

PART

From the Data at Hand to the World at Large

Chapter 18

Sampling Distribution Models

Chapter 19

Confidence Intervals for Proportions

Chapter 20

Testing Hypotheses About Proportions

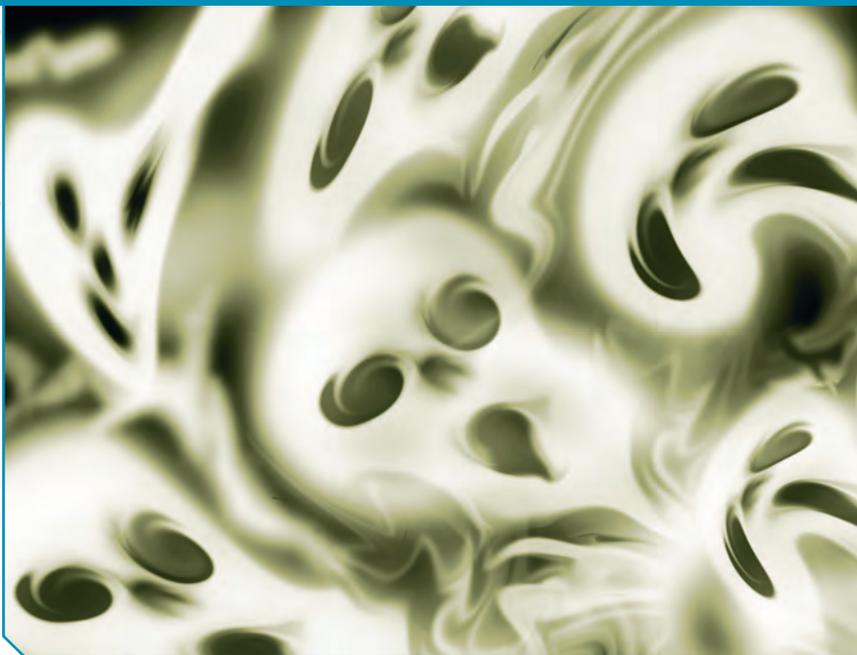
Chapter 21

More About Tests and Intervals

Chapter 22

Comparing Two Proportions

Sampling Distribution Models



WHO	U.S. adults
WHAT	Belief in ghosts
WHEN	November 2005
WHERE	United States
WHY	Public attitudes

In November 2005 the Harris Poll asked 889 U.S. adults, “Do you believe in ghosts?” 40% said they did. At almost the same time, CBS News polled 808 U.S. adults and asked the same question. 48% of their respondents professed a belief in ghosts. Why the difference? This seems like a simple enough question. Should we be surprised to find that we could get proportions this different from properly selected random samples drawn from the same population? You’re probably used to seeing that observations vary, but how much variability among polls should we expect to see?

Why do sample proportions vary at all? How can surveys conducted at essentially the same time by organizations asking the same questions get different results? The answer is at the heart of Statistics. The proportions vary from sample to sample because the samples are composed of different people.

It’s actually pretty easy to predict how much a proportion will vary under circumstances like this. Understanding the variability of our estimates will let us actually use that variability to better understand the world.

The Central Limit Theorem for Sample Proportions

Imagine

We see only the sample that we actually drew, but by simulating or modeling, we can *imagine* what we might have seen had we drawn other possible random samples.

We’ve talked about *Think*, *Show*, and *Tell*. Now we have to add *Imagine*. In order to understand the CBS poll, we want to imagine the results from all the random samples of size 808 that CBS News didn’t take. What would the histogram of all the sample proportions look like?

For people’s belief in ghosts, where do you expect the center of that histogram to be? Of course, we don’t *know* the answer to that (and probably never will). But we know that it will be at the true proportion in the population, and we can call that p . (See the Notation Alert.) For the sake of discussion here, let’s suppose that 45% of all American adults believe in ghosts, so we’ll use $p = 0.45$.

How about the *shape* of the histogram? We don’t have to just imagine. We can simulate a bunch of random samples that we didn’t really draw. Here’s a histogram of the proportions saying they believe in ghosts for 2000 simulated independent samples of 808 adults when the true proportion is $p = 0.45$.

A S **Activity: Sampling Distribution of a Proportion.** You don't have to imagine—you can simulate.

TI-*n*spire

Sample Proportions. Generate sample after sample to see how the proportions vary.

NOTATION ALERT:

The letter p is our choice for the *parameter* of the model for proportions. It violates our “Greek letters for parameters” rule, but if we stuck to that, our natural choice would be π . We could use π to be perfectly consistent, but then we'd have to write statements like $\pi = 0.46$. That just seems a bit weird to us. After all, we've known that $\pi = 3.1415926 \dots$ since the Greeks, and it's a hard habit to break.

So, we'll use p for the model parameter (the probability of a success) and \hat{p} for the observed proportion in a sample. We'll also use q for the probability of a failure ($q = 1 - p$) and \hat{q} for its observed value.

But be careful. We've already used capital P for a general probability. And we'll soon see another use of P in the next chapter! There are a lot of p 's in this course; you'll need to think clearly about the context to keep them straight.



Pierre-Simon Laplace, 1749–1827.

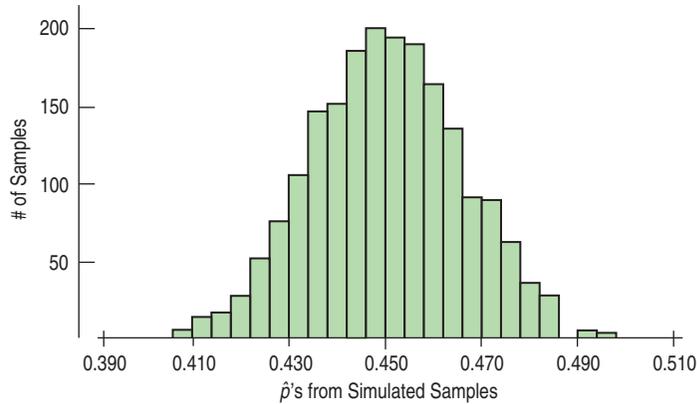


FIGURE 18.1

A histogram of sample proportions for 2000 simulated samples of 808 adults drawn from a population with $p = 0.45$. The sample proportions vary, but their distribution is centered at the true proportion, p .

It should be no surprise that we don't get the same proportion for each sample we draw, even though the underlying true value is the same for the population. Each \hat{p} comes from a different simulated sample. The histogram above is a simulation of what we'd get if we could see *all the proportions from all possible samples*. That distribution has a special name. It is called the **sampling distribution of the proportions**.¹

Does it surprise you that the histogram is unimodal? Symmetric? That it is centered at p ? You probably don't find any of this shocking. Does the shape remind you of any model that we've discussed? It's an amazing and fortunate fact that a Normal model is just the right one for the histogram of sample proportions.

As we'll see in a few pages, this fact was proved in 1810 by the great French mathematician Pierre-Simon Laplace as part of a more general result. There is no reason you should guess that the Normal model would be the one we need here,² and, indeed, the importance of Laplace's result was not immediately understood by his contemporaries. But (unlike Laplace's contemporaries in 1810) we know how useful the Normal model can be.

Modeling how sample proportions vary from sample to sample is one of the most powerful ideas we'll see in this course. A **sampling distribution model** for how a sample proportion varies from sample to sample allows us to quantify that variation and to talk about how likely it is that we'd observe a sample proportion in any particular interval.

