COMPARISON OF POPULATION MEANS FOR SKEWED DATA WITH UNEQUALVARIANCE

Evren ÖZKİP[1], Berna YAZICI[1], Ahmet SEZER[1], Betül KAN[1]

[1]Anadolu University, Science Faculty, Department of Statistics, Eskişehir/TURKEY
e.ozkip@anadolu.edu.tr , bbaloglu@anadolu.edu.tr, a.sezer@anadolu.edu.tr, bkan@anadolu.edu.tr

Abstract

Faced with a positively right-skewed data, lognormal distribution is the most widely used. Lognormal distributions play a central role in biological and health research. The aim of this study is first show exact confidence intervals and tests for a single lognormal mean using generalized p-value and generalized confidence interval approach and the procedure is also used for obtaining confidence intervals and tests for the difference of two lognormal means. Simulation results demonstrate that the coverage accuracy of proposed confidence interval and power of proposed exact tests are satisfactory.

Keyword: Generalized p-value, Generalized confidence interval, Coverage probability.

1.Introduction

Let X be a random variable having a lognormal distribution, and let μ and σ^2, respectively, denote the mean and variance of $In(X)$ so that $Y = In(X) \sim N(\mu, \sigma^2)$. Many of the parameters of interest concerning the lognormal distribution (for example the mean of X) turn out to be functions of both μ and σ^2 it appears difficult to obtain exact and/or optimum tests and confidence intervals. In particular, the mean of the lognormal distribution is given by

$$E(X) = E(\exp(Y)), \qquad \text{where } \eta = \mu + \frac{\sigma^2}{2} \tag{1.1}$$

Clearly, the computation of confidence intervals and test procedures concerning the mean of X is equivalent to the computation of the corresponding quantities for η. A problem of interest in this context is statistical inference concerning the mean of the lognormal distribution. For obtaining confidence intervals and tests for a single lognormal mean, the available small sample procedures are based on a certain conditional distribution, and are computationally very involved and this makes the procedure somewhat difficult to use in practice (Krishnamoorthy and Mathew 2003).

Some simple procedures for obtaining confidence intervals for a single lognormal mean are viewed and compared by Zhou and Gao (1997). These include a large sample method due to Cox, reported in Land (1972), a conservative method due to Angus (1988), and a parametric bootstrap method, also due to Angus (1994). The numerical results in Zhou and Gao (1997) show that in terms of coverage probability, all of these procedures are too conservative or too liberal, unless the sample size is big, in which case, the procedure due to Cox is satisfactory.

The concepts of generalized confidence interval and generalized *P*-value have been widely applied to a variety of practical setting where standard solutions do not exist for confidence intervals and hypothesis testing. The proposed generalized variable approach not only can provide confidence intervals, but also can provide *P*-values for hypothesis testing (Tian and Wu 2007). The generalized p-value has been introduced by Tsui and Weerahandi (1989) and the generalized confidence interval by Weerahandi (1993); see the book by Weerahandi (1995) for a detailed discussion. The concept of generalized confidence interval and generalized P-value have been widely applied to a variety of

practical setting where standard solutions do not exist for confidence intervals and hypothesis testing, see Weerahandi(1995), Krishnamoorthy and Mathew (2003) and Tian and Wu (2007).

2. The generalized pivots for interval estimation

Suppose that $\mathbf{X}= X_1, \ldots, X_n$ form a random sample from a distribution which depends on the parameters (θ, υ) where $\theta$the parameter of interest is and $\boldsymbol{\upsilon}^T$ is a vector of nuisance parameters. A generalized pivot $R(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon})$, where $\mathbf{x}$ is an observed value of $\mathbf{X}$, for interval estimation has the following two properties:

1. $R(\mathbf{X}; \boldsymbol{x}, \theta, \boldsymbol{\upsilon})$ has a distribution free of unknown parameters.
2. The value of $R(\mathbf{x}; \mathbf{x}, \theta, \boldsymbol{\upsilon})$ is θ.

(2.1)

Let that R_α be the 100αth percentile of R. Then R_α is an estimate of $100(1-\alpha)$ per cent lower bound for θ and $(R_{\alpha/2}, R_{1-\alpha/2})$ is a $100(1-\alpha)$ per cent two-sided confidence interval for θ.

