Sains Malaysiana 52(1)(2023): 295-304
http://doi.org/10.17576/jsm-2023-5201-24

Approximation of the Sum of Independent Lognormal Variates using Lognormal Distribution by Maximum Likelihood Estimation Approached

(Penghampiran terhadap Jumlah Variat Tak Bersandar menggunakan Taburan Lognormal Berdasarkan Pendekatan Penganggaran Kebolehjadian Maksimum)

ABDUL RAHMAN OTHMAN[1], LAI CHOO HENG[2], SONIA AÏSSA[3] & NORA MUDA[4,*]

[1]*School of Distance Education, Universiti Sains Malaysia, 11800 Pulau Pinang, Malaysia*
[2]*Kolej Vokasional Nibong Tebal, Jalan Bukit Panchor, 14300 Nibong Tebal, Pulau Pinang, Malaysia*
[3]*Institut National de la Recherche Scientifique, Énergie Matériaux Télécommunications Research Centre, 800, De La Gauchetière Ouest, Bureau 6900, Montréal, Québec H5A 1K6, Canada*
[4]*Department of Mathematical Sciences, Faculty of Science and Technology, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor Darul Ehsan, Malaysia*

Received: 11 March 2022/Accepted: 10 October 2022

ABSTRACT

Three methods of approximating the sum of lognormal variates to a lognormal distribution were studied. They were the Wilkinson approximation, the Monte Carlo version of the Wilkinson approximation and the approximation using estimated maximum likelihood lognormal parameters. The lognormal variates were generated empirically using Monte Carlo simulation based on several conditions such as number of lognormal variates in the sum, number of sample points in the variates, the variates are independent and identically distributed (IID) and also not identically distributed (NIID) with lognormal parameters. Evaluation of all three lognormal approximation methods was performed using the Anderson Darling test. Results show that the approximation using estimated maximum likelihood lognormal parameters produced Type I errors close to the 0.05 target and is considered the best approximation.

Keywords: Anderson-Darling test; lognormal approximation; maximum likelihood; sum of lognormal variates; Wilkinson

ABSTRAK

Tiga kaedah penghampiran bagi jumlah variat lognormal terhadap taburan lognormal telah dikaji. Tiga kaedah penghampiran tersebut adalah kaedah penghampiran Wilkinson, kaedah versi Monte Carlo bagi penghampiran Wilkinson dan kaedah penghampiran dengan penganggaran kebolehjadian maksimum bagi parameter lognormal. Pemboleh ubah lognormal dijana secara empirik melalui simulasi Monte Carlo dengan beberapa keadaan simulasi iaitu bilangan jumlah pemboleh ubah lognormal, bilangan sampel bagi pemboleh ubah lognormal, pemboleh ubah lognormal tak bersandar dan tertabur secara secaman mengikut taburan (IID) dan juga tidak secaman mengikut taburan (NIID) berdasarkan parameter lognormal. Penilaian bagi ketiga-tiga kaedah penghampiran lognormal tersebut dijalankan menggunakan ujian Anderson Darling. Hasil menunjukkan penghampiran menggunakan penganggaran kebolehjadian maksimum terhadap parameter lognormal telah menghasilkan ralat Jenis 1 menghampiri nilai sasaran ralat 0.05 dan dikatakan sebagai penghampiran terbaik.

Kata kunci: Jumlah variat lognormal; kebolehjadian maksimum; penghampiran lognormal; ujian Anderson-Darling; Wilkinson

INTRODUCTION

The lognormal distribution is a continuous distribution. It is a probability distribution that the logarithm of a random variable is distributed to normal. According to Limpert, Stahel and Abbt (2001), a random variable y is said to be lognormally distributed if and only if $y \sim \log(x)$ and indirectly the variable x is also lognormally distributed. In other words, if variable y is normally distributed, then $x = e^{-y}$ will be lognormally distributed.

In general, the lognormal distribution has two parameters, namely the mean μ and standard deviation σ. Assume a random variable X is lognormally distributed or a variable $\ln(X)$ is normally distributed, that is $X \sim LN(\mu,\sigma^2)$, or $\ln(X) \sim N(\mu,\sigma^2)$, then the probability density function for a lognormal distribution with parameters μ and σ is given by Bromideh (2012)

$$g(x \mid \mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}x}\exp(\frac{(\ln(x)-\mu)^2}{2\sigma^2}), \qquad x>0$$

with $\mu > 0$ and σ>0. The maximum likelihood estimation for μ and σ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln(x_i) \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(\ln(x_i)-\hat{\mu})^2$$

The lognormal distribution has been used to model a variety of phenomena such as the terminal fragment length of type II polyketide synthase (PKS) genes found in soil bacteria collected from New Jersey and Uzbekistan by Wawrik et al. (2007), the productivity of collaboration at research institutes in Germany by Havemann, Heinz and Kretschme (2006), the stationary distribution of spine sizes of individual neurons by Loewenstein, Kuras and Rumpel (2011), the time taken for gastric cancer tumor to develop into a clinical case after passing the point of no return, thus, hypothesizing the number of years to eradicate *Helicobacter pylori*, the risk factor for gastric cancer, infection in a population by Osborn et al. (2013), rate distributions in paleontology by Wagner (2011), customer demand in inventory management by Cobb, Rumí and Salmerón (2012), fuzzy number approaches to describe life time data which are more suitable to describe by lognormal distribution with three estimator (Shafiq, Alamgir & Atif 2016) and modeling the periodic change in the interest rate of a given maturity in actuarial science by Becker (1991). Muhammad Farouk, Nazrina and Zakiyah (2020) have applied the lognormal distribution on the new two-sided group chain sampling plan which operates with four acceptance criteria while Abdul Majid and Ibrahim (2021) have found that the income distributions in Malaysia can be best described by the lognormal-Pareto(II) model.

