Vol.11.Issue.4.2023 (Oct-Dec)

http://www.bomsr.com

Email:editorbomsr@gmail.com

RESEARCH ARTICLE

BULLETIN OF MATHEMATICS AND STATISTICS RESEARCH

A Peer Reviewed International Research Journal

INTERNATIONAL STANDARD SERIAL NUMBER 2348-0580

COVERAGE PROPORTION, BALANCE AND WIDTH OF NONPARAMETRIC BOOTSTRAP CONFIDENCE INTERVALS: SIMULATIONS FOR SOCIAL SCIENCE DATA

SACHA VARIN

Professor, Department of Mathematics and Statistics, Collège Villamont, Lausanne, Switzerland

Email: varinsacha@yahoo.fr

DOI:10.33329/bomsr.11.4.21

ABSTRACT

Social science studies are very often based on small sample sizes with skewed distributions. We simulate four different distributions (Gaussian, Weibull, Lognormal and Gamma) and four sample sizes (n=5, n=20, n=100 and n=300) to estimate the population mean and median, two location measures widely used in social sciences. Specifically, we take into consideration the coverage probability, the balance and the width of four nonparametric confidence intervals. The results show that in Gaussian populations the Normal bootstrap confidence interval performs well as expected. In Weibull and Lognormal distributions, the Studentized and BCa intervals provide the best results. In the very skewed Gamma distribution, the Studentized and percentile confidence intervals give the best results for the mean. For the median the Normal confidence interval often provides higher coverage, but at the cost of reduced precision.

Keywords: coverage probability; balance; width/length; nonparametric bootstrap confidence intervals

1 Introduction

Social science data are very often skewed, asymmetric and sometimes even heavy-tailed. Moreover, the sample sizes are very often small (n<100 or even n<30). Therefore the Normal distribution often does not reflect "real world" data in social sciences (Nair et al., 2022). The focus of

our investigation is on estimating two statistics of interest in social sciences - the mean and the median - of several distributions. For sample sizes as small as 20 the mean may be approximately Normal since the Central Limit Theorem (CLT) can often be surprisingly accurate. However, the CLT is based on assumptions that may not hold for any arbitrary process. When these assumptions are violated, there is no guarantee that the distribution of sample averages (or total sums) is Normal. Moreover, even when the assumptions may be satisfied, there are very specialized counter-examples where application of the CLT may require thousands of observations. Finally, the CLT for the median can hold in some specialized situations but not generally. In its most general form, the sample average is a unique type of statistic. Of course the required sample size *n* will depend on the shape of the underlying distribution, and skewed distributions generally require larger *n*.

So for analyzing datasets in the social sciences, we use skewed, asymmetric, and heavy-tailed datasets in this paper, but the Normal distribution is included as a reference point. Specifically, the Normal, Weibull, Gamma and Lognormal distributions are replicated 2,000 times to compute the four most popular nonparametric bootstrap confidence intervals in practice: the Normal, the percentile, the bias corrected and accelerated (BCa) and the Studentized confidence intervals (Efron and Tibshirani, 1993; Chernick, 2011) around the mean and around the median, for several sample sizes (*n*=5, *n*=20, *n*=100 and *n*=300). Using 2,000 replications is a trade-off between having too few (high variance) simulations and the time constraint since bootstrap resampling is a slow process.

We note that the bootstrap method basically assumes that the sample is mimicking the population. Our goal is to compare the performance of these four nonparametric bootstrap confidence intervals based on three key measures: the coverage proportion, the balance and the width of the mean and the median of these four data distributions for different sample sizes. We want to test if the intervals are close in coverage to the nominal level 1-α.

We first present the theoretical aspects in Section 2. In Section 3, we review the evaluation criteria. In section 4, we describe the simulation design. We present the results of our analysis in Section 5, discuss the main findings in Section 6 and conclude with limitations and future work.

2 Theoretical aspects

2.1 Normal distribution

The normal distribution, also known as the Gaussian distribution, is well-known in statistics. Most people recognize its familiar bell-shaped curve. The normal distribution is symmetric around its mean, most of the observations cluster around the central peak, and the probabilities for values further away from the mean decrease exponentially in both directions. Extreme values in both tails of the distribution are similarly unlikely. While the normal distribution is symmetric, not all symmetric distributions are normal. For example, the Student's t, Cauchy, and logistic distributions are symmetric.

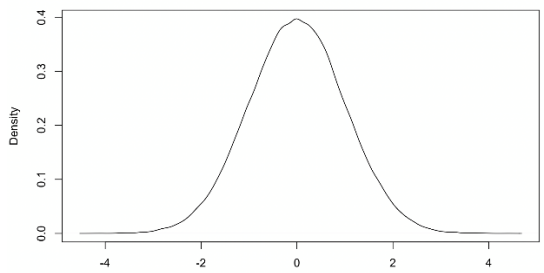

Figure 1: Normal/Gaussian distribution with mean=0 and variance=1

2.2 Skewed distributions

A skewed distribution occurs when one tail is longer than the other. Skewness defines the asymmetry of a distribution. These distributions are encountered in many subject areas when natural limits skew the results away from a boundary and when the data are disproportionally distributed. Specifically, the majority of the data are clustered in one area, and there are one or more more extreme observations – a common situation in social sciences datasets –. Weibull and Lognormal distributions are the most popular distributions for modeling skewed data, and the Gamma distribution is a generalization of the exponential distribution.

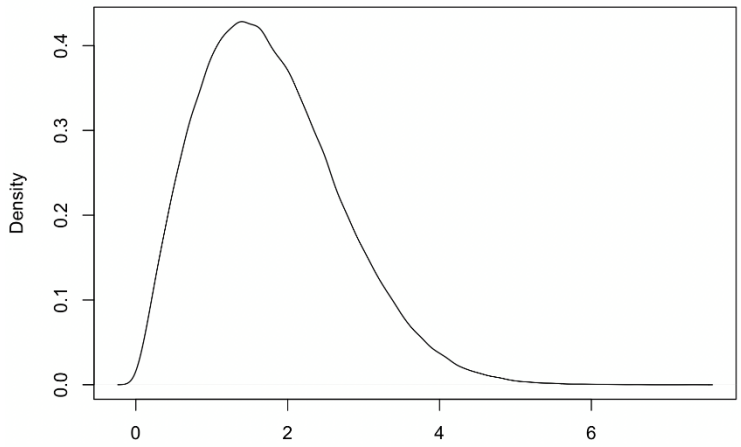

Figure 2: Weibull distributions with shape=2 and scale=2

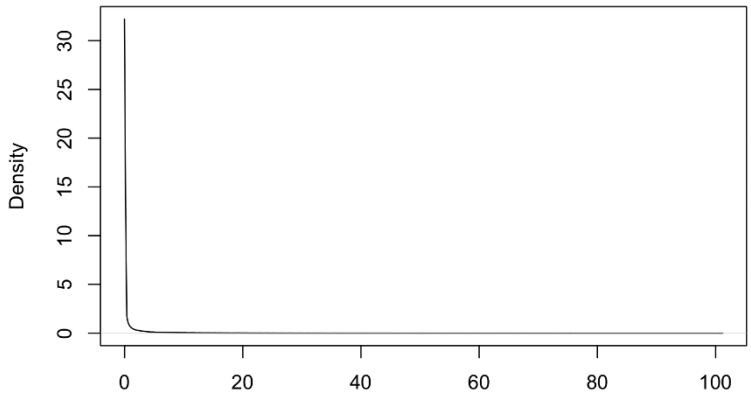

Figure 3: Gamma distributions with shape=1/16 and rate=1/16

2.3 Heavy tailed distributions

A distribution with excess kurtosis is sometimes said to be heavy-tailed compared to the Normal distribution; a familiar example is the t-distribution with low degrees of freedom. In economics and finance, we consider distributions as heavy-tailed if their tails are heavier than exponential distributions, such as those with Pareto (power-law) tails. In our opinion, the most widely accepted definition of a heavy-tailed distribution is one that is heavier than the exponential distribution, a perspective supported by numerous recent literature references (Merz et al., 2022; Bryson already, 1974).

