Performance of Bootstrap for Obtaining 95% Confidence Intervals on the Mean

Tatiana Matejovicova
University of St Andrews

Abstract

In this study performance of different bootstrapping methods when extracting the 95% confidence interval is evaluated. Two versions of methods from each of pivotal and percentile method families were considered for symmetric and non-symmetric distributions. Minimum sample size for percentile method to perform reasonably well was determined. We show the difference between the performance of the standard versions of both methods for different sample sizes and how this is improved when using the advanced versions. Finally the accuracy of the more advanced methods from the two families are compared and we conclude that the pivotal method performs better especially for small sample sizes.

Introduction

To specify uncertainty about the inferred distribution statistic confidence intervals are used to determine an interval that the true value of the statistic lies in with the given probability. However sometimes the distribution of the original data is hard or impossible to determine. Bootstrapping methods are a computationally heavy alternative to analytical methods to determine confidence interval when the distribution of the original data source is not known.

We will look at performance of different bootstrapping methods when extracting the 95% confidence interval of the mean. Methods from two families, percentile and pivotal, will be considered, each with a standard and an enhanced version.

Bootstrap principle notation

Let's define the notation of the standard Bootstrap terms, similarly as in Carpenter and Bithell (2000):

- $x_1, x_2, x_3, \dots x_n$ - n data points sample drawn from an unknown distribution $F(\theta)$ with an unknown mean θ
- $\hat{\theta}$ - the mean of our original sample
- $\hat{\sigma}$ - standard error of the mean of our original sample
- $\hat{\Theta}$ - the distribution of sample means
- $x*_1, x*_2, x*_3, \dots x*_n$ - a nonparametric bootstrap resample of the same size n
- $\hat{\theta}*: \hat{\theta}*_1, \hat{\theta}*_2, \dots \hat{\theta}*_b$ - the means of the b bootstrap resamples
- $\hat{\Theta}*$ - the distribution of the bootstrap resampled means
- $\hat{\sigma}*: \hat{\sigma}*_1, \hat{\sigma}*_2, \dots \hat{\sigma}*_b$ - the standard errors of the b bootstrap resamples
- $\alpha = 5\%$ - i.e. the $100\% - \alpha = 95\%$ confidence interval we are looking for

Bootstrap methods

Percentile methods

Two types of Percentile methods are implemented.

1. Percentile method

The traditional Percentile method is an attractive way to calculate bootstrapped confidence intervals because of its simplicity. We calculate $\widehat{\theta*}$, the parameter of interest i.e. the mean for each bootstrapped sample. We assume that the distribution of sample means can be approximated by the distribution of resampled bootstrapped means i.e. $\widehat{\Theta} \approx \widehat{\Theta*}$ Therefore the 95% confidence interval is calculated as estimated quantiles $(\widehat{\theta*}_{\frac{\alpha}{2}}, \widehat{\theta*}_{1-\frac{\alpha}{2}})$. Even though this method is appealing it is argued by Efron Tibshirani (1993) that the coverage error is substantial for distributions where $\widehat{\Theta}$ is non-symmetric.

2. Bias corrected and accelerated method (BCa)

In BCa method, the bias correction and acceleration terms are added to Percentile Method to make a better estimate of quantiles α_{lower} and α_{upper} to use. The 95% confidence interval is then calculated as estimated quantiles $(\widehat{\theta*}_{\alpha_{lower}}, \widehat{\theta*}_{\alpha_{upper}})$. The bias correction term accounts for the lack of symmetry in $\widehat{\Theta}$, the distribution of sample means. The acceleration term accounts for the change of shape of $\widehat{\Theta}$ as the true mean θ of the original distribution $F(\theta)$ changes.

Pivotal methods

Similarly as with percentile methods, two types of pivotal methods were implemented.

1. Non-studentized pivotal method

In pivotal methods the distribution $W = \widehat{\Theta} - \theta$ is approximated by $W* = \widehat{\Theta*} - \widehat{\theta}$. We have that:

$P(w_{\frac{\alpha}{2}} \le W \le w_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$

$P(w_{\frac{\alpha}{2}} \le \widehat{\Theta} - \theta \le w_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$

And so the 95% confidence interval for θ is $(\widehat{\theta} - w_{1-\frac{\alpha}{2}}, \widehat{\theta} - w_{\frac{\alpha}{2}})$. We approximate this by the interval $(\widehat{\theta} - w*_{1-\frac{\alpha}{2}}, \widehat{\theta} - w*_{\frac{\alpha}{2}})$.

