The Mathematics Enthusiast

Volume 18
Number 1 *Numbers 1 & 2*

Article 8

1-2021

The History of Bootstrapping: Tracing the Development of Resampling with Replacement

Denise LaFontaine

Recommended Citation

LaFontaine, Denise (2021) "The History of Bootstrapping: Tracing the Development of Resampling with Replacement," *The Mathematics Enthusiast*: Vol. 18 : No. 1 , Article 8.
DOI: https://doi.org/10.54870/1551-3440.1515
Available at: https://scholarworks.umt.edu/tme/vol18/iss1/8

The History of Bootstrapping: Tracing the Development of Resampling with Replacement

Denise LaFontaine[1]

University of Montana- Missoula

Abstract: Sampling is one of the most fundamental concepts in statistics, as the quality and accuracy of the statistical inferences made, heavily depend on the method used to obtain the sample and the sample's ability to represent the population of inference. Despite being a simple concept, sampling presents researchers with many challenges. Generally, due to monetary and time constraints, researchers must take a smaller sample size than they would ideally use. Using statistics from these small samples, estimates for population parameters can be made, typically in the form of a confidence interval. However, the validity of these confidence intervals depends on three basic assumptions that are difficult to meet with small sample sizes. This paper traces the development of the sampling method known as bootstrapping that helps small samples to meet these assumptions. The paper touches on previous methods used before the development of bootstrapping and shows how bootstrapping has evolved over the last four decades and become widely used in the field of statistics.

Keywords: variation, resampling with replacement, parameter, empirical distribution, jackknifing, bootstrapping

Bootstrapping was developed in the 20th century by Bradley Efron, an American statistician (Efron, 1979). This method assumes that the sample has the same relationship to the population as it has to an empirical distribution that is created by resampling with replacement from the original distribution N samples of the same size as the original sample. By creating this empirical distribution and comparing the sample statistic to it, the researcher can gauge the accuracy of the inferences on the population parameter. Over the last four decades, bootstrapping has become widely used and has been expanded to include various types of bootstrapping, such as parametric

[1] denise.lafontaine@umconnect.umt.edu

and Bayesian bootstrapping. This paper examines this evolution of the bootstrap resampling method, focusing more on its conception and tracing it to its modern statistical use and some of its current variations.

To understand the evolution of bootstrapping and the concept itself, a look into the history of sampling is necessary, as it is the foundation of the statistical method. Looking through history, it is not clear as to when sampling was first used. According to the American Statistical Association's timeline on statistics, sampling was used as far back as the 5th century in the Peloponnesian War when soldiers were selected to count the number of bricks that made up the height of the wall surrounding the areas their army was planning to attack (n.d). The counts these soldiers came up with created a sample, from which the mode was selected and used to calculate the total height of the wall. The first physical evidence of a sample dates back to 2 C.E. and is actually a complete sample of the population—also known as a census—of the Han Dynasty (American Statistical Association [ASA], n.d.). Samples gradually became more common as time went on, and people developed ways to improve them. An example of this comes from the United Kingdom in 1150 C.E. when the Trial of the Pyx began. In order to test that the coins being produced by the Royal Mint met the compositional standards, coins were selected randomly as they were minted and tested to see if they had the correct weight and composition (ASA, n.d.). The randomness of the selection ensured that the sample of coins was more representative of the whole population since coins minted on various days and at various times would be a part of the sample. This led to what is now typically considered the best method of sampling—simple random sampling. Other variations of sampling have been developed as well for specific cases. However, the goal of all of these methods is the same—accurately represent the population of inference.

Once a representative sample is obtained, inferences can be made. The accuracy of these inferences depends upon three assumptions being made (Graham, 2018). The assumption of normality of the sampling distribution of the parameter is the first of these assumptions. The second assumption is that the standard error of the estimated parameter is a close estimate to the standard deviation of the sampling distribution of the estimated parameter. The last of the assumptions is the estimated parameter has little bias in its estimate. For some parameters, these assumptions can be met relatively easily, while other parameters require a different set of methods in order to meet these assumptions. The median is an example of the latter while the mean would be an example of the former.

