METHODOLOGY
European Journal of Research Methods for the Behavioral and Social Sciences

PsychOpen GOLD

Original Article

Check for updates

The A Priori Procedure for Estimating the Mean in Both Log-Normal and Gamma Populations and Robustness for Assumption Violations

Lixia Cao [1], Tingting Tong [2], David Trafimow [3], Tonghui Wang [4], Xiangfei Chen [5]

[1] *School of Sciences, Xian Technological University, China.* **[2]** *Department of Mathematical Sciences, New Mexico State University, USA.* **[3]** *Department of Psychology, New Mexico State University, USA.* **[4]** *School of Sciences, Xian Technological University, China; Department of Mathematical Sciences, New Mexico State University, USA.* **[5]** *Department of Mathematical Sciences, New Mexico State University, USA.*

Methodology, 2022, Vol. 18(1), 24–43, https://doi.org/10.5964/meth.7321

Received: 2021-08-10 • **Accepted:** 2022-02-21 • **Published (VoR):** 2022-03-31

Corresponding Author: David Trafimow, Department of Psychology, MSC 3452, New Mexico State University, P. O. Box 30001, Las Cruces, NM 88003-8001, USA. E-mail: dtrafimo@nmsu.edu

Abstract

Although the literature on the a priori procedure, designed to help researchers determine the sample sizes they should use in their substantive research, is expanding rapidly, there are two important limitations. First, there is a need to expand to new popular distributions, log-normal and gamma distributions, and the present work provides those expansions. Second, there is a need to test the consequences of wrong distributional assumptions; for example, assuming a log-normal distribution when the population follows a gamma distribution, or the reverse. The present work addresses the limitations with respect to estimating population means, and it includes computer simulations, links to free and user-friendly programs that researchers can utilize for their own research, and two examples involving real data sets for illustrations of our main results.

Keywords

a priori procedure, log-normal distribution, skew normal, gamma distribution, robustness, minimum sample size

Although the a priori procedure (APP) can be used post-data, it was designed to be used pre-data to determine the sample sizes researchers should collect to simultaneously consider two issues: precision and confidence. The precision issue concerns how close sample statistics are to their corresponding population parameters and the confidence issue concerns the probability of meeting the precision criterion. For example, a researcher

might be interested in having 95% confidence of obtaining a sample mean difference that is within one-tenth of a standard deviation of the population mean difference, and an APP equation could provide that answer. APP equations have been devised for a number of purposes such as estimating a single population mean under normality or under skew normality (Trafimow, Wang, & Wang, 2019; Trafimow, Wang, & Wang, 2020), estimating population mean differences for matched or independent samples under normality or under skew normality (Trafimow et al., 2020; Wang, Wang, Trafimow, & Myüz, 2019b; Wang, Wang, Trafimow, & Chen, 2019), estimating population scale values (Wang, Wang, Trafimow, & Myüz, 2019b), estimating population shape values (Wang, Wang, Trafimow, & Myüz, 2019a), and correlation coefficients (Wang et al., 2021).

But not surprisingly, for a new literature, there are limitations and one of them is the need for expansion to additional distribution families. The present work addresses that limitation with respect to estimating population means under log-normal and gamma distributions. Both distributions are continuous probability distributions that are widely used in different fields of science to model continuous variables that are always positive and have skewed distributions. They can be used to fit the data collected in (i) human behaviors such as the length of comments posted in Internet discussion forums; (ii) biology and medicine such as measures of size of living tissue and blood pressure; (iii) social sciences and demographics such as the household income; (iv) reliability analysis, wireless communications, and computer networks and internet traffic analysis, etc.

An additional limitation is that APP calculations require the researcher to make a distributional assumption, but there has been no APP work exploring the consequences when the distributional assumption is wrong. The present work will be the first of such explorations. In brief, suppose that a population follows a log-normal distribution, and the researcher assumes a gamma distribution; or suppose the population follows a gamma distribution and the researcher assumes a log-normal distribution; either way, what are the consequences for being wrong? One possibility is that the sample size determinations are similar for both log-normal and gamma distributions. In that case, the consequence of choosing the wrong distribution could be considered minor because there is little loss in choosing the wrong distribution. In contrast, if sample size determinations are dissimilar for the two distributions, then using the wrong distribution could entail major consequences such as having much less precision than the calculation implies. A caveat is that because of the newness of the APP, procedures have not yet been developed for distinguishing when a discrepancy is minor or major in a APP context.

In summary, the present work was designed to make three main contributions. First, our goal was to expand the APP to two new distributions: log-normal and gamma. Second, we desired to provide the first exploration of the consequences of wrong distributional assumptions, in an APP context, based on the APP expansion to log-normal and gamma distributions. Third, the simulation results show that the coverage rates matched

our specified precision and confidence level very well in both log-normal and gamma distributions.

Properties of Log-Normal and Gamma Distributions

For deriving the APP for estimating the population mean, we need the following results about the log-normal and gamma distribution.

Definition 1: *A positive random variable X is said to be log-normally distributed with parameters μ and σ^2, denoted by $X \sim LN(\mu, \sigma^2)$, if the logarithm of X is normally distributed with mean μ and variance σ^2, $\log X \sim N(\mu, \sigma^2)$. The probability density function (pdf) of X is given by*

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}x\sigma} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\} & if\ x > 0 \\ 0 & if\ x \leq 0 \end{cases} \quad (1)$$

where $\mu \in \Re$ and $\sigma > 0$.

Note that it is easy to show that the *k*th moment of *X* exists and is given by

$$E(X^k) = \exp\left(k\mu + \frac{k^2\sigma^2}{2}\right), \quad k = 1, 2, \ldots$$

so that the mean and the variance of *X* are

$$v_x = E(X) = \exp(\mu + \sigma^2/2) \quad and \quad \sigma_x^2 = Var(X) = (e^{\sigma^2} - 1)\exp(2\mu + \sigma^2)$$

respectively.

