The Sample Variance Formula: A Detailed Study of an Old Controversy

Ky M. Vu, PhD. AuLac Technologies Inc. ©2010
Email: kymvu@aulactechnologies.com

Abstract

The two biased and unbiased formulae for the sample variance of a random variable are under scrutiny. New research result proves that the formula with a smaller divisor is unbiased and the formula with a larger divisor is biased. This fact agrees with the current belief in statistics literature. Many mathematical proofs for both formulae contain errors: This is the reason for the controversy and this new research. The final verdict comes from simulation results.

Keywords: Estimator, expectation operator, Helmert's transformation, sample variance.

1 Introduction

Statisticians know, for a long time, that there are two formulae for the sample variance of a random variable, different by a divisor. For decades, a number of statisticians seem to be happy with a formula, called the unbiased formula, with a smaller divisor. This formula is correct but the proofs for its unbiasedness by a number of authors are incorrect; for example, see P.G. Hoel (1971) and I. Guttman, S.S. Wilks and J.S. Hunter (1971). The other formula with a larger divisor, called the biased formula, is believed wrong but supported by a smaller number of statisticians; for example, see R.V. Hogg and A.T. Craig (1978). In this paper, we will give a detailed study of the formulae with their supported arguments. The paper is organized as follows. Section one is the introduction section. In section two, we present the formulae for the sample variance. In section three, we present three incorrect proofs, collected from textbooks, for the unbiased formula. In section four, we present some arguments for the biased formula. In section five, we discuss the Helmert's transformation and its application to the sample variance formula. In section six, we report the simulation results, which give the verdict as to which formula is the right one to use. Finally, section seven concludes the discussion of the paper.

2 The Formulae for the Sample Variance

Suppose that we have a random variable X has a distribution with mean μ and variance σ^2, which gives values x_t's $(t = 1 \cdots n)$ in a sample. The estimated value for the first moment, known as the mean, is given as

$$\bar{x} = \frac{1}{n}\sum_{t=1}^{n} x_t. \tag{1}$$

There is no controversy about this formula. There is, however, a controversy for the estimator of the second moment about the mean, known as the variance. There are two estimators known by two formulae:

$$\hat{\sigma}^2_{unbiased} = \frac{1}{n-1}\sum_{t=1}^{n}(x_t - \bar{x})^2 \tag{2}$$

and

$$\hat{\sigma}^2_{biased} = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2 \tag{3}$$

for this statistic, called the sample variance.

By developing the square in the last equation, we can write

$$\begin{aligned}
\hat{\sigma}^2_{biased} &= \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2, \\
&= \frac{1}{n}\sum_{t=1}^{n}(x_t^2 - 2x_t\bar{x} + \bar{x}^2), \\
&= \frac{1}{n}\left[\sum_{t=1}^{n} x_t^2 - 2n\bar{x}^2 + n\bar{x}^2\right], \\
&= \frac{1}{n}\left[\sum_{t=1}^{n} x_t^2 - n\bar{x}^2\right], \\
&= \frac{1}{n}\left[\sum_{t=1}^{n} x_t^2 - n(\frac{1}{n}\sum_{t=1}^{n} x_t)^2\right], \\
&= \frac{1}{n}\sum_{t=1}^{n} x_t^2 - (\frac{1}{n}\sum_{t=1}^{n} x_t)^2.
\end{aligned}$$

From the last equation, statisticians call the variance the second moment about the mean.

3 Arguments for the Unbiased Formula

In this section, we will cite three proofs from proponents for Equation (2). In one old textbook, which the author of this paper can no longer find, there is an argument as follows. If we take the expectation of both sides of the last

equation in the last section, we have

$$\begin{aligned} E\{\hat{\sigma}^2_{biased}\} &= E\{\frac{1}{n}\sum_{t=1}^{n} x_t^2 - \left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)^2\}, \\ &= E\{\frac{1}{n}\sum_{t=1}^{n} x_t^2\} - E\{\left(\frac{1}{n}\sum_{t=1}^{n} x_t\right)^2\}, \\ &= \frac{1}{n}\sum_{t=1}^{n} E\{x_t^2\} - \frac{1}{n^2} E\{\left(\sum_{t=1}^{n} x_t\right)^2\} \quad (4) \end{aligned}$$

If x_t's are independent observations, the expected values of the cross-products of the second term on the right hand side of the above equation are zeros and we obtain

$$\begin{aligned} E\{\hat{\sigma}^2_{biased}\} &= \frac{1}{n} n\sigma^2 - \frac{1}{n^2} n\sigma^2, \\ &= \sigma^2 - \frac{1}{n}\sigma^2, \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