The lognormal distribution is also of great importance in wireless communication, e.g., total co-channel interference signal received at a given location (Cardieri & Rappaport 2000), reduction in signal strength caused by signal shadowing (Beaulieu & Xie 2004), fading in ultra wideband communication channels (Saleem, Sieskul & Kaiser 2006), and cognitive radio networks (Di Renzo et al. 2009). Specific to the discipline, the interest is in the sum of lognormal variates. The sum of the lognormal distribution is numerically difficult to compute as it has no known closed form. It is well known that the sum of lognormal variates do not result in another lognormal variate. However, in wireless communication, this sum should be approximated as such Schwartz and Yeh (1982) proposed several approximations of the distribution including Wilkinson's, Schwartz's and Farley's methods. There has also been numerous publications in the discipline with regard to this. To name a few are Beaulieu and Xie (2004), Cardieri and Rappaport (2000), Santos Filho, Cardieri and Yacoub (2005), and Santos Filho, Yacoub and Cardieri (2006).

According to Cardieri and Rappaport (2000), the sum of lognormal variates has some important applications in the field of wireless communication. For example, in a cellular communication system, if a special shadowing effect is considered, then the total received co-channel interference signal for a location that sent from unwanted co-channel base stations is usually modeled as the sum of signal and distributed to lognormal. Furthermore, the exact distribution for the signal which is the sum of the lognormal variates of signal is unknown. However, it is also acceptable that the distribution of this signal can be estimated well with other lognormal distributions.

In general, through this description, it can be emphasized that this sum of lognormal variates is used a lot in wireless communication. However, in Santos Filho, Cardieri and Yacoub (2005), and Santos Filho, Yacoub and Cardieri (2006), it was found that the sum of lognormal variate does not show the same results as the other lognormal variate although the sum of the lognormal variates should be estimated equal to other lognormal variates in the wireless relationship. Therefore, through this research, the problem related to the sum of lognormal variate which is not estimated to be equal to the other lognormal variates is needed to be studied and examined.

A thorough discussion of the lognormal approximation of the sum of lognormal distributions can be found in Cardieri and Rappaport (2000). The approach is as follows. Consider $Y_i \sim N(\mu_i,\sigma_i^2)$ for i = 1, 2, …, N. Let $X_i = \exp(Y_i)$. Then, X_i is a lognormal variate with parameters μ_i and σ_i, where $-\infty < \mu_i < \infty$, $\sigma_i > 0$. Note that X_1, …, X_N are independent and non-identical. Consider

$$W = \sum_{i=1}^{N} X_i . \qquad (1)$$

According to Wilkinson cited by Beaulieu and Xie (2004), Cardieri and Rappaport (2000), Santos Filho,

Cardieri and Yacoub (2005), and Santos Filho, Yacoub and Cardieri (2006), W is approximately lognormal with

$$E(W) = E\left(\sum_{i=1}^{N} X_i\right) \tag{2}$$

and

$$E(W^2) = E\left[\left(\sum_{i=1}^{N} X_i\right)^2\right]. \tag{3}$$

The parameter values of μ_W and σ_W are then calculated from (2) and (3). When $W \sim$ lognormal (μ_W, σ_W), the left-hand sides of (2) and (3) become

$$E[W] = \exp\left(\mu_W + \frac{1}{2}\sigma_W^2\right) \tag{4}$$

and

$$E\left[W^2\right] = \exp\left(2\left(\mu_W + \sigma_W^2\right)\right), \tag{5}$$

respectively. The right-hand sides of (2) and (3) can be expressed as

$$E\left[\sum_{i=1}^{N} X_i\right] = \sum_{i=1}^{N} \exp\left(\mu_i + \frac{1}{2}\sigma_i^2\right) = u_1 \tag{6}$$

and

$$E\left[\left(\sum_{i=1}^{N} X_i\right)^2\right] = E\left[\sum_{i=1}^{N} X_i^2 + 2\sum_{i<<j} X_i X_j\right] \tag{7}$$

$$= \sum_{i=1}^{N} \exp\left(2\left(\mu_i + \sigma_i^2\right)\right) + 2\sum_{i<j} \exp\left(\mu_i + \mu_j + \frac{1}{2}\left(\sigma_i^2 + \sigma_j^2\right)\right)$$
$$= u_2,$$

respectively. Set (4) = (6) and (5) = (7) and solve for μ_W and σ_W in terms of u_1 and u_2. We get

$$\mu_W = 2\ln u_1 - \frac{1}{2}\ln u_2 \tag{8}$$

and

$$\sigma_W = \sqrt{\ln u_2 - 2\ln u_1} \tag{9}$$

Cardieri and Rappaport (2000) continued on to obtain μ_W and σ_W through Monte Carlo simulations and compared them against the two parameters calculated from the Wilkinson approximation.

