QUESTIONS PUPILS ASK
Why do we divide by $n - 1$?

By **Colin Foster**

In these 'Questions pupils ask' articles, I often address unusual – even obscure – questions that pupils *sometimes* ask, because I find them interesting and thought-provoking for reflecting on teaching. But here I want to address a question that is not like this – I think this is a question *every* thoughtful student of statistics is bound to ask. But perhaps this is a feature of how I teach standard deviation, and if I taught it better maybe no one would ask this question?

Consider a small data set consisting of just 5 observations ($n = 5$):

1 7 8 9 15

These numbers have been carefully chosen to lead to integer mean and standard deviation. Although this is 'unrealistic', I think using 'nice' numbers like this when introducing ideas allows students to focus on the structure of what's going on, and hopefully build strong conceptual understanding, before applying what they have learned to the 'messiness' of more realistic situations later (Dudek, 1981; Read, & Riley, 1983).

Calculating the mean of these numbers is straightforward. There are 5 observations, and so we divide the sum by 5:

$$\frac{1 + 7 + 8 + 9 + 15}{5} = 8.$$

To find the variance, we first find the deviations (residuals) of each observation from this mean (8), and then square them and add them up. The sum of squared deviations is

$$(1 - 8)^2 + (7 - 8)^2 + (8 - 8)^2 + (9 - 8)^2 + (15 - 8)^2 = 100.$$

Now, we need to *divide by something*, because the more observations there are in our dataset the higher our sum of squared deviations is likely to be, and we would want to have a measure of variance that is independent of sample size. But exactly what number we divide by is the complicated part here, and this is the issue which leads to the question posed in the title.

If we are only interested in *this particular dataset* of 5 observations, then this is our entire population, and we simply divide the sum of squared deviations by 5. It is just a fact that their mean is 8, and the squared deviations from 8 sum to 100, and so the variance comes to $\frac{100}{5} = 20$. This is a descriptive, not inferential, statistic, and so there is no question of doing any statistical tests of significance, since we're not treating it as a sample taken from any larger population. It's a precise value, not an estimate of anything else, and it doesn't come with any error bounds. It just is what it is.

However, if we *do* view this dataset as a sample of 5 observations, pulled out randomly from some larger population, and our purpose in doing calculations with this sample is to be able to say something about the population from which it comes, then we're in the business of *inferential* statistics. In that case, we want an *unbiased estimate of the population variance*, meaning that if we keep pulling out these $n = 5$ samples, working out the variance of each one, and then averaging these variance values across all of these samples, we will get a value that gets as close as we wish to the true population variance as we take more and more of these $n = 5$ samples. (However close you decide you wish to get, you can always get closer than that just by taking enough $n = 5$ samples.) To do this, it turns out that instead of dividing our 20 by 5, we need to divide it by 4 (i.e., 5 – 1), and so we obtain a *sample variance* of 25 (and so a sample standard deviation of $\sqrt{25} = 5$). I find that this difference is generally puzzling to students: "Why do we divide by $n - 1$ instead of n?" I think they are right to be puzzled about this, and I find that it isn't the easiest thing to explain.

The problem for me is that the concepts that we might want to draw on to explain this discrepancy (degrees of freedom, unbiased estimators) are harder than the concepts that we are trying to introduce (variance, standard deviation). Students can build a good intuitive sense of standard deviation as 'measure of spread' – just like the range, except that it takes account of *all* of the data points, not just the most extreme ones at each end. It can be a nice task to devise data sets with identical ranges but for which the standard deviations are dramatically different. This shows the added value of a standard deviation over a simple calculation of range, which could be entirely dependent on two highly untypical outliers (the most unrepresentative data points). But this nice thinking easily gets side-tracked into concerns about dividing by n versus

dividing by $n-1$ and attempts at justifying this. Students might wonder how it can possibly matter *why* we are working out the variance. How can we have two different formulae, depending on what we are planning to do with this value that we calculate? And why didn't we have this problem with the sample mean?

The way I like to think about this is to appreciate that once we get into inferential statistics there are now *two* means in play. There's the population mean, μ, which we treat as a fixed but unknown quantity, and there's the sample mean, $\bar{x}$, of our particular sample. If we are lucky, $\bar{x}$ will be close to μ. This is more likely to be the case if we have a large sample, so, with a sample of just 5 observations, it could well be that there is a bit of a difference between $\bar{x}$ and μ. Although we know this, we of course have no idea whether our $\bar{x}$ is less than or greater than the true μ. However, because the mean is an unbiased estimator, if we take *lots* of $n=5$ samples, calculate their means, and average these values, we will get arbitrarily close to the population mean μ as we take more and more of these $n=5$ samples. I tend to keep saying "$n=5$ samples" because students can be confused between taking *lots of* samples and taking *larger* samples, both of which improve our estimate of the mean, but are distinct.