The second proof is taken from P.G. Hoel (1971) , pages 191-192. From the properties of E and the definition of σ^2, it follows that

$$\begin{aligned} E\{\hat{\sigma}^2_{biased}\} &= E\left[\frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2\right], \\ &= E\left[\frac{1}{n}\sum_{t=1}^{n}[(x_t - \mu) - (\bar{x} - \mu)]^2\right], \\ &= E\left[\frac{1}{n}\sum_{t=1}^{n}(x_t - \mu)^2 - (\bar{x} - \mu)^2\right], \\ &= \frac{1}{n}\sum_{t=1}^{n} E(x_t - \mu)^2 - E(\bar{x} - \mu)^2, \\ &= \frac{1}{n}\sum_{t=1}^{n} \sigma^2 - \sigma^2_{\bar{x}}, \\ &= \sigma^2 - \frac{1}{n}\sigma^2, \\ &= \frac{n-1}{n}\sigma^2. \end{aligned}$$

The first proof does not use the mean and its estimator, but the second proof uses both. The expected value of the estimator or sample variance $\hat{\sigma}^2_{biased}$ is not the same as the population parameter σ^2, so the proofs claim that the estimator $\hat{\sigma}^2_{biased}$ is biased, hence the reason for its name.

The following argument in another textbook, I. Guttman, S.S. Wilks and J.S. Hunter (1971) pages 181-182, is a more laborious argument and stronger proof for the formula given by Equation (2).

Suppose that $(x_1, \cdots, x_n)$ is a random sample of a random variable X having mean μ and variance σ^2. Then, the random variable

$$L = c_1 x_1 + \cdots + c_n x_n$$

where the c_i's are constants, has the expectation $\mu \sum_{i=1}^{n} c_i$ and variance $\sigma^2 \sum_{i=1}^{n} c_i^2$. In the special case consider

$$\bar{x} = \frac{1}{n}x_1 + \cdots + \frac{1}{n}x_n,$$

we obtain

$$\begin{aligned} E\{\bar{x}\} &= \mu, \\ Var(\bar{x}) &= \frac{\sigma^2}{n}. \end{aligned}$$

The last equation states that

$$E\{n(\bar{x} - \mu)^2\} = \sigma^2.$$

Now, consider the algebraic identity

$$\sum_{i=1}^{n}(x_i - \mu)^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu)^2.$$

Using properties of the expectation operator, we have

$$\sum_{i=1}^{n} E(x_i - \mu)^2 = E\{\sum_{i=1}^{n}(x_i - \bar{x})^2\} + En(\bar{x} - \mu)^2. \quad (5)$$

And making use of a previous result, we find

$$\sum_{i=1}^{n} \sigma^2 = E\{\sum_{i=1}^{n}(x_i - \bar{x})^2\} + \sigma^2,$$

and that means

$$E\{\sum_{i=1}^{n}(x_i - \bar{x})^2\} = (n-1)\sigma^2.$$

Finally, we have

$$E\{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\} = \sigma^2. \quad (6)$$

The first and second proofs try to prove that the estimator $\hat{\sigma}^2_{biased}$ is biased; the third proof, the estimator $\hat{\sigma}^2_{unbiased}$ is unbiased. Unfortunately, all the proofs are wrong as the discussion in the next section will reveal.

4 Arguments for the Biased Formula

In this section, we will give some arguments for the formula given by Equation (3). These arguments are listed below.

4.1 Common Sense

Once in a while, a mathematician gives a proof for his thinking on some problem that is so contrary to common sense thinking, and he believes in his proof. But unfortunately, his proof fails him and does not solve the problem. The story of the Greek mathematician Zeno of Elea (490-430BC) with the solved paradox of *Achilles and the Tortoise* is one example. The story of the sample variance formula might be another.

From common sense thinking, we know that to obtain the variance we must divide the sum of squares of the values deviated from the mean by the total number of values.

4.2 The Probability

In R.V. Hogg and A.T. Craig (1978) pages 124-125, the authors of the reference advocated for the variance formula given by Equation (3). These authors explained that the ratio $1/n$ is the probability of each event for the random variable X to have the value x_t. With such an insight, it is too clear to see that the correct sample variance formula must have the divisor n, not $(n-1)$. This is because the formula is a reflection of an expectation operator E, which is a probability-weighted average. However, these authors could neither prove their preferred formula nor discredit the other one.

