Estimating the mean and variance from the five-number summary of a log-normal distribution

Shi, Jiandong; Tong, Tiejun; Wang, Yuedong; Genton, Marc G.

Published in:
Statistics and its Interface

DOI:
10.4310/SII.2020.V13.N4.A9

Published: 31/07/2020

Document Version:
Publisher's PDF, also known as Version of record

Statistics and Its Interface Volume 13 (2020) 519–531

Estimating the mean and variance from the five-number summary of a log-normal distribution

Jiandong Shi, Tiejun Tong*, Yuedong Wang, and Marc G. Genton

In the past several decades, meta-analysis has been widely used to pool multiple studies for evidence-based practice. To conduct a meta-analysis, the mean and variance from each study are often required; whereas in certain studies, the five-number summary may instead be reported that consists of the median, the first and third quartiles, and/or the minimum and maximum values. To transform the five-number summary back to the mean and variance, several popular methods have emerged in the literature. However, we note that most existing methods are developed under the normality assumption; and when this assumption is violated, these methods may not be able to provide a reliable transformation. In this paper, we propose to estimate the mean and variance from the five-number summary of a log-normal distribution. Specifically, we first make the log-transformation of the reported quantiles. With the existing mean estimators and newly proposed variance estimators under the normality assumption, we construct the estimators of the log-scale mean and variance. Finally, we transform them back to the original scale for the final estimators. We also propose a bias-corrected method to further improve the estimation of the mean and variance. Simulation studies demonstrate that our new estimators have smaller biases and smaller relative risks in most settings. A real data example is used to illustrate the practical usefulness of our new estimators.

Keywords and phrases: Bias correction, Five-number summary, Log-normal distribution, Meta-analysis, Variance.

1. INTRODUCTION

In the past several decades, meta-analysis has been widely used to pool multiple studies for evidence-based practice. With accumulated evidence based on meta-analysis, more reliable and convincing conclusions are able to be drawn for scientific questions. In medical studies, the mean and variance (or standard deviation) are the most commonly reported summary statistics, especially when the data are normally distributed. In certain situations, however, researchers may instead report the whole or part of the five-number summary, which consists of the minimum value a, the first quartile q_1, the median m, the third quartile q_3, and the maximum value b. For convenience, we define the three common scenarios as follows:

$$\mathcal{S}_1 = \{a, m, b; n\},$$
$$\mathcal{S}_2 = \{q_1, m, q_3; n\},$$
$$\mathcal{S}_3 = \{a, q_1, m, q_3, b; n\},$$

where n is the sample size of the data.

To our knowledge, most existing meta-analytical methods, as well as the associated softwares, have been developed to analyze the studies with the mean and variance estimates. For the studies reported with the five-number summary, early researchers often excluded them for further analysis by claiming no sufficient data available. Needless to say, such a procedure will often lose valuable information from the literature and, consequently, the final conclusion is less reliable or is subject to publication bias, especially when a large number of studies were reported with the five-number summary. To avoid such information loss, a few methods for estimating the mean and variance (or standard deviation) from the five-number summary have been developed in the recent literature, including, for example, [11], [22], [23], [1], [16] and [21]. To show their popularity, we note that [23] and [16] have been cited 1185 and 162 times in Google Scholar as of 30 April 2020, respectively. However, it is also known that most existing methods are developed under the normality assumption, and in case if this assumption is violated, the existing methods may not be able to provide a reliable estimation. As an example, [20] and [9] showed that the index values of the vitamin D level tend to be positively skewed and so are unlikely to follow a normal distribution. In medical practice, skewed data are often modeled by the log-normal distribution; see Section 2 for more discussion.

In this paper, we propose to estimate the mean and variance from the reported five-number summary of a log-normal distribution with parameters μ and σ^2, $LN(\mu, \sigma^2)$. Specifically, we first make the log-transformation of the five-number summary; we then apply the existing estimators for normal data to estimate the log-scale mean and variance; and lastly, we transform them back to the original scale to achieve the final estimates of the mean and variance of the log-normal distribution. Note that the above three-step estimators are straightforward and easy to implement, yet on

*Corresponding author.

the other side, they may not be guaranteed to be unbiased. Inspired by this, we further propose a method to improve the three-step estimators by bias correction, and the simulation results show that the bias-corrected estimators are nearly unbiased and have smaller relative risks.

The following notations will be used throughout the paper. Let $X_1, X_2, \ldots, X_n$ be an independent and identically distributed random sample of size n from $LN(\mu, \sigma^2)$. Then by definition, $Y_i = \ln(X_i)$, $i = 1, 2, \ldots, n$, follow a normal distribution with mean μ and variance σ^2. We further define $Z_i = (Y_i - \mu)/\sigma$ as the standardized normal random variables. Then we have $a = \exp(\mu + \sigma a_z)$, $q_1 = \exp(\mu + \sigma q_{1,z})$, $m = \exp(\mu + \sigma m_z)$, $q_3 = \exp(\mu + \sigma q_{3,z})$ and $b = \exp(\mu + \sigma b_z)$, where $\{a_z, q_{1,z}, m_z, q_{3,z}, b_z\}$ represents the five-number summary of the standardized normal sample $Z_1, Z_2, \ldots, Z_n$.

The rest of the paper is organized as follows. In Section 2, we first propose the intermediate estimators for the log-scale mean and variance, and then transform them back to the original scale to achieve the final estimates of the mean and variance. In Section 3, we propose to further improve the three-step estimators in Section 2 by bias correction. We then demonstrate in Section 4 through simulation studies that our bias-corrected estimators are able to eliminate biases and achieve smaller relative risks in most settings. Section 5 presents a real data analysis to illustrate the usefulness of our new estimators. The paper is concluded in Section 6 with some discussion and future directions.

2. ESTIMATING THE MEAN AND VARIANCE FROM A LOG-NORMAL DISTRIBUTION

Let X be a random variable that follows a log-normal distribution $LN(\mu, \sigma^2)$, or equivalently, $Y = \ln(X)$ follows a normal distribution $N(\mu, \sigma^2)$. Then by definition, the mean and variance of X can be derived as

$$\mu_X = E(X) = \exp\left(\mu + \frac{\sigma^2}{2}\right) \tag{1}$$

and

$$\sigma_X^2 = \mathrm{Var}(X) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2). \tag{2}$$

For estimating the mean and standard deviation from the five-number summary, several popular methods have emerged in the recent literature, including, for example, [23], [16] and [21]. However, we note that most existing methods are developed under the normality assumption, and they may not be directly applicable for the log-normal data. In particular, the log-normal data are known to be positively skewed so that the mid-range $(a+b)/2$ does not serve as the center information as that for the normal data. As a remedy, if the raw data are transformed to the log-scale, then by noting that $\ln(a)$ and $\ln(b)$ are the minimum and maximum values of a normal sample, we can apply $(\ln(a) + \ln(b))/2$ to estimate the center information of the log-transformed data.

The logarithmic transformation has been widely used in statistics and related areas to transform the non-normal data to the normal data. As a well-known example, [3] introduced the Box-Cox transformation as

$$y^{(\lambda)} = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0, \\ \ln(y), & \text{if } \lambda = 0, \end{cases}$$

where the parameter λ is to determine the transformation formula for the non-normal data. Since this seminal paper, the Box-Cox transformation has been widely used in many different aspects of data-analysis, in which the log-transformation with $\lambda = 0$ is always among the most commonly used [15, 2, 14, 10]. Another well-known example of the log-transformation is for microarray data, where the gene expression data from the raw intensities are often highly skewed with nearly a half of data being within the interval (0,1) and the other half being distributed in the interval (1, ∞). After the log-transformation, the raw intensities are compressed to a narrower yet more symmetric range and the variance of the intensities is also stabilized. This procedure is also known as the normalization of microarray data.

Following the spirit of the log-transformation, to estimate the mean and variance from a log-normal distribution in Sections 2.1 to 2.3, we first transform the five-number summary under the three scenarios to the log-scale. Based on the existing standard deviation estimators, we propose unbiased variance estimators under three scenarios. Together with the existing mean estimators, we apply them to estimate the mean and variance from the log-transformed data. Lastly, we apply formulas (1) and (2) to transform the intermediate estimates back to achieve the final estimates of the mean and variance.

2.1 Estimation under scenario $\mathcal{S}_1$

For the log-normal data under scenario $\mathcal{S}_1 = \{a, m, b; n\}$, we take the log-scale of the median and the minimum and maximum values. Then by [16], we estimate the mean of the log-transformed data as

$$\hat{\mu}_1 = w_1 \left(\frac{\ln(a) + \ln(b)}{2}\right) + (1 - w_1)\ln(m), \tag{3}$$

where $w_1 = 4/(4 + n^{0.75})$. Further by [23], we propose an unbiased variance estimator of the log-transformed data as follows:

$$z_1^{-1}\left(\frac{\ln(b) - \ln(a)}{\xi}\right)^2,$$

where $z_1 = E[(b_z - a_z)/\xi]^2$ and $\xi = 2\Phi^{-1}[(n - 0.375)/(n + 0.25)]$ with Φ^{-1} being the quantile function of the standard normal distribution.