EVALUATION OF THE DISTRIBUTION OF THE SUM OF LOGNORMAL VARIATES

On the left-hand side of Equation (1), the cumulative distribution function of the lognormal distribution with the parameters obtained in Equations (8) and (9) is then derived. On the right-hand side of Equation (1), the exact or close to the exact cumulative distribution function of the sum of the lognormal variates is derived analytically using the characteristic function approach. The derivation of either one of these functions becomes a numerical problem. Several approaches to solve this numerical problem are discussed in Beaulieu and Xie (2004). In addition to the numerical work done on the derivation of the exact or close to exact cumulative distribution function of the sum of lognormal variates, there has been approximate work involving a mixture of truncated exponential functions on this side of the equation proposed by Cobb, Rumí and Salmerón (2012). The lognormal cumulative distribution function of W is then compared against the exact or close to the exact or the approximated cumulative distribution function of $\sum_{i=1}^{N} X_i$.

The next step involves the evaluation of the closeness of the lognormal distribution of W and the cumulative distribution of $\sum_{i=1}^{N} X_i$. Beaulieu and Xie (2004), Santos Filho, Cardieri and Yacoub (2005), and Santos Filho, Yacoub and Cardieri (2006) simply drew both cumulative distributions and if one is superimposed on the other then the approximation is deemed excellent. Cobb, Rumí and Salmerón (2012) evaluated both cumulative distributions by using a cost efficiency equation. If the fit is good at a low cost then the approximation is deemed excellent. Recently, in a similar work on mixed Gaussian distributions, Selim et al. (2016) used two other evaluation methods: Mean square error of cumulative distribution functions and the Kullback-Liebler divergence index.

ANDERSON-DARLING TEST

The standard statistical methodology that is available in this regard is the test of hypothesis. The goodness-of-fit test, in general, measures how well the data corresponds to the fitted model. There are several goodness-of-fit tests available for testing the lognormal distribution. Based on recent studies on the goodness-of-fit tests on preliminary testing of normality by Keselman, Othman and Wilcox (2014, 2013), and Othman, Keselman and Wilcox (2015), the Anderson-Darling test was found to be the powerful test compared to the Kolmogorov-Smirnov and Cramer-von Mises tests. Hence, we proposed to use the Anderson-Darling test on $\sum_{i=1}^{N} X_i$ to see if they are lognormal or not.

Let $U_1 \leq \ldots \leq U_n$ be an ordered sample of size n from any distribution. Let

$$S = \sum_{i=1}^{n} \frac{(2i-1)}{n} \left[\ln F(U_i) + \ln\left(1 - F(U_{n+1-i})\right) \right], \quad (10)$$

where F is the cumulative distribution function of the lognormal(μ_W, σ_W). The test of hypothesis is set up as

H_0: $\sum_{i=1}^{N} X_i$ is lognormal(μ_W, σ_W) versus

H_1: $\sum_{i=1}^{N} X_i$ is not lognormal(μ_W, σ_W),

where the sample size of $\sum_{i=1}^{N} X_i$ is n. The test statistic is given by

$$A^2 = -n - S. \quad (11)$$

The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. Tabulated values and formulas have been published by Stephens (1979, 1977, 1977a, 1976, & 1974) for a few specific distributions such as normal, lognormal, exponential, Weibull, logistic, extreme value type 1 and others. The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A^2, is greater than the critical value. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. The null hypothesis, H_0 will be rejected when the p-value of the statistic, A^2 is less than 0.05. In our paper, we will test whether $\sum_{i=1}^{N} X_i$, generated empirically using Monte Carlo simulation, is lognormal with two different sets of parameters. The first set of parameters were obtained via Equations (8) and (9) of the Wilkinson's approximation using the known parameter values of $X_i \sim$ lognormal(μ_i, σ_i).

The second set of parameters were calculated from the Monte Carlo simulations also using Equations (8) and (9). In this case, M (where M is a very large positive integer) data sets of $W = \sum_{i=1}^{N} X_i$ are generated. Hence there will be M $E[W]$s and M $E[W^2]$s. Subsequently, solving for the Monte Carlo lognormal parameters of W, μ_W and σ_W resulting in M μ_Ws and M σ_W. Thus $\mu_{WMC} = \sum_{j=1}^{M} \mu_{W_j} \Big/ M$ and $\sigma_{WMC} = \sum_{j=1}^{M} \sigma_{W_j} \Big/ M$.

ANDERSON-DARLING TESTS WITH MAXIMUM LIKELIHOOD ESTIMATES OF PARAMETERS

The statistical analysis of this study is done using SAS 9.4 (SAS Institute Inc 2015) which conducted the following test of hypothesis.

H_0: $\sum_{i=1}^{N} X_i$ is lognormal versus

H_1: $\sum_{i=1}^{N} X_i$ is not lognormal.