This method is simple to understand but distributions W and $W*$ can differ significantly resulting in a coverage error.

2. Studentized pivotal method (Bootstrap-t)

The difference between the distributions W and $W*$ can occur in their variances. Studentized pivotal method accounts for that by approximating a transformed distribution; $T = \frac{\widehat{\Theta} - \theta}{\widehat{\sigma}}$ is approximated by $T* = \frac{\widehat{\Theta*} - \widehat{\theta}}{\widehat{\sigma*}}$. And so the 95% confidence interval $(\widehat{\theta} - \widehat{\sigma} t_{1-\frac{\alpha}{2}}, \widehat{\theta} - \widehat{\sigma} t_{\frac{\alpha}{2}})$ is approximated by $(\widehat{\theta} - \widehat{\sigma} t*_{1-\frac{\alpha}{2}}, \widehat{\theta} - \widehat{\sigma} t*_{\frac{\alpha}{2}})$.

Methods

Choosing simulation variables

In order to run a bootstrap simulation, four variables need to be determined; distribution to sample from, sample size, number of bootstrap resamples and number of simulations to evaluate a bootstrap confidence interval.

Throughout all the experiments, the number of bootstrap resamples and number of simulations for confidence interval evaluations are kept constant. This is because these numbers purely depend on computational resources and therefore are not as interesting as other variables. Therefore we fix:

- Number of bootstrap resamples = 1000, that is a standard number of bootstrap resamples, suggested for example by Efron and Tibshirani (1993).
- Number of simulations = 1000, that turned out to be a favourable number for a fair precision and a reasonable running time. The same number of simulations was used in a similar study by Lei and Smith (2003).

As suggested by Carpenter and Bithell (2000), the fact whether a distribution $\widehat{\Theta}$ of sample means is symmetric can greatly affect the performance of a confidence interval method. Therefore the following continuous distributions are used:

- Standard Normal (mean = 0, sd = 1) as a canonical example of a symmetric distribution.
- Exponential (rate = 1) as an example of a non-symmetric distribution. This distribution will be used to contrast the performance of methods on a non-symmetric distribution.

Because BCa and studentized pivotal method are trying to account for this property we would expect to improve the accuracy compared to their standard counterparts.

Finally, sample size is perhaps the most important variable to influence the correctness of a confidence interval. Sample size in range 4 - 100 will be considered similarly as in experiments performed by Lei and Smith (2003).

Evaluating the methods and obtaining data sets

When evaluating how good a bootstrapping method to generate confidence interval is, we call the method for 1000 different data inputs and calculate the percentage of times when the true data source mean is included within the confidence interval. To obtain an accurate comparison of performance of different methods, we only generate one data input for each part of the study.

Finally, to generate the data inputs a similar strategy is followed as in Lei and Smith (2003). Firstly we create a large data source of size 1000 for both the normal (symmetric) and exponential (non-symmetric) distributions that remain the same throughout the study and represents certain empirical distributions. For each sample we resample with replacement from the given data source. This is to imitate the imperfections of data from the real world.

Results

All the resulting data frames are stored in the folder "code/output" as csv files. All the produced figures are stored in the folder "code/plots". Finally all parts of the study are run in the file "code/SimulationStudy.R".

1. Sample size and percentile method performance

In the first simulation we try to estimate what is the effect of a sample size on the performance on percentile method when sampling from a symmetric and a non-symmetric distribution. Having larger sample size should improve the percentile coverage. Moreover, as explained in the introduction, percentile method should perform worse when the distribution $\widehat{\Theta}$ of sample means is not symmetric which is the case for non-symmetric distributions.

First inclusion percentage is calculated for sample sizes 5 - 100 to get a general idea of what the method performance is.

