Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 7 (13): 1790-1794, 2007
ISSN 1812-5654

On Efficient Confidence Intervals for the Log-Normal Mean

Peter Chami, Robin Antoine and Ashok Sahai
Department of Mathematics and Computer Science,
The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

Abstract: Data obtained in biomedical research is often skewed. Examples include the incubation period of diseases like HIV/AIDS and the survival times of cancer patients. Such data, especially when they are positive and skewed, is often modeled by the log-normal distribution. If this model holds, then the log transformation produces a normal distribution. We consider the problem of constructing confidence intervals for the mean of the log-normal distribution. Several methods for doing this are known, including at least one estimator that performed better than Cox's method for small sample sizes. We also construct a modified version of Cox's method. Using simulation, we show that, when the sample size exceeds 30, it leads to confidence intervals that have good overall properties and are better than Cox's method. More precisely, the actual coverage probability of our method is closer to the nominal coverage probability than is the case with Cox's method. In addition, the new method is computationally much simpler than other well-known methods.

Key words: Normal population mean, confidence interval, simulation study

INTRODUCTION

Data obtained in biomedical research is often skewed. Examples include the incubation period of diseases like HIV and survival times of cancer patients. Since statistical inference based on the normal distribution is well known, an established way to deal with non-normal data is to apply a transformation that makes them normally distributed. The log transformation is one of those most commonly used for this purpose. It is especially recommended when the data come from a population that is positive and skewed. A variable X is said to have a log-normal distribution with parameters μ and σ^2 if $Y = \log X$ is normally distributed with mean μ and variance σ^2. In this case the mean θ of X is

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

We consider the problem of constructing confidence intervals for the mean θ of the log-normal distribution.

Zhou and GAO (1997) did simulations to construct and compare confidence intervals using the four accepted methods at that time. These are the naive method, Angus's conservative method, Angus's Parametric Bootstrap (PB) method and Cox's method. Their simulation study revealed that the naive method was wholly inappropriate and contrary to what one would expect, produced an increase in coverage area with an increase in sample size. When the sample size was fairly small (n = 11), coverage error was overall the smallest for the (PB) method but this result was only obtained when the variance was small. It was also found that the (PB) method was negatively biased.

Cox's method yielded confidence intervals that had comparatively the smallest coverage error for moderate sample sizes (as small as 50). However, unlike the (PB)'s method, Cox's method provided coverage error values that did not significantly increase as σ^2 increased.

Wu *et al.* (2003) derived a modified signed log-likelihood ratio method that, for small samples of size less than 30, outperformed both Cox's method and the (PB) method for all the comparative criteria by Zhou and Gao (1997).

In the current study, we revisit this classical problem and derive a modified version of Cox's method to provide a more efficient estimator. Unlike all the previous papers mentioned, our approach will deal exclusively with samples of size greater than 30. The proposed estimator is compared with the existing methods via the three criteria used by Zhou and Goa (1997), coverage error, interval width and relative bias. We show that coverage error is smaller than any of the other methods, including the modified signed log-likelihood ratio method of Wu *et al.* (2003), for sample size n>30.

Corresponding Author: Peter Chami, Department of Mathematics and Computer Science, The University of the West Indies, St. Augustine, Trinidad and Tobago, West Indies

THE FIVE MAIN APPROACHES

Here, we review the five existing methods for constructing two-sided 1-α level confidence intervals for a log-normal mean θ. Let $X_1, X_2,...,X_n$ be a random sample from a log-normal distribution with parameters μ and σ^2, let $Y = \log X_i$ for I = 1,2,...n and let

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

be the mean of the log-normal.

The Naïve method: This method constructs a confidence interval for μ, the mean of the log-transformed data, using the normal theory as

$$\overline{Y} \pm Z_{\left(1-\frac{\alpha}{2}\right)}\frac{S}{\sqrt{n}}.$$

Next an antilogarithm function is applied to transform the confidence limits back to the original scale to obtain a confidence interval for

$$\theta : \exp\left(\overline{Y} \pm Z_{\left(1-\frac{\alpha}{2}\right)}\frac{S}{\sqrt{n}}\right).$$

For large n, this method leads to biased estimators.