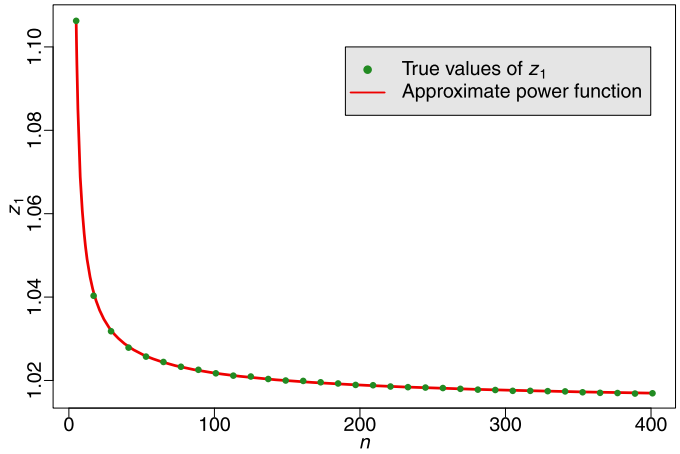

Figure 1. The green points represent the true values of the coefficient z_1 for n from 5 to 400, and the red line represents the approximate function of z_1.

We note, however, that the analytical form of the coefficient z_1 involves some complicated computation of the order statistics and may not be readily accessible to practitioners. To derive an approximation formula of z_1 for practical use, we compute the numerical values of z_1 for n up to 400 and plot them in Figure 1. Observing that z_1 is a monotonically decreasing function of the sample size n, we propose to approximate it by the power function $c_0+c_1[\ln(n)]^{c_2}$ with $c_1 > 0$ and $c_2 < 0$, and further derive the best approximation as $z_1 \approx 1.01 + 0.25[\ln(n)]^{-2}$. We also plot the approximate function of z_1 in Figure 1, which shows that our approximation is quite accurate for the sample size up to 400. Then with the proposed approximation, we estimate the variance of the log-transformed data as

$$\hat{\sigma}_1^2 = \left(\frac{\ln(b) - \ln(a)}{\xi}\right)^2 \left(1.01 + \frac{0.25}{(\ln(n))^2}\right)^{-1}. \tag{4}$$

Finally, with the intermediate estimators (3) and (4) for the log-transformed data, we transform them back to the original scale and get the mean and variance estimators as

$$\hat{\mu}_{X,1} = \exp\left(\hat{\mu}_1 + \frac{\hat{\sigma}_1^2}{2}\right) \tag{5}$$

and

$$\hat{\sigma}_{X,1}^2 = \exp(2\hat{\mu}_1 + 2\hat{\sigma}_1^2) - \exp(2\hat{\mu}_1 + \hat{\sigma}_1^2). \tag{6}$$

2.2 Estimation under scenario $\mathcal{S}_2$

For the log-normal data under scenario $\mathcal{S}_2 = \{q_1, m, q_3; n\}$, as in Section 2.1 we take the log-scale of the median and the first and third quartiles. By [16], we estimate the mean of the log-transformed data as

$$\hat{\mu}_2 = w_2\left(\frac{\ln(q_1) + \ln(q_3)}{2}\right) + (1 - w_2)\ln(m), \tag{7}$$

where $w_2 = 0.7 + 0.39/n$. And also by [23], we propose an unbiased variance estimator of the log-transformed data as follows:

$$z_2^{-1}\left(\frac{\ln(q_3) - \ln(q_1)}{\eta}\right)^2,$$

where $z_2 = E[(q_{3,z} - q_{1,z})/\eta]^2$ and $\eta = 2\Phi^{-1}[(0.75n - 0.125)/(n + 0.25)]$. It is noteworthy that $z_2\hat{\sigma}_2^2$ has been proposed in [18] to estimate the variance of the log-transformed data under scenario $\mathcal{S}_2$. It is evident that [18]'s estimator is biased and has a larger variance than our new estimator.

Similar to z_1, the theoretical values of z_2 is complicated to be computed for practitioners. Following the same spirit in Section 2.1, we approximate z_2 with a power function and conduct the best approximation formula as $z_2 \approx 1 + 1.58/n$. With the approximate formula, we estimate the variance of the log-transformed data as

$$\hat{\sigma}_2^2 = \left(\frac{\ln(q_3) - \ln(q_1)}{\eta}\right)^2 \left(1 + \frac{1.58}{n}\right)^{-1}. \tag{8}$$

Finally, with the intermediate estimators (7) and (8), we transform the data back to the original scale and get the mean and variance estimators as

$$\hat{\mu}_{X,2} = \exp\left(\hat{\mu}_2 + \frac{\hat{\sigma}_2^2}{2}\right) \tag{9}$$

and

$$\hat{\sigma}_{X,2}^2 = \exp(2\hat{\mu}_2 + 2\hat{\sigma}_2^2) - \exp(2\hat{\mu}_2 + \hat{\sigma}_2^2). \tag{10}$$

2.3 Estimation under scenario $\mathcal{S}_3$

For the log-normal data under scenario $\mathcal{S}_3 = \{a, q_1, m, q_3, b; n\}$, we take the log-scale for all the values from the five-number summary. We then apply [16] to estimate the mean of the log-transformed data as

$$\begin{aligned}\hat{\mu}_3 =& w_{3,1}\left(\frac{\ln(a) + \ln(b)}{2}\right) + w_{3,2}\left(\frac{\ln(q_1) + \ln(q_3)}{2}\right) \\ &+ (1 - w_{3,1} - w_{3,2})\ln(m),\end{aligned} \tag{11}$$

where $w_{3,1} = 2.2/(2.2 + n^{0.75})$ and $w_{3,2} = 0.7 - 0.72n^{-0.55}$. By [21], we propose an unbiased variance estimator of the log-transformed data as follows:

$$z_3^{-1}\left[w_3\left(\frac{\ln(b) - \ln(a)}{\xi}\right) + (1 - w_3)\left(\frac{\ln(q_3) - \ln(q_1)}{\eta}\right)\right]^2,$$

where $z_3 = E[w_3(b_z - a_z)/\xi + (1 - w_3)(q_{3,z} - q_{1,z})/\eta]^2$ and $w_3 = 1/(1 + 0.07n^{0.6})$. Again for practical use, we approximate z_3 with a power function and conduct the best approximation formula as $z_3 \approx 1 + 0.28[\ln(n)]^{-2}$. With the

approximate formula, we estimate the variance of the log-transformed data as

$$\hat{\sigma}_3^2 = \left[w_3 \left(\frac{\ln(b) - \ln(a)}{\xi} \right) + (1 - w_3) \left(\frac{\ln(q_3) - \ln(q_1)}{\eta} \right) \right]^2 \left(1 + \frac{0.28}{(\ln(n))^2} \right)^{-1}. \tag{12}$$

Finally, with the intermediate estimators (11) and (12) for the log-transformed data, we transform them back to the original scale that yield the mean and variance estimators as

$$\hat{\mu}_{X,3} = \exp\left(\hat{\mu}_3 + \frac{\hat{\sigma}_3^2}{2} \right) \tag{13}$$

and

$$\hat{\sigma}_{X,3}^2 = \exp(2\hat{\mu}_3 + 2\hat{\sigma}_3^2) - \exp(2\hat{\mu}_3 + \hat{\sigma}_3^2). \tag{14}$$

3. BIAS-CORRECTED ESTIMATION

The estimation methods in Section 2 for the mean and variance are simple and easy to implement for the log-normal data. One problem is that, with a direct plug-in for the estimated parameters, the final estimates of the mean and variance may have non-negligible bias. In this section, we propose to further improve the estimation by eliminating the bias introduced by the plug-in method.

3.1 Bias-corrected estimation under scenario $\mathcal{S}_1$

Under scenario $\mathcal{S}_1 = \{a, m, b; n\}$, we derive in Appendix A that the expected value of the mean estimator (5) is approximately

$$E(\hat{\mu}_{X,1}) \approx \phi_1 \exp\left(\mu + \frac{\sigma^2}{2} \right), \tag{15}$$

where $\phi_1 = 1 + 0.565\sigma^2/n + 0.37\sigma^4/n$. By (15), it is then natural to consider the bias-corrected estimator for the mean as $\hat{\mu}_{X,1}/\hat{\phi}_1$, where $\hat{\phi}_1 = 1 + 0.565\hat{\sigma}_1^2/n + 0.37\hat{\sigma}_1^4/n$. We estimate σ^2 in ϕ_1 by $\hat{\sigma}_1^2$ in (4). For σ^4 in ϕ_1, following the same spirit as in the derivation of $\hat{\sigma}_1^2$, we propose the estimator $\hat{\sigma}_1^4 = [(\ln(b) - \ln(a))/\xi]^4/[1 + 2.23(\ln(n))^{-2}]$. Finally, we have

$$\tilde{\mu}_{X,1} = \exp\left(\hat{\mu}_1 + \frac{\hat{\sigma}_1^2}{2} \right) \left(1 + \frac{0.565}{n}\hat{\sigma}_1^2 + \frac{0.37}{n}\hat{\sigma}_1^4 \right)^{-1}. \tag{16}$$

