Journal of Applied Mathematics & Bioinformatics, vol.5, no.1, 2015, 53-65
ISSN: 1792-6602 (print), 1792-6939 (online)
Scienpress Ltd, 2015

An Improved Minimum Mean Squared Error Estimate of the Square of the Normal Population Variance Using Computational Intelligence

Grant H. Skrepnek[1], Ashok Sahai[2] and Robin Antoine[2]

Abstract

Building upon the commonly-employed approach by Searls, substantial work has addressed the use of the known coefficient of the normal population mean and the normal population variance. Subsequently, several attempts have also sought to formulate estimators for the population mean and variance for a more probable case of the population coefficient of variation being unknown. Across numerous real-world applications within basic science, economic, and medical research, an analyst is required to have an efficient estimator of the square of the population variance. As such, the purpose of the current investigation was to develop and test a more efficient estimator of the square of the population variance for a normal distribution, beyond that of the Minimum Mean Squared Error (*MMSE*) for the square of the population variance. The proposed approach, which

[1] College of Pharmacy and Stephenson Cancer Center, The University of Oklahoma Health Sciences Center, Oklahoma City, OK, USA.
E-mail: Grant-Skrepnek@ouhsc.edu

[2] Department of Mathematics and Statistics, The University of The West Indies, Faculty of Science and Agriculture, St. Augustine Campus, Trinidad and Tobago, West Indies. E-mail: robin.antoine@sta.uwi.edu, Ashok.Sahai@sta.uwi.edu

Article Info: *Received* : October 20, 2014. *Revised* : December 12, 2014.
Published online : January 15, 2015.

incorporated a metaheuristic optimization algorithm of Computational Intelligence in its derivation, captures the information in the sample more fully by including the sample coefficient of variation with the sample mean and sample variance. Results of an empirical simulation study found comprehensive improvement in the relative efficiency of the proposed estimator versus the *MMSE* estimator compared to the square of the sample variance across all defined sample sizes and population standard deviations.

Mathematics Subject Classification: 62H12

Keywords: Empirical-simulation study; Minimum mean-squared-error estimator; Sample coefficient of variation

1 Introduction

Substantial work accompanies the initial research conducted by Searls [1] regarding the estimator for the normal population mean with a known coefficient of variation. In these extensions, to illustrate, focus has often been directed toward utilizing the known coefficient of variation and kurtosis as presented in Khan [2], Gleser and Healy [3], Searls and Intarapanich [4], Arnholt and Hebert [5], and Sahai [6]. Subsequently, numerous approaches have been motivated by a need to formulate estimators for a population mean and variance for a more probable case of the population coefficient of variation being unknown, appearing in Sahai et al. [7], Richards et al. [8], Sahai et al. [9], and Lovric and Sahai [10].

Several research applications exist wherein the analyst requires an efficient estimator of the square of the population variance, particularly within basic science, economic, and medical research (e.g., randomized clinical trials, comparative or cost-effectiveness analyses). In the context of an efficient confidence interval estimation problem for the mean of a lognormal distribution which is commonplace

in these studies, the usual point estimator of the lognormal mean is $\bar{x} + s^2/2$, as described in Verma and Sahai [10], Skrepnek [11], and Skrepnek et al. [12]. This estimator utilizes the sample mean, $\bar{x}$, and the sample variance, s^2, based upon a random sample from the resultant normal population, following log-transformation of the data, as: $x = log(y) \sim N(\theta,\sigma^2)$. The variance of the usual point estimator, $\bar{x} + s^2/2$, is:

$$\frac{\sigma^2}{n} + \frac{\sigma^4}{2 \cdot (1+n)} \tag{1}$$

Notably, to estimate the variance expressed in **(1)**, the analyst is required to have an efficient estimator of the square of the normal population variance, σ^4, which ultimately may lead to an efficient confidence internal estimation of the lognormal mean. In this context, following the approach of Searls [1], a class of estimators, $k \cdot s^4$, may be considered for estimating the square of the normal population variance to establish the Minimum Mean Squared Error (*MMSE*) for the square of the normal population variance, σ^4.