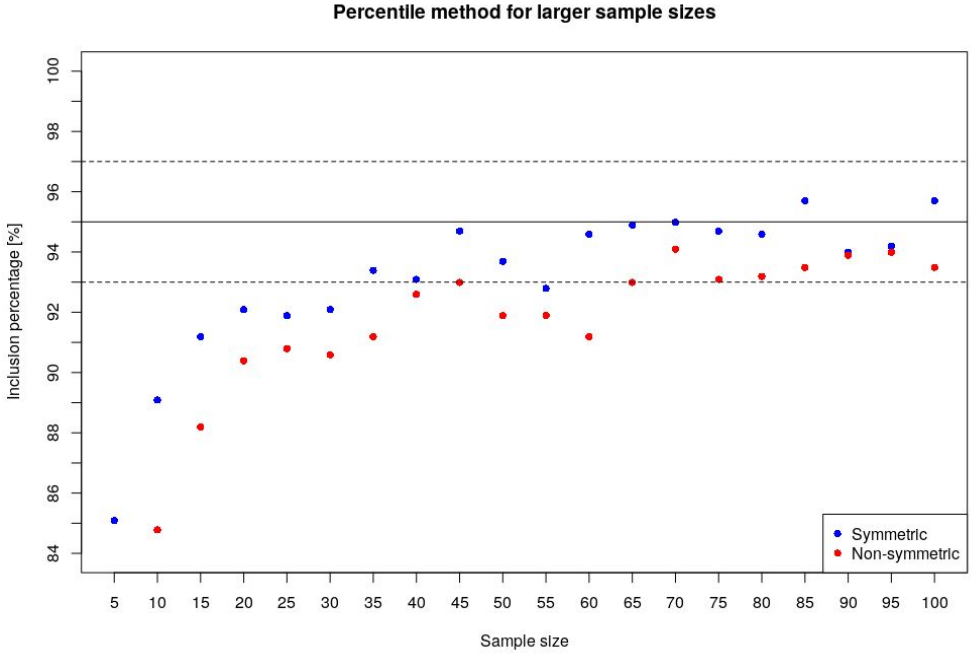

Figure 1

In Figure 1 we see that for symmetric distribution, sample sizes of 60 and higher result in inclusion percentage within $95\% \pm 2\%$ and so for these sizes percentile method performs reasonably well. For non-symmetric distribution the method requires a sample size of 70 or more. Moreover we can see that almost for all the values (38 out of 40 data points) the inclusion percentage is below the desired value.

In the second part of the Simulation 1 we will restrict the simulation with the smaller sample sizes 4 - 42 for which the percentile method appears to be much less accurate.

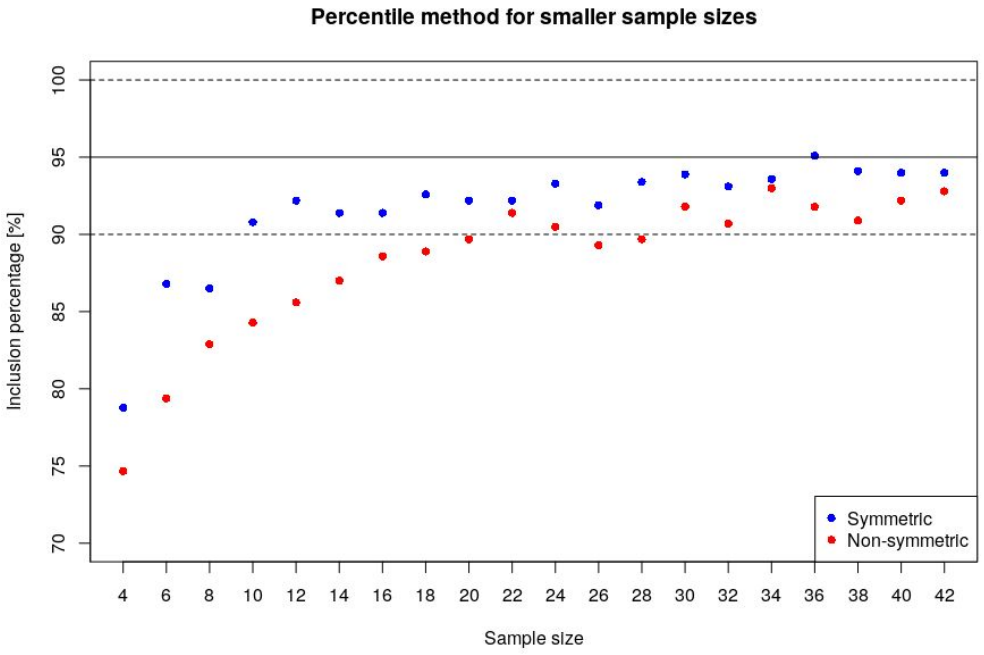

Figure 2

From Figure 2 we can make similar conclusions as from Figure 2. For the inclusion percentage $95\% \pm 5\%$ we require the sample size to be 10 or more and 30 or more for the symmetric distribution and non-symmetric distributions respectively.