Looking at the mean μ as a population parameter, the first assumption can typically be met by invoking the Central Limit Theorem. The Central Limit Theorem argues that although observation themselves may not be normally distributed, the means of the observations will follow a distribution close to normal if the number of observations is large enough (typically greater than 30) (Graham, 2018).

Central Limit Theorem: $If\ n \geq 30, then\ \bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

$where\ \bar{x}\ is\ the\ mean\ of\ the\ sample, \mu\ is\ the\ population\ mean, \sigma\ is\ the\ population\ standard\ deviation, and$

$n\ is\ the\ sample\ size.$

The second assumption can be met as well when working with means because a closed form expression exists that shows that the standard error for the mean can be approximated by $\frac{\sigma}{\sqrt{n}}$. If σ is unknown, as it typically is, it can be approximated by $\frac{s}{\sqrt{n}}$ where s is the standard deviation of the sample as long as one uses the t-distribution rather than the z-distribution or n is sufficiently large. The third assumption is known to be true by the Central Limit Theorem as long as the sample size is greater than or equal to 30 since the sampling distribution of the sample statistic is centered around the population parameter μ.

As just shown, these assumptions are met when making inferences on the population mean. Other statistics, however, cannot meet these assumptions as easily. For example, when working with the sample's trimmed mean $\bar{y}_{TM}$[2] to estimate the population mean or the sample median m to estimate the population median M, issues arise in meeting the necessary assumptions. This is troublesome because the validity of inferences made on these statistics depends on the assumptions being met. Unlike the mean, these parameters do not have something like the Central Limit Theorem to establish normality in their sampling distributions. Similarly, no closed form expression exists to give the standard error of the estimates or the standard deviation of the sampling distribution to compare it to and unbiasedness cannot be easily established. It is for inferences on parameters such as these that separate methods must be used to argue the assumptions are met and therefore, the inferences are accurate and valid.

Bootstrapping was one of the methods developed for these types of cases. However, it was not the first. Rather, bootstrapping developed as an expansion and improvement upon a previously developed method known as jackknife resampling (Efron, 1979). The jackknife method was first developed by British statistician Maurice Quenouille in 1949 in the paper "Problems in Plane Sampling". In this paper, Quenouille presented expressions of the accuracy of measuring linear sampling error and sampling error in systematic and stratified sampling of an area (Quenouille, 1949). An American mathematician by the name of John Tukey expanded on these expressions in

[2] A trimmed mean is the mean of the remaining data after some percentile is trimmed from each end. For example, a ten-percent trimmed mean (the most common percentile used) takes the upper and lower ten percent of the data away, looking only at the middle eighty percent to calculate the mean. This is often used when outliers are present in the dataset.

the math and science progress-oriented atmosphere of 1959 as America competed against the Soviet Union in the Cold War. The expression became known as the Quenouille-Tukey jackknife. Tukey named the method the “jackknife” to symbolize the roughness of the statistical tool, referencing a folding knife that many men carried around at the time that was seen as a useful tool but not an ideal one (Champkin, 2010).

The idea behind the Quenouille-Tukey jackknife resampling method is to take the original sample, exclude one of the observations, and calculate the desired statistic with this new sample. After systematically excluding every observation one at a time and calculating the statistic, there would be n sample statistics—based on samples of size $(n-1)$. This concept of using one sample to create an entire sampling distribution would be what Efron based his bootstrap sampling method off of. The next step in the jackknife method was to average the n sample statistics to find the center of this sampling distribution (Quenouille, 1949). Similar to a point estimate of the parameter, the statistic of the original sample can be given as an estimate of the parameter, but unlike a simple point estimate, the jackknife method provides a measure of accuracy and validity of this estimate by providing the necessary tools to assess whether the three assumptions stated previously in the paper are met. This jackknife method, providing a view of the sampling distribution, allows for the assumption of normality to be assessed either visually or through statistical tests such as the Shapiro-Francia normality test. Also, importantly, the standard error of the estimate of the parameter could be given by comparing each sample statistic in the generated sampling distribution to the mean of the sampling distribution that was previously calculated (Quenouille, 1949). The jackknife method also gives an estimate of the amount of bias in the estimated parameter by comparing it to the center of the sampling distribution. With an insight into how well these three assumptions are met, the jackknife method can justify that a confidence interval for a difficult parameter to work with is valid.

