Estimating the mean of a small sample under the two parameter lognormal distribution

Peter Hingley

Abstract Lognormally distributed variables are found in biological, economic and other systems. Here the sampling distributions of maximum likelihood estimates (MLE) for parameters are developed when data are lognormally distributed and estimation is carried out either by the correct lognormal model or by the mis-specified normal distribution. This is designed as an aid to experimental design when drawing a small sample under an assumption that the population follows a normal distribution while in fact it follows a lognormal distribution. Distributions are derived analytically as far as possible by using a technique for estimator densities and are confirmed by simulations. For an independently and identically distributed lognormal sample, when a normal distribution is used for estimation then the distribution of the MLE of the mean is different to that for the MLE of the lognormal mean. The distribution is not known but can be well enough approximated by another lognormal. An analytic method for the distribution of the mis-specified normal variance uses computational convolution for a sample of size 2. The expected value of the mis-specified normal variance is also found as a way to give information about the effect of the model misspecification on inferences for the mean. The results are demonstrated on an example for a population distribution that is abstracted from a survey.

Peter Hingley
European Patent Office, Munich, Germany
e-mail: phingley@epo.org

Citation: Peter Hingley, Estimating the mean of a small sample under the two parameter lognormal distribution, in R. Anguelov, M. Lachowicz (Editors), Mathematical Methods and Models in Biosciences, Biomath Forum, Sofia, 2018, pp. 100-121,
http://dx.doi.org/10.11145/texts.2018.02.027

1 Introduction

Here some analytic expressions are developed for the distributions of maximum likelihood estimators (MLEs) of parameters of samples from the lognormal distribution. These are described both under a correct lognormal estimation model (EM) and under an incorrect normal EM. The latter situation can occur either because of lack of knowledge of the data generating model (DGM) or because of the simplicity of carrying out statistical inference under the assumption of normality. It may also be that, for a small sample where the statistical assumptions behind the central limit theorem do not apply, asymmetry of the data around the mean is not apparent. Therefore a scientist may be unaware that a variable has a lognormal distribution and so be tempted to measure the arithmetic mean and standard deviation of the sample data in order to use normal inference.

The ideas can be applied in the experimental design phase, when considering the possibility of a different DGM to an EM. By making presumptions about the likely form of the population distribution, then the EM and the sample size can be chosen to give the desired precision of the resulting estimate. At the data analysis stage, other ways to deal with a lack of knowledge of the population distribution include using a robust estimator like the median, or a Student's t test for the mean in the case of a normal distribution with unknown variance.

Lognormally distributed variables are found in biological, economic and other systems. Sometimes it is convenient to calculate statistics directly on the log metric [1] [2] [3] [4]. In this case, straightforward normal theory applies for estimating means and standard errors. It can happen however that the original scale is important. The expression for the MLE of the lognormal mean includes the mean and variance of the associated normal distribution on the log scale. The MLEs are neither unbiased nor efficient in this case [5] and some other estimators are available [6] [10]. But we consider here the situation of straightforward data analysis where the MLEs for mean and variance are used, either under the lognormal EM or under the normal EM. The arithmetic mean, which is the MLE of a normal mean, does not include a variance term.

Exact analytic probability density functions (PDFs) for MLEs under the lognormal EM will be obtained by using a technique for estimator densities (TED) [7] [8] [9]. On the other hand, only approximate forms are developed for the distribution of the arithmetic mean under a lognormal DGM. The analytic PDFs are compared to the empirical PDFs obtained by making simulations with random numbers. Examples are given in the development, firstly for a theoretical illustration and then for a reported distribution of numbers of employees at companies applying for patents.

Section 2 explains TED as an algebraic formula for the PDF of a MLE. Section 3 reviews the exact PDFs of the MLEs of the parameters of the lognormal distribution on lognormally distributed data. Section 4 considers the approximate PDFs of the MLEs of the parameters under the normal distribution on lognormally distributed data. Since this leads to some difficulties even for a sample of size 2, an alternative approach is shown to calculate the expected value of the normal variance estimate. This allows the expected 95 percent range limits for the mean to be found. Section

V discusses an example involving data on the numbers of employees at companies making patent applications from a survey. Section 6 concludes and suggests avenues for further research. Computations were made with R programs.

2 The technique for estimator densities (TED)

This is an exact model based approach to find the density of a MLE, rather than an approximate data based approach such as density estimation where the observed data are used to estimate the distribution [11].

In the following, a term $g()$ indicates a PDF. Consider independently and identically distributed (iid) data that are gathered into a $(n \times 1)$ vector w. In order to obtain the MLEs of the parameters of $g()$, the likelihood of the data is $\prod_{i=1}^{n} g(w_i)$. This is maximised by using the logarithm of the likelihood [12].

Say that $l(\theta, w)$ is the log likelihood of the data under the EM, with p estimable parameters in a $p \times 1$ vector θ. Let $'$ and $''$ indicate differentiation by θ, once or twice respectively. Consider cases where $l(\theta, w)$ is continuous, differentiable and has a single maximum with no other turning point. Then the MLEs $\hat{\theta}$ are given by $l'(\theta, w)|_{\theta=\hat{\theta}} = 0$. There is also the further requirement that $l(\theta, w)$ is differentiable for a second time. It is desired to find $g(\hat{\theta})$. Following [7], consider a $(p \times 1)$ vector T.

$$T(\theta, \theta^*, w) = l'(\theta^*, w) - l'(\theta, w), \tag{1}$$

where θ^* is fixed at an arbitrary value and θ is yet to be specified. Under the regularity conditions that were mentioned above, the exact PDF for $\hat{\theta}$ is given as follows.

$$g(\hat{\theta}) = E_w[|j(\theta, w)||_{\theta=\hat{\theta}}] \cdot g_{[T(\hat{\theta}, \theta^*=\hat{\theta}, w)]}(0) \tag{2}$$

Here $j(\theta, w) = -l''(\theta, w)$ is the observed information. The term $E_w[|j(\theta, w)||_{\theta=\hat{\theta}}]$ describes a conditional expectation, that is conditional on $\theta = \hat{\theta}$ and is taken with respect to w over the EM. The second term represents the value of the PDF $g_{[T(\hat{\theta}, \theta^*, w)]}(t)$, for which $\theta^* = \hat{\theta}$ and $\theta = \hat{\theta}$, so that $t = 0$ by (1).

TED allows for a distinction to be made between the functional forms of the PDFs of the data $g_0(w)$ on the DGM and $g_1(w|\theta)$ on the EM. It can also be used when the functional form of the EM is the same as the DGM.

While TED is useful because it gives the exact PDF of the MLE, from a practical point of view it can only be applied to simple enough models for which the components in equation (2) can be calculated. In order to illustrate how this works, Table 1 shows some previously described examples (from [8]), where a normal EM is used to estimate the mean when the DGM is either normal (with known variance) or negative exponential. In the former case it turns out that $g(\hat{\theta})$ is normal, as is already well known from elementary statistical theory, while in the latter case $g(\hat{\theta})$ has a gamma distribution. The table indicates the terms that combine to give $g(\hat{\theta})$ according to equation (2).

TED is not a panacea, in that the problem of calculating the analytic PDF for $\hat{\theta}$ is transformed into the problem of finding the analytic PDF $g_{[T(\hat{\theta},\theta^*,z)]}(t)$. This can be done for simple PDFs such as those in Table 1. In the cases that are discussed in this paper, the situation is further simplified because $E_w[|j(\theta,w)||_{\theta=\hat{\theta}}] = |j(\theta,w)||_{\theta=\hat{\theta}}$.

Table 1 Examples of the use of TED by equation (2) (from [8]).

