(N-n)$. The standard deviation of the hypergeometric is the square root of the variance, which is $\sqrt{\frac{nM}{N^2(N-1)}(N-M)(N-n)}$.

Using the committee example I review in the previous section, you have $N = 20$, $M = 12$, $N - M = 8$, and $n = 5$. Putting these values into the formulas for variance and standard deviation, you get a variance equal

to $\frac{nM}{N^2(N-1)}(N-M)(N-n) = \frac{5 * 12}{20^2(20-1)}(20-12)(20-5) = \frac{60}{7,600}(8)(15) = \frac{7,200}{7,600} = 0.947$. You get a standard deviation equal to the square root of 0.947, which is 0.973. Now you can state the average expected deviation in the number of men on the committee over repeated trials of the experiment.



Fishing for a population estimate

You often know the total population size when you must work with a distribution, but one of the biggest uses of the hypergeometric distribution is estimating the total size of a population. The hypergeometric is especially helpful in situations where the population is hidden (doesn't identify itself to the world) or rare (not many individuals make up the population). The technique that you use to estimate the size of a population is called the *capture-recapture* technique. For example, suppose you want to estimate the number of fish that live in your farm pond. You go to the pond, "capture" some fish, mark them, and then release them. The next day, you take another sample of fish and count the number of marked fish that you "recapture," and then you release them.

In terms of the hypergeometric distribution, you divide the pond into two populations on the first day: the marked population (M) and the unmarked population ($N - M$). On the second day, you have a sample of size n , and you essentially categorize those fish into two groups: those that you marked (R , the recaptures) and those that you didn't ($n - R$). You *could* work on finding probabilities for certain outcomes to happen, but the most common use of the hypergeometric in

this situation is to estimate N , the total population size. How do you do it?

Here's what happens (assuming the conditions of the hypergeometric hold true; see the section "Zooming In on the Conditions for the Hypergeometric Model" earlier in this chapter). You can expect the number of recaptured fish (R) compared to the sample size (n) to be proportionate to the number of marked fish (M) compared to the total number of fish in the pond (N); in other words, $R/n = M/N$. This setup is good news, because you know the values of n (sample size on the second day), R (number of marked fish recaptured on the second day), and M (the number of fish marked on the first day). All you have to do is cross-multiply and solve for N : $N = (M * n) \div R$.

For example, say you mark 20 fish on the first day (M). On the second day, you sample 50 fish (n) and find that you marked 5 (R) — these 5 are the recaptured fish. Now, to estimate N , the total number of fish in the pond, you take $N = (M * n) \div R = [(20)(50)] \div 5 = 200$. You estimate that the pond holds about 200 fish. Capture-recapture is a backdoor way of making a pretty good guess at the population size.