3. The generalized test variables for hypothesis testing

Consider testing $H_0: \theta < \theta_0$ versus $H_1: \theta > \theta_0$ where θ_0 is a speci_ed quantity. A generalized test variable of the form $T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon})$, where $\mathbf{x}$ is an observed value of $\mathbf{X}$, is chosen to satisfy the following three conditions:

1. For fixed $\mathbf{x}$, the distribution of $T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon})$ is free of the vector of nuisance parameters $\boldsymbol{\nu}$.
2. The value of $T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon})$ at $\mathbf{X}=\mathbf{x}$ is free of any unknown parameters.
3. For fixed $\mathbf{x}$ and $\boldsymbol{\nu}$, and for all t; $Pr[T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon}) > t]$ is either an increasing or a decreasing function of θ. (3.1)

A generalized extreme region is defined as $C = [\mathrm{X}: T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon}) \geq T(\mathbf{x}; \mathbf{x}, \theta, \boldsymbol{\upsilon})]$ (or $C = [\mathrm{X}: T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon}) \leq T(\mathbf{x}; \mathbf{x}, \theta, \boldsymbol{\upsilon})]$) if $T(\mathbf{X}; \mathbf{x}, \theta, \boldsymbol{\upsilon})$ is stochastically decreasing (or increasing) in θ. The generalized *p*-value is defined as $sup_{H_0} P(C/H_0)$. (Weerahandi, 1995)

4. A generalized pivot for lognormal mean

Krishnamoorthy and Mathew (2003) have proposed a generalized confidence interval approach for η. Let X be a random variable following a lognormal distribution, and let μ and σ^2 denote the mean and the variance of $ln(X)$, respectively, so that $Y = In(X) \sim N(\mu, \sigma^2)$. Then the mean of X is as defined in (1.1). Consider the problem of testing

$$H_0: \eta \leq \eta_0 \text{ vs. } H_1: \eta > \eta_0 \tag{4.1}$$

where $\eta = \mu + \sigma^2/2$ and η_0 is a specified quantity.

$$\begin{aligned} T_1 &= \bar{y} - \frac{\bar{Y} - \mu}{S/\sqrt{n}} s/\sqrt{n} + \frac{1}{2}\frac{\sigma^2}{S^2} s^2 - \eta \\ &= \bar{y} - \frac{Z}{U/\sqrt{n-1}} \frac{s}{\sqrt{n}} + \frac{1}{2}\frac{s^2}{U^2/(n-1)} - \eta \end{aligned} \tag{4.2}$$

where $Z = \sqrt{n}(\bar{Y} - \mu)/\sigma \sim N(0,1)$ independently of $U^2 = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. Here χ^2_r denotes the central chisquare distribution with $df = r$. where $\bar{Y}$ and S^2 are the sample mean and variance of the log transformed data, $\bar{y}$ and s^2 are the observed sample mean and variance of the log transformed data. It is clear that T_1 satisfies the conditions in (3.1) and the distribution of T_1 is stochastically decreasing

in . The generalized p-value for testing the hypotheses in (4.1) is thus given by $P(T_1 \leq 0 \backslash \eta = \eta_0)$. The test based on the generalized *p*-value rejects H_0 if the generalized *p*-value is less than some specified (say, $\alpha = 0.05$). Let

$$T_2 = \bar{y} - \frac{\bar{Y} - \mu}{S/\sqrt{n}} s/\sqrt{n} + \frac{1}{2}\frac{\sigma^2}{S^2} s^2$$
$$= \bar{y} - \frac{Z}{U/\sqrt{n-1}} \frac{s}{\sqrt{n}} + \frac{1}{2}\frac{s^2}{U^2/(n-1)} \quad (4.3)$$

Notice that T_2 reduces to η when $\bar{Y} = \bar{y}$ and $S^2 = s^2$, and the distribution of T_2 is free of any unknown parameters.

Both the generalized p-value and the generalized confidence interval can be computed using the following algorithm.

Algorithm 1.

For a given data set $x_1, \ldots, x_n$, set $y_i = In(x_i)$, $i = 1, \ldots, n$.

