CHAPTER 17

Probability Models

Suppose a cereal manufacturer puts pictures of famous athletes on cards in boxes of cereal, in the hope of increasing sales. The manufacturer announces that 20% of the boxes contain a picture of Tiger Woods, 30% a picture of David Beckham, and the rest a picture of Serena Williams.

Sound familiar? In Chapter 11 we simulated to find the number of boxes we'd need to open to get one of each card. That's a fairly complex question and one well suited for simulation. But many important questions can be answered more directly by using simple probability models.

Searching for Tiger

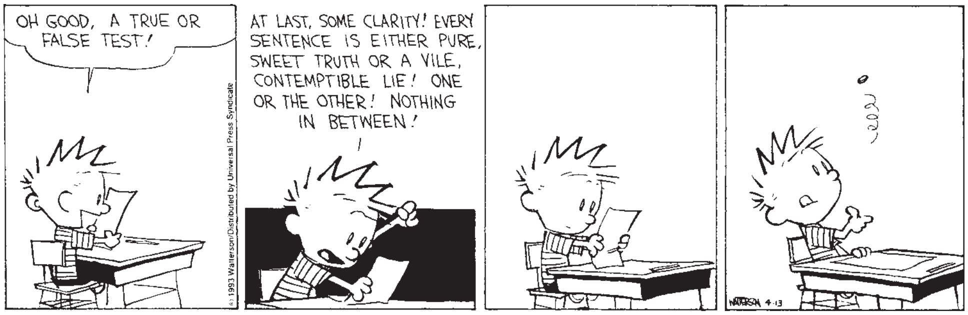
You're a huge Tiger Woods fan. You don't care about completing the whole sports card collection, but you've just *got* to have the Tiger Woods picture. How many boxes do you expect you'll have to open before you find him? This isn't the same question that we asked before, but this situation is simple enough for a probability model.

We'll keep the assumption that pictures are distributed at random and we'll trust the manufacturer's claim that 20% of the cards are Tiger. So, when you open the box, the probability that you succeed in finding Tiger is 0.20. Now we'll call the act of opening *each* box a trial, and note that:

- There are only two possible outcomes (called *success* and *failure*) on each trial. Either you get Tiger's picture (success), or you don't (failure).
- In advance, the probability of success, denoted p, is the same on every trial. Here $p = 0.20$ for each box.
- As we proceed, the trials are independent. Finding Tiger in the first box does not change what might happen when you reach for the next box.

Situations like this occur often, and are called **Bernoulli trials.** Common examples of Bernoulli trials include tossing a coin, looking for defective products rolling off an assembly line, or even shooting free throws in a basketball game. Just as we found equally likely random digits to be the building blocks for our simulation, we can use Bernoulli trials to build a wide variety of useful probability models.

Daniel Bernoulli (1700–1782) was the nephew of Jacob, whom you saw in Chapter 14. He was the first to work out the mathematics for what we now call Bernoulli trials.


OH GOOD, A TRUE OR FALSE TEST!
AT LAST, SOME CLARITY! EVERY SENTENCE IS EITHER PURE, SWEET TRUTH OR A VILE, CONTEMPTIBLE LIE! ONE OR THE OTHER! NOTHING IN BETWEEN!
© 1993 Watterson/Distributed by Universal Press Syndicate
WATERSN 4-13

Calvin and Hobbes © 1993 Watterson. Reprinted with permission of UNIVERSAL PRESS SYNDICATE. All rights reserved.

A S ***Activity:* Bernoulli Trials.** Guess what! We've been generating Bernoulli trials all along. Look at the Random Simulation Tool in a new way.

Back to Tiger. We want to know how many boxes we'll need to open to find his card. Let's call this random variable Y = # boxes, and build a probability model for it. What's the probability you find his picture in the first box of cereal? It's 20%, of course. We could write $P(Y = 1) = 0.20$.

How about the probability that you don't find Tiger until the second box? Well, that means you fail on the first trial and then succeed on the second. With the probability of success 20%, the probability of failure, denoted q, is $1 - 0.2 = 80\%$. Since the trials are independent, the probability of getting your first success on the second trial is $P(Y = 2) = (0.8)(0.2) = 0.16$.

Of course, you could have a run of bad luck. Maybe you won't find Tiger until the fifth box of cereal. What are the chances of that? You'd have to fail 4 straight times and then succeed, so $P(Y = 5) = (0.8)^4(0.2) = 0.08192$.

How many boxes might you expect to have to open? We could reason that since Tiger's picture is in 20% of the boxes, or 1 in 5, we expect to find his picture, on average, in the fifth box; that is, $E(Y) = \frac{1}{0.2} = 5$ boxes. That's correct, but not easy to prove.