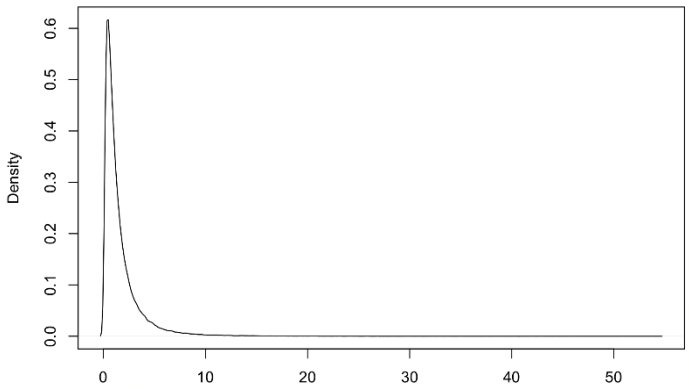

Figure 4: Lognormal distribution with log mean=0 and log standard deviation=1

2.4 Alpha and beta parameters

For the Weibull distribution, the parameter beta is called the shape parameter, while alpha is known as the scale parameter. For the Gamma distribution, the shape parameter specifies the number of events we are modeling. The three distributions considered in this article are skewed and the Lognormal is heavy-tailed. Indeed, using exponential tails as a cutoff, Weibull is heavy-tailed only if beta < 1 which is not the case as beta=2. Moreover, Gamma is heavy-tailed only if alpha > 1 which is not the case as 0 < alpha < 1 in this article.

2.5 Skewness and kurtosis

According to a recent paper (l'Ecuyer, Nakayama, Owen and Tuffin, 2023), skewness and kurtosis are two main factors that affect inference. For high kurtosis the coverage is near the nominal level or even too high, but confidence interval widths can be quite long and variable. When there is high skewness the bootstrap methods should perform better than the Normal and Student's t intervals (l'Ecuyer, Nakayama, Owen and Tuffin, 2023).

The Gaussian distribution has a skewness of 0 and a kurtosis of 3. The Weibull has a skewness of 0.63 and a kurtosis of 3.25. The Lognormal has a skewness of 5.51 and a kurtosis of 68.32 and finally the Gamma has a skewness of 8.38 and a kurtosis of 112.05.

In our opinion, Lognormal is not heavy-tailed enough to cause serious difficulties with inference for the mean and the median via bootstrap. The sort of kurtosis that would be of concern certainly does not include the Lognormal although it can sometimes be so heavy-tailed that the population mean nearly always exceeds all of our sample, which can make bootstrap inference difficult (Mitchell, 1968; Dufresne, 2008). In our opinion, Pareto, Cauchy and other extreme t distributions, e.g. Student *t* with 2 df are in the highly problematic category for inference.

2.6 The two statistics of interests

The two statistics of interests are the mean and the median. In social sciences, these central tendency statistics are the most popular. The preferred measure of central tendency often depends on the shape of the distribution. The mean is most heavily influenced by extreme values, skewness and heavy-tailedness. The mean is preferred when the distribution is symmetric; for more skewed distributions, the median is preferred because it is less sensitive to extreme values. The mean is pulled in the direction of the skewness (i.e., the direction of the tail). In our simulations we consider the Gaussian and more skewed and heavy-tailed distributions, so the median should be preferred but we

take into consideration the mean as well to see how it and its confidence intervals behave with skewed distributions.

In Figure 5 representing a skewed distribution, we see that the mean (blue) is rather far away from the bulk of the data and the median (red) is closer to the "typical" value.

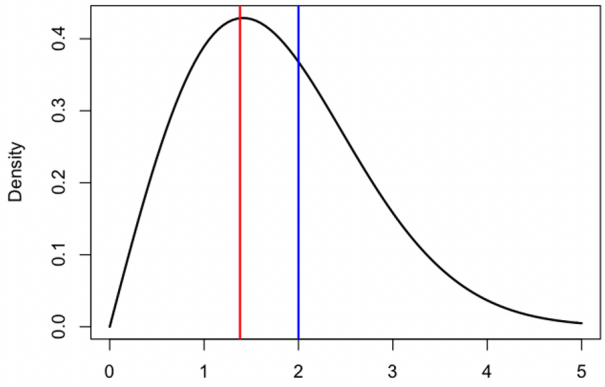

Figure 5: The mean in blue and the median in red in Weibull distribution with shape=2 and scale=2

In statistics, we are using a sample of data to estimate an interval for the parameter value of interest. Under the frequentist framework, different confidence intervals are proposed. Also, there is a common form of interval estimation, such as fiducial intervals, tolerance intervals and prediction intervals. Confidence intervals can be applied in many situations in parametric and nonparametric frameworks (Algarni, Almarashi, Kundu, Abd-Elmougod, Abdel-Khalek, 2022).

2.7 Different confidence intervals

Quantifying uncertainty is a fundamental task for the statistician. We estimate a real-valued parameter θ with some point estimator $\hat{\theta}$ that is based on observed sample data. For the frequentist inference, confidence intervals help to make inferences about hypothesized values of θ. The idea is to use the sample data to generate an interval $I = (\hat{\theta}[\alpha], \hat{\theta}[1-\alpha])$ with the hope that over repeated experiments our intervals will contain the true θ with probability $(1-2\alpha)$, for some pre-chosen α, such as 0.05. A confidence interval *I* then informs us how well we can estimate θ. Confidence intervals are random quantities, varying from sample to sample. Sometimes these random intervals cover the true population parameter and sometimes they do not. Narrow or wide confidence intervals respectively indicate a kind of high or low precision in our estimate of θ. We will speak longer about the precision later (point 3.3).

The gold standard method is the exact confidence interval, which is based on the exact sampling distribution for an estimator $\hat{\theta}$. Unfortunately sampling distributions can be extremely difficult to compute, or we are not willing to make the distributional assumptions that allow their computation. When we lack the means to construct exact confidence intervals, we turn to approximate confidence interval techniques. For many commonly used estimators, we rely on asymptotic results or on the bootstrap to generate approximate confidence intervals.

2.7.1 Exact confidence intervals

It is conceptually straightforward – though the computation might be non-trivial – to construct exact confidence intervals for any parameter (mean and median in our case) and for the distributions of interest in this paper (Martin, 2015). The intervals do not have simple formulas like "estimate

plus/minus margin of error"; the solution will often require numerical methods (some combination of optimization, Monte Carlo, and root-finding).

For any function of the two parameters, the minimum and maximum of the function over the confidence define the confidence interval for the function. This methodology would work for any parametric distribution (Hayter and Kiatsupaibul, 2014). Moreover, it is the best 95% confidence interval because it is strictly based on statistical theory. We will not construct exact confidence intervals here because our goal is to compare the performance of the nonparametric bootstrap confidence intervals in terms of coverage probability, balance and interval width.

2.7.2 Asymptotic confidence intervals

The limiting properties of maximum likelihood estimators (MLEs) are used to obtain the asymptotic distribution of $\hat{\theta}$. Then the asymptotic distribution of $\hat{\theta}$ is asymptotically normally distributed under certain regularity conditions. For example, these conditions are often not met for the MLEs of the extreme-value distributions. It is valid for very large sample size (hence the alternate name is "large-sample" confidence interval).

2.7.3 Bootstrap confidence intervals

Efron (1979) introduces a method for estimating the sampling distribution of an estimator, called the bootstrap distribution. The idea is that by repeatedly sampling our data with replacement and recomputing our estimator, we can approximate the sampling distribution of $\hat{\theta}$. Bootstrapping is a statistical method for inference about a population using sample data. Efron introduced three different methods for generating confidence intervals based on bootstrap distribution: the percentile method, the bias-corrected (*BC*) method, and the bootstrap-T method also called the studentized interval (Efron, 1981).

The *BC* and bootstrap-T method use extra information to correct for bias or skewness in the distribution of $\hat{\theta}$. However, these methods could still perform poorly for small sample sizes. In an attempt to produce a procedure that generates high quality intervals with minimal tedium, Efron (1987) introduced the *bias corrected and accelerated* (BCa) method. The BCa automatically incorporates bias correction and non-constant variance correction to produce high quality intervals (Imholte, 2012). We emphasize that the fundamental assumption required for a valid bootstrap is that the replicates tend to the limiting distribution of the original sample. Sometimes we must use an m<n bootstrap where m/n goes to zero as both m (the size of each resample) and n (the size of the original sample) tend to infinity.