Compute $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $s^2 = (1/(n-1))\sum_{i=1}^{n}(y_i - \bar{y})^2$

For $i = 1$ to m

Generate $Z \sim N(0,1)$ and $U^2 \sim \chi^2_{n-1}$

Set $T_{2i} = \bar{y} - \left(Z/\left(U/\sqrt{n-1}\right)\right) s/\sqrt{n} + \frac{1}{2} s^2/U^2/(n-1)$

(end i loop)

Let $K_i = 1$ if $T_{2i} \leq \eta_0$, else $K_i = 0$

$(1/m)\sum_{i=1}^{m} K_i$ is a Monte Carlo estimate of the generalized p-value for testing (1.3) The $100(1-\alpha)$th percentile of $T_{21}, \ldots, T_{2m}$, denoted by $T_2(1-\alpha)$, is a Monte Carlo estimate of the $100(1-\alpha)$% generalized upper confidence limit for $\eta = \mu + \sigma^2/2$

In order to understand the performance of the generalized p-value and the generalized confidence limits, we estimated the coverage probabilities of the generalized confidence interval as follows:

Algorithm 2. For specified values of n, μ, σ and $0 < \alpha < 1$:

For $i = 1, m_1$

Generate $\bar{y}$ from $N(\mu, \sigma^2/n)$

Generate Q from χ^2_{n-1}, and set $s^2 = \sigma^2 Q/(n-1)$

For $j = 1, m_2$

Generate $Z \sim N(0,1)$ and $U^2 \sim \chi^2_{n-1}$

Set $T_{2i} = \bar{y} - \left(Z/\left(U/\sqrt{n-1}\right)\right) s/\sqrt{n} + \frac{1}{2} s^2/U^2/(n-1)$

(end j loop)

If the $100(1-\alpha)$th percentile $T_2(1-\alpha)$ of $\{T_{21}, \ldots, T_{2m_2}\}$ is greater than $\eta = \mu + \frac{\sigma^2}{2}$.

Set $K_i = 1$; else set $K_i = 0$

(end i loop)

$(1/m)\sum_{i=1}^{m} K_i$ is an estimate of the coverage probability of the generalized upper confidence limit.

All computer simulation were carried out in the R statistical programming environment (R Development Core Team, 2012). We conducted a simulation study for generalized confidence interval and coverage probability. For simulation design, we randomly generated 50 000 random samples. Simulation was computed for the parameter value $\mu = 1$, $\sigma^2 = 0.1,1,5,10$, $n = 5,25,100$ and

$1-\alpha=0.95$. From the numeric results in Table 1, generalized confidence interval and coverage probability are much more satisfactory even when sample sizes are small. The coverage probabilities of generalized of generalized intervals are always close to the nominal level when sample sizes are small and the σ^2's are large.

Table 1
Confidence interval and coverage probability of two sided limits for $\eta=\mu+\sigma^2/2$ based on generalized *p*-value test (GP) at 5% significance level

n	s	CP	Upper	Lower	η_0
5	0.1	0.946	0.893	1.147	1.005
5	1	0.936	0.667	5.590	1.5
5	5	0.952	4.737	108.49	13.5
5	10	0.953	18.087	410.91	51
25	0.1	0.946	0.964	1.047	1.005
25	1	0.946	1.088	2.147	1.5
25	5	0.946	8.277	25.301	13.5
25	10	0.950	24.307	168.45	51
100	0.1	0.943	0.985	1.025	1.005
100	1	0.946	1.280	1.770	1.5
100	5	0.950	10.483	18.029	13.5
100	10	0.946	39.375	68.449	51

6. Comparing the means of two lognormal distribution

Krishnamoorthy and Mathew (2003) have developed a generalized variable approach for comparing the means of two lognormal distribution. They use the following generalized pivotal quantity; let X_1 and X_2 be two independent lognormal random variable, so that $Y_1=In(X_1)\sim N(\mu_1,\sigma_1^2)$ and $Y_2=In(X_2)\sim N(\mu_2,\sigma_2^2)$. Let $\eta_1=\mu_1+\sigma_1^2/2$ and $\eta_2=\mu_2+\sigma_2^2/2$ so that $E(X_1)=exp(\eta_1)$ and $E(X_2)=exp(\eta_2)$. The problem of testing

$$H_0: exp(\eta_1)\le exp(\eta_2)\ vs\ H_1: exp(\eta_1)>exp(\eta_2) \tag{6.1}$$

is equivalent to

$$H_0: \eta_1\le\eta_2 vs\ H_1: \eta_1>\eta_2 \tag{6.2}$$

let

$$T_i=\bar{y}_i-\frac{Z_i}{U_i/\sqrt{n_i-1}}\frac{s_i}{\sqrt{n_i}}+\frac{1}{2}\frac{s_i^2}{U_i^2/(n_i-1)}, i=1,2 \tag{6.3}$$