Cox's method: One way to estimate θ is to estimate μ and σ^2 and then to make use of the relationship

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

If we estimate μ and σ^2 by the sample mean $\overline{Y}$ and the sample variance S^2, respectively of the observations $\overline{Y}$ then we get the point estimator of $\hat{\beta}$ of $\beta = \log\theta$ to be

$$\hat{\beta} = \overline{Y} + \frac{1}{2}S^2$$

Since $(\overline{Y}, S^2)$ is a complete, sufficient statistic for (μ, σ^2) and

$$\hat{\beta} = \overline{Y} + \frac{1}{2}S^2$$

is an unbiased estimator of

$$\log\theta = \mu + \frac{1}{2}S^2,$$

it follows from a well-known theorem (Lehmann, 1983) that $\hat{\beta}$ is an UMVUE of log θ. From the independence of Y and S^2 we get that the variance of $\hat{\beta}$ is

$$\frac{\sigma^2}{n} + \frac{\sigma^4}{2(n-1)}$$

Here we note that, at this point in their discussion of the relevant point estimators, Zhou and Gao (1997) made an error in stating that

$$\frac{S^2}{n} + \frac{S^4}{2(n-1)}$$

is an unbiased estimator of

$$\frac{\sigma^2}{n} + \frac{\sigma^4}{2(n-1)}.$$

Although $E(S^2) = \sigma^2$, S^4 is not an unbiased estimator of σ^4. In fact, because

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

it can be easily shown that

$$E\left(S^4\right) = \frac{(n+1)}{(n-1)}\sigma^4.$$

Thus the correct unbiased estimator of $\mathrm{Var}(\hat{\beta})$ of this form is:

$$\frac{S^2}{n} + \frac{S^4}{2(n+1)}$$

This is also the UMVUE of

$$\frac{\sigma^2}{n} + \frac{\sigma^4}{2(n-1)}.$$

Assuming approximate normality for $\hat{\beta}$, the approximate confidence limits for θ may be obtained in the form $\varphi_\alpha = \exp(\beta + Z_{\alpha\zeta})$.

Angus's conservative method: Angus proposed a conservative method for construction of a confidence interval for lnθ based on the following approximate pivotal statistic:

$$\eta(\theta) = \frac{\left(\overline{Y} + \frac{e^2}{2} - \ln\theta\right)\sqrt{n}}{\sqrt{\left\{S^2\left(1-\frac{S^2}{2}\right)\right\}}} \quad (1)$$

when the sample is finite however, (1) has the same distribution as:

$$T(\sigma) = \frac{N + \sigma\frac{\sqrt{n}}{2}\left(\frac{\chi^2(n-1)}{(n-1)} - 1\right)}{\sqrt{\left\{\frac{\chi^2(n-1)}{n-1}\left(1 + \frac{\sigma^2}{2}\frac{\chi^2(n-1)}{n-1}\right)\right\}}} \quad (2)$$

where, N and χ^2 (n-1) are independent, N is the standard normal and ^{2}(n-1) is a χ^2(n-1) is a χ^2- distribution with n-1 degrees of freedom. This leads (Zhou and Gao, 1997), to the lower and upper limits respectively of the (1-α) confidence interval for ln θ.

