Demo 25: Does *n* – 1 Really Work in the SD?

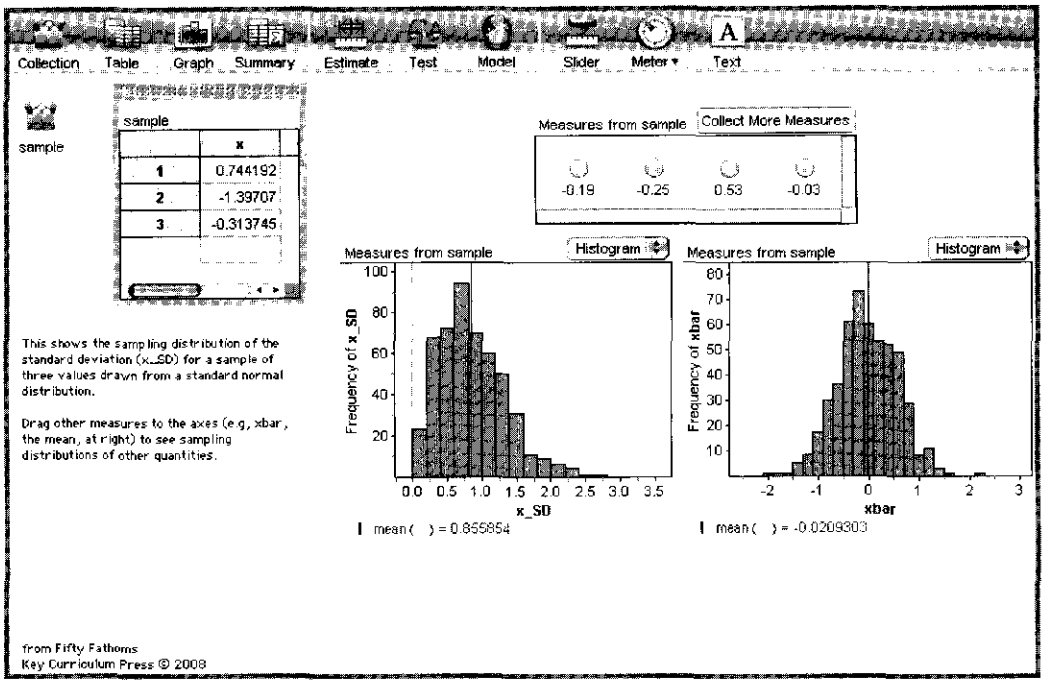
Unbiased estimators • How the familiar formula for sample standard deviation is not unbiased • Why we should care about variance

Amazingly, no. Not exactly. Here's a common misconception. See if it sounds familiar:

Suppose you draw a sample from a population and measure some continuous attribute for every case. Estimating the mean of the population is easy: The mean of the sample is an unbiased estimate of the mean. Estimating the standard deviation of the population is easy, too: You use the standard deviation of the sample—almost. Instead of n in the denominator, though, inside the square root sign, you use $n - 1$. You can't use n because that SD is not an unbiased estimator of the population standard deviation, whereas the one computed with $n - 1$ is. That is,

$$SD_{sample} = \sqrt{\sum \frac{(x-\bar{x})^2}{n-1}}$$

What's not to like? Where's the problem? It's true that we're supposed to use $n - 1$, but *it is not true that the SD so calculated is an unbiased estimator of the population SD.* If this is old news, more power to you. You don't need this.



What To Do

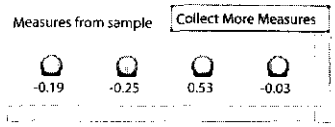
▹ Open **Estimating SD from a Sample.ftm.** It should look something like the illustration.

On the left you can see three values (**sample**) drawn from a normal distribution with a mean of 0 and a standard deviation of 1. The left-hand graph is a distribution of standard deviations (**x_SD**) calculated from 500 different samples, and the mean of those 500 SDs. On the right is a sampling distribution of the mean, just like the sampling distributions we have seen, for example, in Demo 18, "The Road to Student's *t*." But the graph on the left is a sampling distribution of the SD, rather than a sampling distribution of the mean.

Notice that the mean of **x_SD** is less than 1.0, which is the "true" value. Also, the distribution is skewed. What's going on?

First, let's see whether the mean is really as low as it looks.

▷ Click **Collect More Measures** in the **Measures from sample** collection.



The graph will update with the SDs from 500 new samples. It should still be low.

Next, let's see what the "plain" SD (made with n instead of $n - 1$) would look like graphed.

▷ Double-click the measures collection to open its inspector. Drag **x_popSD** from the **Cases** panel to the horizontal axis of the right-hand graph, replacing **xbar**. Its mean should be even lower!

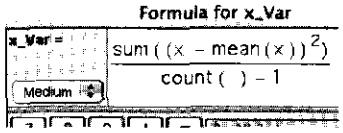
Now, let's see what's really going on:

▷ Double-click the original sample collection (the one with three cases) to open its inspector. Click the **Measures** tab to bring up that panel.

▷ Make a new measure by clicking in the **<new>** cell. Enter **x_Var** (for "x variance").

▷ Double-click **x_Var**'s formula cell (two boxes to the right of the name) to open the formula editor.

▷ Enter **sum((x - mean(x))^2) / (count( ) - 1)**, as shown. (That's the variance of x, using $n - 1$ because this is a sample.)



If you have trouble entering the formula, just type the characters in order (with all parentheses). Or you could simply use **Variance(x)**, since **Variance** is a built-in function—but writing out the formula may be more appropriate here.

▷ Close the formula editor with **OK**. Verify that Fathom computed a variance (it should be the square of the first standard deviation). Then close the inspector to clean up the screen.

▷ Again, click **Collect More Measures** in the measures collection. This time Fathom collected variances as well.

▷ Double-click the measures collection to open its inspector.

▷ Drag the new attribute, **x_Var**, from the inspector of the measures collection to the horizontal axis of the right-hand graph, replacing **x_popSD**.

▷ Click on one of the bars to show where these points are in the **x_SD** graph.

▷ Resample the variances (that is, click **Collect More Measures**) a few times, until you're convinced that, unlike the standard deviations, they really do center around 1.0.

That is, the sample standard deviation (the one with $n - 1$) is not an unbiased estimate of the population standard deviation. *But the sample variance is an unbiased estimate of the population variance,* even though the variance is just the square of the SD.

This helps answer the question of why we should ever care about the variance. Most important, though, is that it shows how a simple transformation of a variable can radically change how a distribution looks.

Questions

1 Why do you suppose we did this with a sample size of 3 as opposed to, say, 100?

2 How could we have predicted that the mean of the "plain" SD distribution would be smaller than the first one we tried? **Sol**

3 How is it possible that the distribution of the SD can look so different from that of its square, the variance?