Using the probability concept, K. Vu (2007) claimed that the expectation operator E can be represented as

$$E\{()_t\} = \lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n}()_t$$

where $()_t$ is a function of a random variable with the index t. Let us apply this technique to find the unbiased estimator for the mean μ of a distribution. We have

$$\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} x_t = \lim_{n\to\infty} \sum_{t=1}^{n} \frac{1}{n} x_t.$$

Now assume that in the sample of n values x_t's, there is a value x and k_x is the number of times x_t's have this value. Then we can write the last equation as below

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} x_t &= \lim_{n\to\infty} \sum_{t=1}^{n} \frac{1}{n} x_t, \\
&= \lim_{n\to\infty} \sum_{x} \frac{k_x}{n} x, \\
&= \lim_{n\to\infty} \sum_{x} x\ Pr\{X = x\}, \\
&= E\{X\}, \\
&= \mu.
\end{aligned}$$

When the value n approaches infinity, the range of x expands to cover the range of the random variable X, and at the value of infinity the ratio k_x/n simply becomes the probability density for the random variable X to have the value x. And depending on the nature of discrete or continuous type of the random variable X, the sum on the right hand side of the last equation will remain a sum or change to an integral. In either case, the right hand side simply becomes the expected value of x_t. For this to be true, however, the ratio k_x/n must be a probability value.

Using this approach, we can find the unbiased estimator for the variance easily. We are now ready to present and prove our research result, in a theorem.

Theorem 4.1 *The unbiased estimator for the variance of some distribution of a random variable X with probability density function $p(x)$, mean μ, variance σ^2 and sample values x_t's $(t = 1, \cdots n)$ is*

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2$$

where $\bar{x}$ is the unbiased mean estimator given by Equation (1).

Proof. *From the equation of the estimator, we write*

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n}\sum_{t=1}^{n}(x_t - \bar{x})^2, \\
&= \sum_{t=1}^{n}\frac{1}{n}(x_t - \bar{x})^2. \qquad (7)
\end{aligned}$$

Now assume that in the sample of n values x_t's, we have some x_t's with the value x and the number of these x_t's is k_x times. Since there is only a finite number of values of x, we can write

$$\begin{aligned}
\hat{\sigma}^2 &= \sum_{x}\frac{k_x}{n}(x - \bar{x})^2, \\
&= \sum_{x}(x - \bar{x})^2\{Pr\ X = x\}, \\
&= \sum_{x}(x - \bar{x})^2 p_x.
\end{aligned}$$

By taking the limit when n approaches infinity, the last equation becomes

$$\begin{aligned}
\sigma^2 &= \sum_{x}(x-\mu)^2 p(x), \quad if\ X\ is\ discrete, \\
&= \int (x-\mu)^2 p(x)dx, \quad if\ X\ is\ continuous, \\
&= E\{(X - E(X))^2\}.
\end{aligned}$$

At the limit of infinity, the sample becomes the population. Since at the limit the estimator gives the definition of the population variance, the estimator must be an unbiased estimator. This fact proves the theorem.

QED

Having proven that the formula with the divisor n is unbiased, we must now prove that the formula with the divisor $(n-1)$ is biased or the proofs in the last section prove nothing. This is what we are going to do next.

4.3 The Errors of the Incorrect Proofs

Now we will investigate the proofs in the last section. The first proof does not use the mean μ and its estimator $\bar{x}$ as the second and third proofs, so it does not make the mistake caused by the estimator, but it has a flaw that is easy to find. The flaw is about the definition of the variance. The variance is defined as the second moment about the mean not about zero. This means that $\sigma^2 \neq E\{x_t^2\}$. The argument has $E\{x_t^2\} = \sigma^2$. This can be only true if the values x_t's

in the argument have a zero mean. However, if the values x_t's have a zero mean, then the second term in Equation (4) will be zero. This will result in the correct expected value σ^2 for the right hand side of Equation (4). The author of the argument set out to prove something that is correct but obtained a wrong result.

