Comparisons between the percentile and bootstrap-t interval estimation methods.

Joseph Free

Department of Mathematics, University of North Florida

August 17, 2020

Abstract

In this report we consider two approaches to estimating confidence intervals about the mean of an artificial data set. The methods compared are the percentile and bootstrap-t methods. These methods are used to estimate confidence intervals using $B = 500$ bootstrap samples, and the estimated intervals are compared to their theoretical equivalents. Source code is included with this report in the appendix along with annotations.

1 Introduction

The Bootstrap is a popular and highly powerful computational statistical tool. Originally developed by Efron [ET93, CB01] in the 70s, the bootstrap has found widespread use in both parametric and non-parametric estimation problems. The general idea of the Bootstrap is that one can emulate the process of sampling additional data by repeated resampling from the original data with replacement, meaning observations can be repeated. To estimate a parameter θ a large number of bootstrap samples are collected and for each sample an estimate $\hat{\theta}^*$, called a bootstrap replicate, is calculated. From these bootstrap replicates, different methods can be employed to arrive at more general and accurate estimates for θ. The real power of the bootstrap is that the θ of interest can generally be any quantity calculated from the data including correlation coefficients and standard errors.

In this report, we consider the bootstrap procedure for estimating confidence intervals using the *percentile* and *bootstrap-t* methods.

1.1 Percentile Method

Confidence intervals are constructed using the percentile method by first generating a set of bootstrap replicates $\theta^*(b)$ for $b = 1, \ldots, B$ and determining their empirical distribution function $\hat{F}$. The $\frac{\alpha}{2}$th and $(1 - \frac{\alpha}{2})$th percentiles $\hat{\theta}_{\frac{\alpha}{2}}, \hat{\theta}_{(1-\frac{\alpha}{2})}$ are then found. By definition, $100 \times (1 - \alpha)\%$ of the data must be between these points. Thus, the interval is taken to be $\left(\hat{\theta}_{\frac{\alpha}{2}}, \hat{\theta}_{(1-\frac{\alpha}{2})}\right)$

1.2 Bootstrap-t Method

To generate confidence intervals using the bootstrap-t method, bootstrap replicates are first studentized and used to generate a studentized replicate distribution. This distribution is used to obtain critical values similar to those in a regular t-table to construct the desired interval.

More specifically, B bootstrap samples are taken and replicates $\hat{\theta}^*(b)$ obtained. For each sample,

$$T^* = \frac{\hat{\theta}^*(b) - \hat{\theta}}{s_{\hat{\theta}}(b)}, b = 1, \ldots, n$$

is calculated, where $s_{\hat{\theta}}(b)$ is the estimated standard error for the bth sample. The $\frac{\alpha}{2}$th and $(1 - \frac{\alpha}{2})$th percentiles $t^*_{1-\frac{\alpha}{2}}$,$t^*_{\frac{\alpha}{2}}$ are found and used to construct the estimated interval $\left(\hat{\theta} - s_{\hat{\theta}} t^*_{1-\frac{\alpha}{2}}, \hat{\theta} - s_{\hat{\theta}} t^*_{\frac{\alpha}{2}}\right)$

2 Problem

We are given the data from 15 students in an intensive statistical computing class on how much their last haircut cost (table 1). Using the percentile and bootstrap-t procedures, we wish to obtain a 90% confidence interval about the mean haircut cost and compare the differences between the two methods.

Table 1: Haircut Data

Student	Price	Student	Price
1	65	9	20
2	32	10	18
3	12	11	40
4	20	12	24
5	60	13	22
6	14	14	25
7	16	15	25
8	21		

3 Method

In what follows, we take $\hat{\theta} = \bar{x}$. Algorithms for each method can be found in the Appendix, section 1.

