Confidence Intervals for the Ratio of Means of Two Independent Log-Normal Distributions

[1]LAPASRADA SINGHASOMBOON, [2]WARARIT PANICHKITKOSOLKUL, [3]ANDREI VOLODIN

[1,2]Department of Mathematics and Statistics, Thammasat University, Pathum Thani, THAILAND
[3]Department of Mathematics and Statistics, University of Regina, Saskatchewan, CANADA

Abstract— In this paper, we investigate confidence intervals for the ratio of means of two independent log-normal distributions. The normal approximation (NA) approach was proposed. We compared the proposed with another approaches, the ML, GCI, and MOVER. The performance of these approaches were evaluated in terms of coverage probabilities and interval widths. The Simulation studies and results showed that the GCI and MOVER approaches performed similar in terms of the coverage probability and interval width for all sample sizes. The ML and NA approaches provided the coverage probability close to nominal level for large sample sizes. However, our proposed method provided the interval width shorter than other methods. Overall, our proposed is conceptually simple method. We recommend that our proposed approach is appropriate for large sample sizes because it is consistently performs well in terms of the coverage probability and the interval width is typically shorter than the other approaches. Finally, the proposed approaches are illustrated using a real-life example.

KeyWords: Confidence intervals, Log-normal, Normal approximation, Simulation

Received: February 8, 2021. Revised: March 5, 2021. Accepted: March 8, 2021. Published: March 17, 2021.

1. Introduction

The log-normal distribution is important in describing positively skewed data. Therefore, the log-normal distribution is used as a model in various real life applications. , for example in medicine where latency periods (the time between the infection and the first symptoms) are log-normal as in [1], in environmental engineering where the probability distribution of contaminant concentrations are often modeled by the log-normal as in [2], in Atmospheric Science, many atmospheric physical and chemical properties are modeled by the log-normal as in [3], and in economics where it can be used to model markets, for example, incomes asin [4] and closing prices on stocks as in [5].

Let the random variable X follow a log-normal distribution with parameter μ and σ^2. Then the random variable $Y = \ln(X)$ follows the normal distribution $N(\mu, \sigma^2)$. The probability density function (pdf) of X is

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}; x > 0, \mu \in \mathbb{R}, \sigma > 0.$$

The mean and variance of the log-normal distribution are

$$\tau = \exp\left(\mu + \frac{\sigma^2}{2}\right) \; and \; \tau_2 = (\exp(\sigma^2) - 1)\exp(2\mu + \sigma^2),$$

respectively.

The inference of mean of the log-normal distribution are frequently interested. Many studies have examined to estimate the confidence interval of the mean as in [6]-[8].

When the interested data follow a skewed distribution, a log-transformation can be used to normalize the distribution of the original. The two well-known methods, the t-test and the Wilcoxon test have been used to study for comparing the means of two independent log-normal distributions in [9]-[12]. However, the Wilcoxon test and the t-test had type I error rates that were very different from the nominal levels when the two population variances were not equal. Then, Reference [13] proposed two new methods: one is a traditional maximum likelihood test which is based on the parametric procedure and the other is the bootstrap test which is based on the nonparametric procedure to overcome this problem. Their purpose is deriving the news methods to testing the difference of the means of two independent log-normal distributions and compare with the t-test and the Wilcoxon test in terms of type I error and power of the test. Their simulation study showed that the traditional maximum likelihood test was the best in terms of the type I error rate and power of the test when the variances of the log-transformed data were unequal and the sample size were large.

Reference [14] shows a new approach was proposed for constructing hypothesis tests using the concept of generalized p-values and this idea was later extended to a method of constructing generalized confidence intervals using generalized pivotal quantity as in [15]. The concepts of the generalized p-value and generalized pivotal quantity have been applied to the problem where standard solutions do not exist for hypothesis testing and confidence intervals. Constructing the test of the parameter of interest is based on the conditions of the generalized p-value. Similarly, when interesting in the problem of the confidence intervals for the parameter of interest is based on the condition's generalized pivotal quantity. Reference [16] proposed the generalized p-value and generalized pivotal quantity concepts to propose the hypotheses tests and generalized confidence intervals (GCI) for a log-normal mean and the ratio of means of two log-normal distributions. In a study comparing the ratio of means of two log-normal distributions, they compared the performance of the proposed with the traditional maximum likelihood method in terms of type I error and power of the test and showed via simulation studies that the test based on the generalized p-value is satisfactory in terms of type I error, it is suitable for all sample sizes, whereas the traditional maximum likelihood method is too conservative or too liberal even for large sample sizes.

The Method of Variance Estimates Recovery (MOVER) was introduced as in [17]. The concept of this method requires only confidence limits for a single parameter to derive the closed-form confidence interval for the function of the parameter. Reference [18] derived the confidence interval for a single log-normal mean and a difference between two log-normal means based on the MOVER method. The single log-normal mean and a difference between two log-normal means can be assumed as the function of parameter, respectively. They showed this closed-form procedure, requiring only confidence limits for a normal mean and variance. In simulation studies, they compared the MOVER with the GCI. The results exhibited the MOVER performs at least as well as the GCI method in terms of the interval width for all sample sizes.

The purpose of this paper is to present a simple approach, namely a Normal approximation (NA) to confidence interval estimation for ratio of means of two independent log-normal distributions. The main statistical tool that we are using is the famous Delta method. It can be explained briefly in the following way; for details we refer interested reader to any advanced textbook on Mathematical Statistics, for example as in [19]. And also to determine the coverage probabilities and interval widths of the confidence interval by using the NA compare with the above three main confidence interval estimation approaches (ML, GCI and MOVER) to see the performance of them in simulation studies and the approaches are also applied to a real-life example.