A parametric bootstrap method: The bootstrap interval described by Angus applies the parametric t-percentile bootstrap method to the approximate pivotal statistics η(θ) Eq. 1 By letting t_0 and t_1 be the $\frac{\alpha}{2}$ percentile and the $1-\frac{\alpha}{2}$ percentile of η(θ), respectively. Hence a theoretical 1-α level confidence level for lnθ is

$$\ln = \left(\frac{\bar{Y} + \frac{S^2}{2} - t_1}{\sqrt{\left\{ \frac{S^2\left(1+\frac{S^2}{2}\right)}{n} \right\}}}, \bar{Y} + \frac{S^2}{2} - t_0 \sqrt{\left\{ \frac{S^2\left(1+\frac{S^2}{2}\right)}{n} \right\}} \right) \quad (3)$$

The unknown quantiles t_0 and t_1 can be estimated by a parametric bootstrap sample.

Modified signed log-likelihood ratio method: Wu *et al.* (2003) asserted that Cox's method did not perform well in small sample settings due its nonquadratic and asymmetric shape of the likelihood profile for small n. They instead considered the modified signed log-likelihood ratio introduced by Banndroff-Nielsen (1986 and 1991), generally known as the r* formula:

$$r^* = r^*(\psi) + r(\psi) + r(\psi)^{-1} \log\left\{ \frac{u(\psi)}{r(\psi)} \right\} \quad (4)$$

where, u (Ψ) is a quantity and the general form of r* is given in the Appendix. r* being asymptotically distributed as a standard normal variate with third order accuracy. Therefore, an approximate 100 (1-α)% confidence interval based on r* is

$$\left\{ \psi, \left| r^*(\psi) \right| \le Z_{\frac{\alpha}{2}} \right\} \quad (5)$$

where unlike Cox's interval, this r* interval calculates the confidence limit from the observed asymmetric likelihood - based function r* (Ψ) which theoretically should have achieved a more accurate coverage probability than Cox's. The modified signed log-likelihood ratio method produced zero coverage errors and almost negligible average biases and both the coverage probabilities and average biases remained nearly constant as the variance increased.

THE IMPROVED COX'S CONFIDENCE INTERVAL(C I) ESTIMATORS

Firstly, we note a simple fact regarding the concept of estimation. An estimator, say t of the parameter say, θ is examined for its efficiency on the following counts:

Bias (t) = E[t - E(t)] = B (t), say.
Variance (t) = E [t-E (t)]^2 = V (t), say
Mean square Error (t) = MSE (t) = E [t - θ]^2 = M (t), say.
It could easily be checked that:
M (t) = V (t) + B (t).
Let us consider:

$$t^* = \bar{x}\left(1 + s^2/\left(n\bar{x}^2\right)\right) = \bar{x}(1+v) \quad (6)$$

Where, $v = \left(s^2/\left(n\bar{x}^2\right)\right)$

Since $(\bar{x}, s^2)$ a jointly complete sufficient statistic for (230, σ^2), it suffices to find an unbiased estimator of M (t*) = V (t*) + B (t*), as a function of ($\bar{x}$, s^2) alone. This estimator would be a UMVUE. From (5), we have

$$M(t^*)/M(\bar{x}) = R, \text{ say} = \left(\frac{n}{\sigma^2}\right) E\left[(\bar{x}-\mu) + \frac{s^2}{n\bar{x}} \right]^2$$

$$= 1+2A+B$$

Where, $A = (1/\sigma^2) E(s^2(\bar{x}-\mu)/\bar{x})$ and B = $(1/(\sigma^2))$ $E = (s^4/(n\bar{x}))$ Using the results that $\bar{x} \sim N(\mu, \sigma^2/n)$, $(n-1)s^2/\sigma^2 \sim \chi^2(n-1)$ and that $\bar{x}$ and s^2 are independent, we can find unbiased estimators of A and B as follows.