To use a Normal model, we need to specify two parameters: its mean and standard deviation. The center of the histogram is naturally at p , so we'll put μ , the mean of the Normal, at p .

What about the standard deviation? Usually the mean gives us no information about the standard deviation. Suppose we told you that a batch of bike helmets had a mean diameter of 26 centimeters and asked what the standard deviation was. If you said, "I have no idea," you'd be exactly right. There's no information about σ from knowing the value of μ .

But there's a special fact about proportions. With proportions we get something for free. Once we know the mean, p , we automatically also know the standard deviation. We saw in the last chapter that for a Binomial model the standard deviation of the *number* of successes is \sqrt{npq} . Now we want the standard deviation

¹ A word of caution. Until now we've been plotting the *distribution of the sample*, a display of the actual data that were collected in that one sample. But now we've plotted the *sampling distribution*; a display of summary statistics (\hat{p} 's, for example) for many different samples. "Sample distribution" and "sampling distribution" sound a lot alike, but they refer to very different things. (Sorry about that—we didn't make up the terms. It's just the way it is.) And the distinction is critical. Whenever you read or write something about one of these, think very carefully about what the words signify.

² Well, the fact that we spent most of Chapter 6 on the Normal model might have been a hint.

of the *proportion* of successes, \hat{p} . The sample proportion \hat{p} is the number of successes divided by the number of trials, n , so the standard deviation is also divided by n :

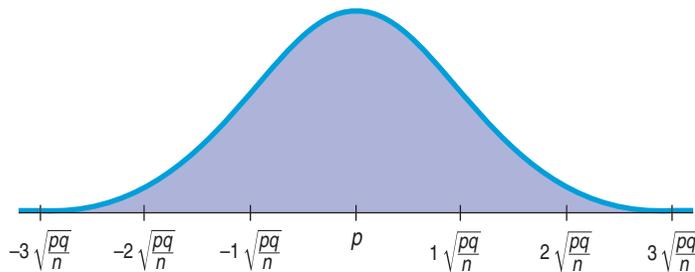
$$\sigma(\hat{p}) = SD(\hat{p}) = \frac{\sqrt{npq}}{n} = \sqrt{\frac{pq}{n}}$$

When we draw simple random samples of n individuals, the proportions we find will vary from sample to sample. As long as n is reasonably large,³ we can model the distribution of these sample proportions with a probability model that is

$$N\left(p, \sqrt{\frac{pq}{n}}\right).$$

AS **Simulation: Simulating Sampling Distributions.** Watch the Normal model appear from random proportions.

FIGURE 18.2
A Normal model centered at p with a standard deviation of $\sqrt{\frac{pq}{n}}$ is a good model for a collection of proportions found for many random samples of size n from a population with success probability p .



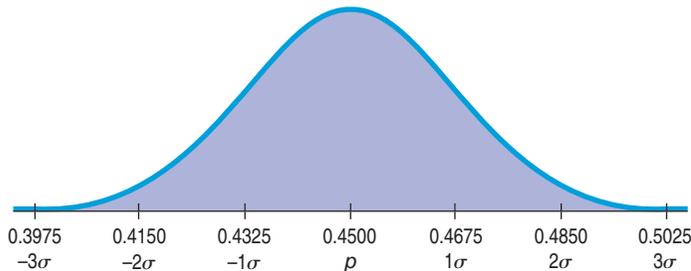
NOTATION ALERT:
In Chapter 8 we introduced \hat{y} as the predicted value for y . The “hat” here plays a similar role. It indicates that \hat{p} —the observed proportion in our data—is our *estimate* of the parameter p .

Although we’ll never know the true proportion of adults who believe in ghosts, we’re supposing it to be 45%. Once we put the center at $p = 0.45$, the standard deviation for the CBS poll is

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.45)(0.55)}{808}} = 0.0175, \text{ or } 1.75\%.$$

Here’s a picture of the Normal model for our simulation histogram:

FIGURE 18.3
Using 0.45 for p gives this Normal model for Figure 18.1’s histogram of the sample proportions of adults believing in ghosts ($n = 808$).



AS **Simulation: The Standard Deviation of a Proportion.** Do you believe this formula for standard deviation? Don’t just take our word for it—convince yourself with an experiment.

Because we have a Normal model, we can use the 68–95–99.7 Rule or look up other probabilities using a table or technology. For example, we know that 95% of Normally distributed values are within two standard deviations of the mean, so we should not be surprised if 95% of various polls gave results that were near 45% but varied above and below that by no more than two standard deviations. Since $2 \times 1.75\% = 3.5\%$,⁴ we see that the CBS poll estimating belief in ghosts at 48% is *consistent* with our guess of 45%. This is what we mean by **sampling error**. It’s not really an *error* at all, but just *variability* you’d expect to see from one sample to another. A better term would be **sampling variability**.

³ For smaller n , we can just use a Binomial model.
⁴ The standard deviation is 1.75%. Remember that the standard deviation always has the same units as the data. Here our units are %. But that can be confusing, because the standard deviation is not 1.75% of anything. It is 1.75 percentage points. If that’s confusing, try writing the units as “percentage points” instead of %.

How Good Is the Normal Model?

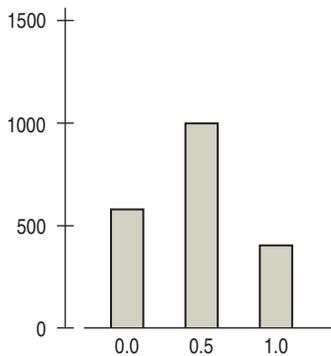


FIGURE 18.4

Proportions from samples of size 2 can take on only three possible values. A Normal model does not work well.

Stop and think for a minute about what we've just said. It's a remarkable claim. We've said that if we draw repeated random samples of the same size, n , from some population and measure the proportion, \hat{p} , we see in each sample, then the collection of these proportions will pile up around the underlying population proportion, p , and that a histogram of the sample proportions can be modeled well by a Normal model.

There must be a catch. Suppose the samples were of size 2, for example. Then the only possible proportion values would be 0, 0.5, and 1. There's no way the histogram could ever look like a Normal model with only three possible values for the variable.

Well, there *is* a catch. The claim is only approximately true. (But, that's OK. After all, models are only supposed to be approximately true.) And the model becomes a better and better representation of the distribution of the sample proportions as the sample size gets bigger.⁵ Samples of size 1 or 2 just aren't going to work very well. But the distributions of proportions of many larger samples do have histograms that are remarkably close to a Normal model.

Assumptions and Conditions

To use a model, we usually must make some assumptions. To use the sampling distribution model for sample proportions, we need two assumptions:

The Independence Assumption: The sampled values must be independent of each other.

The Sample Size Assumption: The sample size, n , must be large enough.

Of course, assumptions are hard—often impossible—to check. That's why we *assume* them. But, as we saw in Chapter 8, we should check to see whether the assumptions are reasonable. To think about the Independence Assumption, we often wonder whether there is any reason to think that the data values might affect each other. Fortunately, we can often check *conditions* that provide information about the assumptions. Check these conditions before using the Normal to model the distribution of sample proportions:

Randomization Condition: If your data come from an experiment, subjects should have been randomly assigned to treatments. If you have a survey, your sample should be a simple random sample of the population. If some other sampling design was used, be sure the sampling method was not biased and that the data are representative of the population.