Finally the real difference between inclusion percentage for symmetric and non-symmetric distributions using a t-test for the sample sizes 4 - 42.

$$H_0 : \mu_{sym} - \mu_{non.sym} = 0$$
$$H_1 : \mu_{sym} - \mu_{non.sym} \neq 0$$

```
        Paired t-test

data:  incl.sym and incl.non.sym
t = 8.0307, df = 19, p-value = 1.583e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 2.445477 4.169523
sample estimates:
mean of the differences
                 3.3075
```

Figure 3 - T-test for difference in percentile method accuracy with symmetric and non-symmetric distribution

From Figure 3 we see that the 95% confidence interval for the difference of mean inclusion percentage for symmetric distribution minus mean inclusion percentage for non-symmetric distribution is (2.7, 4.65) which indicates that there is a difference in performance of the percentile algorithm for symmetric and non-symmetric distribution and the percentile method seems to perform better on symmetric distribution.

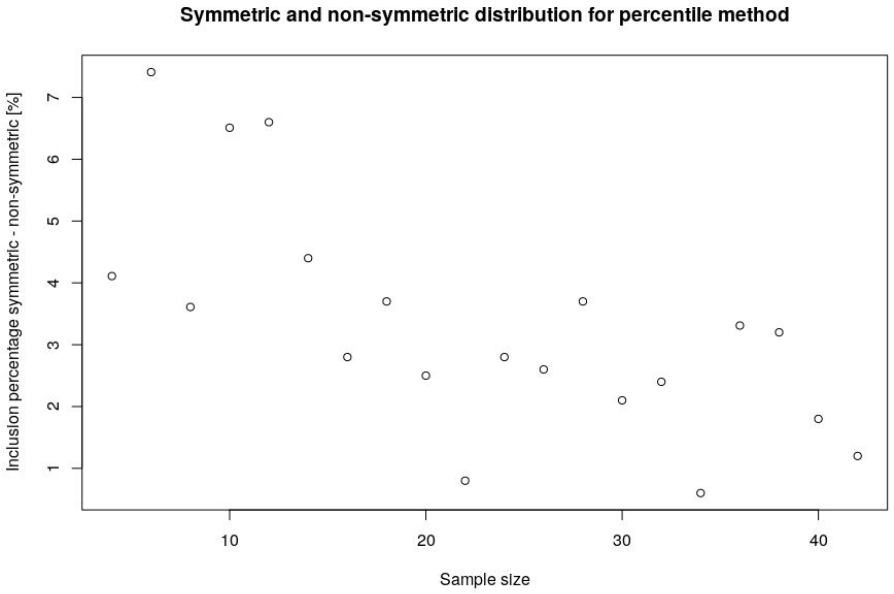

Figure 4

Figure 4 shows a decreasing trend of the difference between inclusion percentage on symmetric and non-symmetric distribution with increasing sample size. This acts as an indication that non-symmetric distribution has a greater effect on accuracy of percentile method for smaller sample sizes.

2. Comparison of percentile method and BCa method

As suggested in introduction BCa method should act as an enhancement to Percentile method. In this simulation we will explore how bias-reduction and acceleration terms can improve performance of the percentile method for symmetric and non-symmetric distributions. Since in simulation 1 the sample sizes 4 - 42 resulted in a much less accurate inclusion percentage we will restrict the simulations on those.