2.8 Parametric versus nonparametric bootstrap confidence intervals

If we know the population distribution for the statistic or underlying random variables, then we can make a 95% parametric bootstrap confidence interval. In practice we need to be reasonably certain about the distribution and the possibility of estimating its parameters. Then we use those estimates to draw replicate samples.

The main difference between the parametric and the nonparametric bootstrap is that the former generates its samples from the assumed distribution of the data or test statistic using the estimated parameter values, whereas the latter generates its replicates by sampling with replacement from the observed data. The nonparametric bootstrap makes fewer assumptions and provides less information than the parametric bootstrap. It is less precise when the parametric assumptions are true but more accurate otherwise. The parametric bootstrap confidence interval is narrower than the

nonparametric bootstrap confidence interval because it is based on the additional information about the distribution of the population.

When samples come from unknown skewed distribution families, a nonparametric confidence can estimate the confidence interval empirically.

There are several ways to compute boostrap confidence intervals. Which confidence interval to use depends in practice on the characteristics of the distribution of the bootstrap estimates.

We thus use the sample distribution as an estimate of the population distribution. In this paper, we mainly use known skewed distributions so the parametric bootstrap confidence intervals is feasible. Usually, however, the distributions are unknown; they are merely approximations of Gaussian, of Weibull, of Gamma, etc. That is of course the motive for nonparametric bootstrap confidence intervals; we want to see the performance of these intervals with known skewed and long-tailed distributions. We focus on four methods that are the most popular in practice: the percentile, the bias corrected and accelerated (BCa), the Studentized and the Normal confidence intervals (Efron and Tibshirani, 1993; Chernick, 2011).

2.8.1 The percentile bootstrap

The percentile intervals naturally extend of the idea of using the bootstrap distribution to estimate the sampling distribution. The percentile bootstrap interval is just the interval between the $100 \times (\frac{\alpha}{2})$ and $100 \times (1 - \frac{\alpha}{2})$ percentiles of the distribution of θ estimates obtained from resampling, where θ represents a parameter of interest and α is the level of significance (e.g., =.05 for 95% confidence intervals) (Efron, 1982). A bootstrap percentile confidence interval of $\hat{\theta}$ (an estimator of θ) can be obtained as follows: (1) *B* random bootstrap samples are generated, (2) a parameter estimate is calculated from each bootstrap sample, (3) the *B* estimates are ordered from the lowest to highest, and (4) the confidence interval is then

$\left[\hat{\theta}_{lower\,\lim}, \hat{\theta}_{upper\,\lim}\right] = \left[\hat{\theta}_j^*, \hat{\theta}_k^*\right]$, where $\hat{\theta}_j^*$ denotes the *j*th quantile (lower limit), and $\hat{\theta}_k^*$ denotes the *k*th quantile (upper limit) ; $j = \left[\frac{\alpha}{2} \times B\right]$, $k = \left[(1 - \frac{\alpha}{2}) \times B\right]$.

For example, a 95% percentile bootstrap confidence interval with 10,000 bootstrap samples is the interval between the 25th quantile value and the 975th quantile value of the 10,000 bootstrap parameter estimates.

2.8.2 The nonparametric bias corrected and accelerated (BCa) bootstrap

The bias corrected and accelerated (BCa) method implicitly incorporates bias correction, corrections for deviation from non-normality, and variance stabilization. The acceleration deals with skew. Efron named the confidence interval the BCa interval because it corrects for both bias and "acceleration" of the variance (Efron, 1987; Efron and Tibshirani, 1993).

The BCa interval is second-order correct, meaning that although none of the bootstrap intervals are guaranteed to have exact 95% coverage probability, the BCa coverage gets close to 95% with smaller sample sizes than any of the first-order correct methods such as the normal, basic, and percentile intervals. So, if we are concerned about coverage probability due to small sample size and asymetric distributions, the BCa is preferred. BCa intervals require the estimation of a bias term and an acceleration term.

The bias-correction factor $\hat{z}_0$ is estimated as the proportion of the bootstrap estimates less than the original parameter estimate,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\left\{\hat{\theta}^* < \hat{\theta}\right\}}{\mathrm{B}}\right),$$

where Φ^{-1} is the inverse function of a standard normal cumulative distribution function (e.g., Φ^{-1} (.975)=1.96) and # is the counting operator. The acceleration factor â is estimated through jackknife resampling (i.e., leave-one-out resampling), which involves generating *n* replicates of the original sample, where *n* is the number of observations in the sample. The average of these estimates is,

$$\hat{\theta}_{(.)} = \sum_{i=1}^{n} \frac{\hat{\theta}_{(-i)}}{n}$$

Then, the acceleration factor â is estimated as follows,

$$\hat{a} = \frac{\sum_{i=1}^{n} (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^3}{6\left\{\sum_{i=1}^{n} (\hat{\theta}_{(.)} - \hat{\theta}_{(i)})^2\right\}^{3/2}}.$$

where $\hat{\theta}_{(.)}$ is the mean of the bootstrap estimates and $\hat{\theta}_{(i)}$ the estimate after deleting the *i*th case. Armed with the values of $\hat{z}_0$ and â, we now estimate the quantiles α_1 and α_2 we will use for establishing the confidence limits,

$$\alpha_1 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right\},$$

$$\alpha_2 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right\},$$

where α is the Type-I error rate, usually .05. The confidence limits are $95\%\ \mathrm{CI} = \left[\theta^{*(\alpha_1)}, \theta^{*(\alpha_2)}\right]$. Based on this formula, it should be obvious that α_1 and α_2 reduces to the percentile intervals when the bias and acceleration terms are zero. The effect of the bias and acceleration corrections is to change the percentiles we use to establish our limits. The BCa interval is also transformation-invariant and range-preserving, meaning that, if the statistic only falls in a certain range [*a,b*], then the BCa interval is not going to fall outside that range [*a,b*].

2.8.3 The Studentized bootstrap

The Studentized bootstrap confidence interval – also called "bootstrap-*t*" by DiCiccio and Efron (1996) – assumes that $\frac{\hat{\theta} - \theta}{s\hat{e}}$ is approximately *t*-distributed, where θ is the true parameter value, $\hat{\theta}$ is an estimate of the true parameter, and $s\hat{e}$ is the sample estimate of the standard error. The bootstrap Studentized confidence interval for a certain α level (e.g., =.05 for 95% confidence intervals) is constructed as follows**:**

$$\left[\hat{\theta} - t_{n-1}^{(\frac{\alpha}{2})} \cdot s\hat{e}, \hat{\theta} + t_{n-1}^{(1-\frac{\alpha}{2})} \cdot s\hat{e}\right].$$

The bootstrap standard error of each estimate, $s\hat{e}(\hat{\theta}^*)$, is used for $s\hat{e}$. We remark that the Studentized interval is second-order correct like BCa, although it requires more computation. The bootstrap-*t* intervals are neither transformation-invariant nor range-preserving.

2.8.4 The Normal bootstrap

The Normal bootstrap uses the standard formula for calculating a confidence interval with the normal distribution value (e.g., 1.96) as the multiplier of the standard error. However, there are two differences. First, we use the bootstrap estimate of the standard error in the formula. Second, there is an adjustment for the estimated bias, -0.005. Accordingly, the estimate of a 95% nonparametric Normal confidence interval around a statistic θ is

95%*CI*= θ - *bias* $\pm$ 1.96$\times$*SE*$_{\theta}$.

The bootstrap Normal approximation and percentile 95% confidence intervals will have approximately the correct coverage probability – not too different from the nominal 95% – if the bootstrap distribution appears to be normally distributed. However, if the bootstrap distribution is asymmetric then these 95% confidence intervals may have coverage probability different from 95%, especially if the asymmetry is substantial. If the bootstrap distribution is asymmetric then it is recommended to use Efron's BCa interval (Davison and Hinkley, 1997). We could even propose the use of the Studentized bootstrap in case of asymmetry.

3 Evaluation criteria

We now evaluate three properties of a confidence interval using Monte Carlo simulations with different sample sizes: (a) coverage accuracy proportion, (b) the balance (another terminology is the central coverage) and (c) the width for two statistics of interests: the mean and the median. We draw 10,000 bootstrap replications since the BCa can be unstable when the acceleration factor is computed from the jackknife.

3.1 Coverage accuracy proportion

The coverage proportion of a confidence interval is the proportion that the true parameter value is included within the confidence interval across replicated samples. The coverage proportion (the chance that the interval covers the parameter) is called the confidence level.