The μ and σ parameters of a lognormal distribution are estimated first from available data using maximum likelihood (Cohen 1951) by assuming that the data is lognormal. Recalling Equation (1) where $W = \sum_{i=1}^{N} X_i$ Assume $W_1,\ldots,W_n$ be a random sample from a lognormal(μ,σ) distribution with mean, θ and variance, η^2. As stated earlier, the transformation $V_i = \ln W_i$. will result in $V_i \sim N(\mu, \sigma^2)$. From Cohen (1951), the relationships between the mean, θ and variance, η^2 of a lognormal distribution and its parameters m and s are given by

$$\theta = \exp\left(\mu + \frac{1}{2}\sigma^2\right) \quad (12)$$

$$\eta^2 = \exp\left(2\mu + \sigma^2\right)\left\{\exp\left(\sigma^2\right) - 1\right\} \quad (13)$$

Let

$$\bar{V} = \frac{1}{n}\sum_{i=1}^{n} V_i \quad (14)$$

and

$$S_V^2 = \sum_{i=1}^{n} \left(V_i - \bar{V}\right)^2. \quad (15)$$

The maximum likelihood estimators (MLE) of θ and η^2 in terms of $\bar{V}$ and S_V^2 are

$$\hat{\theta} = \exp\left(\bar{V} + \frac{1}{2n} S_V^2\right) \quad (16)$$

and

$$\hat{\eta}^2 = \exp\left(2\bar{V} + \frac{1}{n} S_V^2\right)\left\{\exp\left(\frac{1}{n} S_V^2\right) - 1\right\} \quad (17)$$

Equating (12) to (16) and (13) to (17), the MLE of μ and σ are given by

$$\hat{\mu}_{Ws} = \bar{V} = \frac{1}{n}\sum_{i=1}^{n} \ln W_i \quad (18)$$

and

$$\hat{\sigma}_{Ws} = \sqrt{\frac{1}{n} S_V^2} = \sqrt{\frac{1}{n}\left[\sum_{i=1}^{n} \left(\ln W_i - \bar{\mu}\right)^2\right]}. \quad (19)$$

Then the actual null hypothesis becomes H_0: $\sum_{i=1}^{N} X_i$ is lognormal($\widehat{\mu}_{Ws}$, $\widehat{\sigma}_{Ws}$).

In our paper, we compared the approximations distribution to the sum of lognormal variates; Wilkinson approximation and Monte Carlo Wilkinson approximation and Monte Carlo Maximum Likelihood estimation of Wilkinson approximation. The differences are on the parameter estimation as stated in equations (8) and (9) for the Wilkinson approximation, Monte Carlo lognormal parameters of W (Wilkinson approximation); $\mu_{WMC} = \sum_{j=1}^{M} \mu_{W_j} \Big/ M$ and $\sigma_{WMC} = \sum_{j=1}^{M} \sigma_{W_j} \Big/ M$ and Monte Carlo maximum likelihood estimation as in equations (18) and (19). The Monte Carlo simulation were generated for M=10,000 times. Here M is a data set that we generated and it is a very large positive integer. Therefore, we have considered the issue of less accuracy in Wilkinson approximation towards the larger values of $\Sigma XiNi$=1 and improves greatly in the parameter estimation of the sum of lognormal variates.

METHODS

SIMULATION CONDITIONS

The lognormal variates X_i were generated empirically using Monte Carlo simulation. The simulation conditions are: 1) The number of lognormal variates in the sum, N = 2, 10. 2) The number of sample points in the variates, n = 5, 15, 25. 3) The variates are independent and identically distributed (IID) with lognormal parameters, $\mu_i = 0$ and $\sigma_i = 4,\ 12$. 4) The variates are independent but not identically distributed (INID) with lognormal parameters assigned as in Table 1.

TABLE 1. Assignment of μ_i and σ_i when the variates are not identically distributed

N =2	N=10
Set 1: $\mu_1 = 0 \quad \mu_2 = 10$ $\sigma_1 = 4 \quad \sigma_2 = 8$	Set 3: $\mu_1 = \cdots = \mu_5 = 0 \quad \mu_6 = \cdots = \mu_{10} = 20$ $\sigma_1 = \cdots = \sigma_5 = 4 \quad \sigma_6 = \cdots = \sigma_{10} = 12$
Set 2: $\mu_1 = 0 \quad \mu_2 = 20$ $\sigma_1 = 4 \quad \sigma_2 = 12$	Set 4: $\mu_1 = \cdots = \mu_3 = 0 \quad \mu_4 = \cdots = \mu_6 = 10 \quad \mu_7 = \cdots = \mu_{10} = 20$ $\sigma_1 = \cdots = \sigma_3 = 4 \quad \sigma_4 = \cdots = \sigma_6 = 8 \quad \sigma_7 = \cdots = \sigma_{10} = 12$

For both IID and the INID cases, the number of distinct sets of simulation conditions are given by the expression; (number of variates × the number of sample sizes × the number of parameter sets × the number of parameter derivations for W). For each case, there will be = 2 × 3 × 2 × 3 = 36 sets. Altogether, there are 72 distinct sets of simulation conditions.

The procedure to collect Type I error rates for three approximations are as follows:

1. For each distinct set, the values of μ_W and σ_W are calculated using the Wilkinson approximation. Let us denote them as μ_{Wt} and σ_{Wt}, respectively.

2. For each distinct set generate $\sum_{i=1}^{N} X_i$, M = 10,000 times. Here M is a simulated data set that we generated and it is a very large positive integer that we have considered to overcome the issue of less accuracy towards the larger values of $\Sigma XiNi$=1.

