Bootstrapping Inference

Tom Blain

Data Science Portfolio A - Section I

Abstract

Bootstrapping, first introduced by Bradley Efron in 1979 [1], is a powerful tool which allows us to conduct inference about an unknown population from sample data by iterative resampling with replacement of the sample data. In this paper we will provide an introduction to bootstrapping techniques, introduce percentile and Bias-corrected accelerated (BCa) confidence intervals from the bootstrap distribution and conclude with an example on real world data bootstrapping and producing confidence intervals for a mean statistic given a sample.

1 Motivation

Bootstrapping is a widely used statistical method cited across all areas of scientific study. In the lectures of this course, it is mentioned as an option for resampling tests to estimate variance of parameter estimates, however, use cases do not seem clear cut and it seems like a promising method, lacking context. It is therefore of importance to further study this technique, to be able to confidently implement this tool in areas where resampling can be powerful as classical methods may not be applicable. It is harder than needed to be able to formally justify the use of this method in a context, and feel confidence that the results can be trusted.

A topic note on bootstrapping will allow us to have a much more rigorous understanding of this method, and give a base for sampling methods in general. In practical data science, good datasets will rarely exist - bootstrapping methods are most powerful when handling small sample sizes, which is common when data is hard to gather, or expensive. It will be inevitable that this situation will arise as we progress in our studies.

A starting point will be to study Efrons original paper of the proposed method, before moving further into modern applications and refinements made on the algorithm. There is therefore an overlap in many research fields where statistical testing might be used and studied. We might find the best approach will be to combine insights from many areas to best answer the proposed question on the justification for bootstrapping. Combining a mathematical and practical approach will best enable us to understand and test the benefits and drawbacks of the method, and feel confidence applying this method in our own models.

2 Introduction

Boos(2003) [2] describes the popularity of the bootstrap as follows: 'In this world, the data analyst can obtain any quantity of interest by simulation. For example, if the variance of a complicated parameter estimate in this world is desired, just computer generate B replicate samples (bootstrap samples or resamples), compute the estimate for each resample and then use the sample variance of the B estimates as an approximation to the variance...In effect this bootstrap world simulation approach opened up complicated statistical methods to anybody with a computer and a random number generator. Random variable calculus can be replaced by computing power.' We might find this to be an ambitious claim - we must be careful not to misuse statistical methods without being sure the selected method is appropriate with mathematical guarantees.

Since its introduction, the straightforwardness of the method combined with increasingly powerful computation has proved invaluable for estimates of standard errors or confidence intervals. Experiments may have large costs associated with sampling and in real world applications, the population dataset will rarely be possible to obtain.

A simple bootstrapping algorithm can be defined as follows
1. Obtain a sample of size n from the population data $\{x_1, x_2, ..., x_n\}$
2. Resample n data points at random from our original sample with replacement to obtain new dataset $\{x_1^*, x_2^*, ..., x_n^*\}$
3. Calculate the statistic on this resampled set $\hat{\theta^*}$
4. Repeat steps 2-3 for a large number of iterations

3 Bootstrap Confidence Interval

The idea behind the bootstrap is that if the original sample is representative of the population, then the bootstrap distribution of the mean will look approximately like the sampling distribution of the mean, that is, have roughly the same spread and shape. However, the mean of the bootstrap distribution will be same as the mean of the original sample, not necessarily that of the original population [3].

The bootstrap procedure may be used with a wide variety of statistics we can write a function to compute, with very little change in procedure. Allowing statistical inferences such as confidence intervals to be calculated even for statistics for which there are no straightforward formulas.

3.1 Percentile Intervals

Since we wish to be using a large number of iterations in the bootstrapping process, we can use ordered bootstrap statistics to find a confidence interval simply [4].