For example, ideally, a 95% confidence interval (a nominal coverage) should include the true parameter of interest 95% of the time over replications. If we increase the coverage proportion by using a 99% calculation it becomes more accurate because the true parameter value is more likely to be within the range.

Using bootstrap methods we get an approximate interval. An approximate interval could undercover (i.e. advertised 95% when its actual coverage is only 91%) or in the rare but less serious case overcover (i.e. advertised coverage is 95% but actual is 98%). If the actual confidence level is close we call it accurate. Accuracy is important with bootstrap confidence intervals which are never exact**;** but some variants may be more accurate than others.

We use simulation to estimate the empirical coverage proportion for the confidence interval, which should be close to predefined 95% confidence level.

3.2 Balance

The balance (also called central coverage) of a confidence interval refers to how the non-coverage is split. That is, how many times the population value is greater than the upper limit of the interval and how many times the population value is smaller than the lower limit of the interval. In an ideal situation, a confidence interval should be balanced such that the population value is greater than the upper limit or smaller than the lower limit for the same number of times across replications (e.g., 2.5% of the time for a 95% confidence interval), while achieving the desired level of coverage (Aguirre-Urreta and Rönkkö, 2018).

While working with an asymmetric distribution there is no obligation to have asymmetric confidence intervals. Any interval that covers the true value with probability 0.95 is a valid 95% confidence interval, and there will always be a unique symmetric interval that achieves this. Researchers construct two one-sided bounds with equal confidence level and then put them together, providing balance (or approximate balance if the method is approximate) (Meeker, Hahn and Escobar, 2017). However, it is usual to use intervals with half the probability in each tail of the distribution, and these will be asymmetric about the sample mean (or median). Such asymmetric intervals will typically be shorter (more precise) than the symmetric one, which is the reason for using them. We note a particularity regarding the Lognormal distribution. The confidence interval of the log of the Lognormal mean and median is symmetric, but we consider the antilogs, so the interval will not be symmetric.

3.3 Width

We want to know how close the calculated value is to the true value. Frequentist confidence intervals do not strictly provide that measure of precision. While there is no direct connection between the width of a confidence interval and the precision, there is an almost universal connection between standard errors and precision, and in most cases the width of a confidence interval is proportional to a standard error. The width tells something about the range of plausible values given the data (i.e., how well we can bound the true value), hence it acts like a measure of precision. A narrower confidence interval may be more precise but, when calculated the same way, such as the 95% method, they all have the same accuracy. They capture the true parameter value the same proportion of the time (95%). A 99% confidence interval is wider than a 95%. Therefore, it is more likely that it will contain the true parameter value. If we make a confidence interval narrower with lower variability and higher sample size, it becomes more precise: the likely values cover a smaller range. If we increase the coverage by using a 99% calculation, it becomes more accurate: the true value is more likely to be within the range. So a narrower confidence interval may be more precise (shorter width), but its accuracy is fixed by the procedure backing it, be it 90%, 95%, etc. As long as that procedure was correctly designed, the true value will be contained in the interval at the prescribed rate. So, in the 95% of cases where the confidence interval does cover the actual parameter, then the width/length tells us something about the range of plausible values given the data. In the 5% of cases where it does not, then the confidence interval is misleading since the sample is misleading.

3.4 How to choose the "best" nonparametric bootstrap technique?

The problem of testing the performance of interval estimation procedures involves approximations of various kinds, and there is a need to check that the actual performance of a procedure is closed to what is claimed (Algarni, Almarashi, Kundu, Abd-Elmougod, Abdel- Khalek, 2022).

The coverage proportion and the interval width are related: longer confidence intervals have higher coverage proportion and shorter confidence intervals have lower coverage proportion. Since these two key concepts are in opposition to each other, that is, better coverage proportion goes with weaker width/length and vice versa, it is useful to combine these measures.

Whatever the inferential framework, we think the most important criterion is that the coverage proportion, whether this be a frequentist or bayesian version, should be correct. Our approach is based on the frequentist viewpoint of Cox, Fraser and Reid (Fraser, 1991; Fraser and Reid, 1995; Reid and Cox, 2015): when we are comparing various inferential methods, the most important criterion is the coverage proportion. Beyond that there is scope for different views. If the underlying distribution is symmetric, then it would be very unusual not to use the symmetric interval, which is also the shortest interval with a given coverage proportion. If the underlying distribution is not symmetric, then there are two obvious choices: put equal probability in each tail or choose the shortest interval. These two options, both of which used in practice, will give different intervals. On the whole, bayesians opt for the shortest interval (highest posterior density or HPD interval) while frequentists tend for equal probability in the tails. This difference is probably due more to tradition or ease of computation than to any explicit reasoning. Either version can be defended.

Theoretically, the methods should be as close to (1-α) as possible in the center and with equal tail, a good balance (α/2 on each side). If the methods are close in coverage then we can choose the one having the shortest average width: the shorter the better because the more precise. Indeed, ideally, we would like a narrow confidence interval because we will have a much better idea of the actual population value.

In general, classical statistics puts first priority on getting the coverage proportion correct and secondary priority on precision (small interval width). We note as well that classical statistics seeks intervals whose coverage equals or exceeds the 95% level, so an interval that undercovers is undesirable while a conservative interval satisfies this criterion. Nonetheless, a conservative interval may be undesirable because it is too wide and therefore not precise. In principle, it is desirable for an interval to be as small as possible, while still having the correct coverage proportion.

4 Simulation design

Researchers in the social sciences often work with small data sets, even with just a few observations, so we consider a range of sample sizes: n=5; n=20; n=100 and n=300. What will simulation reveal about the behavior of confidence intervals with very small samples and with larger samples? A good sample is strongly representative of the population; but if the underlying distribution is long-tailed, a small sample will seldom be adequate to construct a reliable confidence interval. Moreover, it is important to consider whether the mean is an appropriate measure of location. If we are looking for a typical or central value of a skewed distribution, the median is likely preferable to the mean.

We run simulations with R software (R Core Team, 2023) for four two-sided nonparametric bootstrap confidence intervals (percentile, BCa, Studentized and Normal) around the mean and the median of the Normal distribution and of three skewed distributions (Weibull, Lognormal and Gamma).

The simulation method for estimating the coverage proportion of a confidence interval has four steps:

(1) Select the number of replications, in this article, 2,000 times and the sample sizes *n*, in this article *n*=5; *n*=20; *n*=100 and *n*=300.
(2) Simulate 2,000 samples of size *n* from the population.
(3) Compute the confidence interval for each sample.
(4) Compute the proportion of samples for which the known population parameters (mean and median) are contained in the confidence interval.

These proportions are estimates for the empirical coverage proportions for the confidence interval. As an example, for *n*=5, we randomly select 5 values inside the population of *N*=100,000 values. We replicate the experiment 2,000 times and thus get 2,000 different confidence intervals. We finally count the proportion of samples for which the known mean or the known median is contained in the 2,000 confidence intervals.

Accordingly the simulations allow us to calculate the coverage proportion, the balance and the width/length of these four confidence intervals for different sample sizes.

We evaluate the performance of the four nonparametric bootstrap methods in the Normal distribution with zero mean and unit variance, in Weibull with shape=2 and scale=2, in Gamma with shape=1/16 and rate=1/16 and in Lognormal with log mean=0 and log standard deviation=1.

5 Results

This section provides the results of the simulation study showing the performance of the four two-sided bootstrap confidence intevals methods in terms of coverage, balance and width. Our smallest sample size is *n*=5 because with *n*=3 the three bootstrap confidence intervals break down; indeed the coverage proportion is between 73% and 83%. These coverage proportions are too low for reliable inference. With *n*=4, only the Studentized interval achieves 91% coverage; the coverage of BCa and the percentile interval is just 79%. Given that coverage should be at least 90%, the smallest sample in our simulations is *n*=5.

We remark that, under appropriate conditions, the coverage of a statistic will converge asymptotically to the nominal confidence level; and the average width of the intervals will decrease. However, convergence fails for some statistics, for example the sample maximum or minimum.

In the following tables, the largest value for the coverage proportion, the more balanced values for the balance and the lowest value for the interval width/length are shown with a bold font for the mean and for the median.