Furthermore, we derive in Appendix A that the expected value of the variance estimator (6) can be approximated as

$$E(\hat{\sigma}_{X,1}^2) \approx \phi_{1,1} \exp(2\mu + 2\sigma^2) - \phi_{1,2} \exp(2\mu + \sigma^2), \tag{17}$$

where $\phi_{1,1} = 1 + 2.26\sigma^2/n + 5.92\sigma^4/n$ and $\phi_{1,2} = 1 + 2.26\sigma^2/n + 1.48\sigma^4/n$. This then suggests the bias-corrected estimator of the variance as

$$\tilde{\sigma}_{X,1}^2 = \exp(2\hat{\mu}_1 + 2\hat{\sigma}_1^2) \left(1 + \frac{2.26}{n}\hat{\sigma}_1^2 + \frac{5.92}{n}\hat{\sigma}_1^4 \right)^{-1} - \exp(2\hat{\mu}_1 + \hat{\sigma}_1^2) \left(1 + \frac{2.26}{n}\hat{\sigma}_1^2 + \frac{1.48}{n}\hat{\sigma}_1^4 \right)^{-1}. \tag{18}$$

3.2 Bias-corrected estimation under scenario $\mathcal{S}_2$

Under scenario $\mathcal{S}_2 = \{q_1, m, q_3; n\}$, we derive in Appendix B that the expected value of the mean estimator (9) is approximately

$$E(\hat{\mu}_{X,2}) \approx \phi_2 \exp\left(\mu + \frac{\sigma^2}{2} \right), \tag{19}$$

where $\phi_2 = 1 + 0.57\sigma^2/n + 0.75\sigma^4/n$. By (19), it is natural to consider the bias-corrected estimator of the mean as $\hat{\mu}_{X,2}/\hat{\phi}_2$, where $\hat{\phi}_2 = 1 + 0.57\hat{\sigma}_2^2/n + 0.75\hat{\sigma}_2^4/n$. We estimate σ^2 in ϕ_2 by $\hat{\sigma}_2^2$ in (8). For σ^4 in ϕ_2, following the same spirit as in the derivation of $\hat{\sigma}_2^2$, we propose the estimator $\hat{\sigma}_2^4 = [(\ln(q_3) - \ln(q_1))/\eta]^4/(1 + 19.2/n^{1.2})$. Finally, we have

$$\tilde{\mu}_{X,2} = \exp\left(\hat{\mu}_2 + \frac{\hat{\sigma}_2^2}{2} \right) \left(1 + \frac{0.57}{n}\hat{\sigma}_2^2 + \frac{0.75}{n}\hat{\sigma}_2^4 \right)^{-1}. \tag{20}$$

Furthermore, we derive in Appendix B that the expected value of the variance estimator (10) can be approximated as

$$E(\hat{\sigma}_{X,2}^2) \approx \phi_{2,1} \exp(2\mu + 2\sigma^2) - \phi_{2,2} \exp(2\mu + \sigma^2), \tag{21}$$

where $\phi_{2,1} = 1 + 2.28\sigma^2/n + 12\sigma^4/n$ and $\phi_{2,2} = 1 + 2.28\sigma^2/n + 3\sigma^4/n$. Thus we propose the bias-corrected estimator of the variance as

$$\tilde{\sigma}_{X,2}^2 = \exp(2\hat{\mu}_2 + 2\hat{\sigma}_2^2) \left(1 + \frac{2.28}{n}\hat{\sigma}_2^2 + \frac{12}{n}\hat{\sigma}_2^4 \right)^{-1} - \exp(2\hat{\mu}_2 + \hat{\sigma}_2^2) \left(1 + \frac{2.28}{n}\hat{\sigma}_2^2 + \frac{3}{n}\hat{\sigma}_2^4 \right)^{-1}. \tag{22}$$

3.3 Bias-corrected estimation under scenario $\mathcal{S}_3$

Under scenario $\mathcal{S}_3 = \{a, q_1, m, q_3, b; n\}$, we derive in Appendix C that the expected value of the mean estimator is

$$E(\hat{\mu}_{X,3}) \approx \phi_3 \exp\left(\mu + \frac{\sigma^2}{2} \right), \tag{23}$$

where $\phi_3 = 1 + 0.405\sigma^2/n + 0.315\sigma^4/n$. By (23), it is natural to consider the bias-corrected estimator of the mean as $\hat{\mu}_{X,3}/\hat{\phi}_3$, where $\hat{\phi}_3 = 1 + 0.405\hat{\sigma}_3^2/n + 0.315\hat{\sigma}_3^4/n$. We estimate σ^2 in ϕ_3 by $\hat{\sigma}_3^2$ in (12). For σ^4 in ϕ_3, following the same spirit as in the derivation of $\hat{\sigma}_3^2$, we propose the estimator

Table 1. The normal-based mean and variance estimators under the three scenarios

Scenario	Mean estimator	Variance estimator
$\mathcal{S}_1$	$w_1\left(\frac{a+b}{2}\right)+(1-w_1)m$	$\left(\frac{b-a}{\xi}\right)^2\left(1.01+\frac{0.25}{(\ln(n))^2}\right)^{-1}$
$\mathcal{S}_2$	$w_2\left(\frac{q_1+q_3}{2}\right)+(1-w_2)m$	$\left(\frac{q_3-q_1}{\eta}\right)^2\left(1+\frac{1.58}{n}\right)^{-1}$
$\mathcal{S}_3$	$w_{3,1}\left(\frac{a+b}{2}\right)+w_{3,2}\left(\frac{q_1+q_3}{2}\right)(1-w_{3,1}-w_{3,2})m$	$\left[w_3\left(\frac{b-a}{\xi}\right)+(1-w_3)\left(\frac{q_3-q_1}{\eta}\right)\right]^2\left(1+\frac{0.28}{(\ln(n))^2}\right)^{-1}$

$\hat{\sigma}_3^4=[w_3(\ln(b)-\ln(a))/\xi+(1-w_3)(\ln(q_3)-\ln(q_1))/\eta]^4/(1+3.93/n)$. Finally, we have

(24)
$$\tilde{\mu}_{X,3}=\exp\left(\hat{\mu}_3+\frac{\hat{\sigma}_3^2}{2}\right)\left(1+\frac{0.405}{n}\hat{\sigma}_3^2+\frac{0.315}{n}\hat{\sigma}_3^4\right)^{-1}.$$

Furthermore, we derive in Appendix C that the expected value of the variance estimator (14) can be approximated as

(25)
$$E(\hat{\sigma}_{X,3}^2)\approx\phi_{3,1}\exp(2\mu+2\sigma^2)-\phi_{3,2}\exp(2\mu+\sigma^2),$$

where $\phi_{3,1}=1+1.62\sigma^2/n+5.04\sigma^4/n$ and $\phi_{3,2}=1+1.62\sigma^2/n+1.26\sigma^4/n$. This suggests the bias-corrected estimator of the variance as

$$\tilde{\sigma}_{X,3}^2=\exp(2\hat{\mu}_3+2\hat{\sigma}_3^2)\left(1+\frac{1.62}{n}\hat{\sigma}_3^2+\frac{5.04}{n}\hat{\sigma}_3^4\right)^{-1}$$
(26)
$$-\exp(2\hat{\mu}_3+\hat{\sigma}_3^2)\left(1+\frac{1.62}{n}\hat{\sigma}_3^2+\frac{1.26}{n}\hat{\sigma}_3^4\right)^{-1}.$$

4. SIMULATION STUDY

In this section, we conduct simulation studies to assess the finite sample performance of the proposed estimators, including the plug-in (PI) estimators in (5), (6), (9), (10), (13) and (14), and the bias-corrected (BC) estimators in (17), (18), (21), (22), (25) and (26). In addition, our simulations also include the existing normal-based (NB) estimators (see Table 1) for comparison, as well as to explore their potential consequence.

For each of the three scenarios, we consider two different log-normal distributions: $LN(3,0.3^2)$ and $LN(3,0.7^2)$. As shown in Figure 2, $LN(3,0.3^2)$ is a less right-skewed distribution compared to $LN(3,0.7^2)$. Then for each setting with the sample size up to 400, we generate $T=100{,}000$ random samples and apply the NB, PI and BC methods to estimate the mean and variance of the log-normal distribution. Finally, to compare the performance of the three estimation methods, we apply the relative bias (RB) defined as

$$\mathrm{RB}(\hat{\mu}_X)=\frac{1}{T}\sum_{i=1}^{T}\frac{\hat{\mu}_{X,i}-\mu_X}{\mu_X}$$

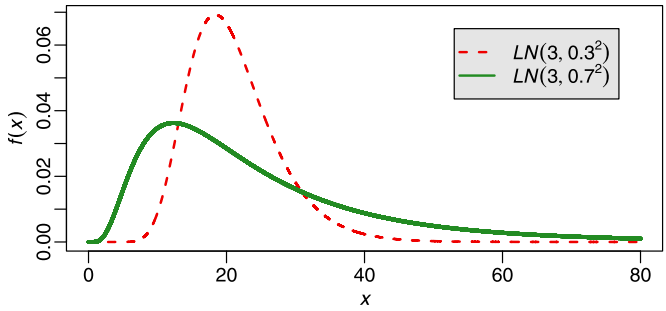

Figure 2. The dashed line represents the probability density function of $LN(3,0.3^2)$ and the solid line represents the probability function of $LN(3,0.7^2)$.