10% Condition: The sample size, n , must be no larger than 10% of the population. For national polls, the total population is usually very large, so the sample is a small fraction of the population.

Success/Failure Condition: The sample size has to be big enough so that we expect at least 10 successes and at least 10 failures. When np and nq are at least 10, we have enough data for sound conclusions. For the CBS survey, a "success" might be believing in ghosts. With $p = 0.45$, we expect $808 \times 0.45 = 364$ successes and $808 \times 0.55 = 444$ failures. Both are at least 10, so we certainly expect enough successes and enough failures for the condition to be satisfied.

The terms "success" and "failure" for the outcomes that have probability p and q are common in Statistics. But they are completely arbitrary labels. When we say that a disease occurs with probability p , we certainly don't mean that getting sick is a "success" in the ordinary sense of the word.

⁵ Formally, we say the claim is true in the limit as n grows.

These last two conditions seem to conflict with each other. The **Success/Failure Condition** wants sufficient data. How much depends on p . If p is near 0.5, we need a sample of only 20 or so. If p is only 0.01, however, we'd need 1000. But the **10% Condition** says that a sample should be no larger than 10% of the population. If you're thinking, "Wouldn't a larger sample be better?" you're right of course. It's just that if the sample were more than 10% of the population, we'd need to use different methods to analyze the data. Fortunately, this isn't usually a problem in practice. Often, as in polls that sample from all U.S. adults or industrial samples from a day's production, the populations are much larger than 10 times the sample size.

A Sampling Distribution Model for a Proportion

We've simulated repeated samples and looked at a histogram of the sample proportions. We modeled that histogram with a Normal model. Why do we bother to model it? Because this model will give us insight into how much the sample proportion can vary from sample to sample. We've simulated many of the other random samples we might have gotten. The model is an attempt to show the distribution from *all* the random samples. But how do we know that a Normal model will really work? Is this just an observation based on some simulations that *might* be approximately true some of the time?

It turns out that this model can be justified theoretically and that the larger the sample size, the better the model works. That's the result Laplace proved. We won't bother you with the math because, in this instance, it really wouldn't help your understanding.⁶ Nevertheless, the fact that we can think of the sample proportion as a random variable taking on a different value in each random sample, and then say something this specific about the distribution of those values, is a fundamental insight—one that we will use in each of the next four chapters.

We have changed our point of view in a very important way. No longer is a proportion something we just compute for a set of data. We now see it as a random variable quantity that has a probability distribution, and thanks to Laplace we have a model for that distribution. We call that the **sampling distribution model** for the proportion, and we'll make good use of it.

A S *Simulation: Simulate the Sampling Distribution Model of a Proportion.* You probably don't want to work through the formal mathematical proof; a simulation is far more convincing!

We have now answered the question raised at the start of the chapter. To know how variable a sample proportion is, we need to know the proportion and the size of the sample. That's all.

THE SAMPLING DISTRIBUTION MODEL FOR A PROPORTION

Provided that the sampled values are independent and the sample size is large enough, the sampling distribution of \hat{p} is modeled by a Normal model

with mean $\mu(\hat{p}) = p$ and standard deviation $SD(\hat{p}) = \sqrt{\frac{pq}{n}}$.

Without the sampling distribution model, the rest of Statistics just wouldn't exist. Sampling models are what makes Statistics work. They inform us about the amount of variation we should expect when we sample. Suppose we spin a coin 100 times in order to decide whether it's fair or not. If we get 52 heads, we're probably not surprised. Although we'd expect 50 heads, 52 doesn't seem particularly unusual for a fair coin. But we would be surprised to see 90 heads; that might really make us doubt that the coin is fair. How about 64 heads? Harder to say. That's a case where we need the sampling distribution model. The sampling model quantifies the variability, telling us how surprising any sample proportion is. And

⁶ The proof is pretty technical. We're not sure it helps *our* understanding all that much either.

it enables us to make informed decisions about how precise our estimate of the true proportion might be. That's exactly what we'll be doing for the rest of this book.

Sampling distribution models act as a bridge from the real world of data to the imaginary model of the statistic and enable us to say something about the population when all we have is data from the real world. This is the huge leap of Statistics. Rather than thinking about the sample proportion as a fixed quantity calculated from our data, we now think of it as a random variable—our value is just one of many we might have seen had we chosen a different random sample. By imagining what *might* happen if we were to draw many, many samples from the same population, we can learn a lot about how close the statistics computed from our one particular sample may be to the corresponding population parameters they estimate. That's the path to the *margin of error* you hear about in polls and surveys. We'll see how to determine that in the next chapter.

FOR EXAMPLE

Using the sampling distribution model for proportions

The Centers for Disease Control and Prevention report that 22% of 18-year-old women in the United States have a body mass index (BMI)⁷ of 25 or more—a value considered by the National Heart Lung and Blood Institute to be associated with increased health risk.

As part of a routine health check at a large college, the physical education department usually requires students to come in to be measured and weighed. This year, the department decided to try out a self-report system. It asked 200 randomly selected female students to report their heights and weights (from which their BMIs could be calculated). Only 31 of these students had BMIs greater than 25.

Question: Is this proportion of high-BMI students unusually small?

First, check the conditions:

- ✓ **Randomization Condition:** The department drew a random sample, so the respondents should be independent and randomly selected from the population.
- ✓ **10% Condition:** 200 respondents is less than 10% of all the female students at a “large college.”
- ✓ **Success/Failure Condition:** The department expected $np = 200(0.22) = 44$ “successes” and $nq = 200(0.78) = 156$ “failures,” both at least 10.

It's okay to use a Normal model to describe the sampling distribution of the proportion of respondents with BMIs above 25.

The phys ed department observed $\hat{p} = \frac{31}{200} = 0.155$.

The department expected $E(\hat{p}) = p = 0.22$, with $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.22)(0.78)}{200}} = 0.029$,

so $z = \frac{\hat{p} - p}{SD(\hat{p})} = \frac{0.155 - 0.22}{0.029} = -2.24$.

By the 68–95–99.7 Rule, I know that values more than 2 standard deviations below the mean of a Normal model show up less than 2.5% of the time. Perhaps women at this college differ from the general population, or self-reporting may not provide accurate heights and weights.

⁷ BMI = weight in kg / (height in m)².



JUST CHECKING

1. You want to poll a random sample of 100 students on campus to see if they are in favor of the proposed location for the new student center. Of course, you'll get just one number, your sample proportion, \hat{p} . But if you imagined all the possible samples of 100 students you could draw and imagined the histogram of all the sample proportions from these samples, what shape would it have?
2. Where would the center of that histogram be?
3. If you think that about half the students are in favor of the plan, what would the standard deviation of the sample proportions be?

STEP-BY-STEP EXAMPLE

Working with Sampling Distribution Models for Proportions

Suppose that about 13% of the population is left-handed.⁸ A 200-seat school auditorium has been built with 15 “lefty seats,” seats that have the built-in desk on the left rather than the right arm of the chair. (For the right-handed readers among you, have you ever tried to take notes in a chair with the desk on the left side?)

Question: In a class of 90 students, what's the probability that there will not be enough seats for the left-handed students?



Plan State what we want to know.