Considering the prior issues, the purpose of the current research endeavor was to develop and test a more efficient estimator of the square of the population variance for a normal distribution, beyond that of the existing Minimum Mean Squared Error (*MMSE*) for the square of the population variance. The proposed approach, which incorporated a metaheuristic optimization (bat) algorithm of Computational Intelligence (CI) in its derivation, utilizes the information in the sample more fully by incorporating the sample coefficient of variation with the sample mean and sample variance. An empirical simulation study was conducted to assess the relative efficiency of the new proposed estimator, an Improved Mixed Minimum Squared Error (*IMMMSE*), and the *MMSE* compared to the square of the sample variance, s^4 (i.e., the sample counterpart estimator of the estimator of the square of the population variance, σ^4). The empirical investigation consisted of the calculations of the actual Mean Squared Error (*MSE*) of the estimators $MMSE(s^4)$, $IMMMSE(s^4)$, and s^4. For comprehensibility, results concerning the relative

efficiency of *MMSE(s^4)* and *IMMMSE(s^4)* versus s^4 were expressed in percentage terms.

2 The Proposition of an Efficient Estimator of the Square of a Normal Population Variance, σ^4

When considering a normal population with a population variance, *N(θ, σ^2)*, the process of obtaining the most efficient estimator of the square of the population variance, σ^4, seeks to utilize to the fullest extent possible information contained in the random sample from this population of size $n \sim x_1, x_2, \ldots, x_n$ that is summarized via the following two population statistics:

sample mean: $\overline{x} = \dfrac{\sum_{i=1}^{i=n} x_i}{n}$

sample variance: $s^2 = \dfrac{\sum_{i=1}^{i=n} \left(x_i - \overline{x}\right)^2}{\left(n-1\right)}$ (2)

By applying the well-known approach of using the sampling distribution of the sample variance, namely, $\left(n-1\right)\cdot\left(s^2/\sigma^2\right) \sim \chi^2$, and given that the degrees of freedom, *df = (n – 1)*, the following are obtained:

$$E(s^2) = \sigma^2 \tag{3}$$

$$E(s^4) = \frac{\sigma^4}{cn_1} \text{ where } cn_1 = \frac{(n-1)}{(n+1)} \tag{4}$$

$$E(s^6) = \frac{\sigma^6}{cn_2} \text{ where } cn_2 = \frac{(n-1)^2}{\{(n+1)(\mathrm{n}+3)\}} \tag{5}$$

$$E(s^8) = \frac{\sigma^8}{cn_3} \text{ where } cn_3 = \frac{(n-1)^3}{\{(n+1)(\mathrm{n}+3)(\mathrm{n}+5)\}} \tag{6}$$

As such, the subsequent lemma may be established.

Lemma. In the class of estimators $k \cdot s^4$, the Minimum Mean Squared Error (*MMSE*) estimator of σ^4 (i.e., the square of the normal population variance, σ^2) is $MMSE_1(s^4) = k^* \cdot s^4$, given that:

$$k^* = \frac{cn_3}{cn_1} = \frac{(n-1)^2}{\{(n+3)(n+5)\}} \tag{7}$$

Proof. For the *MMSE* in the class of estimators $k \cdot s^4$, the optimal value of k is:

$$k^* = E(s^4) \cdot \frac{\sigma^4}{E(s^8)}$$

via the straightforward application of (4) and (6). □

Contextually, it is also important to note, too, that based upon (4) and (6), the Relative Variance, $RV(s^4) = V(s^4)/\sigma^8$, of the estimator $MMSE_1(s^4)$ is:

$$\begin{aligned} V_1 &\equiv RV\left(MMSE_1\left(s^4\right)\right) \\ &= \frac{V\left(MMSE_1\left(s^4\right)\right)}{\sigma^8} \\ &= \left(k^*\right)^2 \cdot RV\left(s^4\right) \\ &= \frac{8 \cdot \left(k^*\right)^2 \cdot (n+1) \cdot (n+2)}{(n-1)^3} \end{aligned} \tag{8}$$

Building upon the aforementioned, an estimator based upon the sample coefficient of variation intended to more fully utilize the information within a given sample (i.e., the proposed estimator) may be formulated based upon notation set as $a = s^2 / \left(\bar{x}\right)^2$ and the square of the sample coefficient of variation designated as V. In considering a new class of estimators, $C \cdot \left(\bar{x}\right)^4$, it may be noted that:

$$\bar{x} \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

$$E\left(\bar{x}\right) = \theta$$

$$E\left(\bar{x}\right)^2=\left(1+\frac{a}{n}\right)\cdot\theta^2$$

$$E\left(\bar{x}\right)^3=\left(1+3\cdot\frac{a}{n}\right)\cdot\theta^3$$

$$E\left(\bar{x}\right)^4=\left(1+6\cdot\frac{a}{n}+3\cdot\left(\frac{a}{n}\right)^2\right)\cdot\theta^4$$

