Comparative Study of Confidence Intervals for Population Median

Submitted by Sagnika Chakraborty and Shwetab Sinha

November 5, 2010

Abstract

Here our objective is to make a comparative study among the three method as (i) Asymptotic Normal Approximation (ii) Bootstrap - t (ii) Bootstrap Percentile Method for determining the confidence interval for **Population Median**. For comparison purpose we have taken random samples of different sizes from Normal, Uniform, Exponential, Lognormal, t_5 and χ^2_5 distributions. Finally we want to comment on the efficiency of the three methods with respect to two aspects **(i) length of confidence intervals (ii) observed coverage probabilities**.

1 Introductoin:

In point estimation, when the random variables are X_1,X_2,...,X_n and θ(may be scalar or vector) is the unknown parameter, we try to estimate a parametric function $\gamma(\theta)$ by means of a single value , say t, the value of a statistic (an estimator) T corresponding to the observed value x_1,x_2,...,$x_n$of the random variables. In Inerval Estimation, consider two limits t_1 and t_2 ($t_1 < t_2$) computed from the set of observation x_1,x_2,...,x_n. It is claimed with a certain degree of confidence(measured in probabilistic terms) that the true value of $\gamma(\theta)$ lies between t_1 and t_2. Ideally a confidence interval should reflect the shape of a distribution, specially when the distribution is skewed.

A random interval $CI(\underline{X}) = [l(\underline{X}), u(\underline{X})]$ is a said be a level $(1-\alpha)$ confidence interval for $\gamma(\theta)$ if $P[\gamma(\theta) \in [l(\underline{X}), u(\underline{X})]] \geq 1-\alpha \; \forall \; \gamma(\theta)$. We can have one-sided confidence intervals as well, where $l(\underline{X}) = -\infty$ or $u(\underline{X}) = \infty$. We need the knowledge of the underlying distributuin for consructing exact confidence interval. If we don't have any underlying assumptions then also we can give confidence interval by some different methods viz asymptotic normal approximation approach and bootstrap apporach. Here, we shall give a description and analysis of those methods considering the parametric function as the population median.

1.1 Basic set-up:

Here, we shall state some basic assumptions and notations. They are as follows:

□ Let X_1,X_2,...,X_n is a random sample with distribution function F_θ.

□ Here our parameter of inerest is $F^{-1}(0.5)$,the population median. For our discussion we have chosen n to be odd, say $n = 2k+1$.

□ Sample median based on X_1,X_2,...,X_n is defined as $X_{(k+1)}$ where $X_{(.)}$ denotes the ordered observations.

☐ Here our objective is to find confidence interval for $F^{-1}(0.5)$ using three different approach as (i) The Asymptotic Normal Approximation (ii) Bootstrap – t and (iii) Bootstrap Percentile method.

☐ We vary the sample size, n as 11, 21, 51, 101, 501, 1001.

☐ We have taken the level $(1 - \alpha)$ as 95%.

1.2 Comparing tools:

Our motive is to compare the three methods mentioned above by means of **(i) Length of the Confidence Inerval** and **(ii) Observed Coverage Probility**.

We have already defined what a confidence inerval is. By our notation, if $[l(\underline{X}), u(\underline{X})]$ be the confidence interval for $\gamma(\theta)$, then the **length of confidence inerval** for $\gamma(\theta)$ can trivially be defined as the simple differnece between upper confidence limit and lower confidence limit i.e. $u(\underline{X}) - l(\underline{X})$.

The probability that the confidence interval contains the true parameter value (here, population median) is called the Coverage Probability. Here we want to estimate the Coverage Probability by the Observed Coverage Probability. For a fixed sample size and a fixed distribution simulate the confidence limits for a large number of times. Then compute the proportion of how many times the population median falls within the confidence limits. This gives the **Observed Coverage Probability**.

2 Different Methods:

Since we don't have any knowledge about the parent distribution of the sample observations. We use the following three methods:

2.1 Asymptotic Normal Approximation:

Let ξ_p and x_p be the p-th population and sample quantile respectively. We can define ξ_p as $\xi_p = inf\{x|F(x) \geq p\}$, $0 < p < 1$. We are interested in $p = 1/2$. We have the folowing asymptotic result as

$\sqrt{n}(x_{1/2} - \xi_{1/2}) \xrightarrow{d} N(0, 1/(4f^2(\xi_{1/2}))$.