The Geometric Model

TI-*nspire*

Geometric probabilities. See what happens to a geometric model as you change the probability of success.

We want to model how long it will take to achieve the first success in a series of Bernoulli trials. The model that tells us this probability is called the **Geometric probability model.** Geometric models are completely specified by one parameter, p, the probability of success, and are denoted Geom(p). Since achieving the first success on trial number x requires first experiencing $x - 1$ failures, the probabilities are easily expressed by a formula.

NOTATION ALERT:

Now we have two more reserved letters. Whenever we deal with Bernoulli trials, p represents the probability of success, and q the probability of failure. (Of course, $q = 1 - p$.)

GEOMETRIC PROBABILITY MODEL FOR BERNOULLI TRIALS: Geom(p)

p = probability of success (and $q = 1 - p$ = probability of failure)
X = number of trials until the first success occurs

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

Expected value: $E(X) = \mu = \dfrac{1}{p}$

Standard deviation: $\sigma = \sqrt{\dfrac{q}{p^2}}$

FOR EXAMPLE Spam and the Geometric model

Postini is a global company specializing in communications security. The company monitors over 1 billion Internet messages per day and recently reported that 91% of e-mails are spam!

Let's assume that your e-mail is typical—91% spam. We'll also assume you aren't using a spam filter, so every message gets dumped in your inbox. And, since spam comes from many different sources, we'll consider your messages to be independent.

Questions: Overnight your inbox collects e-mail. When you first check your e-mail in the morning, about how many spam e-mails should you expect to have to wade through and discard before you find a real message? What's the probability that the 4th message in your inbox is the first one that isn't spam?

There are two outcomes: a real message (success) and spam (failure). Since 91% of e-mails are spam, the probability of success $p = 1 - 0.91 = 0.09$.

Let X = the number of e-mails I'll check until I find a real message. I assume that the messages arrive independently and in a random order. I can use the model Geom(0.09).

$$E(X) = \frac{1}{p} = \frac{1}{0.09} = 11.1$$

$$P(X = 4) = (0.91)^3(0.09) = 0.0678$$

On average, I expect to have to check just over 11 e-mails before I find a real message. There's slightly less than a 7% chance that my first real message will be the 4th one I check.

Note that the probability calculation isn't new. It's simply Chapter 14's Multiplication Rule used to find $P(\text{spam} \cap \text{spam} \cap \text{spam} \cap \text{real})$.

MATH BOX

We want to find the mean (expected value) of random variable X, using a geometric model with probability of success p.

First, write the probabilities:

x	1	2	3	4	⋯
$P(X = x)$	p	qp	q^2p	q^3p	⋯

The expected value is: $E(X) = 1p + 2qp + 3q^2p + 4q^3p + \cdots$

Let $p = 1 - q$: $= (1 - q) + 2q(1 - q) + 3q^2(1 - q) + 4q^3(1 - q) + \cdots$

Simplify: $= 1 - q + 2q - 2q^2 + 3q^2 - 3q^3 + 4q^3 - 4q^4 + \cdots$

That's an infinite geometric series, with first term 1 and common ratio q: $= 1 + q + q^2 + q^3 + \cdots$

$$= \frac{1}{1 - q}$$

So, finally . . . $E(X) = \dfrac{1}{p}.$

Independence

One of the important requirements for Bernoulli trials is that the trials be independent. Sometimes that's a reasonable assumption—when tossing a coin or rolling a die, for example. But that becomes a problem when (often!) we're looking at situations involving samples chosen without replacement. We said that whether we find a Tiger Woods card in one box has no effect on the probabilities in other boxes. This is *almost* true. Technically, if exactly 20% of the boxes have Tiger Woods cards, then when you find one, you've reduced the number of remaining Tiger Woods cards. If you knew there were 2 Tiger Woods cards hiding in the 10 boxes of cereal on the market shelf, then finding one in the first box you try would clearly change your chances of finding Tiger in the next box. With a few million boxes of cereal, though, the difference is hardly worth mentioning.

If we had an infinite number of boxes, there wouldn't be a problem. It's selecting from a finite population that causes the probabilities to change, making the trials not independent. Obviously, taking 2 out of 10 boxes changes the probability. Taking even a few hundred out of millions, though, makes very little difference. Fortunately, we have a rule of thumb for the in-between cases. It turns out that if we look at less than 10% of the population, we can pretend that the trials are independent and still calculate probabilities that are quite accurate.

The 10% Condition: Bernoulli trials must be independent. If that assumption is violated, it is still okay to proceed as long as the sample is smaller than 10% of the population.

