Demo 44: Power

How power—the chance that you reject the null hypothesis—changes with the population parameters

Here's one way to think about it: Power is the chance that you will get what you want.

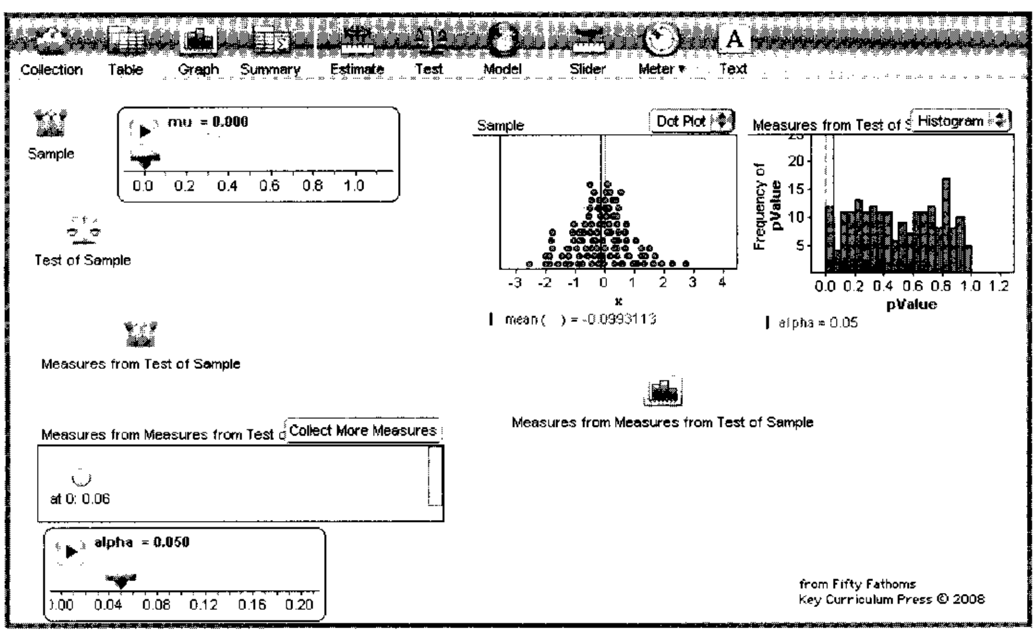
According to the dominant scientific paradigm, our desires do not enter into it. In fact, a lot of the cautionary machinery of statistics exists in order to ensure that our results are honest and dispassionate. In many cases, a null result is worthwhile and illuminating. But let's face it: You just spent a year—or ten—of your life, or $50 million, working on something that matters to you. And it all comes down, at the end, to a *P*-value. If it's small, what you did worked. If it's not, you may not be a failure as a person, but you still feel bad.

So *before* you do this work, you should know the chance that you will get that low *P*-value—the chance that you will reject the null hypothesis. Finding that probability—the power of your test—is one of the most useful ideas in elementary statistics. But it's conceptually difficult, too.

First, it depends a great deal on what the truth is. Suppose your test boils down to a null hypothesis that the mean of some quantity is 0, and you test whether you can reject that hypothesis with an α level of 0.05. If the true mean is 0.0001, the null hypothesis is false. But if the spread of the data is about 1 unit, you won't reject the null hypothesis any more than if the mean were 0. On the other hand, if the true mean is 10, you'll reject it every time.

Power also depends on sample size. The bigger the sample, the better you know the true mean—and the more likely you are to detect a difference in that mean from the null value. Finally, power depends on the α level, as we will see. If you are willing to accept more false positives (Type I errors), you can reduce your false negatives (Type II errors), thereby increasing your power.

So power depends on the true value, on sample size, and on the significance level α. It also depends on the spread of the population. While you'll often see it reported as a single number, it's important to remember that power is a *function*.

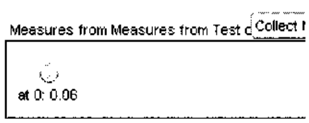


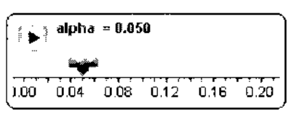
What To Do

- Open **Power.ftm.** It will look something like the illustration.

The machinery in this demo is complex, and some objects have been shrunken into icons. Be sure you have looked at Demo 43, "The Distribution of *P*-Values."