As we don't have any idea of f, we need to estimate the asymtotic variance, for which go for bootstraping. The way we have computed the estiamte of the asymptotic variance is explained in section 2.2. Take that estiamte be v. So, the confidence interval obtained in this way is as follows

$[x_{1/2} - \sqrt{v}\tau(\alpha/2), x_{1/2} + \sqrt{v}\tau(\alpha/2)]$ where, $\tau(\alpha/2)$ is the upper $\alpha/2$ point of a standard normal distribution.

We have different choices of n and different distributions. From each distribution we generate a random sample of size n and based on that we obtain a confidence interval and the correspondig confidence length.

2.2 Bootstrap-t:

Let $X_1, X_2, ..., X_n \sim$ iid F. $X_1^*, X_2^*, ..., X_n^*$ is an SRSWR from X_1, X_2,..., X_n. Then, X_1^*, X_2^*,..., X_n^* F_n, F_n being the empirical cumulative distribution function(ECDF) of X_1, X_2,..., X_n. By Bootstrap principle we can make inferences about the sampling distribution of $T(X_1, X_2, ..., X_n)$ by studying the sampling distribution of $T(X_1^*, X_2^*, ..., X_n^*)$.

Construct the pivotal quantities $z_i^* = \frac{(T(\underline{X}_i^*)-T(\underline{X}_i))}{\sqrt{V(T(\underline{X}_i^*))}}$ $i = 1(1)B$.

Where, $T(\underline{X}) = X_{(k+1)}$; $V(T(\underline{X}_i*)) = \sum_{j=1}^{n} X_{i(j)}^2 p_j - \left(\sum_{j=1}^{n} X_{i(j)} p_j\right)^2$; $p_j = P(X_{(k+1)}^* = X_{(j)}) = \sum_{t=0}^{k}[b(t|n, \frac{j-1}{n}) - b(t|n, \frac{j}{n})]$; $b(x|n,p) = ^n C_x p^x (1-p)^{n-x}$

Now, take $Z_1^*, Z_2^*, ..., Z_n^*$ and obtain their ECDF G_{Z^*}. The Bootstrap-t confidence interval is then given by

$CI(\underline{X}) = [T(\underline{X}) - G_{Z^*}^{-1}(\alpha/2)\sqrt{V^*(T(\underline{X}))}, T(\underline{X}) - G_{Z^*}^{-1}(1-\alpha/2)\sqrt{V^*(T(\underline{X}))}]$

2.3 Bootstrap Percentile:

Here we work directly on the distribution of $T(\underline{X}) = x_{1/2} = X_{(k+1)}$. For each of the B bootstrap samples $X_{i1}^*, X_{i2}^*, ..., X_{in}^*$ generated from the original random sample. We compute $T_{ni}^* = T(X_{i1}^*, X_{i2}^*, ..., X_{in}^*) = T(\underline{X_i}^*)$ and rank them; $i = 1(1)B$. The $\alpha/2$-th percentile $(1-\alpha/2)th$ percentile of the bootstrap statistics determine the lower and upper confidence limits with a $100(1-\alpha)$ % confidence coefficient.

3 Analysis:

We have discussed so far the three stated methods to compute the confidence interval and hence confidence length. Once we have obtained the confidence interval we can obtain the observed coverage probility(methods already mentioned in the section 1.2). Here shall give our compuational results, plots and analysis. For Normal Approximation and Bootstrap-t we have taken $B = 2500$ and for bootstrap Percentile we have taken $B = 500$.

3.1 Compuation:

The following tables show the **length of confidence interval** for different distributions and different sample sizes

Normal Approximation Method

Distribution	Normal	Uniform	Exponential	Lognormal	t_5	χ_5^2
$n = 11$	1.6572838	0.60314926	1.3250820	3.6476840	1.9567719	4.5090940
$n = 21$	1.2017001	0.5231143	0.4703594	1.8104927	1.5640626	2.2686832
$n = 51$	0.7631804	0.3394571	0.4281044	0.4651753	0.7753996	1.7857013
$n = 101$	0.4716088	0.19694416	0.425705	0.6489667	0.6562063	1.7396733
$n = 501$	0.2569842	0.0799056	0.1591069	0.2027807	0.2085941	0.6366813
$n = 1001$	0.2058368	0.0715141	0.1156479	0.1348823	0.1756398	0.4960418

