Journal of Statistics Education

Confidence Intervals for the Mean of a Log-Normal Distribution

Ulf Olsson
Swedish University of Agricultural Sciences

Journal of Statistics Education Volume 13, Number 1 (2005), jse.amstat.org/v13n1/olsson.html

Copyright © 2005 by Ulf Olsson, all rights reserved. This text may be freely shared among individuals, but it may not be republished in any medium without express written consent from the authors and advance notification of the editor.

Key Words: Generalized confidence interval.

Abstract

Methods for calculating confidence intervals for the mean are reviewed for the case where the data come from a log-normal distribution. In a simulation study it is found that a variation of the method suggested by Cox works well in practice. An approach based on Generalized confidence intervals also works well. A comparison of our results with those of Zhou and Gao (1997) reveals that it may be preferable to base the interval on t values, rather than on z values.

1. The problem

In applied statistics classes we sometimes come across data that need to be transformed prior to analysis. For example, income data can often be considered to be log-normal. One way of analyzing such data is to log-transform the original variable X and to base the inference on the transformed variable $Y = \log(X)$. This means that we assume that the distribution from which our data emerges can be approximated with a log-normal distribution. In this paper we will discuss interval estimation of the arithmetic mean value of X in a log-normal distribution. It is true that the median is often used to describe the average of skewed distributions like income distributions. However, there are situations when the arithmetic mean is a parameter of interest. For example, in a sample survey, a confidence interval for the average income can be used to calculate a confidence interval for the total income in the population.

Note that if X is log-normal, then the median of Y is equal to the log of the median of X. In this paper we will assume that it is the arithmetic mean of X, and not the median of X, that we want to make inference about.

It is a rather straight-forward task to use the log-transformed data Y to calculate a confidence interval for the expected value (mean value) of Y. We will discuss how this result can be used to calculate a confidence interval for the expected value of X.

2. Theory and notation

Let X denote the original variable that follows a log-normal distribution. X has expected value $E(X)=\theta$ and variance $Var(X)=\delta^2$. We let Y denote the log-transformed, normally distributed variable $Y = \log(X)$, that has mean value $E(Y)=\mu$, and variance $Var(Y)=\sigma^2$. Denote the sample mean of Y with $\overline{Y}$, and the sample variance of Y with s^2.

It holds (see e.g. Zhou and Gao, 1997) that

$$\theta = e^{\left(\mu+\frac{\sigma^2}{2}\right)} \tag{1}$$

This means that the mean value of X is not equal to the antilog of the mean value of Y. An estimator of $\log(\theta)$ can be calculated from sample data as

$$\widehat{\log(\theta)} = \left(\overline{Y} + \frac{S^2}{2}\right) \tag{2}$$

An estimator of the variance of $\widehat{\log(\theta)}$ is given by

$$\widehat{Var}\left(\widehat{\log(\theta)}\right) = \frac{S^2}{n} + \frac{S^4}{2(n-1)} \tag{3}$$

see e.g. Zhou and Gao, (1997).

3. Confidence intervals for $E(X)=\theta$

3.1 A numerical example

We will illustrate a number of methods for computing a confidence interval for θ using a small numerical example. The methods include a naïve method based on transformation and back-transformation; a method proposed by Cox; a modified version of the Cox method; a method motivated by large-sample theory; and a method based on generalized confidence intervals (Weerahandi, 1993; Krishnamoorthy and Mathew, 2003). Other methods that have been suggested for the same purpose are reviewed in Zhou and Gao (1997), but according to their simulation results the Cox method works well in large samples, and reasonably well even in small samples.

One sample of n=40 observations was generated, using SAS (1997) software, from a

log-normal distribution with parameters μ = 5 and σ = 1. The population mean of X is $\theta = e^{\left(\mu+\frac{\sigma^2}{2}\right)}$ = 244.69. The observations were transformed as Y=log(X). The raw sample data are given in Table 1. The sample data are summarized in Table 2.

Table 1. A sample of data from a log-normal distribution.