$$E\left(\bar{x}\right)^5=\left(1+10\cdot\frac{a}{n}+15\cdot\left(\frac{a}{n}\right)^2\right)\cdot\theta^5$$

$$E\left(\bar{x}\right)^6=\left(1+15\cdot\frac{a}{n}+45\cdot\left(\frac{a}{n}\right)^2+15\cdot\left(\frac{a}{n}\right)^3\right)\cdot\theta^6$$

$$E\left(\bar{x}\right)^7=\left(1+21\cdot\frac{a}{n}+105\cdot\left(\frac{a}{n}\right)^2+105\cdot\left(\frac{a}{n}\right)^3\right)\cdot\theta^7$$

$$E\left(\bar{x}\right)^8=\left(1+28\cdot\frac{a}{n}+210\cdot\left(\frac{a}{n}\right)^2+525\cdot\left(\frac{a}{n}\right)^3\right)\cdot\theta^8 \qquad (9)$$

Thus, the *MMSE* estimator in the class $C\cdot\left(\bar{x}\right)^4$ is:

$$C^*=\frac{E\left(\bar{x}\right)^4\cdot\theta^4}{E\left(\bar{x}\right)^8}$$

and, via (9), the following is obtained:

$$C^*=\frac{\left(1+6\cdot\frac{a}{n}+3\cdot\left(\frac{a}{n}\right)^2\right)}{\left(1+28\cdot\frac{a}{n}+210\cdot\left(\frac{a}{n}\right)^2+525\cdot\left(\frac{a}{n}\right)^3\right)} \qquad (10)$$

Therefore, in the class of estimators $C\cdot\left(\bar{x}\right)^4$, the *MMSE* of σ^4 (i.e., the square of the normal population variance σ^2) would be:

$$MMSE_2\left(s^4\right)=\frac{C^*\cdot s^4}{a^2} \qquad (11)$$

The relative variance, $RV\left(s^4\right) = V\left(s^4\right) \big/ \sigma^8$, of the estimator $MMSE_2(s^4)$ is:

$$
\begin{aligned}
V_2 &\equiv RV\left(MMSE_2\left(s^4\right)\right) \\
&= \frac{V\left(MMSE_2\left(s^4\right)\right)}{\theta^8} \\
&= \left(\frac{C^*}{a^2}\right)^2 \cdot RV\left(s^4\right) \\
&= \frac{8 \cdot \left(\frac{C^*}{a^2}\right)^2 \cdot (n+1) \cdot (n+2)}{(n-1)^3}
\end{aligned}
\tag{12}
$$

Given the aforementioned, a 'Mixed' *MMSE* estimator of σ^4 may be expressed as:

$$
MMMSE_2\left(s^4\right) = \frac{V_1 \cdot \left(MMSE_1\left(s^4\right)\right) + V_2 \cdot \left(MMSE_2\left(s^4\right)\right)}{\left(V_1 + V_2\right)} \tag{13}
$$

Subsequently, we may consider a class of estimators as follows, with '*m*' defined as a non-negative integer:

$$
MmMMSE_2\left(s^4\right) = MMSE_1\left(s^4\right) + m \cdot \left\{MMSE_1\left(s^4\right) - MMMSE_2\left(s^4\right)\right\} \tag{14}
$$

With the application of Computational Intelligence (CI) and simulation through a bat-inspired metaheuristic optimization algorithm described in Yang [13] and Khan et al. [14], it may be observed that optimum results for the non-negative integer in (14) are achieved with m = 10. Hence, it is proposed that an *Improved Mixed Minimum Mean Squared Error*, $IMMMSE(s^4)$, is defined as follows, which is ultimately the focus of the empirical simulation study to evaluate its relative efficiency versus the Minimum Mean Squared Error, $MMSE_1(s^4)$, for the square of the population variance σ^4:

$$
IMMMSE\left(s^4\right) = MMSE_1\left(s^4\right) + 10 \cdot \left\{MMSE_1\left(s^4\right) - MMMSE_2\left(s^4\right)\right\} \tag{15}
$$

3 Empirical Simulation Study

3.1 Methodology

An empirical simulation study was undertaken to assess the efficiency of the proposed estimators developed in the current investigation relative to the square of the sample variance, which is the sample counterpart estimator of the square of the population variance, s^4. This simulation consists in the calculations of the actual *MSEs* (Mean Squared Error) of the estimators $MMSE(s^4)$, $IMMMSE(s^4)$, and (s^4); results are presented as a Relative Efficiency for $MMSE(s^4)$ and $IMMMSE(s^4)$ versus s^4, expressed in percentage terms.