Let $\hat{\theta^*}_{(b)}$ represent the ordered bootstrap estimates, and suppose we wish to construct a $(100-a)\%$ confidence interval. If iterations i is large, then the $\alpha/2$ and $1-\alpha/2$ quantiles of $\hat{\theta^*}_{(b)}$ are approximately $\hat{\theta^*}_{(lower)}$ and $\hat{\theta^*}_{(upper)}$, s.t. $lower = i\alpha/2$, $upper = i(1-\alpha/2)$. if *lower* and *upper* are not integers, we can interpolate or round.

3.2 Bias-corrected and accelerated bootstrap interval (BCa interval)

The advantages to the BCa interval is that it corrects for bias and skewness in the distribution of bootstrap estimates [5]. The bias-correction refers to the adjustment made to the sample estimate to correct for the tendency of the estimate to be too low or too high when the sample size is small. Acceleration refers to the adjustment made to account for the skewness of the underlying distribution. The result of these adjustments is a confidence interval that can be more accurate, especially when the sample size is small or the underlying distribution is not normal.

A BCa (bias-corrected and accelerated) confidence interval can be computed using the following steps:

Obtain a sample of size n from the population and calculate the sample mean, $\bar{x}$. Using bootstrapping, generate B resamples of size n from the sample, and calculate the mean of each resample, $\bar{x}^*(b)$.

Calculate the bias-correction factor, z^*, where s is the standard deviation of the resampled means.

$$z^* = \frac{\bar{x}^* - \bar{x}}{s} \tag{1}$$

Calculate the acceleration factor, a^*, where n is the sample size.

$$a^* = -1 * \frac{z^* * (n-1)}{n + (z^*)^2} \tag{2}$$

Compute the lower and upper limits of the BCa interval:

$$\text{lower limit} = \bar{x} - z^* s(1 + \frac{1}{n} - a^*) \tag{3}$$

$$\text{upper limit} = \bar{x} + z^* s(1 - a^*) \tag{4}$$

The BCa interval can be represented mathematically as

$$\bar{x} \pm z^* s(1 + \frac{1}{n} - a^*) \tag{5}$$

for lower and upper limit respectively. [6]

4 Titanic example

This section is based on the code which is available to view at
[github.com/TBlainUoB/Bootstrapping-titanic-dataset]
alongside a pdf markdown document.

In this example we take the titanic dataset [7], and wish to produce a confidence interval for the average age of an adult passenger (≥ 18) from a small sample using bootstrapping techniques. Since we have access to the population data, we are able to evaluate performance for different sample sizes.

```
data <- subset(data, data$Age >= 18)
age_data <- data$Age
age_data <- age_data[!is.na(age_data)]
c("True mean", mean(age_data))
[1] "True mean"        "33.5831946755408"
```

This is the true population mean which we will now make unknown to our bootstrap model

```
n <- 50
sample_age <- sample(age_data, n, replace = FALSE)
c("Sample mean", mean(sample_age))
[1] "Sample mean"      "33.11"
```

Take a sample of size n=30

```
bootstrap <- function(sample_age, mean_subsamples){
  subsample <- sample(sample_age, n, replace = TRUE)
  mean_subsamples <- append(mean_subsamples, mean(subsample))
  return(mean_subsamples)
  }

iterations <- 10000
bootstrap_means <- c()
for(i in 1:iterations){
  bootstrap_means <- bootstrap(sample_age, bootstrap_means)
}
```

The bootstrap function takes our sample and subsamples with replacement n times where each mean from the subsample is appended to our meansubsamples vector. A histogram is created from the meansubsample vector with a red line as the true population mean. Figure 1

```
CI <- 0.95
orderedmean <- sort(bootstrap_means)
lower <- orderedmean[iterations*((1-CI)/2)]
upper <- orderedmean[iterations - (iterations*(1- CI)/2)]
bca <- bca(bootstrap_means, CI)
```

