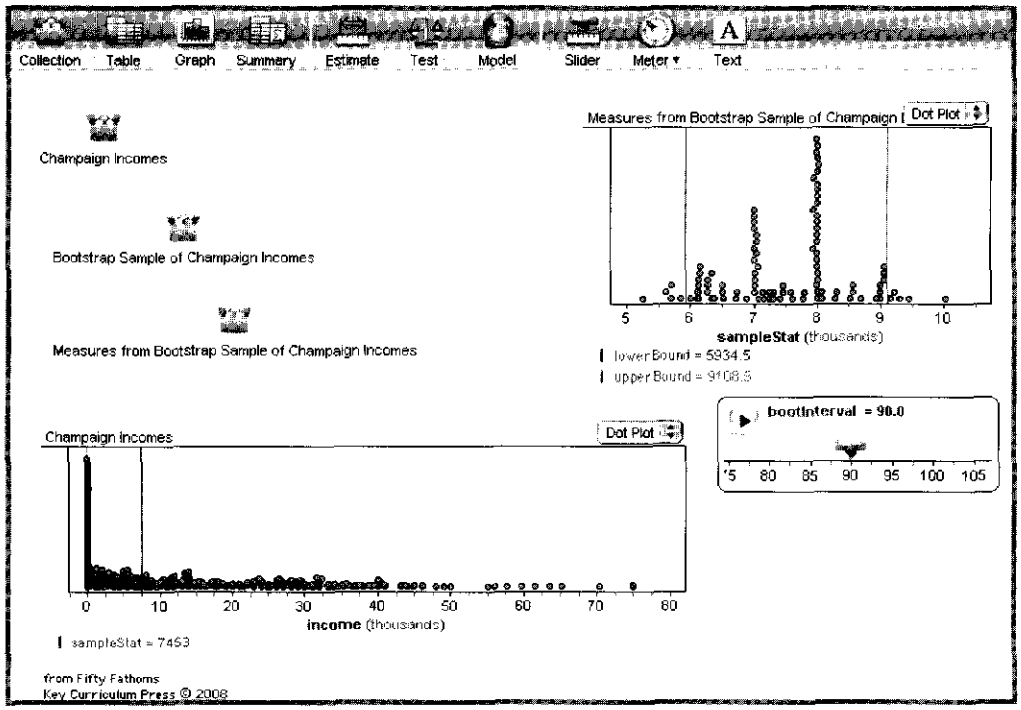
Demo 33: Using the Bootstrap to Estimate a Parameter

The bootstrap • Using resampling (with replacement) to create an interval for a parameter

The traditional way to estimate a mean—take a sample from the population and make a confidence interval—assumes that your sample is either large or normal. The Central Limit Theorem makes this process more or less correct for many distributions of data and for some parameters other than the mean, but sometimes (as we might see in Demo 27, "The Central Limit Theorem") the distribution is so odd as to make things problematic. Or perhaps you don't want to estimate the mean but, rather, some other statistic, and you're not sure that the Central Limit Theorem applies. Or maybe you're just skeptical whether your situation is appropriate for the traditional CI and are looking for something that doesn't make those assumptions. If any of these is true, the bootstrap may be for you.

The key idea behind bootstrapping is this: When you take a sample, that's all the data you have. So you assume that *the distribution of the values in the population is identical to the distribution in the sample.* That's the key point. You don't assume anything about normality.

And then, to find out how far off you might be, you resample from that distribution, calculating the parameter you're interested in. Let's try it, and estimate the *median* income of a population. We have a sample of incomes from the 1990 Census in Champaign County, Illinois.



What To Do

▷ Open **Bootstrap.ftm.** It should look like the illustration.

The top collection is the sample of 500 incomes. You can see them in a graph at the bottom. The median (called **sampleStat**) shows up as a vertical line on the graph. The middle collection, **Bootstrap Sample of Champaign Incomes,** is a collection (we did not make a graph of these incomes, but you could plot them if you wanted) where we have sampled from the former collection 500 times with replacement. That

is, it's a lot like the original, but some cases will be duplicated and some omitted.