EM:- Normal $N(\delta,\eta_0{}^2)$, see equation (3)
Log likelihood $l(\theta,z), \theta=\delta$, see equation (14), set w to z
$\frac{\delta l(\theta,z)}{\delta\delta} = l'(\theta,z) = \frac{1}{\eta_0{}^2}(\sum z_i - [n\delta])$
$T(\hat{\theta},\theta^*) = \frac{1}{\eta_0{}^2}(\sum z_i - [n\delta^*])$
A. $E_z[\vert j(\theta,z)\vert\vert_{\theta=\hat{\theta}}] = j(\theta,z) = \frac{n}{\eta_0{}^2}$
DGM:- Normal $N(\delta_0,\eta_0{}^2)$, see equation (3)
$g_{[T(\hat{\theta},\theta^*,z)]}(t) = \frac{1}{\sqrt{2\pi\frac{n}{\eta_0{}^2}}}\exp[\frac{-\eta_0{}^2}{2n}[t-(\frac{n}{\eta_0{}^2}(\delta_0-\delta^*)]^2]$
B. $g_{[T(\hat{\theta},\theta^*=\hat{\theta},z)]}(0) = \frac{1}{\sqrt{2\pi\frac{n}{\eta_0{}^2}}}\exp[\frac{-n}{2\eta_0{}^2}[(\hat{\delta}-\delta_0)^2]$
A x B. $g(\hat{\theta}) = \frac{1}{\sqrt{2\pi\frac{\eta_0{}^2}{n}}}\exp[\frac{-n}{2\eta_0{}^2}[(\hat{\delta}-\delta_0)^2] = N(\delta_0,\frac{\eta_0{}^2}{n})$
DGM:- Negative Exponential $\frac{1}{v_0}\exp[\frac{-1}{v_0}z]$, $z\geq 0$
$g_{[T(\hat{\theta},\theta^*,z)]}(t) = \frac{\eta_0{}^2(\eta_0{}^2t+n\delta^*)^{n-1}}{v_0{}^n(n-1)!}\exp[\frac{-1}{v_0}(\eta_0{}^2t+n\delta^*)]$, $(\eta_0{}^2t+n\delta^*)\geq 0$
C. $g_{[T(\hat{\theta},\theta^*=\hat{\theta},z)]}(0) = \frac{\eta_0{}^2(n\hat{\delta})^{n-1}}{v_0{}^n(n-1)!}\exp[\frac{-n\hat{\delta}}{v_0}]$, $\hat{\delta}\geq 0$
A x C. $g(\hat{\theta}) = \frac{(n^n)(\hat{\delta}^{n-1})}{v_0{}^n(n-1)!}.\exp[\frac{-n\hat{\delta}}{v_0}]$, $\hat{\delta}\geq 0$. This is Gamma$(\frac{v_0}{n})$

3 Densities of estimators for the lognormal distribution

In this section, results are described when the data are generated by the lognormal distribution and estimated using the MLEs for the lognormal distribution. Most of the results are already known but are redeveloped here using TED to give an integrated approach.

3.1 The lognormal distribution

If a variable z, $-\infty < z < \infty$, has a normal distribution $N(\delta,\eta^2)$, with mean δ and variance η^2, then its PDF is,

$$g(z) = \frac{1}{\sqrt{2\pi\eta^2}} \exp[\frac{-1}{2\eta^2}(z-\delta)^2], \tag{3}$$

with $-\infty < z < \infty$.

If $w = \exp(z)$, $0 < w < \infty$, use of a Jacobian gives the two parameter lognormal PDF $LN(\mu,\sigma^2)$ for w.

$$g(w) = \frac{1}{\sqrt{2\pi\sigma^2}}\frac{1}{w} \exp[\frac{-1}{2\sigma^2}(log(w)-\mu)^2], \tag{4}$$

with $0 < w < \infty$.

The two parameters can be gathered into a parameter vector $\Delta^T{}_{(2x1)} = (\mu,\sigma^2)$. The expected value of w is $\exp(\mu + \frac{\sigma^2}{2})$ [3]. The mean is a function of b as well as of a in $LN(a,b)$, unlike the case of $N(a,b)$ where the mean a is not a function of b.

As an illustration, consider the distribution $LN(-1.5,3)$. The expected value of an observation w from this distribution is $\exp(-1.5+1.5) = 1$. This is an asymmetric PDF, as is shown in Fig. 1.

3.2 The maximum likelihood estimate of the sample mean

Here the MLE of the sample mean of a lognormal distribution is shown, assuming that the variance is known.

For an iid sample, $w_i, i = 1,...,n$, from $LN(\mu,\sigma^2)$, the log likelihood is,

$$\begin{aligned} l(\Delta,w) = &\quad log[\prod_{i=1}^{n} g(w_i)] \\ = &\quad \frac{-n}{2}log(2\pi\sigma^2) - \sum log(w_i) - \frac{1}{2\sigma^2}\sum(log(w_i)-\mu)^2, \end{aligned} \tag{5}$$

where summation signs apply to the sample members from 1 to n and $\Delta^T = (\mu, \sigma^2)$.

Reparameterise from Δ^T to $\theta^T = (\gamma, \sigma^2)$, where $\gamma = \exp(\mu + \frac{\sigma^2}{2})$. This is the mean of the lognormal variable that was given in Section 3.1. We do not bother to parameterise the lognormal variance explicitly. In terms of the new parameters, the log likelihood is

$$l(\theta, w) = \frac{-n}{2} log(2\pi\sigma^2) - \sum log(w_i) - \frac{1}{2\sigma^2} \sum (log(w_i) + \frac{\sigma^2}{2} - log(\gamma))^2 \quad (6)$$

In order to obtain the MLE for γ, the derivative of the log likelihood is taken wrt γ.

$$\frac{\delta l(\theta, w)}{\delta \gamma} = l'(\theta, w) = \frac{1}{\sigma^2 \gamma} . \sum (log(w_i) + \frac{\sigma^2}{2} - log(\gamma)) \quad (7)$$

Assuming that σ^2 is known, the mle $\hat{\gamma}$ is given by $l'(\theta, w)|_{\gamma = \hat{\gamma}} = 0$.

$$\hat{\gamma} = exp(\frac{\sum log(w_i)}{n} + \frac{\sigma^2}{2}) \quad (8)$$

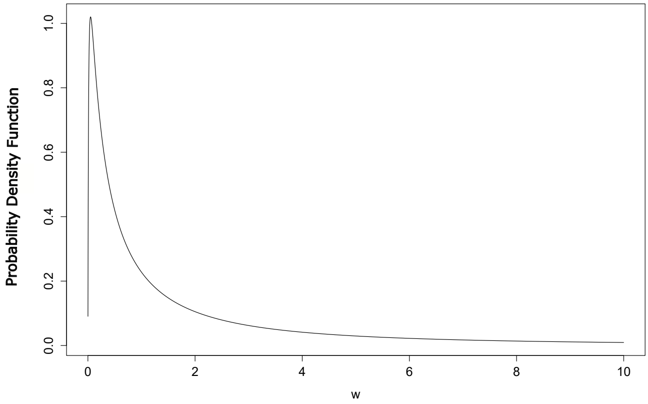

Fig. 1 The lognormal PDF with mean $\gamma = 1$ and lognormal variance term $\sigma^2 = 3$, which is written in terms of $LN(\mu, \sigma^2)$ as $LN(-1.5, 3)$, with $\gamma = \exp(\mu + \frac{\sigma^2}{2})$. This is used for the illustrations in Sections 3 and 4.

3.3 The PDF of the MLE of the sample mean

Here the PDF of $\hat{\gamma}$ is described, assuming that the variance term σ^2 is known. TED will be used to find $g(\hat{\gamma}|\sigma^2)$.

For the lognormal distribution, equation (6) gives,

$$T(\hat{\theta},\theta^*,w) = l'(\theta^*,w) = \frac{1}{\sigma^2\gamma^*}.\sum(log(w_i) + \frac{\sigma^2}{2n} - log(\gamma^*)) \qquad (9)$$

To develop $g_{[T(\hat{\theta},\theta^*,w)]}(t)$, consider the DGM where $\log(w_i) \sim N(\mu_0,\sigma^2) = N(\log(\gamma_0) - \frac{\sigma^2}{2n},\sigma^2)$.