To analyze the data, for both the percentile and bootstrap-t methods, $B = 500$ bootstrap samples and sample mean replicates were generated. In order to isolate the effects of the individual procedures, the methods were applied on *identical samples* from the original data. That is, the same sample of bootstrap replicates $\hat{\theta}(b)$ used for the percentile method was also used for the bootstrap-t method. This guaranteed a consistent experimental unit in our analysis and allowed for a fair comparison

Figure 1: Distribution of original data

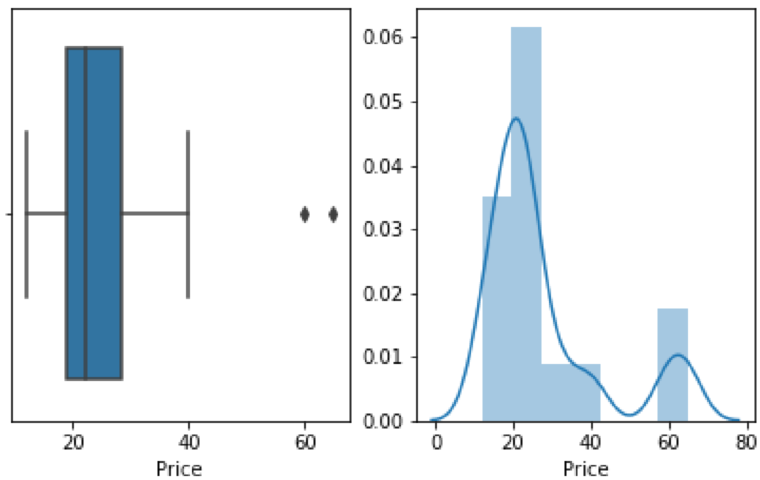

Figure 1: Distribution of the original 15 data points. *Left:* Box plot of the data. Here it is apparent there are two outlying data points, and a slight skew right. **Right:** Histogram of the data with a KDE overlay. Here the outlying points can be seen as a second bulge to the right.

between the methods by eliminating potential noise introduced by using different samples for each method. The 90% confidence intervals for each are found, and distribution plots are obtained with the intervals overlayed. Additionally, the empirical coverage probability for each interval is obtained. For the purposes of this report, we have taken the empirical coverage probability to be

$$\hat{P}(\hat{\theta}_{lo} \leq \theta \leq \hat{\theta}_{hi}) = \frac{\#\{\hat{\theta}(b) \in [\hat{\theta}_{lo}, \hat{\theta}_{hi}]\}}{B}$$

where $\hat{\theta}_{lo}$ is the lower limit of the estimated bootstrapped confidence interval and $\hat{\theta}_{hi}$ is the upper limit.

Because the parameter of interest is the mean for this problem, and the standard error and sampling distribution of $\bar{x}$ is well known, we can compare the results obtained to conventional theory to see is there are any significant departures. Thus, both bootstrapped intervals are compared with the approximate theoretical confidence intervals one might calculate by hand: specifically the normal 90% confidence interval $(\bar{x} - 1.645\frac{s}{\sqrt{n}}, \bar{x} + 1.645\frac{s}{\sqrt{n}})$ and the student-t confidence interval calculated using 14 degrees of freedom $(\bar{x} - 1.761\frac{s}{\sqrt{n}}, \bar{x} + 1.761\frac{s}{\sqrt{n}})$. The theoretical intervals are compared to the bootstrapped intervals graphically. In fact, we expect from the central limit theorem that the distribution of the bootstrapped replicates approaches a normal distribution for larger bootstrap samples. Hence, one barometer of the performance of the bootstrapped intervals will be how much they agree with the normal 90% interval.

To analyze the general behavior of the methods, bootstrapped intervals are also estimated for varying sample sizes $B = 50, 250, 500, 1000, 10000$ and the results are plotted for each procedure.

Lastly, the behavior of the procedures under a transformation is considered. The original data are converted to logarithmic scale, the bootstrap methods are then applied to the log-data to obtain intervals in transformed scale, and the end-points are then remapped to the original scale. The output is then compared to that of the initial estimates.

