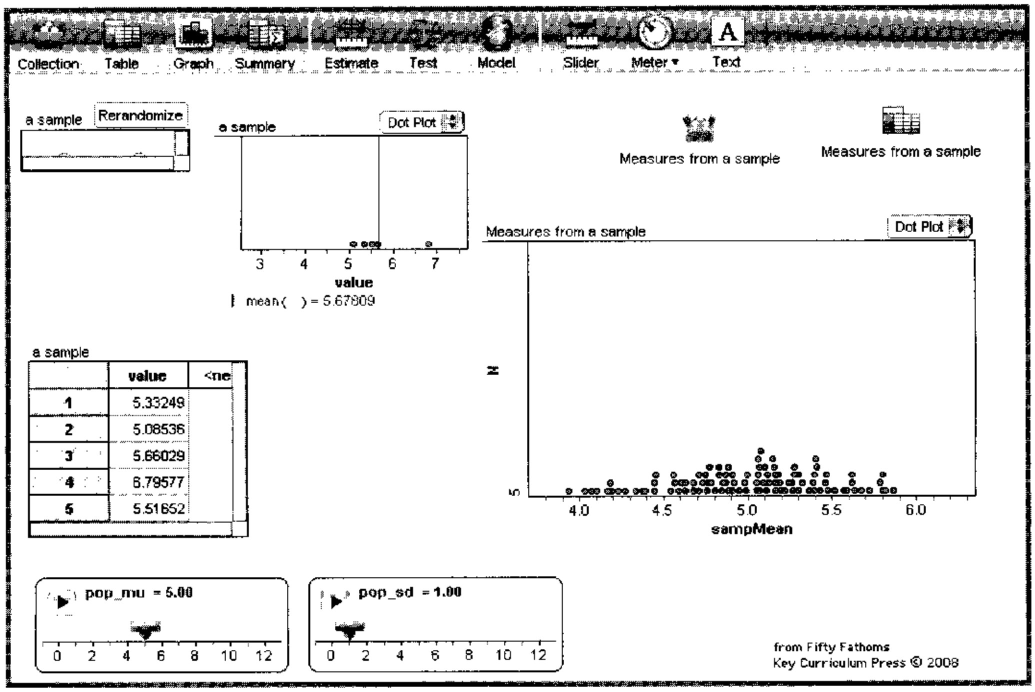
Demo 23: Sampling Distributions and Sample Size

How sampling distributions (of the mean) get narrower as you increase sample size

It makes intuitive sense that you get a better idea of a statistic—a mean, say, or a proportion—if you use a larger sample to estimate that statistic. But you seldom get to see the effect of sample size in a single graph. In this demo, you get to see it: We will build sampling distributions of the mean for samples of different sizes; and as usual, you will control the mean and standard deviation of the population using sliders.

Note: Why do we assume the population is normally distributed? In general, it's not. But for this demo, it's a convenience: You could use any distribution, as we do in Demo 27, "The Central Limit Theorem," but that requires extra machinery on the screen that is beside the point of this demo. With a distribution that we have described with a random number formula instead of by actually sampling, we need only the two collections—the sample and its measures—instead of three (population, sample, and measures).



mean( ) = 5.67809

a sample

	value	<ne
1	5.33249	
2	5.08536	
3	5.66029	
4	6.79577	
5	5.51652	

pop_mu = 5.00

pop_sd = 1.00

from Fifty Fathoms
Key Curriculum Press © 2008

What To Do

- ▹ Open **Sampling Dists Sample Size.ftm.** It will look something like the illustration.

The collection on the left, called **a sample,** has five points in it (though some points may be off the graph). The values are random, drawn from a normal distribution where the mean is 5.0 and the standard deviation is 1.0. On the right is a collection of measures from that sample—at this point, 100 means of that sample. That is, we chose 100 samples of five, computed their means, and plotted them. The vertical axis—you can see the "5" at the bottom—is the sample size.

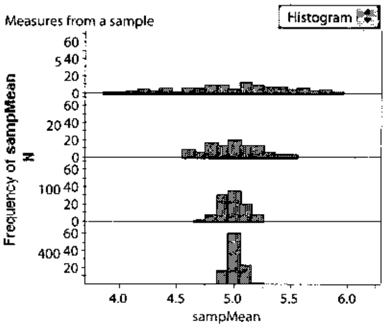
Before we explore sample size, we should look briefly at the left-hand collection (**a sample**) alone.

- ▹ Press the **Rerandomize** button in the left-hand collection. Note how the graph and table change. You can see the mean in the graph. See how it changes from press to press. (Here people often wonder if there is an automated way to collect all of those means. That's what *measures* are for; in fact, that's exactly what the right-hand graph is: 100 of those means.)

