Demo 18: The Road to Student's *t*

Using standard error as the scale for measuring how far a sample mean is from the true mean • How these quantities are not normally distributed; in fact they follow a t-distribution

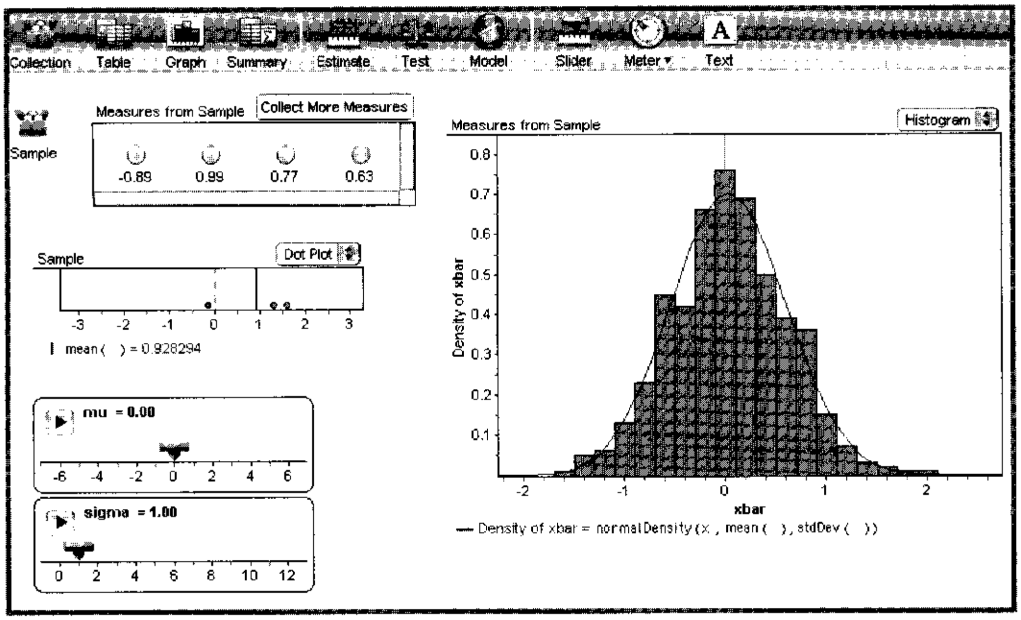
Where have the last two demos led us? Why do we care about this difference between standard deviation and standard error? Here's what we have seen:

- Standard deviation measures the spread in the sample and therefore reflects the spread in the population.
- The standard error of the mean measures the spread of a sampling distribution and is therefore a measure of how well we know the mean of the population if we have only a sample.

This second point sounds obvious, but it isn't. You can remember sampling distribution this way:

The SE—the standard deviation divided by $\sqrt{N}$—is the standard deviation of the sampling distribution. The standard deviation is a measure of how far a given data value is likely to be from the mean. So, if you take a sample and compute its mean, how far is that mean likely to be from the true mean of the population? The standard error.

That's what we said in Demo 17, "What Is Standard Error, Really?" Now we look more deeply and uncover one of the great subtleties of statistics.



| mean( ) = 0.928294

— Density of xbar = normalDensity(x , mean( ), stdDev( ))

What To Do

▷ Open **Road to t.ftm.** It will look something like the illustration.

This is based on **What Is SE.ftm,** if it looks familiar. We see the **Sample** collection (three cases, normally distributed, sliders for **mu** and **sigma,** plus a graph) and its derived **Measures from Sample** collection.

The histogram shows the sampling distribution of 500 means (called **xbar**), with the relevant[2] normal curve superimposed. It looks pretty normal.

[2] We're using the data's mean and SD to define a normal curve. We could have used slider parameters in our formula, for example, **normalDensity(x, mu, sigma /** $\sqrt{count( )}$ **),** but using the data works well here.

▹ To show that it wasn't a fluke, press the **Collect More Measures** button on the upper-middle collection, and generate a new set of 500 **xbar**s.

We can predict the shape of the sampling distribution if we sample repeatedly. We take the standard deviation, divide by $\sqrt{N}$, and use that as the standard deviation for our normal distribution. If we know about the Central Limit Theorem (Demo 27), we even know it's normal.

A Longer Explanation Than Usual

Here comes the hard part, conceptually: In real life, we only get the one sample, and we don't know what the true mean and standard deviation are. When we get that sample, we'd like to know how far that sample is likely to be from the mean.