$$A = \left(\frac{1}{\sigma^2}\right) E_{s^2} E_{\bar{x}} \left[\left(\frac{s^2}{\bar{x}}\right)(\bar{x}-\mu) \right]$$

$$= \left(\frac{c}{\sigma^2}\right) E_{s^2} \left[\int_{-\infty}^{+\infty} \left(\frac{s^2}{\bar{x}}\right)(\bar{x}-\mu) \exp\left(\frac{-n(\bar{x}-\mu)^2}{2\sigma^2}\right) d\bar{x} \right],$$

where, $c = \left(\frac{n}{2\pi\sigma^2}\right)^{1/2}$

$$= \left(\frac{-c}{n}\right) E_{s^2} \left[\int_{-\infty}^{+\infty} \left(\frac{s^2}{\bar{x}}\right) \frac{d}{d\bar{x}} \left[\exp\left(\frac{-n(\bar{x}-\mu)^2}{2\sigma^2}\right) \right] d\bar{x} \right]$$

$$= \left(\frac{-c}{n}\right) E_{s^2} \left[\int_{-\infty}^{+\infty} \left(\frac{s^2}{\bar{x}^2}\right) \exp\left(\frac{-n(\bar{x}-\mu)^2}{2\sigma^2}\right) d\bar{x} \right]$$

$$= -\frac{1}{n} E\left(\frac{s^2}{\bar{x}^2}\right) = -E\left(\frac{s^2}{n\bar{x}^2}\right)$$

$$= -E(v)$$

Hence $\hat{A} = -v$ is an unbiased estimator of A. Further,

$$B = \left(\frac{1}{n\sigma^2}\right)E\left(\frac{s^4}{\bar{x}^2}\right)$$

$$= \left(\frac{1}{n\sigma^2}\right)E_{\bar{x}}E_{s^2}\left(\frac{s^4}{\bar{x}^2}\right)$$

$$= \frac{\sigma^2}{n(n-1)^2}E_{\bar{x}}E_{s^2}\left(\left(\frac{(n-1)s^2}{\sigma^2}\right)^2\frac{1}{\bar{x}^2}\right),$$

$$= \frac{\sigma^2}{n(n-1)^2}E_{\bar{x}}\left(\frac{n^2-1}{\bar{x}^2}\right)E\left((n-1)s^2/\sigma^2\right)^2$$

$$= E\left(\chi^2(n-1)\right)^2 = n^2-1$$

$$= \frac{n+1}{n-1}E\left(\frac{s^2}{n\bar{x}^2}\right), E(s^2) = \sigma^2$$

$$= \frac{n+1}{n-1}E(v)$$

Hence $\hat{B}$ = ((n+1)/(n-1)v is an unbiased estimator of B. Let $\hat{R} = 1 - 2\hat{A} + \hat{B}$. Then $\hat{R}$ is an unbiased estimator of R and hence an UMVU estimator of R. We can write $\hat{R}$ as

$$\hat{R} = 1 - 2v + \left(\frac{n+1}{n-1}\right)v$$

$$= 1 - \left(\frac{n-3}{n-1}\right)v \quad (7)$$

Hence,

$$\hat{M}(t^*) = (s^2/n).[1 - (n-3)v/(n-1)] \quad (8)$$

and

$$\hat{B}(t^*) = \left(s^2/(n\bar{x})\right) \quad (9)$$

Therefore,

$$\hat{V}(t^*) = \hat{M}(t^*) + (\hat{B}(t^*))^2 \quad (10)$$

Now, we are ready to propose our improved CIs, as follows:

Let's call the Cox's CI (Lower CI(:CI low) and Upper CI (:CI high) limits), as Estimator 1, i.e.,

Estimator 1 low: $\bar{x} + s^2/2 - Z_{(1-\alpha/2)}*[s_2/n + s^4/2(n-1)]^{1/2}$
Estimator 1 high: $\bar{x} + s^2/2 + Z_{(1-\alpha/2)}*[s_2/n + s^4/2(n-1)]^{1/2}$

As per our proposition the efficient CI's (Lower CI(CI low) and Upper CI(CI high) limits) for the Lognormal mean θ:

Estimator 2 low = $t^* + s^2 - Z_{(1-\alpha/2)}*[\hat{V}(t^*)]^{1/2}$

And

Estimator 2 high = $t^* + Z_{(1-\alpha/2)}*[\hat{V}(t^*)]^{1/2}$

SIMULATION AND CONCLUSIONS

As mentioned in the beginning, we emulate the Simulation Framework of Zhou and Gao (1997) for the good reasons explained in their paper. Adopting the same structure of their simulation study, we also have carried out a simulation study in this section, of which the results are reported in the Tables in the Appendix.