Numerical convolution of log-normal distributions has shown that the sum of such distributions is a distribution which follows the log-normal law with a fair approximation (but not exactly). Therefore, it can be assumed that the sum of two (or more) log-normal distributions is, in a first approximation, another log-normal distribution. The problem is to find this distribution without the tedious work of numerical convolution. The basic idea is to find a log-normal distribution which has the same moments as the exact sum-distribution. Fenton (1960) and Schwartz and Yeh (1982) estimate the pdf for a sum of log-normal random variables using another log-normal pdf with the same mean and variance. The Fenton approximation, referred to as the Fenton-Wilkinson (FW) method, is simple to apply, and for a wide range of log-normal parameters has been shown to be reasonably accurate in comparison to the Schwartz-Yeh (SY) method. We

introduce the FW method in the following Lemma, which will be used to derive our main results in next section.

Lemma 1: *Consider the sum of n independent and identically distributed (i.i.d.) log-normal random variables* $X_1, \ldots, X_n$, *where each* $X_i \sim LN(\mu, \sigma^2)$. *The Fenton-Wilkinson (FW) approximation of the sum* $T = \sum_{i=1}^{n} X_i$ *is a log-normal distribution with parameters* μ_n *and* σ_n^2, *where*

$$\mu_n = \log n + \mu + \frac{1}{2}(\sigma^2 - \sigma_n^2) \qquad \text{and} \qquad \sigma_n^2 = \log\left(\frac{e^{\sigma^2} - 1}{n} + 1\right) \tag{2}$$

Proposition 1: *Let* $X, X_1, \ldots, X_n$ *be a random sample from the log-normal distribution* $LN(\mu, \sigma^2)$. *Then*

i. *for a positive constant a,* $cX \sim LN(\mu + \log a, \sigma^2)$.
ii. *The sample mean* $\bar{X}$ *is approximately log-normal distributed with parameters* $\mu_n - \log n$ *and* σ_n^2 *given in (2).*

Density curves of $\bar{X}$ for $\mu = 0$, $\sigma = 0.7$, with different sample sizes $n = 30$, 50, and 100 are given in Figure 1, while the density curves of $\bar{X}$ for $\mu = 0.5$, $n = 50$, with different $\sigma = 0.5$, 0.75, and 1 are given in Figure 2. From Figure 1, we can see that the density shapes are toward symmetric as the sample size n increases, but they change a lot as the parameter σ increases.

Figure 1

Density Curves of $\bar{X}$ *for Values of n = 30, 50, 100,* $\mu = 0$ *and* $\sigma = 0.75$

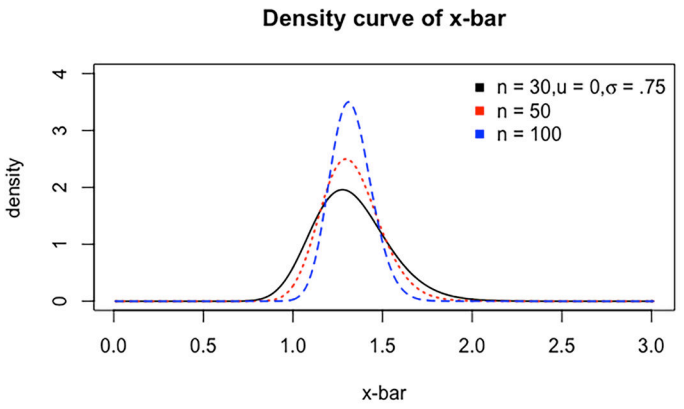
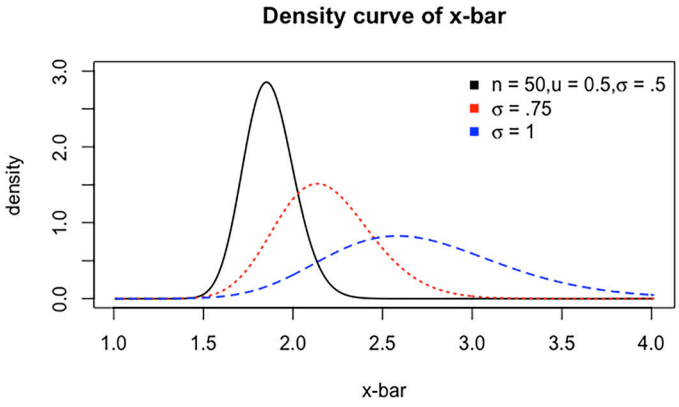


Figure 2

Density Curves of $\bar{X}$ for Values of σ = 0.5, 0.75, 1, μ = 0.5 and n = 50



In order to compare the log-normal distribution and the gamma distribution, we need the following definition and properties.

Definition 2: *A random variable Y is said to have a gamma distributed with the shape parameter k and the scale parameter θ, denoted by Y ~ Gamma(k, θ) if its pdf is given by log-normal pdf*

$$f_Y(y) = \begin{cases} \frac{1}{\Gamma(k)\theta^k} y^{k-1} \exp\left(-\frac{k}{\theta}\right) & if\ y > 0 \\ 0 & if\ y \le 0 \end{cases} \tag{3}$$

It is easy to show the following properties of gamma distribution.