4 Results

The distribution of the original data can be seen in Figure 1. Both the boxplot and histogram indicate the presence of two outlying observations in the data. With the exception of these two observations, the data is otherwise fairly well behaved. The mean of the original data is $\bar{x} = \$27.6$.

The initial results of the first round of bootstrapping is summarized in Figure 2, which contains the distribution of the $B = 500$ bootstrap replicates, the distribution of the bootstrap-t statistics, the overlayed estimated confidence intervals for the methods, and the empirical coverage probability. Here it can be seen that distribution of replicates is almost symmetric, and that the percentile method yields an interval that is nearly centered about the sample mean $\bar{x} = \$27.6$. Conversely, the bootstrap-t interval is highly asymmetric with a long right tail. The distribution of T^* statistics gives an indication on why this is the case: it has an incredibly long left tail stretching as far out as -12 on the T^* axis. It can also be seen that the percentile method covers approximately 90% of the distribution as we should expect, but the bootstrap-t method is actually covering nearly 95%. See table 2 for the estimated intervals corresponding to Figure 1. Note the drastic difference in the right end points.

Table 2: Bootstrapped Confidence Intervals

Method	LCL ($)	UCL ($)	Coverage
Percentile	21.333	34.933	90.2%
Bootstrap-t	21.582	41.393	94.5%
Δ	0.249	6.460	4.3%

Table 3: Theoretical Confidence Intervals

Method	LCL ($)	UCL ($)
t-Interval	20.417	34.783
Normal interval	20.890	34.309

In Figure 3, the intervals from table 2 are given alongside 1) the 90% normal confidence interval and 2) the 90% student-t interval that one might calculate from the original data. It is immediately evident from the plot that the percentile interval is a decent approximation of the normal interval; whereas the bootstrap-t interval only appears to agree with the left end point of the normal interval, and departs from the right end point. These same observations also hold for the student-t interval as it is surprisingly close to the normal interval.

To assess the stability of the methods, bootstrap samples of size $B = 50, 250, 500, 1000, 10000$ were made, and the procedures applied. The results are seen in Figure 4, were it can be seen that the

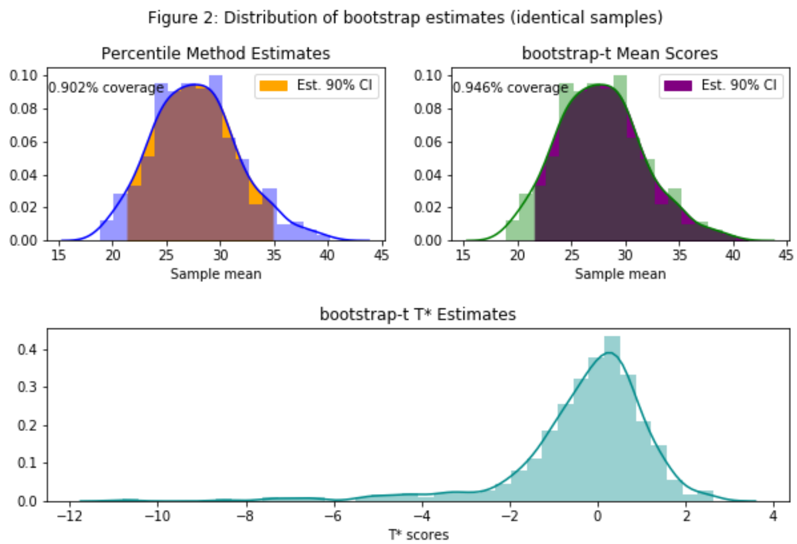

Figure 2: *Left*: Estimated confidence interval using percentile method and empirical data coverage. *Right*: Estimated confidence interval using bootstrap-t method and empirical data coverage. *Bottom:* Distribution of the studentized bootstrap replications. Note the very severe skew left.

---

percentile method exhibits a surprising level of stability for varying sample sizes. That is, for each distribution of replicates, the interval appears to be centered very nearly symmetrically about the mean of the distribution. The bootstrap-t distribution on the other hand appears to behave as before with elongated tails. It also appears the empirical coverages follow a similar pattern: the percentile method consistently covers 90% of the distribution, while the bootstrap-t method fluctuates.