This paper is organized as follows. Section II, we first review the existing approaches for the problem of constructing confidence intervals for ratio of two log-normal means, and then we proposed a Normal approximation (NA) approach for this problem. In Section III, simulation studies and rerults are conducted to study the performance of the NA, and the results are compared with those of other approaches in terms of coverage probabilities and interval widths, a real-life example is analyzed for illustration purposes is presented in section IV. In section V, the discussion is presented. Finally, section VI is provided conclusion.

2. Preliminaries

Suppose that $X_{i1}, X_{i2}, \dots, X_{in_i}$ be two independent random samples from log-normal population with parameters μ_i and $\sigma_i^2, i = 1,2$.

The mean of the i-th log-normal distribution is

$$M_i = \exp\left(\mu_i + \frac{\sigma_i^2}{2}\right).$$

The problem of interest is to constructing confidence intervals for ratio of means of the two log-normal populations; that is

$$\theta = \frac{M_1}{M_2} = \exp(\eta_1 - \eta_2),$$

where $\eta_i = \mu_i + \frac{\sigma_i^2}{2}, i = 1,2$.

The random variables $Y_{ij} = \ln(X_{ij}), i = 1,2, j = 1, \dots, n_i$ follow the normal distribution $N(\mu_i, \sigma_i^2)$. Based on this fact, the following unbiased maximum likelihood estimators (MLEs) for μ_i and σ_i^2, respectively, are well known

$$\hat{\mu} = \overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad and \quad \hat{\sigma}_i^2 = S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2$$

In this Section, we first review the three of existing approaches for the problem confidence interval estimation of the ratio of means of two log-normal populations. Then, we present the new approach to this problem using the Normal approximation (NA).

2.1. Traditional maximum likelihood (ML)

Zhou, Gao, and Hui [13] proposed a traditional maximum likelihood (ML) approach for testing the difference log-normal means. Thus, this method can be extended to construct the confidence interval for the ratio of two log-normal means for this study.

Firstly, they derived the point estimator for testing the difference log-normal means. The maximum likelihood method was used. Then, the maximum likelihood estimators for $\eta_1 - \eta_2$ is given by

$$\hat{\eta}_1 - \hat{\eta}_2 = \left(\overline{Y}_1 + \frac{S_1^2}{2}\right) - \left(\overline{Y}_2 + \frac{S_2^2}{2}\right).$$

Note that $\overline{Y}_i$ and S_i^2 are independent, $\overline{Y}_i$ is distributed as $N\left(\mu_i, \frac{\sigma_i^2}{n_i}\right)$ and $\frac{(n_i-1)S_i^2}{\sigma_i^2}$ is distributed as χ^2 with $n_i - 1$ degrees of freedom.

Secondly, they obtained the variance of the difference log-normal means $\eta_1 - \eta_2$.

$$Var(\hat{\eta}_1 - \hat{\eta}_2) = Var\left[\left(\overline{Y}_1 + \frac{S_1^2}{2}\right) - \left(\overline{Y}_2 + \frac{S_2^2}{2}\right)\right] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_1^4}{2(n_1-1)} + \frac{\sigma_2^2}{n_2} + \frac{\sigma_2^4}{2(n_2-1)}.$$

After estimating the variance for $Var(\hat{\eta}_1 - \hat{\eta}_2)$ by substituting estimates S_1^2 and S_2^2 for σ_1^2 and σ_2^2, respectively. They applied the central limit theorem and the asymptotic property of the maximum likelihood estimator to build the confidence interval for the difference means, $\eta_1 - \eta_2$.

Then, the confidence interval (CI) for the difference means, $\eta_1 - \eta_2$ takes the form

$$CI = \left(\overline{Y}_1 + \frac{S_1^2}{2}\right) - \left(\overline{Y}_2 + \frac{S_2^2}{2}\right) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_1^4}{2(n_1-1)} + \frac{S_2^2}{n_2} + \frac{S_2^4}{2(n_2-1)}}.$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile value of the standard normal distribution.

Finally, taking the exponentiation into the lower and upper limits of the CI for $\eta_1 - \eta_2$.

Therefore, the $100(1-\alpha)\%$ two-sided CI for ratio of two log-normal means θ based on the ML method is given by

$$CI_{ML} = \exp\left[\left(\overline{Y}_1 + \frac{S_1^2}{2}\right) - \left(\overline{Y}_2 + \frac{S_2^2}{2}\right) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^4}{2(n_1-1)} + \frac{S_2^2}{n_2} + \frac{S_2^4}{2(n_2-1)}}\right]$$

2.2. Generalized confidence interval (GCI)

The basic idea of the Generalized confidence interval (GCI) was originally introduced by Weerahandi [15]. Let X be a random sample whose distribution involve the θ, the parameter of interest, and λ, a nuisance parameter, and let x denote the observed value of X.

To find confidence interval for θ, it is first need to find a generalized pivotal quantity $R(X; x, \theta, \lambda)$, which is a function of the random sample X, the observed data x and the unknown parameters θ, λ and should be satisfy the following conditions:

1. The distribution of $R(X; x, \theta, \lambda)$ is free of unknown parameters;
2. The observed value of $R(X; x, \theta, \lambda)$ is equal to the parameter of interest(θ).