STEP-BY-STEP EXAMPLE Working with a Geometric Model

People with O-negative blood are called "universal donors" because O-negative blood can be given to anyone else, regardless of the recipient's blood type. Only about 6% of people have O-negative blood.

Questions:

1. If donors line up at random for a blood drive, how many do you expect to examine before you find someone who has O-negative blood?
2. What's the probability that the first O-negative donor found is one of the first four people in line?

THINK

Plan State the questions.

I want to estimate how many people I'll need to check to find an O-negative donor, and the probability that 1 of the first 4 people is O-negative.

Check to see that these are Bernoulli trials.

✓ There are two outcomes:
success = O-negative
failure = other blood types

✓ The probability of success for each person is $p = 0.06$, because they lined up randomly.

✓ **10% Condition:** Trials aren't independent because the population is finite, but the donors lined up are fewer than 10% of all possible donors.

Variable Define the random variable.

Let X = number of donors until one is O-negative.

Model Specify the model.

I can model X with Geom(0.06).

SHOW

Mechanics Find the mean.

Calculate the probability of success on one of the first four trials. That's the probability that $X = 1, 2, 3,$ *or* 4.

$$E(X) = \frac{1}{0.06} \approx 16.7$$

$$\begin{aligned} P(X \le 4) &= P(X = 1) + P(X = 2) + \\ &\quad P(X = 3) + P(X = 4) \\ &= (0.06) + (0.94)(0.06) + \\ &\quad (0.94)^2(0.06) + (0.94)^3(0.06) \\ &\approx 0.2193 \end{aligned}$$

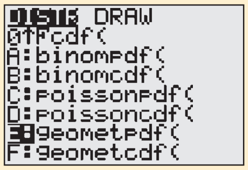
TELL

Conclusion Interpret your results in context.

Blood drives such as this one expect to examine an average of 16.7 people to find a universal donor. About 22% of the time there will be one within the first 4 people in line.

TI TIPS

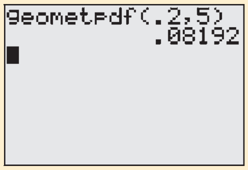
Finding geometric probabilities



Your TI knows the geometric model. Just as you saw back in Chapter 6 with the Normal model, commands to calculate probability distributions are found in the `2nd DISTR` menu. Have a look. After many others (Don't drop the course yet!) you'll see two Geometric probability functions at the bottom of the list.

- `geometpdf(.`
  The "pdf" stands for "probability density function." This command allows you to find the probability of any *individual* outcome. You need only specify *p*, which defines the Geometric model, and *x*, which indicates the number of trials until you get a success. The format is `geometpdf(p,x)`.

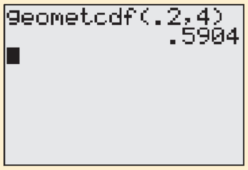
  For example, suppose we want to know the probability that we find our first Tiger Woods picture in the fifth box of cereal. Since Tiger is in 20% of the boxes, we use $p = 0.2$ and $x = 5$, entering the command `geometpdf(.2,5)`. The calculator says there's about an 8% chance.



```
geometpdf(.2,5)
               .08192
```

- `geometcdf(.`
  This is the "cumulative density function," meaning that it finds the sum of the probabilities of several possible outcomes. In general, the command `geometcdf(p,x)` calculates the probability of finding the first success *on or before* the *x*th trial.

  Let's find the probability of getting a Tiger Woods picture by the time we open the fourth box of cereal—in other words, the probability our first success comes on the first box, or the second, or the third, or the fourth. Again we specify $p = 0.2$, and now use $x = 4$. The command `geometcdf(.2,4)` calculates all the probabilities and adds them. There's about a 59% chance that our quest for a Tiger Woods photo will succeed by the time we open the fourth box.



```
geometcdf(.2,4)
                .5904
```

The Binomial Model

We can use the Bernoulli trials to answer other questions. Suppose you buy 5 boxes of cereal. What's the probability you get *exactly* 2 pictures of Tiger Woods? Before, we asked how long it would take until our first success. Now we want to find the probability of getting 2 successes among the 5 trials. We are still talking about Bernoulli trials, but we're asking a different question.

***Activity:* The Binomial Distribution.** It's more interesting to combine Bernoulli trials. Simulate this with the Random Tool to get a sense of how Binomial models behave.

This time we're interested in the *number of successes* in the 5 trials, so we'll call it X = number of successes. We want to find $P(X = 2)$. This is an example of a **Binomial probability.** It takes two parameters to define this **Binomial model:** the number of trials, n, and the probability of success, p. We denote this model Binom(n, p). Here, $n = 5$ trials, and $p = 0.2$, the probability of finding a Tiger Woods card in any trial.