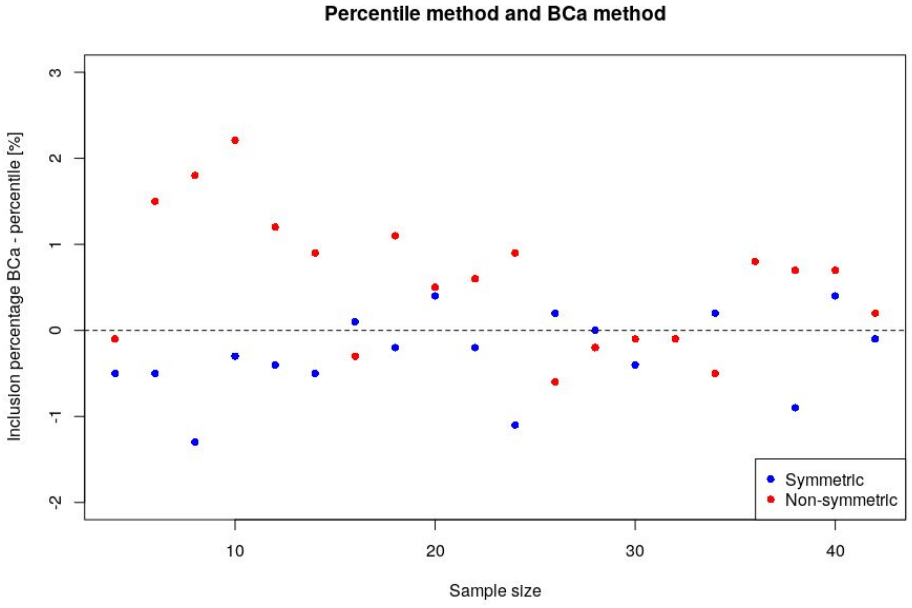

Figure 5

In Figure 5 we plot the difference between inclusion in percentage BCa - percentile method for symmetric and non-symmetric distribution. Most of the red points lies above the horizontal line indicating a better performance of BCa method for non-symmetric distribution. Most of blue points lie below the horizontal line indicating better performance of percentile method for symmetric distribution. To measure the true difference we perform a t-test.

$$H_0 : \mu_{sym} - \mu_{non.sym} = 0$$
$$H_1 : \mu_{sym} - \mu_{non.sym} \neq 0$$

```
        Paired t-test

data:  non.sym.diff and sym.diff
t = 3.179, df = 19, p-value = 0.004942
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.26662 1.29438
sample estimates:
mean of the differences
                 0.7805
```

Figure 6 - T-test for difference in accuracy between BCa and percentile method for symmetric and non-symmetric distribution

The 95% confidence interval for the true difference is $(0.48, 1.36)$ which indicates that BCa method is indeed more beneficial compared to percentile method when the distribution is non-symmetric rather than symmetric. For a symmetric distribution the usage of the BCa method seems to result in a less accurate inclusion percentage.

3. Comparison of non-studentized and studentized pivotal methods

In this simulation we analyse the performance of pivotal methods. We compare the performance of studentized and non-studentized versions of the algorithm for both symmetric and non-symmetric distribution.

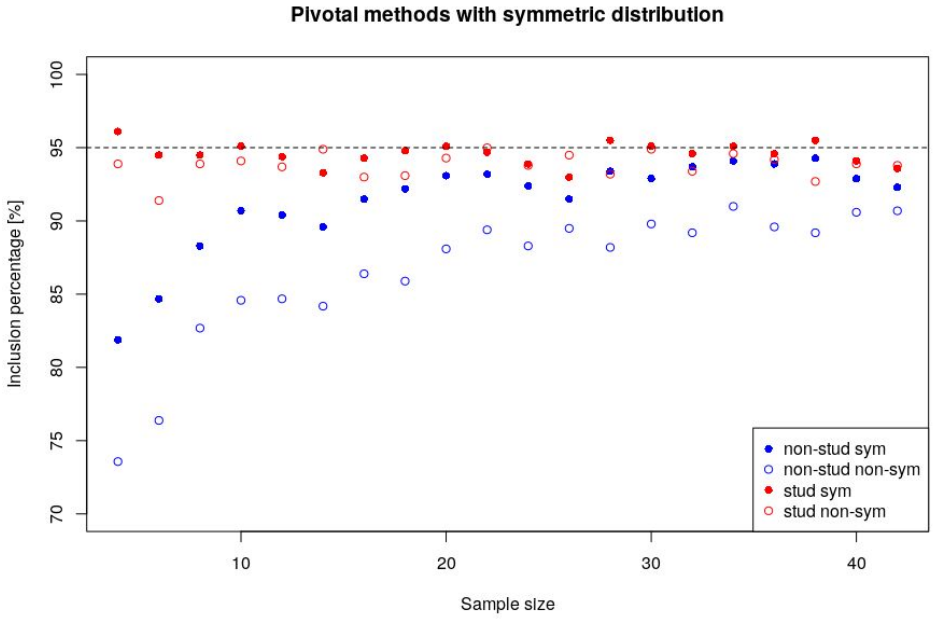

Figure 7

Figure 7 shows that especially for small sample values, full circles occur above empty circles. For larger sample values they occur closer together. Therefore similarly as with percentile method, the method is more accurate when the underlying distribution is symmetric. Red circles are above blue circles for both full and empty circles and so for both symmetric and non-symmetric distribution, the method is more accurate when it is studentized, reaching very close to 95% even for very small sample sizes.

4. Comparison of percentile and pivotal methods

In this simulation we compare the performance of non-studentized and studentized pivotal methods for symmetric and non-symmetric distributions.