Finally, the bottom collection contains measures from that sample, where, in this case, the measure (**sampleStat**) is the sample *median*. That statistic is plotted in the upper right—a sampling distribution, for 100 repeated bootstrap samples. We have also plotted two values: an upper bound and a lower bound of what's called a bootstrap interval. They are, in this case, the 5th and 95th percentiles of that distribution. The percentiles are controlled, in turn, by the slider **bootInterval**, currently set at 90. That is, the two lines encompass the middle 90% of the distribution. So the bootstrap interval is kind of like a confidence interval.

- ▷ Let's collect another set of bootstrap measures. Click once on the measures collection (the bottom one) to select it. Then choose **Collect More Measures** from the **Collection** menu. Fathom gradually collects 100 measures and replaces the data on the graph with the new sample medians.
- ▷ Repeat as necessary. Become convinced that, while the upper and lower bounds move around a little, they stay roughly in the same place.
- ▷ Drag the **bootInterval** slider (or edit the number) to see what happens.

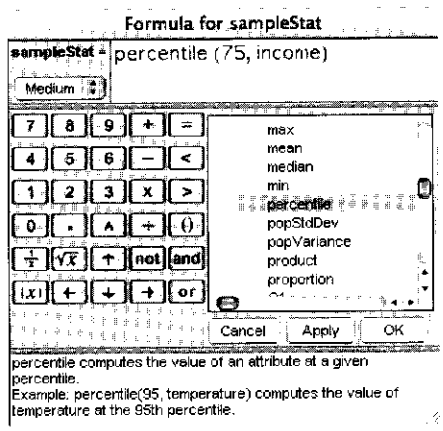
Questions

1. If the median of the original sample is 7453, how can it be that so many of the bootstrap samples have a median of 8000? Shouldn't 7453 be, if not the most popular value, at least close? **Sol**
2. How can you predict, before you move the slider, which way the bounds will move?

Onward!

Let's make a bootstrap of something else. How about the 75th percentile of income?

- ▷ Double-click the source collection—**Champaign Incomes**—to open its inspector.
- ▷ Click on the **Measures** tab to bring that panel to the front.
- ▷ Double-click the formula for **sampleStat** (it's currently **median(income)**) to open the formula editor.
- ▷ Edit the formula to read **percentile(75, income)**, as shown.[5]



- ▷ Press **OK** to close the editor. Note how the line moved in the bottom graph to indicate the new value of the sample statistic.
- ▷ Collect bootstrap measures again as before. Note the width and position of the 90% interval. (Remember, you control the interval with the **bootInterval** slider.)
- ▷ Try any other statistic that interests you. If you can't think of any, here's one:

  median(income, sex = "M") − median(income, sex = "F")

[5]The illustration also shows part of the *formula browser*. Although you can simply type the name of a function to enter it into a formula, you can also look for functions in the browser and double-click to enter them. **Percentile** is under **Functions | Statistical | One attribute.** Notice the help text that describes how to use the function.

Challenges

3 Explain why the median income of Champaign County seems to be so low. Can it be that half the people earned less than $7500 per year, even way back in 1989? **Sol**

4 Related task: Look up the median income of Champaign County, Illinois, from the 1990 Census. (Use the Internet. Start at www.census.gov.) Compare it to your interval and explain why it's so far off.

5 Make bootstrap estimates of men and women separately. Do the 90% bootstrap intervals of the median overlap?

Make a bootstrap estimate of the *mean* income at the 90% level. Compare that to a traditional confidence interval. (Use a Fathom estimate; choose **Estimate Mean** from the pop-up menu in the estimate, and drag the **income** attribute from the source collection to the place indicated. Also be sure to make it a 90% interval instead of the default 95%.) How close are they?

6 Explain why a bootstrap interval is pretty much the same as a confidence interval. Refer to the definition of a confidence interval if you have trouble.

7 Describe some advantages and disadvantages of using this bootstrap technique for making estimates of parameters.