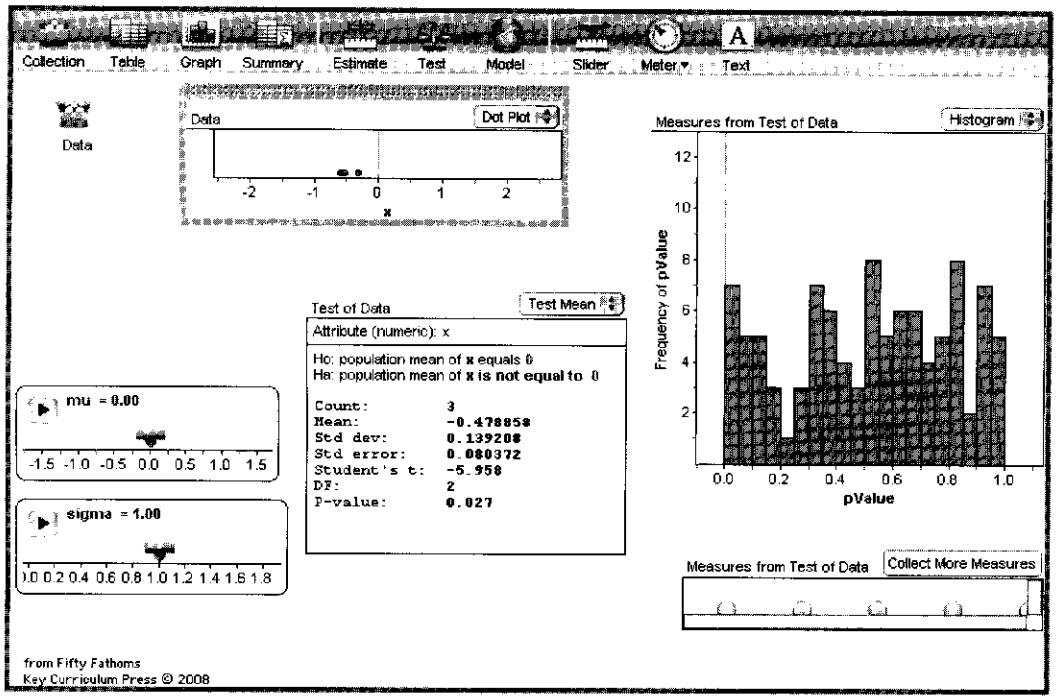
Demo 43: The Distribution of *P*-Values

How the distribution of P is flat if the null hypothesis is true • How it changes if the null hypothesis is false

This demo assumes that you know about hypothesis tests and what *P*-values are. You can think of it as an extension of Demo 39, "Another Look at a *t*-Test"; in fact, the file uses a lot of the same machinery. Here we focus on the distribution of *P*-values, which leads to discussions of power.



▷ Open **Distribution of P-values.ftm.** It should look something like the illustration.

In this document, the data—the collection in the upper left—are once again three points drawn from a normal distribution with mean **mu** and standard deviation **sigma.** You can see the data in the small graph at the top. A *t*-test of the mean appears in the middle of the screen (with **Verbose** turned off, to save space). At right, you can see a distribution of 100 *P*-values from 100 repeated tests, each test a new sample drawn from that normal distribution.

▷ Look at the test in the middle of the screen and the **data** above it; in particular, look at the data values and the *P*-value from the test; make sure they seem to correspond. That is, the data probably straddle zero, and the *P*-value is probably high.

▷ Drag the **mu** slider to the right, say, to 1.0, and see how the *P*-value changes. (Note: The histogram of the distribution of *P*-values will not change.)

▷ Drag **sigma** to make it smaller; again, see how *P* changes.

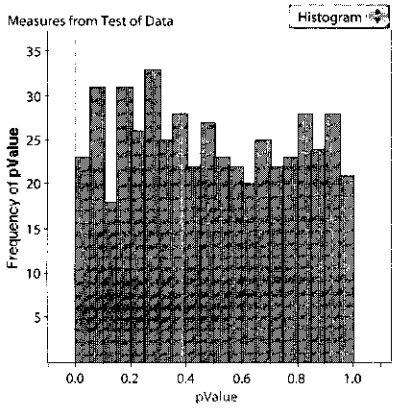
▷ This completes our orientation; return the sliders to **mu = 0** and **sigma = 1**—a standard normal distribution, and one where the null hypothesis ($\mu = 0$) is *true.*

▷ Click the **Collect More Measures** button in the collection at lower right. Fathom empties the graph and starts performing *t*-tests. Observe how the top graph updates with new samples, the test updates with each analysis, and the histogram updates with each new *P*-value. Do this repeatedly until you understand what's happening. See if you can figure out the distribution of *P*.

Probably, with only 100 points, it is not clear what that distribution is. In fact, it's *flat,* which is a very important result you may never have seen before. Now, in order to see this better, we're going to increase the number of tests we do from 100 to 500. But with the animation, it will take too long. So the next two steps are for speed; you can omit them if you have a really fast computer.

When you do these steps, however, you will no longer see the test itself, the data, or the gradual updating of the histogram. Slow it down again if you need to for understanding.

- *Speed step 1:* Turn the graph of the original data (the one with three points) and the *t*-test into icons by dragging their corners until they're small.
- *Speed step 2:* Double-click the collection **Measures from Test of Data** (the box, not its name) to open its inspector. Turn *off* animation, and increase the number of measures from 100 to 500. Close the inspector.
- Now click **Collect More Measures** to get a fresh, highly populated plot. It should look something like the illustration (and more convincingly flat).



- Put your "hand" over the left-hand bar and look in the status bar at the bottom of the window. That will show you how many cases—out of 500—are in the bar, which includes all of the *P*-values between 0.00 and 0.05. That is, it's the number of Type I errors, where we would erroneously reject the (true) null hypothesis at the 5% level. That number should be about 500 ÷ 20, or 25 cases.
- Predict what will happen when you increase **mu.** (Remember what happened when you increased **mu** before we collected measures.)
- Increase **mu** to 0.2 and collect measures again. Observe how the graph changes, and how the count of rejections in the left-hand bar changes as well.
- Keep increasing **mu** and collecting measures, until you reject the null hypothesis in almost all of the tests.
- Predict first, then start increasing **sigma,** collecting measures as you go. What happens?

What You Should Take Away

Two things happen. First, accepting or rejecting a hypothesis is not the clear-cut decision we might think it is. In Fathom, we can control the highly variable truth, and test repeatedly, so we see a range of results. And look how mushy the distribution of *P*-values is! Even at $\mu = 1.5$, we reject $\mu = 0$ much less than half the time. So suppose you do a study and you get $P = 0.07$. What should you make of it? Whatever you decide, remember the wide variety of "truths" that could plausibly give you that result.

Second, statistics is at its most useful in that region between the null hypothesis and where it's obvious you should reject it. That's the transition we just explored: between only 5% of the tests being in the region $P < 0.05$ and almost all of the tests falling there. Each test has a characteristic function that describes how the probability of rejection changes as the population parameter changes: this is *power.* Ideally, your test will not reject the null hypothesis when it is true, but just stray a millimeter from that knife edge, and it rejects. Alas, you can't make such a test—there are always tradeoffs—but some tests come closer than others to this ideal.