A comparison of variance estimators with known and unknown population means

Jennifer Ng and Julie Horrocks

This manuscript was prepared while Jennifer Ng was a summer research assistant under the supervision of Prof. Julie Horrocks, Department of Mathematics and Statistics, College of Physical and Engineering Sciences.

Variance is a concept that is key, yet often difficult to estimate in statistics. In this paper, we consider the problem of estimating the population variance when the population mean is known. We compare two estimators, one that incorporates the known population mean and another which estimates the population mean. The standard normal, standard exponential, and t distribution with 3 degrees of freedom are considered, with sample sizes of 5, 20, 50, and 100. It is determined that both estimators are unbiased. For the normal and exponential distributions, both estimators have similar variances; however, the estimator that incorporates the known mean has marginally lower variance, and thus is recommended. For the t(3) distribution, the variances of the estimators do not exist.

BACKGROUND

The Central Limit Theorem (CLT) states that the distribution of the sample mean approaches a normal distribution as the sample size increases regardless of the underlying distribution of the sample observations. Consequently, variance estimators discussed here will be based on the normal distribution. The maximum likelihood method developed by R.A. Fisher in 1922 will be primarily used, as this method is the most widely used statistical estimation technique.

If $X_1, X_2, ..., X_n$ is a random sample from a normal population with mean $\mu$ and variance $\sigma^2$ (denoted $N(\mu, \sigma^2)$), then its likelihood and log-likelihood functions are respectively given as follow:

$$L(\mu, \sigma^2 | x) = (2\pi\sigma^2)^{-n/2}\exp[-1/(2\sigma^2)\sum_{i=1}^{n}(x_i - \mu)^2]$$

and

$$\log L(\mu, \sigma^2 | x) = -n/2 \log(2\pi) - n/2 \log(\sigma^2) - 1/(2\sigma^2)\sum_{i=1}^{n}(x_i - \mu)^2$$

The partial derivatives are as follow:

$$\frac{\partial}{\partial \mu} \log L(\mu, \sigma^2 | x) = \sum_{i=1}^{n}(x_i - \mu) / \sigma^2$$

and
\[ \frac{\partial}{\partial \sigma^2} \log L(\mu, \sigma^2 | \mathbf{x}) = -n/(2\sigma^4) + \sum_{i=1}^{n} (x_i - \mu)^2 / (2\sigma^4) \]

By setting the partial derivatives to zero and solving for the parameter in question, the maximum likelihood estimator of the variance can be obtained. In the case where the population mean \( \mu \) is known, only the partial derivative with respect to \( \sigma^2 \) needs to be considered and the estimate obtained for the variance is

\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \]  \hspace{1cm} (1)

When \( \mu \) is unknown, the partial derivative with respect to \( \mu \) must also be set to zero and the resulting equation solved for \( \mu \), yielding a solution of \( \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i / n = \bar{x} \), which is the sample mean. This result is then substituted into equation (1) to obtain the estimator \( \hat{\sigma}^2 \).

It can be proven for any distribution of \( x_i \) with mean \( \mu \) and variance \( \sigma^2 \) that the variance estimator \( \hat{\sigma}^2 \) is unbiased, in that its expected value is equal to the parameter it is estimating, \( \sigma^2 \). The estimator \( \hat{\sigma}^2 \), on the other hand, is biased (also for any distribution), exhibited as follows:

\[ \text{E}[(\hat{\sigma}^2 - \sigma^2)] = \frac{1}{n} \text{E} \left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 \right\} / n \]

\[ = \left\{ \frac{1}{n} \left[ \text{E} \left\{ \sum_{i=1}^{n} x_i^2 \right\} - \text{E} \left\{ 2 \bar{x} \sum_{i=1}^{n} x_i \right\} + \text{E} \left\{ \sum_{i=1}^{n} \bar{x}^2 \right\} \right] \right\} / n \]

\[ = \frac{1}{n} \left\{ \sum_{i=1}^{n} \text{E} \left\{ x_i^2 \right\} - \text{E} \left\{ 2 \bar{x} \sum_{i=1}^{n} x_i \right\} + \text{E} \left\{ \sum_{i=1}^{n} \bar{x}^2 \right\} \right\} / n \]

\[ = \left\{ \frac{1}{n} \left[ \sum_{i=1}^{n} \left( \text{Var} \{ x_i \} + (\text{E} \{ x_i \})^2 \right) - 2n \text{E} \{ \bar{x}^2 \} + n \text{E} \{ \bar{x}^2 \} \right] \right\} / n \]

\[ = \left\{ \frac{1}{n} \left[ \sum_{i=1}^{n} \left( \sigma^2 + \mu^2 \right) - n \text{E} \{ \bar{x}^2 \} \right] \right\} / n \]

\[ = \left\{ \frac{n \sigma^2 + n \mu^2 - n(\sigma^2 / n + \mu^2) \right\} / n \]

\[ = \frac{n \sigma^2 + n \mu^2 - \sigma^2 - n \mu^2)} / n \]

\[ = \frac{n-1}{n} \sigma^2 \]