Bootstrap - t

Distribution	Normal	Uniform	Exponential	Lognormal	t_5	χ_5^2
$n = 11$	1.32184488	0.54265885	1.5030063	1.6729722	0.909526	5.349940
$n = 21$	2.7717161	0.4934628	1.936393	1.0499008	1.8746418	3.985369
$n = 51$	0.684698	0.52483492	0.3782448	0.7627589	0.7104776	2.2178439
$n = 101$	0.545602	0.19701634	0.2304753	0.5148220	0.6158574	1.2999361
$n = 501$	0.2102102	0.1112084	0.2304573	0.2175360	0.2881911	0.4851781
$n = 1001$	0.2137893	0.06362625	0.1666144	0.1511102	0.1235170	0.595531

Bootstrap Percentile

Distribution	Normal	Uniform	Exponential	Lognormal	t_5	χ^2_5
$n = 11$	1.5986627	0.63288102	1.0825161	2.8478768	1.0890002	3.9250231
$n = 21$	0.8261523	0.39771752	0.7934934	0.8511795	0.7414386	3.1779971
$n = 51$	0.6766108	0.27574842	0.4955611	0.7170813	0.4735087	2.0602902
$n = 101$	0.4563483	0.21783961	0.4568604	0.4607078	0.3160581	1.3519523
$n = 501$	0.1922252	0.06524238	0.1377893	0.1832444	0.1770720	0.6207911
$n = 1001$	0.1628343	0.04779884	0.1075508	0.1526250	0.1510305	0.4569633

It is important to note that Bootstrap-t method gives better result(shorter length) than Percentile Method for symmetric distribution and Percentile Method shows better than Bootstrap-t Method for Asymmetric distribution.

The following tables show the **observed coverage probability** for different distributions and different sample sizes

Normal Approximation Method

Distribution	Normal	Uniform	Exponential	Lognormal	t_5	χ^2_5
$n = 11$	0.926	0.917	0.936	0.949	0.942	0.93
$n = 21$	0.931	0.938	0.93	0.946	0.947	0.941
$n = 51$	0.934	0.937	0.933	0.943	0.939	0.934
$n = 101$	0.943	0.933	0.941	0.947	0.944	0.933
$n = 501$	0.944	0.942	0.942	0.951	0.947	0.939
$n = 1001$	0.952	0.96	0.951	0.953	0.967	0.95

Bootstrap-t Method

Distribution	Normal	Uniform	Exponential	Lognormal	t_5	χ^2_5
$n = 11$	0.927	0.908	0.926	0.899	0.948	0.938
$n = 21$	0.936	0.915	0.932	0.907	0.946	0.924
$n = 51$	0.942	0.936	0.931	0.918	0.95	0.937
$n = 101$	0.951	0.937	0.934	0.935	0.953	0.935
$n = 501$	0.947	0.929	0.942	0.931	0.951	0.938
$n = 1001$	0.956	0.942	0.946	0.949	0.954	0.947

Bootstrap Percentile Method

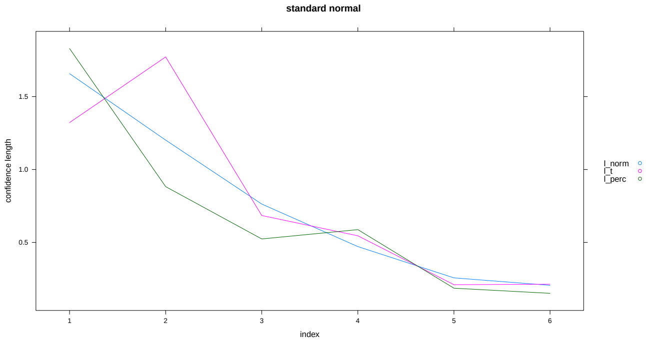
Distribution	Normal	Uniform	Exponential	Lognormal	t_5	χ^2_5
$n = 11$	0.932	0.935	0.926	0.927	0.935	0.922
$n = 21$	0.944	0.932	0.933	0.935	0.933	0.942
$n = 51$	0.944	0.949	0.942	0.937	0.934	0.946
$n = 101$	0.952	0.941	0.945	0.932	0.95	0.943
$n = 501$	0.952	0.953	0.945	0.945	0.946	0.939
$n = 1001$	0.953	0.962	0.947	0.947	0.945	0.948

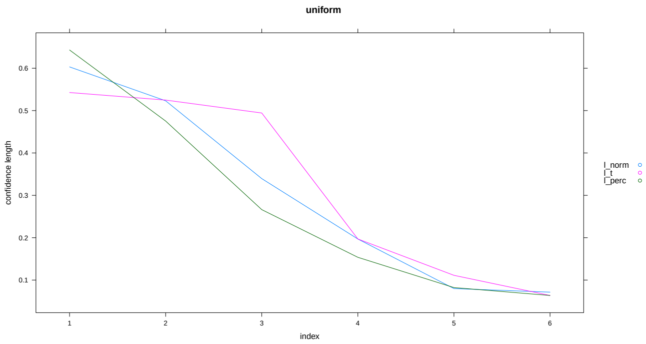
3.2 Comaprison:

We compare the three methods on the basis for different distributions on the basis of (i) length of confidence interval and (ii) observed covareage probability. For comaparison purpose we have taken the bootstrap sample size, $B = 2500$ for all cases.