The problem with the variance comes because we want to calculate the deviations of each data point 'from the mean' – but which mean? Is it $\bar{x}$ or μ? We would like it to be μ, but we don't know μ, and so we have to use $\bar{x}$, as that's all we know. So now when we calculate the mean of our sample, and get 8, this is *not* necessarily the *population* mean – it's our best *estimate* of the population mean, but it's probably wrong. It's all we have, so we have to use it, and, so when we calculate

$$(1-8)^2+(7-8)^2+(8-8)^2 \\ +(9-8)^2+(15-8)^2=100,$$

in each bracket we are subtracting the *sample* mean $\bar{x}=8$.

Why should this matter? It's probably close to the population mean, after all. The issue is that when we get to the final bracket in our sum of squares, the 15 isn't providing any new information. In the expression below, the blank box *has to be* 15, because we can work it out from the other data points and the 8, which we've had to calculate already:

$$(1-8)^2+(7-8)^2+(8-8)^2 \\ +(9-8)^2+(\blacksquare-8)^2.$$

Even if we just know that *four* of the data points are 1, 7, 8 and 9, because we know that the mean is 8, we can calculate what the final data point must be:

$$\blacksquare=5\times 8-(1+7+8+9)=15.$$

This means that the fifth data point (it could equally apply to any other of the data points) is bringing in no new information, once you've worked out that the mean is 8. So, the number of *degrees of freedom* here is 4, not 5, and this means that 4 is the appropriate number to divide by. Once you've calculated the 8, you've used up one of your degrees of freedom, and so you only have 4 independent pieces of information left. We can think of this as the 'cost' of pretending that the sample mean is really the population mean. Another way to say this is that although there are n independent observations in the original sample, there are only $n-1$ independent *deviations from the mean*, as those deviations necessarily (by definition) have a sum of zero. (We could equally work out $\blacksquare$ by summing the deviations from the other four values. We can work out that -7, -1, 0 and 1 sum to -7, and so the final deviation must be $+7$, which means that $\blacksquare-8=7$ and so $\blacksquare=15$, as before.)

I still tend to find students unconvinced at this point. They might agree that the number of degrees of freedom is 4 rather than 5, but why do we divide by the number of degrees of freedom, rather than the total number of data points? So I also like to offer a more qualitative argument (Note 1). It is fairly intuitive that deviations from the sample mean are likely to be *smaller* than they would be from the true population mean. After all, the sample mean has been calculated to be precisely the mean of this particular data set, whereas the population mean in general won't be the same. So the variance will almost always be *smaller* when it's calculated by using the sum of the squared deviations from the *sample* mean than when it's calculated using the sum of the squared deviations from the true *population* mean. So, the variance we calculate from our sample is on average going to be smaller than it should be, and it turns out that we can correct this by dividing the sum of squared deviations by a *smaller* number than n, so that the answer comes out *bigger*. Of course, this doesn't tell us that $n-1$ is the right number to divide by (as it could be $n-\frac{1}{2}$ or something, or it could even vary with n), but it at least points us in the right direction, and highlights that dividing by is certainly *not* going to give us the best estimate for, n the population variance. It means that the 'sample variance' (i.e., the *unbiased* estimate, with the $n-1$ denominator) is *not* the 'sample's variance' (i.e., the *biased* estimator, with the denominator of n)!

A formal proof involves carefully exposing the distinctions between things like $E(X^2)$ and $E(\bar{X}^2)$, and is quite algebra-heavy (see Toller, 2009, p 18, for a careful treatment). But we can extend this qualitative argument to make it a bit more quantitative while avoiding the fiddly manipulation. When we use the sample mean instead of the population mean, the magnitude of each residual is going to be wrong by $|\mu-\bar{x}|$. This means that the sum of the squares of the residuals will be too low by $n(\mu-\bar{x})^2$, and so the difference between the expected values of the biased and unbiased variances will be the expected value of $(\mu-\bar{x})^2$. This is just the variance of the sampling distribution of the mean, which is $\frac{\sigma^2}{n}$, and so, when we calculate the variance of the

sample (the biased estimator), we underestimate $\widehat{\sigma^2}$ by $\frac{\widehat{\sigma^2}}{n}$:

$\widehat{\sigma^2} - \frac{\widehat{\sigma^2}}{n} =$ biased estimator.

Rearranging,

$\widehat{\sigma^2}\left(1 - \frac{1}{n}\right) =$ biased estimator,

$\widehat{\sigma^2}\left(\frac{n-1}{n}\right) =$ biased estimator,

$$\widehat{\sigma^2} = \left(\frac{n}{n-1}\right) \times \text{biased estimator.}$$

The variance of the sample underestimates the variance of the population, so we need to scale it up by $\frac{n}{n-1}$ (which is always greater than 1), and this is known as *Bessel's correction* (Note 2).

