Why is the sample variance a biased estimator?

Stephen So, PhD, MIEEE
Signal Processing Laboratory, Griffith School of Engineering,
Griffith University, Brisbane, QLD, Australia, 4111.
s.so@griffith.edu.au

September 11, 2008

Abstract

While Bessel's correction for the sample variance is well known and quoted abundantly in statistics' texts, a detailed treatment of why the correction is needed at all, does not appear to be prominently mentioned, other than a cursory statement to the effect that the estimator is biased without the correction. In this article, we present a mathematical treatment of the 'uncorrected' sample variance and explain why it is a biased estimator of the true variance of a population. This will be of interest to readers who are studying or have studied statistics but whom cannot find the real reason for Bessel's correction. The reader is assumed to have some prior knowledge of statistics and probability, particularly in relation to random variables and the mathematical expectation operator.

Index terms: sample variance, Bessel's correction, biased estimator

Contents

1 Introduction

The variance of a population σ^2 is an important second-order statistical measure since it gives an indication of the spread of data around the population mean μ. Assuming that ith datum in the population is represented as x_i and the number of data in the entire population is N_p, then the population variance is defined as:

$$\sigma^2 = \frac{1}{N_p}\sum_{i=1}^{N_p}(x_i - \mu)^2 \tag{1}$$

where the population mean is given by:

$$\mu = \frac{1}{N_p}\sum_{i=1}^{N_p} x_i \tag{2}$$

It may be impractical to calculate the population variance directly, perhaps due to N_p being very large or due to the data of the entire population being unavailable. In which case, we may only be given a smaller subset of the population, i.e. N samples (where $N < N_p$). Given the N samples of the population, we may *estimate* the population mean and variance by using the same expressions:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{3}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{4}$$

However, as quoted by many statistics' texts, the 'correct' sample variance should be estimated instead as:

$$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2 \tag{5}$$

The modification from $\frac{1}{N}$ to $\frac{1}{N-1}$ is called *Bessel's correction*. The reason that is usually given for why one should use Eq. (5) and not (4) is because the latter estimate is *biased*. The problem here is that not much more is given to the reader as to why Eq. (4) is considered biased.

This article will attempt to give a mathematical treatment of the sample variance as expressed in Eq. (4). The structure of the article is as follows. The next section will deal with the concept of bias in estimators when dealing with random variables. In Section, we present the mathematical treatment by finding the expected value of the biased sample variance estimator and showing that it is not equal to the population variance.

2 Mathematical notation

As this article will present several mathematical derivations, it is particularly helpful to present the variable notation that will be used.

X	random variable
x_i	ith sample or realisation of the random variable X
$E\{\bullet\}$	expectation operator
μ	population mean
σ^2	population variance
$\hat{\mu}$	sample mean
$\hat{\sigma^2}$	sample variance

3 The concept of bias in estimators

It is common place for us to estimate the value of a quantity that is related to a random population. Often this estimate needs to be obtained without all the necessary information available. For example, in order to find the average height of the human population on Earth, we can only estimate this quantity by taking a smaller sample set in practice. Therefore, this sample mean is an *estimator* of the quantity that we wish to find, namely the average height of the population.

Because the sample sets are picked randomly, then we cannot expect the sample mean to be exactly the same in each case. So it too will be a random variable. Therefore our next best hope is that these sample means should, *on the average*, fall on the true population mean. If the average of the sample means is equal to the population mean, then this suggests that our estimator is *unbiased*:

$$E\{\hat{\mu}\} = \mu \tag{6}$$

If an estimator is a biased one, that implies that the average of all the estimates is away from the true value that we are trying to estimate:

$$B = E\{\hat{\mu}\} - \mu \tag{7}$$

Therefore, the aim of this paper is to show that the average or expected value of the sample variance of (4) is *not equal* to the true population variance:

$$E\{\hat{\sigma^2}\} \neq \sigma^2 \tag{8}$$

4 Mathematical derivation of the bias in the uncorrected sample variance

Note that we assume that $\{x_i; i = 1, 2, \ldots, N\}$ are independent and identically distributed (iid).

$$
\begin{aligned}
E\{\hat{\sigma}_X^2\} &= E\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i-\hat{\mu}_X)^2\right\} && (9)\\
&= E\left\{\frac{1}{N}\sum_{i=1}^{N}(x_i^2-2x_i\hat{\mu}_X+\hat{\mu}_X^2)\right\} && (10)\\
&= E\left\{\frac{1}{N}\sum_{i=1}^{N}x_i^2-2\hat{\mu}_X\frac{1}{N}\sum_{i=1}^{N}x_i+\hat{\mu}_X^2\right\} && (11)\\
&= E\left\{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\hat{\mu}_X^2\right\} && (12)\\
&= E\left\{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\left(\frac{1}{N}\sum_{i=1}^{N}x_i\right)^2\right\} && (13)\\
&= E\left\{\frac{1}{N}\sum_{i=1}^{N}x_i^2-\frac{1}{N^2}\left(\sum_{i=1}^{N}x_i^2+\sum_{\substack{i=1\\i\neq j}}^{N}\sum_{j=1}^{N}x_ix_j\right)\right\} && (14)\\
&= \frac{1}{N}\sum_{i=1}^{N}E\{x_i^2\}-\frac{1}{N^2}\left(\sum_{i=1}^{N}E\{x_i^2\}+\sum_{\substack{i=1\\i\neq j}}^{N}\sum_{j=1}^{N}E\{x_ix_j\}\right) && (15)\\
&= \frac{1}{N}\sum_{i=1}^{N}E\{x_i^2\}-\frac{1}{N^2}\left(\sum_{i=1}^{N}E\{x_i^2\}+\sum_{\substack{i=1\\i\neq j}}^{N}\sum_{j=1}^{N}\mu_X^2\right) && (16)\\
&= \frac{1}{N}\sum_{i=1}^{N}E\{x_i^2\}-\frac{1}{N^2}\left(\sum_{i=1}^{N}E\{x_i^2\}+(N^2-N)\mu_X^2\right) && (17)\\
&= \frac{N-1}{N^2}\sum_{i=1}^{N}E\{x_i^2\}-\frac{N-1}{N}\mu_X^2 && (18)\\
&= \frac{N-1}{N}E\{x_i^2\}-\frac{N-1}{N}\mu_X^2 && (19)\\
&= \frac{N-1}{N}(\sigma_X^2+\mu_X^2)-\frac{N-1}{N}\mu_X^2 && (20)\\
&= \frac{N-1}{N}\sigma_X^2 && (21)\\
&\neq \sigma_X^2 && (22)
\end{aligned}
$$

Therefore, the sample variance without Bessel's correction is a biased estimator of the population variance. Note that we have used the following expressions to simplify (15) and (19), respectively:

$$
\begin{aligned}
E\{x_ix_j\} &= \mu_X^2 \text{ (iid assumption)} && (23)\\
E\{x_i^2\} &= \sigma_X^2+\mu_X^2 && (24)
\end{aligned}
$$