Lemma 2: *Let Y, Y_1, ..., Y_n be a random sample from the gamma distribution Gamma(k, θ). Then*

i. *the mean and the variance of Y are*

$$\nu_y = E(Y) = k\theta \qquad and \qquad \sigma_y^2 = Var(Y) = k\theta^2$$

respectively,

ii. *the sampling distribution of the sample mean $\bar{Y}$ is also gamma distributed with the shape parameter nk and the scale parameter θ/n: $\bar{Y}$ ~ Gamma(nk, θ/n), and*

iii. *the mean and the variance of $\bar{Y}$ are given by $E(\bar{Y}) = k\theta$, $Var(\bar{Y}) = k\theta^2/n$.*

The APP for Estimating the Population Mean

In this section, we will set up the APP procedures for (i) estimating the population mean by a random sample from the log-normal distribution with parameters μ and σ^2, and (ii) estimating population mean by a random sample from the gamma distribution with the shape parameter k and the scale θ.

The Necessary Sample Size for Estimating ν_x With Known σ^2 Under the Log-Normal Setting

In order to determine the necessary sample size n to be $c \times 100\%$ confident for the given precision, we consider the distribution of the unbiased estimator $\hat{\nu}_x = \bar{X}$ for ν_x given in Definition 1 for known standard deviation σ.

Theorem 1: *Suppose that $X_1, \ldots, X_n$ forms a random sample from the log-normal distribution $LN(\mu, \sigma^2)$. Let c be the confidence level and f be the precision which are specified such that the error associated with estimator $\bar{X}$ is $f\sigma_x$. More specifically, if*

$$P\left(f_1\sigma_x \le \bar{X} - \nu_x \le f_2\sigma_x\right) = c \tag{4}$$

where f_1 and f_2 are the left and right precision which are restricted by $\max\left\{\left|f_1\right|, f_2\right\} \le f$, and σ_x is the population standard deviation. Then the minimum sample size n required can be obtained by

$$\int_{L_x}^{U_x} f(z)dz = c \tag{5}$$

such that $U_x - L_x$ is minimized, where

$$L_x = f_1\left(e^{\sigma^2} - 1\right)^{1/2}, \qquad U_x = f_2\left(e^{\sigma^2} - 1\right)^{1/2}$$

and the approximated pdf f(z) is

$$f(z) = \frac{1}{\sqrt{2\pi}\sigma_n(z+1)} \exp\left\{-\frac{\left[\log(z+1) + \frac{1}{2}\sigma_n^2\right]^2}{2\sigma_n^2}\right\}$$

Remark 1: The required sample size n, f_1 and f_2 such that $\max\left\{\left|f_1\right|, f_2\right\} \le f$ are obtained simultaneously, given specified precision f and the confidence level $c \times 100\%$. The corresponding confidence interval for ν_x based on Theorem 1 is

$$\left[\bar{X} - \sigma_x f_2, \quad \bar{X} - \sigma_x f_1\right]$$

The Necessary Sample Size for Estimating ν_y With Known k Under the Gamma Setting

Similarly, the necessary sample size n can be obtained for a given confidence level $c \times 100\%$ and precision f. Consider the distribution of the unbiased estimator $\widehat{\nu_y} = \bar{Y}$ for ν_y given in Definition 2 for known shape parameter k. The second main result is given below.

Theorem 2: *Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables from the gamma distribution Gamma(k, θ). Let c be the confidence level and f be the precision which are specified such that the error associated with estimator $\bar{Y}$ is $f\sigma_y$. More specifically, if*

$$P\left(f_1\sigma_y \le \bar{Y} - \nu_y \le f_2\sigma_y\right) = c \tag{6}$$

where f_1 and f_2 are the left and right precision which are restricted by $max\{|f_1|, f_2\} \le f$, and σ_y is the population standard deviation. Then the minimum sample size n required can be obtained by

$$\int_{L_y}^{U_y} f(w)dw = c \tag{7}$$

such that $U_y - L_y$ is minimized, where $L_y = \sqrt{n}f_1, U_y = \sqrt{n}f_2$, and the pdf $f(w)$ is

$$f(w) = \frac{(kn)^{kn/2}}{\Gamma(kn)}\left(w + \sqrt{kn}\right)^{kn-1} \exp\left[-\sqrt{kn}(w + \sqrt{kn})\right]$$

Remark 2: The required sample size n, f_1, and f_2 can be obtained simultaneously, given the precision f and the confidence level $c \times 100\%$. The corresponding confidence interval for ν_y based on Theorem 2 is

$$\left[\bar{Y} - \sigma_y f_2, \quad \bar{Y} - \sigma_y f_1\right]$$

The Robustness of APP to Some Sorts of Assumption Violations

Like any inferential statistics, the APP assumes a statistical model. For example, we assume log-normal distributions and gamma distributions in this paper. What if particular assumptions are wrong? For example, what if we get the distribution wrong? An important question is: How robust is the APP to various sorts of assumption violations? If one could show robustness to at least some sorts of assumption violations, that would

be very helpful. Even if one finds some assumption violations to which the APP is not robust, that would still be useful because we would know where it is important to be careful. Based on this idea, we first consider sample size. As a trivial example, suppose we assume a gamma distribution and the truth is that there is a log-normal population. We want to see what difference it makes in the estimated necessary sample size.

When we determine the necessary sample sizes for estimating the population mean, we assume σ^2 is known in log-normal setting and assume k is known in gamma setting. To see the difference in the necessary sample sizes, we would control the relationship between σ^2 and k. A natural idea is to render the first two moments of the sample mean, which is an estimator of the population mean in both models. Thus we get the following equation:

$$e^{\sigma^2} - 1 = 1/k \tag{8}$$

For example if we let $\sigma = .75$ then $k = 1.32$ and the corresponding necessary sample size for specific precision values and confidences by Equations (5) and (7) given in Theorem 1 and 2, respectively. Here we consider $c = 0.95, 0.9$ with different values of f, and the corresponding necessary sample sizes in log-normal setting and gamma setting are given in Table 1 and Table 2, respectively, using the programs linked in the table notes.