When estimating the parameter θ *with sample size* n *the Quenouille − Tukey jackknife can be carried out as follows*

1.) *Calculate* $\hat{\theta}$ *for the original sample*
2.) *Calculate* $\hat{\theta}_i$ *for the sample leaving out observation* x_i *for* $i = 1,2,\ldots,n$

$$\hat{\theta}_\mu = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i$$

$$Bias = (n-1)(\hat{\theta}_\mu - \hat{\theta})$$

$$SE(\hat{\theta}) = \left\{\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_i - \hat{\theta}_\mu)^2\right\}^{\frac{1}{2}}$$

The method detailed above quickly grasped the attention of the statistical community due to its potential to justify inferences made on parameters, specifically on nonparametric parameters[3] that had proven difficult for statisticians up until that point. One important statistician that took an interest was Rupert Miller, a professor at Stanford University. Miller researched the jackknifing technique and wrote many papers on the subject, trying to point out its flaws and help resolve them. This research likely impacted the career of Miller's Ph.D. student, Bradley Efron, the founder of the bootstrapping methods. In fact, after receiving his doctorate and working at Stanford for a few years, Efron went on sabbatical to Imperial College where Miller gave a lecture revolving around his 1964 paper on the method of jackknifing (Holmes, 2003). With the encouraging push of a colleague, Efron began looking at the jackknife method. The influence of Miller on Efron in this early research is evident in his references to Miller's previous work, specifically "The Jackknife: A Review" that attempted to detail all of the research and findings on the jackknife

[3] Nonparametric parameters are parameters that do not have a known distribution and typically involve analyzing the data set based on the rank of the observations in the data set rather the numerical values of the observations.

method from its inception through 1974 (Efron, 1979; Efron & Stein, 1981). Over the next few years, Efron worked on developing a method that would accomplish the same thing but be more randomized and less systematic. In January of 1979, he published his paper on bootstrap methods, claiming that they were actually more applicable and dependable than the jackknife methods (Efron, 1979). In fact, Efron explained that jackknifing is a linear approximation of bootstrapping (Efron, 1981a). Thus, although bootstrapping was developed later, it found that its predecessor in estimating bias and variance was really a subset of its own methods.

The type of bootstrapping method proposed by Efron is relatively simple, but its implications are great. Bootstrapping can do things that other methods cannot, and it can do them better. While jackknifing notoriously fails to accurately estimate the variance in the sample median, Efron's method can (Efron, 1979). Also, rather than it being a systematic method for estimating variance and bias, bootstrapping is randomized. Starting with the original sample of size n, Efron proposed assuming the sample to be representative of the population of inference and resampling from that sample with replacement. Resampling with replacement involves taking the original sample and using a random number generator to pick a number from the dataset. This number is then kept in the dataset so that it can be chosen again, and a new number is randomly selected. This process is repeated until a new sample of size n is created. Doing this a multitude of times results in many "bootstrap" samples also of size n that have been randomly drawn from the assumed population. By running the same statistics on these "bootstrap" samples as the original sample, a sampling distribution can be created to understand its shape, center, and spread (Graham, 2018). From this, a confidence interval can be generated for the population parameter based on the original sample statistic, and the accuracy and validity of this interval can be justified with the information about the sampling distribution given by the bootstrap methods.

To help display the difference in the two methods and the difference in the accuracy of the two methods, an example will be done using the jackknife method and then the bootstrap method to estimate the variance of the median of a data set.