While the second and third proofs appear to be mathematically flawless, they actually prove nothing. Their faults are in using the value $\bar{x}$. The errors in the second and third proofs is in the probability density distributions of X and $\bar{X}$. They are different. Both have mean μ; but the variances are σ^2 and σ^2/n, respectively. Assuming that the probability density distribution of $\bar{X}$ is $p(\bar{x})$; then to obtain the variance about the mean of this variable, we must write

$$\begin{aligned} E\{(\bar{x}-\mu)^2\} &= \int (\bar{x}-\mu)^2 p(\bar{x}) d\bar{x}, \\ &= \frac{\sigma^2}{n}. \end{aligned}$$

We only investigate the case when $\bar{X}$ is of continuous type; because it is the usual case for $\bar{x}$ to be the average of n values x_t's. But the argument applies to the case when $\bar{X}$ is of discrete type as well. Now if we apply the same procedure, *taking expectation*, to the quantity $(x_t-\mu)^2$, we get

$$\begin{aligned} E\{(x_t-\mu)^2\} &= \int (x_t-\mu)^2 p(\bar{x}) d\bar{x}, \\ &\neq \sigma^2. \end{aligned}$$

This is because x_t has a wrong probability density value. The result will lead to

$$E\{\frac{1}{n-1}\sum_{t=1}^{n}(x_t-\bar{x})^2\} \neq \sigma^2,$$

which is contrary to what Equation (6) claims.

4.4 The Number Degrees of Freedom

Another argument for the unbiased formula comes from the number of degrees of freedom. Proponents for this formula claim that the sum of squares $\sum_{t=1}^{n}(x_t-\bar{x})^2$ has only $(n-1)$ degrees of freedom because one is lost with the calculation of the average $\bar{x}$. In n values $(x_t-\bar{x})$, $t=1,\cdots n$, there is one that does not have a free value; it can be calculated from the other $(n-1)$ values. This argument appears to be correct first, and it has many statisticians, including the author of this paper, under its spell for some time. But it has an error like the formula it supports. The question here is: How many degrees of freedom are there in the sum of squares $\sum_{t=1}^{n}(x_t-\bar{x})^2$? To answer this question, we write the sum of squares as

$$\sum_{t=1}^{n}(x_t-\bar{x})^2 = \sum_{t=1}^{n}x_t^2 - n(\frac{1}{n}\sum_{t=1}^{n}x_t)^2$$

and count the number of degrees of freedom of the quantity on the right hand side of the last equation. The number of degrees of freedom must be n. Looking at the sum on the left hand side, we accept the fact that it has only $(n-1)$ free quantities $(x_t-\bar{x})$. However, each of these quantities has two degrees of freedom: One is given by x_t; the other, $\bar{x}$. Therefore, the sum should have $2(n-1)$ degrees of freedom. However, we cannot count one degree of freedom more than once. The result is: We have a total of n degrees of freedom. This is why when we remove $\bar{x}$, the sum shows that it has n degrees of freedom as the right hand side tells us.

5 The Helmert Transformation

A brilliant transformation that can be used to solve the variance formula problem is the Helmert's transformation. It is an attempt to show the correct number of degrees of freedom in the sum of squares of a sample variance formula. If we apply the Helmert's transformation to the variables $x_t's$ $(t=1,\cdots n)$, we obtain a new set of variables as follows:

$$\begin{aligned} y_1 &= \frac{x_1-x_2}{\sqrt{1\times 2}}, \\ y_2 &= \frac{x_1+x_2-2x_3}{\sqrt{2\times 3}}, \\ &\vdots \ , \\ y_{n-1} &= \frac{x_1+x_2+x_3+\cdots-(n-1)x_n}{\sqrt{(n-1)n}}, \\ y_n &= \frac{x_1+x_2+\cdots+x_n}{\sqrt{n}} = \sqrt{n}\bar{x}. \end{aligned}$$

The new variables y_i's $(i=1,\cdots n-1)$ have a zero mean and the same variance σ^2 of the variable x_i.

The Helmert's transformation is an orthogonal transformation; therefore, the Jacobian of the transformation is one and we have

$$\sum_{t=1}^{n}x_t^2 = \sum_{t=1}^{n}y_t^2.$$

Therefore, we can write

$$\begin{aligned} \sum_{t=1}^{n-1}y_t^2 &= \sum_{t=1}^{n}y_t^2 - y_n^2, \\ &= \sum_{t=1}^{n}x_t^2 - n\bar{x}^2, \\ &= \sum_{t=1}^{n}(x_t-\bar{x})^2, \\ &= S. \end{aligned}$$

Since the sum S has $(n-1)$ independent variables y_i's $(i=1,\cdots n-1)$, each with variance σ^2, we can obtain the variance for the random variable by taking S divided by $(n-1)$. This means that this is another argument for the unbiased formula (2).