Table 1

The Value of Sample Size n in Log-Normal Setting, Left Precision f_1 and Right Precision f_2 Under Different f for the Given c = 0.95, 0.9 and σ = .75

Precision (f)	Confidence level (c)	Sample size[a] (n)	Left precision (f_1)	Right precision (f_2)
$f = 0.1$	0.95	391	−0.0978204	0.0996861
	0.9	275	−0.0998779	0.0989713
$f = 0.15$	0.95	177	−0.1453921	0.1496078
	0.9	124	−0.1493012	0.1473061
$f = 0.2$	0.95	105	−0.1921884	0.1998908
	0.9	69	−0.1967918	0.1934145
$f = 0.25$	0.95	67	−0.2378956	0.2497299
	0.9	48	−0.2474282	0.2422409

[a]To find the sample size needed to have a particular probability that the sample mean will be within a desired distance of the population mean, assuming the population is lognormally distributed $LN(\mu, \sigma^2)$, the lognormal program can be used at https://probdiffgamma.shinyapps.io/lognormal/. To use the lognormal program it is necessary to make three entries. In the first box, type in the value of (σ). In the second box, type in the desired degree of precision (f). In the third box, type in the desired confidence level (c). Then click 'update' to obtain the sample size needed to meet your specifications for precision and confidence.

Table 2

The Value of Sample Size n in Gamma Setting, Left Precision f_1 and Right Precision f_2 Under Different f for the Given c = 0.95, 0.9 and k = 1.32

Precision (f)	Confidence level (c)	Sample size[a] (n)	Left precision (f_1)	Right precision (f_2)
f = 0.1	0.95	392	−0.0986157	0.0998879
	0.9	271	−0.0998625	0.0992504
f = 0.15	0.95	174	−0.1470710	0.1499018
	0.9	124	−0.1500000	0.1486157
f = 0.2	0.95	98	−0.1940000	0.1991398
	0.9	71	−0.1986367	0.1962319
f = 0.25	0.95	66	−0.2408662	0.2487501
	0.9	45	−0.2487469	0.2450630

[a]To find the sample size needed to have a particular probability that the sample mean will be within a desired distance of the population mean, assuming the population is gamma distributed, the gamma program can be used at https://probdiffgamma.shinyapps.io/app-gamma/. To use the gamma program it is necessary to make three entries. In the first box, type in the shape parameter of the population distribution (k). In the second box, type in the desired degree of precision (f). In the third box, type in the desired confidence level (c). Then click 'update' to obtain the sample size needed to meet your specifications for precision and confidence, assuming the shape parameter of the log-arithmetically transformed population that you entered in the first box.

From the results given in Tables 1 and 2, we can see that the sample sizes derived under two different populations with same confidence, precision and paired values of parameter are similar. For example, when f = 0.1, c = 0.95 we get n = 391 in Table 1 and n = 392 in Table 2. That is to say the APP is robust to the population assumption violations because the required sample sizes are very close in both tables.

Simulation Results

In this section, we conduct two simulations. First we process a simulation to see how big a difference we have when we use the same sample size for the estimation of parameters in both models. For the comparison of two models, we use measures of the model, such as log-likelihood, AIC, and BIC values, respectively.

The Akaike information criterion (AIC) is an estimator of prediction error and relative quality of statistical models for a given set of data (see Akaike, 1974 and Aho, Derryberry, & Peterson, 2014). Let m be the number of estimated parameters in the model. Let $\widehat{L}$ be the maximum value of the likelihood function for the model. Then the AIC value of the model is given by

$$\mathrm{AIC} = 2m - 2\log(\widehat{L})$$

The formula for the Bayesian information criterion (BIC) is similar to the formula for AIC, but with a different penalty for the number of parameters, (see Schwarz, 1978). With AIC the penalty is $2m$, whereas with BIC the penalty is $m\log(n)$. In practice, the model with the lowest AIC or BIC is preferred, while the model with the highest log-likelihood value is preferred too.

The simulation results are listed in Tables 3 and 4, respectively, using the sample size n required for precision $f = 0.2$ and confidence level $c = 0.95$.

Table 3

The Mean, Absolute Bias, and the Standard Deviation of the MLEs of Parameters in Both the Log-Normal and Gamma Models Are Listed with Sample Size n = 105 and M = 10000 Simulated Data Sets

	Log-normal (True model)		**Gamma**	
***n* = 105**	$\hat{\mu}$	$\hat{\sigma}$	$\hat{k}$	$\hat{\theta}$
True value	1.0000	0.7500	1.3244	2.7190
Mean	0.9993	0.7450	1.9914	1.8563
|Bias|	0.0007	0.0050	0.6666	0.8627
Std. Dev.	0.0727	0.0514	0.2563	0.0820
Log-L	**−222.7483**		−227.1051	
AIC	**449.4967**		458.2102	
BIC	**454.8046**		463.5181	
p_{AIC}	**0.9168**		0.0832	

Table 4

The Mean, Absolute Bias, and the Standard Deviation of the MLE for Parameters in Both the Log-Normal and Gamma Models Are Given with Sample Size n = 98 and M = 10000 Simulated Data Sets

	Log-normal		**Gamma (True model)**	
***n* = 98**	$\hat{\mu}$	$\hat{\sigma}$	$\hat{k}$	$\hat{\theta}$
True value	1.5183	0.4270	5.0000	1.0000
Mean	1.5057	0.4659	5.1665	0.9813
|Bias|	0.0135	0.0372	0.1665	0.0187
Std. Dev.	0.0470	0.0332	0.7155	0.1507
Log-L	−211.4705		**−209.8472**	
AIC	426.9410		**423.6944**	
BIC	432.1109		**428.8643**	
p_{AIC}	0.1887		**0.8113**	

In Table 3, we first generate the required $n = 105$ sample data points from a log-normal distribution $LN(\mu, \sigma^2)$ with $\mu = 1$ and $\sigma = 0.75$, and use this data sample to fit both the log-normal and gamma models. Then we performed the $M = 10000$ simulated data sets and calculated means, absolute bias, and standard deviations of estimators, together with log-likelihood, AIC, and BIC values. From Table 3, we can see that if we use the gamma model to fit the generated data points, both bias and standard errors of estimates are larger than those in fitted log-normal model. Also the Log-likelihood, AIC and BIC values indicate the support of the log-normal models.