Krishnamoorthy and Mathew [16] defined a GCI for $\eta_1 - \eta_2$ as

$$R_{\eta_1-\eta_2} = R_{\eta_1} - R_{\eta_2} \qquad (1)$$

This result is based on the observation that

$$R_{\eta_i} = \overline{y}_i - \frac{\overline{Y}_i - \mu_i}{S_i/\sqrt{n_i}} \frac{s_i}{\sqrt{n_i}} + \frac{1}{2}\frac{\sigma_i^2}{S_i^2} s_i^2$$
$$= \overline{y}_i - \frac{Z_i}{\sqrt{n_i}} \frac{s_i\sqrt{n_i-1}}{U_i} + \frac{1}{2}\frac{s_i^2}{U_i^2/(n_i-1)}, i = 1,2,$$

where Z_i and U_i are independent and $Z_i = \frac{\overline{Y}_i - \mu_i}{\sigma_i/\sqrt{n_i}} \sim N(0,1), U_i^2 = \frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1}$.

We note first of all that the second expression suggests that the distribution of R_{η_i} is free of unknown parameters.

Secondly, the first expression is equal to η, if we substitute $\overline{y}_i$ and s_i^2 for $\overline{Y}_i$ and S_i^2.

Therefore, the generalized confidence interval for θ may be obtained using the follow algorithm:

The algorithm:

1. For a given data set, compute the sample means and sample variances $\overline{y}_1, \overline{y}_2, s_1^2$ and s_2^2 using the log-transform data.
2. For $i = 1$ to m
3. Generate value for Z_1, Z_2, U_1^2, U_2^2 from the standard normal distribution and the chi-squared distribution with $n-1$ degree of freedom, respectively.
4. compute $R_{\eta_1-\eta_2} = R_{\eta_1} - R_{\eta_2}$ as in (1).
5. End loop for i
6. Order the m values for $R_{\eta_1-\eta_2}$ and compute the $100(\alpha/2)$ and $100(1-\alpha/2)$ percentiles of $R_{\eta_1-\eta_2}$, denoted by $R_{\eta_1-\eta_2}(\alpha/2)$ and $R_{\eta_1-\eta_2}(1-\alpha/2)$, respectively.
7. the $(1-\alpha)100\%$ two-sided generalized confidence interval for θ is obtained by take the exponentiated for $R_{\eta_1-\eta_2}(\alpha/2)$ and $R_{\eta_1-\eta_2}(1-\alpha/2)$.

2.3 Method of Variance Estimates Recovery (MOVER)

Zou and Donner [17] offered a Method of Variance Estimates Recovery (MOVER) for a confidence intervalconstruction. This method provides a closed-form CI and easy to compute. The idea of deriving the closed-form interval is based on only estimates confidence limits for a linear combination of parameters from confidence limits for the individual or single parameters based on the recovery of variance estimates.

To describe the MOVER concept, let $\hat{\theta}_i, i = 1,2$ be point estimates and assume that θ_1 and $\theta_2$are independently distributed. In general, the Wald's confidence interval for $\theta_1 + \theta_2$ is given by

$$(L, U) = \left(\hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_2)}, \hat{\theta}_1 + \hat{\theta}_2 + z_{\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta}_1) + \mathrm{Var}(\hat{\theta}_2)}\right),$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile value from the standard normal distribution.

The Wald's CI does not perform well in small sample sizes. Its performance can be improved by obtaining $Var(\hat{\theta}_i)$ for individual parameter $\hat{\theta}_i, i = 1,2$ at the neighborhood of the confidence limits L and U separately.

Let (l_1, u_1) and (l_2, u_2) be confidence intervals for θ_1 and θ_2, respectively, where

$$l_i = \hat{\theta}_i - z_{\alpha/2}\sqrt{Var(\hat{\theta}_i)}, u_i = \hat{\theta}_i + z_{\alpha/2}\sqrt{Var(\hat{\theta}_i)},$$

which yield the estimated variances is

$$\widehat{Var_l}(\hat{\theta}_i) = \frac{(\hat{\theta}_i - l_i)^2}{z_{\alpha/2}^2}, \widehat{Var_u}(\hat{\theta}_i) = \frac{(u_i - \hat{\theta}_i)^2}{z_{\alpha/2}^2}.$$

The confidence limits L have plausible values $l_1 + l_2$ as the minimum value and U or $u_1 + u_2$ as the maximum value for $\theta_1 + \theta_2$, respectively. This implies that constructing of confidence interval for $\theta_1 + \theta_2$, they substitute corresponding variance estimators, $\widehat{Var}(\hat{\theta}_i)$ in the confidence limits L and U, respectively.

Therefore, the two-sided confidence interval (L, U) for $\theta_1 + \theta_2$ given by

$$L = \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{Var_l}(\hat{\theta}_1) + \widehat{Var_l}(\hat{\theta}_2)} = \hat{\theta}_1 + \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (\hat{\theta}_2 - l_2)^2}$$

and

$$U = \hat{\theta}_1 + \hat{\theta}_2 - z_{\alpha/2}\sqrt{\widehat{Var_u}(\hat{\theta}_1) + \widehat{Var_u}(\hat{\theta}_2)} = \hat{\theta}_1 + \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (u_2 - \hat{\theta}_2)^2}.$$

Considering the construction of confidence interval for $\theta_1 - \theta_2$, we can apply this concept by writing $\theta_1 - \theta_2$ as $\theta_1 + (-\theta_2)$, and recognizing that the CI for $-\theta_2$ is $(-u_2, -l_2)$. Then, the two-sided confidence interval (L, U) for $\theta_1 - \theta_2$ is given by

$$(L, U) = \left(\hat{\theta}_1 - \hat{\theta}_2 - \sqrt{(\hat{\theta}_1 - l_1)^2 + (u_2 - \hat{\theta}_2)^2}, \hat{\theta}_1 - \hat{\theta}_2 + \sqrt{(u_1 - \hat{\theta}_1)^2 + (\hat{\theta}_2 - l_2)^2}\right).$$

Zou, Huo, and Taleban [18] applied the MOVER method to construct confidence interval for the single log-normal mean and difference two log-normal means.