Model Think about the assumptions and check the conditions.

You might be able to think of cases where the **Independence Assumption** is not plausible—for example, if the students are all related, or if they were selected for being left- or right-handed. But for a random sample, the assumption of independence seems reasonable.

I want to find the probability that in a group of 90 students, more than 15 will be left-handed. Since 15 out of 90 is 16.7%, I need the probability of finding more than 16.7% left-handed students out of a sample of 90 if the proportion of lefties is 13%.

- ✓ **Independence Assumption:** It is reasonable to assume that the probability that one student is left-handed is not changed by the fact that another student is right- or left-handed.
- ✓ **Randomization Condition:** The 90 students in the class can be thought of as a random sample of students.
- ✓ **10% Condition:** 90 is surely less than 10% of the population of all students. (Even if the school itself is small, I'm thinking of the population of all possible students who could have gone to the school.)
- ✓ **Success/Failure Condition:**

$$np = 90(0.13) = 11.7 \geq 10$$

$$nq = 90(0.87) = 78.3 \geq 10$$

⁸ Actually, it's quite difficult to get an accurate estimate of the proportion of lefties in the population. Estimates range from 8% to 15%.

State the parameters and the sampling distribution model.

The population proportion is $p = 0.13$. The conditions are satisfied, so I'll model the sampling distribution of \hat{p} with a Normal model with mean 0.13 and a standard deviation of

$$SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.13)(0.87)}{90}} \approx 0.035$$

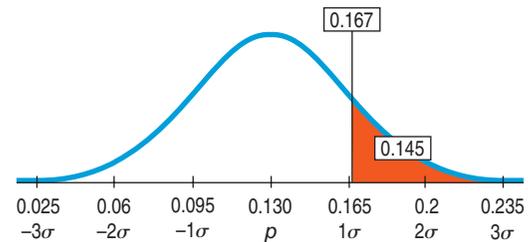
My model for \hat{p} is $N(0.13, 0.035)$.

SHOW

Plot Make a picture. Sketch the model and shade the area we're interested in, in this case the area to the right of 16.7%.

Mechanics Use the standard deviation as a ruler to find the z-score of the cutoff proportion. We see that 16.7% lefties would be just over one standard deviation above the mean.

Find the resulting probability from a table of Normal probabilities, a computer program, or a calculator.



$$z = \frac{\hat{p} - p}{SD(\hat{p})} = \frac{0.167 - 0.13}{0.035} = 1.06$$

$$P(\hat{p} > 0.167) = P(z > 1.06) = 0.1446$$

TELL

Conclusion Interpret the probability in the context of the question.

There is about a 14.5% chance that there will not be enough seats for the left-handed students in the class.

What About Quantitative Data?

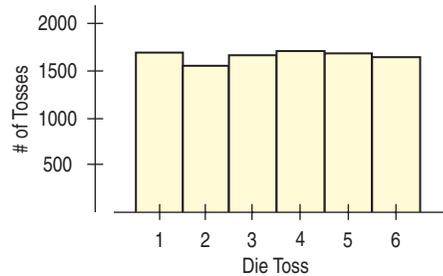
Proportions summarize categorical variables. And the Normal sampling distribution model looks like it is going to be very useful. But can we do something similar with quantitative data?

Of course we can (or we wouldn't have asked). Even more remarkable, not only can we use all of the same concepts, but almost the same model, too.

What are the concepts? We know that when we sample at random or randomize an experiment, the results we get will vary from sample-to-sample and from experiment-to-experiment. The Normal model seems an incredibly simple way to summarize all that variation. Could something that simple work for means? We won't keep you in suspense. It turns out that means also have a sampling distribution that we can model with a Normal model. And it turns out that Laplace's theoretical result applies to means, too. As we did with proportions, we can get some insight from a simulation.

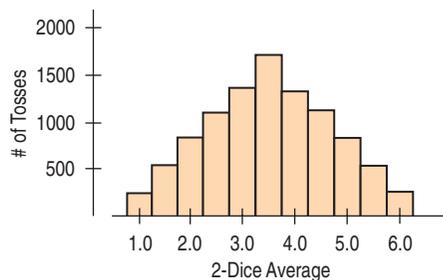
Simulating the Sampling Distribution of a Mean

Here's a simple simulation. Let's start with one fair die. If we toss this die 10,000 times, what should the histogram of the numbers on the face of the die look like? Here are the results of a simulated 10,000 tosses:



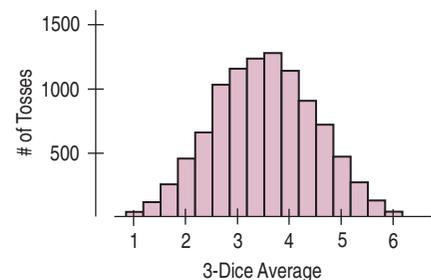
Now let's toss a *pair* of dice and record the average of the two. If we repeat this (or at least simulate repeating it) 10,000 times, recording the average of each pair, what will the histogram of these 10,000 averages look like? Before you look, think a minute. Is getting an average of 1 on *two* dice as likely as getting an average of 3 or 3.5?

Let's see:



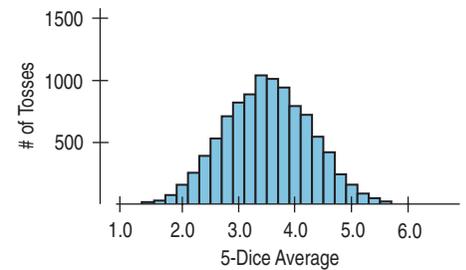
We're much more likely to get an average near 3.5 than we are to get one near 1 or 6. Without calculating those probabilities exactly, it's fairly easy to see that the *only* way to get an average of 1 is to get two 1's. To get a total of 7 (for an average of 3.5), though, there are many more possibilities. This distribution even has a name: the *triangular* distribution.

What if we average 3 dice? We'll simulate 10,000 tosses of 3 dice and take their average:



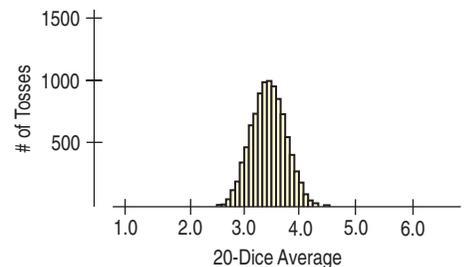
What's happening? First notice that it's getting harder to have averages near the ends. Getting an average of 1 or 6 with 3 dice requires all three to come up 1 or 6, respectively. That's less likely than for 2 dice to come up both 1 or both 6. The distribution is being pushed toward the middle. But what's happening to the shape? (This distribution doesn't have a name, as far as we know.)

Let's continue this simulation to see what happens with larger samples. Here's a histogram of the averages for 10,000 tosses of 5 dice:



The pattern is becoming clearer. Two things continue to happen. The first fact we knew already from the Law of Large Numbers. It says that as the sample size (number of dice) gets larger, each sample average is more likely to be closer to the population mean. So, we see the shape continuing to tighten around 3.5. But the shape of the distribution is the surprising part. It's becoming bell-shaped. And not just bell-shaped; it's approaching the Normal model.