Exactly 2 successes in 5 trials means 2 successes and 3 failures. It seems logical that the probability should be $(0.2)^2(0.8)^3$. Too bad! It's not that easy. That calculation would give you the probability of finding Tiger in the first 2 boxes and not in the next 3—*in that order*. But you could find Tiger in the third and fifth boxes and still have 2 successes. The probability of those outcomes in that particular order is $(0.8)(0.8)(0.2)(0.8)(0.2)$. That's also $(0.2)^2(0.8)^3$. In fact, the probability will always be the same, no matter what order the successes and failures occur in. Anytime we get 2 successes in 5 trials, no matter what the order, the probability will be $(0.2)^2(0.8)^3$. We just need to take account of all the possible orders in which the outcomes can occur.

Fortunately, these possible orders are *disjoint*. (For example, if your two successes came on the first two trials, they couldn't come on the last two.) So we can use the Addition Rule and add up the probabilities for all the possible orderings. Since the probabilities are all the same, we only need to know how many orders are possible. For small numbers, we can just make a tree diagram and count the branches. For larger numbers this isn't practical, so we let the computer or calculator do the work.

Each different order in which we can have k successes in n trials is called a "combination." The total number of ways that can happen is written $\binom{n}{k}$ or ${}_nC_k$ and pronounced "n choose k."

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \text{ where } n! \text{ (pronounced "}n\text{ factorial") } = n \times (n-1) \times \cdots \times 1$$

NOTATION ALERT:

Now punctuation! Throughout mathematics $n!$, pronounced "n factorial," is the product of all the integers from 1 to n. For example, $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$.

For 2 successes in 5 trials,

$$\binom{5}{2} = \frac{5!}{2!(5-2)!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = \frac{5 \times 4}{2 \times 1} = 10.$$

So there are 10 ways to get 2 Tiger pictures in 5 boxes, and the probability of each is $(0.2)^2(0.8)^3$. Now we can find what we wanted:

$$P(\#\text{success} = 2) = 10(0.2)^2(0.8)^3 = 0.2048$$

In general, the probability of exactly k successes in n trials is $\binom{n}{k}p^kq^{n-k}$.

Using this formula, we could find the expected value by adding up $xP(X = x)$ for all values, but it would be a long, hard way to get an answer that you already know intuitively. What's the expected value? If we have 5 boxes, and Tiger's picture is in 20% of them, then we would expect to have $5(0.2) = 1$ success. If we had 100 trials with probability of success 0.2, how many successes would you expect? Can you think of any reason not to say 20? It seems so simple that most people wouldn't even stop to think about it. You just multiply the probability of success by n. In other words, $E(X) = np$. Not fully convinced? We prove it in the next Math Box.

The standard deviation is less obvious; you can't just rely on your intuition. Fortunately, the formula for the standard deviation also boils down to something simple: $SD(X) = \sqrt{npq}$. (If you're curious about where that comes from, it's in the Math Box too!) In 100 boxes of cereal, we expect to find 20 Tiger Woods cards, with a standard deviation of $\sqrt{100 \times 0.8 \times 0.2} = 4$ pictures.

Time to summarize. A Binomial probability model describes the number of successes in a specified number of trials. It takes two parameters to specify this model: the number of trials n and the probability of success p.

TI-nspire™

Binomial probabilities.
Do-it-yourseif binomial models! Watch the probabilities change as you control *n* and *p*.

BINOMIAL PROBABILITY MODEL FOR BERNOULLI TRIALS: Binom(*n*, *p*)

n = number of trials
p = probability of success (and $q = 1 - p =$ probability of failure)
X = number of successes in n trials

$$P(X = x) = {}_nC_x\, p^x q^{n-x}, \textit{where } {}_nC_x = \frac{n!}{x!(n-x)!}$$

Mean: $\mu = np$
Standard Deviation: $\sigma = \sqrt{npq}$

MATH BOX

To derive the formulas for the mean and standard deviation of a Binomial model we start with the most basic situation.

Consider a single Bernoulli trial with probability of success p. Let's find the mean and variance of the number of successes.