914.9	1568.3	50.5	94.1	199.5	23.8	70.5	213.1
44.1	331.7	139.3	115.6	38.4	357.1	725.9	253.2
905.6	155.4	138.1	95.2	75.2	275.0	401.1	653.8
390.8	483.5	62.6	128.5	81.5	218.5	308.2	41.2
60.3	506.9	221.8	112.5	93.7	199.3	210.6	39.2

Table 2. Summary statistics for the sample data.

Variable	**Mean**	**Median**	**St. dev.**
X	274.963	177.350	310.343
Y = log(X)	5.127	5.170	1.004

3.2 Naïve method

It would seem natural to use the following "naïve" approach for calculating a confidence interval for θ. A confidence interval for μ is calculated using standard methods. The limits of the confidence interval are back-transformed to give the limits in a confidence interval for θ.

For our example data, the naïve approach would produce the point estimate $\tilde{\theta}$ = $e^{5.127}$=168.51. A standard 95% confidence interval for μ is calculated as $5.127 \pm 2.02\sqrt{\frac{1.010}{40}}$ with limits [4.806, 5.448]. This would give limits for θ as $e^{4.806}$ = 122.24 and $e^{5.448}$ = 232.29. Note that this confidence interval does not cover the population mean value, which is 244.69. Of course, this can occur because of chance; after all, we have only studied one single sample so far. However, it is noteworthy that the interval does not even cover the *sample* mean, which is 275.0. This illustrates the fact that the naïve method gives a biased estimator of θ.

3.3 Cox method

Cox (quoted as "personal communication" in Land, 1971) has suggested that a

confidence interval for $E(X)=\theta$ can be calculated in the following way:

Calculate a confidence interval for $\log(\theta)$ as

$$\bar{Y} + \frac{S^2}{2} \pm z\sqrt{\frac{S^2}{n} + \frac{S^4}{2(n-1)}} \qquad (4)$$

where z is the appropriate percentage point of the standard Normal distribution. The limits in this confidence interval are back-transformed to give a confidence interval for θ . The method is valid for large samples. A similar approach has been suggested by Zhou, Gao, and Hui (1997) for the two-sample case.

For the sample data, $\bar{y}$ =5.127 and s^2=1.010. The 95% confidence interval for $\log(X)$ is $5.127 + \frac{1.010}{2} \pm 1.96\sqrt{\frac{1.010}{40} + \frac{1.010^2}{2(40-1)}}$ with confidence limits [5.248, 6.016]. Taking anti-logs we obtain the limits in the 95% confidence interval for θ as $e^{5.248}$ = 190.24 and $e^{6.016}$ = 409.82, respectively. A point estimate of θ is $\hat{\theta} = e^{\left(\bar{Y}+\frac{S^2}{2}\right)} = e^{\left(5.127+\frac{1.010}{2}\right)} = 279.22$.

3.4 Cox method: a modified version

In the version of (4) that was given in Zhou and Gao (1997), the standard normal variate z was used. We propose to use t, with degrees of freedom based on the d.f. for the estimate of σ^2. There are several reasons for this suggestion. One reason is that a confidence interval for μ would base the interval on t. A second reason is simply that this will produce confidence intervals with coverage closer to the nominal level. The use of z instead of t might explain the rather poor performance of the Cox interval, for small n, in the simulations in Zhou and Gao (1997); our results presented below are considerably better.

For the sample data, $\bar{y}$ = 5.127 and s^2 = 1.010. The 95% confidence interval for $\log(X)$ is $5.127 + \frac{1.010}{2} \pm 2.02\sqrt{\frac{1.010}{40} + \frac{1.010^2}{2(40-1)}}$ with confidence limits [5.237, 6.027]. Taking anti-logs we obtain the limits in the 95% confidence interval for θ as $e^{5.237}$ = 188.0 and $e^{6.027}$ = 414.7, respectively. For this sample size, the difference compared to the standard Cox method is small.