3. Then, calculate the Monte Carlo values of μ_W and σ_W from 2, $\mu_{WMC} = \left(\sum_{j=1}^{M} \mu_{W_j} \right) / M$ and $\sigma_{WMC} = \left(\sum_{j=1}^{M} \sigma_{W_j} \right) / M$ respectively. Run 2 against the lognormal(μ_{Wt} , σ_{Wt}) distribution using the Anderson-Darling test. There will be 10,000 p-values. Let the p-value be denoted by p and #(p < 0.05) be the number of p-values less than 0.05 from the 10,000 p-values. Then the p-value of the Anderson-Darling test for any particular set of simulation conditions = #(p < 0.05)/M.

4. Repeat 4 with the lognormal(μ_{WMC}, σ_{WMC}).

5. For each data set j in 2, assume that they are lognormal. Calculate the maximum likelihood estimates of μ and σ : $\widehat{\mu}_{Ws_j}$ and $\widehat{\sigma}_{Ws_j}$ as stated in Equations (18) and (19). There will be M = 10,000 parameter estimates of $\widehat{\mu}_{Ws_j}$ and $\widehat{\sigma}_{Ws_j}$ Report $\widehat{\mu}_{Ws} = \left(\sum_{j\ 1}^{M} \widehat{\mu}_{Ws_j} \right) / M$ and $\widehat{\sigma}_{Ws} = \left(\sum_{j=1}^{M} \widehat{\sigma}_{Ws_j} \right) / M$.

6. At the same time run 2 against lognormal($\hat{\mu}_{Ws_j}, \hat{\sigma}_{Ws_j}$) distribution using the Anderson-Darling test. Collect the 10,000 *p*-values of this test. Summarize the *p*-values with #($p < 0.05$)/*M*.

From this procedure, then Type I error is measured using Bradley's liberal criterion with the condition that the probability of Type I error must be in the interval of $0.5\alpha < \hat{\alpha} < 1.5\alpha$ so that the test is considered robust and good to use by meeting all the conditions as stated in Table 1.

Results and Discussion

In order to do comparison between the three-approximation method to the sum of lognormal variates, we need to estimate the parameter values of *W*. The parameter values were estimated according to the various conditioned as stated in Table 1 and the results are shown in Table 2. The parameter values in Table 2 were used in our hypothesis testing to test the H_0 according to the parameter we estimated.

TABLE 2. The parameters of *W* (μ and σ) for three approximation methods

N	$X_1, \ldots, X_N$ parameters	$W \sim$ lognormal (μ_{Wt}, σ_{Wt})	$W \sim$ lognormal (μ_{WMC} , σ_{WMC}) $W \sim$ lognormal ($\hat{\mu}_{Ws}, \hat{\sigma}_{Ws}$)[a]		
			$n = 5$	$n = 15$	$n = 25$
2	IID lgn(0, 4)	(1.04, 3.91)	(4.41, 2.66) (2.37, 3.07)	(4.31, 2.83) (2.36, 3.20)	(4.38, 2.80) (2.36, 3.22)
2	IID lgn(0, 12)	(1.04, 11.97)	(33.65, 3.26) (6.82, 9.33)	(36.43, 3.33) (6.79, 9.73)	(35.98, 3.36) (6.79, 9.80)
2	INID lgn Set 1[b]	(10, 8)	(27.26, 3.34) (10.65, 6.75)	(28.21, 3.33) (10.64, 7.03)	(28.04, 3.32) (10.62, 7.07)
2	INID lgn Set 2[b]	(20, 12)	(53.40, 3.29) (20.35, 10.83)	(55.22, 3.42) (20.35,11.25)	(54.78, 3.45) (20.31, 11.32)
10	IID lgn(0, 4)	(3.45, 3.70)	(6.72, 2.52) (6.50, 2.03)	(6.77, 2.54) (6.50, 2.13)	(6.76, 2.53) (6.50, 2.14)
10	IID lgn(0, 12)	(3.45, 11.90)	(38.32, 3.14) (18.57, 6.52)	(38.76, 3.29) (18.56, 6.83)	(38.60, 3.31) (18.56, 6.87)
10	INID lgn Set 3[b]	(22.41,11.93)	(58.00, 3.19) (34.01, 7.42)	(58.52, 3.32) (34.02, 7.80)	(57.87, 3.34) (34.03, 7.86)
10	INID lgn Set 4[b]	(22.08,11.94)	(57.99, 3.19) (32.70, 7.50)	(57.71, 3.17) (32.69, 7.88)	(57.21, 3.22) (32.70, 7.93)

Note: lgn is referred to lognormal. [a]MC estimation of the maximum likelihood estimates of ($\hat{\mu}_{Ws}, \hat{\sigma}_{Ws}$). [b]See Table 1

The Type I error rates of all 72 tests are presented in Tables 3 and 4. In both tables the Type I error rates are presented in sets of three. The first entries being rates from tests involving parameters obtained from the Wilkinson approximation, μ_{Wt} and σ_{Wt}. The second entries are the rates from tests involving parameters of the Wilkinson approximation obtained via Monte Carlo, μ_{WMC} and σ_{WMC}. While the third entries are rates obtained from tests of lognormality whose parameters are based on the MLE $\widehat{\mu}_{Ws}$ and $\widehat{\sigma}_{Ws}$. Note that the target Type I error rate is 0.05, therefore the Bradley's criterion interval is (0.025, 0.075) used to measure the robustness of the test as suggested by Bradley (1978) to deem whether the Type I error rates obtained are close to 0.05 or not. Any test with its Type I error rate falling outside of this interval means that the $\sum_{i=1}^{N} X_i$ is lognormal with different μ_W and σ_W values than the ones stated in the null hypothesis.