Table 1: Gaussian distribution (mean=0; standard deviation=1). The population mean = 0 and the population median = 0

			Percentile	**BCa**	**Studentized**	**Normal**
coverage proportion	*n*=5	mean *median*	85.1% *94.2%*	83.5% *75.9%*	**94.9%** ***94.6%***	85.1% *91.1%*
Balance	*n*=5	mean *median*	7.8% LL; 7.1% UL *3% LL; 2.8% UL*	7.8% LL; 8.7% UL *3.4% LL; 20.7% UL*	**2.4% LL; 2.7% UL** ***2.7% LL; 2.7% UL***	7.7% LL; 7.2% UL *4.8% LL; 4.1% UL*
Interval width	*n*=5	mean *median*	1.49 *2.38*	1.52 ***1.67***	3.76 *3.15*	**1.48** *2.33*
coverage proportion	*n*=20	mean *median*	91.9% ***93.1%***	**93.8%** *92.8%*	93.7% *85.8%*	93.7% *90.6%*
Balance	*n*=20	mean *median*	**3.8% LL; 4.3% UL** *3.1% LL; 3.8% UL*	2.8% L; 3.4% UL *2.2% LL; 5% UL*	**3.4% LL; 2.9% UL** *7.6% LL; 6.6% UL*	3.6% LL; 2.7% UL ***5% LL; 4.4% UL***
Interval width	*n*=20	mean *median*	**0.84** ***1.04***	0.85 *1.06*	0.93 *1.11*	**0.84** ***1.04***
coverage proportion	*n*=100	mean *median*	94.4% *93.6%*	95% ***95%***	95.6% *88.3%*	**96.1%** *91.8%*
Balance	*n*=100	mean *median*	2.4% LL; 3.2% UL *4% LL; 2.4% UL*	2% LL; 3% UL *1.9% LL; 3.1% UL*	2.5% LL; 1.9% UL *6.6% LL; 5.1% UL*	**2.1% LL; 1.8% UL** ***4.5% LL; 3.7% UL***
Interval width	*n*=100	mean *median*	0.39 *0.49*	**0.38** *0.49*	0.40 ***0.48***	0.39 *0.50*
coverage proportion	*n*=300	mean *median*	92.7% ***94.3%***	92.7% *94%*	**96.6%** *89.8%*	93.9% *94.2%*
Balance	*n*=300	mean *median*	3% LL; 4.3% UL *2% LL; 3.7% UL*	3% LL; 4.3% UL *2% LL; 4% UL*	**1.5% LL; 1.9% UL** *4.6% LL; 5.6% UL*	2% LL; 4.1% UL ***3.1% LL; 2.7% UL***
Interval width	*n*=300	mean *median*	0.23 *0.28*	**0.22** *0.28*	0.23 ***0.27***	**0.22** *0.29*

Table 2: Weibull distribution (shape =2 ; scale = 2). The population mean = 1.77 and the population median = 1.66

			Percentile	**BCa**	**Studentized**	**Normal**
coverage proportion	*n*=5	mean *median*	83.8% ***93.6%***	84% *76.9%*	**95.2%** *93.1%*	84.7% *89.5%*
Balance	*n*=5	mean *median*	4.5%LL; 11.7%UL ***3.2% LL; 3.2% UL***	5.5% LL; 10.5% UL *3% LL; 20.1% UL*	**1.9% LL; 2.9% UL** *3.1% LL; 3.8% UL*	4.8%LL; 10.5%UL *4.6% LL; 5.9% UL*
Interval width	*n*=5	mean *median*	**1.33** *2.1*	1.39 ***1.41***	3.74 *2.82*	1.35 *2.12*
coverage proportion	*n*=20	mean *median*	91.6% ***94.5%***	94.4% *94.4%*	**95.4%** *87.6%*	92.2% *91.8%*
Balance	*n*=20	mean *median*	3% LL; 5.4% UL ***2.4% LL; 3.1% UL***	**2.3% LL; 3.3% UL** *1.8% LL; 3.8% UL*	1.3% LL; 3.3% UL *5.2% LL; 7.2% UL*	2.1% LL; 5.7% UL *3.7% LL; 4.5% UL*
Interval width	*n*=20	mean *median*	0.79 ***0.97***	**0.78** *1*	0.89 *1.03*	0.79 *1.06*

coverage proportion	*n*=100	mean *median*	94.6% *93.8%*	92.4% ***96%***	**96%** *85.6%*	94.4% *93.6%*
Balance	*n*=100	mean *median*	2.2% LL; 3.2% UL ***2.7% LL; 3.5% UL***	4.2% LL; 3.4% UL *2.6% LL; 1.4% UL*	**1.8% LL; 2.2% UL** *6.4% LL; 8% UL*	1.8% LL; 3.8% UL *2.6% LL; 3.8% UL*
Interval width	*n*=100	mean *median*	**0.36** *0.47*	**0.36** *0.47*	0.37 ***0.46***	**0.36** *0.47*

coverage proportion	*n*=300	mean *median*	94.2% ***95%***	**96.6%** *93%*	96.4% *89%*	94.7% *93.2%*
Balance	*n*=300	mean *median*	2.8% LL; 3% UL ***2.5% LL; 2.5% UL***	**1.7% LL; 1.7% UL** *3.4% LL; 3.6% UL*	1.2% LL; 2.4% UL *5.7% LL; 5.3% UL*	1.9% LL; 3.4% UL *3.9% LL; 2.9% UL*
Interval width	*n*=300	mean *median*	**0.21** ***0.27***	**0.21** *0.28*	**0.21** *0.28*	**0.21** *0.28*

Table 3: Lognormal distribution (meanlog=0 sdlog=1). The population mean = 1.66 and the population median = 1

			Percentile	**BCa**	**Studentized**	**Normal**
coverage proportion	*n*=5	mean *median*	74.9% *93.4%*	72.7% *77%*	**89.7%** ***93.6%***	73% *90.3%*
Balance	*n*=5	mean *median*	0.8%LL; 24.3%UL ***3.8% LL; 2.8% UL***	1,4% LL; 25.9% UL *3.5% LL; 19.5% UL*	**0.1%LL; 10.2%UL** *2% LL; 4.4% UL*	1.3%LL; 25.7%UL *3.3% LL; 6.4% UL*
Interval width	*n*=5	mean *median*	**2.34** *3.84*	2.53 ***1.51***	14.3 *3.94*	2.37 *3.59*

coverage proportion	*n*=20	mean *median*	84.3% *94%*	86.8% ***94.5%***	**90.4%** *89.6%*	83.9% *92.5%*
Balance	*n*=20	mean *median*	1.2%LL; 14.5%UL ***3% LL; 3% UL***	2.4% LL; 10.8% UL *2.2% LL; 3.3% UL*	**1.4% LL; 8.2% UL** *3.9% LL; 6.5% UL*	0.1% LL; 16% UL *2.5% LL; 5% UL*
Interval width	*n*=20	mean *median*	**1.52** *1.15*	1.77 ***1.13***	2.64 *1.39*	1.55 *1.25*

coverage proportion	*n*=100	mean *median*	91.4% *96%*	92.2% ***97.6%***	**94.3%** *92.9%*	91.1% *91.2%*
Balance	*n*=100	mean *median*	1.4% LL; 7.2% UL ***2.1% LL; 1.9% UL***	2.8% LL; 5% UL *0.6% LL; 1.8% UL*	**2.1% LL; 3.6% UL** *2.4% LL; 4.7% UL*	0.9% LL; 8% UL *2.5% LL; 6.3% UL*
Interval width	*n*=100	mean *median*	**0.79** ***0.50***	0.84 ***0.50***	0.94 *0.53*	0.80 *0.51*

coverage proportion	*n*=300	mean *median*	93.7% *95%*	93% ***95.6%***	**94.2%** *92.1%*	93.1% *93.5%*
Balance	*n*=300	mean *median*	0.5% LL; 5.8% UL ***2.3% LL; 2.7% UL***	**3.4% LL; 3.6% UL** *1.4% LL; 3% UL*	1.4% LL; 4.4% UL *2.1% LL; 5.8% UL*	0.1% LL; 6.8% UL *2.3% LL; 4.2% UL*
Interval width	*n*=300	mean *median*	0.49 *0.29*	**0.48** ***0.28***	0.53 *0.31*	0.49 *0.29*

