Teaching Bits: "Random Thoughts on Teaching"

Deborah J. Rumsey
The Ohio State University

"Let's Just Eliminate the Variance"

The sample variance, s^2, is a common staple in the traditional introductory statistics course and textbook when presenting options to measure the amount of 'spread' or 'variability' within a data set. Once the formula, calculations, and examples involving the sample variance have been presented, one then moves on to the sample standard deviation, s, by taking the square root of the variance. And we never look back. From there we only use the standard deviation when calculating, measuring, interpreting, and comparing the amount of variability in one or more data sets.

But what do we leave behind for the students to sort out as a result of discussing sample variance? My answer is confusion. If you're like everyone else, you will normally get many (legitimate) questions such as "But what does variance really mean?"; "What's the difference between variance and standard deviation?"; "Which one should we use when you ask us to find an appropriate measure of spread for a data set?"; and the ultimate question: "Why do we even use variance anyway?" Indeed! My answer to the last question is: we shouldn't use it. We shouldn't just tolerate it as a step toward finding the standard deviation. We should just break with tradition, let go of the variance, and focus on the standard deviation.

As we all know, the formula for the sample variance is $s^2 = \dfrac{(x-\overline{x})^2}{n-1}$. The typical intuition behind it begins by saying we are looking to measure variation in the data around the mean and one way to do it is to find the sample variance. The steps involve taking the differences between each data value and the mean, squaring them to make them positive (so they don't all cancel out) and dividing the total by "n-1" which is "similar-to-but-not-exactly-the-same-as" finding the average. (Many discussions and arguments have ensued regarding the explanation of dividing by n-1, degrees of freedom, and all that, but that is a different topic.) In the end you get a result representing, in a very rough sense, the average squared distance from the mean (although that interpretation is also up for debate because it can be misleading for certain data sets.)

So what's the problem? The problem with variance is that the final result is in terms of square units. Unless you are measuring carpet, square units make no sense. If your data set involves exam scores, the variance is in terms of square points. If your data is temperatures in Fahrenheit, the variance is in terms of square degrees. And if your data represents number of people per family, the variance is technically in terms of square people (or people squared, which further confuses students into thinking you square the result and you stay in units of

people.) We could get around this issue by just not reporting any units for variance, but we want to make sure students stay within the context of the problem, and to me that means focusing on units. With standard deviation the units have meaning, so why bother with the variance?

In the end, sample variance merely confuses students and frustrates teachers who have to try to 'sell it' to their students. So why has it been incorporated into the intro stat courses and textbooks by default? One 'justification' is that that the calculations are easier to make and understand if you find the variance first, and then take the square root to find the standard deviation – which is ultimately what you are REALLY looking for. My response to this reasoning is two-fold. First, I'm betting that our students can take a square root as part of the calculations for standard deviation, without having to pause first and label s^2, and then take the square root. Any formula that involves the sums of squared differences divided by n-1 is certainly not complicated much more by adding a square root. Second, it is more important and meaningful to interpret standard deviation that has already been calculated by software rather than to calculate it by hand, so ease of calculation of the variance is no longer an issue.

I looked at the description of variance on Wikipedia, and found the following description, regarding the standard deviation (expected deviation) and variance of random variables. This description encapsulates the problem I have with the traditional approach.

> Unlike expected deviation, variance has different units from the variable, for example being in square inches when the variable is in inches. This inconvenience is eliminated with the notion of standard deviation of a variable as the square root of its variance, or root mean square deviation. In the dice example the standard deviation is $\sqrt{35/12} \approx 1.7$, slightly larger than the expected deviation of 1.5.

And what about notation? We use s^2
for sample variance and s for sample standard deviation, which makes it appear that you find the standard deviation first, then square it to get the variance. However, the calculations go the other way around of course. I'm not in favor of using s for variance and $\sqrt{s}$
for standard deviation, but our current notation only adds to the confusion. Better to just refer to s and leave it at that. (Variance diehards can then simply square the standard deviation and call it s^2.)

I'm not trying to eliminate variance from the entire field of statistics; it has a huge and justified role in theoretical and mathematical statistics for sure. But for the intro course, I would be willing to bet that teachers would gladly trade in the sample variance formula and focus on standard deviation any day, and I see no reason why they shouldn't. So if you are one of those teachers who has had to deal with sample variance simply as a means to an end, you have my permission to bypass it and just focus on the standard deviation.

Those are my random thoughts on teaching for this time around. Now what do YOU think?