

The Perverse Nature of Standard Deviation

Denton Bramwell

Standard deviation is simpler to understand than you think, but also harder to deal with. If you understand it, you can use it for sensible decision making, while avoiding the foolishness that often goes on as a result of having it readily available.

The Grim and Theoretical Prelude

There are four things that will be helpful to you, if you understand them.

1. Standard deviation is one measure of how spread out your data are. The higher the standard deviation is, the more spread out your data are.
2. Standard deviation is hard to estimate with precision.
3. Changes in standard deviation are devilishly difficult to reliably detect.
4. Standard deviations don't simply add. This causes some processes to behave in ways that can be puzzling to the uninitiated.

Standard deviation is just a fancy way of averaging each point's distance from the mean. You take the distance from each point to the mean, square this number, add up all the squared distances (that's what we call the "sum of squares"), and divide by $n-1$, where n is the number of data in your sample. Once you have that number, you take the square root. In spite of all the squaring and square rooting, and subtracting 1 from the number of data, it's still just a dolled-up average distance from the mean.

For normally distributed data, that is, data that looks like a bell-shaped curve, 68% of cases will fall within one standard deviation of the mean, 95% within two, and 99.7% within three. If the standard deviation of your muzzle velocity is 30 fps, then 95% of your shots will be within plus or minus 60 fps of the mean.

For small collections of data, range (largest value minus the smallest, or what some shooters call "extreme spread") can be reliably converted to standard deviation. In fact, this estimate of standard deviation is probably a little better than the sum of squares method when n is small, since it does not have a tendency to underestimate variation in small samples, as the sum of squares method does.

If Your Sample is This Many Items	Divide Your Range by This to Get Standard Deviation
2	1.128
3	1.693
4	2.059
5	2.326
6	2.534
7	2.704

For example, if you shoot five shots, and your highest muzzle velocity is 2900 fps, and your lowest muzzle velocity is 2800 fps, then the standard deviation of your muzzle velocity is estimated by $(2900 - 2800)/2.326 = 42.99$ fps.

Practical Application #1: Evaluating Muzzle Velocities

Standard deviation is hard to estimate with precision.

When you shoot five shots, and measure something about them, you are taking a sample. From the sample, you hope to make some estimate of what the firearm will do in the long run, over many shots. Since you are dealing with a sample, your estimate of the long term performance will be imperfect, though it may be precise enough to be useful.

Suppose that a reloader is unwisely carried away with getting the standard deviation of his handloads down to nothing. He fires five shots, and chronographs each at 2960, 3002, 2982, 2976, and 2981 fps, calculates the standard deviation, 15.04 fps, and feels very pleased that his handloads are so consistent. But are they?

It is true enough that his sample of five has a standard deviation of 15.04, but what does that tell us about the long-term performance of his loading technique? If we repeated this same test 100 times, using exactly the same components and methods, then about 95 times out of that 100, we would find a standard deviation between 9.77 and 35.68. Statistically, we say that the true, long term standard deviation could easily be anywhere in that range. So, based on a sample of five, the shooter who thinks he has a superb standard deviation could actually have a standard deviation as high as 35.68, which is about typical for commercial ammunition. It was just his lucky day. The five shots he fired happened to be very close to each other, just by luck of the draw.

The reloader that does not recognize this can easily end up chasing phantoms. One day, he shoots test shots, and is very happy with his result. The next, things seem to have “gone to pot”, and he can’t figure out what he is doing wrong. The fact is that nothing has necessarily changed.

Practical Application #2: Evaluating Targets

Targets are a bit more complex than muzzle velocities, because a group is two-dimensional, not one-dimensional.

However, similar to the muzzle velocity case, group size is variation, and is a slippery devil to estimate using small samples. If you are using five-shot groups to evaluate group size, you should expect that groups will naturally vary plus or minus about 50% with no change whatever in the load, rifle, or shooter’s performance.

So our shooter goes to the range, and shoots a 5/8” group with one of his loads. Very pleased with himself, he tries another load, and gets 1 3/8”. Now, he’s frustrated, and

trying to figure out what's wrong with the second load. After all, his first group was much better, so he must have done something wrong.

Such joy and frustration is a waste of energy.

If the rifle is truly a 1" machine, then the shooter should expect that 95% of his five-shot groups will fall between ½" and 1 ½", **with absolutely no change in real performance.** You cannot reliably estimate the long-term characteristic of the rifle with just one or two five-shot groups. Groups within plus or minus 50% of the true long term average do not indicate any real change.

Practical Application #3: Evaluating Change

From rule 3, changes in variation are devilishly difficult to detect.

Our handloader has, through 25 test shots, estimated that his long-term standard deviation is actually closer to 25 fps. Our reloader carefully adjusts his reloading process, and fires another 25 shots, with a standard deviation of 20. Obviously, a nice improvement, right? The data don't support that conclusion.

One of the best tests for standard deviation change is the F Testⁱⁱ. In this test, we look at the ratio of the two standard deviations, squared. 25 shots in each test group is not enough to reliably detect a 4:5 ratio of standard deviations, as we have here. Something just shy of 100 in each of the two groups is required. Otherwise, what we think is real change might easily be normal random variation.