Are you convinced? Let's skip ahead and try 20 dice. The histogram of averages for 10,000 throws of 20 dice looks like this:



Now we see the Normal shape again (and notice how much smaller the spread is). But can we count on this happening for situations other than dice throws? What kinds of sample means have sampling distributions that we can model with a Normal model? It turns out that Normal models work well amazingly often.

A S **Activity: The Sampling Distribution Model for Means.** Don't just sit there reading about the simulation—do it yourself.

The Fundamental Theorem of Statistics

The dice simulation may look like a special situation, but it turns out that what we saw with dice is true for means of repeated samples for almost every situation. When we looked at the sampling distribution of a proportion, we had to check only a few conditions. For means, the result is even more remarkable. *There are almost no conditions at all.*

Let's say that again: The sampling distribution of *any* mean becomes more nearly Normal as the sample size grows. All we need is for the observations to be independent and collected with randomization. We don't even care about the shape of the population distribution!⁹ This surprising fact is the result Laplace proved in a fairly general form in 1810. At the time, Laplace's theorem caused quite a stir (at least in mathematics circles) because it is so unintuitive. Laplace's result is called the **Central Limit Theorem**¹⁰ (CLT).

"The theory of probabilities is at bottom nothing but common sense reduced to calculus."

—Laplace, in *Théorie analytique des probabilités*, 1812

⁹ OK, one technical condition. The data must come from a population with a finite variance. You probably can't imagine a population with an infinite variance, but statisticians can construct such things, so we have to discuss them in footnotes like this. It really makes no difference in how you think about the important stuff, so you can just forget we mentioned it.

¹⁰ The word "central" in the name of the theorem means "fundamental." It doesn't refer to the center of a distribution.

Laplace was one of the greatest scientists and mathematicians of his time. In addition to his contributions to probability and statistics, he published many new results in mathematics, physics, and astronomy (where his nebular theory was one of the first to describe the formation of the solar system in much the way it is understood today). He also played a leading role in establishing the metric system of measurement.

His brilliance, though, sometimes got him into trouble. A visitor to the Académie des Sciences in Paris reported that Laplace let it be widely known that he considered himself the best mathematician in France. The effect of this on his colleagues was not eased by the fact that Laplace was right.

TI-*n*spire

The Central Limit Theorem. See the sampling distribution of sample means take shape as you choose sample after sample.

Why should the Normal model show up again for the sampling distribution of means as well as proportions? We're not going to try to persuade you that it is obvious, clear, simple, or straightforward. In fact, the CLT is surprising and a bit weird. Not only does the distribution of means of many random samples get closer and closer to a Normal model as the sample size grows, *this is true regardless of the shape of the population distribution!* Even if we sample from a skewed or bimodal population, the Central Limit Theorem tells us that means of repeated random samples will tend to follow a Normal model as the sample size grows. Of course, you won't be surprised to learn that it works better and faster the closer the population distribution is to a Normal model. And it works better for larger samples. If the data come from a population that's exactly Normal to start with, then the observations themselves are Normal. If we take samples of size 1, their "means" are just the observations—so, of course, they have Normal sampling distribution. But now suppose the population distribution is very skewed (like the CEO data from Chapter 5, for example). The CLT works, although it may take a sample size of dozens or even hundreds of observations for the Normal model to work well.

For example, think about a really bimodal population, one that consists of only 0's and 1's. The CLT says that even means of samples from this population will follow a Normal sampling distribution model. But wait. Suppose we have a categorical variable and we assign a 1 to each individual in the category and a 0 to each individual not in the category. And then we find the mean of these 0's and 1's. That's the same as counting the number of individuals who are in the category and dividing by n . That mean will be . . . the *sample proportion*, \hat{p} , of individuals who are in the category (a "success"). So maybe it wasn't so surprising after all that proportions, like means, have Normal sampling distribution models; they are actually just a special case of Laplace's remarkable theorem. Of course, for such an extremely bimodal population, we'll need a reasonably large sample size—and that's where the special conditions for proportions come in.

THE CENTRAL LIMIT THEOREM (CLT)

The mean of a random sample is a random variable whose sampling distribution can be approximated by a Normal model. The larger the sample, the better the approximation will be.

Assumptions and Conditions

AS **Activity: The Central Limit Theorem.** Does it really work for samples from non-Normal populations?

The CLT requires essentially the same assumptions as we saw for modelling proportions:

Independence Assumption: The sampled values must be independent of each other.

Sample Size Assumption: The sample size must be sufficiently large.

We can't check these directly, but we can think about whether the **Independence Assumption** is plausible. We can also check some related conditions:

Randomization Condition: The data values must be sampled randomly, or the concept of a sampling distribution makes no sense.

10% Condition: When the sample is drawn without replacement (as is usually the case), the sample size, n , should be no more than 10% of the population.

Large Enough Sample Condition: Although the CLT tells us that a Normal model is useful in thinking about the behavior of sample means when the

sample size is large enough, it doesn't tell us how large a sample we need. The truth is, it depends; there's no one-size-fits-all rule. If the population is unimodal and symmetric, even a fairly small sample is okay. If the population is strongly skewed, like the compensation for CEOs we looked at in Chapter 5, it can take a pretty large sample to allow use of a Normal model to describe the distribution of sample means. For now you'll just need to think about your sample size in the context of what you know about the population, and then tell whether you believe the **Large Enough Sample Condition** has been met.

But Which Normal?

A S

Activity: The Standard Deviation of Means. Experiment to see how the variability of the mean changes with the sample size.

The CLT says that the sampling distribution of any mean or proportion is approximately Normal. But which Normal model? We know that any Normal is specified by its mean and standard deviation. For proportions, the sampling distribution is centered at the population proportion. For means, it's centered at the population mean. What else would we expect?

What about the standard deviations, though? We noticed in our dice simulation that the histograms got narrower as we averaged more and more dice together. This shouldn't be surprising. Means vary less than the individual observations. Think about it for a minute. Which would be more surprising, having *one* person in your Statistics class who is over 6'9" tall or having the *mean* of 100 students taking the course be over 6'9"? The first event is fairly rare.¹¹ You may have seen somebody this tall in one of your classes sometime. But finding a class of 100 whose mean height is over 6'9" tall just won't happen. Why? Because *means have smaller standard deviations than individuals*.

How much smaller? Well, we have good news and bad news. The good news is that the standard deviation of \bar{y} falls as the sample size grows. The bad news is that it doesn't drop as fast as we might like. It only goes down by the *square root* of the sample size. Why? The Math Box will show you that the Normal model for the sampling distribution of the mean has a standard deviation equal to

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population. To emphasize that this is a standard deviation *parameter* of the sampling distribution model for the sample mean, \bar{y} , we write $SD(\bar{y})$ or $\sigma(\bar{y})$.

A S

Activity: The Sampling Distribution of the Mean. The CLT tells us what to expect. In this activity you can work with the CLT or simulate it if you prefer.

THE SAMPLING DISTRIBUTION MODEL FOR A MEAN (CLT)

When a random sample is drawn from any population with mean μ and standard deviation σ , its sample mean, \bar{y} , has a sampling distribution

with the same *mean* μ but whose *standard deviation* is $\frac{\sigma}{\sqrt{n}}$ (and we write

$\sigma(\bar{y}) = SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$). No matter what population the random sample comes

from, the *shape* of the sampling distribution is approximately Normal as long as the sample size is large enough. The larger the sample used, the more closely the Normal approximates the sampling distribution for the mean.