TABLE 3. Type I error rates of the Anderson Darling tests whether $\sum_{i=1}^{N} X_i$ is lognormal when $X_1, \ldots, X_N$ are IID

	$N = 2$		$N = 10$	
	$(\mu_i = 0,\ \sigma_i = 4)$	$(\mu_i = 0,\ \sigma_i = 12)$	$(\mu_i = 0,\ \sigma_i = 4)$	$(\mu_i = 0,\ \sigma_i = 12)$
	0.076[a]	0.144	0.496	0.966
$n = 5$	0.433[b]	1.000	**0.032**	1.000
	0.051[c]	**0.049**	**0.051**	**0.051**
	0.226	0.451	0.998	1.000
$n = 15$	0.738	1.000	**0.056**	1.000
	0.057	**0.056**	0.086	0.077
	0.390	0.722	1.000	1.000
$n = 25$	0.930	1.000	0.089	1.000
	0.058	**0.056**	0.111	0.099

Note: [a]Tested against lognormal(μ_{Wt} , σ_{Wt}). [b]Tested against lognormal(μ_{WMC} , σ_{WMC}). [c]Tested against lognormal. Parameters estimated by maximum likelihood

From Table 3, it shows that for the tests involving parameters obtained via the Wilkinson approximation μ_{Wt} and σ_{Wt} (first entries), none of the Type 1 errors falls in (0.025, 0.075) interval for both N=2 and N=10 and all sample sizes. As for the tests involving parameters obtained via Monte Carlo, μ_{WMC} and σ_{WMC} (second entries), only 2 of Type 1 errors are within the interval for the conditions of N =10 and n=5, 15 with $(\mu_i = 0,\ \sigma_i = 4)$. The best Type I error rates were obtained from tests of lognormality whose parameters are based on the MLE $\widehat{\mu}_{Ws}$ and $\widehat{\sigma}_{Ws}$ (third entries) with N=2 at all sample size. For N=10, only Type 1 error rates with sample size n=5 that fall within the interval.

For the case of INID sum of lognormal variates in Table 4, there are 6 out of 12 Type I errors collected for tests involving parameters obtained via the Wilkinson approximation μ_{Wt} and σ_{Wt} (first entries), are within the (0.025, 0.075) interval; that are the tests involving $N = 2$ lognormal variates only but not for N=10. As for the tests involving parameters obtained via Monte Carlo, μ_{WMC} and σ_{WMC} (second entries), none of the Type I errors close to 0.05 and fall within interval. Therefore, that tests are not robust compared to other tests. The best Type I error rates were obtained from tests of lognormality whose parameters are based on the MLE $\widehat{\mu}_{Ws}$ and $\widehat{\sigma}_{Ws}$ (third entries). Nine out of 12 tests achieved Type I error

TABLE 4. Type I error rates of the Anderson Darling tests whether $\sum_{i=1}^{N} X_i$ is lognormal when $X_1, \ldots, X_N$ are INID

	$N = 2$		$N = 10$	
	Set 1[d]	Set 2	Set 3	Set 4
$n = 5$	**0.034**[a]	**0.043**	0.663	0.548
	1.000[b]	1.000	1.000	1.000
	0.051[c]	**0.046**	**0.053**	**0.050**
$n = 15$	**0.038**	**0.046**	0.999	0.996
	1.000	1.000	1.000	1.000
	0.071	**0.053**	**0.065**	**0.072**
$n = 25$	**0.041**	**0.042**	1.000	1.000
	1.000	1.000	1.000	1.000
	0.092	**0.053**	0.077	0.086

Note: [a]Tested against lognormal(μ_{Wt}, σ_{Wt}). [b]Tested against lognormal (μ_{WMC}, σ_{WMC}). [c]Tested against lognomal. Parameters estimated by maximum likelihood. [d]See Table 1

rates that are close to 0.05 and were observed for n = 5 and n = 15 when N = 2 and N=10 for all Set 1 to Set 4 parameters and only one Type 1 error rates is within the interval which is for N=2 and n =25 for Set 2 parameter only. None of the Type 1 error for N=10 and n=25 falls in the interval.