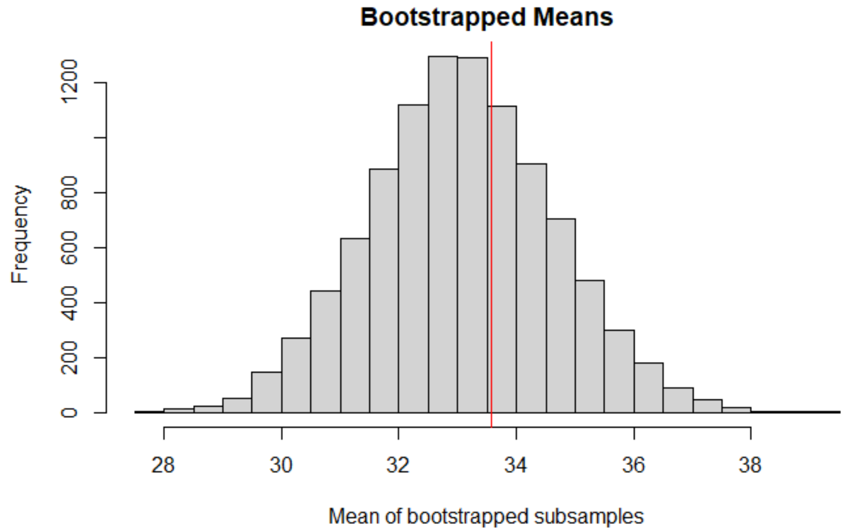

Figure 1: Histogram representing the bootstrapped estimates for the mean (n=50). True population mean in red.

```
[1] "Percentile CI"     "30.06"     "36.22"
[1] "BCa CI"            "30.12"     "36.3"
```

Here we have computed confidence intervals for both the quantile CI and BCa CI, which have been plotted on the histogram. As you can see, there is not too much difference between our intervals, this may be because of little bias and skewness on the sample. Figure 2 [8]

Our sample size was picked arbitrarily, Figure 3 produces a plot comparing the density of the bootstrapping algorithm on various n values. Results here are roughly as we would expect with low sample size producing a density with mean far from the true mean and higher variance. There will always be a trade off with collecting enough samples for a good result, and feasibility/cost in a real world scenario.

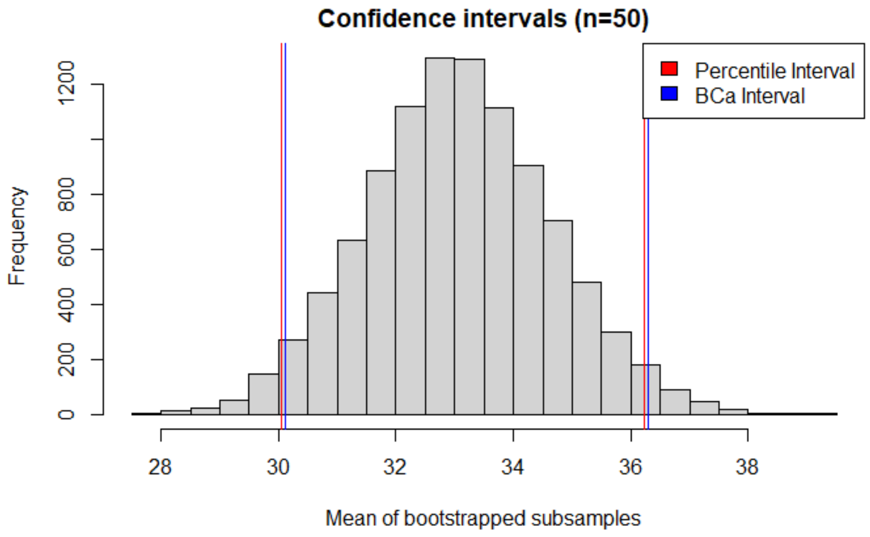

Figure 2: Percentile CI and BCa CI for bootstrap estimates of mean (n=50)