Similarly in Table 4, we generate the required $n = 98$ random data points from the gamma model, $Gamma(k, \theta)$, with shape $k = 5$ and scale $\theta = 1$, Then we calculate the Maximum likelihood estimates of the parameters in both models, together with log-likelihood, AIC and BIC values. From Table 4, we can see that the fitted gamma model is better than fitted log-normal model.

For the effectiveness of comparison between two models we use p_{AIC}, the proportion of the true model selected by using AIC among $M = 10000$ runs of the simulated data. We can see that $p_{AIC} = 0.9168$ in Table 3 indicates the log-normal (true model) is more frequently selected model by AIC than the gamma model. Similarly, in Table 4 $p_{AIC} = 0.8113$ indicates the gamma (true model) is more frequently selected model by AIC than the log-normal model.

For investigating the changes of AIC and BIC values to parameter σ in log-normal model, and to the scale θ in gamma model, the results are listed in Figures 3 and 4, and Figures 5 and 6. Figures 3 and 4 show that the values of AIC and BIC are changing as the value of (σ) of log-normal distribution is being changed with $\mu = 1$. Similarly, Figure 5 and 6 show that the values of AIC and BIC are changing as the value of scale (θ) of gamma distribution is being changed with shape parameter $k = 1$. Here the sample size is 100. We can see that the difference of both AICs and BICs are getting bigger as the σ and θ parameters change from 1 to 5, respectively, in Figures 3 and 4 (log-normal distribution) and Figures 5 and 6 (gamma distribution).

Figure 3

AIC with Mean 1 and Different Values of Standard Variation of Log-Normal Distribution

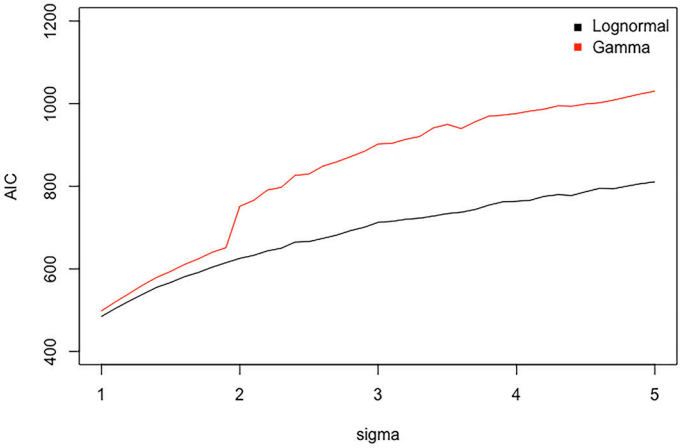


Figure 4

BIC with Mean 1 and Different Values of Standard Variation of Log-Normal Distribution

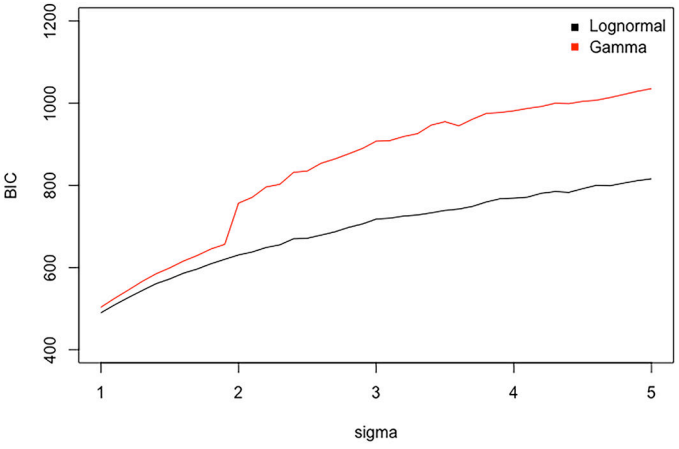


Figure 5

AIC with Mean 1 and Different Values of Standard Variation of Gamma Distribution

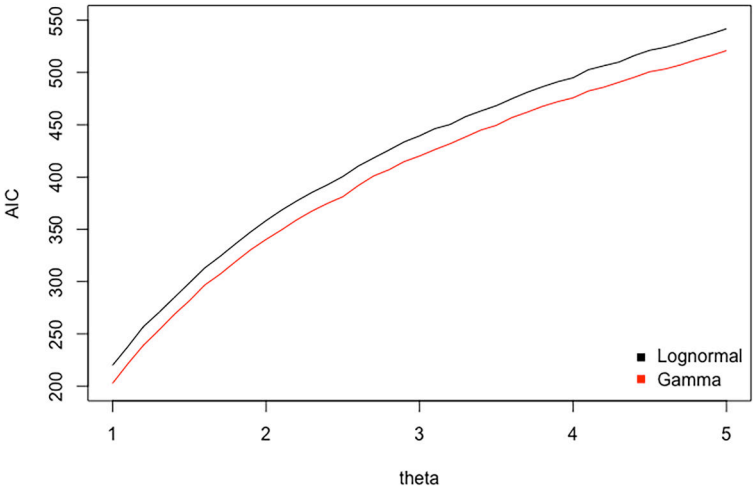
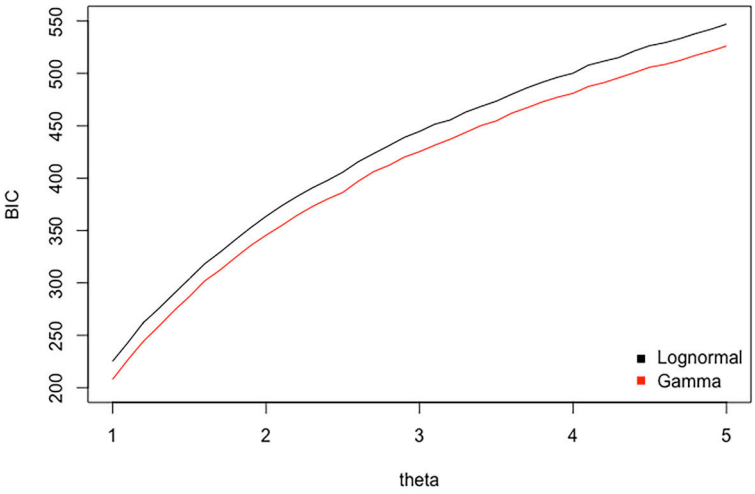