Based on (l_1, u_1) and (l_2, u_2), the confidence intervals for $\theta_1 = \mu$ and $\theta_2 = \frac{\sigma^2}{2}$ are

$$(l_1, u_1) = \left(\bar{y} - z_{\alpha/2}\sqrt{s^2/n}, \bar{y} + z_{\alpha/2}\sqrt{s^2/n}\right)$$

and

$$(l_2, u_2) = \left[\frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{2\chi^2_{\alpha/2,n-1}}\right],$$

respectively.

where

$$\hat{\mu} = \bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad and \quad \hat{\sigma}_i^2 = S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2,$$

and $\bar{y}$, s^2 are the observed values of $\bar{Y}_i$, S_i^2, respectively.

These limits can be applied to derive the CI for the single log-normal mean, where assume $\theta_1 = \mu$ and $\theta_2 = \frac{\sigma^2}{2}$ and consider in construction of confidence interval for $\theta_1 + \theta_2$.

So, the two-sided confidence interval (L, U) for the single log-normal mean, $\theta_1 + \theta_2 = \mu + \frac{\sigma^2}{2}$ based on the MOVER is given by

$$L = \bar{y} + \frac{s^2}{2} - \sqrt{\frac{z_{\alpha/2}^2 s^2}{n} + \left(\frac{s^2}{2} - \frac{(n-1)s^2}{2\chi^2_{1-\alpha/2,n-1}}\right)^2} \tag{2}$$

and

$$U = \bar{y} + \frac{s^2}{2} + \sqrt{\frac{z_{\alpha/2}^2 s^2}{n} + \left(\frac{(n-1)s^2}{2\chi^2_{\alpha/2,n-1}} - \frac{s^2}{2}\right)^2}. \tag{3}$$

Similarly, for the difference two log-normal means

$$\eta_1 - \eta_2 = \left(\mu_1 + \frac{\sigma_1^2}{2}\right) - \left(\mu_2 + \frac{\sigma_2^2}{2}\right),$$

the estimator is

$$\hat{\eta}_1 - \hat{\eta}_2 = \left(\bar{Y}_1 + \frac{S_1^2}{2}\right) - \left(\bar{Y}_2 + \frac{S_2^2}{2}\right).$$

They let $\eta_1 = \theta_1$, $\eta_2 = \theta_2$, where $\eta_1 = \mu_1 + \frac{\sigma_1^2}{2}$ and $\eta_2 = \mu_2 + \frac{\sigma_2^2}{2}$ and consider in construction of confidence interval for $\theta_1 - \theta_2$.

Then, the $(1 - \alpha)100\%$ two-sided confidence interval for $\eta_1 - \eta_2$ based on the MOVER is as follows

$$CI = [\hat{\eta}_1 - \hat{\eta}_2 - \sqrt{(\hat{\eta}_1 - L_1)^2 + (U_2 - \hat{\eta}_2)^2}, \hat{\eta}_1 - \hat{\eta}_2 + \sqrt{(U_1 - \hat{\eta}_1)^2 + (\hat{\eta}_2 - L_2)^2}],$$

where

$$L_i = \bar{y}_i + \frac{s_i^2}{2} - \sqrt{\frac{z_{\alpha/2}^2 s_i^2}{n_i} + \left(\frac{s_i^2}{2} - \frac{(n_i - 1)s_i^2}{2\chi^2_{1-\alpha/2,n_i-1}}\right)^2}$$

and

$$U_i = \bar{y}_i + \frac{s_i^2}{2} + \sqrt{\frac{z_{\alpha/2}^2 s_i^2}{n_i} + \left(\frac{(n_i - 1)s_i^2}{2\chi^2_{\alpha/2,n_i-1}} - \frac{s_i^2}{2}\right)^2}$$

for $i = 1,2$.

Finally, taking exponential function into the lower and upper limits of the CI for the difference two log-normal means, $\eta_1 - \eta_2$.

Therefore, the confidence interval for ratio of two log-normal means, θ based on the MOVER is given by

$$CI_{MOVER} = \exp\left[\hat{\eta}_1 - \hat{\eta}_2 - \sqrt{(\hat{\eta}_1 - L_1)^2 + (U_2 - \hat{\eta}_2)^2}, \hat{\eta}_1 - \hat{\eta}_2 + \sqrt{(U_1 - \hat{\eta}_1)^2 + (\hat{\eta}_2 - L_2)^2}\right],$$

where the lower limit is $\exp[\hat{\eta}_1 - \hat{\eta}_2 - \sqrt{(\hat{\eta}_1 - L_1)^2 + (U_2 - \hat{\eta}_2)^2}]$, and the upper limit is $\exp[\hat{\eta}_1 - \hat{\eta}_2 + \sqrt{(U_1 - \hat{\eta}_1)^2 + (\hat{\eta}_2 - L_2)^2}]$.