Table 4: Higly skewed Gamma distribution (shape =1/16; rate = 1/16). The population mean = 0.99 and the population median = 0.00014

			Percentile	**BCa**	**Studentized**	**Normal**
coverage proportion	*n*=5	mean *median*	46% *93.4%*	48.4% *78.4%*	**86.8%** ***96.5%***	42.7% *95%*
Balance	*n*=5	mean *median*	0.1%LL; 53.9%UL ***2.9% LL; 3.7% UL***	0.2% LL; 51.4% UL *2.5% LL; 19.1% UL*	**0.1%LL; 13.1%UL** *0.5% LL; 3% UL*	0% LL; 57.3% UL *0.9% LL; 4.1% UL*
Interval width	*n*=5	mean *median*	**2.62** *4.25*	3.30 ***0.70***	1.7e+16 *1.64*	2.79 *4.61*

coverage proportion	*n*=20	mean *median*	71.9% *93.2%*	73.5% *93.5%*	**92.9%** *82.2%*	69.3% ***99.4%***
Balance	*n*=20	mean *median*	0.4%LL; 27.7%UL *4.9% LL; 1.9% UL*	1.1% LL; 25.4% UL *4.3% LL; 2.2% UL*	**0.7% LL; 6.4% UL** *0% LL; 17.8% UL*	0% LL; 30.7% UL **0% LL; 0.6% UL**
Interval width	*n*=20	mean *median*	**2.30** *0.14*	3.11 *0.13*	343.5 ***0.05***	2.53 0.28

coverage proportion	*n*=100	mean *median*	89.4% *95.2%*	89.2% *94.4%*	**93.9%** *76.6%*	85.2% ***96.8%***
Balance	*n*=100	mean *median*	0.5%LL; 10.1%UL ***2.4% LL; 2.4% UL***	2.2% LL; 8.6% UL *2.4% LL; 3.2% UL*	**1.2% LL; 4.9%UL** *0% LL; 23.4% UL*	0.1%LL; 14.7%UL *0% LL; 3.2% UL*
Interval width	*n*=100	mean *median*	**1.43** *0.005*	1.68 *0.004*	2.34 ***0.003***	1.45 *0.007*

coverage proportion	*n*=300	mean *median*	92.7% ***95.1%***	**95%** *94.8%*	94.5% *79.4%*	90.9% *94.1%*
Balance	*n*=300	mean *median*	1.1% LL; 6.2% UL ***2.3% LL; 2.6% UL***	1.3% LL; 3.7% L *2% LL; 3.2% UL*	**2.2% LL; 3.3% UL** *0% LL; 20.6% UL*	0.5% LL; 8.6% UL *0% LL; 5.9% UL*
Interval width	*n*=300	mean *median*	0.90 *0.0009*	0.97 *0.0009*	1.07 ***0.0008***	**0.85** *0.001*

6 Discussion

We start with general discussions. The accuracy of a confidence interval method depends largely on the skewness and kurtosis of the sampling distribution. For example, for the two distributions (Gaussian and Weibull) with quite low skewness we clearly see that the Normal interval performs very well. With higher skewness (Lognormal and Gamma) the other bootstrap methods perform much better than the Normal. High kurtosis (Lognormal and Gamma) causes different problems. The coverage is quite good – sometimes even too high – but interval widths are quite large and variable, in agreement with (l'Ecuyer, Nakayama, Owen and Tuffin, 2023).

Our results show that the more sophisticated methods – Studentized and BCa – are often conservative for both the mean and the median, especially for large sample sizes (*n*=100 and *n*=300). An exception is the higly skewed Gamma distribution.

Because samples from social science data are often small, it is noteworthy that the coverage for n = 5 and n = 20 is too optimistic except for the Studentized in Gaussian and Weibull distributions. There the coverage proportion is very often equal the theoretical level (95%), and sometimes it is even conservative, which makes sense since it is more like the standard t-test. More precisely, for *n*=5 and *n*=20, the coverage of intervals is not different from the nominal for the mean and the median with Studentized and BCa in Gaussian and Weibull distributions. For more skewed Lognormal and Gamma distributions, the coverage of intervals is not always close to nominal for the mean and the median.

However all the other methods tend to undercover – some dramatically –, especially for the mean and for small sample size.

For larger sample sizes (*n*=100 and *n*=300) the BCa and the Studentized do well. Other numerical studies have led to similar conclusions: in small samples, nonparametric bootstrap methods typically undercover somewhat, and the Studentized bootstrap method can work better than BCa method (Davison and Hinkley, 1997).

The percentile method has a good balance for the median while the Studentized intervals have the best balance for the mean.

No method performs very well overall, but the Studentized interval works best for small sample sizes (*n*=5 and *n*=20) at the cost of a lack of precision (large interval width); to a lesser extent the BCa works relatively well in small samples.

Even if Studentized and BCa perform quite well in small sample sizes, the results of these simulations should raise some concern about using bootstrapping as a panacea for small sample sizes.

Far from being a drawback, the fact that the Studentized bootstrap method can produce long confidence intervals (a lack of precision) is precisely what gives it the best coverage of the methods considered in our simulation study, and the "conservativeness" of the BCa method is what may lead it to undercover.

The interval widths are very similar for all methods in large samples (*n*=100 and *n*=300), which is absolutely not the case for small sample sizes as we could expect. Indeed, the most common way researchers make the confidence intervals narrow is by increasing the sample size.

As for the balance, the results are very mixed. Based on the lower and upper tails some methods are not different from the theoretical value, others are quite different. There is no method that is always more balanced than the others.

We now report more precise results for each of the four population distributions together with brief summaries.

6.1 Gaussian distribution

As we might expect the Normal confidence intervals perform well in the Gaussian distribution for every sample size and for the mean and the median.

For very small sample size (*n*=5), the Studentized interval has the best balance and the largest coverage proportion due to the very large interval width. So the Studentized is very accurate but not

very precise. Based on the central coverage proportion and based on the lower and upper tails, the Studentized version is not significantly different from the theoretical value of 95%.

Fo the mean and *n*=20, the BCa provides the larger coverage proportion although the Normal and Studentized give nearly the same performance as the BCa. Except for the percentile, all three methods are not significantly different from the theoretical value of 95% based on the central coverage proportion.

The percentile and the Studentized intervals have good balance. The percentile and the Normal provide the smallest interval width. For the median, the percentile and the Normal perform well.

For the mean and *n*=100 the Normal interval is conservative (96.1%) and offers the largest coverage proportion and the best balance. Based on the lower and upper tails, the Normal is not significantly different from the theoretical value of 95%.

The BCa has the smallest interval width, but all the other methods perform nearly as well. For the median, the BCa has the largest coverage proportion; its coverage proportion for the mean and the median is exactly 95%. The Normal produces the best balance while the Studentized has the smallest interval width.

For the mean and *n*=300 the Studentized is conservative (96.6%), has the largest coverage proportion and the best balance. Based on the lower and upper tails, the Studentized is not significantly different from the theoretical value of 95%. The Normal and the BCa have the smallest interval width. For the median, the percentile and the Normal offer the largest coverage proportion, the Normal has the best balance and the Studentized has the smallest interval width.

6.1.1 Small synthesis

In summary, for the mean in Gaussian distributions, the Studentized and the Normal confidence intervals provide the best results in general. The Studentized very often has a larger interval width while the Normal is the more precise. For the median, the Normal is the more balanced. So, in case of Gaussian distributions, the Normal intervals are preferable.

6.2 Weibull distribution

For *n*=5 and for the mean, the Studentized is conservative (95.2%), provides the largest coverage proportion and the best balance but again at the cost of a wide interval, so there is again a lack of precision. Based on the lower and upper tails, the Studentized is not significantly different from the theoretical value of 95%.

The percentile has the smaller interval width. For the median, the percentile gives the largest coverage proportion and the best balance. The BCa provides the smaller interval width.

For *n*=20 and the mean, the Studentized is conservative (95.4%) and offers the larger coverage proportion. The BCa gives the best balance and the smaller interval width. For the median, the percentile is the best in all three categories (coverage proportion, balance and interval width). The percentile's level (94.5%) is not statistically different from the theoretical value of 95%.

For *n*=100 and the mean, the Studentized provides the largest coverage proportion and the best balance. Based on the lower and upper tails the Studentized is not significantly different from the theoretical value of 95%. They all have nearly the same interval width. For the median, the BCa is conservative (96%) and gives the largest coverage proportion while the percentile provides the best balance. By a narrow margin the Studentized offers the smallest interval width.