and

$$\mathrm{RB}(\hat{\sigma}_X^2)=\frac{1}{T}\sum_{i=1}^{T}\frac{\hat{\sigma}_{X,i}^2-\sigma_X^2}{\sigma_X^2},$$

where $\hat{\mu}_{X,i}$ and $\hat{\sigma}_{X,i}^2$ are the estimates from three estimation methods from the ith sample. In addition, we also compute the relative mean squared error (RMSE) for the mean estimators and compute the relative Stein's loss (RSL) [8] for the variance estimators defined as

$$\mathrm{RMSE}(\hat{\mu}_X)=\frac{\sum_{i=1}^{T}(\hat{\mu}_{X,i}-\mu_X)^2}{\sum_{i=1}^{T}(\bar{X}_i-\mu_X)^2}$$

and

$$\mathrm{RSL}(\hat{\sigma}_X^2)=\frac{\sum_{i=1}^{T}\left(\hat{\sigma}_{X,i}^2/\sigma_X^2-\ln(\hat{\sigma}_{X,i}^2/\sigma_X^2)-1\right)}{\sum_{i=1}^{T}\left(S_i^2/\sigma_X^2-\ln(S_i^2/\sigma_X^2)-1\right)},$$

where $\bar{X}_i$ is the sample mean and S_i^2 is the sample variance of the ith sample. Note that Stein's loss penalizes the underestimation as equally as the overestimation, which is more appropriate for evaluating the variance estimators, in particular for the skewed distributions. With 100,000 simulations for each setting, the simulation results under scenario $\mathcal{S}_1$ are reported in Figure 3 for the mean estimators, and in Figure 4 for the variance estimators.

From the reported RBs and RMSEs of the mean estimators in Figure 3, it is evident that the PI and BC estimators perform better than the NB estimator in most settings.

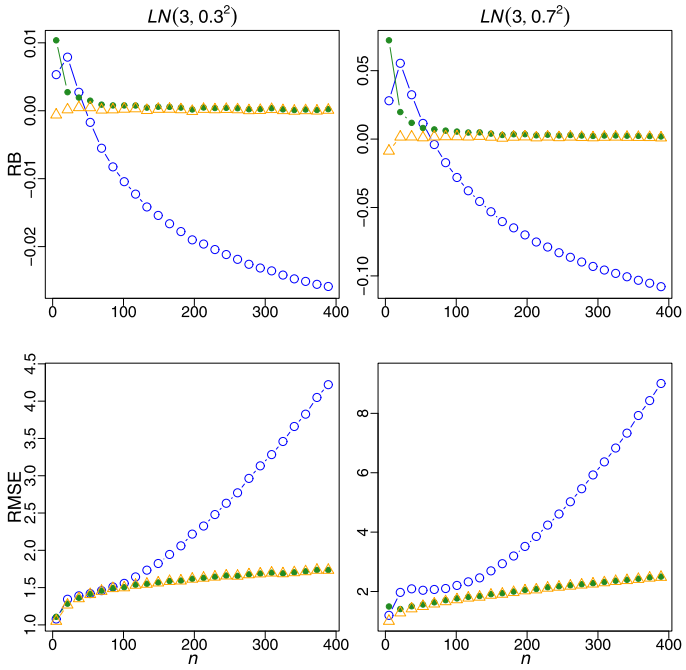

Figure 3. The RBs and RMSEs of three types of mean estimators under scenario $\mathcal{S}_1$, where the blue empty points represent the NB estimator, the green solid points represent the PI estimator, and the orange empty triangles represent the BC estimator.

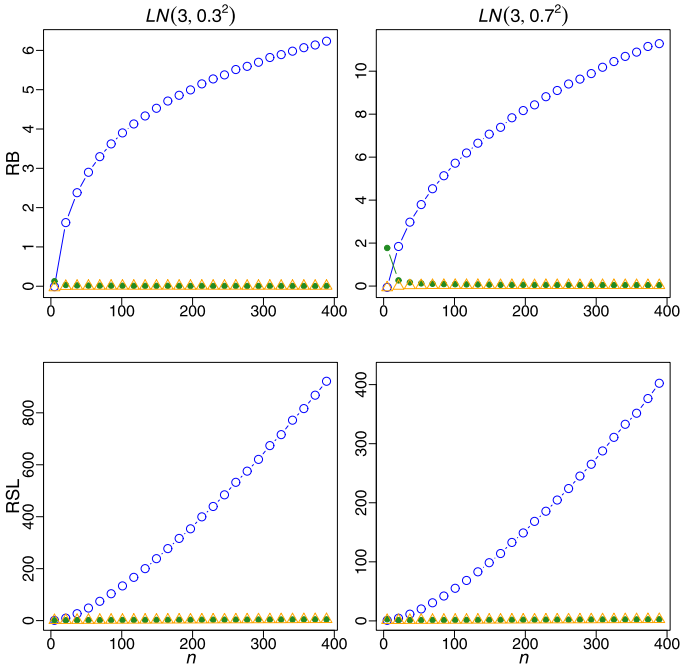

Figure 4. The RBs and RSLs of three types of variance estimators under scenario $\mathcal{S}_1$, where the blue empty points represent the NB estimator, the green solid points represent the PI estimator, and the orange empty triangles represent the BC estimator.

Specifically, in view of the RBs, the NB estimator is significantly biased in particular for the large sample sizes, which also leads to the larger RMSEs. Furthermore, for the newly proposed PI and BC estimators of the mean, we note that the BC estimator always provides smaller RBs and smaller RMSEs than the PI estimator, and such improvements get more evident when n is small. For the simulation results on the variance estimators reported in Figure 4, it is noted that the NB estimator yields unacceptably large RBs and RSLs. This indicates that the NB estimator is not applicable in practice. For the newly proposed PI and BC estimators of the variance, the BC estimator always yields smaller RBs and smaller RSLs than the PI estimator. To conclude, with the RB, RMSE and RSL as the criteria, the BC method performs better than the NB and PI methods and can be recommended for practical use.

To avoid the main text being too lengthy, we report the simulation results under scenarios $\mathcal{S}_2$ and $\mathcal{S}_3$ in Appendix D and Appendix E, where the comparative results remain similar as those under scenario $\mathcal{S}_1$.

5. REAL DATA ANALYSIS

Through the simulation studies, it has been shown that our new estimators offer more accurate estimates of the mean and variance if the data follow a log-normal distribution. In this section, we apply the proposed methods to a real data example and compare the results with those based on the NB methods.

[17] studied the relationship between the low serum vitamin D levels and tuberculosis. They included seven studies in meta-analysis, where two of them reported the mean and standard deviation for cases and controls, three of them reported the median and range, one of them reported the mean and range, and the other one reported the odds ratio. Noting that the odd ratio is not able to be synthesized with the mean and standard deviation, we exclude that study from our meta-analysis and present the summary statistics of the other six studies in the following table.

From Table 2 we note that, for the studies reported with the medians (or mean) and ranges, the median (or mean) values are closer to the minimum values than to the maximum values. It is known that similar patterns have also been observed in the literature, see, for example, [20] and [9], in which the data are positively skewed so that the NB estimates may lead to misleading results. For the study reported with the mean and range, we apply the mean directly in the meta-analysis. Recall that the SD is equal to the multiplication of the mean and $\sqrt{\exp(\sigma^2)-1}$ under a log-normal distribution. To estimate the SD, we first estimate $\exp(\sigma^2)$ by following the same spirits of the PI estimation in Section 2 and the BC estimation in Section 3. Then with the estimate of $\exp(\sigma^2)$ plugged into $\sqrt{\exp(\sigma^2)-1}$, the final SD estimate is achieved by taking the multiplication of the re-

Table 2. The summary statistics of the six studies included in the meta-analysis

Study	Cases	Controls
Davies et al. (1985) [5]	Median (range): 16 (2.25-74.25)	Median (range): 27.25 (9-132.5)
Grange et al. (1985) [12]	Median (range): 65.75 (43.75-130.5)	Median (range): 69.5 (48.5-125)
Davies et al. (1987) [7]	Median (range): 39.75 (16.75-89.25)	Median (range): 65.5 (26.25-114.75)
Davies et al. (1988) [6]	Mean (SD): 69.5 (24.5)	Mean (SD): 95.5 (29.25)
Chan et al. (1994) [4]	Mean (SD): 46.5 (18.5)	Mean (SD): 52.25 (15.75)
Sasidharan et al. (2002) [19]	Mean (range): 26.75 (2.5-75)	Mean (range): 48.5 (22.5-145)

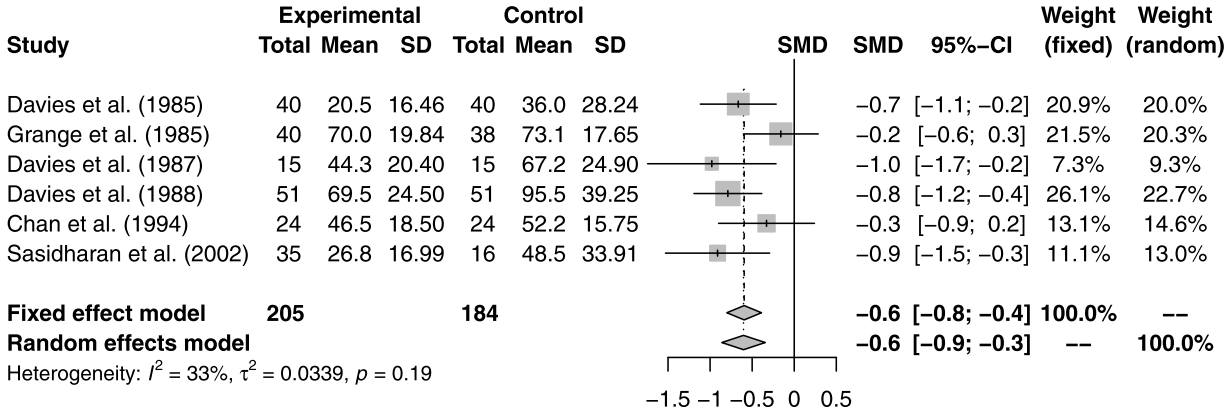

Figure 5. The forest plot based on the normal-based (NB) estimates.