2.4 Normal approximation method (NA)

The parameter of interest is

$$\theta = \exp\left(\mu_1 + \frac{\sigma_1^2}{2} - \mu_2 - \frac{\sigma_2^2}{2}\right),$$

and the suggested plug-in estimator

$$\hat{\theta} = \exp\left(\bar{Y}_1 + \frac{S_1^2}{2} - \bar{Y}_2 - \frac{S_2^2}{2}\right).$$

Since the ratio of means, θ, is a complex function. The exact mean and exact variance may be difficult to obtain. Then in this study, we would like to propose the CI for θ based on the normal approximation method with the estimated mean and the variance by using the Delta method and also apply the Delta method for proving the asymptotic normality of this estimator.

In the Delta method, function $g(V_1, V_2, \ldots, V_k)$ of k variables is expanded into the Taylor series at the point

$\theta_1 = E(V_1), \theta_2 = E(V_2), ldots, \theta_k = E(V_k)$:

$$g(V_1, V_2, \dots, V_k) = g(\theta_1, \theta_2, \dots, \theta_k) + \sum_{j=1}^{k} \frac{\partial g(\theta_1, \theta_2, \dots, \theta_k)}{\partial v_j}(V_j - \theta_j) + Remainder.$$

It is possible to prove that $\sqrt{n} Remainder \to 0$ in probability as the sample sizes $n_1, n_2 \to \infty$. For details we refer to the famous monograph in [19].

In our case we have two samples $X_{11}, X_{12}, \dots, X_{1n_1}$ and $X_{21}, X_{32}, \dots, X_{2n_2}$ from independent log-normal populations with parameters μ_1, σ_1^2 and μ_2, σ_2^2, respectively. The random variables $Y_{ij} = \ln(X_{ij}), j = 1, \dots, n_i$ follow the normal distribution $N(\mu_i, \sigma_i^2), i = 1,2$. Based on this fact, the following estimators of μ_i and σ_i^2, respectively, are considered

$$\hat{\mu} = \overline{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \quad and \quad \hat{\sigma}_i^2 = S_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(Y_{ij} - \overline{Y}_i)^2, i = 1,2.$$

Note that all four estimators under consideration $\overline{Y}_1, \overline{Y}_2, S_1^2$ and S_2^2 are independent because sample mean and sample variance for a normal population are independent and, moreover, the populations are independent.

We also need the following facts:

1. $E(\overline{Y}_i) = \mu_i$ and $E(S_i^2) = \sigma_i^2$.
2. $Var(\overline{Y}_i) = \frac{\sigma_i^2}{n_i}$ and $Var(S_i^2) = \frac{2\sigma_i^4}{n_i - 1}, i = 1,2$.

If we denote the basic statistics

$$V_1 = \overline{Y}_1, V_2 = S_1^2, V_3 = \overline{Y}_2, \ and \ V_4 = S_2^2,$$

then $\hat{\theta} = g(V_1, V_2, V_3, V_4)$, where the function

$$g(v_1, v_2, v_3, v_4) = \exp\left(v_1 + \frac{v_1}{2} - v_3 - \frac{v_4}{2}\right).$$

Partial derivatives are as follows:

$$\frac{\partial g(v_1, v_2, v_3, v_4)}{\partial v_1} = \exp\left(v_1 + \frac{v_1}{2} - v_3 - \frac{v_4}{2}\right) = g(v_1, v_2, v_3, v_4),$$
$$\frac{\partial g(v_1, v_2, v_3, v_4)}{\partial v_2} = \frac{1}{2} g(v_1, v_2, v_3, v_4),$$
$$\frac{\partial g(v_1, v_2, v_3, v_4)}{\partial v_3} = -g(v_1, v_2, v_3, v_4),$$
$$\frac{\partial g(v_1, v_2, v_3, v_4)}{\partial v_4} = -\frac{1}{2} g(v_1, v_2, v_3, v_4).$$

Remind that

$$\theta = g(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2) = \exp\left(\mu_1 + \frac{\sigma_1^2}{2} - \mu_2 - \frac{\sigma_2^2}{2}\right),$$

then the values of the partial derivatives at the point of means are:

$$\frac{\partial g(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)}{\partial v_1} = \theta,$$
$$\frac{\partial g(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)}{\partial v_2} = \frac{\theta}{2},$$
$$\frac{\partial g(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)}{\partial v_3} = -\theta,$$
$$\frac{\partial g(\mu_1, \sigma_1^2, \mu_2, \sigma_2^2)}{\partial v_4} = -\frac{\theta}{2}.$$

The linear terms of the Taylor expansion of the statistics $\hat{\theta}$ take form

$$\hat{\theta} = g(V_1, V_2, V_3, V_4) \approx \theta + \theta(V_1 - \mu_1) + \frac{\theta}{2}(V_2 - \sigma_1^2) - \theta(V_3 - \mu_2) - \frac{\theta}{2}(V_4 - \sigma_2^2)$$
$$= \theta\left(1 + (\overline{Y}_1 - \mu_1) + \frac{1}{2}(S_1^2 - \sigma_1^2) - (\overline{Y}_2 - \mu_2) - \frac{1}{2}(S_2^2 - \sigma_2^2)\right).$$