For *n*=300, the BCa is conservative (96.6%) and performs the best in the three categories. All the methods give the same interval width. For the median, the percentile has the theoretical coverage proportion (95%) and performs the best in the three categories. Based on the lower and upper tails the percentile is true (2.5% lower and 2.5% upper). All the methods give nearly the same interval width (precision).

6.2.1 Small synthesis

We can summarize these results in Weibull distribution by saying that for the mean, the Studentized and BCa are the best in general while the Studentized suffers from a lack of precision (very large interval width) while the BCa is the most precise interval. For the median the percentile is the best in general. The Weibull distribution being asymetrical, the Normal intervals perform quite poorly.

6.3 Lognormal distribution

For *n*=5 and the mean, the Studentized has an actual coverage of 89.7%. It provides the best coverage proportion and the best balance but again at the cost of imprecision. Indeed the interval width is very large. The percentile is the most precise interval. Nevertheless, we note that, based on the central coverage proportion, all methods differ significantly from the theoretical value of 95%.

For the median, the Studentized has an actual coverage of 93.6% – close to the theoretical central coverage proportion – but again has a wide interval. The percentile provides the best balance and the BCa has the lowest interval width.

For *n*=20 and the mean, the Studentized has actual coverage of 90.4%, provides the largest coverage proportion and the best balance with again a lack of precision. The percentile is the more precise interval. For the median, the BCa achieves almost the nominal coverage (94.5%) while providing the largest coverage proportion and the smallest interval width. The percentile provides the best balance.

For *n*=100 and the mean, the Studentized is very close to the nominal value (94.3%), gives the largest coverage proportion and the best balance but again with a lack of precision. The percentile provides the lowest interval width. For the median, the BCa is conservative (97.6%), offers the largest coverage proportion and the smallest interval width. The percentile is the more balanced. Indeed, the lower and upper tails are not significantly different from the theoretical values.

For *n*=300 and the mean, the Studentized is very close to the nominal value (94.2%) and provides the largest coverage proportion. The BCa is the most balanced and the most precise. For the median, the BCa is conservative (95.6%), has the largest coverage proportion and is the most precise. The percentile is the most balanced. Again, the lower and upper tails are not significantly different from the theoretical values.

6.3.1 Small synthesis

We can summarize these results in Lognormal distribution by saying that the Studentized is the best for the mean but always with a lack of precision. For the median, the BCa and the percentile are the best. The BCa is conservative, has largest coverage proportion, is the most precise while the percentile is the most balanced. Again, the Lognormal distribution being an asymetrical distribution, the Normal intervals perform quite poorly.

6.4 Gamma distribution

For *n*=5 and the mean, the Studentized has an actual coverage of 86.8%, provides the larger coverage proportion and the best balance at the cost of a huge interval width (1.7e+16). However, we note that all the methods are very liberal, indeed, based on the central coverage proportion, they are all significantly different from the theoretical value of 95%. The percentile is the more precise interval. For the median, the Studentized is conservative (96.5%), gives the larger coverage proportion, the percentile is the more balanced and the BCa is the more precise.

For *n*=20 and the mean, we have exactly the same namely the Studentized has the larger coverage proportion and the best balance again at the cost of a very wide interval (343.5) and the percentile is the more precise interval. For the median, the Normal provides the larger coverage proportion and the best balance at the cost of a lack of precision, indeed the interval width is large, while the Studentized is the more precise.

For *n*=100 and the mean we have again the same situation: the Studentized has the largest coverage proportion –nearly equal to the theoretical level 95%; moreover it has the best balance. The percentile is the most precise. For the median, the larger coverage is offered by the Normal again at the cost of some imprecision, while the percentile gives the best balance and the Studentized is the most precise.

For *n*=300 and the mean, the BCa has exactly the theoretical coverage proportion (95%), offers the largest coverage proportion while the Studentized is the best balanced and the Normal the most precise. For the median, the percentile has exactly the theoretical coverage proportion (95.1%), provides the largest coverage proportion and the best balance while the Studentized is the most precise.

How could we explain the big difference between the coverage proportions for the mean and for the median? The issue may have something to do with the extreme Gamma distribution that generates our data. When the shape parameter is small (e.g., less than 0.1), most of the Gamma samples will be very close to 0 – numerically almost exactly 0 –; only a few will be large. Since most of the bootstrap samples will contain only these very small values, the estimate of the mean will be very far from the true value of 1. In that case, it makes sense that the bootstrap intervals could miss the target value with high probability. This does not affect the median because the median of the Gamma distribution is very small, much smaller than the mean of 1. So even if we get bootstrap samples that only contain small observations, their small median will be close to the true Gamma median.

By the way, how could we explain the superiority of the Normal bootstrap CIs compared to Studentized in such a skewed Gamma distribution? In other words, why would the Studentized version be affected more than the Normal bootstrap? The Studentized version divides by an estimated standard error which could be very small since the bootstrap sample tends to have only very small values.

Therefore the ratio (x.bar.boot - x.bar) / se.boot could vary wildly since the numerator is large relative to the denominator.

6.4.1 Small synthesis

We summarize saying that for the mean the Studentized very often offers the largest coverage proportion and the best balance while the percentile gives the most precise intervals. For large *n*=300, BCa is good. For the median, the Normal often delivers the largest coverage proportion at the cost of

a lack of precision, the percentile often offers the best balance and the Studentized often provides the best precision.

7 Conclusion

When we compute a 95% confidence interval by bootstrap the coverage is based on large sample theory. Thus, depending on the sample size and the parameter that we are attempting to estimate, the actual coverage of the unknown parameter may either overstate the actual coverage or it may be too conservative. Of course we will not know for sure which condition holds in practice, but we can simulate the coverage by repeatedly generating data from the assumed distribution and recording the percentage of bootstrap confidence intervals that actually contain the parameter of interest for that distribution.

The simulation should address several key issues: which type of interval has coverage closest to the nominal value (95%); and which interval has good balance and a short width for different sample sizes (*n* from 5 to 300)? Evidently we cannot expect any confidence interval method to be "automatic," if only because there are trade-offs among these criteria. There is a "no free lunch" bootstrap technique. Simulation methods facilitate the computation of good confidence intervals without recourse to advanced mathematics, but they have not removed the need for thoughtful data analysis.

The simulations indicate that the Studentized bootstrap is very effective for the construction of 95% confidence intervals for the mean and the median of a small sample. It closely approximates the nominal coverage for a diverse collection of sampling distributions. In view of its desirable large sample properties (Hall, 1988) and superior small sample behaviour, the Studentized should be more widely used to compute 95% confidence intervals for the mean and the median in social sciences research. Nevertheless, the Studentized bootstrap has two serious drawbacks: it is not scale invariant, and the interval width can be very large, so the Studentized intervals can suffer from a lack of precision. To a lesser extent the BCa method provides also better results than the two other methods.

The Studentized and the BCa intervals are second-order correct (Efron and Tibshirani, 1993), so their coverage approximates 95% with smaller sample sizes than any of the first-order correct methods such as the Normal and percentile intervals. So if we are concerned about coverage due to small sample size and asymmetrical distributions, the Studentized and BCa intervals seem to be good choices.

The BCa method works well under a wide range of situations, including skewness; the more non-Normal the sampling distribution is, the more reason to use the BCa. The only reason not to use BCa is that the sample size is too small to estimate the BCa parameters reliably.

While the Studentized interval is a viable alternative to the non-parametric BCa method and is just as easy to implement, the Studentized interval can break down when the standard error of the estimator is relatively unstable. As a practical recommendation, the BCa method will often perform well where the Studentized interval might have trouble, but the BCa method also has a tendency to produce intervals that do not achieve the nominal coverage (DiCiccio and Efron, 1996; Hall, 1988).

We wanted to show the strength of the Studentized nonparametric bootstrap confidence interval around the mean and the median for the very skewed Gamma distribution with shape=rate=1/16. We ran the simulation 2,000 times with n=4 and n=3. For n=4, it is remarkable that the coverage proportion of the mean is quite high: 86.2% (BCa is 45.4%).