¹¹ If students are a random sample of adults, fewer than 1 out of 10,000 should be taller than 6'9". Why might college students not really be a random sample with respect to height? Even if they're not a perfectly random sample, a college student over 6'9" tall is still rare.

MATH BOX

We know that \bar{y} is a sum divided by n :

$$\bar{y} = \frac{y_1 + y_2 + y_3 + \cdots + y_n}{n}.$$

As we saw in Chapter 16, when a random variable is divided by a constant its variance is divided by the *square* of the constant:

$$\text{Var}(\bar{y}) = \frac{\text{Var}(y_1 + y_2 + y_3 + \cdots + y_n)}{n^2}.$$

To get our sample, we draw the y 's randomly, ensuring they are independent. For independent random variables, variances add:

$$\text{Var}(\bar{y}) = \frac{\text{Var}(y_1) + \text{Var}(y_2) + \text{Var}(y_3) + \cdots + \text{Var}(y_n)}{n^2}.$$

All n of the y 's were drawn from our population, so they all have the same variance, σ^2 :

$$\text{Var}(\bar{y}) = \frac{\sigma^2 + \sigma^2 + \sigma^2 + \cdots + \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

The standard deviation of \bar{y} is the square root of this variance:

$$SD(\bar{y}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}.$$

We now have two closely related sampling distribution models that we can use when the appropriate assumptions and conditions are met. Which one we use depends on which kind of data we have:

- ▶ When we have categorical data, we calculate a sample proportion, \hat{p} ; the sampling distribution of this random variable has a Normal model with a mean at the true proportion (“Greek letter”) p and a standard deviation of $SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \frac{\sqrt{pq}}{\sqrt{n}}$. We’ll use this model in Chapters 19 through 22.
- ▶ When we have quantitative data, we calculate a sample mean, \bar{y} ; the sampling distribution of this random variable has a Normal model with a mean at the true mean, μ , and a standard deviation of $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$. We’ll use this model in Chapters 23, 24, and 25.

The means of these models are easy to remember, so all you need to be careful about is the standard deviations. Remember that these are standard deviations of the *statistics* \hat{p} and \bar{y} . They both have a square root of n in the denominator. That tells us that the larger the sample, the less either statistic will vary. The only difference is in the numerator. If you just start by writing $SD(\bar{y})$ for quantitative data and $SD(\hat{p})$ for categorical data, you’ll be able to remember which formula to use.

FOR EXAMPLE

Using the CLT for means

Recap: A college physical education department asked a random sample of 200 female students to self-report their heights and weights, but the percentage of students with body mass indexes over 25 seemed suspiciously low. One possible explanation may be that the respondents “shaded” their weights down a bit. The CDC reports that the mean weight of 18-year-old women is 143.74 lb, with a standard deviation of 51.54 lb, but these 200 randomly selected women reported a mean weight of only 140 lb.

Question: Based on the Central Limit Theorem and the 68–95–99.7 Rule, does the mean weight in this sample seem exceptionally low, or might this just be random sample-to-sample variation?

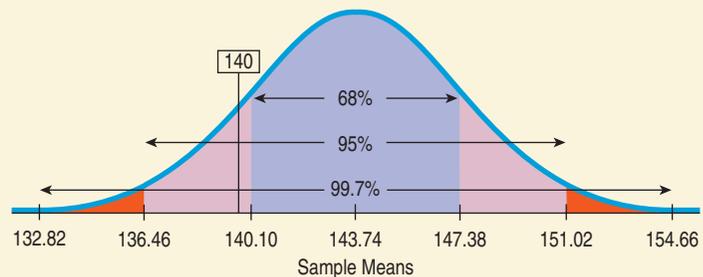
The conditions check out okay:

- ✓ **Randomization Condition:** The women were a random sample and their weights can be assumed to be independent.
- ✓ **10% Condition:** They sampled fewer than 10% of all women at the college.
- ✓ **Large Enough Sample Condition:** The distribution of college women’s weights is likely to be unimodal and reasonably symmetric, so the CLT applies to means of even small samples; 200 values is plenty.

The sampling model for sample means is approximately Normal with $E(\bar{y}) = 143.7$ and

$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{51.54}{\sqrt{200}} = 3.64. \text{ The expected}$$

distribution of sample means is:



The 68–95–99.7 Rule suggests that although the reported mean weight of 140 pounds is somewhat lower than expected, it does not appear to be unusual. Such variability is not all that extraordinary for samples of this size.

STEP-BY-STEP EXAMPLE

Working with the Sampling Distribution Model for the Mean

The Centers for Disease Control and Prevention reports that the mean weight of adult men in the United States is 190 lb with a standard deviation of 59 lb.¹²

Question: An elevator in our building has a weight limit of 10 persons or 2500 lb. What’s the probability that if 10 men get on the elevator, they will overload its weight limit?

THINK

Plan State what we want to know.

Asking the probability that the total weight of a sample of 10 men exceeds 2500 pounds is equivalent to asking the probability that their mean weight is greater than 250 pounds.

¹² Cynthia L. Ogden, Cheryl D. Fryar, Margaret D. Carroll, and Katherine M. Flegal, *Mean Body Weight, Height, and Body Mass Index, United States 1960–2002, Advance Data from Vital and Health Statistics Number 347*, Oct. 27, 2004. <https://www.cdc.gov/nchs>

Model Think about the assumptions and check the conditions.

Note that if the sample were larger we'd be less concerned about the shape of the distribution of all weights.

State the parameters and the sampling model.

- ✓ **Independence Assumption:** It's reasonable to think that the weights of 10 randomly sampled men will be independent of each other. (But there could be exceptions—for example, if they were all from the same family or if the elevator were in a building with a diet clinic!)
- ✓ **Randomization Condition:** I'll assume that the 10 men getting on the elevator are a random sample from the population.
- ✓ **10% Condition:** 10 men is surely less than 10% of the population of possible elevator riders.
- ✓ **Large Enough Sample Condition:** I suspect the distribution of population weights is roughly unimodal and symmetric, so my sample of 10 men seems large enough.

The mean for all weights is $\mu = 190$ and the standard deviation is $\sigma = 59$ pounds. Since the conditions are satisfied, the CLT says that the sampling distribution of \bar{y} has a Normal model with mean 190 and standard deviation

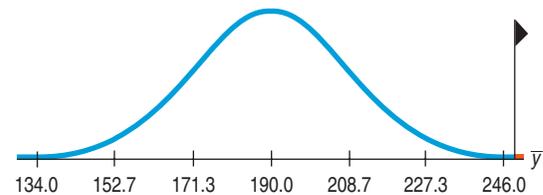
$$SD(\bar{y}) = \frac{\sigma}{\sqrt{n}} = \frac{59}{\sqrt{10}} \approx 18.66$$



Plot Make a picture. Sketch the model and shade the area we're interested in. Here the mean weight of 250 pounds appears to be far out on the right tail of the curve.

Mechanics Use the standard deviation as a ruler to find the z-score of the cutoff mean weight. We see that an average of 250 pounds is more than 3 standard deviations above the mean.

Find the resulting probability from a table of Normal probabilities such as Table Z, a computer program, or a calculator.



$$z = \frac{\bar{y} - \mu}{SD(\bar{y})} = \frac{250 - 190}{18.66} = 3.21$$

$$P(\bar{y} > 250) = P(z > 3.21) = 0.0007$$



Conclusion Interpret your result in the proper context, being careful to relate it to the original question.