As a rule of thumb, 22 data in each group is needed to detect a 1:2 ratio between two standard deviations, about 35 data in each group are required for a 2:3 ratio, and about 50 in each group are required for a 3:4 ratio.

In a practical example, if you think you have lowered the standard deviation of your muzzle velocity from 20 to 15 fps, you'll need to chronograph 100 rounds, 50 from each batch. If the ratio of the two standard deviations comes out 3:4 or greater, you're justified in saying the change is real.

As if this were not a hard enough task, the F Test is notoriously sensitive to any non-normality in the data.

Given the twin burdens of large sample size and sensitivity to non-normality, it is much more difficult than most people expect to tell if a change in standard deviation is real, or if it was just our lucky day. Consequently, a lot of people perform tests, and draw conclusions on truly insufficient data.

If you're doing a really rigorous test, you need statistical software, such as QuikSigma™ or Minitab®, to evaluate changes in variationⁱⁱⁱ. Statistical software won't get you past

the sample size barrier, but it will help you make an intelligent evaluation of the results you get.

Practical Application #4: Stop Fiddling With Things That Don't Matter

Standard deviations don't simply add. If you don't understand the consequences of this fact, you might spend a lot of time working on process variables that don't matter. This is probably best shown with a personal illustration.

When I started handloading, I individually hand weighed the powder for my rifles.

My little 223 likes to give me five-shot groups of about ½" at 100 yards. It is a very dependable performer.

Some of my very early handloads had a muzzle velocity standard deviation of about 40 fps, which is not great. Commercial ammunition seems to be around 35 fps, more or less. My 223 didn't care. It just kept on giving me nice groups, for a sporter. I improved my methods, and found that with very little effort, I could get into the mid 20's, and with a bit of effort I could get into the teens.

So let's take the example of reloads done with modest attention to standard deviation, and use 25 fps as the standard deviation in muzzle velocity in our example.

At the moment, I like Varget in my 223. So I ran a little test on my Lee Perfect Powder Measure, to see how consistent it is with Varget. I dumped a large number of charges, and weighed each. The result was a standard deviation of .11 grains.

In that cartridge, a grain of powder is about 100 fps of muzzle velocity. So the standard deviation supplied by a .11 grain standard deviation of powder charge is equivalent to 11 fps in muzzle velocity. Now, we have to combine that variation with the existing 25 fps, because we're checking to see if hand weighing makes a real difference.

To combine the 25 fps and 11 fps, we square each, add them, and take the square root of the total.

Square root $(25^2 + 11^2) = 27.31$ fps.

For a 30-06, where a grain of powder is about 50 fps, the results are even less encouraging.

Square root $(25^2 + 5.5^2) = 25.59$ fps

So if I just use the powder dump, instead of individually hand weighing, the standard deviation of my muzzle velocity will increase from 25 to 27.31 fps or 25.59 fps, depending on which rifle I'm loading for. Detecting a change that small would require many hundreds of rounds. The day that I did this calculation, I quit individually weighing rifle loads. It's a total waste of time in my situation.

When you combine several sources of variation, if one source is much larger than the others, it alone will almost completely determine that total variation. There is practically no payout in working on the lesser sources of variation. You have to find the largest single source, kill it, and work your way down.

Conclusions:

1. Standard deviation is simply a glorified average of how far each point in your collection of data is from the mean of the data. Bigger standard deviations, and larger ranges indicate the data that is more spread out.
2. For samples as small as five or so, use range instead of standard deviation. For small samples, standard deviation will almost always underestimate variation.
3. Base estimates of standard deviation on small samples only if you are content to have a large amount of uncertainty in your estimate. It takes a lot of data to precisely estimate a standard deviation.
4. Do not interpret small changes in variation as real change, unless you have the large sample size required to support such a conclusion.
5. Evaluate firearm accuracy based on many groups. Do not be distracted by changes in group size that are within plus or minus 50% of your firearm's long term average group size. Such variations are completely explainable by nothing but normal random variation, and do not indicate any change in the firearm, loads, or shooting technique.
6. If you are trying to compare standard deviations, use ratios of standard deviations, and the sample sizes shown.
7. Since standard deviations add by the square root of the sum of the squares, the largest standard deviation will have disproportionate influence. One practical application of this is that if your powder dump is fairly consistent, and the standard deviation of your muzzle velocity is in the mid 20's, hand weighing each rifle charge is a waste of time. Another application is that if you're trying to improve something, the largest single source of variation must be isolated, and reduced. Working on the lesser sources is rarely worthwhile.

ⁱ Before 1925, everybody divided by n , but, then, Fischer came out with his book on ANOVA. $n-1$ makes the math in ANOVA work cleanly, and the influence of ANOVA on the science of statistics was profound. After 1925, the world pretty much switched to $n-1$. The choice of n or $n-1$ is simply a matter of convenience, anyway, and does not spring from some great, secret truth, known only to statisticians.

ⁱⁱ Named after Sir Ronald Fischer, who invented ANOVA.

ⁱⁱⁱ www.pmg.cc, www.minitab.com

© 2004 Denton Bramwell