where $Z_i=\sqrt{n_i}(\bar{Y}_i-\mu_i)/\sigma_i\sim N(0,1)$ and $U_i^2=(n_i-1)S_i^2/\sigma_i^2\sim\chi^2_{n_i-1}$, for $i=1,2$ and $\bar{Y}_i, S_i^2$ are defined as:

$$Y_i=\frac{1}{n_i}\sum_{j=1}^{n_i}Y_{ij},\qquad S_i^2=\frac{1}{n_i-1}\sum_{j=1}^{n_i}\left(Y_{ij}-\bar{Y}_i\right)^2,\qquad i=1,2$$

The simulation study was carried out along the lines of Algorithm 2 given for the one-sample case. For the simulations, we have taken μ_2, without loss of generality. Various combinations of σ_1^2,σ_2^2 and μ_1 are considered in Table 2, and all the results correspond to a nominal level of %5. The numeric results in Krishnamoorthy and Mathew (2003) show that sample large test has a small power when the sample sizes are small and $\sigma_1^2\neq\sigma_2^2$. The power to generalized *p*-value test even when

sample sizes are small and $\sigma_1^2 \neq \sigma_2^2$ is extremely satisfactory and it is applicable regardless of the sample size.

Table 2
Power of the generalized *p*-value test (GP test) at 5% significance level when $\mu_2 = 0$ and $H_1: \eta_1 > \eta_2$

n_1	n_2	μ_1	σ_1^2	σ_2^2	Power
5	5	0	10	2	0.3248
5	5	0	20	2	0.5684
5	5	3	4	2	0.4952
10	10	0	10	4	0.2830
10	10	0	20	4	0.6680
10	10	3	20	4	0.9000
25	25	1	1	1	0.8292
25	25	0	4	1	0.7444
25	50	1	1	1	0.9568
25	50	0	4	1	0.8248
100	25	1	5	4	0.4396
100	25	0	3	1	0.8100

7. Conclusion

In this paper, we propose generalized confidence intervals for concerning the mean of single lognormal distribution and procedure for power of the difference of two lognormal means. By simulation, we demonstrate that the proposed confidence interval provides satisfactory coverage probabilities and good balance between upper and lower limit. The procedures are easy to compute and are applicable to small samples.

Acknowledgment

This study is supported by the Research Foundation of Anadolu University. (Project Number: 1202F38).

REFERENCES

Angus, J.E., (1988). '*Inferences on the lognormal mean for complete samples*.' Comm. Statistical Simulation Computation 17, 1307–1331.

Angus, J.E., (1994). '*Bootstrap one-sided confidence intervals for the lognormal mean*.' Statistician 43, 395–401.

Gill, P. S., (2004). '*Small-sample inference for the comparison of means of log-normal distributions*.' Biometrics 60, 525–527.

Krishnamoorthy, K. and Mathew, T., (2003). '*Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals*.' Journal of Statistical Planning and Inference 115, 103–121.

Koch, L. A., (1966). The logarithm in biology. J. Theoret. Biol. 12, 276-290.

Land, C.E., (1972). '*An evaluation of approximate confidence interval estimation methods for lognormal means*.' Technometrics 14, 145–158.

R Development Core Team. (2012) R: A language and environment for statistical computing. *R foundation for Statistical Computing.*

Tian, L., and Wu, J., (2007). '*Inferences on the Common Mean of Several Log-Normal Populations: The Generalized Varianle Approach.*' Biometrical Journal 49, 944-951.

Tsui, K.W., Weerahandi, S., (1989). '*Generalized p-values in sigini&cance testing of hypotheses in the presence of nuisance parameters*.' J. Am. Statist. Assoc. 84, 602–607.

Weerahandi, S., (1993). '*Generalized confidence intervals*.' J. Am. Statist. Assoc. 88, 899–905.

Weerahandi, S., (1995). '*Exact Statistical Methods for Data Analysis*.' Springer, New York.

Zhou, L., Mathew, T.,(1994). '*Some tests for variance components using generalized p-values.*' Technometrics 36, 394–402.