6 Simulation Results

As there are n x_i's $(i = 1, \cdots n)$ in the sum of squares S, we can still argue that the sum has n degrees of freedom (an earlier argument). If the variance can be considered as a sum of squares divided by its number of degrees of freedom, the divisor for the sample variance formula should be n. This means that this is another argument for the biased formula (3). The controversy starts again with the Helmert's transformation. The controversy of the variance formula becomes the controversy of the number of degrees of freedom: Is it n because of the variable x_i or is it $(n-1)$ because of the variable y_i? As there are no criteria to choose the right number of degrees of freedom, the author of this paper decided to take the proof-of-the-pudding approach to solve this controversy with software simulation.

In this simulation, the author of this paper wrote a small software program using the MATLAB[1] language. Ten thousand observations of a standardized normally distributed random variable were created. The observations were multiplied by a constant to give the random variable a variance of value 10 then added by another constant to give it a mean of value 10. Ten thousand observations are a big number, big enough to be considered as a population. The sample size n, taken from this population, has various values from 2 to 8. Then the sample variances were calculated from these samples by the two formulae (2) and (3). The variances were calculated repeatedly ten thousand times, each time with a new population of a standardized normally distributed random variable, to create populations of the variances. Then the means of these variance populations were calculated and reported. Table 1 is the result of these simulation runs.

While the number of degrees of freedom in the sum of squares S can hardly be agreed upon, the simulation results support the unbiased formula, ie. Equation (2). This result comes as a surprise to the author of this paper because of his support for the biased formula and his finding of the incorrect proofs for the unbiased formula. The result also tells us that the number of degrees of freedom in the sum of squares S is $(n-1)$, not n. With the Helmert's transformation, the proof for the unbiasedness of formula (2) can be established with the transformed variable y_i and the approach used in the theorem in section four.

Since the simulation supports the unbiased formula, we must find the error in the proof for the biased formula. This error can be found in Equation (7). One entry in the mean $\bar{x}$ will join with the variable x_t to create a ratio value that will not give a probability value. The proof is perfect for the mean, but it fails for the variance, with the way it is defined.

[1] MATLAB is a trade mark of The MathWorks, Inc.

Table 1. Sample Variances

Sample Size	Run	Variance	$\hat{\sigma}^2_{biased}$	$\hat{\sigma}^2_{unbiased}$
$n = 2$	1	10.0	4.9887	9.9774
	2	10.0	4.9949	9.9897
	3	10.0	4.9654	9.9308
$n = 3$	1	10.0	6.6668	10.0002
	2	10.0	6.6670	10.0004
	3	10.0	6.7573	10.1360
$n = 4$	1	10.0	7.5757	10.1009
	2	10.0	7.5195	10.0260
	3	10.0	7.5315	10.0420
$n = 5$	1	10.0	7.9661	9.9576
	2	10.0	8.0407	10.0509
	3	10.0	7.8863	9.8579
$n = 6$	1	10.0	8.2688	9.9226
	2	10.0	8.3285	9.9942
	3	10.0	8.3732	10.0449
$n = 7$	1	10.0	8.5002	9.9168
	2	10.0	8.5657	9.9932
	3	10.0	8.5416	9.9652
$n = 8$	1	10.0	8.6824	9.9228
	2	10.0	8.7385	9.9869
	3	10.0	8.7744	10.0279

7 Conclusion

In this paper, the sample variance formulae are studied again to determine the right one for computation. The old belief is the formula with a smaller divisor is the unbiased one. While the problem appears to be easy, the verdict must be decided with simulation results as there are incorrect proofs for both formulae in contention and no criteria for determining the number of degrees of freedom. Simulation results support the formula with a smaller divisor.

References

Irwin Guttman, Samuel S. Wilks and J. Stuart Hunter (1971). *Introductory Engineering Statistics*. John Wiley & Sons, Inc., New York, NY, USA, ISBN 0-471-33770-6.

Paul G. Hoel (1971). *Introduction to Mathematical Statistics*. John Wiley & Sons, Inc., New York, NY, USA, 4th Edition, ISBN 0-471-40365-2.

Robert V. Hogg and Allen T. Craig (1978). *Introduction to Mathematical Statistics*. Macmillan Publishing Co., Inc., New York, NY, USA, Fourth Edition, ISBN 0-02-355710-9.

Ky M. Vu (2007). *The ARIMA and VARIMA Time Series: Their Modelings, Analyses and Applications*. AuLac Technologies Inc., Ottawa, ON, Canada, ISBN 978-0-9783996-1-0.