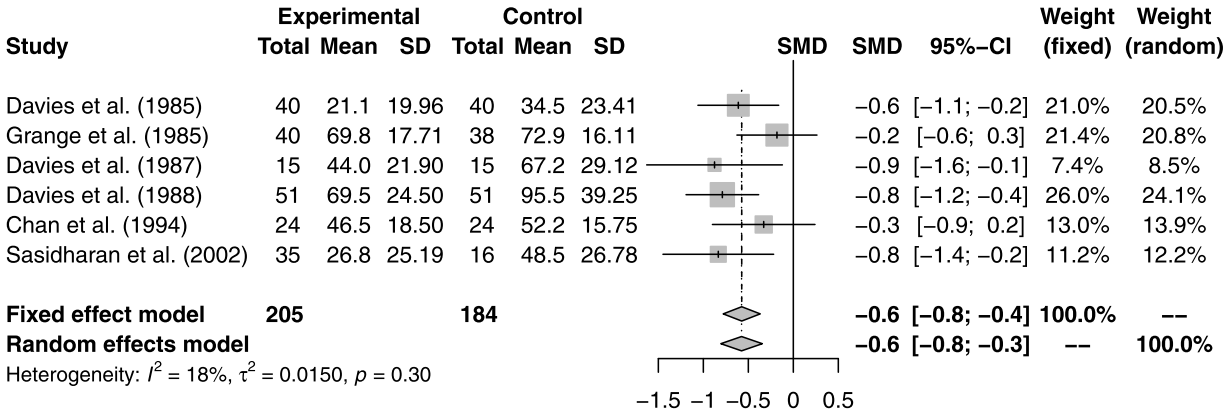

Figure 6. The forest plot based on the plug-in (PI) estimates.

ported mean and the estimate of $\sqrt{\exp(\sigma^2) - 1}$ from each estimation method. In addition by [13], the effect sizes are measured with the standardized mean difference (SMD). We fit the fixed-effect and random-effects models to the estimates based on the NB, PI and BC methods. The estimates of the mean and standard deviation from the three methods and their meta-analytical results are reported in Figures 5, 6 and 7, respectively.

As shown in the forest plots, the three methods lead to similar estimates of the means and standard deviations, and therefore yield similar effect sizes and confidence intervals. We note, however, that the values of the heterogeneity index I^2 are different in the three meta-analyses. Specifically, the value of I^2 in the meta-analysis based on the NB estimates is as large as 33% which is of the moderate heterogeneity according to [13] with the threshold as 25%. While for the I^2 values from the other two meta-analyses, they are 18% and 21%, respectively, both indicating a low heterogeneity. This coincides with the p-values in the three meta-analyses, where the p-value in the first meta-analysis is smallest, indicating a more significant heterogeneity among the included studies. Together with the simulation results under scenario $\mathcal{S}_1$ that the NB estimates are not reliable for skewed data, we conclude that the meta-analysis based on the NB estimates is not reliable, but instead our new estimates should be adopted for further meta-analysis.

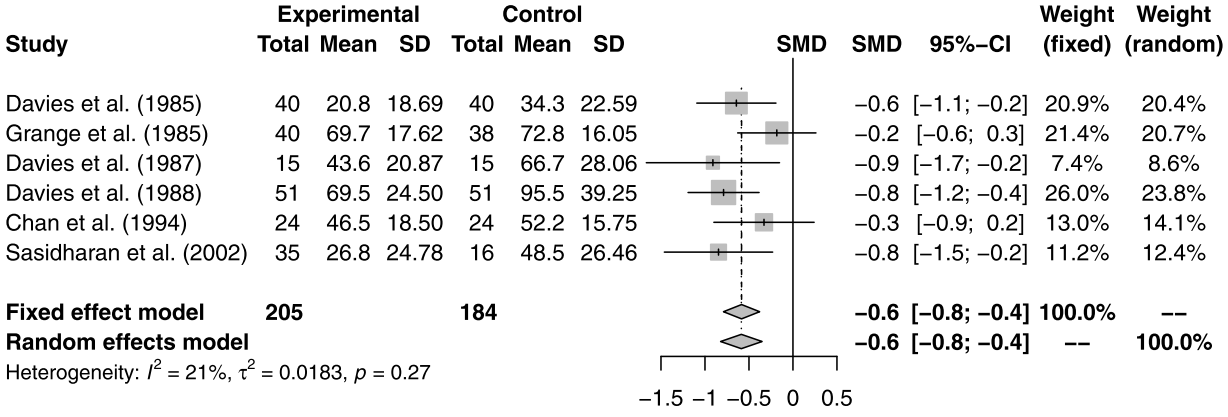

Figure 7. The forest plot based on the bias-corrected (BC) estimates.

6. DISCUSSION

Meta-analysis has become increasingly popular in the past several decades, especially in evidence-based practice. By synthesizing the evidence from multiple studies, researchers can achieve more reliable conclusions for a specific scientific question. In clinical trials with continuous outcomes, the mean and variance are most commonly reported; whereas some other studies may report the five-number summary or part of it as the summary statistics. Several popular methods have been proposed in the recent literature to estimate the mean and variance (or standard deviation) from the five-number summary. However, most existing methods are developed under the normality assumption, and in case if this assumption is violated, the existing methods may not be able to provide a reliable estimation. Thus blindly applying the existing methods to skewed data may lead to misleading or erroneous conclusions.

In this paper, we estimate the mean and variance from the five-number summary of a log-normal distribution. Firstly, we make the log-transformation of the five-number summary; we then apply the existing estimators for normal data to estimate the log-scale mean and variance; and lastly, we transform them back to the original scale to achieve the final estimates. The above three-step estimators are straightforward and easy to implement, yet they may not be guaranteed to be unbiased. This then motivates us to further improve the plug-in estimators by bias correction. Through simulation studies, we demonstrate that our new estimators perform better than the normal-based estimators in most settings.

In addition to the plug-in and bias-corrected methods considered in this paper, a potentially novel method can also be the quantile least squares (QLS) estimation [25, 24] as follows:

$$\hat{\theta}_{\mathrm{QLS}} = \arg\min_{\theta}\{[\hat{\xi}_k - \xi_k(\theta)]^\top[\hat{\xi}_k - \xi_k(\theta)]\}$$
$$= \arg\min_{\theta}\sum_{i=1}^{k}\{\hat{\xi}_{p_i} - \xi_{p_i}(\theta)\}^2,$$

where $\hat{\xi}_k = (\hat{\xi}_{p_1}, \ldots, \hat{\xi}_{p_k})^\top$ and $\xi_k(\theta) = (\xi_{p_1}(\theta), \ldots, \xi_{p_k}(\theta))^\top$ with $\top$ the transpose of a vector, and $\hat{\xi}_{p_i}$ and $\xi_{p_i}(\theta)$ are the sample and theoretical p_i quantiles of the assumed distribution for $i = 1, \ldots, k$. For the log-normal distribution, the classical parameters are μ and σ^2 where $\theta = (\mu, \sigma^2)^\top$ is the vector form. Thus we have $\xi_{p_i}(\theta) = \exp(\mu + \sigma z_{p_i})$, where $z_i = (1, z_{p_i})^\top$ and z_{p_i} is the p_i quantile of the standard normal distribution. However, by solving the minimization problem, we can only obtain the estimates of μ and σ^2. For the final mean and variance estimates, one more step that plugs in $\hat{\theta}_{\mathrm{QLS}}$ to formulas (1) and (2) is needed, where some biases can be involved. The problem can be resolved by reparameterizing the classical parameter vector $\theta = (\mu, \sigma^2)^\top$ as $\theta' = (\mu_X, \sigma_X^2)^\top$. It then follows directly from formulas (1) and (2) that

$$\mu = \ln(\mu_X) - \frac{1}{2}\ln\left(\frac{\sigma_X^2}{\mu_X^2} + 1\right),$$
$$\sigma^2 = \ln\left(\frac{\sigma_X^2}{\mu_X^2} + 1\right).$$

With μ_X and σ_X^2 as the new parameters, we have

$$\hat{\theta}'_{\mathrm{QLS}} = \arg\min_{\theta}\sum_{i=1}^{k}(\hat{\xi}_{p_i} - \xi_{p_i}(\theta'))^2,$$

where $\xi_{p_i}(\theta') = \exp(\ln(\mu_X) - \ln(\sigma_X^2/\mu_X^2 + 1)/2 + \sqrt{\ln(\sigma_X^2/\mu_X^2 + 1)}z_{p_i})$. Without any intermediate estimates, the QLS estimates of the mean and variance can be directly obtained. However, the QLS estimation may have a drawback that the sample quantiles may not guarantee to be close to the population quantiles in the sense of expectation, in particular for skewed data. Moreover, since the data available from the five-number summary is very limited, it may not be sufficient to conduct the QLS estimation and so future research may be warranted.