The chance that a random collection of 10 men will exceed the elevator's weight limit is only 0.0007. So, if they are a random sample, it is quite unlikely that 10 people will exceed the total weight allowed on the elevator.

About Variation

“The n 's justify the means.”

—Apocryphal
statistical saying

Means vary less than individual data values. That makes sense. If the same test is given to many sections of a large course and the class average is, say, 80%, some students may score 95% because individual scores vary a lot. But we'd be shocked (and pleased!) if the *average* score of the students in any section was 95%. Averages are much less variable. Not only do group averages vary less than individual values, but common sense suggests that averages should be more consistent for larger groups. The Central Limit Theorem confirms this hunch; the fact that $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ has n in the denominator shows that the variability of sample means decreases as the sample size increases. There's a catch, though. The standard deviation of the sampling distribution declines only with the square root of the sample size and not, for example, with $1/n$.

The mean of a random sample of 4 has half $\left(\frac{1}{\sqrt{4}} = \frac{1}{2}\right)$ the standard deviation of an individual data value. To cut the standard deviation in half again, we'd need a sample of 16, and a sample of 64 to halve it once more.

If only we had a much larger sample, we could get the standard deviation of the sampling distribution *really* under control so that the sample mean could tell us still more about the unknown population mean, but larger samples cost more and take longer to survey. And while we're gathering all that extra data, the population itself may change, or a news story may alter opinions. There are practical limits to most sample sizes. As we shall see, that nasty square root limits how much we can make a sample tell about the population. This is an example of something that's known as the Law of Diminishing Returns.

A Billion Dollar Misunderstanding? In the late 1990s the Bill and Melinda Gates Foundation began funding an effort to encourage the breakup of large schools into smaller schools. Why? It had been noticed that smaller schools were more common among the best-performing schools than one would expect. In time, the Annenberg Foundation, the Carnegie Corporation, the Center for Collaborative Education, the Center for School Change, Harvard's Change Leadership Group, the Open Society Institute, Pew Charitable Trusts, and the U.S. Department of Education's Smaller Learning Communities Program all supported the effort. Well over a billion dollars was spent to make schools smaller.

But was it all based on a misunderstanding of sampling distributions? Statisticians Howard Wainer and Harris Zwerling¹³ looked at the mean test scores of schools in Pennsylvania. They found that indeed 12% of the top-scoring 50 schools were from the smallest 3% of Pennsylvania schools—substantially more than the 3% we'd naively expect. But then they looked at the *bottom* 50. There they found that 18% were small schools! The explanation? Mean test scores are, well, means. We are looking at a rough real-world simulation in which each school is a trial. Even if all Pennsylvania schools were equivalent, we'd expect their mean scores to vary. How much? The CLT tells us that means of test scores vary according to $\frac{\sigma}{\sqrt{n}}$. Smaller schools have (by definition) smaller n 's, so the sampling distributions of their mean scores naturally have larger standard deviations. It's natural, then, that small schools have both higher and lower mean scores.

¹³ Wainer, H. and Zwerling, H., “Legal and empirical evidence that smaller schools do not improve student achievement,” *The Phi Delta Kappan* 2006 87:300–303. Discussed in Howard Wainer, “The Most Dangerous Equation,” *American Scientist*, May–June 2007, pp. 249–256; also at www.Americanscientist.org.

On October 26, 2005, *The Seattle Times* reported:

[T]he Gates Foundation announced last week it is moving away from its emphasis on converting large high schools into smaller ones and instead giving grants to specially selected school districts with a track record of academic improvement and effective leadership. Education leaders at the Foundation said they concluded that improving classroom instruction and mobilizing the resources of an entire district were more important first steps to improving high schools than breaking down the size.

The Real World and the Model World

Be careful. We have been slipping smoothly between the real world, in which we draw random samples of data, and a magical mathematical model world, in which we describe how the sample means and proportions we observe in the real world behave as random variables in all the random samples that we might have drawn. Now we have *two* distributions to deal with. The first is the real-world distribution of the sample, which we might display with a histogram (for quantitative data) or with a bar chart or table (for categorical data). The second is the math world *sampling distribution model* of the statistic, a Normal model based on the Central Limit Theorem. Don't confuse the two.

For example, don't mistakenly think the CLT says that the *data* are Normally distributed as long as the sample is large enough. In fact, as samples get larger, we expect the distribution of the data to look more and more like the population from which they are drawn—skewed, bimodal, whatever—but not necessarily Normal. You can collect a sample of CEO salaries for the next 1000 years,¹⁴ but the histogram will never look Normal. It will be skewed to the right. The Central Limit Theorem doesn't talk about the distribution of the data from the sample. It talks about the sample *means* and sample *proportions* of many different random samples drawn from the same population. Of course, the CLT does require that the sample be big enough when the population shape is not unimodal and symmetric, but the fact that, even then, a Normal model is useful is still a very surprising and powerful result.



JUST CHECKING

4. Human gestation times have a mean of about 266 days, with a standard deviation of about 16 days. If we record the gestation times of a sample of 100 women, do we know that a histogram of the times will be well modeled by a Normal model?
5. Suppose we look at the *average* gestation times for a sample of 100 women. If we imagined all the possible random samples of 100 women we could take and looked at the histogram of all the sample means, what shape would it have?
6. Where would the center of that histogram be?
7. What would be the standard deviation of that histogram?

¹⁴ Don't forget to adjust for inflation.

Sampling Distribution Models

Let's summarize what we've learned about sampling distributions. At the heart is the idea that *the statistic itself is a random variable*. We can't know what our statistic will be because it comes from a random sample. It's just one instance of something that happened for our particular random sample. A different random sample would have given a different result. This sample-to-sample variability is what generates the sampling distribution. The sampling distribution shows us the distribution of possible values that the statistic could have had.

We could simulate that distribution by pretending to take lots of samples. Fortunately, for the mean and the proportion, the CLT tells us that we can model their sampling distribution directly with a Normal model.

The two basic truths about sampling distributions are:

1. Sampling distributions arise because samples vary. Each random sample will contain different cases and, so, a different value of the statistic.
2. Although we can always simulate a sampling distribution, the Central Limit Theorem saves us the trouble for means and proportions.

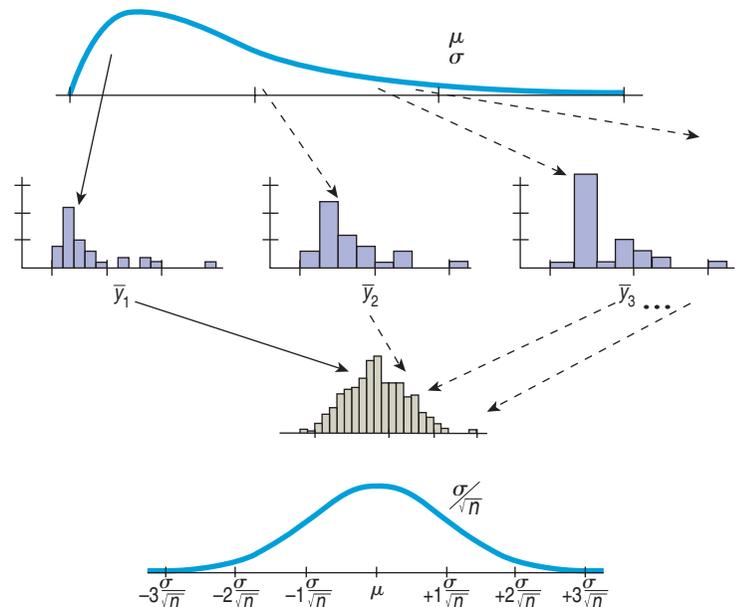
Here's a picture showing the process going into the sampling distribution model:

A S **Simulation: The CLT for Real Data.** Why settle for a picture when you can see it in action?