3.5 Generalized confidence intervals

Generalized confidence intervals (Weerahandi, 1993) can be used for inference about parameters where the sampling distribution is complicated. As noted in equation (2), the lognormal mean is a function of $\bar{y}$, which can be assumed to be Normally distributed, and S^2, which is a function of a χ^2 variate. Krishnamoorthy and Mathew

(2003, p. 108) suggested the following procedure for computing a confidence interval for the lognormal mean:

Calculate $\bar{y}$ and s^2 from the data.

For $i = 1$ to m (where m is large, for example m=10000)

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

For each i, calculate $T_{2i} = \bar{y} - \frac{Z}{U/\sqrt{(n-1)}}\frac{s}{\sqrt{n}} + \frac{1}{2}\frac{s^2}{U^2/(n-1)}$.

(end i loop)

For a 95% confidence interval, the 2.5% and 97.5% percentiles for T_2 are calculated from the 10000 simulated values. These are the lower and upper limits in a confidence interval for $\mu + \frac{\sigma^2}{2}$. This means that a 95% confidence interval for the lognormal mean is obtained as [exp($T_{2;0.025}$), exp($T_{2;0.975}$)].

3.6 An approach based on large-sample theory

Instead of basing the calculations on transformed data the confidence interval may be calculated from the sample mean and sample variance of X directly, without using any transformations. According to the Central limit theorem, the distribution of a sample mean $\bar{X}$ can be approximated with a normal distribution if n is reasonably large, for a large class of distributions. Thus, for large samples we can calculate the confidence interval as

$$\bar{X} \pm z\sqrt{\frac{S_x^2}{n}} \qquad (5)$$

In our example, the 95% confidence interval can be calculated as $275.0 \pm 1.96\sqrt{\frac{96286.1}{40}}$, which gives the limits as [178.84, 371.16].

4. An application

The data in Table 3 are nine measurements of carbon monoxide levels in the air. The measurements were made close to a California oil refinery in 1990 - 1993. We will use these data to obtain confidence intervals for the mean carbon monoxide level. Initial investigations of these data, and of other similar datasets, indicates that a log-normal model may be appropriate. The data are posted at lib.stat.cmu.edu/DASL/.

Table 3. Carbon monoxide levels at an oil refinery in California.

CO level	Date
12.5	9/11/90
20	10/4/90
4	12/3/91
20	12/10/91
25	5/7/92
170	8/6/92
15	9/10/92
20	9/22/92
15	3/30/93

The 95% confidence intervals for the example data, using the different methods we have discussed, are given in Table 4. It may be noted that our modified Cox method gives a somewhat wider interval than the Cox method, as expected. The generalized confidence interval has an upper limit that is well above the others, for these data.

Table 4. Lower and upper limits in confidence intervals using the different methods.

Method	Lower limit	Upper limit
Naïve approach	9.15	40.95
Cox method	14.15	68.49
Modified Cox method	12.31	78.72
Large-sample approach	-6.11	73.11
Generalized confidence interval	16.65	153.19

5. A simulation study

Samples of sizes 5 to 500 were generated from a log-normal distribution with parameters $\mu = 5$ and $\sigma = 1$. 1000 replications were used. The SAS (1997) software was used for simulation and analysis. Confidence intervals for the mean value θ were calculated according to the methods discussed above, in each sample.

The confidence intervals included are:

- the naïve approach.
- the Cox approach (equation (4), using z as multiplier.
- the modified Cox method with t instead of z as multiplier.
- the generalized confidence intervals. The simulation of the sampling distribution was based on 10000 replications.

- the Large-sample approach, i.e. $\bar{X} \pm z\sqrt{\frac{S_x^2}{n}}$

Each interval was compared to the population mean value θ = 244.69, and the number of intervals below, covering, or above θ was calculated. The results that are summarized in Table 5 give the percentage of the samples that cover θ, and the percentage of the samples that produce intervals above or below θ.

Table 5. Results of the simulation study: percent of all intervals that cover the true parameter value.