To sum up, we recommend the bias-corrected method for estimating the mean and variance of a log-normal dis-

Table 3. The recommended mean and variance estimators under the three scenarios

Scenario	Mean	Variance
$\mathcal{S}_1$	estimator (16)	estimator (18)
$\mathcal{S}_2$	estimator (20)	estimator (22)
$\mathcal{S}_3$	estimator (24)	estimator (26)

tribution, and to be more specific, we summarize the recommended estimators under the three scenarios in Table 3. Also for practical implementation, we have provided an Excel spreadsheet which is available in the Supplementary Materials, http://intlpress.com/site/pub/files/_supp/sii/2020/0013/0004/SII-2020-0013-0004-s005.xlsx.

APPENDIX A

Derivation of (15). To derive (15), we apply the second-order Taylor expansion around $\mu+\sigma^2/2$ for $\hat{\mu}_{X,1}$ as follows:

$$\hat{\mu}_{X,1} \approx \exp\left(\mu+\frac{\sigma^2}{2}\right)\left[1+\left(\hat{\mu}_1+\frac{\hat{\sigma}_1^2}{2}-\mu-\frac{\sigma^2}{2}\right)+\frac{1}{2}\left(\hat{\mu}_1+\frac{\hat{\sigma}_1^2}{2}-\mu-\frac{\sigma^2}{2}\right)^2\right].$$

Then by taking the expectation on both sides and the fact that $\hat{\mu}_1+\hat{\sigma}_1^2/2$ is an unbiased estimator of $\mu+\sigma^2/2$, we have

$$\begin{aligned} E(\hat{\mu}_{X,1}) &\approx \exp\left(\mu+\frac{\sigma^2}{2}\right)\left[1+\frac{1}{2}\mathrm{Var}\left(\hat{\mu}_1+\frac{\hat{\sigma}_1^2}{2}\right)\right] \\ &= \exp\left(\mu+\frac{\sigma^2}{2}\right)\left(1+\frac{1}{2}V_{1,1}\sigma^2+\frac{1}{8}V_{1,2}\sigma^4+\frac{1}{2}V_{1,3}\sigma^3\right), \end{aligned} \tag{27}$$

where

$$\begin{aligned} V_{1,1} &= \mathrm{Var}\left(w_1\left(\frac{a_z+b_z}{2}\right)+(1-w_1)m_z\right), \\ V_{1,2} &= \mathrm{Var}\left(z_1^{-1}\left(\frac{b_z-a_z}{\xi}\right)^2\right), \\ V_{1,3} &= \mathrm{Cov}\left(w_1\left(\frac{a_z+b_z}{2}\right)+(1-w_1)m_z, z_1^{-1}\left(\frac{b_z-a_z}{\xi}\right)^2\right). \end{aligned}$$

Note that $V_{1,1}$, $V_{1,2}$ and $V_{1,3}$ are the functions of n only. Through numerical computation, we observe that $V_{1,3}$ is always much smaller than $V_{1,1}$ and $V_{1,2}$ for any given n, and so the covariance term is nearly negligible. We further follow Section 2.1 and derive the approximate formulas of $V_{1,1}$ and $V_{1,2}$ as $V_{1,1} \approx 1.13/n$ and $V_{1,2} \approx 2.96/n$. Finally, by (27) we have

$$E(\hat{\mu}_{X,1}) \approx \exp\left(\mu+\frac{\sigma^2}{2}\right)\left(1+\frac{0.565}{n}\sigma^2+\frac{0.37}{n}\sigma^4\right).$$

Derivation of (17). Similar to the derivation of (15), we can apply the Taylor expansion on each term of estimator (6) and then take the expectation on both sides. By doing so, we have

$$\begin{aligned} E(\hat{\sigma}_{X,1}^2) \approx& \exp(2\mu+2\sigma^2)\left(1+\frac{1}{2}\mathrm{Var}(2\hat{\mu}_1+2\hat{\sigma}_1^2)\right) \\ &-\exp(2\mu+\sigma^2)\left(1+\frac{1}{2}\mathrm{Var}(2\hat{\mu}_1+\hat{\sigma}_1^2)\right) \\ \approx& \exp(2\mu+2\sigma^2)\left(1+2V_{1,1}\sigma^2+2V_{1,2}\sigma^4\right) \\ &-\exp(2\mu+\sigma^2)\left(1+2V_{1,1}\sigma^2+\frac{1}{2}V_{1,2}\sigma^4\right). \end{aligned}$$

Then with the approximate formulas for $V_{1,1}$ and $V_{1,2}$, it yields the result in (17).

APPENDIX B

Derivation of (19). To derive (19), we apply the second-order Taylor expansion around $\mu+\sigma^2/2$ for $\hat{\mu}_{X,2}$ as

$$\hat{\mu}_{X,2} \approx \exp\left(\mu+\frac{\sigma^2}{2}\right)\left[1+\left(\hat{\mu}_2+\frac{\hat{\sigma}_2^2}{2}-\mu-\frac{\sigma^2}{2}\right)+\frac{1}{2}\left(\hat{\mu}_2+\frac{\hat{\sigma}_2^2}{2}-\mu-\frac{\sigma^2}{2}\right)^2\right].$$

Then by taking the expectation on both sides and the fact that $\hat{\mu}_2+\hat{\sigma}_2^2/2$ is an unbiased estimator of $\mu+\sigma^2/2$, we have

$$\begin{aligned} E(\hat{\mu}_{X,2}) &\approx \exp\left(\mu+\frac{\sigma^2}{2}\right)\left[1+\frac{1}{2}\mathrm{Var}\left(\hat{\mu}_2+\frac{\hat{\sigma}_2^2}{2}\right)\right] \\ &= \exp\left(\mu+\frac{\sigma^2}{2}\right)\left(1+\frac{1}{2}V_{2,1}\sigma^2+\frac{1}{8}V_{2,2}\sigma^4+\frac{1}{2}V_{2,3}\sigma^3\right), \end{aligned} \tag{28}$$

where

$$\begin{aligned} V_{2,1} &= \mathrm{Var}\left(w_2\left(\frac{q_{1,z}+q_{3,z}}{2}\right)+(1-w_2)m_z\right), \\ V_{2,2} &= \mathrm{Var}\left(z_2^{-1}\left(\frac{q_{3,z}-q_{1,z}}{\eta}\right)^2\right), \\ V_{2,3} &= \mathrm{Cov}\left(w_2\left(\frac{q_{1,z}+q_{3,z}}{2}\right)+(1-w_2)m_z, z_2^{-1}\left(\frac{q_{3,z}-q_{1,z}}{\eta}\right)^2\right). \end{aligned}$$

Through numerical computation, the true values of $V_{2,3}$ is much smaller than those of $V_{2,1}$ and $V_{2,2}$ for any given n and thus the covariance term in (28) is nearly negligible. By fitting the true values of $V_{2,1}$ and $V_{2,2}$ as in Section 2.2, we derive the approximate formulas of $V_{2,1}$ and $V_{2,2}$ as $V_{2,1} \approx$

$1.14/n$ and $V_{2,2} \approx 6/n$. With the approximate formulas, we finally derive (19).

Derivation of (21). Similar to the derivation of (19), we apply the Taylor expansion on each term of estimator (10) and then take the expectation on both sides so that

$$\begin{aligned} E(\hat{\sigma}^2_{X,2}) \approx & \exp(2\mu + 2\sigma^2)\left(1 + \frac{1}{2}\mathrm{Var}(2\hat{\mu}_2 + 2\hat{\sigma}^2_2)\right) \\ & - \exp(2\mu + \sigma^2)\left(1 + \frac{1}{2}\mathrm{Var}(2\hat{\mu}_2 + \hat{\sigma}^2_2)\right) \\ \approx & \exp(2\mu + 2\sigma^2)\left(1 + 2V_{2,1}\sigma^2 + 2V_{2,2}\sigma^4\right) \\ & - \exp(2\mu + \sigma^2)\left(1 + 2V_{2,1}\sigma^2 + \frac{1}{2}V_{2,2}\sigma^4\right). \end{aligned}$$

With the approximate formulas for $V_{2,1}$ and $V_{2,2}$, it yields the result in (21).

