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, we never actually draw all those samples, so the CLT is talking about an imaginary distribution—the sampling distribution model.

And the CLT does require that the sample be big enough when the population shape is not unimodal and symmetric. But it is still a very surprising and powerful result.

But Which Normal?

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?

A S The Standard Deviation of Means. Experiment to see how the variability of the mean changes with the sample size. Here's another surprising result you don't have to trust us for. Find out for yourself.

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.[9] 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 does it work that way? 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 $\sigma(\bar{y})$.

"The n's justify the means."

—Apocryphal statistical saying

A S 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

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.

[9] 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

Why is $SD(\bar{y}) = \frac{\sigma}{\sqrt{n}}$? 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:

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

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

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

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

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

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, $\hat{p}$; its sampling distribution 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 throughout Chapters 19 through 22.
- When we have quantitative data, we calculate a sample mean, $\bar{y}$; its sampling distribution 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 throughout 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.

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?