Finally, in Figure 5 the results of applying the bootstrap to transformed data are given. Here the natural logarithm was applied to the original data and the log-data was bootstrapped as before. Confidence intervals for the log data were constructed then transformed back to the original scale by exponentiation. In the Figure, the normal 90% confidence interval is given along with the transformed bootstrap intervals and empirical coverages. It can be seen that in this situation, the percentile interval no longer agrees well with the normal interval, and actually covers less of the distribution. However, the bootstrap-t method now appears nearly symmetric about the mean with the end points of the interval close to the normal end points.

5 Discussion

The results above seem to indicate that, while the goals of the percentile and bootstrap-t methods are similar, the interval estimates they provide differ from one another. Most notable is that the percentile method seems to be stable in its coverage of the bootstrap distribution, whereas the bootstrap-t method has a blown out tail.

The instability of the bootstrap-t method is actually a result of the outlying observations 60 and

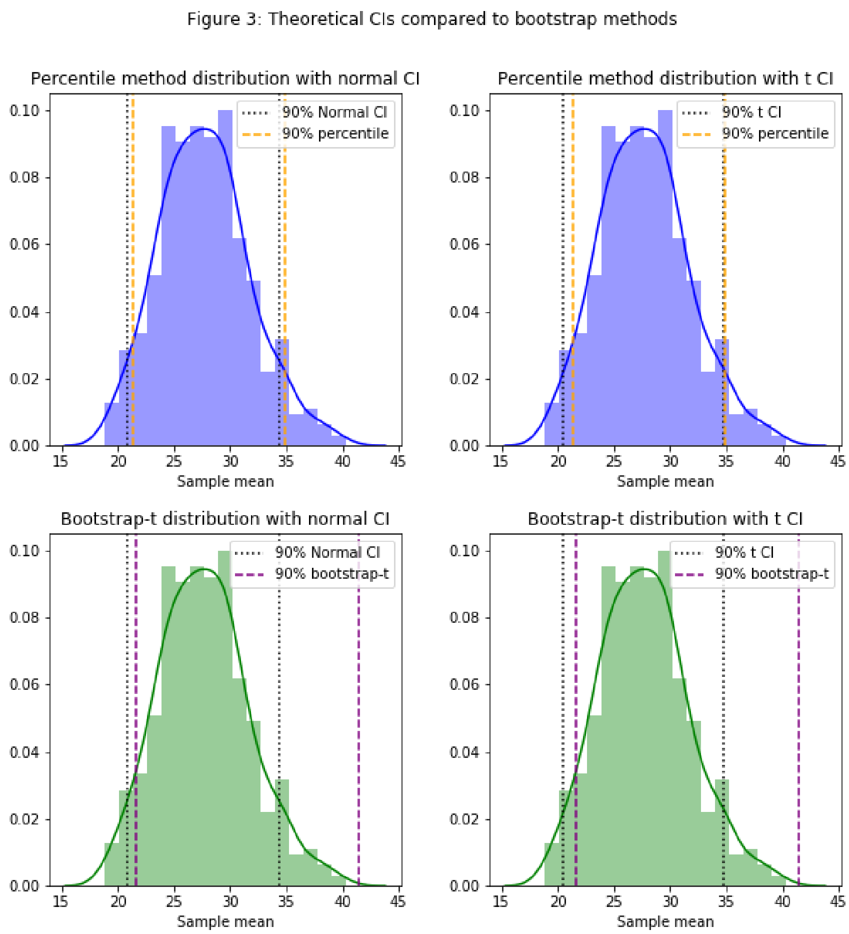

Figure 3: *Top*: Percentile method with theoretical normal and t confidence intervals. *Bottom*: Bootstrap-t method with theoretical normal and t confidence intervals. Note that dashed lines indicate bootstrapped intervals while the dotted lines indicate theoretical intervals.