APPENDIX C

Derivation of (23). To derive (23), we apply the second-order Taylor expansion around $\mu + \sigma^2/2$ for $\hat{\mu}_{X,3}$ as

$$\begin{aligned} \hat{\mu}_{X,3} \approx & \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 + \left(\hat{\mu}_3 + \frac{\hat{\sigma}^2_3}{2} - \mu - \frac{\sigma^2}{2}\right)\right. \\ & \left. + \frac{1}{2}\left(\hat{\mu}_3 + \frac{\hat{\sigma}^2_3}{2} - \mu - \frac{\sigma^2}{2}\right)^2\right]. \end{aligned}$$

Then by taking the expectation on both sides and the fact that $\hat{\mu}_3 + \hat{\sigma}^2_3/2$ is an unbiased estimator of $\mu + \sigma^2/2$, we have

$$\begin{aligned} E(\hat{\mu}_{X,3}) &\approx \exp\left(\mu + \frac{\sigma^2}{2}\right)\left[1 + \frac{1}{2}\mathrm{Var}\left(\hat{\mu}_3 + \frac{\hat{\sigma}^2_3}{2}\right)\right] \\ &= \exp\left(\mu + \frac{\sigma^2}{2}\right)\left(1 + \frac{1}{2}V_{3,1}\sigma^2 + \frac{1}{8}V_{3,2}\sigma^4 + \frac{1}{2}V_{3,3}\sigma^3\right), \end{aligned} \tag{29}$$

where

$$\begin{aligned} V_{3,1} &= \mathrm{Var}\left(w_{31}\left(\frac{b_z + a_z}{2}\right) + w_{32}\left(\frac{q_{1,z} + q_{3,z}}{2}\right)\right. \\ &\quad \left. + (1 - w_{31} - w_{32})m_z\right), \\ V_{3,2} &= \mathrm{Var}\left(z_3^{-1}\left[w_3\left(\frac{b_z - a_z}{\xi}\right) + (1 - w_3)\left(\frac{q_{3,z} - q_{1,z}}{\eta}\right)\right]^2\right), \\ V_{3,3} &= \mathrm{Var}\left(w_{31}\left(\frac{b_z + a_z}{2}\right) + w_{32}\left(\frac{q_{1,z} + q_{3,z}}{2}\right)\right. \\ &\quad + (1 - w_{31} - w_{32})m_z, \\ &\quad \left. z_3^{-1}\left[w_3\left(\frac{b_z - a_z}{\xi}\right) + (1 - w_3)\left(\frac{q_{3,z} - q_{1,z}}{\eta}\right)\right]^2\right). \end{aligned}$$

Through numerical computation, the true values of $V_{3,3}$ is much smaller than those of $V_{3,1}$ and $V_{3,2}$ for any given n and the covariance term in (29) is nearly negligible. By fitting the true values of $V_{3,1}$ and $V_{3,2}$ as in Section 2.3, we derive the approximate formulas of $V_{3,1}$ and $V_{3,2}$ as $V_{3,1} \approx 0.81/n$ and $V_{3,2} \approx 2.52/n$. With the approximate formulas, we finally derive (23).

Derivation of (25). Similar to the derivation of (23), we apply the Taylor expansion on each term of estimator (14) and then take the expectation on both sides. Specifically, it yields that

$$\begin{aligned} E(\hat{\sigma}^2_{X,3}) \approx & \exp(2\mu + 2\sigma^2)\left(1 + \frac{1}{2}\mathrm{Var}(2\hat{\mu}_3 + 2\hat{\sigma}^2_3)\right) \\ & - \exp(2\mu + \sigma^2)\left(1 + \frac{1}{2}\mathrm{Var}(2\hat{\mu}_3 + \hat{\sigma}^2_3)\right) \\ \approx & \exp(2\mu + 2\sigma^2)\left(1 + 2V_{3,1}\sigma^2 + 2V_{3,2}\sigma^4\right) \\ & - \exp(2\mu + \sigma^2)\left(1 + 2V_{3,1}\sigma^2 + \frac{1}{2}V_{3,2}\sigma^4\right). \end{aligned}$$

With the approximate formulas for $V_{3,1}$ and $V_{3,2}$, it yields the result in (25).

APPENDIX D

Under scenario $\mathcal{S}_2$, from the reported RBs and RMSEs of the mean estimators in Figure 8, it is evident that the PI

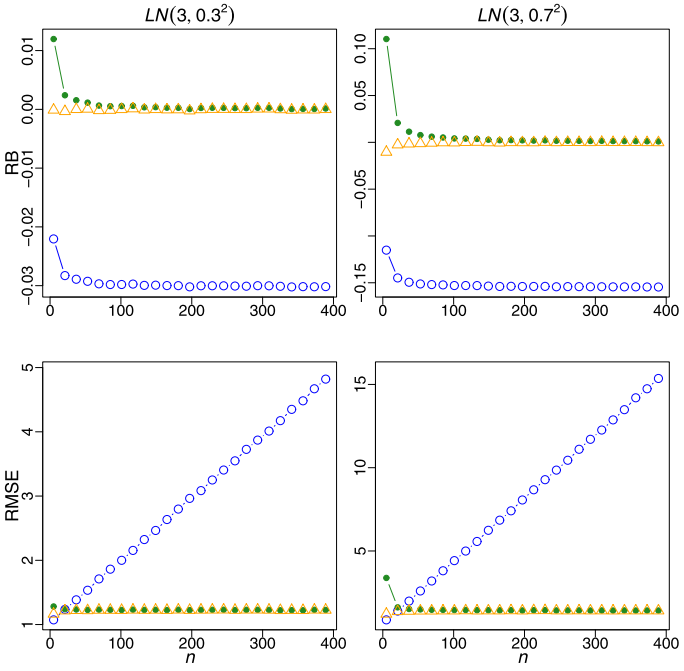

Figure 8. The RBs and RMSEs of three types of mean estimators under scenario $\mathcal{S}_2$, where the blue empty points represent the NB estimator, the green solid points represent the PI estimator, and the orange empty triangles represent the BC estimator.

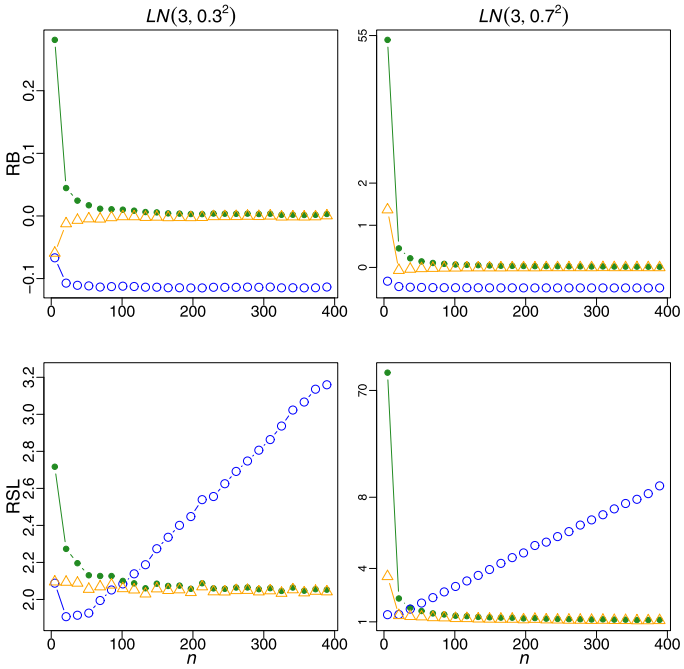

Figure 9. The RBs and RSLs of three types of variance estimators under scenario $\mathcal{S}_2$, where the blue empty points represent the NB estimator, the green solid points represent the PI estimator, and the orange empty triangles represent the BC estimator.

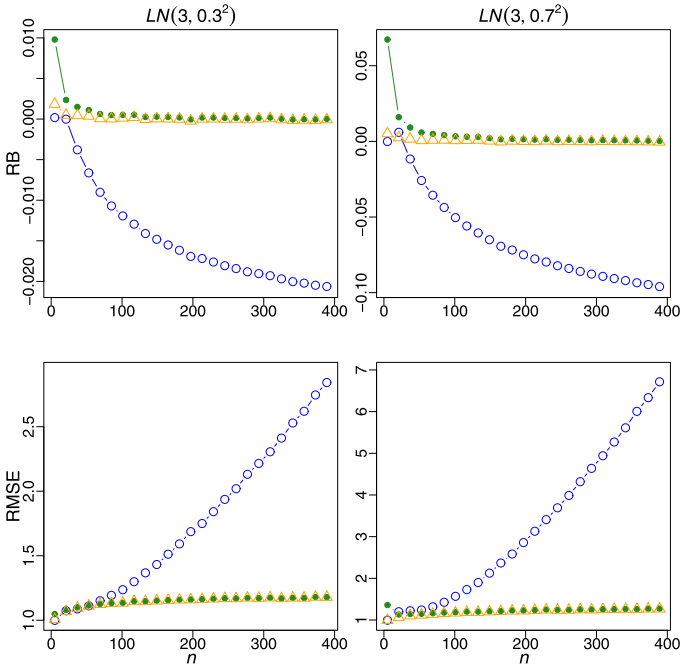

Figure 10. The RBs and RMSEs of three types of mean estimators under scenario $\mathcal{S}_3$, where the blue empty points represent the NB estimator, the green solid points represent the PI estimator, and the orange empty triangles represent the BC estimator.

and BC estimators perform better than the NB estimator in most settings. Specifically, in view of the RBs, the NB estimator is always biased regardless of the sample size n. This also leads to the larger RMSEs of NB estimator, which gets more evident as n increases. Furthermore, for the newly proposed PI and BC estimators of the mean, we note that the BC estimator always provides smaller RBs and smaller RMSEs than the PI estimator, and such improvements get more evident when n is small. This suggests that the PI method may not be capable to provide the accurate estimates for small sample sizes, which also demonstrates the necessity of developing the BC method. For the simulation results on the variance estimators reported in Figure 9, it is noted that the NB estimator yields the small RSLs when n is small. The main reason is because the NB estimator often underestimates the true variance with the smaller variances, whereas as n is small, the BC estimator yields the larger variances in spite of smaller biases. However, as n gets large, our BC estimator performs thoroughly better than the NB estimator. For the newly proposed PI and BC estimators of the variance, the BC estimator always yields smaller RBs and smaller RSLs than the PI estimator. To conclude, with the RB, RMSE and RSL as the criteria, the BC method generally performs better than the NB and PI methods and can be recommended for practical use.