We also observed that the Type I errors of the Anderson-Darling tests are closed to 0.05 for W being approximately lognormal(μ_{Wt}, σ_{Wt}) when μ_{Wt} is close to $\widehat{\mu}_{Ws}$ and σ_{Wt} is close to $\widehat{\sigma}_{Ws}$ (Table 2). For example, N = 2 INID lognormal variates Sets 1 and 2. In Set 1, (μ_{Wt} = 10, σ_{Wt} = 8) while the maximum likelihood estimates are ($\widehat{\mu}_{Ws}, \widehat{\sigma}_{Ws}$) = {(10.65, 6.75), (10.64, 7.03), (10.62, 7.07)} for n = 5, 15 and 25, respectively. In Set 2, (μ_{Wt} = 20, σ_{Wt} = 12) while the maximum likelihood estimates are ($\widehat{\mu}_{Ws}, \widehat{\sigma}_{Ws}$) = {(20.35, 10.83), (20.35, 11.25), (20.31, 11.32)} for n = 5, 15 and 25, respectively. The same can be said for the two cases for W being approximately lognormal (μ_{WMC}, σ_{WMC}), namely when N = 10 IID lognormal variates with $(\mu_i = 0,\ \sigma_i = 4)$ with sample sizes n = 5 and n = 15 (also see Table 2). In this case (μ_{WMC}, σ_{WMC}) = {(6.72, 2.52), (6.77, 2.54)} compared to ($\widehat{\mu}_{Ws}, \widehat{\sigma}_{Ws}$) = {(6.50, 2.03), (6.50, 2.13)} for n = 5 and 15, respectively.

Type I error rates that are greater than 0.05 indicated that the W are not lognormal at the stated parameter values, be they (μ_{Wt}, σ_{Wt}), (μ_{WMC}, σ_{WMC}) or ($\widehat{\mu}_{Ws}, \widehat{\sigma}_{Ws}$). However, the maximum likelihood estimates ($\widehat{\mu}_{Ws}, \widehat{\sigma}_{Ws}$) tended to produce Type I error rates that are closer to 0.05. Even when they are higher than 0.075, they are still not excessively high. The summarization of the results in Tables 3 and 4 is shown in Table 5. This summary shows the approximation of the sum of lognormal variates that close to lognormal distribution according to their specific conditions such as number of variates, no of sample size and the parameter of mean and standard deviation.

The huge discrepancy between the values of (μ_{Wt}, σ_{Wt}) and (μ_{WMC}, σ_{WMC}) is contrary to what was obtained in Cardieri and Rappaport (2000) where (μ_{Wt}, σ_{Wt}) was contained in a mesh of (μ_{WMC}, σ_{WMC}). The square mesh was created from a large range of μ_{WMC} by a small range in σ_{WMC}. Hence, the Monte Carlo method cannot produce good estimates of the W through the Wilkinson approximation.

TABLE 5. The summary of approximation of the sum of lognormal variates that close to lognormal distribution according to specific conditions

Approximation of $\sum_{i=1}^{N} X_i$ to Lognormal	$X_1, \ldots, X_N$ is IID	$X_1, \ldots, X_N$ is NIID
Wilkinson		No. of variate, N=2 No. of sample, n=5,15,25 Parameter: Set 1[a] & Set 2[a]
Monte Carlo Wilkinson	No. of variate, N=10 No. of sample, n=5,15 Parameter: $(\mu_i = 0,\ \sigma_i = 4)$ No. of variate, N=2 & 10 No. of sample, n=5 Parameter: $(\mu_i = 0,\ \sigma_i = 4)$ & $(\mu_i = 0,\ \sigma_i = 12)$	
Maximum Likelihood		No. of variate, N=2 & 10 No. of sample, n=5 & 15 Parameter: Set 1[a], Set 2[a], Set 3[a] & Set 4[a]
	No. of variate, N=2 No. of sample, n=15 & 25 Parameter: $(\mu_i = 0,\ \sigma_i = 4)$ & $(\mu_i = 0,\ \sigma_i = 12)$	No. of variate, N=2 No. of sample, n=25 Parameter: Set 2[a]

Note: [a] See Table 1

The Wilkinson approximation of the lognormal distribution of W which was shown to provide reasonable closeness with the actual or close to actual distribution of $\sum_{i=1}^{N} X_i$ in Beaulieu and Xie (2004) and Santos Filho, Cardieri and Yacoub (2005) was shown to exhibit this phenomenon only for the sum of two lognormal variates that are not identical. It failed for the other cases, i.e., the sum of two and ten identical lognormal variates and the sum of ten non-identical lognormal variates.

CONCLUSIONS

Based upon the results, the best method of approximating the lognormal distribution to the sum of lognormal variates is to assume that the sum is lognormal, estimate the parameter values using maximum likelihood estimation and then run the Anderson-Darling test of lognormality with the estimated values in the null hypothesis. The Wilkinson approximation in this particular study only works for the sum of two non-identical lognormal variates. The Monte Carlo version of the Wilkinson approximation did not work at all, producing parameter values of W higher than the usual version of the approximation or the estimated parameters obtained through maximum likelihood. This resulted in complete rejection of the null hypothesis in all 10,000 trials leading to a Type I error rate of 1.000.

ACKNOWLEDGEMENTS

This study was supported by the Research University Grant (GGP-2020-023).

REFERENCES

Abdul Majid, M.H. & Ibrahim, K. 2021. Composite pareto distributions for modelling household income distribution in Malaysia. *Sains Malaysiana* 50(7): 2047-2058.

Beaulieu, N.C. & Xie, Q. 2004. An optimal lognormal approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology* 53: 479-489.

Becker, D.N. 1991. Statistical tests of the lognormal distribution as a basis for interest rate changes. *Transactions of the Society of Actuaries* 43: 7-72.

Bradley, J.V. 1978. Robustness? *British Journal of Mathematical and Statistical Psychology* 31: 144-152.

Bromideh, A.A. 2012. Discriminating between Weibull and Log-Normal distributions based on Kullback-Leibler divergence. *Istanbul University Econometrics and Statistics e-Journal* 16(1): 45-54.