For this example, say a researcher is interested in the oxygen level of a nearby river after a mine was established in order to assess the river's ability to support fish. Fish typically can survive if the oxygen level is above five parts per million. In trying to answer the research question, the researcher ideally would want to sample as many places as possible along the river. However, due to cost constraints, time constraints, and the habitat impact that the surveying equipment has, the researcher must limit the number of observations to fifteen randomly selected test spots within fifty miles of the mine. The results of this testing is $\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$. Due to mild outliers, the researcher chooses to look at the median rather than the mean. The median of the sample is found to be 5.4 ppm. In order to use this statistic to create a confidence interval for the median oxygen

$Example\ of\ Bootstrapping$:

1.) Calculate $\hat{\theta}$ for $Original\ Sample = \{x_1, x_2, x_3, \dots, x_n\}$ $(n\ values\ total)$
2.) $Resample\ with\ replacement$: $\{x_2, x_3, x_5, x_5, x_1, x_2, x_n, \dots. x_2\}$ $(n\ values\ total)$
3.) Repeat 10,000 times, calculating $\widehat{\theta_\imath}$ for each bootstrap sample
4.) Create a histogram of the $\hat{\theta}_i$'s to see the shape of the sampling distribution
5.) Calculate the average of $\hat{\theta}_i$'s (symbolized by $\hat{\theta}_\mu$)

$$Bias\ = \hat{\theta}_\mu - \hat{\theta}$$

$$SE(\theta) = \left\{\frac{1}{n-1}\sum_{i=1}^{10000}(\hat{\theta}_i - \hat{\theta}_\mu)^2\right\}^{\frac{1}{2}}$$

level of that whole stretch of the river, the researcher needs to use resampling methods in order to meet the assumptions necessary to ensure a valid confidence interval.

Jackknife:

Let X be a vector representing our sample. Then,

$X = \{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$ with $n = 15$ where $m(X) = 5.4$.

Observation being left out		$m(X_i)$	$\left(m(X_i) - m(X)\right)^2$
x_1	$\{4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_2	$\{2.3, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_3	$\{2.3, 4.5, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_4	$\{2.3, 4.5, 4.8, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_5	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_6	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_7	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.45	.0025
x_8	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.35	.0025
x_9	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.3	.0100
x_{10}	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$	5.3	.0100
x_{11}	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.8, 5.8, 5.9, 7.8\}$	5.3	.0100
x_{12}	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.9, 7.8\}$	5.3	.0100
x_{13}	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.9, 7.8\}$	5.3	.0100
x_{14}	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 7.8\}$	5.3	.0100
x_{15}	$\{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9\}$	5.3	.0100

Jackknife (continued):

$$\hat{\theta}_\mu = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \frac{1}{15}(7(5.45) + 5.35 + 7(5.3))$$

$$\hat{\theta}_\mu = 5.373333$$

$$Bias \ = (n-1)\hat{\theta}_\mu - \hat{\theta} = (14)5.373333 - 5.4 = -.373338$$

$$SE(\theta) = \left\{\frac{1}{n-1}\sum_{i=1}^{15}(\hat{\theta}_i - \hat{\theta}_\mu)^2\right\}^{\frac{1}{2}} = \sqrt{\left(\frac{14}{15}\right)\left(7(.0025) + .0025 \ + 7(.0100)\right)}$$

$$SE(m) = .2898275$$

From the jackknife method, the bias in the estimate is given to be -.373338 and a standard error of the median is calculated to be .2898275. When compared to the bias and standard error estimates given from the bootstrap method, it will be obvious why Efron's method was a great improvement upon the Quenouille-Tukey jackknife method.[4]

[4] It should be noted here that improvements have been made upon the jackknife resampling method in order to make it more effective. There is now a deleted-d method of jackknifing, in which d observations are left out for each recalculation of the sample statistic. This has been shown to resolve many issues of the jackknife method in estimating the standard error and bias of non-smooth parameters (Shao & Wu, 1989).