Here's the probability model for the number of successes:

x	0	1
$P(X = x)$	q	p

Find the expected value:

$$\begin{aligned} E(X) &= 0q + 1p \\ E(X) &= p \end{aligned}$$

And now the variance:

$$\begin{aligned} Var(X) &= (0 - p)^2 q + (1 - p)^2 p \\ &= p^2 q + q^2 p \\ &= pq(p + q) \\ &= pq(1) \\ Var(X) &= pq \end{aligned}$$

What happens when there is more than one trial, though? A Binomial model simply counts the number of successes in a series of n independent Bernoulli trials. That makes it easy to find the mean and standard deviation of a binomial random variable, Y.

$$\begin{aligned} \text{Let } Y &= X_1 + X_2 + X_3 + \cdots + X_n \\ E(Y) &= E(X_1 + X_2 + X_3 + \cdots + X_n) \\ &= E(X_1) + E(X_2) + E(X_3) + \cdots + E(X_n) \\ &= p + p + p + \cdots + p \text{ (There are } n \text{ terms.)} \end{aligned}$$

So, as we thought, the mean is $E(Y) = np$.

And since the trials are independent, the variances add:

$$\begin{aligned} Var(Y) &= Var(X_1 + X_2 + X_3 + \cdots + X_n) \\ &= Var(X_1) + Var(X_2) + Var(X_3) + \cdots + Var(X_n) \\ &= pq + pq + pq + \cdots + pq \text{ (Again, } n \text{ terms.)} \\ Var(Y) &= npq \end{aligned}$$

Voilà! The standard deviation is $SD(Y) = \sqrt{npq}$.

FOR EXAMPLE Spam and the Binomial model

Recap: The communications monitoring company *Postini* has reported that 91% of e-mail messages are spam. Suppose your inbox contains 25 messages.

Questions: What are the mean and standard deviation of the number of real messages you should expect to find in your inbox? What's the probability that you'll find only 1 or 2 real messages?

I assume that messages arrive independently and at random, with the probability of success (a real message) $p = 1 - 0.91 = 0.09$. Let X = the number of real messages among 25. I can use the model Binom(25, 0.09).

$$\begin{aligned} E(X) &= np = 25(0.09) = 2.25 \\ SD(X) &= \sqrt{npq} = \sqrt{25(0.09)(0.91)} = 1.43 \\ P(X = 1 \text{ or } 2) &= P(X = 1) + P(X = 2) \\ &= \binom{25}{1}(0.09)^1(0.91)^{24} + \binom{25}{2}(0.09)^2(0.91)^{23} \\ &= 0.2340 + 0.2777 \\ &= 0.5117 \end{aligned}$$

Among 25 e-mail messages, I expect to find an average of 2.25 that aren't spam, with a standard deviation of 1.43 messages. There's just over a 50% chance that 1 or 2 of my 25 e-mails will be real messages.

STEP-BY-STEP EXAMPLE Working with a Binomial Model

Suppose 20 donors come to a blood drive. Recall that 6% of people are "universal donors."

Questions:

1. What are the mean and standard deviation of the number of universal donors among them?
2. What is the probability that there are 2 or 3 universal donors?

THINK

Plan State the question.

I want to know the mean and standard deviation of the number of universal donors among 20 people, and the probability that there are 2 or 3 of them.

Check to see that these are Bernoulli trials.

✓ There are two outcomes:

success = O-negative
failure = other blood types

✓ $p = 0.06$, because people have lined up at random.

✓ ***10% Condition:*** Trials are not independent, because the population is finite, but fewer than 10% of all possible donors are lined up.

Variable Define the random variable.

Let X = number of O-negative donors among $n = 20$ people.

Model Specify the model.

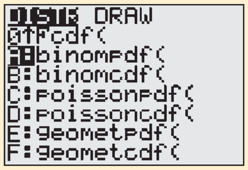
I can model X with Binom(20, 0.06).

Mechanics Find the expected value and standard deviation.	$E(X) = np = 20(0.06) = 1.2$ $SD(X) = \sqrt{npq} = \sqrt{20(0.06)(0.94)} \approx 1.06$ $P(X = 2 \text{ or } 3) = P(X = 2) + P(X = 3)$ $= \binom{20}{2}(0.06)^2(0.94)^{18}$ $+ \binom{20}{3}(0.06)^3(0.94)^{17}$ $\approx 0.2246 + 0.0860$ $= 0.3106$
Conclusion Interpret your results in context.	In groups of 20 randomly selected blood donors, I expect to find an average of 1.2 universal donors, with a standard deviation of 1.06. About 31% of the time, I'd find 2 or 3 universal donors among the 20 people.

TI Tips

Finding binomial probabilities

Remember how the calculator handles Geometric probabilities? Well, the commands for finding Binomial probabilities are essentially the same. Again you'll find them in the `2nd DISTR` menu.