Therefore, the statistic $\hat{\theta}$ is asymptotically normal with the mean

$$Asymptotic\ Mean\ = \theta$$

Variance (remind that all four statistics $\overline{Y}_i$ and $S_i^2, i = 1,2$ are independent:

$$Asymptotic\ Varaince\ = \tau^2$$
$$= var\left\{\theta\left(1 + (\overline{Y}_1 - \mu_1) + \frac{1}{2}(S_1^2 - \sigma_1^2) - (\overline{Y}_2 - \mu_2) - \frac{1}{2}(S_2^2 - \sigma_2^2)\right)\right\}$$
$$= \theta^2\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \frac{1}{4}\frac{2\sigma_1^4}{n_1 - 1} + \frac{1}{4}\frac{2\sigma_2^4}{n_2 - 1}\right)$$
$$= \theta^2\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + \frac{\sigma_1^4}{2(n_1 - 1)} + \frac{\sigma_2^4}{2(n_2 - 1)}\right).$$

To obtain the plug-in estimator of the variance, we substitute estimations for μ_i and $\sigma_i^2, i = 1,2$.

$$\hat{\theta} = \exp\left(\overline{Y}_1 + \frac{S_1^2}{2} - \overline{Y}_2 - \frac{S_2^2}{2}\right),$$

$$\widehat{\tau^2} = \hat{\theta}^2\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \frac{S_1^4}{2(n_1 - 1)} + \frac{S_2^4}{2(n_2 - 1)}\right).$$

In the following we will deal with the second component of this formula, which we can call the variance component. Hence we use the formula

$$T = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + \frac{S_1^4}{2(n_1 - 1)} + \frac{S_2^4}{2(n_2 - 1)}}.$$

If the sample sizes for both sampling schemes tend to infinity, then

$$P(|\theta - \hat{\theta}| \le z_{\alpha/2}\tau) \sim 1 - \alpha,$$

where $z_{\alpha/2}$ is $(1 - \alpha/2)$-quantile of the standard normal distribution. Replacing τ^2 by its plug-in estimators $\widehat{\tau^2}$ presented in above, we obtain the same asymptotic equality.

Simple algebra shows that if the sample sizes in both sample schemes tend to infinity, then the intervals with the following end-points,

$$\hat{\theta}(1 \mp z_{\alpha/2}T) \tag{4}$$

are the asymptotic $(1 - \alpha)$-confidence sets for the ratio of log-normal means θ.

3. Simulation Studies and Results

A simulation study is performed to evaluate the coverage probability and interval width of the NA in comparison to three existing approaches: ML, MOVER and GCI for constructing 95% confidence intervals for ratio of means of two independent log-normal distributions. We use the nominal level $\alpha = 0.05$ and $N = 10,000$ simulated samples for each parameter setting. All simulations were carried out in R statistical software.

For our simulations, we used the combination of parameter values as follows. Without loss of generality, we set $\mu_1 = \mu_2 = 0$ and three values of σ_i^2 were 0.3, 0.5 and 1.0. Three values for n_i were 10, 50 and 100, $i = 1,2$.

Table I. Coverage probability and interval width of the approaches for constructing 95% confidence intervals for ratio of two log-normal means for homogeneity variances and balance designs

σ_1^2	σ_2^2	Methods	(n1,n2)=(10,10)		(n1,n2)=(50,50)		(n1,n2)=(100,100)	
			Coverage probability	Interval width	Coverage probability	Interval width	Coverage probability	Interval width
0.3	0.3	ML	0.9400	1.1267	0.9390	0.4676	0.9590	0.3296
		MOVER	0.9780	1.3960	0.9900	0.9742	0.9900	0.9199
		GCI	0.9610	1.4737	0.9430	0.4838	0.9610	0.3343
		NA	0.9330	1.0724	0.9360	0.4634	0.9600	0.3281
0.5	0.5	ML	0.9450	1.6497	0.9380	0.6379	0.9620	0.4471
		MOVER	0.9770	2.0023	0.9900	1.3448	0.9990	1.2692
		GCI	0.9630	2.4071	0.9390	0.6644	0.9600	0.4550
		NA	0.9260	1.5037	0.9360	0.6276	0.9560	0.4434
1.0	1.0	ML	0.9500	2.5652	0.9400	0.8740	0.9610	0.6061
		MOVER	0.9750	3.0297	0.9900	1.8874	0.9990	1.7765
		GCI	0.9650	4.6746	0.9470	0.9185	0.9620	0.6194
		NA	0.9120	2.1565	0.9360	0.8486	0.9500	0.5973

We report the results of the coverage probability and interval width of all approaches for CIs for ratio of means of two log-normal distributions. Table I for homogeneity variances and balance designs, and Table II for both homogeneous and heterogeneous variances under unbalance designs.

Table I for the homogeneity variances and balance designs, when $(\sigma_1^2, \sigma_2^2) = (0.3, 0.3), (0.5, 0.5)$, the coverage probability of the MOVER and GCI were greater than the nominal level for small to moderate sample sizes. For sample sizes were large, the coverage probability of all approaches performed well in terms of the coverage probability. However, the interval width of the NA was always shorter than other approaches. When $(\sigma_1^2, \sigma_2^2) = (1.0, 1.0)$, the coverage probability of all approaches were performed well for small to moderate sample sizes. The NA approach performed well in terms of the interval width.