As it is highly desirable for an estimator to be unbiased, \( \hat{\sigma}^2 \) is modified to produce an unbiased estimator, denoted \( s^2 \):

\[ s^2 = \frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1) \]  \hspace{1cm} (2)

The two unbiased estimators of the variance \( \hat{\sigma}^2 \) and \( s^2 \) given by equations (1) and (2) will be examined in this study. It can be shown that if \( x_i \sim N(\mu, \sigma^2) \) for all \( i = 1, 2, \ldots, n \), and the \( x_i \) are independent, then

\[ u = \frac{n}{\sigma^2} \hat{\sigma}^2 - \chi^2_n \quad \text{and} \quad v = \frac{n-1}{\sigma^2} s^2 - \chi^2_{n-1}^4 \]

Below, these transformations will be applied to the estimators to verify their distributions. Upon simplification, the following results emerge:

\[ u = \sum_{i=1}^{n} (x_i - \mu)^2 / \sigma^2 - \chi^2_n \]  \hspace{1cm} (3)

and

\[ v = \sum_{i=1}^{n} (x_i - \bar{x})^2 / \sigma^2 - \chi^2_{n-1} \]  \hspace{1cm} (4)

Note that a \( \chi^2_n \) random variable has variance \( 2n \), \( \sigma^2 \) so \( \text{Var}(u) = 2n \) and \( \text{Var}(v) = 2(n-1) \) and \( \text{Var}(v) < \text{Var}(u) \) for all \( n \). Then

\[ \text{Var}(\hat{\sigma}^2) = \text{Var}(\sigma^2 u / n) = \sigma^4 \text{Var}(u) / n^2 \]

\[ = 2\sigma^4 n / n^2 = 2\sigma^4 / n \]  \hspace{1cm} (5)

Similarly,

\[ \text{Var}(s^2) = \text{Var}(\sigma^2 v / (n-1)) = \sigma^4 \text{Var}(v) / (n-1)^2 \]

\[ = 2\sigma^4 (n-1) / (n-1)^2 = 2\sigma^4 / (n-1) \]  \hspace{1cm} (6)

Consequently, given a sample of independent random variables \( x_i, i = 1, 2, \ldots, n \), where each \( x_i \) has distribution \( N(\mu, \sigma^2) \), theory establishes that \( \hat{\sigma}^2 \) always has smaller variance than \( s^2 \). For independent non-normal random variables \( x_i \), the distributions given in (3) and (4) will be achieved asymptotically due to the CLT. As the variance formulae in (5) and (6) are approximations for independent non-normal random variables, we will use simulations to compare the two estimators when the samples are from non-normal distributions. We chose the exponential distribution because it is frequently used in applied statistics, especially reliability and survival analysis. The \( t \) distribution with 3 degrees of freedom was selected for its heavy tail, as real data such as those pertaining to economics and telecommunications often have heavier tails than allowed for data such as those pertaining to economics and telecommunications often have heavier tails than allowed for under the normal distribution. \(^3\) In addition, simulations from the normal distribution are shown for comparison.
SIMULATION METHODS

The simulation study was conducted using the statistical software R version 2.4.1. The simulation was designed as a fully-crossed two-way factorial design with the following factors: three distributions and four sample sizes, for a total of 12 combinations. The standard normal (Normal\((\mu = 0, \sigma^2 = 1)\)) and standard exponential (Exponential\((\text{mean } \theta = 1)\)) distributions were examined, along with the t distribution with three degrees of freedom. Sample sizes of \(n = 5, 20, 50,\) and 100, representing a wide range from small to large, were considered.

One thousand iterations of the following process were repeated. Independent random variables \(X_1, X_2, \ldots, X_n\), with \(n\) taking on the values listed above, were generated from the three distributions and were then substituted into the equations for \(\widehat{\sigma}^2\) and \(s^2\). The sample mean and sample variance of the values obtained for \(\widehat{\sigma}^2\) and \(s^2\) were then calculated over the 1000 iterations. The estimated absolute relative bias, defined as the absolute value of the difference between the sample mean of the variance estimates and the true variance, divided by the true variance was then calculated. The estimated absolute relative bias and sample variance were used as performance measures to determine the superior estimator. The values generated were also substituted into equations (3) and (4) to obtain estimates for \(u\) and \(v\). Quantile-quantile (QQ) plots of \(u\) and \(v\) were then constructed to illustrate their distributions.