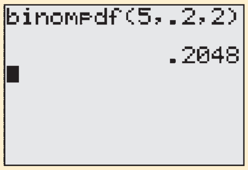


```
DISTR DRAW
0↑Fcdf(
A:binompdf(
B:binomcdf(
C:poissonpdf(
D:poissoncdf(
E:geometpdf(
F:geometcdf(
```

- `binompdf(`

  This probability density function allows you to find the probability of an *individual* outcome. You need to define the Binomial model by specifying n and p, and then indicate the desired number of successes, x. The format is `binompdf(n,p,X)`.

  For example, recall that Tiger Woods' picture is in 20% of the cereal boxes. Suppose that we want to know the probability of finding Tiger exactly twice among 5 boxes of cereal. We use $n = 5$, $p = 0.2$, and $x = 2$, entering the command `binompdf(5,.2,2)`. There's about a 20% chance of getting two pictures of Tiger Woods in five boxes of cereal.

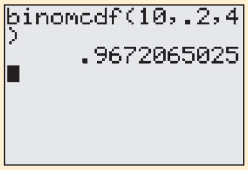


```
binompdf(5,.2,2)
               .2048
```

- `binomcdf(`

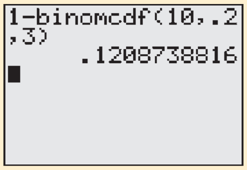
  Need to add several Binomial probabilities? To find the total probability of getting x or fewer successes among the n trials use the cumulative Binomial density function `binomcdf(n,p,X)`.

  For example, suppose we have ten boxes of cereal and wonder about the probability of finding up to 4 pictures of Tiger. That's the probability of 0, 1, 2, 3 or 4 successes, so we specify the command `binomcdf(10,.2,4)`. Pretty likely!



```
binomcdf(10,.2,4
)
         .9672065025
```

  Of course "up to 4" allows for the possibility that we end up with none. What's the probability we get at least 4 pictures of Tiger in 10 boxes? Well, "at least 4" means "not 3 or fewer." That's the complement of 0, 1, 2, or 3 successes. Have your TI evaluate `1-binomcdf(10,.2,3)`. There's about a 12% chance we'll find at least 4 pictures of Tiger in 10 boxes of cereal.



```
1-binomcdf(10,.2
,3)
         .1208738816
```

The Normal Model to the Rescue!

Suppose the Tennessee Red Cross anticipates the need for at least 1850 units of O-negative blood this year. It estimates that it will collect blood from 32,000 donors. How great is the risk that the Tennessee Red Cross will fall short of meeting its need? We've just learned how to calculate such probabilities. We can use the Binomial model with $n = 32{,}000$ and $p = 0.06$. The probability of getting *exactly* 1850 units of O-negative blood from 32,000 donors is $\binom{32000}{1850} \times 0.06^{1850} \times 0.94^{30150}$. No calculator on earth can calculate that first term (it has more than 100,000 digits).[1] And that's just the beginning. The problem said *at least* 1850, so we have to do it again for 1851, for 1852, and all the way up to 32,000. No thanks.

When we're dealing with a large number of trials like this, making direct calculations of the probabilities becomes tedious (or outright impossible). Here an old friend—the Normal model—comes to the rescue.

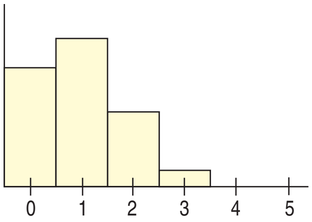
The Binomial model has mean $np = 1920$ and standard deviation $\sqrt{npq} \approx 42.48$. We could try approximating its distribution with a Normal model, using the same mean and standard deviation. Remarkably enough, that turns out to be a very good approximation. (We'll see why in the next chapter.) With that approximation, we can find the *probability:*

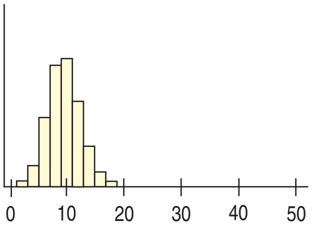
$$P(X < 1850) = P\left(z < \frac{1850 - 1920}{42.48}\right) \approx P(z < -1.65) \approx 0.05$$

There seems to be about a 5% chance that this Red Cross chapter will run short of O-negative blood.

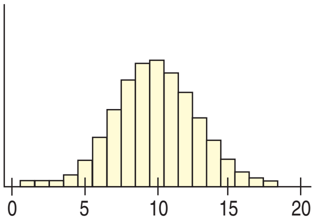
A S **Activity: Normal Approximation.** Binomial probabilities can be hard to calculate. With the Simulation Tool you'll see how well the Normal model can approximate the Binomial—a much easier method.