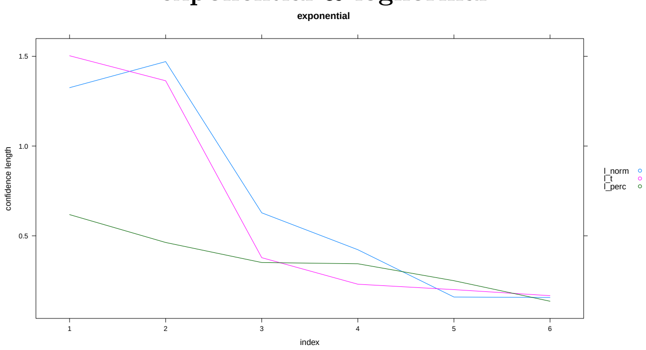
The following plots show the length of confidence interval for population median for a given distribution for different sample sizes and different methods.

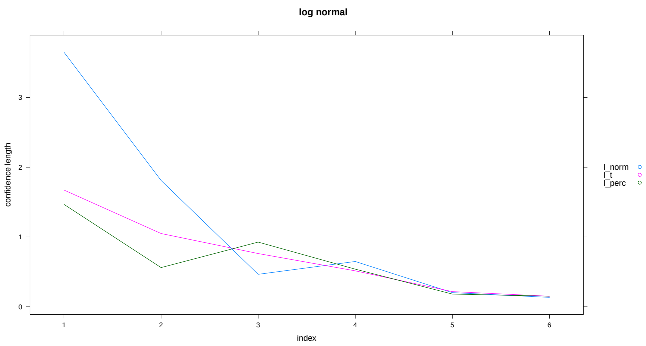
normal & uniform



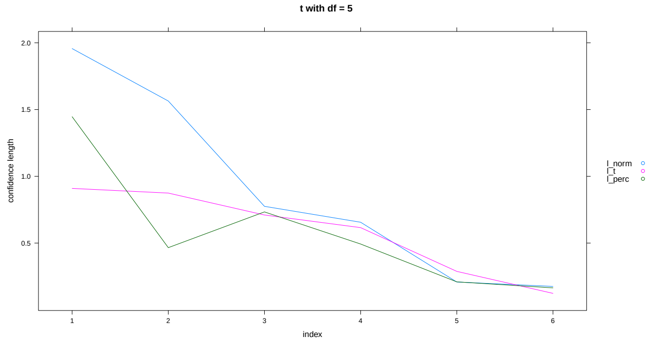


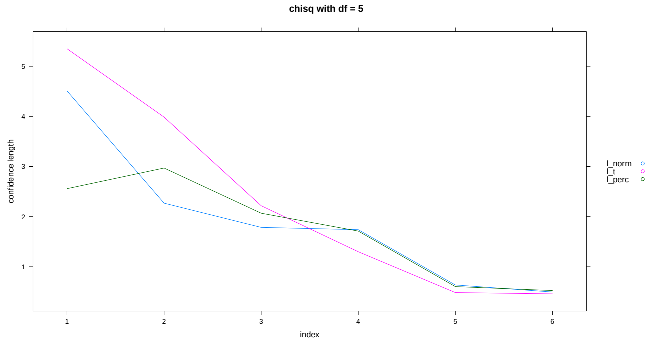
exponential & lognormal





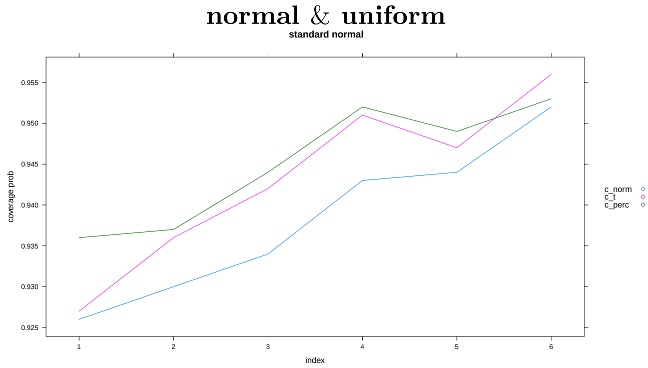
t_5 & χ^2_5

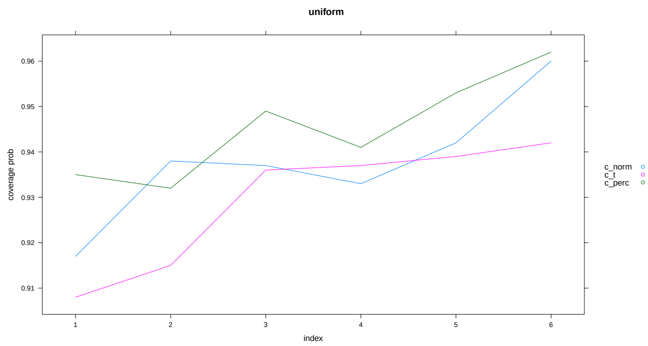




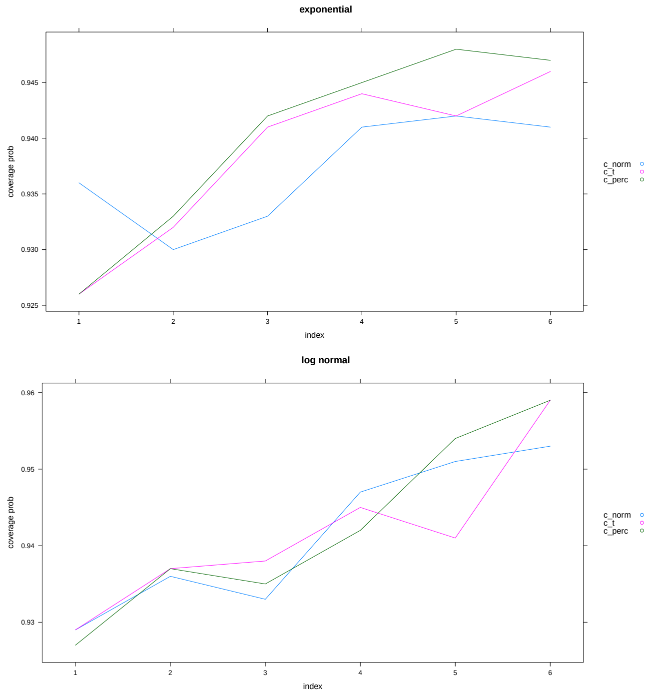
The following plots show the observed coverage probabilities for population median for a given distribution for different sample sizes and different methods.