Table II. Coverage probability and Interval width of four approaches for constructing 95% confidence intervals for ratio of two log-normal means for both homogeneous and heterogeneous variances under unbalance designs

σ_1^2	σ_2^2	Methods	(n1,n2)=(50,10)		(n1,n2)=(100,10)		(n1,n2)=(100,50)	
			Coverage probability	Interval width	Coverage probability	Interval width	Coverage probability	Interval width
0.3	0.3	ML	0.9340	0.8350	0.9190	0.7811	0.9510	0.4021
		MOVER	0.9930	1.1895	0.9930	1.1466	0.9990	0.9423
		GCI	0.9560	0.9045	0.9510	0.8485	0.9510	0.4093
		NA	0.9380	0.8100	0.9220	0.7601	0.9490	0.3995
0.5	0.5	ML	0.9340	1.1661	0.9190	1.0862	0.9510	0.5468
		MOVER	0.9920	1.6502	0.9910	1.5878	0.9990	1.3000
		GCI	0.9540	1.2201	0.9470	1.1336	0.9520	0.5561
		NA	0.9320	1.1031	0.9220	1.0336	0.9510	0.5403
1.0	1.0	ML	0.9330	1.6607	0.9180	1.5360	0.9490	0.7448
		MOVER	0.9900	2.3399	0.9840	2.2436	0.9990	1.8210
		GCI	0.9550	1.6481	0.9460	1.5060	0.9490	0.7555
		NA	0.9410	1.5034	0.9140	1.4054	0.9510	0.7288
0.3	0.5	ML	0.9270	1.0045	0.9120	0.9566	0.9490	0.4543
		MOVER	0.9810	1.3686	0.9750	1.3273	0.9990	1.0568
		GCI	0.9510	1.0467	0.9420	0.9985	0.9550	0.4597
		NA	0.9340	0.9533	0.9180	0.9117	0.9510	0.4496
0.5	1.0	ML	0.9280	1.4467	0.9160	1.3820	0.9400	0.6022
		MOVER	0.9760	1.8588	0.9670	1.8081	0.9990	1.4069
		GCI	0.9560	1.3543	0.9390	1.2906	0.9470	0.6012
		NA	0.9360	1.2794	0.9090	1.2336	0.9466	0.5883
0.3	1.0	ML	0.9270	1.2743	0.9140	1.2329	0.9580	0.3831
		MOVER	0.9620	1.5935	0.9520	1.5603	0.9990	1.1491
		GCI	0.9550	1.1833	0.9400	1.1488	0.9630	0.3930
		NA	0.9290	1.1302	0.9070	1.1022	0.9530	0.3785

Table II for both homogeneous and heterogeneous variances under unbalance designs. When the sample sizes were small to moderate $(n_1, n_2) = (50,10), (100,10)$, all approaches performed well in terms of the coverage probability , except when $(\sigma_1^2, \sigma_2^2) = (0.5, 1.0)$ the coverage probability of the ML was less than the nominal level Furthermore, the NA approach always performed well in terms of the interval width for all values of the variances.

4. A Real Life Example

According to Tai et al. [20], the illustrative example deals with survival times in months for patients who died from a particular cancer. Data of the first group were constructed for 184 patients who had limited stage small-cell lung cancer (LC) and the survival time for the second group were constructed from 38 patients who died of cervical cancer (CC). After taking the natural logarithm, the calculated statistics for these two group of log-transformed measurements are as follows Table III.

Table III. Sample size, sample means and sample variances for the two data sets

Groups	n_i	μ_i	σ_i^2
Small-Cell Lung cancer (LC)	184	2.8591	0.2461
Cervical cancer (CC)	38	3.3900	0.6454

The histogram and corresponding density plots, Q-Q plot and fitted density for log-transformed data for both groups were showed in Fig. 1(a)-(b).

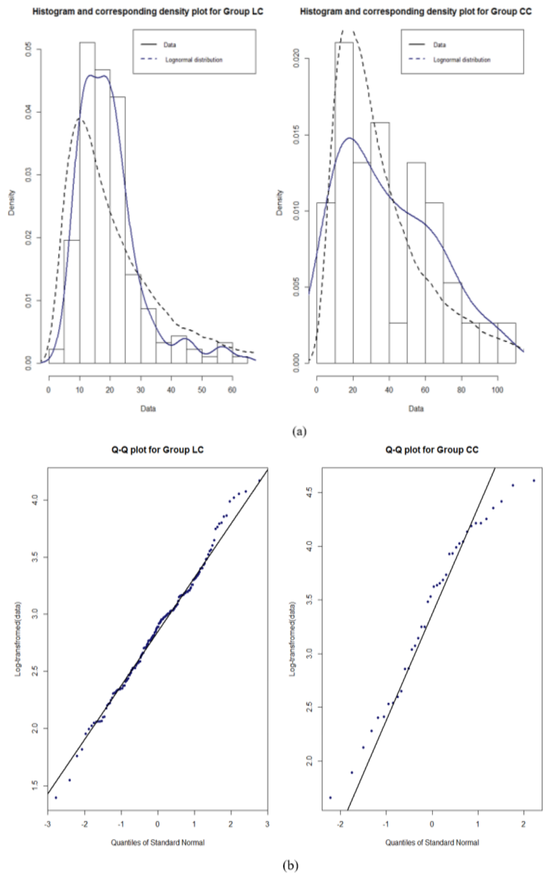

Fig. 1 (a) The histogram and corresponding density plots for both groups, (b) and the Q-Q plot and fitted density for log-transformed data for both groups

Moreover, To confirm that the log-normal distributions appropriated for two datasets. The Shapiro-Wilk test was used and p-values for the first group was 0.5306 and for the second group was 0.1186. These results indicate that the log-normal models is adequate for two datasets.

To compare the ratio of two log-normal means, the 95% confidence interval (lower, upper) and the average width by using the ML, MOVER, GCI, and NA approaches are demonstrated in Table IV.

