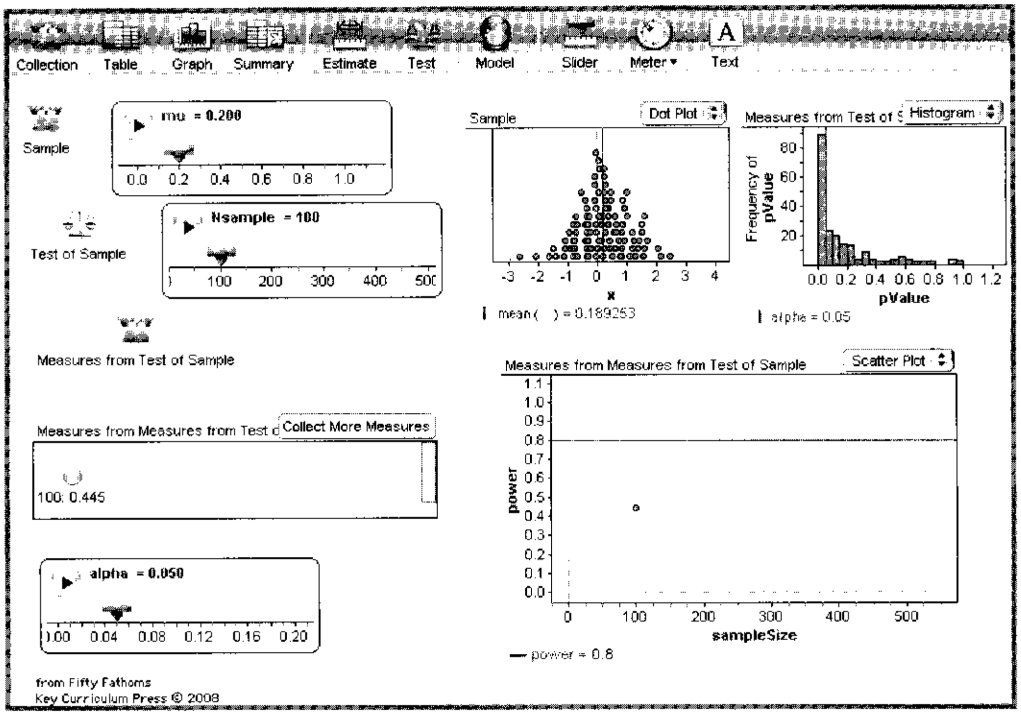
Demo 45: Power and Sample Size

How power—the chance that you reject the null hypothesis—changes with sample size

In Demo 44, "Power," we saw how power—the chance that we'll get what we want (a rejection of the null hypothesis)—depends on the population parameter and on the significance level α of the test. But it also depends on sample size; that's what this demo is all about. This issue is critically important for designing experiments. You have to figure out how big your sample will need to be in order to demonstrate what you want to show. If your sample is too small, the whole experiment may be useless. Put in economic terms, if you have low power, you're certain to spend money but are unlikely to get a significant result. On the other hand, samples can be expensive: Too much data and you spent more than you needed to.

We begin with a file based on the one from Demo 44, "Power."



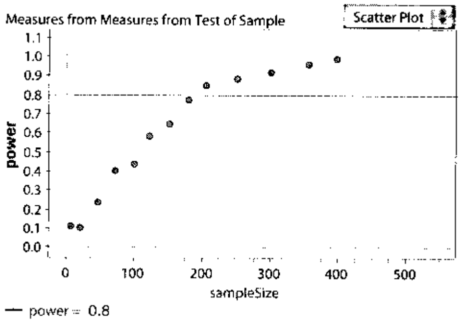
What To Do

▹ Open **Power and Sample Size.ftm.** It looks something like the illustration.

This file is like the last one except that we have a slider for the sample size, **Nsample,** currently set to 100. We have also set the population mean **mu** to 0.20. The test is still against $\mu = 0$ (which is false). So the power we see, 0.445, is the chance that we will correctly reject the null hypothesis; it corresponds to the 89 cases in the left-hand bin of the histogram (which still summarizes the P-values from 200 t-tests). Notice that the graph is now **power** as a function of **sampleSize.**

That power, 0.445, is not enough to convince us that we should go ahead with the experiment. How big does our sample need to be for us to have an 80% chance of rejecting $\mu = 0$ if the population is at 0.2?

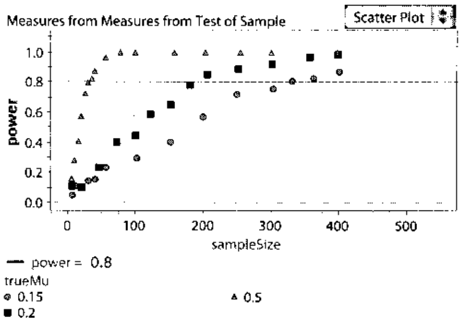
▹ Change **N** to 150. Then click **Collect More Measures** in the **Measures from Measures of Test of Sample** collection. Fathom resamples from the population, performs 200 tests on the samples of 150, and reports the P-values in the histogram. Another point appears on the scatter plot.



- ▹ Do the same for other values of **N** ranging from 5 to 400. You should end up with a graph looking something like the illustration; you'll see that to get a power of 0.8 requires a sample of around 200.

Of course, that is true only if the population mean is really 0.20. What if it's different?

- ▹ Set **mu** at 0.5 (that is, what if the true population were further from zero?). Then repeat the process, changing **N** and pressing **Collect More Measures** alternately until you can see the shape of the new curve—and roughly where it reaches a power of 0.8.
- ▹ Do the same with **mu** set at 0.15.



When you change **mu** and collect measures, a legend appears. When you're done, the graph will look like the one in the illustration.

What does this tell us? If the effect is a lot larger (**mu = 0.50**), we can get by with a much smaller sample of about 35 cases, and still have that 80% chance of showing that the mean is not zero at the 5% level. On the other hand, if we're wrong in the other direction by only a little (**mu = 0.15** instead of **0.20**), we'll need roughly twice the sample size to get that power of 0.8.

Challenges

1 In this situation, it looks as if the curves do not have a slope of zero when they hit the axis (**sampleSize = 0**). Does that make sense? Try to explain why a positive slope makes sense (or doesn't).

2 What would happen to these curves if the population standard deviation were 2.0 instead of 1.0? **Sol**

3 Suppose we made a graph of the sample size you need for a power of 0.80 as a function of the population mean. What would that look like, roughly? How could you use it to plan experiments?

4 What do you think about looking for a power of 80%? What considerations would make you want it to be higher? Or lower? How high should it be before you do an experiment?

5 As at the end of the previous demo, re-create this, but instead of using the *t*-test, reject the null hypothesis if the box in a box plot of the sample data does not overlap zero (use the functions **Q1** and **Q3** to determine that formulaically). Compare that to the *t*-test. Which is more powerful?