It follows from (9) that, $T(\hat{\theta},\theta^*,w) \sim N\left[\frac{n}{\sigma^2\gamma^*}\log\left(\frac{\gamma_0}{\gamma^*}\right), \frac{n}{\sigma^2\gamma^{*2}}\right]$.

$$g_{[T(\hat{\theta},\theta^*,w)]}(t) = \frac{\sqrt{\sigma^2}\gamma^*}{\sqrt{2\pi n}}\exp\left[\frac{-\sigma^2\gamma^{*2}}{2n}\left[t - \frac{n}{\sigma^2\gamma^*}\log\left(\frac{\gamma_0}{\gamma^*}\right)\right]^2\right]$$

Following on from (7),

$$j(\theta,w) = l''(\theta,w) = \frac{-1}{\sigma^2\gamma^2}.[n + \sum(\log(w_i)) + n\frac{\sigma^2}{2} - n\log(\gamma)]$$

To obtain $E_w[|j(\theta,w)||_{\theta=\hat{\theta}}]$, note that $\sum(\log(w_i)) = nlog(\hat{\gamma}) - n\frac{\sigma^2}{2}$ by equation (8). So $E_w[|j(\theta,w)||_{\theta=\hat{\theta}}] = \frac{n}{\sigma^2\hat{\gamma}^2}$.

For $g(\hat{\gamma}|\sigma^2)$, according to equation (2), set $t = 0$, $\gamma^* = \hat{\gamma}$, in $g_{[T(\hat{\theta},\theta^*,w)]}(t)$ and multiply by $\frac{n}{\sigma^2\hat{\gamma}^2}$.

$$g(\hat{\gamma}|\sigma^2) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma^2}}.\frac{1}{\hat{\gamma}}exp[\frac{-n}{2\sigma^2}(log(\hat{\gamma}) - log(\gamma_0))^2] \sim LN(log(\gamma_0),\frac{\sigma^2}{n}) \qquad (10)$$

Comparison of expressions (10) and (4) shows that the mean of an iid sample from a lognormal distribution with known σ^2 has a lognormal distribution, with mean $\exp(\log(\gamma_0) + \frac{\sigma^2}{2n})$ and a variance term $\frac{\sigma^2}{n}$.

For the illustration that was introduced in 3.1, the middle diagram in Fig. 3 (below) shows a comparison of the PDF specified by equation (10) and a probability histogram derived from simulated data for samples of size $n = 2$ from $LN(-1.5,3)$. This distribution is $LN(1,1.5)$ and has mean $\exp(0 + \frac{1.5}{2}) = 2.12$. Equivalence of the analytic PDF to the simulations is indicated.

3.4 The PDF of the associated sample variance $\hat{\sigma^2}$

By applying the derivative $\frac{\delta l(\theta,w)}{\delta\sigma^2}$ to equation (6), the MLE of σ^2 is $\hat{\sigma^2} = \frac{\sum(log(w_i)-\hat{\mu})^2}{n}$. The analytic unconditional PDF $g(\hat{\sigma^2})$ can be found from equation (3). Say that the true value of σ^2 is ${\sigma_0}^2$. Standard theory [12] shows that the quantity $\frac{n\hat{\sigma^2}}{{\sigma_0}^2}$ has a chi squared distribution with $n-1$ degrees of freedom. So, by transformation,

$$g(\hat{\sigma^2}) = \frac{\frac{n}{{\sigma_0}^2}[\frac{n\hat{\sigma^2}}{{\sigma_0}^2}]^{\frac{n-1}{2}-1}.exp[\frac{-n}{2}\frac{\hat{\sigma^2}}{{\sigma_0}^2}]}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}, \tag{11}$$

where $\Gamma()$ is the Gamma function.

Equation (11) indicates that $g(\hat{\sigma^2})$ does not have to be written as a conditional PDF, because it is independent of γ_0 and $\hat{\gamma}$. Fig. 2 shows this PDF for the illustration with $n = 2$, again comparing the PDF specified by equation (11) with a probability histogram derived from simulated sets of samples. Equivalence of the analytic PDF to the simulations is indicated.

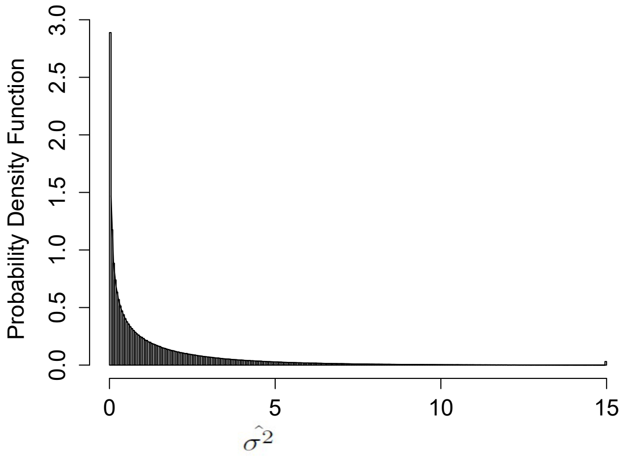

Fig. 2 $g(\hat{\sigma^2})$ for $LN(-1.5,3)$ with $n = 2$. Histogram of one million sample estimates of $\hat{\sigma^2}$. The analytic curve (11) is included as a solid line.

3.5 *The conditional PDF* $g(\hat{\gamma}|\hat{\sigma^2})$

Here the approach in Section 3.3 is extended to obtain the PDF of the sample mean $\hat{\gamma}$ when it is conditional on the sample variance $\hat{\sigma^2}$. TED will be used to find $g(\hat{\gamma}|\hat{\sigma^2})$, that is conditional on the estimate $\hat{\sigma^2}$ from the same data set.

Say that the underlying parameters are γ_0 and ${\sigma_0}^2$. The log likelihood (6) for the estimation model is now written with $\sigma^2 = \hat{\sigma^2}$, while for the DGM the same likelihood is written with $\gamma = \gamma_0$ and $\sigma^2 = {\sigma^2}_0$. The conditional MLE is derived in an analogous way to equation (8).

$$\hat{\gamma}_{|\hat{\sigma^2}} = exp(\frac{\sum log(w_i)}{n} + \frac{\hat{\sigma^2}}{2})$$

T is now obtained as in equation (9).

$$T(\hat{\theta}, \theta^*, w) = \frac{1}{\hat{\sigma^2}\gamma^*} . \sum(log(w_i) + \frac{\hat{\sigma^2}}{2} - log(\gamma^*))$$

To find $g_{[T(\hat{\theta},\theta^*,w)]}(t)$, the DGM gives $\log(w_i) \sim N(\mu_0, {\sigma_0}^2) = N(\log(\gamma_0) - \frac{{\sigma^2}_0}{2}, {\sigma^2}_0)$.
It follows that,

$$T(\hat{\theta}, \theta^*, w) \sim N(\frac{n}{\hat{\sigma^2}\gamma^*}[log(\frac{\gamma_0}{\gamma^*}) + \frac{\hat{\sigma^2} - {\sigma_0}^2}{2}], \frac{n{\sigma_0}^2}{(\hat{\sigma^2})^2\gamma^{*2}})$$

So,

$$g_{[T(\hat{\theta},\theta^*,w)]}(t) = \frac{\hat{\sigma^2}\gamma^*}{\sqrt{2\pi n{\sigma_0}^2}} exp[\frac{-(\hat{\sigma^2})^2\gamma^{*2}}{2n{\sigma_0}^2}[t - [\frac{n}{\hat{\sigma^2}\gamma^*}(log(\gamma_0) - \frac{{\sigma_0}^2}{2}) + \frac{n}{\gamma^*}(\frac{1}{2} - \frac{log(\gamma^*)}{\hat{\sigma^2}})]]^2]$$

The multiplicative expectation term is $E_w[|j(\theta, w)||_{\theta=\hat{\theta}}, \sigma^2 = \hat{\sigma^2}] = \frac{n}{\hat{\sigma^2}\hat{\gamma^2}}$.