---

65 in the original data. As it turns out, the bootstrap-t is highly sensitive to outlying observations and variability in the data, and generally only performs well under ”nice” conditions or variance stabilizing transformations [ET93]. However, this is actually both a strength and weakness of the bootstrap-t method, as the resulting increase in coverage also increases the chances of covering the true population parameter. Meanwhile, the percentile method is generally robust to distributional quirks in the data because the confidence limits are determined strictly by the empirical percentiles of the bootstrap distribution: there is no calculation or dependence on pivotal quantities like the bootstrap-t's studentized statistic. In practice, though, the percentile method has been known to underestimate the true coverage probability. [ET93, EH16].

6 Conclusion

The percentile and bootstrap-t methods are the two simplest methods for estimating confidence intervals. As such these methods like other computational methods have their relative strengths and limitations. As seen above, the percentile method appears to agree with the standard normal interval, but at the potential cost of miscovering the data by not incorporating information about the underlying distribution. Contrast this with the bootstrap-t interval, which appears to drastically overestimate the data in the presence of outlying observations.

In most applied situations, the choice of which method to use is often depends on the context of the problem. So there is no way to definitively state that one method is superior to the other *in general*. But in this circumstance we know the sampling distribution of the sample mean. Thus, we should expect that the superior method should be the one that agrees most with prevailing statistical theory. In this light it appears that the percentile method performed better than the bootstrap-t in that the interval estimates it provided were consistent with both the t-interval one might have calculated using the original 15 data points and the large sample normal confidence interval one might use when considering the distribution of the bootstrap replicates.

References

[CB01] George Casella and Roger Berger. *Statistical Inference*. Duxbury Resource Center, June 2001.

[EH16] Bradley Efron and Trevor Hastie. *Computer Age Statistical Inference: Algorithms, Evidence, and Data Science*. Cambridge University Press, USA, 1st edition, 2016.

[ET93] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA, 1993.