- ▹ Verify informally that the variation of the means you get is roughly the same as the variation of the means (named **sampMean**) in the right-hand graph.
- ▹ Click the collection **a sample** once to select it. Then choose **New Cases** from the **Collection** menu. Give the collection—our sample—15 more cases (for a total of 20).
- ▹ Click the **Measures from a sample** collection once to select it. Then choose **Collect More Measures** from the **Collection** menu.

Notice how the right-hand graph updates. It collected 100 more samples—this time of size 20—computed their means, and plotted them. The graph is set up to separate the different sample sizes.

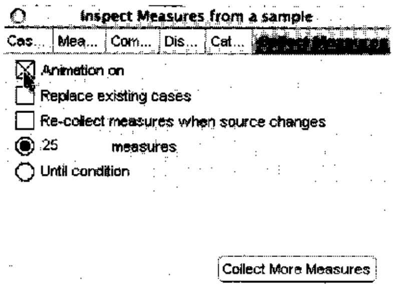
- ▹ Repeat the last two steps, adding 80 cases (for a total of 100); collecting measures; and then adding 300 cases (for a total of 400) and collecting measures again. There are so many dots now, they overlap.
- ▹ Change the dot plots to histograms by choosing **Histogram** from the pop-up menu in the graph. Your graph should show four histograms:



To Slow Down the Process

This way of collecting measures can seem mysterious. It might help to slow it down and see what happens in more detail. Fathom lets you do that. Here's how:

- ▹ Double-click the measures collection to bring up its inspector.
- ▹ Click the **Collect Measures** tab to bring up its panel. This is where you control how the measures are collected.



☒ Animation on
□ Replace existing cases
□ Re-collect measures when source changes
● 25 measures
○ Until condition

- ▹ Change the panel to turn on animation, as shown. You may also want to reduce the number of measures you collect (for the sake of speed) from 100 to 25.
- ▹ Close the inspector and proceed as before: Select the **Measures from a sample** collection, then choose **Collect More Measures** from the **Collection** menu. You'll see the new cases appear (on the graph of the sample means) one at a time instead of all at once.

Questions

1 Why does the mean change when you press **Rerandomize**? Isn't the mean just 5.0?

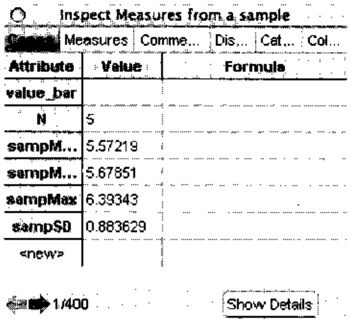
2 What do you notice about the distributions of the sample means when you increase the sample size? **Sol**

Extension

You don't have to collect sample *means.* You can make sampling distributions of any statistic. Try sample *medians,* for example. This will be easy, because these collections have actually been collecting sample medians (and sample maxima) all along:

▹ Double-click the measures collection to open its inspector.

▹ Click the **Cases** tab. You'll see a window like the one in the illustration.



Attribute	Value	Formula
value_bar		
N	5	
sampM...	5.57219	
sampM...	5.67851	
sampMax	6.39343	
sampSD	0.883629	
<new>		

▹ Drag the attribute name **sampMedian** to the graph to replace **sampMean.**

More Questions

3 How does the median graph look the same as (or different from) the graph for means?

4 In general, the bigger the sample size in the first collection, the narrower the distribution in the second, measures collection. Explain why that happens.

Another Extension

Do the same as in the previous extension, but for *standard deviation* (use **sampSD**) instead of median. In this case, the distribution gets narrower, as before, but something else happens as well. What is it? **Sol**

We discuss this in detail in Demo 25, "Does $n - 1$ Really Work in the SD?"

Note

The distributions on the right, the ones in the measures collection, are called *sampling distributions.* In many books, they care only about sampling distributions of the mean, which is what we did first. But you can make sampling distributions of any statistic you like. In Fathom, make it a measure in your original collection: Open the inspector by double-clicking the collection; click the **Measures** tab to go to that panel; and define the measure there by giving it a name and formula. Then, when you collect measures, it will be an ordinary, graphable attribute in the measures collection. (We prepared the collections in this demo specially to make mean, median, and max easy to study.)

The Central Limit Theorem is about sampling distributions becoming more normal as n increases. But not all sampling distributions do that. If you make a sampling distribution of **max( ),** for example, it will not gradually become normal—or even converge to a particular value. (You can see it as we did with the median; use **sampMax** instead of **sampMean** or **sampMedian.**)