Can we always use a Normal model to make estimates of Binomial probabilities? No. Consider the Tiger Woods situation—pictures in 20% of the cereal boxes. If we buy five boxes, the actual Binomial probabilities that we get 0, 1, 2, 3, 4, or 5 pictures of Tiger are 33%, 41%, 20%, 5%, 1%, and 0.03%, respectively. The first histogram shows that this probability model is skewed. That makes it clear that we should not try to estimate these probabilities by using a Normal model.





Now suppose we open 50 boxes of this cereal and count the number of Tiger Woods pictures we find. The second histogram shows this probability model. It is centered at $np = 50(0.2) = 10$ pictures, as expected, and it appears to be fairly symmetric around that center. Let's have a closer look.



The third histogram again shows Binom(50, 0.2), this time magnified somewhat and centered at the expected value of 10 pictures of Tiger. It looks close to Normal, for sure. With this larger sample size, it appears that a Normal model might be a useful approximation.

A Normal model, then, is a close enough approximation only for a large enough number of trials. And what we mean by "large enough" depends on the probability of success. We'd need a larger sample if the probability of success were very low (or very high). It turns out that a Normal model works pretty well if we expect to see at least 10 successes and 10 failures. That is, we check the **Success/ Failure Condition.**

The Success/ Failure Condition: A Binomial model is approximately Normal if we expect at least 10 successes and 10 failures:

$$np \geq 10 \text{ and } nq \geq 10.$$

TI-*nspire*

How close to Normal? How well does a Normal curve fit a binomial model? Check out the Success/ Failure Condition for yourself.

[1] If your calculator *can* find Binom(32000,0.06), then it's smart enough to use an approximation. Read on to see how you can, too.

MATH BOX

It's easy to see where the magic number 10 comes from. You just need to remember how Normal models work. The problem is that a Normal model extends infinitely in both directions. But a Binomial model must have between 0 and n successes, so if we use a Normal to approximate a Binomial, we have to cut off its tails. That's not very important if the center of the Normal model is so far from 0 and n that the lost tails have only a negligible area. More than three standard deviations should do it, because a Normal model has little probability past that.

So the mean needs to be at least 3 standard deviations from 0 and at least 3 standard deviations from n. Let's look at the 0 end.

We require:	$\mu - 3\sigma > 0$
Or in other words:	$\mu > 3\sigma$
For a Binomial, that's:	$np > 3\sqrt{npq}$
Squaring yields:	$n^2p^2 > 9npq$
Now simplify:	$np > 9q$
Since $q \leq 1$, we can require:	$np > 9$

For simplicity, we usually require that np (and nq for the other tail) be at least 10 to use the Normal approximation, the Success/Failure Condition.[2]

FOR EXAMPLE Spam and the Normal approximation to the Binomial

Recap: The communications monitoring company *Postini* has reported that 91% of e-mail messages are spam. Recently, you installed a spam filter. You observe that over the past week it okayed only 151 of 1422 e-mails you received, classifying the rest as junk. Should you worry that the filtering is too aggressive?

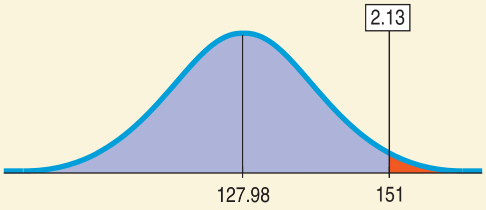
Question: What's the probability that no more than 151 of 1422 e-mails is a real message?

I assume that messages arrive randomly and independently, with a probability of success (a real message) $p = 0.09$. The model Binom(1422, 0.09) applies, but will be hard to work with. Checking conditions for the Normal approximation, I see that:

- ✓ These messages represent less than 10% of all e-mail traffic.
- ✓ I expect $np = (1422)(0.09) = 127.98$ real messages and $nq = (1422)(0.91) = 1294.02$ spam messages, both far greater than 10.

It's okay to approximate this binomial probability by using a Normal model.

$$\mu = np = 1422(0.09) = 127.98$$
$$\sigma = \sqrt{npq} = \sqrt{1422(0.09)(0.91)} \approx 10.79$$
$$P(x \leq 151) = P\left(z \leq \frac{151 - 127.98}{10.79}\right)$$
$$= P(z \leq 2.13)$$
$$= 0.9834$$



Among my 1422 e-mails, there's over a 98% chance that no more than 151 of them were real messages, so the filter may be working properly.