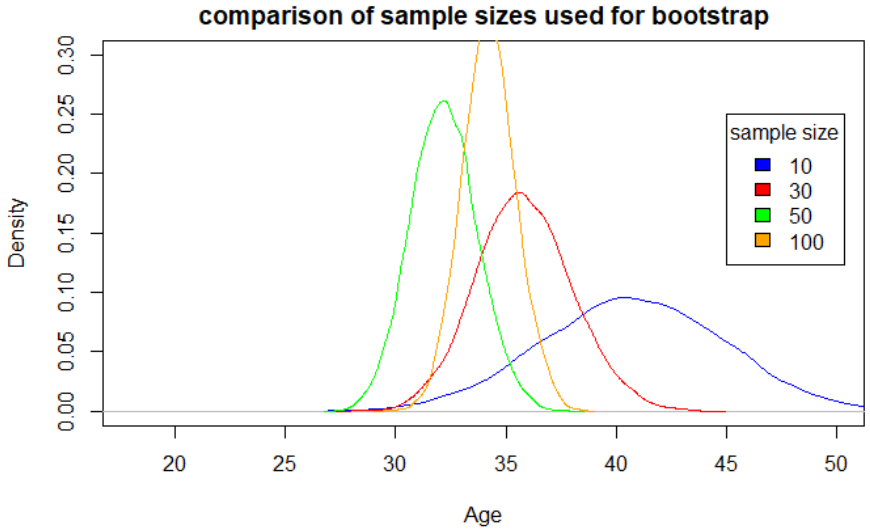

Figure 3: Various density plots of bootstrapping with different sample size n

4.1 Male/Female adult survival age analysis

To demonstrate the versitility of our bootstrapping method, let's now apply the same algorithm to some different data within the dataset.

We first filter the data by survived and split into male/female data sets, again removing data entries where age is N/A. We can then apply the same bootstrapping function ($n = 50$) to the new datasets and plot histograms 4.

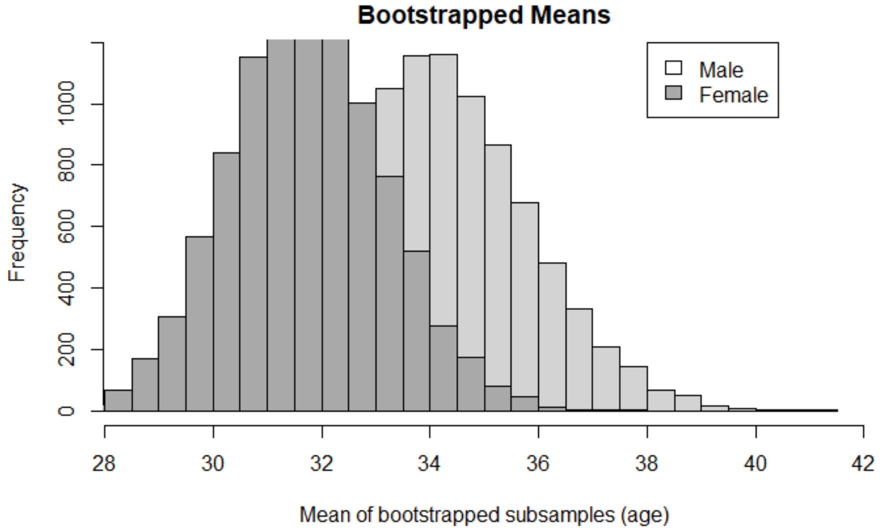

Figure 4: Male and Female survived bootstrap histogram of the mean statistic

We now compute confidence intervals for the male survival age bootstrapped data using both our quantile and BCa methods we introduced in section 2. 5.

There appears to be a large difference in the BCa and percentile intervals. This may indicate bias and/or skewness in the data. More research into the bias and skewness should be conducted.

Our analysis here doesn't actually tell us much more than we could have guessed about the dataset, however, the purpose of the exercise was to demonstrate the relative ease to apply a bootstrapping algorithm to a dataset and only very minor adjustments would have to be made to apply the same code to something else we might wish to study.