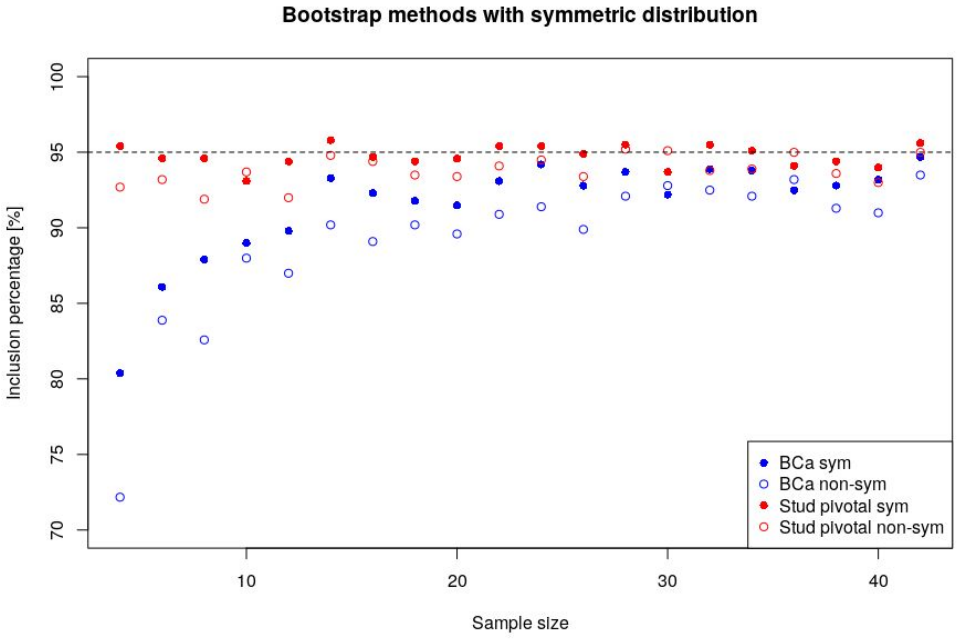

Figure 8

Figure 8 shows red points above blue points for both symmetric and non-symmetric distribution. Therefore the non-studentized pivotal method performs more accurately. Especially for small sample sizes the studentized pivotal method seems more suitable.

Discussion

First we showed that the percentile method becomes more accurate with increased sample size. This is probably because with more data points the sample is a better reflection of what the underlying distribution is. To obtain inclusion percentage $95\% \pm 5\%$ that we need a sample size of 10 or more and 20 or more for symmetric and non-symmetric distribution respectively. Percentile method is less accurate for non-symmetric distribution but the difference diminishes with increasing sample size. This difference can probably be explained by the fact that in non-symmetric distribution the standard error varies more when sampled, resulting in a non-symmetric shape of $\widehat{\Theta}$.

In second and third simulation we showed the for both percentile and non-studentized pivotal methods, the effect of non-symmetry of $\widehat{\Theta}$ can be mitigated by using the more advanced versions of the methods, BCa and studentized pivotal respectively. In second simulation

however we showed that for symmetric distributions, BCa sometimes performs slightly worse than standard percentile method which might be because of its excessive complexity.

Finally we compared the accuracy of BCa method and studentized pivotal method. Studentized pivotal method turned out to be more accurate than BCa for both symmetric and non-symmetric distribution. However for large sample sizes the difference diminishes. Perhaps we could conclude that the way that studentized pivotal method accounts for non-symmetry of $\widehat{\Theta}$ is more effective than that of BCa. Moreover the assumptions that the studentized pivotal method makes are probably more applicable for our data.

Conclusion

We found that for percentile method to perform reasonably well we need at least a sample size of 10 data points. As expected both percentile and non-studentized pivotal methods perform worse when applied to data from non-symmetric distribution. We showed that this can be mitigated by the more advanced versions of both methods BCa and studentized pivotal respectively. We concluded that this is because of the non-symmetry of the distribution of sample means however more investigation should be done towards this direction. For example the shape distribution of sample means could be investigated for various non-symmetric and symmetric distributions. Finally we also showed that studentized pivotal method achieves better accuracy than BCa method even for very small sample sizes. This is likely because the underlying assumptions that the studentized pivotal method makes is more suitable to our data.

References

1. Efron B, Tibshirani RJ: An Introduction to the Bootstrap, 1993.
2. Carpenter J, Bithell J:, Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians, 2000.
3. Lei S, Smith MR: Evaluation of Several Nonparametric Bootstrap Methods to Estimate Confidence Intervals for Software Metrics, 2003.
4. Singh K, Xie M: Bootstrap: A Statistical Method, 2010.