MAC-CPTM Situations Project

Situation 67: Sample and Population Variance

Prepared at University of Georgia
Center for Proficiency in Teaching Mathematics
15 July 2008 – Ken Montgomery
04 December 2008 – Sarah Donaldson
26 January 2009 – Sarah Donaldson
27 February 2009 – Sarah Donaldson

Prompt

A student in a statistics class has observed that the definition of sample variance $\left(s^2\right)$ is the average of the squares of the deviations from the mean of the data set.

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}$$

Yet, to compute the average of these squared deviations, they are summed and then divided by (*n*-1). “Why,” asked the student, “do you not divide by *n*? Is this an actual mean or some sort of ‘pseudo-mean’?”

Commentary

The concept in statistics known as variance is closely related to standard deviation: both indicate the spread of the data distribution about the mean. In fact, the standard deviation is simply the square root of the variance. There are a number of reasons why the variance is calculated the way it is. As the formula for sample variance shows, the sum of the squared deviations is divided by *n* – 1. In the following Foci, we investigate reasons why division by *n* – 1, rather than division by *n*, is necessary to calculate the sample variance.

Mathematical Foci

Mathematical Focus 1

Since the deviations sum to zero, only n – 1 of the data values may “vary freely.”

One way to think about a data set is in terms of variability. That is, we might consider how much the data points in a sample differ (or deviate) from the sample mean, which may be interpreted as the balance point of the distribution of the data. The deviation of a single data point, x, from the sample mean, $\bar{x}$, is simply the difference $x - \bar{x}$. The sum of all these differences must be 0 since the sample mean, $\bar{x}$, is the balance point. This can be seen algebraically. For example, consider a data set of 3 values, x_1, x_2, and x_3. The value of $\bar{x}$ is the arithmetic mean of these values: $\bar{x} = \frac{x_1 + x_2 + x_3}{3}$. This can be rewritten as $x_1 + x_2 + x_3 = 3\bar{x}$. Now consider the sum of the deviations: $(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x})$. This can be rewritten as $x_1 + x_2 + x_3 - 3\bar{x}$. Since $x_1 + x_2 + x_3 = 3\bar{x}$ (see above), $x_1 + x_2 + x_3 - 3\bar{x} = 0$, so the sum of the deviations is 0: $(x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) = 0$.

We can use the fact that the deviations sum to 0 to think about the role of n – 1 in the formula for sample variance. Since the deviations sum to 0, if we know all but 1 of them (that is, if we know n – 1 of them), the value of the last (i.e. the n^{th}) term is determined. In other words, only n – 1 of the data values may “vary freely.” In statistics this is known as *degrees of freedom*: the number of data values that have “freedom” to vary such that the mean remains the same.

For example, suppose we had a data set of 5 values and we know that $\bar{x}$ = 9. There are many sets of 5 data values that have a mean of 9, but as soon as we know 4 of them, the 5th is determined. One possibility is that 4 of the data values are 5, 8, 12, and 13, as shown below.

x	$\bar{x}$	$x - \bar{x}$
5	9	-4
8	9	-1
12	9	3
13	9	4
?	9	

Given these 4 data values and their mean, we can determine the last data value. Since the sum of the deviations is 0, we can find the last deviation, d:
-4 + -1 + 3 + 4 + d = 0
d = -2
Therefore the 5th data value is 7 since 7 – 9 = -2.

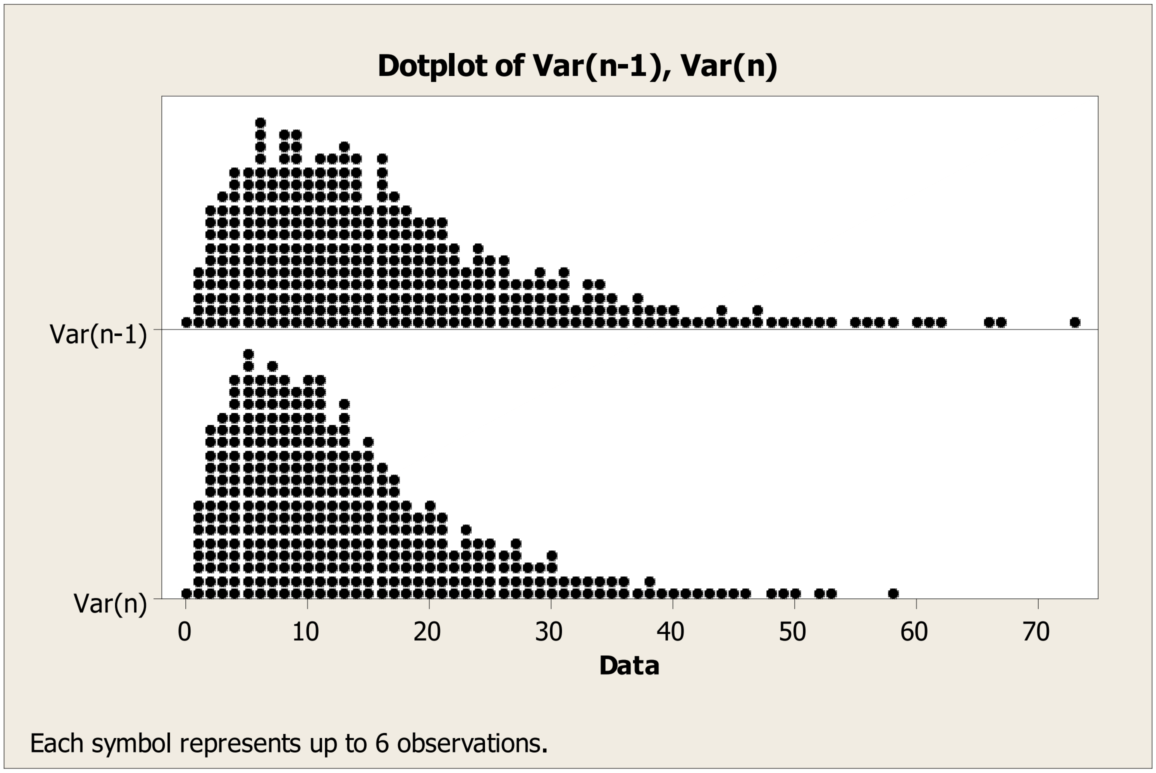
Why can we then conclude that the sample variance should be calculated by dividing by n – 1? It is because we want to divide by the number of data values that are independent (can vary freely)—i.e. we divide by the degrees of freedom. By dividing by the degrees of freedom, the sample variance is an unbiased estimate of the population variance.