Figure 6

BIC with Mean 1 and Different Values of Standard Variation of Gamma Distribution



In order to compare the coverage rates of confidence intervals with specified precision and confidence level, we process the second simulation for the performance of the confidence intervals obtained by using the derived sample sizes obtained. The coverage

rate (cr) of interval estimating for population mean (v_x) from Theorem 1 with parameters $\mu = 1$, and $\sigma = 0.25, 0.5$, respectively, is listed in Table 5. The coverage rate (cr) of interval estimating for population mean (v_y) from Theorem 2 with parameters $k = 5$, and $\theta = 1, 2$, respectively, is given in Table 6. All results are illustrated with a number of simulations runs $M = 500000$. From both Tables 5 and 6, we can see our APP methods are very effective.

Table 5

Coverage Rate (CR) of Interval Estimating for Population Mean (v_x) from Theorem 3.1 with $\mu = 1$, $\sigma = 0.25, 0.5$ and $M = 500000$

Precision (f)	Confidence level (c)	n ($\mu = 1, \sigma = 0.25$)	cr ($\mu = 1, \sigma = 0.25$)	n ($\mu = 1, \sigma = 0.5$)	cr ($\mu = 1, \sigma = 0.5$)
$f = 0.1$	0.95	388	0.9514	392	0.9524
	0.9	273	0.9012	271	0.9003
$f = 0.15$	0.95	173	0.9519	178	0.9548
	0.9	121	0.9011	124	0.9063
$f = 0.2$	0.95	121	0.9011	102	0.9565
	0.9	68	0.9038	70	0.9072
$f = 0.25$	0.95	63	0.9532	65	0.9564
	0.9	44	0.9036	45	0.9088

Table 6

Coverage Rate (CR) of Interval Estimating for Population Mean (v_y) from Theorem 3.2 with $k = 5$, $\theta = 1, 2$ and $M = 500000$

Precision (f)	Confidence level (c)	n ($k = 5, \theta = 1$)	cr ($k = 5, \theta = 1$)	n ($k = 5, \theta = 2$)	cr ($k = 5, \theta = 2$)
$f = 0.1$	0.95	388	0.9518	388	0.9513
	0.9	273	0.9018	273	0.9016
$f = 0.15$	0.95	173	0.9512	173	0.9519
	0.9	121	0.9003	121	0.9012
$f = 0.2$	0.95	98	0.9518	98	0.9526
	0.9	69	0.9045	69	0.9029
$f = 0.25$	0.95	63	0.9531	63	0.9530
	0.9	44	0.9036	44	0.9029

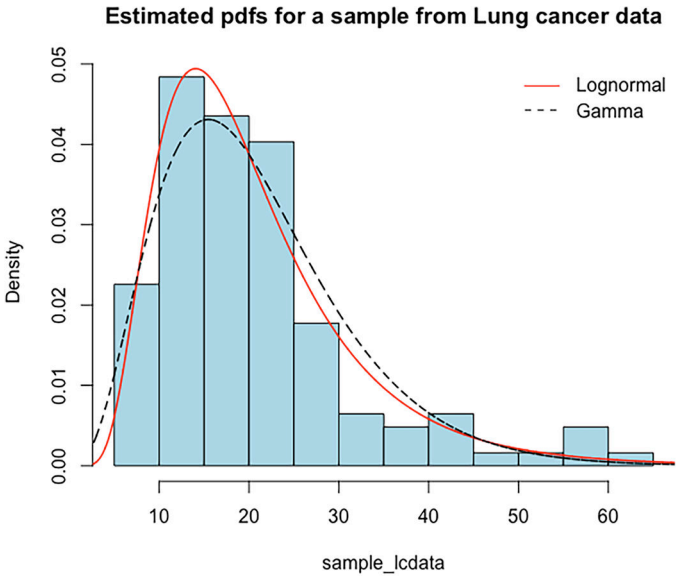
Real Data Examples

In this section we will analyze two real data sets for investigating the performance of our APP methods under log-normal and gamma assumptions.

The first data set is on the survival time (in months) of 184 patients who had limited stage small-cell lung cancer from Overduin (2004). We use both log-normal and gamma distributions to fit this data set, the maximum likelihood estimates for parameters in log-normal model are $\widehat{\mu} = 2.8866$, $\widehat{\sigma} = 0.4835$, and in gamma model are $\widehat{k} = 4.3334$, $\widehat{\theta} = 4.5705$. Using the input of values of $\widehat{\mu}$ and $\widehat{k}$ with $f = 0.15$ and $c = 0.9$ in the links provided in the notes of Table 1 and 2, we obtain the necessary sample sizes for log-normal and gamma are $n = 124$ and 121, respectively. Since the sample sizes are very close we use the larger n. A random sample of simple size $n = 124$ from the lung cancer data set is selected to fit both distributions. The relative histogram, the fitted log-normal and gamma pdfs for the sampled data are plotted in Figure 7. From Table 7, we know that log-normal model fitting is preferable.