normal & uniform

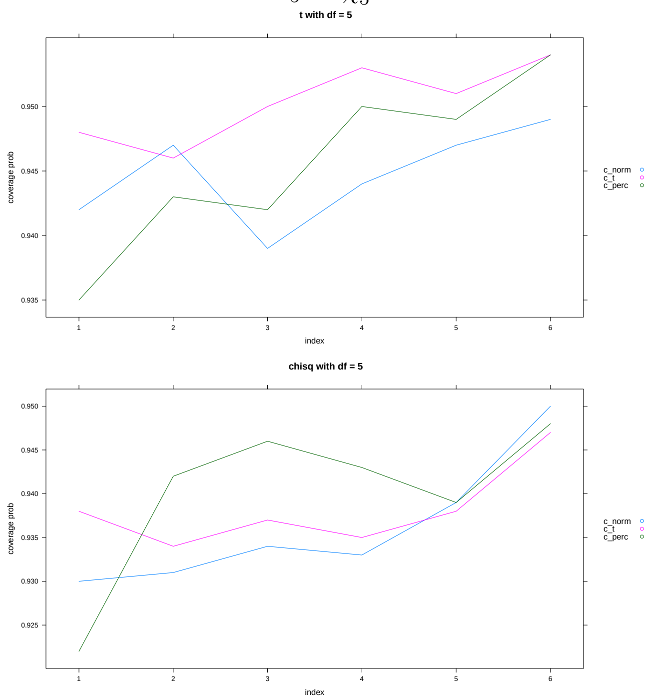




exponential & lognormal



t_5 & χ^2_5



4 Conclusion:

From the computed tables and plots it can be concluded that bootstrapping give significantly better result than normal approximation method for small sample sizes. However, in general, bootstrapping gives satisfactory result. It is to be noted that even in case of normal approximation as we don't have any knowledge about the underlying distribution we have to go for bootstrapping to estimate the variance.

5 Acknowledgement:

☐ Dr. Deepayan Sarkar
☐ Statistical Computing by Kundu & Das