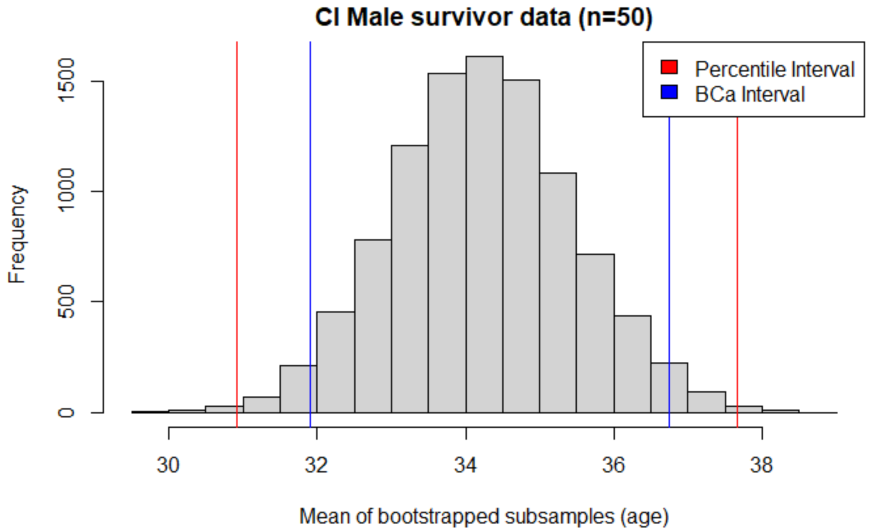

Figure 5: Quantile and BCa CI for Male survived passenger bootstrap

5 Reflection

We have demonstrated the ease of use of bootstrapping specific to generating confidence intervals for a statistic using a sample of a dataset. I am particularly impressed by the ease of implementation and seemingly powerful results (more work would be required to evaluate this). That powerful computation can do this method in seconds to approximate hard to calculate statistics serves a great purpose. However, I am still no clearer to being sure of the validity of results in more complex cases - the basic question I want to ask is what guarantee do we have that our results are good? We did not ask questions of assumptions required for validity - I would like to extend this project in the future to ask questions to the effectiveness and quality of the results bootstrapping provides, as well as studying developments and variations of the classic algorithm such as parametric bootstrapping. I hope this will give insight into appropriate use cases I can be confident applying these methods to.

It will be useful to continue coming across use cases in data science in my future studies, and read more practical applications in studies across a broad range of fields.

Many more further topics of bootstrapping are open to be studied and added to this portfolio, including bayesian bootstrap, smooth bootstrap, parametric bootstrap, as well as further techniques to improve computational efficiency. Bagging, also known as bootstrap aggregating, is a machine learning algorithm based off the concepts of bootstrapping, which relates closely to random forests.

References

[1] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1 – 26, 1979. [Online]. Available: https://doi.org/10.1214/aos/1176344552

[2] D. D. Boos, "Introduction to the bootstrap world," *Statistical Science*, vol. 18, no. 2, pp. 168–174, 2003. [Online]. Available: http://www.jstor.org/stable/3182846

[3] B. Efron and R. Tibshirani, "An introduction to the bootstrap," 1993.

[4] J. Fox, "Bootstrapping regression models 21.1 bootstrapping basics," 2002. [Online]. Available: https://www.sagepub.com/sites/default/files/upm-binaries/21122_Chapter_21.pdf

[5] R. Wicklin, "The bias-corrected and accelerated (bca) bootstrap interval," 2017. [Online]. Available: https://blogs.sas.com/content/iml/2017/07/12/bootstrap-bca-interval.html

[6] T. DiCiccio and R. Tibshirani, "Bootstrap confidence intervals and bootstrap approximations," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 163–170, 1987. [Online]. Available: http://www.jstor.org/stable/2289143

[7] Kaggle, "Titanic dataset," 2012. [Online]. Available: https://www.kaggle.com/competitions/titanic/data

[8] coxed (version 0.3.3), "bca: Bias-corrected and accelerated confidence intervals," 2020. [Online]. Available: https://www.rdocumentation.org/packages/coxed/versions/0.3.3/topics/bca