For $g(\hat{\gamma}|\hat{\sigma^2})$ according to equation (2), set $t = 0$, $\gamma^* = \hat{\gamma}$, in $g_{[T(\hat{\theta},\theta^*,w)]}(t)$ and multiply by $\frac{n}{\hat{\sigma^2}\hat{\gamma^2}}$.

$$g(\hat{\gamma}|\hat{\sigma^2}) = \frac{\sqrt{n}}{\sqrt{2\pi{\sigma_0}^2}} . \frac{1}{\hat{\gamma}} exp[\frac{-n}{2{\sigma_0}^2}(log(\hat{\gamma}) - [log(\gamma_0) + \frac{\hat{\sigma^2} - {\sigma_0}^2}{2}])^2] \sim LN(log(\gamma_0) + \frac{\hat{\sigma^2} - {\sigma_0}^2}{2}, \frac{{\sigma_0}^2}{n}) \qquad (12)$$

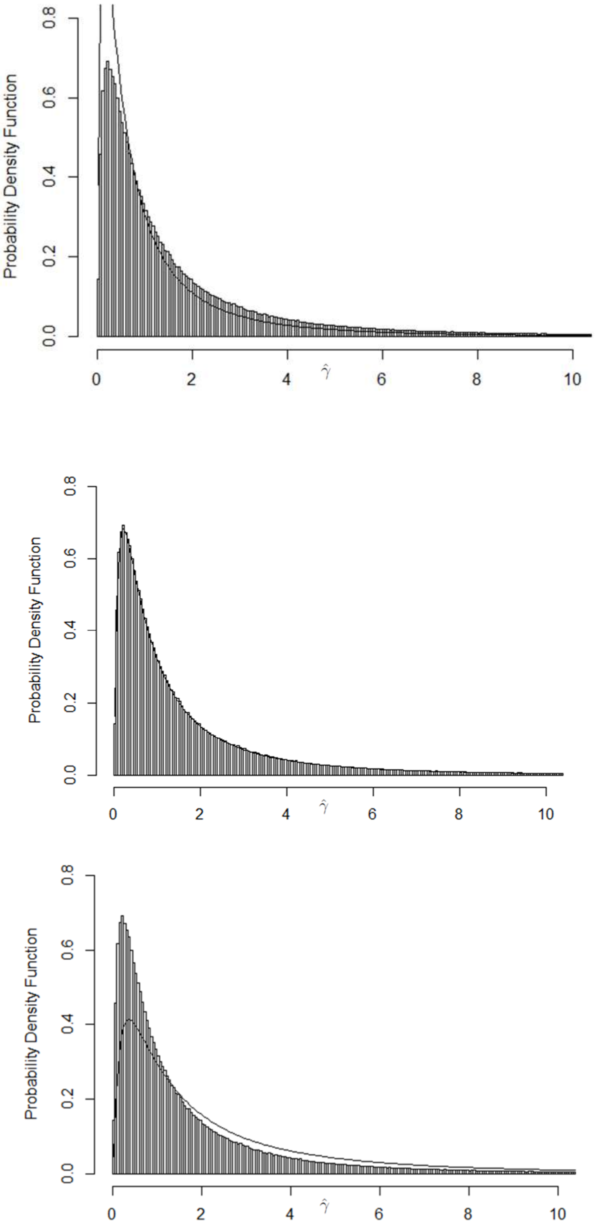

Fig. 3 $g(\hat{\gamma}|\hat{\sigma^2})$ for $LN(-1.5,3)$ with $n = 2$. Histograms of one million sample estimates. The analytic curves (12) are included as solid lines. a) $\hat{\sigma^2} = 2.25$; b) $\hat{\sigma^2} = 3$; c) $\hat{\sigma^2} = 4$.

Equation (12) shows that the mean of an iid sample from a lognormal distribution that is conditional on $\hat{\sigma^2}$ has a lognormal distribution that depends on both ${\sigma_0}^2$ and $\hat{\sigma^2}$. The mean is $\exp(\log(\gamma_0) + \frac{\hat{\sigma^2}}{2n})$, while the variance term $\frac{{\sigma_0}^2}{n}$ does not depend on $\hat{\sigma^2}$. Expression (12) is a generalisation of equation (10).

Figs. 3a to 3c show three variants of $g(\hat{\gamma}|\hat{\sigma^2})$ for the illustration (taking $n = 2$ from LN(-1.5, 3)), corresponding to conditional values for $\hat{\sigma^2}$ of 2.25, 3 and 4 respectively. The data have been generated using ${\sigma_0}^2 = 3$. Agreement of the analytic PDF with the histogram only occurs when $\hat{\sigma^2} = {\sigma_0}^2 = 3$ in the middle diagram (as was discussed in Section 3.3).

3.6 The joint PDF $g(\hat{\gamma},\hat{\sigma^2})$

Here the results of Sections 3.4 and 3.5 are combined to find the joint PDF of $\hat{\gamma}$ and $\hat{\sigma^2}$ from a sample.

The previous section 3.5 showed that, for data sets from the lognormal distribution, in the PDF $g(\hat{\gamma}|\hat{\sigma^2})$ there is a dependency between $\hat{\gamma}$ and $\hat{\sigma^2}$ that needs to be considered. $\hat{\sigma^2}$ will not be the same over several data samples, so the conditional PDF $g(\hat{\gamma}|\hat{\sigma^2})$ may be difficult to interpret. The joint PDF of $\hat{\gamma}$ and $\hat{\sigma^2}$ is of interest in order to better understand the consequences of the model. This is given by multiplying the expressions (11) and (12).

$$g(\hat{\gamma},\hat{\sigma^2}) = g(\hat{\sigma^2}).g(\hat{\gamma}|\hat{\sigma^2}) = \frac{\frac{\sqrt{n}n}{\sqrt{2\pi{\sigma_0}^2\hat{\gamma}{\sigma_0}^2}}}{2^{\frac{n-1}{2}}\Gamma(\frac{n-1}{2})}\left[\frac{n\hat{\sigma^2}}{{\sigma_0}^2}\right]^{\frac{n-1}{2}-1}$$

$$.\exp\left[\frac{-n}{2}\left(\frac{\hat{\sigma^2}}{{\sigma_0}^2}+\frac{1}{{\sigma_0}^2}\left[\log(\hat{\gamma})-\log(\gamma_0)-\frac{{\sigma_0}^2-\hat{\sigma}^2}{2}\right]\right)^2\right] \quad (13)$$

Figure 4 shows this bivariate PDF for the illustration, using simulated data sets (left plot) and the analytic formula (right plot). Agreement of the PDFs is indicated.

For $n = 2$, $g(\hat{\gamma},\hat{\sigma^2})$ descends in both directions with no observable mode. Chi-squared distributions with 1 or 2 degrees of freedom have no mode [3]. This has an effect on the associated PDF $g(\hat{\gamma},\hat{\sigma^2})$ when $n = 2$. Fig. 5 shows that there is a mode for the bivariate PDF with the same model and parameters when $n = 6$, again by simulations (left plot) and by the analytic formula (right plot).

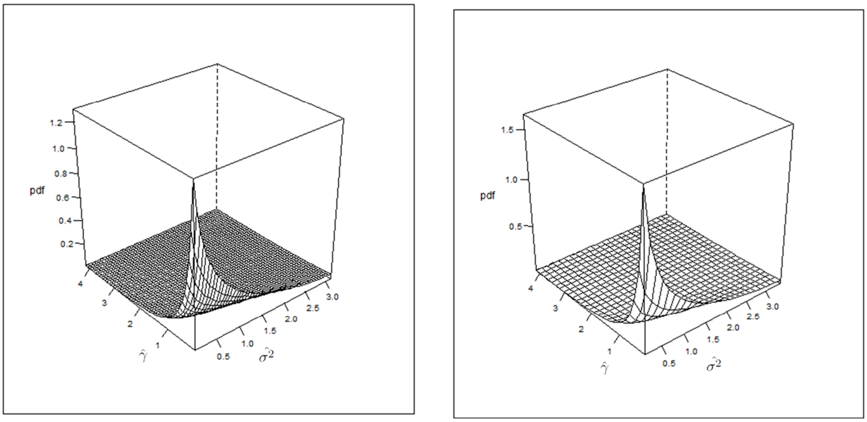

Fig. 4 $g(\hat{\gamma},\hat{\sigma^2})$ for $LN(-1.5,3)$ with $n=2$. a):- based on a histogram of ten million sample estimates; b):- using the analytic formula (13).