The parent population sampled in the simulation study was defined as a normal distribution with illustrative values of its population standard deviation as $\sigma = 0.20$, $\sigma = 0.25$, $\sigma = 0.30$, $\sigma = 0.40$, $\sigma = 0.45$, and $\sigma = 0.50$. Sample sizes were defined as n = 6, n = 11, n = 21, n = 31, n = 41, n = 51, n = 71, n = 101, n = 202, and n = 303. Additionally, population means were also considered. While the population coefficient of variation is typically unknown, the empirical investigation utilized a value of 0.25 as $\theta = 4\sigma$. Matlab 2010b code [The Mathworks Inc., Natick, Massachusetts] was developed and run with 51,000 replications. Again, results were presented as 'Relative Efficiencies':

$$\begin{aligned}
&\mathrm{RelEff}_{\%}\left\{MMSE\left(s^4\right) versus \left(s^4\right)\right\} \\
&\quad = 100 \cdot \frac{MSE\left(s^4\right)}{MSE\left(MMSE\left(s^4\right)\right)} \\
&\mathrm{RelEff}_{\%}\left\{IMMMSE\left(s^4\right) versus \left(s^4\right)\right\} \\
&\quad = 100 \cdot \frac{MSE\left(s^4\right)}{MSE\left(IMMMSE\left(s^4\right)\right)}
\end{aligned} \tag{16}$$

3.2 Results

Presented in Table 1, the relative efficiency of *IMMMSE(s^4)* ranged from a of 105.245 percent (σ = 0.25, n = 303) and a high of 647.400 percent (σ = 0.50, n 6), while the relative efficiency of *MMSE(s^4)* ranged from a low of 103.961 (σ = 0.25, n = 303) to a high of 602.842 percent (σ = 0.50, n = 6). Contingent this observation, the absolute difference between the *IMMMSE(s^4)* and *MMSE(s^4)* was at a maximum under conditions of a large population standard deviation and small sample size (σ = 0.50, n = 6) of 44.558 percentage points (*IMMMSE(s^4)* = 647.400% versus *MMSE(s^4)* = 602.842%), and at a minimum of 1.284 percentage points as sample sizes increased (σ = 0.25, n = 303) (*IMMMSE(s^4)* = 105.245% versus *MMSE(s^4)* =103.961%). Collectively, findings support the proposed efficient estimator, *IMMMSE(s^4)*, under all combinations of sample size and population standard deviations as illustrated by the gains in efficiency compared *MMSE(s^4)*.

Table 1: Relative Efficiencies of *IMMMSE(s^4)* and *MMSE(s^4)* Estimators Across Varying Sample Sizes and Population Standard Deviations

	Relative Efficiency Compared to s^4 (in %)						
	Population Standard Deviation						
Sample Size, ***Estimators***	***σ* = 0.20**	***σ* = 0.25**	***σ* = 0.30**	***σ* = 0.35**	***σ* = 0.40**	***σ* = 0.45**	***σ* = 0.50**
n=6							
MMSE (s^4)	572.647	589.823	581.617	566.441	576.710	575.012	602.842

IMMMSE (s^4)	610.154	632.370	618.729	600.219	613.251	610.602	647.400
n=11							
MMSE (s^4)	279.938	280.518	275.787	276.921	277.612	273.618	278.949
IMMMSE (s^4)	299.897	300.768	293.331	294.867	297.200	291.671	298.286
n=21							
MMSE (s^4)	173.994	173.182	175.231	173.994	174.589	175.884	173.907
IMMMSE (s^4)	185.468	184.363	187.331	185.462	186.445	187.671	185.608
n=31							
MMSE (s^4)	146.215	146.519	147.752	145.377	147.658	147.858	147.194
IMMMSE (s^4)	155.724	156.171	157.709	154.847	157.500	157.798	157.015
n=41							
MMSE (s^4)	133.858	134.665	134.216	134.498	134.065	134.043	133.544
IMMMSE (s^4)	142.700	143.695	142.786	143.485	143.183	142.600	142.017
n=51							
MMSE (s^4)	127.318	127.266	126.125	126.344	127.451	126.449	127.669
IMMMSE (s^4)	135.750	135.616	133.841	134.255	135.640	134.204	136.169
n=71							
MMSE (s^4)	118.928	118.252	118.873	117.912	118.723	119.297	118.862
IMMMSE (s^4)	126.057	125.197	125.937	124.673	125.814	126.713	126.122
n=101							