	Naïve approach			Cox method			Modified Cox method		
n	Below	Covering	Above	Below	Covering	Above	Below	Covering	Above
5	13.5	86.2	0.3	10.6	87.2	2.2	5.9	93.5	0.6
10	31.3	68.5	0.0	8.2	91.1	0.7	5.9	93.9	0.2
20	54.8	45.2	0.0	4.8	94.2	1.0	3.6	95.7	0.7
30	75.9	24.1	0.0	6.5	92.6	0.9	5.4	93.9	0.7
50	94.3	5.7	0.3	4.0	95.4	0.6	3.9	95.5	0.6
100	99.9	0.1	0.0	3.3	95.5	1.2	3.2	95.7	1.1
200	100.0	0.0	0.0	2.6	95.2	2.2	2.6	95.2	2.2
500	100.0	0.0	0.0	3.0	95.1	1.9	3.0	95.1	1.9
1000	100.0	0.0	0.0	3.3	94.4	2.3	3.3	94.4	2.3

	Large sample approach			Generalized C I		
n	Below	Covering	Above	Below	Covering	Above
5	16.8	83.0	0.2	1.3	94.1	4.6
10	16.4	83.6	0.0	2.2	93.7	4.1
20	12.0	87.9	0.1	1.9	95.2	2.9
30	14.0	85.6	0.4	2.1	94.6	3.3
50	9.4	90.4	0.2	2.2	95.0	2.8
100	7.6	92.1	0.3	2.9	93.7	3.4
200	6.5	92.2	1.3	1.3	95.9	2.8
500	4.9	94.0	1.1	2.8	94.2	3.0
1000	4.8	93.8	1.4	2.3	95.8	1.9

6. Discussion

The results for the Cox intervals are similar to the simulations in Zhou and Gao (1997). However, the coverage percentage is improved, especially in small samples, if

the intervals are based on t rather than on z. For the modified Cox approach, the percentage of intervals which cover θ is close to the nominal level, 95%, for all sample sizes. This also holds for the generalized confidence interval approach. Note that the modified Cox intervals are slightly assymetric with a higher percentage to the left. The generalized confidence intervals are also slightly assymetric, but to the right.

The large-sample method, that is based on Central Limit Theorem arguments, gives a consistently lower coverage than 95%. Sample sizes of more than 200 seem to be needed to obtain a confidence level close to the nominal one. As expected, the intervals based on the naïve approach fail, since these intervals are intervals for some other parameter. The simulations were also run with standard deviations 0.5 and 2. All methods performed somewhat worse when the standard deviation increased but the relationships between methods remained unchanged.

It seems that the confidence intervals based on the modified Cox method work well for practical purposes. The calculations are simple and may be performed by hand, if desired. The generalized confidence interval approach also works well; a small disadvantage is that it requires a computer to simulate the sampling distribution.

References

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.

Land, C. E. (1971), “Confidence intervals for linear functions of the normal mean and variance,” *Annals of Mathematical Statistics*, 42, 1187-1205.

SAS Institute Inc. (1997), *SAS/STAT software: Changes and enhancements through Release 6.12,* Cary, NC: SAS Institute Inc.

Weerahandi, S. (1993), “Generalized confidence intervals”. *Journal of the American Statistical Association*, 88, 899-905.

Zhou, X-H., and Gao, S. (1997), “Confidence intervals for the log-normal mean,” *Statistics in Medicine*, 16, 783-790.

Zhou, X-H., Gao, S., and Hui, S. L. (1997), “Methods for comparing the means of two independent log-normal samples,” *Biometrics*, 53, 1129-1135.

Ulf Olsson
Department of Biometry and Engineering
Swedish University of Agricultural Sciences
Box 7032, S-75007
Uppsala
Sweden
Ulf.Olsson@bt.slu.se

Volume 13 (2005) | Archive | Index | Data Archive | Information Service | Editorial Board | Guidelines for Authors | Guidelines for Data Contributors | Home Page | Contact JSE | ASA Publications