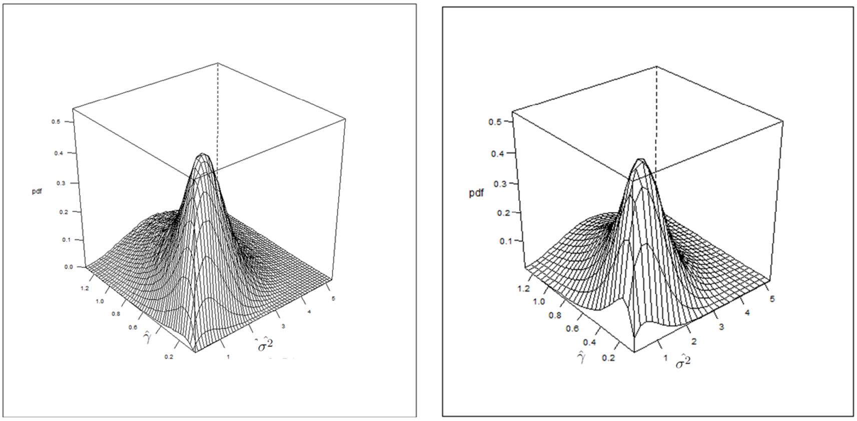

Fig. 5 $g(\hat{\gamma},\hat{\sigma^2})$ for $LN(-1.5,3)$ with $n=6$. a) based on a histogram of ten million sample estimates; b) using the analytic formula (13).

4 Fitting the normal distribution to lognormal data

The consequences will now be described of wrongly using the normal distribution as EM when the DGM is the lognormal distribution. The DGM will be written $LN(\mu_0,{\sigma_0}^2)$ and the EM will be written $N(\delta,\eta^2)$, as in Equation (3) but now in terms of $g(w)$ rather than $g(z)$.

4.1 The conditional PDF $g(\hat{\delta}|\hat{\eta^2})$

Here the PDF of the sample mean will be developed when it is conditional on the sample variance.

For the normal distribution, the log likelihood for δ, conditional on $\hat{\eta^2}$, from equation (3) is

$$l(\delta, w|\hat{\eta^2}) = -nlog(\sqrt{2\pi\hat{\eta^2}}) - \frac{1}{2\hat{\eta^2}}\sum((w_i - \delta)^2) \qquad (14)$$

Differentiating by δ, the MLE for a normal EM is the sample mean $\hat{\delta} = \frac{\Sigma w_i}{n}$. The exact distribution of $\hat{\delta}$ is not known and unfortunately TED does not help here because it also specifies the need to develop an expression for the distribution of $\sum w_i$.

Several methods are available to approximate the distribution. One way is to approximate $g(\hat{\delta}|\hat{\eta}^2)$ by a transformed version of the lognormal distribution $LN(log(\gamma_0), \frac{\sigma^2}{n})$ for $g(\hat{\gamma}|\sigma^2)$. Expression (10) for the distribution of the MLE $\hat{\gamma}$ under the lognormal EM cannot be used directly, because this is for the geometric mean with a correction as at (8), that does not have the same distribution as the arithmetic mean $\hat{\delta}$. The difference can be seen with simulation results for the illustration using the same DGM by comparing the distributions shown in Fig. 6 and Fig. 3b for $n = 2$. The distribution for $\hat{\delta}$ is shifted to the left compared to the one for $\hat{\gamma}$.

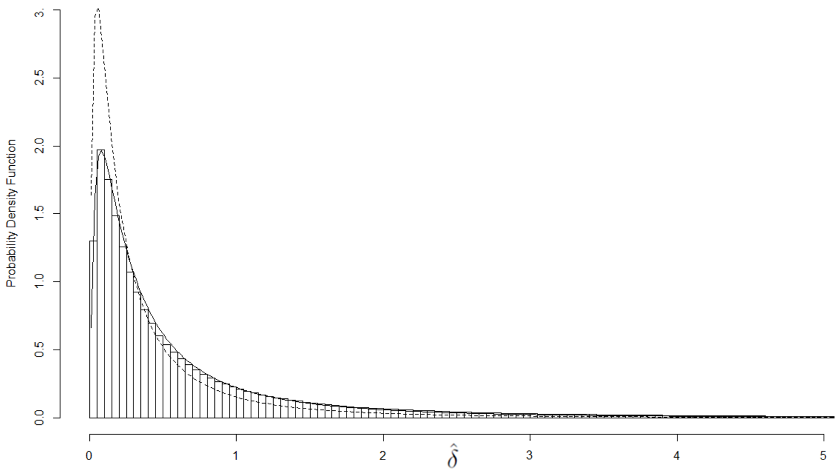

Fig. 6 $g(\hat{\delta}|\sigma^2)$ for the arithmetic mean on data from $LN(-1.5, 3)$ with $n = 2$. Histogram of one million sample estimates. The analytic curve (15) is included as a solid line. The dashed line shows the approximate analytic density $LN(log(\gamma_0) - \frac{\sigma^2}{n}, \frac{\sigma^2}{n})$.

Relating to the lognormal distribution in equation (4), assume that $\hat{\delta}$ estimates $\exp(\mu_0)$ while $\hat{\gamma}$ estimates $\exp(\mu_0).\exp(\frac{{\sigma_0}^2}{2n})$. The choice of a lognormal distribution $g(\hat{\delta}|{\sigma_0}^2) \sim LN(log(\gamma_0) - \frac{{\sigma_0}^2}{n}, \frac{{\sigma_0}^2}{n})$ preserves the same variance $\frac{{\sigma_0}^2}{n}$ as in $g(\hat{\gamma}|{\sigma_0}^2)$ at (10), and gives the mean $\exp(log(\gamma_0) - \frac{{\sigma_0}^2}{2n})$. This is consistent with $\hat{\gamma}$ estimating $\exp(\mu_0).\exp(\frac{{\sigma_0}^2}{2n})$ and $\hat{\delta}$ estimating $\exp(\mu_0)$. However the dashed line in Fig. 6 shows that this gives only an approximate fit when $n = 2$. It was also verified that it gives only an approximate fit to simulations when $n = 6$ with the same parameter values.

Empirical investigation suggests that the following distribution works better for $n = 2$.

$$g(\hat{\delta}|\hat{\eta}^2) = g(\hat{\delta}^2|{\sigma_0}^2) \sim LN(log(\gamma_0) - \frac{{\sigma_0}^2}{2\sqrt{n}}, \frac{{\sigma_0}^2}{n}) \tag{15}$$

This gives the mean $exp(log(\gamma_0) - \frac{{\sigma_0}^2}{2}(\frac{1}{\sqrt{n}} - \frac{1}{n}))$, which is $exp(0 - \frac{3}{2}(\frac{1}{\sqrt{2}} - \frac{1}{2})) = 0.733$ for the illustration with $n = 2$. The solid analytic line in Fig. 6 according to equation (15) closely follows the shape of the histogram. Unlike equation (12) for $g(\hat{\gamma}|\widehat{\sigma^2})$, equation (15) is not directly dependent on its conditional argument $\widehat{\sigma^2}$ and can be written as $g(\hat{\delta})$.

4.2 The pdf of the misfitted sample variance

Here some steps are shown towards developing the PDF of the normal EM sample variance $\hat{\eta}^2$. As in Section 3.6, the intention is then to seek to multiply $g(\widehat{\eta^2})$ by $g(\hat{\delta}|\widehat{\eta^2}) = g(\hat{\delta})$ from Section 4.1, in order to determine the joint PDF $g(\hat{\delta}, \widehat{\eta^2})$.

The log likelihood conditional on δ is written in a similar fashion to equation (14).

$$l(\eta^2, w|\delta) = -nlog(\sqrt{2\pi\eta^2}) - \frac{1}{2\eta^2}\sum((w_i - \delta)^2)$$

Differentiating by η^2,

$$l'(\eta^2, w|\delta) = \frac{-n}{2\eta^2} + \frac{1}{2(\eta^2)^2}\sum((w_i - \delta)^2)$$

From this, the MLE is $\widehat{\eta^2} = \frac{1}{n}\sum((w_i - \delta)^2) = \frac{\Sigma r_i}{n}$, where $r_i = (w_i - \delta)^2$.