[2] Looking at the final step, we see that we need $np > 9$ in the worst case, when q (or p) is near 1, making the Binomial model quite skewed. When q and p are near 0.5—say between 0.4 and 0.6—the Binomial model is nearly symmetric and $np > 5$ ought to be safe enough. Although we'll always check for 10 expected successes and failures, keep in mind that for values of p near 0.5, we can be somewhat more forgiving.

Continuous Random Variables

There's a problem with approximating a Binomial model with a Normal model. The Binomial is discrete, giving probabilities for specific counts, but the Normal models a **continuous** random variable that can take on *any value*. For continuous random variables, we can no longer list all the possible outcomes and their probabilities, as we could for discrete random variables.[3]

As we saw in the previous chapter, models for continuous random variables give probabilities for *intervals* of values. So, when we use the Normal model, we no longer calculate the probability that the random variable equals a *particular* value, but only that it lies *between* two values. We won't calculate the probability of getting exactly 1850 units of blood, but we have no problem approximating the probability of getting 1850 *or more*, which was, after all, what we really wanted.[4]

JUST CHECKING

As we noted a few chapters ago, the Pew Research Center (www.pewresearch.org) reports that they are actually able to contact only 76% of the randomly selected households drawn for a telephone survey.

1. Explain why these phone calls can be considered Bernoulli trials.
2. Which of the models of this chapter (Geometric, Binomial, Normal) would you use to model the number of successful contacts from a list of 1000 sampled households? Explain.
3. Pew further reports that even after they contacted a household, only 38% agree to be interviewed, so the probability of getting a completed interview for a randomly selected household is only 0.29. Which of the models of this chapter would you use to model the number of households Pew has to call before they get the first completed interview?

WHAT CAN GO WRONG?

- **Be sure you have Bernoulli trials.** Be sure to check the requirements first: two possible outcomes per trial ("success" and "failure"), a constant probability of success, and independence. Remember to check the 10% Condition when sampling without replacement.
- **Don't confuse Geometric and Binomial models.** Both involve Bernoulli trials, but the issues are different. If you are repeating trials until your first success, that's a Geometric probability. You don't know in advance how many trials you'll need—theoretically, it could take forever. If you are counting the number of successes in a specified number of trials, that's a Binomial probability.
- **Don't use the Normal approximation with small *n*.** To use a Normal approximation in place of a Binomial model, there must be at least 10 expected successes and 10 expected failures.

[3] In fact, some people use an adjustment called the "continuity correction" to help with this problem. It's related to the suggestion we make in the next footnote and is discussed in more advanced textbooks.

[4] If we really had been interested in a single value, we might have approximated it by finding the probability of getting between 1849.5 and 1850.5 units of blood.

CONNECTIONS

This chapter builds on what we know about random variables. We now have two more probability models to join the Normal model.

There are a number of "forward" connections from this chapter. We'll see the **10% Condition** and the **Success/Failure Condition** often. And the facts about the Binomial distribution can help explain how proportions behave, as we'll see in the next chapter.

WHAT HAVE WE LEARNED?

We've learned that Bernoulli trials show up in lots of places. Depending on the random variable of interest, we can use one of three models to estimate probabilities for Bernoulli trials:

- a Geometric model when we're interested in the number of Bernoulli trials until the next success;
- a Binomial model when we're interested in the number of successes in a certain number of Bernoulli trials;
- a Normal model to approximate a Binomial model when we expect at least 10 successes and 10 failures.

Terms

Bernoulli trials, if . . .

388. 1. there are two possible outcomes.
     2. the probability of success is constant.
     3. the trials are independent.

Geometric probability model

389. A Geometric model is appropriate for a random variable that counts the number of Bernoulli trials until the first success.

Binomial probability model

393. A Binomial model is appropriate for a random variable that counts the number of successes in a fixed number of Bernoulli trials.

10% Condition

391. When sampling without replacement, trials are not independent. It's still okay to proceed as long as the sample is smaller than 10% of the population.

Success/Failure Condition

397. For a Normal model to be a good approximation of a Binomial model, we must expect at least 10 successes and 10 failures. That is, $np \geq 10$ and $nq \geq 10$.

Skills

THINK

- Know how to tell if a situation involves Bernoulli trials.
- Be able to choose whether to use a Geometric or a Binomial model for a random variable involving Bernoulli trials.

SHOW

- Know the appropriate conditions for using a Geometric, Binomial, or Normal model.
- Know how to find the expected value of a Geometric model.
- Be able to calculate Geometric probabilities.
- Know how to find the mean and standard deviation of a Binomial model.
- Be able to calculate Binomial probabilities, perhaps approximating with a Normal model.

TELL

- Be able to interpret means, standard deviations, and probabilities in the Bernoulli trial context.