Figure 7

The Fitted PDFs and the Relative Histogram for a Sample from the Lung Cancer Data



The second data set is the salary data of faculties in the College of Arts and Sciences, New Mexico State University (2018/19). After the same process as for the first data set, a random sample of the larger size $n = 123$ (this value is determined by the estimates of μ and k with $f = 0.15$ and $c = 0.9$) is selected form the second data set. The relative histogram, the fitted log-normal and gamma pdfs for the sampled data are plotted in Figure 8. From Table 7, we know that gamma model fitting is preferable.

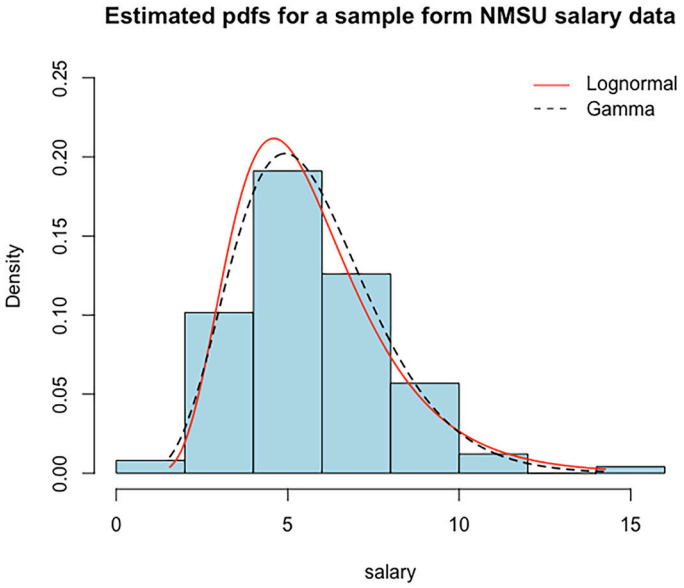
Table 7

Comparison Between the Log-Normal and Gamma

Distribution	Estimator	Log-L	AIC	BIC
		First data set: Lung cancer		
log-normal	$\hat{\mu} = 2.8986$; $\hat{\sigma} = 0.5053$	**−450.7487**	**905.4973**	**911.1379**
gamma	$\hat{k} = 3.9510$; $\hat{\theta} = 5.2405$	−455.3863	914.7727	920.4133
		Second data set: Faculty salary		
log-normal	$\hat{\mu} = 1.6702$; $\hat{\sigma} = 0.3815$	−261.4429	526.8858	532.5102
gamma	$\hat{k} = 7.3854$; $\hat{\theta} = 0.7710$	**−259.7783**	**523.5566**	**529.1810**

Figure 8

The Fitted PDFs and the Relative Histogram for the Salary Data



The values of the log-likelihood, AIC and BIC criteria resulted from fitting log-normal distribution and gamma distribution to the two data sets are presented in Table 7.

Discussion

An important contribution is that the present work, which includes links to free and user-friendly programs, expands the APP so that it can be used under log-normal and gamma distributions. Thus, researchers who wish to know sample sizes needed to have sample statistics that are good estimators of corresponding population parameters, but who worry about not having normally distributed data, need worry no longer. One reason this is important is that most distributions are skewed (Blanca, Arnau, López-Montiel, Bono, & Bendayan, 2013; Ho & Yu, 2015; Micceri, 1989), thereby rendering the family of normal distributions less relevant in estimation contexts. In addition, as there are many ways in which skewness can occur, it is desirable to have the possibility of using many distribution families as potential models, rather than just the skew normal family that has been used earlier (Trafimow et al., 2019). Thus, the present expansion of the APP to the log-normal and gamma families is potentially useful. The potential utility is backed by both computer simulations and worked examples based on real data.

In addition, however, the present work is, to our knowledge, the first APP work that directly addresses the issue of mistakes in identifying the relevant distribution family. To that end, we have explored the consequences of assuming a log-normal distribution in the presence of a gamma distribution, or assuming a gamma distribution in the presence of a log-normal distribution. The results are nuanced. Although the consequences of being wrong are minimal with respect to sample size computations, Figures 3 and 4 show that the difference in AIC and BIC increases as σ increases. Thus, the consequences for being wrong vary depending on the researchers goals. If the goal is sample size determination, the consequences of using the wrong distribution are minimal. In contrast, if the goal is more complex, where AIC or BIC is relevant, the consequences of using the wrong distribution might matter more. An important caveat is that the present work concerns log-normal and gamma distributions. It is not difficult to imagine the possibility of arriving at different conclusions with different distribution families.

In conclusion, the present equations and links to programs successfully expand the APP to log-normal and gamma distributions. And we have seen that the consequences of making a wrong assumption with respect to which family of distributions to use are often, but not always, minimal. We hope and expect that future research will include more APP expansions to distribution families not addressed here.

Funding: This research was partially supported by Shaanxi Province Plan to Improve Public Scientific Literacy of China (NO. 2021PSL122).

Acknowledgments: The authors would like to thank the editor and referees for their useful comments and suggestions, which significantly improved the quality of the present paper.

Competing Interests: The authors have declared that no competing interests exist.

Supplementary Materials

Supplementary materials include three parts. The first one is the R-code for the link of required sample size for Gamma distribution. The second one is the R-code for the link of required sample size for log-normal distribution. The third one is the R-code for simulations and real data analysis. (for access see Index of Supplementary Materials below).