RESULTS AND DISCUSSION

Under the Normal \((\mu = 0, \sigma^2 = 1)\) distribution, the true mean of the data (used in the calculation of \(\widehat{\sigma}^2\)) is \(\mu = 0\), and the true variance of the data is \(\sigma^2 = 1\). Likewise under the Exponential \((\theta = 1)\) distribution, the mean is \(\mu = \theta = 1\) and the variance is \(\sigma^2 = \theta^2 = 1\). For the t distribution with degrees of freedom \(\nu = 3\), the mean and variance are \(\mu = 0\) and \(\sigma^2 = \frac{\nu}{\nu - 2} \approx 3\), respectively. \(^2\)

As the variance estimators \(\widehat{\sigma}^2\) and \(s^2\) are both unbiased, as shown in the Background section, both estimators should have mean values close to the values of the true variances listed above. The mean values over 1000 iterations for \(\widehat{\sigma}^2\) and \(s^2\) that resulted from the simulation study are listed in
Table 1. From Table 1, it is evident that both estimators \( \hat{\sigma}^2 \) and \( s^2 \) are good on average at estimating the variances for all three distributions, as the mean values are all close to the true values.

Figures 1, 2, and 3 represent the estimated absolute relative bias of the estimators, defined above in the Simulation Methods section, for the standard normal, standard exponential, and t (with 3 degrees of freedom) distributions, respectively. These figures show that \( \hat{\sigma}^2 \) has smaller estimated absolute relative bias than \( s^2 \) in more situations than the reverse case; however, the difference in bias is not great (differences often do not appear until the third decimal place).

Recall that for the normal distribution, the true absolute relative bias is zero. Figure 4 combines the plots from Figures 1, 2, and 3 to allow for comparison on the same scale. No clear trend in estimated absolute relative bias is exhibited as sample size increases.

Note that the trend for the standard exponential and t(3) appear to be opposing each other, as one increases when the other decreases, but that their ranges appear to be similar, and that the difference between either distribution and the standard normal is larger.

Table 2 lists the sample variances over 1000 iterations for the estimators \( \hat{\sigma}^2 \) and \( s^2 \). The former estimator attains a lower estimated variance in most cases. This indicates that \( \hat{\sigma}^2 \) is the superior estimator, as lower variance is a desirable quality. This result is somewhat expected, as \( \hat{\sigma}^2 \) uses the actual population mean in its calculations, and \( s^2 \) uses an estimate of the population mean. Intuitively, estimators that make use of more known information should be more accurate than those that ignore the known information and that instead use estimates.

It should also be noted that, for the most part, variability increases as the underlying distribution varies from Normal(0, 1), to Exponential(1) to t(3) and decreases as sample size increases. The exception is the t(3) distribution. Here the variance of both estimators increases as \( n \) increases from 50 to 100. There are several possible explanations for this apparent anomaly, which will be discussed in the next section.

Figure 4 illustrates the QQ plots of \( u \) and \( v \), the transformations of the variance estimators given in equations (3) and (4). These plots show the ordered sample values of \( u \) plotted against the theoretical quantiles of the \( \chi^2_n \) distribution (in a, c, e, and g), and the ordered sample values of \( v \) against the theoretical quantiles of the \( \chi^2_{n-1} \) distribution (in b, d, f, and h). Not all sample sizes are shown, but the results that are shown are typical for each distribution.

The underlying distribution of the values plotted in a), b), c), and d) is the standard normal. In this situation, it can be shown that the sums \( u \) and \( v \) are distributed as \( \chi^2_n \) and \( \chi^2_{n-1} \) respectively, and this is reflected by the approximate linearity of the QQ plots in a), b), c), and d).

The distribution of the values plotted in e) and f) is the standard exponential. In this case, the sums \( u \) and \( v \) do not have an exact chi-square distribution. This is reflected in plots e) and f) where there appears to be some slight curvature in the QQ plots. The values plotted in g) and h) have the t(3) distribution. These plots suggest that \( u \) and \( v \) do not follow the \( \chi^2_n \) and \( \chi^2_{n-1} \) distributions.