Unlike the situation in Section 3.4, the PDF for $\widehat{\eta^2}$ depends on γ_0 as well as ${\sigma_0}^2$. Consider the case $n = 1$. Since $w \sim LN(\mu_0, {\sigma_0}^2)$, with $\mu_0 = log(\gamma_0) - \frac{{\sigma_0}^2}{2}$, the quantity $v_i = \frac{(log(w_i) - \mu_0)^2}{{\sigma_0}^2}$ follows a chi-square distribution on 1 degree of freedom. That is,

$$g(v_i) = \frac{1}{\sqrt{2\pi}}.exp[\frac{-v_i}{2}].v_i{}^{-\frac{1}{2}}, v_i{\geq}0$$

But the analyst assumes that the data follows a normal distribution as EM $N(\delta,\eta^2)$. So he believes that $\frac{(w_i-\delta)^2}{\eta_0{}^2} = \frac{r_i}{\eta_0{}^2}$ follows a chi-square distribution on 1 degree of freedom. However the actual distribution $g(r_i)$ is obtained by writing v_i as $\frac{[log(\sqrt{r_i}+\delta)-\mu_0]^2}{\sigma_0{}^2}$.

In order to transform the chi-square PDF for v_i to the PDF for r_i, use the Jacobian,

$$|\frac{dv_i}{dr_i}| = |\frac{log(\sqrt{r_i}+\delta)-\mu_0}{\sigma_0{}^2(\sqrt{r_i}+\delta)\sqrt{r_i}}|$$

Hence,

$$g(r_i) = \frac{1}{\sqrt{2\pi\sigma_0{}^2}(\sqrt{r_i}+\delta)\sqrt{r_i}}.exp(\frac{-1}{2\sigma_0{}^2}.[log(\sqrt{r_i}+\delta)-\mu_0]^2) \qquad (16)$$

Following (15), set $\delta = exp(log(\gamma_0) - (\frac{\sigma_0{}^2}{2\sqrt{n}}))$. When $n = 1$, Fig. 7 shows this for $\sigma_0{}^2 = 3$, $\delta = exp(\mu_0) = exp(0-\frac{3}{2}) = -1.5$. The left plot shows simulations and the right plot shows $g(r_i)$ according to equation (16). Agreement of these plots seems likely. The distribution has a long upper tail.

$g(r_i)$ in equation (16) can be converted to an analytic PDF for $n = 2$ using computational convolution. The PDF $g(\hat{\eta^2})$ for the MLE $\hat{\eta^2} = \frac{1}{2}\sum((w_i-\delta)^2) = \frac{1}{2}\sum r_i$ is as follows.

$$g(\hat{\eta^2}) = 2g(\sum r_i(\hat{\eta^2})) = 2\int_0^\infty g(s_i)(2\hat{\eta^2}-s_i)ds_i$$

Simulations can also be done by adding two appropriate random numbers for each sample member. The resulting PDFs are shown in Fig. 8, with the simulations on the left and the analytic PDF on the right part of the figure.

There are some difficulties emulating the behaviour at the lower bound with the analytic method when compared to the simulations. No attempt has yet been made to expand this approach to higher values of n.

4.3 The joint PDF $g(\hat{\delta},\hat{\eta^2})$

In this Section, the joint PDF of $\hat{\delta}$ and $\hat{\eta^2}$ from a sample is shown. The two-way plots that represent $g(\hat{\delta},\hat{\eta^2})$ are shown for simulations only in Fig. 9, for $n = 2$ and for $n = 6$. There is a very long upper tail for the variance estimates, most prominently for $n = 2$ with its "torpedo trail", but also with a more extended and diffuse "hill" when $n = 6$.

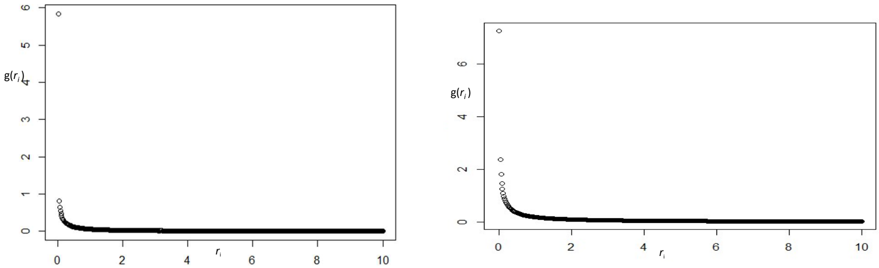

Fig. 7 $g(r_i)$ for the DGM $LN(-1.5,3)$ with $n=1$. a) based on a histogram of ten million simulations; b) using formula (16).

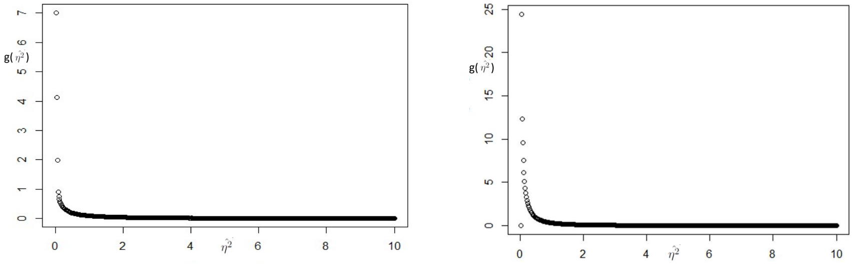

Fig. 8 $g(\hat{\eta^2})$ for the DGM $LN(-1.5,3)$ with $n=2$. a) based on a histogram of ten million simulations; b) using computational convolution from the formula (16).

Fig. 9 can be compared with Fig. 4 and Fig. 5 for the correct lognormal EM. But a direct comparison of the spread of $\hat{\sigma^2}$ with that of $\hat{\eta^2}$ needs to take account of the different scales involved for the variance terms.

4.4 Expected value of the misfitted normal variance

Given the difficulties in constructing the analytical form of the two way plots for $g(\hat{\delta},\hat{\eta^2})$, in this section another analytical approach is taken for the case $n=2$.

Attention will be restricted to the effects of the wrong estimation model on the expected value of the variance term $E[\hat{\eta^2}] = \frac{E[\sum r_i]}{2}$, when $n=2$. This is the same as $E[r_i]$ due to independence of the sample members. Say that $E[r_i]=E[r]$.

From equation (16),

$$E[r] = \frac{1}{\sqrt{2\pi\sigma_0{}^2}}.\int_0^\infty \frac{s}{(\sqrt{s}+\delta)\sqrt{s}}.exp(-\frac{1}{2\sigma_0{}^2}.[log(\sqrt{s}+\delta)-\mu_0]^2)ds$$

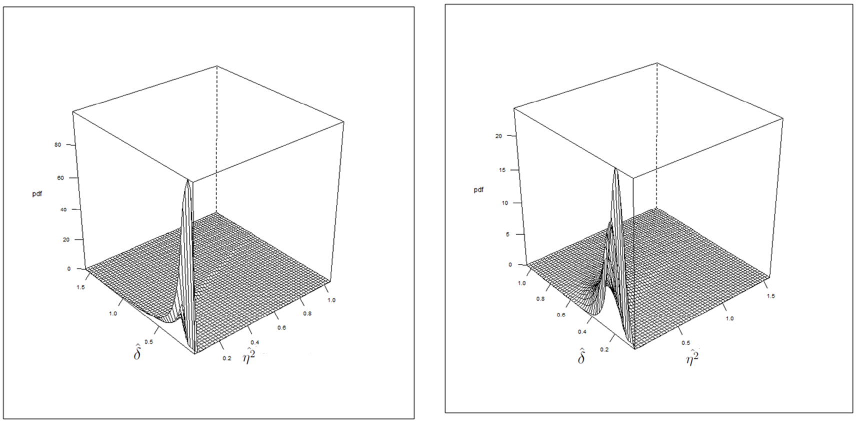

Fig. 9 $g(\hat{\delta},\hat{\eta^2})$ for the numerical illustration with misspecification, based on a histogram of ten million sample estimates. a):- $n=2$; b):- $n=6$.