We're tempted to say, "Let's take the sample standard deviation and divide it by $\sqrt{N}$ to get a standard error. Now we can reverse the traditional logic of the standard error, like this: Just as there's a 95% chance that the sample mean will be within two standard errors of the true mean, we have 95% confidence that the true mean is within two standard errors of the sample mean."

We'd be wrong. That symmetrical logic breaks down, as we will see. In the demo, we'll pretend for a moment we don't know "the truth" when we draw the sample. We pull three numbers **x** from the distribution and calculate the mean, **xbar.** We want to get some idea how far that is likely to be from the true mean.

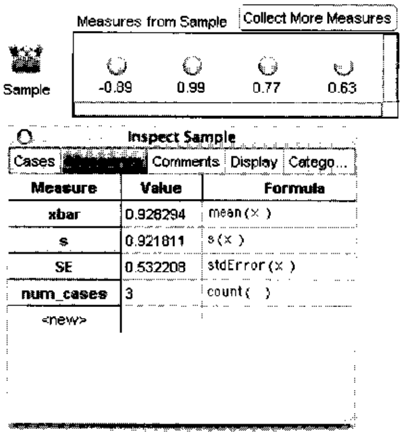
Here is the crux of the matter: It's not fair to calculate that distance in absolute units. If we really don't know how far it is to the true mean, that answer should not depend on the scale—for example, the system of units—we measure with. But we *can* express the distance *in terms of the spread of our sample* (or spread of the *population,* but we don't know that). We could use the standard deviation of our sample as a unit of measurement, but since we know that the distance from the sample mean to the true mean scales like standard error (not standard deviation), let's use that. So far, we've matched the tempting logic above.

Then imagine that we're told the truth, so we can see how far off we really are. Now, instead of plotting the errors in absolute units, we'll make the experiment independent of the particular scale and calculate the errors in *units of sample standard errors.* And plot them.

Note: This trick, scaling the data this way, uses *dimensionless* quantities. If the original measurements were in centimeters, say, this would give us an analysis that had no units at all. That way it would scale properly even when we changed systems of measurement.

▹ First, we have to calculate this dimensionless "difference." Double-click the **Sample** collection (not **Measures from Sample**) to open its inspector.

▹ Click the **Measures** tab to open that panel.



Measure	Value	Formula
xbar	0.928294	mean(x)
s	0.921811	s(x)
SE	0.532208	stdError(x)
num_cases	3	count( )
<new>		

▹ Make a **<new>** measure; call it **error.**

▹ Double-click **error**'s formula box (to the right) to open the formula editor.

▹ Enter **xbar / SE** and press **OK** to close the editor (these attributes are both defined in this same panel, so it's OK to use their names). Note: We assumed **mu = 0**, so **xbar** *is* the deviation.

▹ Close the inspector to save screen space.

We've defined **error** to be a dimensionless number—the distance from the sample mean (**xbar**) to the true mean but in units of standard errors. Now let's see its distribution:

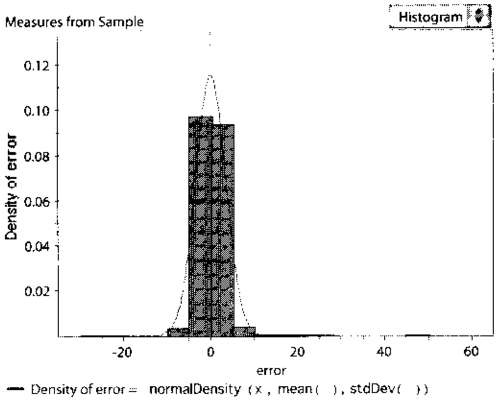
▹ Press the **Collect More Measures** button on the measures collection. Fathom collects 500 new

xbars. Where are the **error**s? We need to tell Fathom to plot them.

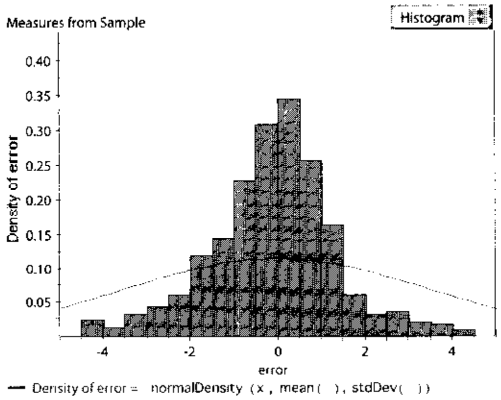
▷ Double-click the **Measures from Sample** collection (the open box, not its name) to open its inspector.