Sample holds 100 cases, graphed in the top-center dot plot. Their true population mean is **mu,** controlled by the slider. They are normally distributed with a standard deviation of 1.0. The test is a one-sample *t*-test of the mean against a null hypothesis that $\mu = 0$. The *P*-values from 200 such tests are graphed in the histogram, upper right. The open, measures collection at left summarizes those 200 tests.
In this case, 0.06 of them (at $\mu = 0$) rejected the null hypothesis at the 0.05 level; that level is controlled by the **alpha** slider just below it.

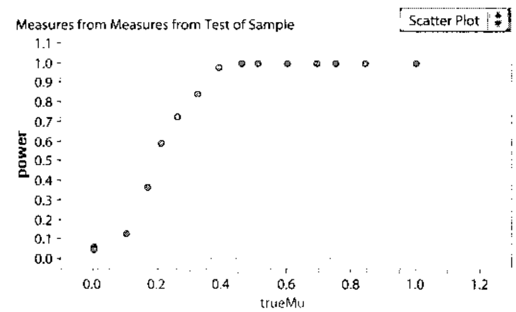




- Press the **Collect More Measures** button in the open collection. Fathom performs 200 tests, resampling every time, and updates the displays. A new summary "ball" appears.

Make sure you understand where the number on the ball comes from: It's the number of cases in the histogram that are less than 0.05, divided by 200 (the total number of tests). Furthermore, make sure you understand that these are tests where we *wrongly* rejected the null hypothesis: Type I errors. Finally, note that these values are not far from 0.05—which is good because that's exactly what the alpha level is supposed to be: the chance that you get a Type I error.

- Change **mu** to a small value, about 0.1. See how the dot plot changes.
- Press **Collect More Measures** again. Graphs update, and you get a new ball.
- Let's start graphing these results: Open the iconified graph by dragging its lower-right corner to fill much of the remaining space. You'll see the proportion of rejections, **power,** plotted against **trueMu,** the population proportion.
- Repeatedly change **mu** and **Collect More Measures,** filling in the graph pretty well between **mu = 0** and **mu = 1.** Make sure at every point that you can see the correspondence between the new point and the histogram above (which you may need to rescale from time to time). You should see a graph like the one in the illustration.



Notice that the height of each point is like the height of the left-hand bar in the histogram of *P*-values—the bar for the region below $P = 0.05$—which represents the tests where we would reject the null hypothesis.

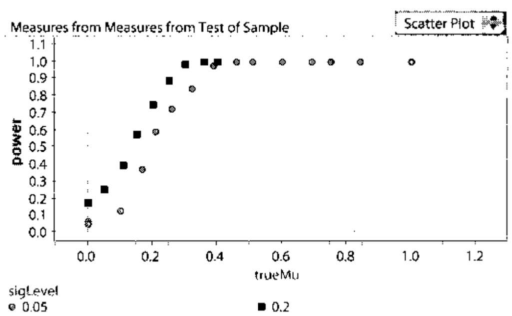
This curve is one way to look at the power function. What does it tell you? That if you're taking samples of 100 from this population with a standard deviation of 1.0, and your standard for detecting an effect is $P < 0.05$, your chances of rejecting that null hypothesis are good if the true population mean is over about 0.4, and bad if it's less than about 0.2, even though the null hypothesis would be false. In practice, many people accept a power of about 0.8 or 0.9—corresponding here to a mean of 0.3 or 0.4.

Let's see what the results would be if we used a different alpha.

▹ Set the slider **alpha** to **0.2.**

▹ Repeat the previous actions, filling in the graph between **mu = 0** and **mu = 0.5**. You'll see the new points appear on the graph, coded to show the different **alpha**s (here called **sigLevel**).

Here you see that we can detect a smaller difference from zero in the mean. For example, if the population mean **trueMu** is only 0.2, we would have a **power** greater than 0.8.



This graph illustrates two important principles:

- The intercept of the power curve is the alpha level of the test.
- If you increase the alpha level, you increase the power.

Challenges

1 Explain why the alpha level is the same as the intercept.

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

3 The points in the graph look as if they don't quite lie on a smooth curve. Why not? **Sol**

4 It looks as if the curve—the one the points don't quite lie on—has a slope of zero at the vertical axis. Is that true? (Zoom in, add points, and see what you think.) Explain why or why not.

5 Re-create this demo, 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?

6 Re-create this demo, but with a test of difference of means. (You'll need to create a "null" sample in your source collection for comparison.) What do you find? How can you relate it to the test of mean?

7 Re-create the previous task (difference of means) except that this time, instead of using *t*, reject the null hypothesis if the boxes in the two box plots do not overlap (use the functions **Q1** and **Q3** to determine that formulaically). Compare that to the *t*-test. Which is more powerful?