By making the substitution $u=\sqrt{s}+\delta$, using the positive square root only, this can be written as follows.

$$E[r]=\frac{1}{\sqrt{2\pi\sigma_0{}^2}}.\int_{\delta}^{\infty}2(u-\delta)^2.\frac{1}{u}.exp(\frac{-1}{2\sigma_0{}^2}.[log(u)-\mu_0]^2))du \tag{17}$$

Let $E_{\delta}[m(u)(\mu,\sigma_0{}^2)]$ indicate the incomplete expectation of $m(u)$ under $LN(\mu_0,\sigma_0{}^2)$, where $u\geq\delta$. Equation (17) can be written as,

$$E[r]=E_{\delta}[2(u-\delta)^2(\mu_0,\sigma_0{}^2)]$$

The following expression for the incomplete moments of the lognormal distribution will be used [2] [13].

$$E_{\delta}[u^k(p,q)]=\int_{\delta}^{\infty}x^kLN_x(p,q)dx=1-\mu_k\Phi(\frac{log(\delta)-p-kq}{\sqrt{r}}) \tag{18}$$

Here k is the order of the incomplete moment, Φ is the cumulative distribution function of the standard normal distribution $N(0,1)$ (from $-\infty$ to the argument), and μ_k is the corresponding complete moment $\exp[kp+\frac{k^2}{2}q]$.

Equation (17) can be split into three terms of this type.

$$\begin{aligned}&E_{\delta}[2(u-\delta)^2(\mu_0,\sigma_0{}^2)]\\ =\quad &E_{\delta}[2u^2(\mu_0,\sigma_0{}^2)]-E_{\delta}[4u\delta(\mu_0,\sigma_0{}^2)]+E_{\delta}[2\delta^2(\mu_0,\sigma_0{}^2)]\end{aligned} \tag{19}$$

The evaluated expression is found by using (18) and (19).

$$
\begin{aligned}
E[\widehat{\eta^2}] = E[r] = 2\exp[2\mu_0 + 2{\sigma_0}^2] & \left[1 - \Phi\left(\frac{\log(\delta) - \mu_0 - 2{\sigma_0}^2}{\sqrt{{\sigma_0}^2}}\right)\right] \\
-4\delta.\exp\left[\mu_0 + \frac{{\sigma_0}^2}{2}\right] & \left[1 - \Phi\left(\frac{\log(\delta) - \mu_0 - {\sigma_0}^2}{\sqrt{{\sigma_0}^2}}\right)\right] \\
+2\delta^2. & \left[1 - \Phi\left(\frac{\log(\delta) - \mu_0}{\sqrt{{\sigma_0}^2}}\right)\right]
\end{aligned}
\tag{20}
$$

$E[\widehat{\eta^2}]$ is the expectation of the estimate of the variance on the incorrect normal EM, although the distribution of $\widehat{\eta^2}$ is asymmetric as can be seen in Fig. 8.

Under the running example with $\gamma_0 = 1$, ${\sigma_0}^2 = 3$ and $n = 2$, the 95 percent range for $\hat{\gamma}$ under the cumulative distribution function (CDF) from the correct lognormal EM, which is $g(\hat{\gamma}|{\sigma_0}^2)$ as in Equation (10), was obtained numerically as approximately (0.064, 14.64). The exact 95 percent range for $\hat{\delta}$ under the CDF from the incorrect normal EM, which is $g(\hat{\delta}|\hat{\eta}^2)$ as in Equation (15), was obtained numerically as (0.019, 4.93).

For $n = 1$, the lognormal density (15) has $\mu_0 = log(\gamma_0) - \frac{\sigma^2}{2} = 0 - \frac{3}{2} = -1.5$. Setting δ to $exp(\frac{-\sigma^2}{2n})\gamma_0 = exp(\frac{-3}{2}).1 = 0.2231$, this gives $\sqrt{E[\widehat{\eta^2}]} = 1.3597$ by equation (20). On the normal EM, normal inference gives expected 95 percent confidence limits as $\hat{\delta}$ +/- $1.96\sqrt{E[\widehat{\eta^2}]/2}$, which are (-1.66, 2.11). These limits for $\hat{\delta}$ would be wider in case Student's t distribution based limits were to be used because of the small sample size. Since the negative value of the lower limit is unrealistic, in the next section only one sided upper range limits and confidence limits will be considered.

5 Example of numbers of employees data

Here the above results are illustrated on a set of survey data about numbers of employees in companies that made applications for patents at the European Patent Office (EPO) in 2015 [14]. This distribution is asymmetric. Fig. 10 shows the data. The mean is estimated in the survey report as 2174 employees, while the median is 95 employees. Based on this and the frequency classes in the figure, a lognormal distribution was fitted with $\gamma_0 = 2174$, ${\sigma_0}^2 = 6.1815$ and $\mu_0 = 4.59$.

Assuming that this lognormal distribution describes the population, consider designing an experiment where data will be collected from a random sample of $n = 2$ companies, say those in some particular country or industry. Fig. 11 shows the joint PDF $g(\hat{\gamma}, \widehat{\sigma^2})$ that was obtained analytically. It also shows the PDF for $n = 6$.

Fig. 12 shows the bivariate PDF $g(\hat{\delta}, \widehat{\eta^2})$ that is found by using the wrong normal estimation model by simulations.

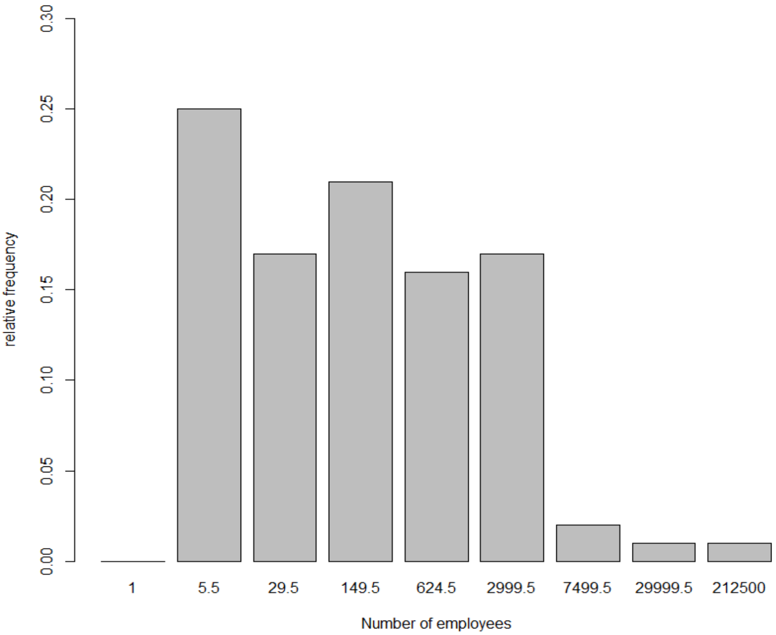

Fig. 10 The frequency distribution of numbers of employees per applicant for patents from survey data [14]. Note that this representation gives equal weight to the grouped classes and is not arithmetic.

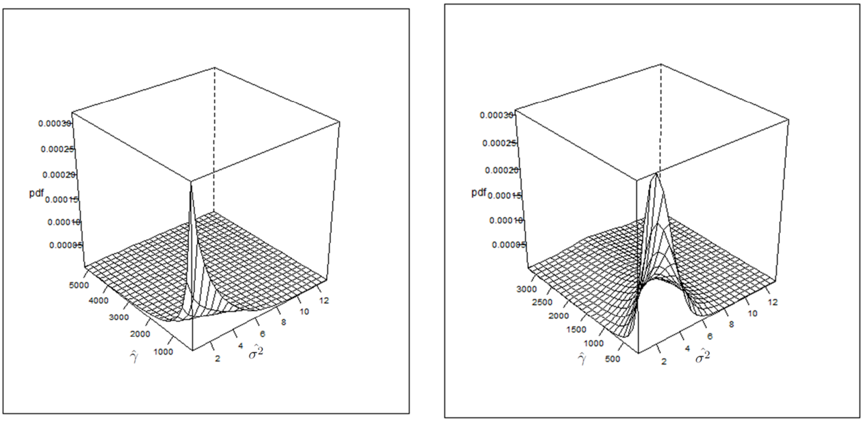

Fig. 11 $g(\hat{\gamma},\hat{\sigma^2})$ for the model for numbers of employees. a):- $n = 2$; b):- $n = 6$.