MMSE (s^4)	112.786	113.153	112.765	113.122	112.425	112.863	112.502
IMMMSE (s^4)	118.881	119.464	118.906	119.393	118.217	119.055	118.513
n=202							
MMSE (s^4)	106.320	106.076	106.493	106.618	106.613	105.908	106.436
IMMMSE (s^4)	110.102	109.589	110.483	110.614	110.391	109.248	110.199
n=303							
MMSE (s^4)	104.403	103.961	104.251	104.342	104.224	104.240	104.307
IMMMSE (s^4)	106.239	105.245	105.808	105.952	105.907	105.793	106.072

MMSE = Minimum Mean Squared Error; *IMMMSE* = Improved Mixed Minimum Mean Squared Error (proposed estimator); s^4 = estimate of the square of the population variance; n = sample size; σ = population standard deviation

4 Conclusion

The current investigation sought to develop and test a more efficient estimator of the square of the population variance for a normal distribution, beyond that of the Minimum Mean Squared Error (*MMSE*) for the square of the population variance. By using the information in the sample more fully by via the sample coefficient of variation with the sample mean and sample variance, and applying a bat-inspired metaheuristic optimization algorithm of Computational Intelligence in its derivation, results of an empirical simulation study found improvements in relative efficiency comprehensively across all defined sample sizes and population standard deviations.

References

[1] D.T. Searls, The Utilization of a Known Coefficient of Variation in the Estimation Procedure, *Journal of the American Statistical Association*, **59**, (1964), 1225-1226.

[2] R.A. Khan, A Note on Estimating the Mean of a Normal Distribution with Known Coefficient of Variation, *Journal of the American Statistical Association*, **63**, (1968), 1039-1041.

[3] L.J. Gleser and J.D. Healy, Estimating the Mean of a Normal Distribution with Known Coefficient of Variation, *Journal of the American Statistical Association*, **71**, (1976), 977-981.

[4] D.T. Searls and P. Intarapanich, A Note on an Estimator for the Variance that Utilizes the Kurtosis, *The American Statistician*, **44**, (1990), 295-296.

[5] A.T. Arnholt and J.E. Hebert, Estimating the Mean with Known Coefficient of Variation, *The American Statistician*, **49**, (1995), 367-369.

[6] A. Sahai, Efficient Estimator of Population Variance of Normal Distribution with Known Coefficient of Variation, *InterStat*, (June, 2011), Article 001.

[7] A. Sahai, R. Antoine, K. Wright, and M.R. Acharya, On Efficient Variance Estimation for Normal Populations, *InterStat*, (November, 2009), Article 002.

[8] W.A. Richards, R. Antoine, A. Sahai and M.R. Acharya, On Efficient Iterative Estimation Algorithm Using Sample Counterpart of the Searls' Normal Mean Estimator, *InterStat*, (January, 2010), Article 007.

[9] A. Sahai, M.R. Acharya and H. Ali, Efficient Estimation of Normal Population Mean, *Journal of Applied Sciences*, **6**, (2006), 1966-1968.

[10] S. Verma and A. Sahai, Efficient Confidence Interval Mean Estimation for Commonly-Used Statistical Distribution Modeling Earth System Sciences, *InterStat*, (October, 2009), Article 002.

[11] G.H. Skrepnek, Regression Methods in the Empirical Analysis of Health Care Data, *Journal of Managed Care Pharmacy*, **11**, (2005), 240-251.

[12] G.H. Skrepnek, E.L. Olvey and A. Sahai, Econometric Approaches in Evaluating Cost and Utilization within Pharmacoeconomic Analyses, Pharmaceutical Policy and Law, **14**, (2012), 105-122.

[13] X.S. Yang, A New Metaheuristic Bat-Inspired Algorithm, In: J.R. Gonzales, et al., editors, *Nature Inspired Cooperative Strategies for Optimization* (NISCO 2010), *Studies in Computational Intelligence*, **284**, (2010), 65-74.

[14] K. Khan and A. Sahai, A Comparison of BA, GA, PSO, BP and LM for Training Feed Forward Neural Networks, *International Journal of Intelligent Systems and Applications*, **4**, (2012), 23-29.