Bootstrap:

Let X be our sample. $X = \{2.3, 4.5, 4.8, 5.1, 5.2, 5.2, 5.2, 5.4, 5.5, 5.5, 5.7, 5.8, 5.8, 5.9, 7.8\}$

with $n = 15$ and $m(X) = 5.4$.

Using a computer, resample with replacement to get the following $X_i's$.

X_i	Random Number Sequence	Resulting Bootstrapped Sample	$m(X_i)$
X_1	{6,4,10,14,11,4,4,2,10,2,10,10,14,11,10}	{5.2,5.1,5.5,5.9,5.7,5.1,5.1,4.5,5.5,4.5,5.5,4.5,5.9,5.7,5.5}	5.5
X_2	{3,8,7,9,11,11,12,8,8,6,8,11,12,10,11}	{4.8,5.4,5.2,5.5,5.7,5.7,5.8,5.4,5.4,5.2,5.4,5.7,5.8,5.5,5.7}	5.5
X_3	{12,14,2,11,13,7,11,3,14,8,3,10,4,1,11}	{5.8,5.9,4.5,5.7,5.8,5.2,5.7,4.8,5.9,5.4,4.8,5.5,5.1,2.3,5.7}	5.5
X_4	{15,13,9,15,10,7,15,12,7,9,6,13,5,2,3}	{7.8,5.8,5.5,7.8,5.5,5.2,7.8,5.8,5.2,5.5,5.2,5.8,5.2,4.5,4.8}	5.5
X_5	{10,8,8,10,3,4,15,12,8,9,13,2,7,10,4}	{5.5,5.4,5.4,5.5,4.8,5.1,7.8,5.8,5.4,5.5,5.8,4.5,5.2,5.5,5.1}	5.4
X_6	{12,14,11,1,3,12,11,10,14,13,12,12,5,5,3}	{5.8,5.9,5.7,2.3,4.8,5.8,5.7,5.5,5.9,5.8,5.8,5.8,5.2,5.2,4.8}	5.7
X_7	{2,5,8,2,7,4,12,4,1,11,6,2,6,12,10}	{4.5,5.2,5.4,4.5,5.2,5.1,5.8,5.1,2.3,5.7,5.2,4.5,5.2,5.8,5.5}	5.2
X_8	{3,12,15,4,8,9,14,4,14,3,3,1,12,6,10}	{4.8,5.8,7.8,5.1,5.4,5.5,5.9,5.1,5.9,4.8,4.8,2.3,5.8,5.2,5.5}	5.4
X_9	{3,12,3,1,8,2,7,8,15,4,14,2,3,6,7}	{4.8,5.8,4.8,2.3,5.4,4.5,5.2,5.4,7.8,5.1,5.9,4.5,4.8,5.2,5.2}	5.2
X_{10}	{12,5,12,11,3,5,1,6,6,12,4,1,5,7,8}	{5.8,5.2,5.8,5.7,4.8,5.2,2.3,5.2,5.2,5.8,5.1,2.3,5.2,5.2,5.4}	5.2
X_{11}	{4,13,12,11,6,7,14,10,4,4,3,10,15,14,13}	{5.1,5.8,5.8,5.7,5.2,5.2,5.9,5.5,5.1,5.1,4.8,5.5,7.8,5.9,5.8}	5.5
X_{12}	{2,8,7,4,3,6,2,9,9,8,13,9,1,15,3}	{4.5,5.4,5.2,5.1,4.8,5.2,4.5,5.5,5.5,5.4,5.8,5.5,2.3,7.8,4.8}	5.2
X_{13}	{12,2,4,2,12,4,5,7,15,10,4,1,3,13,9}	{5.8,4.5,5.1,4.5,5.8,5.1,5.2,5.2,7.8,5.5,5.1,2.3,4.8,5.8,5.5}	5.2
X_{14}	{8,15,12,5,8,3,10,3,3,1,12,15,9,7,15}	{5.4,7.8,5.8,5.2,5.4,4.8,5.5,4.8,4.8,2.3,5.8,7.8,5.5,5.2,7.8}	5.4
X_{15}	{5,5,14,7,9,4,7,3,14,8,11,1,7,13,10}	{5.2,5.2,5.9,5.2,5.5,5.1,5.2,4.8,5.9,5.4,5.7,2.3,5.2,5.8,5.5}	5.2