Index of Supplementary Materials

Cao, L., Tong, T., Trafimow, D., Wang, T., & Chen, X. (2022). *Supplementary materials to "The a priori procedure for estimating the mean in both log-normal and gamma populations and robustness for assumption violations"* [Code]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.5655

References

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology, 95*(3), 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., & Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology, 9*(2), 78–84. https://doi.org/10.1027/1614-2241/a000057

Fenton, L. (1960). The sum of log-normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems, 8*(1), 57–67.

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement, 75*(3), 365–388. https://doi.org/10.1177/0013164414548576

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156–166. https://doi.org/10.1037/0033-2909.105.1.156

Overduin, S. (2004). *Use of the lognormal distribution for survival data: Inference and robustness* [Doctoral dissertation, Simon Fraser University].

Schwartz, S. C., & Yeh, Y.-S. (1982). On the distribution function and moments of power sums with log-normal components. *Bell System Technical Journal, 61*(7), 1441–1462.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Trafimow, D., Wang, C., & Wang, T. (2020). Making the a priori procedure work for differences between means. *Educational and Psychological Measurement, 80*(1), 186–198. https://doi.org/10.1177/0013164419847509

Trafimow, D., Wang, T., & Wang, C. (2019). From a sampling precision perspective, skewness is a friend and not an enemy! *Educational and Psychological Measurement, 79*(1), 129–150. https://doi.org/10.1177/0013164418764801

New Mexico State University. (2018/19). *Budget estimate [Salaries].*

Wang, C., Wang, T., Trafimow, D., & Chen, J. (2019). Extending a priori procedure to two independent samples under skew normal settings. *Asian Journal of Economics and Banking, 3*(02), 29–40.

Wang, C., Wang, T., Trafimow, D., Li, H., Hu, L., & Rodriguez, A. (2021). Extending the a priori procedure (APP) to address correlation coefficients. In N. T. Nguyen, V. Kreinovich, & D. T. Nguyen (Eds.), *Data science for financial econometrics* (pp. 141–149). Springer. https://doi.org/10.1007/978-3-030-48853-6_10

Wang, C., Wang, T., Trafimow, D., & Myüz, H. A. (2019a). Desired sample size for estimating the skewness under skew normal settings [Conference Session]. *International Conference of the Thailand Econometrics Society* (pp. 152–162). Springer.

Wang, C., Wang, T., Trafimow, D., & Myüz, H. A. (2019b). Necessary sample sizes for specified closeness and confidence of matched data under the skew normal setting. *Communications in Statistics-Simulation and Computation.* Advance online publication. https://doi.org/10.1080/03610918.2019.1661473

Appendix

Proof of Proposition 1: (i) If $X \sim LN(\mu, \sigma^2)$, $\log(X) \sim N(\mu, \sigma^2)$. Given any positive constant a, we have $\log(aX) = \log a + \log X \sim N(\mu + \log a, \sigma^2)$, thus $aX \sim LN(\mu + \log a, \sigma^2)$. (ii) The result follows directly from Lemma 1 and (i) with $a = 1/n$. □

Proof of Theorem 1: Consider the standardized random variable

$$Z = \left(e^{\sigma^2} - 1\right)^{1/2} \frac{\bar{X} - \nu_x}{\sigma_x} = \frac{\bar{X} - \exp\left(\mu + \frac{\sigma^2}{2}\right)}{\exp\left(\mu + \frac{\sigma^2}{2}\right)} = \frac{\bar{X}}{\exp\left(\mu + \frac{\sigma^2}{2}\right)} - 1$$

Then equation in (4) is equivalent to

$$P\left(\left(e^{\sigma^2} - 1\right)^{1/2} f_1 \le Z \le \left(e^{\sigma^2} - 1\right)^{1/2} f_2\right) = c$$

so that

$$\int_{L_x}^{U_x} f(z)dz = c$$

where $L_x = \left(e^{\sigma^2} - 1\right)^{1/2} f_1$ and $U_x = \left(e^{\sigma^2} - 1\right)^{1/2} f_2$.

Note that by (i) and (ii) in Proposition 1, we obtain

$$\frac{\bar{X}}{\exp\left(\mu + \frac{\sigma^2}{2}\right)} \sim \text{Approximate } LN\left(-\frac{1}{2}\sigma_n^2, \sigma_n^2\right)$$

so that the pdf of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}\sigma_n(z+1)} \exp\left\{-\frac{\left[\log(z+1) + \frac{1}{2}\sigma_n^2\right]^2}{2\sigma_n^2}\right\} \quad \square$$

Proof of Theorem 2: Consider the standardized random variable

$$W = \frac{\bar{Y} - \nu_y}{\sqrt{Var(\bar{Y})}} = \frac{\bar{Y} - k\theta}{\sqrt{k\theta^2/n}} = \frac{\bar{Y}}{\theta\sqrt{k/n}} - \sqrt{kn}$$

Then equation in (6) is equivalent to

$$P\left(\sqrt{n}f_1 \le W \le \sqrt{n}f_2\right) = c$$

that is

$$\int_{L_y}^{U_y} f(w)dw = c$$

with $L_y = \sqrt{n}f_1$, and $U_y = \sqrt{n}f_2$. Note that by Lemma 2, we have

$$\frac{\bar{Y}}{\theta\sqrt{k/n}} \sim Gamma\left(kn, \frac{1}{\sqrt{kn}}\right)$$

so that the pdf of W is given by

$$f(w) = \frac{(kn)^{kn/2}}{\Gamma(kn)}\left(w + \sqrt{kn}\right)^{kn-1} \exp\left[-\sqrt{kn}\left(w + \sqrt{kn}\right)\right] \quad \square$$