Mathematical Focus 2

The sample variance is an unbiased estimator of the population variance.

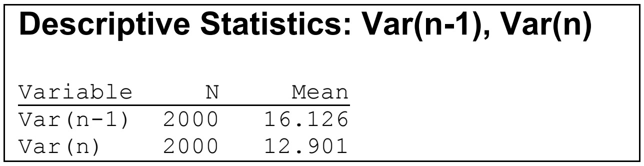
If the expected value of the sample mean, $\bar{x}$, is the same as the value of the mean, μ, of the population from which the sample was taken, we say that the sample mean is an *unbiased estimate* of the population mean. We may also talk about bias as it relates to the *variance* of a sample—that is, the biasness of the sample variance, s^2, as an estimate of the population variance, σ^2. In calculating the sample variance, dividing the sum of the squared deviations by n may seem intuitive, but doing so will result in the sample variance being a biased estimate of the population variance. In order for the sample variance to be an *unbiased* estimate of the population variance, we must divide by $n - 1$ (called the degrees of freedom—see Focus 1 and the Post Commentary).

The following example illustrates what would happen if we were to divide by n when calculating the sample variance, rather than dividing by $n - 1$. The figure below shows two dot plots—one in which $n - 1$ was used in the calculation (the first plot, indicated by "Var(n-1)" and one in which n was used (indicated by "Var(n)").



In this example, a computer program was used to simulate taking 2000 samples, with the size of each sample being $n = 5$. Note that this is a relatively low value for n. This means the difference between n and $n - 1$ is large enough to illustrate the point. (For very large values of n, the difference between n and $n - 1$ is minute.) For the data in this example, it is known that the population variance, σ^2, is 16. So to be an unbiased estimate of this population variance, we expect the value of the sample variance to be $s^2 = 16$. That is, the average value of the variance (i.e. the average of all the data points in the dot plot above) should be at or near 16.

The statistics below indicate the actual average of these sample variances when the variance is calculated using $n - 1$ and when it is calculated using n.



Descriptive Statistics: Var(n-1), Var(n)

```
Variable      N      Mean
Var(n-1)   2000    16.126
Var(n)     2000    12.901
```

We can see that the average sample variance when $n - 1$ is used is s^2 =16.126 and when n is used, the average sample variance is much too low: $s^2 = 12.901$. Clearly the calculation using $n - 1$ gives the better estimate of the true population variance of $\sigma^2 = 16$, whereas a calculation using n will, on average, underestimate the population variance.

Post Commentary

We can investigate the topic of an unbiased estimate by further considering the concept of *expected value*. This may shed light on why we divide by *n* – 1 in calculating the variance.

Definition: A statistic $\hat{\Theta}$ is an unbiased estimator of the parameter θ if and only if $E\left(\hat{\Theta}\right) = \theta$ (in other words if the expected value of the statistic equals the parameter).

Theorem: If S^2 is the variance of a random sample from an infinite population with the finite variance σ^2, then $E\left(S^2\right) = \sigma^2$.

Proof:

$$E(S^2) = E\left(\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}\right) = \frac{1}{n-1}\cdot E\left(\sum_{i=1}^{n}\left[\left(x_i - \mu\right) - \left(\bar{x} - \mu\right)\right]^2\right)$$

$$= \frac{1}{n-1}\cdot\left\{\sum_{i=1}^{n} E\left[\left(x_i - \mu\right)^2\right] - n\cdot E\left[\left(\bar{x} - \mu\right)^2\right]\right\}$$

Then, because $E\left[\left(x_i - \mu\right)^2\right] = \sigma^2$ and since $E\left[\left(\bar{x} - \mu\right)^2\right] = \frac{\sigma^2}{n}$, we have that

$$E\left(S^2\right) = \frac{1}{n-1}\cdot\left\{\sum_{i=1}^{n}\sigma^2 - n\cdot\frac{\sigma^2}{n}\right\} = \sigma^2$$

and thus, by the definition, the sample variance is an unbiased estimator of the population variance.■

References

Agresti, A. & Franklin, C. (2007). *Statistics: The art and science of learning from data.* Upper Saddle River, NJ: Pearson/Prentice Hall.

Freund, J. E. (1992). *Mathematical Statistics* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Mendenhall, W., Beaver, R. J. & Beaver, B. M. (2005). *Introduction to Probability and Statistics (12th ed.).* Pacific Grove, CA: Thomson Brooks/Cole.