Bootstrap (continued):

$$\hat{\theta}_\mu = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i = \frac{1}{15}(5.5 + 5.5 + 5.5 + 5.5 + 5.4 + 5.7 + 5.2 + 5.4 + 5.2 + 5.2 + 5.5 + 5.2 + 5.2 + 5.4 + 5.2)$$

$$\hat{\theta}_\mu = 5.373333$$

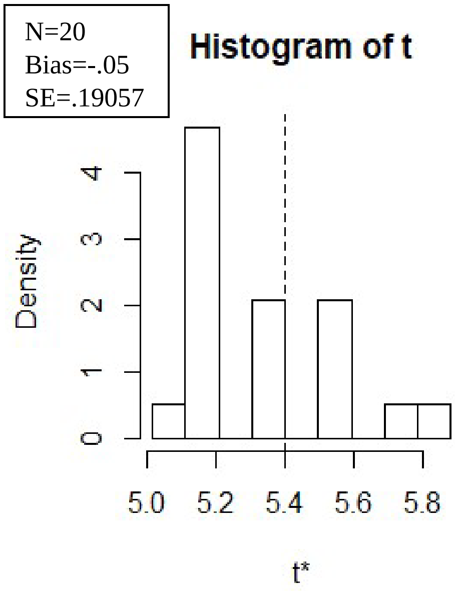
$$Bias \ = \hat{\theta}_\mu - \hat{\theta} = 5.373333 - 5.4 = -.02667$$

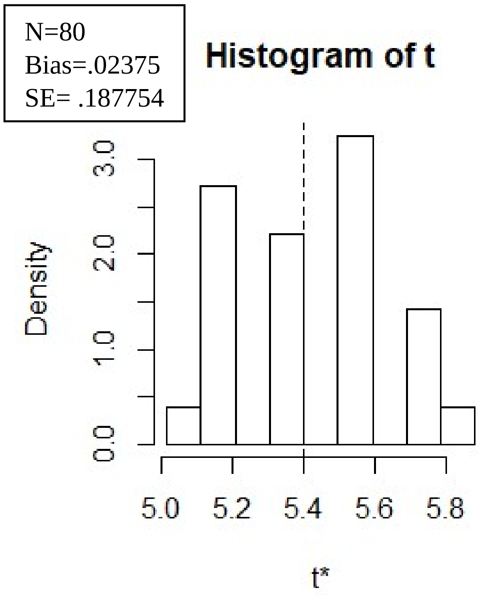
$$SE(\theta) = \left\{\frac{1}{n-1}\sum_{i=1}^{15}(\hat{\theta}_i - \hat{\theta}_\mu)^2\right\}^{\frac{1}{2}}$$

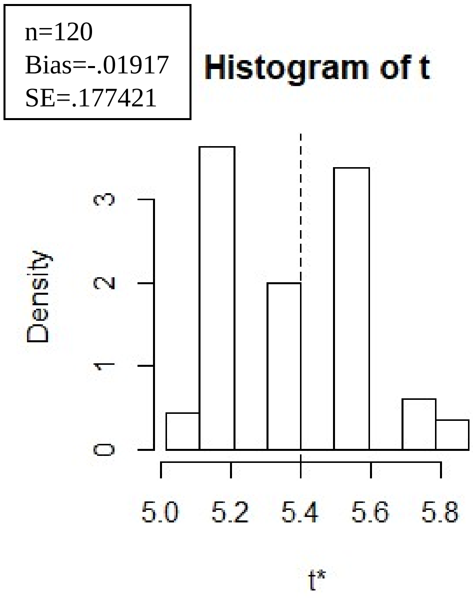
$$= \sqrt{(\frac{1}{14})(6(5.2 - 5.373333)^2 + 3(5.4 - 5.373333)^2 \ + 4(5.5 - 5.373333)^2 + (5.7 - 5.373333)^2}$$