APPENDIX E

Under scenario $\mathcal{S}_3$, from the reported RBs and RMSEs of the mean estimators in Figure 10, it is evident that the PI and BC estimators perform better than the NB estimator in most settings. Specifically, the NB estimator is biased and yields the larger RMSEs, which is similar to the case under scenario $\mathcal{S}_1$. Furthermore, for the newly proposed PI and BC estimators of the mean, we note that the BC estimator always provides smaller RBs and smaller RMSEs than the PI estimator, and such improvements get more evident when n is small. For the simulation results on the variance estimators reported in Figure 11, the PI and BC estimators perform better than the NB estimator in most settings. For the newly proposed PI and BC estimators of the variance, the BC estimator always yields smaller RBs and smaller RSLs than the PI estimator. To conclude, with the RB, RMSE and RSL as the criteria, the BC method performs better than the NB and PI methods and can be recommended for practical use.

ACKNOWLEDGEMENTS

The authors thank the editor, the associate editor, and the reviewer for their helpful comments that have led to a significant improvement of the paper. Tiejun Tong's research was supported by the Initiation Grant for Faculty Niche

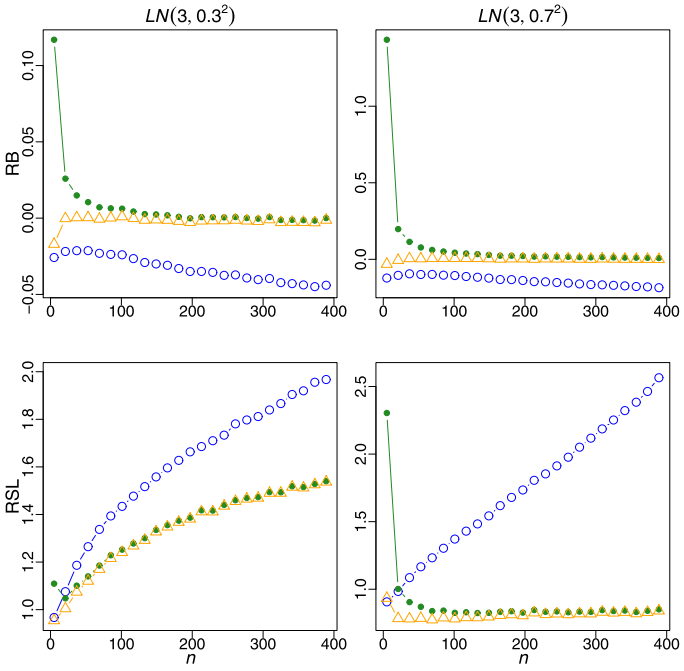

Figure 11. The RBs and RSLs of three types of variance estimators under scenario $\mathcal{S}_3$, where the blue empty points represent the NB estimator, the green solid points represent the PI estimator, and the orange empty triangles represent the BC estimator.

Research Areas (No. RC-IG-FNRA/17-18/13) and the Century Club Sponsorship Scheme of Hong Kong Baptist University, the General Research Fund (No. HKBU12303918), the Health and Medical Research Fund (No. 04150476), and the National Natural Science Foundation of China (No. 11671338).

Received 14 February 2020

REFERENCES

[1] Bland, J. M. (2015). Estimating mean and standard deviation from the sample size, three quartiles, minimum, and maximum. *International Journal of Statistics in Medical Research*, 4, 57–64.

[2] Bland, J. M. and Altman, D. G. (1996). Transformations, means, and confidence intervals. *British Medical Journal*, 312, 1079.

[3] Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B*, 26, 211–243. MR0192611

[4] Chan, T. Y., Poon, P., Pang, J., et al. (1994). A study of calcium and vitamin D metabolism in Chinese patients with pulmonary tuberculosis. *Journal of Tropical Medicine and Hygiene*, 97, 26–30.

[5] Davies, P. D., Brown, R. C. and Woodhead, J. S. (1985). Serum concentrations of vitamin D metabolites in untreated tuberculosis. *Thorax*, 40, 187–190.

[6] Davies, P. D., Church, H. A., Bovornkitti, S. and Charumilind, A. (1988). Altered vitamin D homeostasis in tuberculosis. *Internal Medicine Thailand*, 4, 45–47.

[7] Davies, P. D., Church, H. A., Brown, R. C. and Woodhead, J. S. (1987). Raised serum calcium in tuberculosis patients in Africa. *European Journal of Respiratory Diseases*, 71, 341–344.

[8] Dey, D. K. and Srinivasan, C. (1985). Estimation of a covariance matrix under Stein's loss. *The Annals of Statistics*, 13, 1581–1591. MR0811511

[9] Fang, S., Sui, D., Wang, Y., et al. (2016). Association of vitamin D levels with outcome in patients with melanoma after adjustment for C-reactive protein. *Journal of Clinical Oncology*, 34, 1741–1747.

[10] Feng, C., Wang, H., Lu, N. and Tu, X. M. (2013). Log transformation: application and interpretation in biomedical research. *Statistics in Medicine*, 32, 230–239. MR3041863

[11] Hozo, S. P., Djulbegovic, B. and Hozo, I. (2005). Estimating the mean and variance from the median, range, and the size of a sample. *BMC Medical Research Methodology*, 5, 13.

[12] Grange, J. M., Davies, P. D., Brown, R. C., Woodhead, J. S. and Kardjito, T. (1985). A study of vitamin D levels in Indonesian patients with untreated pulmonary tuberculosis. *Tubercle*, 66, 187–191.

[13] Higgins, J. P. and Green, S. (2008). *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: John Wiley & Sons.

[14] Higgins, J. P., White, I. R. and Anzures-Cabrera, J. (2008). Meta-analysis of skewed data: combining results reported on log-transformed or raw scales. *Statistics in Medicine*, 27, 6072–6092. MR2522311

[15] Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine*, 14, 811–819.

[16] Luo, D., Wan, X., Liu, J. and Tong, T. (2018). Optimally estimating the sample mean from the sample size, median, mid-range, and/or mid-quartile range. *Statistical Methods in Medical Research*, 27, 1785–1805. MR3803266

[17] Nnoaham, K. E. and Clarke, A. (2008). Low serum vitamin D levels and tuberculosis: a systematic review and meta-analysis. *International Journal of Epidemiology*, 37, 113–119. MR2733493

[18] Quan, H., Zhang, J., Zhou, D. and Dukovic, D. (2018). Estimation of standard deviation for a log-transformed variable based on summary statistics in the original scale. *Statistics in Biopharmaceutical Research*, 10, 30–38.

[19] Sasidharan, P. K., Rajeev, E. and Vijayakumari, V. (2002). Tuberculosis and vitamin D deficiency. *Journal of the Association of Physicians of India*, 50, 554–558.

[20] Sheikh, A., Saeed, Z., Jafri, S. A. D., Yazdani, I. and Hussain, S. A. (2012). Vitamin D levels in asymptomatic adults-a population survey in Karachi, Pakistan. *PLoS ONE*, 7, e33452.

[21] Shi, J., Luo, D., Weng, H., et al. (2020). Optimally estimating the sample mean and standard deviation from the five-number summary. *arXiv preprint*arXiv:2003.02130.

[22] Walter, S. D. and Yao, X. (2007). Effect sizes can be calculated for studies reporting ranges for outcome variables in systematic reviews. *Journal of Clinical Epidemiology*, 60, 849–852.

[23] Wan, X., Wang, W., Liu, J. and Tong, T. (2014). Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range. *BMC Medical Research Methodology*, 14, 135.

[24] Xu, G. and Genton, M. G. (2015). Efficient maximum approximated likelihood inference for Tukey's g-and-h distribution. *Computational Statistics & Data Analysis*, 91, 78–91. MR3368007

[25] Xu, Y., Iglewicz, B. and Chervoneva, I. (2014). Robust estimation of the parameters of g-and-h distributions, with applications to outlier detection. *Computational Statistics & Data Analysis*, 75, 66–80. MR3178358

Jiandong Shi
Department of Mathematics
Hong Kong Baptist University
Hong Kong
E-mail address: 18481701@life.hkbu.edu.hk

Tiejun Tong
Department of Mathematics
Hong Kong Baptist University
Hong Kong
E-mail address: tongt@hkbu.edu.hk

Yuedong Wang
Department of Statistics and Applied Probability
University of California – Santa Barbara
Santa Barbara
USA
E-mail address: yuedong@pstat.ucsb.edu

Marc G. Genton
Statistics Program
King Abdullah University of Science and Technology
Thuwal
Saudi Arabia
E-mail address: Marc.Genton@kaust.edu.sa