Using 6000 samples (of illustrative sizes of 51, 101, 151, 201 and 301) from the relevant lognormal distribution with illustrative values of σ^2: 1.00, 1.25, 1.50, 1.75 and 2.00 (assuming like in Land (1971), for the sake of simplicity of illustration and without any loss of generality, that the population mean $\mu = -\sigma^2/2$), we have calibrated the characteristics of the CIs : Coverage Probability (Cvg. Prob.), Coverage Error (Cvg. Error), Length of the CI (Length), Proportion/ Probability of cases of the CI not covering the true value of the actual population mean, when CIs are on the left/right of the true value of the population mean(Left/Right Bs., respectively) and hence the Relative Bias (Rel. Bs.).

The results of the simulation study are tabulated in the five tables given in the appendix, which deal with sample sizes of 51, 101, 151, 201 and 301.

Overall, the Estimator 2 performs better than Cox's method, in the sense that the achieved coverage probability is closer to the nominal probability of 0.90 (in other words, the coverage error is smaller).

Appendix: For n = 51

Variance	Estimator	Cvg. Prob	Cvg. Err	Length	Left Bs.	Right Bs.	Rel. Bs.
1	Estimator 1	0.894500	0.005500	0.564413	0.072333	0.033167	0.371248
1	Estimator 2	0.904767	0.004767	0.564796	0.093317	0.001917	0.959748
1.25	Estimator 1	0.895883	0.004117	0.658311	0.072617	0.031500	0.394910
1.25	Estimator 2	0.904767	0.002767	0.658344	0.091150	0.006083	0.874871
1.5	Estimator 1	0.893917	0.006083	0.747801	0.075983	0.030100	0.432522
1.5	Estimator 2	0.987117	0.002883	0.747823	0.092567	0.010317	0.799449
1.75	Estimator 1	0.895567	0.004433	0.838174	0.075233	0.029200	0.440792
1.75	Estimator 2	0.891170	0.001883	0.838194	0.089683	0.012200	0.760510
2	Estimator 1	0.894000	0.006000	0.924802	0.078817	0.027183	0.487107
2	Estimator 2	0.895250	0.004750	0.924818	0.092350	0.012400	0.763246

For n = 101

Variance	Estimator	Cvg. Prob.	Cvg. Br.	Length	Left Bs.	Right Bs.	Rel. Bs.
1	Estimator 1	0.895817	0.004183	0.401299	0.065717	0.038467	0.261558
1	Estimator 2	0.900767	0.000767	0.401302	0.080150	0.019083	0.615385
1.25	Estimator 1	0.897767	0.002233	0.466858	0.067217	0.035017	0.314966
1.25	Estimator 2	0.900750	0.000750	0.466860	0.080033	0.019217	0.612762
1.5	Estimator 1	0.898383	0.001617	0.531154	0.065483	0.036133	0.288831
1.5	Estimator 2	0.900517	0.000517	0.531156	0.077600	0.021883	0.560060
1.75	Estimator 1	0.897517	0.002483	0.594364	0.067933	0.034550	0.325744
1.75	Estimator 2	0.897533	0.002167	0.594365	0.079183	0.022983	0.550082
2	Estimator 1	0.898117	0.001883	0.655450	0.069967	0.031917	0.373466
2	Estimator 2	0.898533	0.001467	0.655451	0.079583	0.021883	0.568660