Table IV. 95% confidence interval (lower, upper) and the average width by using the ML, MOVER, GCI and NA approaches.

Approach	ML	MOVER	GCI	NA
95% CI	(0.3552, 0.6529)	(0.2791, 0.9124)	(0.3366, 0.6348)	(0.3350, 0.6282)
Average width	0.2977	0.6333	0.2982	0.2932

This is in agreement with our simulation study which indicated that the NA approach is shorter interval width than other approaches.

5. Discussion

In this paper, we wold like to identify potential methods that can be recommended to practitioners for constructing confidence interval for the ratio of means of two independent log-normal distributions. From the simulation results presented in Table I and II.

In Table I for homogeneity variances and balance designs, when small to moderate sample sizes the GCI and MOVER approaches were better than other approaches in terms of the coveragre probability and interval width. The NA approach always shortest interval width especially when sample sizes were larger and variance were decrease. For both homogeneous and heterogeneous variances and unbalance designs were shown in Table II. When small to moderate sample sizes, MOVER leads to coverage probabilities closer to the nominal level while GCI approach are better than the MOVER in terms of interval width. For large sample sizes, NA has better coverage probability and interval width than other appraoches.

6. Conclusion

The purpose of the paper is to propose a simple approach for the problem of the CI for ratio of means of two log-normal distributions and compare performance with the existing approaches. A simulation studies and results were conduct to compare the performance of these methods in terms of the coverage probability and interval width to indicate which method is optimal under diffrence condition of the population of two independent log-normal distributions. The results indicate that the GCI and MOVER approaches perform well for all situations. However, when he sample size were large, the NA approach performed better than other appraoches in terms of interval width.

Acknowledgment

The authors thank the reviewers for suggestions which lead to considerable improvements in the manuscript.

References

[1] K. Kondo, “The log-normal distribution of the incubation time of exogenous diseases,” *Japanese Journal of Human Genetics*, vol. 21, pp. 217-237, 1977.

[2] W.R.Ott, “Environmental Indices,” *Ann Arbor, MI, Ann Arbor Science*, 1978.

[3] C.Di Giorgio, A.Krempff, H.Guiraud, P.Binder, C.Tiret, G.Dumenil, “Atmospheric pollution by airborne microorganisms in the city of Marseilles,” *Atmospheric Environmental*, vol. 30, pp. 155-160, 1996.

[4] Bundesamt fur Statistik, “Statistisches Jahrbuch der Schweiz,” *Verlag Neue Zricher Zeitung*, 1997.

[5] I. Antoniou, V. Ivanov, P. Zrelov, “On the lognormal distribution of stock market data,” *Physica A: Statistical Mechanics and its Applications*, vol. 331, no. 3-4, pp. 617-638, 2004.

[6] J. E. Angus, “Bootstrap one-sided confidence intervals for the log-normal mean,” *The Statistician*, vol. 43, no.3, pp. 395–401, 1994.

[7] C. E. Land, “An evaluation of approximate confidence interval estimation methods for lognormal means,” *Technometrics*, vol. 14, no. 1, pp. 145–158, Feb. 1972.

[8] X.H. Zhou, and S. Gao, “Confidence intervals for the log-normal mean,” *Statistics in Medicine*, vol. 16, pp. 783–790, 1997.

[9] U. Chand, “Distributions related to comparison of two means and two regression coefficients,” *Annals of Mathematical Statistics* ,vol. 21, pp. 507-522, 1950.

[10] H. R. Van der Vaart, “On the robustness of Wilcoxon's two-sample test,” *In Quantitive Methods in Pharmacology, H. D. Jonge (ed),* pp. 140-158, 1961.
[11] J. W. Pratt, “Robustness of some procedures for the two-sample location problem,” *Journal of the American Statistical Association,* vol. 59, pp. 665-680, 1964.
[12] E. L. Lehmann, Nonparametrics: Statistical Methods Based on Ranks. *San Francisco: Holden-Day,* 1975.
[13] X.H. Zhou, S. Gao, and S. L. Hui, “Methods for comparing the means of two independent log-normal samples,” *Biometrics*, vol. 53, no. 3, pp. 1129-1135, Sep. 1997.
[14] K. W. Tsui, and S. Weerahandi, “Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters,” *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 602–607, Jun. 1989.
[15] S. Weerahandi, “Generalized confidence intervals,” *Journal of the American Statistical Association*, vol. 88, pp. 899–905, 1993.
[16] K. Krishnamoorthy, and T. Mathew, “Inferences on the means of lognormal distributions using generalized p-values and generalized condence intervals,” *Journal of Statistical Planning and Inference*, vol. 115, pp. 103–120, 2003.
[17] G.Y. Zou, C.Y. Huo, and J. Taleban, “Simple confidence intervals for lognormal means and their differences with environmental applications,” *Environmetrics*, vol.20, pp. 172-180, 2009.
[18] G.Y. Zou, and A. Donner, “Construction of confidence limits about effect measures: A general approach,” Stat. Med, vol. 27, pp.1693–1702, 2008.
[19] E.L. Lehmann, Elements of Large Sample Theory. *Springer*, 2004, pp.85-91.
[20] P. Tai et al., “Twenty-year follow-up study of long-term survival of limited-stage small- cell lung cancer and overview of prognostic and treatment factors,” *Int. Journal of Radiation Oncology Biol. Phys*, vol. 56, no. 3, pp. 626–633, 2003.