▷ Click the **Cases** tab to open that panel.

▷ Drag the name of the attribute **error** from the inspector to the horizontal axis of the graph, replacing **xbar**. The graph updates, and we see the distribution of this new attribute.



▷ It looks bizarre, like the graph above. We need to rescale the axes and change the bin width. You can do it by hand, or double-click the graph and set the numbers in the graph's inspector. Use a **binWidth** of 0.5 and a range (**xLower** to **xUpper**) of about –5 to +5.



The distribution does not match the curve—a normal curve with the same mean and standard deviation as the data.[3] So there must be some really far-out points to make the standard deviation so large. You probably saw them before you rescaled.

At any rate, we have found that *even though the xbars are normally distributed, these error quantities are not* (at least with $n = 3$).

If it's not normal, what is it? Let's see:

▷ Click the graph once to select it.

▷ Choose **Plot Function** from the **Graph** menu. The formula editor opens.

▷ Enter **tDensity(x, N – 1)**. Press **OK** to close the editor and show the graph.[4]

It should match! This distribution was Gosset's great discovery, what we now call Student's *t*-distribution. What we called *error* we could just as well have called *t*. To emphasize:

$$t=\frac{\bar{x}-x_0}{SE}$$

where x_0 is a value you're comparing the sample to. Of course, in this case we needed to know the true mean, but the symmetrical logic that failed us earlier now holds up. The point is that if you measure these differences in terms of standard errors, you don't need to know the population spread—because it's all done using the scale of the *sample* standard error.

Measuring in terms of spread also helps us clarify this important distinction:

- A difference measured in units of sample standard deviation (effect size) is an indication of how *meaningful* the difference is. That is, it says how much distributions overlap, or how far apart they are in terms of their spreads.

- A difference measured in units of standard error of the mean (Student's *t*) is an indication of how statistically *significant* the difference is. That is,

[3] We could simply have plotted **normalDensity(x)** to get a standard normal curve, which looks (at first glance) much closer to this distribution. It is just as unsuitable, however, because of the data out in the tails. With the curve we have drawn, the difference between the histogram and the function is more dramatic.

[4] Why n – 1? It's a parameter of the distribution called the *degrees of freedom*. The –1 has to do with the idea that once we specify the mean, we can freely choose only n – 1 data values—the last one is determined. The main point, however, is that the distribution actually has a different shape for different values of n.

it helps us understand how likely it is that the difference occurred by chance.

Extensions

These extensions are worth some time and thought if you're studying the *t*-distribution, *t*-tests, or confidence intervals based on *t*.

- ▹ Change the **error** histogram to a normal quantile plot. What does that graph tell you? (Note that when you change back, you may have to remind Fathom to display a *density* histogram. Choose **Scale | Density** from the **Graph** menu.)
- ▹ Change the value of **sigma** (and re-collect measures) to see that the **error** distribution (that is, *t*) stays the same even when the population spread changes.
- ▹ Change the value of **mu** (and re-collect measures) to see the distribution change. That's because our definition of error assumes that the true population mean is zero (it should have been **(xbar – mu) / SE**). What we get is a *noncentral t*-distribution.
- ▹ Add cases to the **Sample** collection. See how the graphs change as sample size changes.

Challenges

1 Look back at the histogram of **error.** Maybe it *is* normal, but the normal curve just needs to be rescaled to fit the graph. After all, it looks about the right shape—it's just that the SD is too big. Try making a slider for a parameter to rescale the SD, and see what you find. (One suggestion for a formula:

normalDensity(x, mean( ), stdDev( ) / K)

where **K** is the slider.) It fits pretty well—what is it about the graph that says it does not fit well enough?

2 One way to characterize this new distribution is that it's kind of normal except that it has longer tails. Why is that so? What do you suppose the tail samples tend to have in common? What about the ones in the hump? (Demo 19, "A Close Look at the *t*-Statistic," might help you look at this.)

3 Why didn't we just measure the difference from the true mean in standard *deviations* instead of standard *errors*? It would be dimensionless—independent of the original units—and in a scale that depended only on the sample we got. Try it, see what happens, and explain your results. **Sol**

4 We made this statement earlier: "Just as there's a 95% chance that the sample mean will be within two standard errors of the true mean, we have 95% confidence that the true mean is within two standard errors of the sample mean." We claimed that this statement (in its context) was wrong—or at least misleading. What's wrong with it?