[10-08-10-RT]

Standard deviation clarified

The purpose of this note clear up a confusion that exists throughout the literature concerning standard deviation.

The average variance of a collection of data points $\{x_1, x_2, \ldots, x_n\}$ is denoted by the symbol σ^2 and defined thus

$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}$ where μ is the arithmetic mean of $\{x_1, x_2, \ldots, x_n\}$, $\mu = \frac{x_1+x_2+\cdots+x_n}{n}$.

The standard deviation denoted by the symbol σ is the square root of the variance; that is,

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \; .$$

So far, so good. Now, suppose we desire to know μ, σ^2, and σ for a collection of data points, but the values of all the data points are impractical to obtain.

For example, suppose that we manufacture a certain kind of fuse for the protection of electrical circuits and the buildings in which such circuits are installed. The fuse is nominally rated at 20 amps, meaning, roughly, that it should blow when the current in the circuit is 20 amps. Now, it is of course practically impossible, or at least extremely expensive, to make thousands of fuses each of which blows at exactly 20 amps. Nor is such an exact fuse typically needed. Rather, a fuse that does not blow too early nor too late is acceptable. Our company need only guarantee that the fuses will blow when the current is somewhere between 19.9 amps and 20.1 amps. But, how can we know for each fuse we make that it has the essential quality "blows between 19.9 and 20.1 amps"? We certainly cannot test each fuse, because the test requires that the fuse blow while we measure the current. If we test every fuse, we have no fuses remaining to sell!

We need a way to say something about all the fuses based on our destructive tests of a small sample of fuses that come off the production line. We wish to know μ, σ, and σ^2 for the entire population of fuses, but our knowledge must be based on μ, σ, and σ^2 of the fuses in our sample.

Naturally, we call in a statistician. She points out that μ, σ, and σ^2 for the sample we test may not have exactly the same values as μ, σ, and σ^2 for the entire population of fuses. To keep things straight, she proposes that we call μ, σ, and σ^2 **parameters** of the population and that the μ, σ, and σ^2 of the sample we call $\overline{x}$, s, and s^2. These $\overline{x}$, s, and s^2 are to be collectively known not as parameters, but as **statistics**. So, we make a distinction between the population mean, μ, and the sample mean, $\overline{x}$. Similarly, we distinguish the population variance, σ^2, and the sample variance, s^2. Finally, we distinguish the population standard deviation, σ, from the sample standard deviation, s.

How do we compute the sample standard deviation, s, given the data points collected by destroying the sample fuses? The same way we compute the standard deviation of anything,

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}} \; .$$

How do we say what the value of σ is, given we know the value of s? It turns out that σ is approximately equal to s. And, we can get an even better approximation for σ by multiplying s by a factor k which is equal to $\sqrt{\frac{n}{n-1}}$. That is,

$\sigma \approx \sqrt{\frac{n}{n-1}}\ s$ or, equivalently, $\sigma \approx \sqrt{\frac{n}{n-1}}\ \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$.

■ Summary

Sample standard deviation, $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}$.

Population standard deviation, $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$.

Approximation of population standard deviation using sample standard deviation, $\sigma \approx \sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$.

■ Yeah, so what's the confusing part?

Well, the mathematics is not the confusing part. The confusing part is that textbooks and the statistical literature get the above mixed up and various authors mix it up in various ways. So, there is no one mix-up that if it were unmixed would make everything OK. There is just a general fiasco.

■ What's a student to do?

Be aware that whatever textbook or statistics you read may well have the above all mixed up. Being aware of this, you will carefully investigate what the book you are reading means by the various symbols.

■ How can statisticians put up with this?

Well, they don't care! And, they do not care because in general it does not matter. Consider the factor $\sqrt{\frac{n}{n-1}}$ whose presence is the only difference between $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}$ and $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$. Ask yourself what happens to $\sqrt{\frac{n}{n-1}}$ as n becomes large. Golly, $\sqrt{\frac{n}{n-1}} \longrightarrow 1$. So for large values of n, $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}$ and $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$ are pretty much equal. So, really, the only people who would worry about this confusing mess would be students if they have the misfortune to be examined by a fussy instructor who thinks this is a big deal. (Your beloved teacher does not think it is a big deal, but he suspects the IB guys do.)

■ How large does n have to be for $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n}}$ and $\sqrt{\frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1}}$ to be practically the same?

Answer: 30. (Hey, I read it in a book.)

■ **I'm always getting confused by the symbols on my calculator for standard deviation. And what's worse, when I borrow a calculator of a different brand the symbols are different again from mine. For example, I see on the screen $\sigma_n = 1.325$ and $\sigma_{n-1} = 1.411$. But my friend's says for the same data, $\sigma = 1.325$ and $S = 1.411$.**

Remember that the population standard deviation equals $\sqrt{\frac{n}{n-1}}$ times the sample standard deviation. Now, $\sqrt{\frac{n}{n-1}}$ must be greater than 1, $n \in \{2, 3, 4, \cdots\}$. So, on your screen, the greater number is the approximate population standard deviation and the lesser number is the sample standard deviation. Always.