Cardieri, P. & Rappaport, T.S. 2000. Statistics of the sum of lognormal variables in wireless communications. In *Spring 2000 Vehicular. Technology Conference: IEEE 51st Vehicular Technology Conference Proceedings* May 15-18, Tokyo, Japan. pp. 1823-1827.

Cobb, B.R., Rumí, R. & Salmerón, A. 2012. Approximating the distribution of a sum of log-normal random variables. In *The Proceedings of the Sixth European Workshop on Probabilistic Graphical Models.* pp. 67-74.

Cohen, A.C. 1951. Estimating parameters of logarithmic-normal distributions by maximum likelihood. *Journal of the American Statistical Association* 46: 206-212.

Di Renzo, M., Imbriglio, L., Graziosi, F. & Santucci, F. 2009. Distributed data fusion over correlated log-normal sensing and reporting channels: Application to cognitive radio networks. *IEEE Transactions on Wireless Communications* 8: 5813-5821.

Havemann, F., Heinz, M. & Kretschme, H. 2006. Collaboration and distances between German immunological institutes - A trend analysis. *Journal of Biomedical Discovery and Collaboration* 1: 6.

Keselman, H.J., Othman, A.R. & Wilcox, R. 2014. Preliminary testing for normality in the multi-group problem: Is this a good practice? *Clinics in Dermatology* 2: 29-43.

Keselman, H.J., Othman, A.R. & Wilcox, R. 2013. Preliminary testing for normality: Is this good practice? *Journal of Modern Applied Statistical Methods* 2: 2-19.

Limpert, E., Stahel, W.A. & Abbt, M. 2001. Log-normal distribution across the sciences: Keys and clue. *Bioscience* 51(5): 341-352.

Loewenstein, Y., Kuras, A. & Rumpel, S. 2011. Multiplicative dynamics underlie the emergence of the log-normal distribution of spine sizes in the neocortex. *Journal of Neuroscience* 31: 9481-9488.

Muhammad Farouk, Nazrina Aziz & Zakiyah Zain. 2020. The application of lognormal distribution on the new two-sided group chain sampling plan. *Sains Malaysiana* 49(5): 1145-1152.

Osborn, J.F., Cattaruzza, M.S., Ferri, A.M., De Angelis, F., Renzi, D., Marani, A. & Vaira, D. 2013. How long it will take to reduce gastric cancer incidence by eradicating *Heliobacter pylori* infection? *Cancer Prevention Research* 6: 695-700.

Othman, A.R., Keselman, H.J. & Wilcox, R. 2015. Assessing normality: Applications in multi-group designs. *Malaysian Journal of Mathematical Sciences* 9: 53-65.

Saleem, M., Sieskul, B.T. & Kaiser, T. 2006. Channel capacity assessments in UWB communication system over lognormal fading. *The Institution of Engineering and Technology Seminar on Ultra Wideband Systems, Technologies and Applications*. pp. 155-159.

Santos Filho, J.C.S., Yacoub, M.D. & Cardieri, P. 2006. Highly accurate range-Adaptive lognormal approximation to lognormal sum distributions. *Electronics Letters* 42: 361-363.

Santos Filho, J.C.S., Cardieri, P. & Yacoub, M.D. 2005. Simple accurate lognormal approximation to lognormal sums. *Electronics Letters* 41: 1016-1017.

SAS Institute Inc. SAS OnlineDoc 9.4. 2015; Cary, NC.

Schwartz, S.C. & Yeh, Y.S. 1982. On the distribution function and moments of power sums with lognormal components. *Bell Labs Technical Journal* 61: 1441-1462.

Selim, B., Alhussein, O., Muhaidat, S., Karagiannidis, G.K. & Liang, J. 2016. Modeling and analysis of wireless channels via the mixture of Gaussian distribution. *IEEE Transactions on Vehicular Technology* 65: 8309-8321.

Shafiq, M., Alamgir & Atif, M. 2016. On the estimation of three parameters lognormal distribution based on fuzzy life time data. *Sains Malaysiana* 45(11): 1773-1777.

Stephens, M.A. 1979. Tests of fit for the logistic distribution based on the empirical distribution function. *Biometrika* 66: 591-595.

Stephens, M.A. 1977. Goodness of fit for the extreme value distribution. *Biometrika* 64: 583-588.

Stephens, M.A. 1977a. *Goodness of Fit with Special Reference to Tests for Exponentiality*. Technical Report No. 262, Department of Statistics, Stanford University, Stanford, CA.

Stephens, M.A. 1976. Asymptotic results for goodness-of-fit statistics with unknown parameters. *Annals of Statistics* 4: 357-369.

Stephens, M.A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69: 730-737.

Wagner, P.J. 2011. Modelling rate distributions using character compatibility: Implications for morphological evolution among fossil invertebrates. *Biology Letters* 8: 143-146.

Wawrik, B., Kutliev, D., Abdivasievna, U.A., Kukor, J.J., Zylstra, G.J. & Kerkhof, L. 2007. Biogeography of actinomycete communities and Type II polyketide synthase genes in soils collected in New Jersey and Central Asia. *Applied and Environmental Microbiology* 73: 2982-2989.

*Corresponding author; email: noramuda@ukm.edu.my