Finally, it can be reassuring to see the difference in a simulation. For example, if you use *R* to take 10^4 samples of size 5 at random from a normal distribution with mean 8 and standard deviation 5, and calculate the variance of each sample, using both the biased estimator (with denominator n) and the unbiased estimator (with denominator $n - 1$), you can compare the two distributions obtained (Figure 1, Note 3). For this simulation, the mean of the variances comes to 20.07 for the biased variances and 25.09 for the unbiased variances (correct to 2 decimal places), nicely illustrating that the biased estimator is lower than the true population value of 25, and the quotient $\frac{25}{20} = \frac{5}{4}$ exactly matches Bessel's correction.

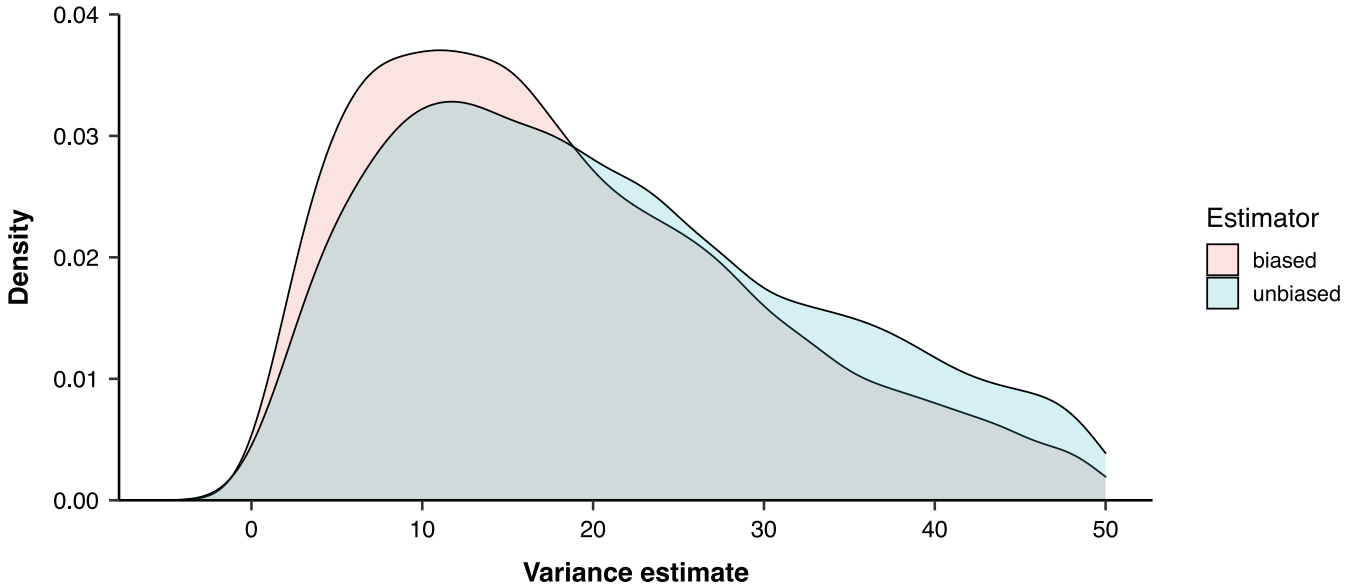

Figure 1: The variance of 10^4 samples with $n = 5$ calculated using both n (biased) and $n - 1$ (unbiased) as denominators.

Notes

1. Another informal argument is that, in the case of an $n = 1$ sample, we should really be *unable* to calculate a variance, and so the division by zero that would result from an $n - 1$ denominator seems appropriate. Using a denominator of n would give us a value that could have no meaning (Rosenthal, 2015). (However, you might argue that zero is a perfectly reasonable answer when $n = 1$.)
2. Note that, although Bessel's correction produces an unbiased estimate of the variance, its square root is *not* an unbiased estimator of the population *standard deviation* (and the bias in the standard deviation is that it is always an *under*estimate).
3. For the *R* code, see: **www.foster77.co.uk/Simulation%20of%20biased%20and%20unbiased%20variance%20estimators.R**

Acknowledgement

I am extremely grateful to Owen Toller for some very helpful comments on a draft of this article.

References

Dudek, F. J. 1981 'Data sets having integer means and standard deviations', *Teaching of Psychology* **8**(1), pp. 51-51.

Read, K. L. Q., & Riley, I. S. 1983 'Statistics problems with simple numbers', *The American Statistician* **37**(3), pp. 229-231.

Rosenthal, J. S. November 17, 2015 'The kids are alright: Divide by *n* when estimating variance' [blogpost]. https://imstat.org/2015/11/17/the-kids-are-alright-divide-by-n-when-estimating-variance/

Toller, O. 2009 *The Mathematics of A-level Statistics*. Mathematical Association.

Keywords: Degrees of freedom, Sample variance, Standard deviation, Unbiased estimators, Variance

Author: Colin Foster, Department of Mathematics Education, Schofield Building, Loughborough University, Loughborough LE11 3TU.

Email: **c@foster77.co.uk**

website: **www.foster77.co.uk**

blog: **https://blog.foster77.co.uk/**