FIGURE 18.5

We start with a population model, which can have any shape. It can even be bimodal or skewed (as this one is). We label the mean of this model μ and its standard deviation, σ .

We draw one real sample (solid line) of size n and show its histogram and summary statistics. We imagine (or simulate) drawing many other samples (dotted lines), which have their own histograms and summary statistics.



We (imagine) gathering all the means into a histogram.

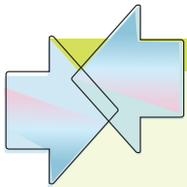
The CLT tells us we can model the shape of this histogram with a Normal model. The mean of this Normal is μ , and the standard deviation is $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$.

WHAT CAN GO WRONG?

- ▶ **Don't confuse the sampling distribution with the distribution of the sample.** When you take a sample, you always look at the distribution of the values, usually with a histogram, and you may calculate summary statistics. Examining the distribution of the sample data is wise. But that's not the sampling distribution. The sampling distribution is an imaginary collection of all the values that a statistic *might* have taken for all possible random samples—the one you got and the ones that you didn't get. We use the sampling distribution model to make statements about how the statistic varies.

(continued)

- ▶ **Beware of observations that are not independent.** The CLT depends crucially on the assumption of independence. If our elevator riders are related, are all from the same school (for example, an elementary school), or in some other way aren't a random sample, then the statements we try to make about the mean are going to be wrong. Unfortunately, this isn't something you can check in your data. You have to think about how the data were gathered. Good sampling practice and well-designed randomized experiments ensure independence.
- ▶ **Watch out for small samples from skewed populations.** The CLT assures us that the sampling distribution model is Normal if n is large enough. If the population is nearly Normal, even small samples (like our 10 elevator riders) work. If the population is very skewed, then n will have to be large before the Normal model will work well. If we sampled 15 or even 20 CEOs and used \bar{y} to make a statement about the mean of all CEOs' compensation, we'd likely get into trouble because the underlying data distribution is so skewed. Unfortunately, there's no good rule of thumb.¹⁵ It just depends on how skewed the data distribution is. Always plot the data to check.



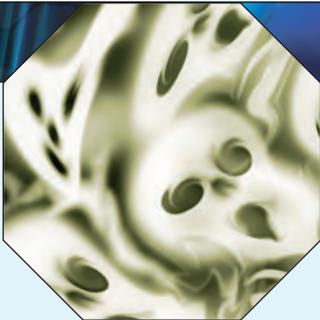
CONNECTIONS

The concept of a sampling distribution connects to almost everything we have done. The fundamental connection is to the deliberate application of randomness in random sampling and randomized comparative experiments. If we didn't employ randomness to generate unbiased data, then repeating the data collection would just get the same data values again (with perhaps a few new measurement or recording errors). The distribution of statistic values arises directly because different random samples and randomized experiments would generate different statistic values.

The connection to the Normal distribution is obvious. We first introduced the Normal model before because it was "nice." As a unimodal, symmetric distribution with 99.7% of its area within three standard deviations of the mean, the Normal model is easy to work with. Now we see that the Normal holds a special place among distributions because we can use it to model the sampling distributions of the mean and the proportion.

We use simulation to understand sampling distributions. In fact, some important sampling distributions were discovered first by simulation.

WHAT HAVE WE LEARNED?



Way back in Chapter 1 we said that Statistics is about variation. We know that no sample fully and exactly describes the population; sample proportions and means will vary from sample to sample. That's sampling error (or, better, sampling variability). We know it will always be present—indeed, the world would be a boring place if variability didn't exist. You might think that sampling variability would prevent us from learning anything reliable about a population by looking at a sample, but that's just not so. The fortunate fact is that sampling variability is not just unavoidable—it's predictable!

¹⁵ For proportions, of course, there is a rule: the **Success/Failure Condition**. That works for proportions because the standard deviation of a proportion is linked to its mean.

We've learned how the Central Limit Theorem describes the behavior of sample proportions—shape, center, and spread—as long as certain assumptions and conditions are met. The sample must be independent, random, and large enough that we expect at least 10 successes and failures. Then:

- ▶ The sampling distribution (the imagined histogram of the proportions from all possible samples) is shaped like a Normal model.
- ▶ The mean of the sampling model is the true proportion in the population.
- ▶ The standard deviation of the sample proportions is $\sqrt{\frac{pq}{n}}$.

And we've learned to describe the behavior of sample means as well, based on this amazing result known as the Central Limit Theorem—the Fundamental Theorem of Statistics. Again the sample must be independent and random—no surprise there—and needs to be larger if our data come from a population that's not roughly unimodal and symmetric. Then:

- ▶ Regardless of the shape of the original population, the shape of the distribution of the means of all possible samples can be described by a Normal model, provided the samples are large enough.
- ▶ The center of the sampling model will be the true mean of the population from which we took the sample.
- ▶ The standard deviation of the sample means is the population's standard deviation divided by the square root of the sample size, $\frac{\sigma}{\sqrt{n}}$.

Terms

Sampling distribution model	413. Different random samples give different values for a statistic. The sampling distribution model shows the behavior of the statistic over all the possible samples for the same size n .
Sampling variability Sampling error	414. The variability we expect to see from one random sample to another. It is sometimes called sampling error, but sampling variability is the better term.
Sampling distribution model for a proportion	416. If assumptions of independence and random sampling are met, and we expect at least 10 successes and 10 failures, then the sampling distribution of a proportion is modeled by a Normal model with a mean equal to the true proportion value, p , and a standard deviation equal to $\sqrt{\frac{pq}{n}}$.
Central Limit Theorem	421. The Central Limit Theorem (CLT) states that the sampling distribution model of the sample mean (and proportion) from a random sample is approximately Normal for large n , <i>regardless of the distribution of the population, as long as the observations are independent.</i>
Sampling distribution model for a mean	423. If assumptions of independence and random sampling are met, and the sample size is large enough, the sampling distribution of the sample mean is modeled by a Normal model with a mean equal to the population mean, μ , and a standard deviation equal to $\frac{\sigma}{\sqrt{n}}$.

Skills

THINK

- ▶ Understand that the variability of a statistic (as measured by the standard deviation of its sampling distribution) depends on the size of the sample. Statistics based on larger samples are less variable.

SHOW

- ▶ Understand that the Central Limit Theorem gives the sampling distribution model of the mean for sufficiently large samples regardless of the underlying population.
- ▶ Be able to demonstrate a sampling distribution by simulation.
- ▶ Be able to use a sampling distribution model to make simple statements about the distribution of a proportion or mean under repeated sampling.

TELL

- ▶ Be able to interpret a sampling distribution model as describing the values taken by a statistic in all possible realizations of a sample or randomized experiment under the same conditions.