The t(3) DISTRIBUTION

As noted above, for the t(3) distribution, the sample variance of the two estimators does not decrease as the sample size increases. One possible explanation is that the R function for generating a random sample from the t(3) distribution is not working correctly. To investigate this we generated a sample of size 100 from the t(3) distribution using the \( rt \) function in R, and plotted this against the theoretical quantiles of the t(3) distribution. This plot is shown in Figure 6. The plot suggests that the sample does indeed follow the t(3) distribution.
Another possible explanation for the erratic behavior of the
sample variance of $\hat{\sigma}^2$ and $s^2$ is that the true variance of the
estimators does not exist, and in fact this turns out to be the
case.

It is well known that not all moments of the t distribution
exist. Recall that the $r^{th}$ moment, $\mu'_r = E[X^r]$, of an
absolutely continuous random variable $X$ with density $f(x)$ is
defined as

$$\mu'_r = \int_{-\infty}^{\infty} x^r f(x) \, dx \tag{1}$$

provided that the integral exists. If the integral does not
exist, we say that the $r^{th}$ moment does not exist. The first
moment, $E[X]$, is of course the mean or expectation of $X$ and
is denoted $\mu$. The $r^{th}$ central moment, $\mu_r = E[X-E[X]^r] = E[X-\mu]^r$, is defined as

$$\mu_r = \int_{-\infty}^{\infty} (x-\mu)^r f(x) \, dx \tag{2}$$

provided that the integral exists and the mean exists. The
second central moment, $\mu_2 = E[X-\mu]^2$, is of course the
variance, $\sigma^2$.

For the t distribution with $k$ degrees of freedom, the $r^{th}$
moment exists only if $r < k$. Thus the $r^{th}$ moment of the t
distribution with $k$ degrees of freedom does not exist for $r \geq k$.

For instance, consider the t distribution with 1 degree of
freedom, which is named the Cauchy distribution. Its mean,
variance, and all higher moments and central moments do not
exist.

Similarly, the t distribution with 2 degrees of freedom has
a finite expectation – namely zero – but its second moment,
and hence its variance, do not exist. All higher moments and
central moments of the t distribution with 2 degrees of
freedom are likewise nonexistent. For the t distribution with
3 degrees of freedom, the 3rd, 4th and all higher moments and
central moments are likewise nonexistent.

Consider a random sample from a distribution with $r^{th}$
central moment $\mu_r$. It can be shown that the variance of $s^2$ is
a function of $\mu_4$, the 4th central moment of the distribution of
the underlying observations. All moments and central
moments of the normal and exponential distribution exist,
including the 4th, so the variance of $s^2$ exists for these
distributions. For the t(3) distribution, however, the 4th
moment does not exist and so the variance of $s^2$ when the
sample is from the t(3) distribution does not exist. Likewise,
the variance of $\hat{\sigma}^2$ also involves the 4th central moment, and
thus when the sample is from the t(3) distribution, the
variance of $\hat{\sigma}^2$ also does not exist. This explains the erratic
behavior of the estimators when the sample is from the t(3)
distribution.
Both estimators of the variance discussed here are unbiased. For a random sample from the normal distribution, the variance of $\hat{\sigma}^2$ is shown theoretically to be lower than that of $s^2$. For the standard exponential distribution, simulations show that the variance of $\hat{\sigma}^2$ is typically lower than that of $s^2$. Thus for these distributions, it is recommended that $\hat{\sigma}^2$ be used when the population mean is known. Although when estimating the variance of a distribution, the mean is usually also unknown, and situations may arise where it may be assumed to be equal to a given value. This situation could arise, for instance, in a study of sources of variability of blood pressure measurements. Suppose two blood pressure measurements are taken in quick succession on a single individual, and let these measurements be denoted $X_1$ and $X_2$. We might assume that $X_1$ and $X_2$ are identically distributed. The difference in the two measurements, $Y=X_1-X_2$, could then be assumed to have a mean of zero. The estimated variance of $Y$ could then be used to infer something about the variance of $X_1$ and $X_2$. In such cases, incorporating the known value of the mean into the variance estimator will yield marginally more accurate results.

However, seeing as the variance of the two variance estimators are similar, it is not imperative that the true mean be known, as $s^2$ is also an acceptable estimator.

The $t$ distribution with 3 degrees of freedom has very heavy tails, which means that very large positive or negative values are common, and the third and higher moments consequently do not exist. In this situation, estimating the variance of the distribution is not recommended, and other approaches should be used.

It should be noted that only three distributions and four sample sizes were considered for this study. In the future, other distributions should be studied. Theoretical expressions for the variance of $\hat{\sigma}^2$ and $s^2$ should be investigated for samples from non-normal distributions. Other variance estimators could be considered.
Acknowledgments
The authors would like to thank the referees of this paper, the University of Guelph, and the Natural Sciences and Engineering Research Council of Canada (NSERC) for providing the funding necessary for the completion of this study.

REFERENCES