However, the Studentized bootstrap interval breaks down for *n*=3 with a coverage proportion of only 70.1%. The simulation confirms that the Studentized interval has a very good second-order correctness confidence interval because even with *n*=4 the coverage proportion is quite good 86.2% (near 90%). Nevertheless, the interval width is very large. The BCa, being second-order too, does not perform as well as the Studentized method with such a small sample size.

For comparison, the coverage proportion of the Normal interval for *n*=4 is only 38.2%.

The current study focused on the 95% confidence level and sample sizes ≤ 300 because they are most frequently encountered in social sciences applications. Future research on the use of different confidence levels and larger samples (i.e., 1,000-5,000) is also warranted to provide more practical implications for applied research.

One of the limitations of our paper is the lack of noise and outliers in the distributions. It would be worthwhile to explore the consequences of some noise and outliers in the distributions. Bootstrap methods require samples to be representative of the population. In a large bootstrap sample a single outlier will not be selected in most replications so the sample data will mimic the majority of the data. However, the bootstrap could fail with small sample sizes, high levels of contamination, or highly skewed distributions.

Future research will examine altenatives to the bootstrap. There are methods that, unlike the bootstrap, can in principle achieve 4^{th} order accuracy but are mathematically more complicated. The likelihood-ratio test (LRT) is a parametric method that dates from the 1930s but is not often used because of its theoretical complexity and problems of numerical computation. Many approaches have been suggested to improve the accuracy of the LRT. Some are problem specific, some are theoretically complicated and some are numerically complicated. One successful approach is the Bartlett correction (Wong, 2023; Li, Lin and Wong, 2020; Shi and Wong, 2018). These three papers applied the LRT to a specific problem. From Meeker, Hahn and Escobar (2017), we observe that the confidence interval based on the LRT is often superior to the confidence interval based on the asymptotic distribution of the MLE. If the population distribution is known, we should as well look at an exact confidence interval that could provide very good results.

In a nonparametric framework, the concept called "empirical likelihood" is an alternative to the nonparametric bootstrap for constructing confidence intervals. Empirical likelihood is based on the principle of maximizing the likelihood function subject to certain constraints imposed by the data. It does not rely on a specific parametric distribution for the data, making it a flexible and distribution-free method. The extension of empirical likelihood to high-dimensional cases, for example, can be found in the work by Chang et al. (2021). Empirical likelihood has several desirable properties, such as asymptotic optimality and coverage accuracy, which means that it provides valid confidence intervals with large samples. It also has the advantage of being Bartlett correctable (DiCiccio et al., 1991), which allows for improving the accuracy of the confidence intervals. These methods are not in the scope of this paper but are topics of this authors ongoing research.

Acknowledgments

We thank Professor Efron (Stanford University), Professor Owen (Stanford University), Professor Olive (Southern Illinois University), Professor Gilleland (University Corporation for Atmospheric Research), Professor Fearn (University College London) and Professor Martin (North Carolina State Univerity) for valuable comments and Professor Blankmeyer (Texas State University) for careful reading.

References

[1]. Algarni, A., Almarashi, A.M., Kundu, D., Abd-Elmougod, G.A., Abdel-Khalek, S. (2022). Comparison of different confidence intervals under Type-I censoring scheme. *Journal of Mathematics,* 2022: 1-9. https://doi.org/10.1155/2022/1272045

[2]. Aguirre-Urreta, M.I., Rönkkö, M. (2018). Statistical Inference with PLSc Using Bootstrap Confidence Intervals. *MIS Quarterly,* 42(3): 1001-1020. https://doi.org/10.25300/MISQ/2018/13587

[3]. Bryson, M. (1974). Heavy-tailed distributions: properties and Tests. *Technometrics,* 16(1): 61-68.

[4]. Chang, J., Chen, S.X., Tang, C.Y., Wu, T.T. (2021). High-dimensional empirical likelihood inference. *Biometrika,* 108(1): 127-147.

[5]. Chernick, M.R. (2011). *Bootstrap Methods: A guide for practitioners and researchers, Vol. 619.* Hoboken, NJ: John Wiley and Sons.

[6]. Davison, A.C., Hinkley, D.V. (1997). *Bootstrap methods and their application*. Cambridge, United Kingdom: Cambridge University Press.

[7]. DiCiccio, T.J., Hall, P., Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann. Statist*, 19(2): 1053-1061. DOI: 10.1214/aos/1176348137.

[8]. DiCiccio, T.J., Efron, B. (1996). Bootstrap confidence intervals. Statistical science. *A Review Journal Of The Institute Of Mathematical Statistics,* 11(3): 189-212.

[9]. Dufresne, D. (2008). Sums of lognormals. Available at: https://www.soa.org/globalassets/assets/files/static-pages/research/arch/2009/arch-2009-iss1-dufresne.pdf

[10]. Efron, B. (1979). Bootstrap Methods: Another look at the jackknife. *Ann. Statist,* 7(1):1-26. https://doi.org/10.1214/aos/117634455

[11]. Efron, B. (1981). Nonparametric Estimates of Standard Error: The Jackknife, the Bootstrap and Other Methods. *Biometrika*, 68(3): 589-599. https://doi.org/10.2307/2335441

[12]. Efron, B. (1982). Society for Industrial and Applied Mathematics. *The jackknife, the bootstrap, and other resampling plans.* Philadelphia, Pa: Society for Industrial and Applied Mathematics. http://www.loc.gov/catdir/enhancements/fy0726/81084708-t.html

[13]. Efron, B. (1987). Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association*, 82(397): 171-185. https://doi.org/10.2307/2289144

[14]. Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York. https://doi.org/10.1007/978-1-4899-4541-9

[15]. Fraser, D.A.S. (1991). Statistical Inference: Likelihood to Significance. *Journal of American Statistical Association*, 86: 258-265. https://doi.org/10.1080/01621459.1991.10475029

[16]. Fraser, D.A.S., Reid, N. (1995). Ancillaries and Third Order Significance. *Utilitas Mathematica*, 7: 33-53.

[17]. Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Ann. Statist,* 16(3): 927-953.

[18]. Hayter, A.J., Kiatsupaibul, S. (2014). Exact Inferences for a Gamma Distribution, *Journal of quality technology: A quarterly journal of methods applications and related topics,* 46(2): 140-149.

[19]. Imholte, G. (2012). Bradley Efron's Better Bootstrap Confidence Intervals, *Journal of the American Statistical Society,* 82(397): 171-185.

[20]. l'Ecuyer, P., Nakayama, M.K., Owen, A., Tuffin, B. (2023). Confidence Intervals for Randomized Quasi-Monte Carlo Estimators. (hal-04088085)

[21]. Li, X., Lin, W., Wong, A. (2020). Accurate inference for the mean of the Poisson-Exponential distribution. *Journal of the Iranian Statistical Society,* 19: 1-19.

[22]. Martin, R. (2015). Plausibility Functions and Exact Frequentist Inference, *Journal of the American Statistical Association,* 110(512): 1552-1561.

[23]. Meeker, W.Q., Hahn, G.J., Escobar, L.A. (2017). Statistical Intervals: A Guide for Practitioners and Researchers (2nd Edition). John Wiley & Sons, Hoboken.

[24]. Merz, B., Basso, S., Fischer, S., Lun, D., Blöschl, G., Merz, R., Guse, B., Viglione, A., Vorogushyn, S., Macdonald, E., Wietzke, L., Schumann, A. (2022). Understanding heavy tails of flood peak distributions. *Water Resources Research,* 58(6): 1-37.https://doi.org/10.1029/2021WR030506

[25]. Mitchell, R.L. (1968). Permanence of the log-normal distribution, *J. Optical Society of America,* 58: 1267-1272.

[26]. Nair, J., Wierman, A., Zwart, B. (2022). *The fundamentals of heavy tails*. Cambridge Series in Statistical and Probabilistic Mathematics.

[27]. R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https//www.R-project.org/.

[28]. Reid, N., Cox, D.R. (2015). On Some Principles of Statistical Inference. *International Statistical Review / Revue Internationale de Statistique*, 83(2), 293–308.http://www.jstor.org/stable/44162424

[29]. Shi, X., Wong, A. (2018). Accurate tests for the equality of coefficients of variation. *Journal of Statistical Computation and Simulation,* 88: 3529-3542.

[30]. Wong, A. (submitted). Comparing several gamma means: an improved log-likelihhod ratio test. *Entropy*. 25.