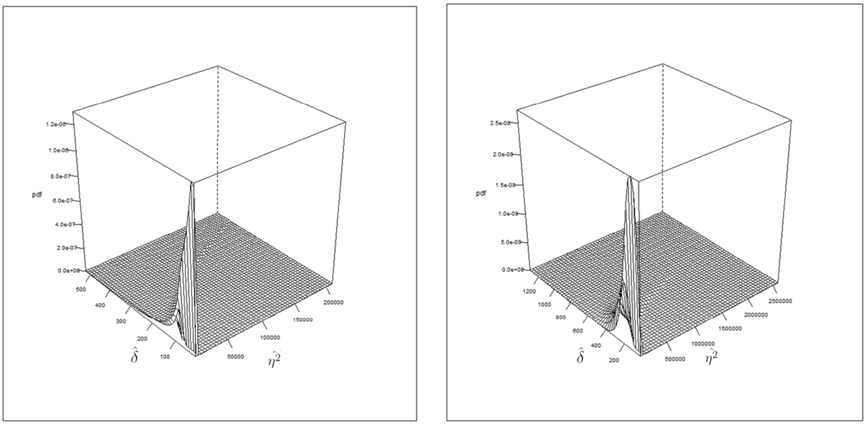

Fig. 12 $g(\hat{\delta}, \hat{\eta^2})$ for the model for numbers of employees. a):- $n = 2$; b):- $n = 6$.

For $n = 2$, the expected one sided upper range limit for 95 percent of the sample means using the correct lognormal EM, under $g(\hat{\gamma}|{\sigma_0}^2)$ as in Equation (10), was obtained numerically as 73019. As was discussed in Section 4.4, for usage in equation (20) δ can be set to $\exp(\frac{-6.1815}{2}) \times 2174 = 98.9$. This value of δ gives $\mu_0 = \log(\delta) - \frac{\sigma^2}{2} = 4.593 - \frac{6.1815}{2} = 1.503$. The expected one sided upper 95 percent range limit for $\hat{\delta}$ under the CDF from the corresponding lognormal distribution (15) was obtained numerically as 5241. The normal variance estimate, $E[\hat{\eta^2}]$ from equation (20), has a square root of 3069, giving an expected 95 percent one sided upper confidence limit of $98.85 + 1.65 \times 3069/\sqrt{2}$, which is 3680. This would be higher in case a Student's t distribution based limit was used. See Fig. 13.

There are other ways to calculate an expected 95 percent one sided upper confidence limit for the mean. The variance of the lognormal distribution is $[\exp(\sigma^2) - 1]\exp(2\mu + \sigma^2)$ [1]. With $\mu = 4.595$ and $\sigma^2 = 6.1815$, the square root of this variance is 2172, which suggests a one sided upper range limit of only 2633, although this could be made larger by using a Student's t distribution based limit.

6 Conclusions

The above approaches demonstrate the effect of misspecifying the normal model for estimation on data that were generated by the lognormal distribution. This can be useful at the experimental design stage where model robustness issues may be of concern. While the context of a data set may sometimes give knowledge about the DGM, in other cases this will not be known. Clearly the distribution of the MLE of the lognormal mean can differ considerably from that of the arithmetic mean with consequences for the statistical inferences from a sample. Inferences that are made

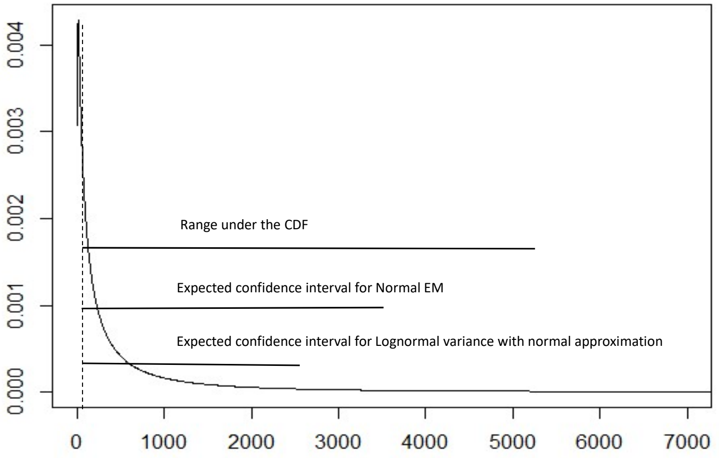

Fig. 13 Ranges for the mean up to one sided 95 percent limits for the employees data (with $n = 2$), by three methods. The distribution (15) is also shown, with the vertical dotted line representing δ.

about the population mean may be more than trivially different under the alternative estimation models.

The analytic results that were obtained give a better handle from which to make calculations than having to depend on simulation results. However neither the pragmatic approximation for the density of the misspecified normal mean nor the analytic method to obtain the density of the MLE for the misspecified normal variance by convolutions have yet been fully developed. The suggested technique in Section 4.4, to calculate $E[\hat{\eta^2}]$ under misspecification, has some promise.

These approaches should be extended to the case $n > 2$. It would also be interesting to extend the techniques to other distributions, in particular the gamma distribution where the shape is explicitly parameterised. The behaviour of estimators other than the MLE could be considered as well.

A further application with the employees data could be to compare the distributions from successive surveys to see whether they differ significantly. In biology, underlying mechanisms that involve multiplicative factors can often justify use of the lognormal distribution for estimation [15]. Even when a data set is only slightly asymmetric, the underlying DGM may be better described by a lognormal distribution than by a normal distribution.

Acknowledgment

The author would like to thank the reviewers and the editor for comments that were used to improve the paper.

References

1. Wikipedia, *Log-normal distribution*, `https://en.wikipedia.org/wiki/Log-normaldistribution`, 2017.
2. J. Aitchison and J. Brown, *The lognormal distribution, with special reference to its use in economics*. Cambridge, 1963.
3. N. Johnson, S. Kotz and N. Balakrishnan, *Continuous Univariate Distributions*, Vol. 1. London: Wiley, 1994.
4. E. Crow and L. Shimizu, *Lognormal distributions, theory and application.* New York: Marcel-Dekker, 1988.
5. N. Longford, *Inference with the lognormal distribution.* Journal of Statistical Planning and Inference, 139, 2329-2340, 2009.
6. B. Ginos, *Parameter estimation for the lognormal distribution*. Brigham Young University Scholars archive. http://scholarsarchive.byu.edu/etd/1928/, 2009.
7. P. Hingley, *Analytic estimator densities for common parameters under misspecified models*, in: M. Hubert, G. Pison and S. Van Aelst (eds), Theory and applications of recent robust methods. Statistics for Industry and Technology, Basel: Birkhauser, 2004.
8. P. Hingley, *Distributions of maximum likelihood estimators and model comparisons*, in A. Korsunsky (ed), Current themes in Engineering Science 2007. American Institute of Physics, 2008.
9. P. Hingley, *Applications and Extensions of a Technique for Estimator Densities.* IAENG Journal of Applied Mathematics, 39:1, 2009.
10. C. Land, *An evaluation of approximate confidence interval estimation methods for lognormal means*. Technometrics, 14, 145-158, 1972.
11. Wikipedia, *Density estimation.* `https://en.wikipedia.org/wiki/Densityestimation`, 2017.
12. P. Hoel, *Introduction to mathematical statistics*. London: Wiley, 1965.
13. *Useful facts about lognormal distribution.* `www.komkon.org/~tacik/science/lognorm.pdf`, 2017.
14. European Patent Office, *Patent Filings Survey 2016.* `http://www.epo.org/service-support/contact-us/surveys/patent-filings.html`, 2017.
15. E. Limpert and W. Stahel, *The log-normal distribution.* Significance, 14, 1, 8-9, 2017.