For n= 151

Variance	Estimator	Cvg. Prob.	Cvg. Err.	Length	Left Bs.	Right Bs.	Rel. Bs.
1	Estimator 1	0.896617	0.003383	0.3280390	0.063350	0.040030	0.225536
1	Estimator 2	0.898217	0.001783	0.3280390	0.076783	0.025000	0.508760
1.25	Estimator 1	0.899867	0.000133	0.3818700	0.062783	0.037350	0.253995
1.25	Estimator 2	0.900083	0.000083	0.3818700	0.074150	0.025567	0.487214
1.5	Estimator 1	0.897733	0.002267	0.4339820	0.065417	0.036850	0.279335
1.5	Estimator 2	0.898483	0.001517	0.4339830	0.075183	0.026333	0.481202
1.75	Estimator 1	0.900400	0.000400	0.4855310	0.063133	0.036467	0.267738
1.75	Estimator 2	0.900533	0.000533	0.4855310	0.071767	0.026700	0.457684
2	Estimator 1	0.899000	0.001000	0.5359874	0.065533	0.035467	0.297690
2	Estimator 2	0.899100	0.000900	0.5359740	0.073850	0.027050	0.463826

For n = 201

Variance	Estimator	Cvg. Prob.	Cvg. Err.	Length	Left Bs.	Right Bs.	Rel. Bs.
1	Estimator 1	0.895875	0.001250	0.284285	0.060717	0.040533	0.199342
1	Estimator 2	0.899750	0.000250	0.284285	0.072183	0.028067	0.440067
1.25	Estimator 1	0.899483	0.000517	0.331042	0.060133	0.040383	0.196485
1.25	Estimator 2	0.900567	0.000567	0.331042	0.069467	0.029967	0.397251
1.5	Estimator 1	0.900667	0.000667	0.376590	0.060967	0.038367	0.227517
1.5	Estimator 2	0.900050	0.000050	0.376359	0.069367	0.040383	0.402055
1.75	Estimator 1	0.900133	0.000133	0.420651	0.062117	0.037750	0.423992
1.75	Estimator 2	0.900000	0.000000	0.420651	0.069333	0.296670	0.400673
2	Estimator 1	0.898983	0.001017	0.464453	0.062900	0.038117	0.245339
2	Estimator 2	0.899317	0.000683	0.464453	0.069983	0.030700	0.390167

For n = 301

Variance	Estimator	Cvg. Prob.	Cvg. Err.	Length	Left Bs.	Right Bs.	Rel. Bs.
1	Estimator 1	0.898133	0.001867	0.232327	0.058417	0.043450	0.146924
1	Estimator 2	0.899217	0.000783	0.232327	0.067567	0.332170	0.340830
1.25	Estimator 1	0.899183	0.000817	0.270332	0.059450	0.041367	0.179368
1.25	Estimator 2	0.899883	0.000117	0.270332	0.067750	0.033367	0.340036
1.5	Estimator 1	0.900883	0.000883	0.307407	0.057467	0.041650	0.159576
1.5	Estimator 2	0.900700	0.000700	0.307407	0.064883	0.034417	0.306814
1.75	Estimator 1	0.899867	0.000133	0.343681	0.060350	0.039783	0.205393
1.75	Estimator 2	0.900117	0.000117	0.343681	0.066400	0.032882	0.337586
2	Estimator 1	0.898933	0.001067	0.379563	0.061100	0.039967	0.209103
2	Estimator 2	0.899783	0.000217	0.379563	0.066250	0.033967	0.322135

REFERENCES

Land, C.E., 1971. Confidence intervals for linear functions of the normal mean and variance. Ann. Math. Stat., 42: 1187-1205.

Lehmann, E.L., 1983. Theory of Point Estimation. John Wiley and Sons, New York.

Wu, J., A.C.M. Wong and G.Y. Jiang, 2003. Likelihood-based confidence interval for log-normal mean. Stat. Med., 22: 1849-1860.

Zhou, X.H. and S. Gao, 1997. Confidence intervals for the log-normal mean. Stats. Med., 16: 783-790.