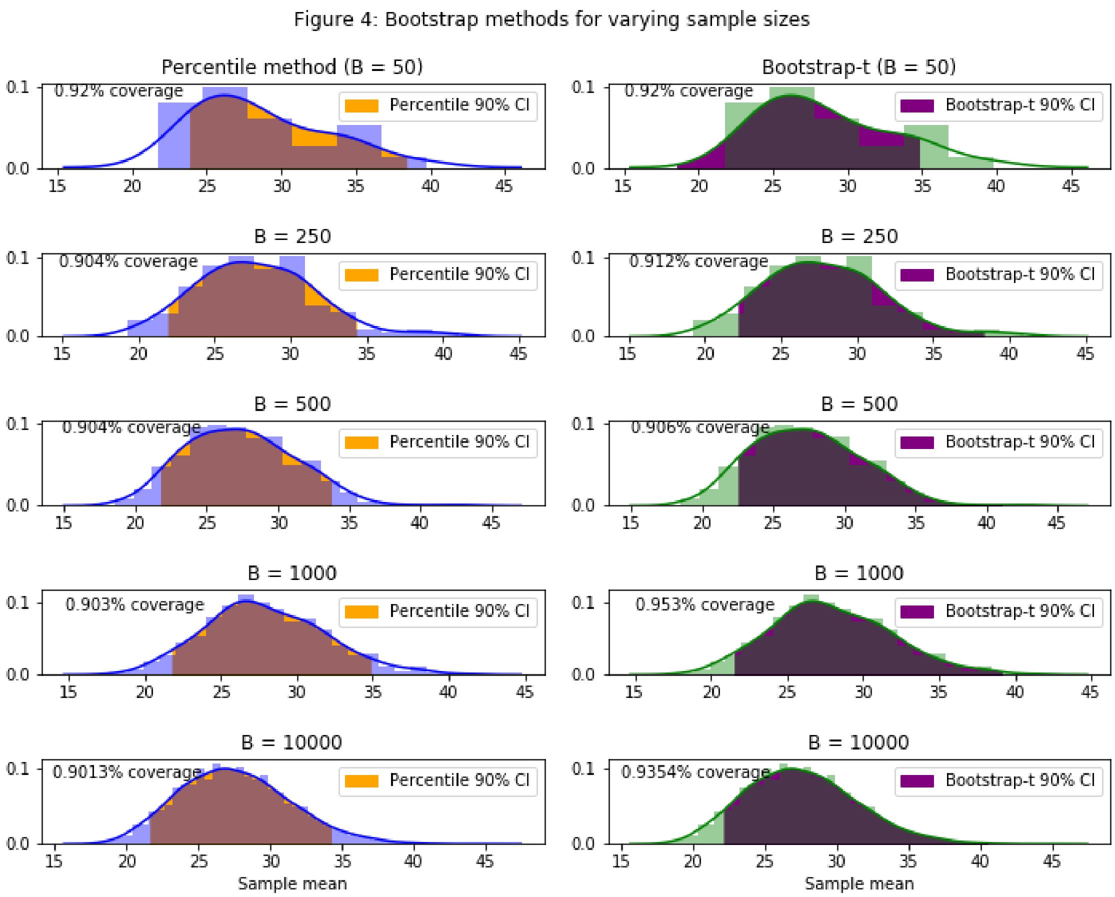

Figure 4: Bootstrap replication distributions for varying sample sizes. $B = 50, 250, 500, 1000, 10000$. To the left, the percentile method and accompanying CIs are given. To the right, the bootstrap-t method. Note the asymmetry in the bootstrap-t interval compared to the percentile method. Empirical coverages are given in the upper lefthand corners of each subplot.

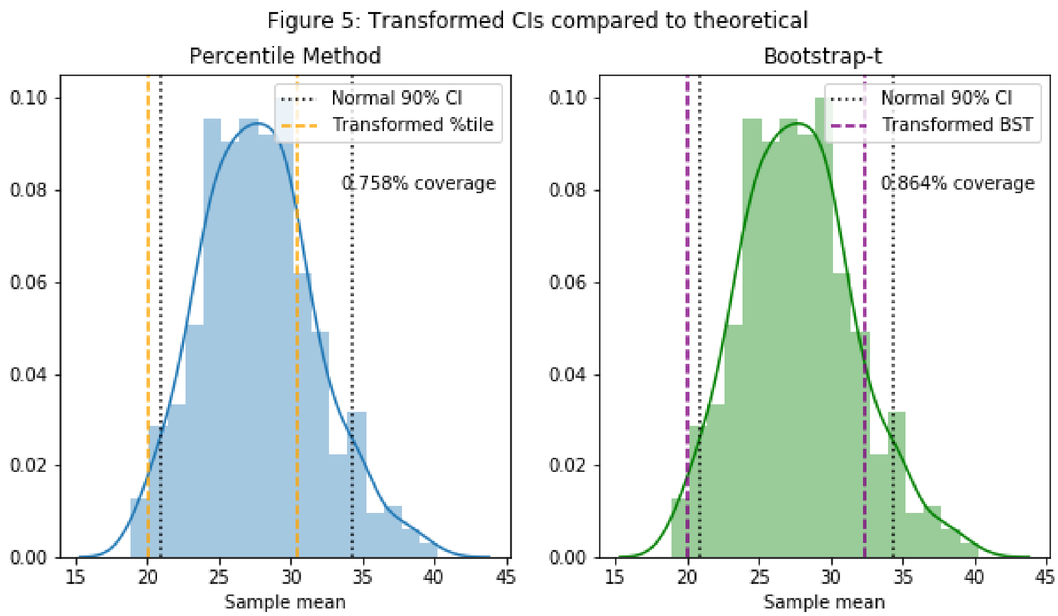

Figure 5: Percentile and bootstrap-t methods following a transformation. The original data was subject to a variance stabilizing transformation (natural logarithm), 500 bootstrap samples were taken, and the two methods used on the log data. The endpoints of the intervals were then remapped to the original scale with the exponential function.