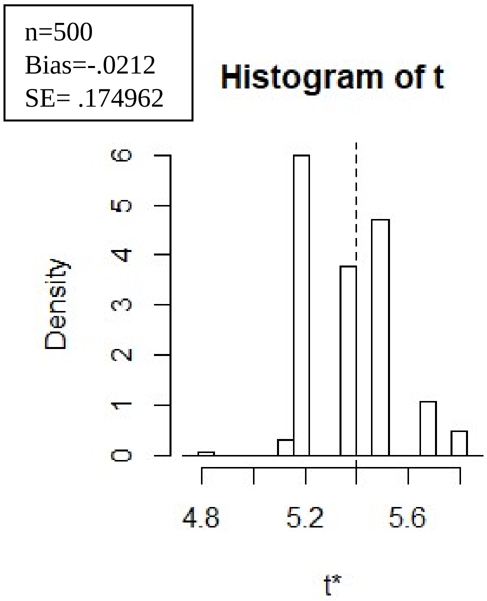
$$SE(m) = .1624221$$

The bias estimate given for the bootstrap sample is less than one-tenth of the bias estimate given from the jackknife method, and this is only when using fifteen bootstrap samples. As the number of bootstrap samples increased, this bias tends to be further decreased. Similarly, the standard error estimate for the bootstrap method is almost half that of the jackknife resampling method, and like the bias, with more bootstrapped samples, this estimate of the standard error will become more accurate. The distribution of the median of the bootstrapped samples can be seen below in the histograms where N represents the number of bootstrapped samples.


N=20
Bias=-.05
SE=.19057


N=80
Bias=.02375
SE= .187754


n=120
Bias=-.01917
SE=.177421


n=500
Bias=-.0212
SE= .174962

One thing that can be seen is that the sampling distribution of the median does not appear to be normal. This is important to note because this means calculating a normal-based confidence interval for the median would give an inaccurate estimate. To resolve this issue, Efron eventually developed a confidence interval that corrects for the skewness and nonnormality of the sampling distribution. This will be discussed later on in the paper as the evolution of the bootstrapping method is traced out. The most important thing to note here is the utility of the bootstrapping method over the jackknife method. Although the bootstrap method may involve more calculation than the jackknife method, with modern technology, it has three advantages that make it the more ideal method: its ability to accurately estimate that standard error of non-smooth parameters such as quantiles or the median (Martin, 1990), the ability to resample as many times as desired, and the reduction of bias through randomization.

These benefits may seem small, but they have greatly expanded the scope of research. Studies where cost previously would have been too great in order to do research, such as having to administer an expensive treatment to many subjects, can now be done with smaller sample sizes due to the technique of bootstrapping. Other studies that would have been too time-intensive, such as sampling many acres in a forest, can now be done relatively quickly by taking a smaller sample and bootstrapping. Efron's bootstrapping techniques have heavily impacted the field of statistics and continue to do so; they allow researchers to work with sample sizes that previously had been too small to make inferences based upon and to work with statistics that have unknown sampling distributions like the median and trimmed mean.

In his original paper "Bootstrap Methods: Another Look at the Jackknife", Efron proposed three primary methods for computing the bootstrap distribution. The first and simplest to understand is the one used in the above example that approximates the bootstrap distribution based upon empirical bootstrapped samples. One of the other methods proposed by Efron involves the theoretical calculation of the bootstrapped distribution (Efron, 1979). These calculations often are difficult but provide a true look at the bootstrap distribution instead of an approximation to it. In his paper, Efron gives the probability distribution for the median of the bootstrap sampling distribution for a sample size of thirteen. Efron also proposed using a Taylor Series to approximate the mean and variance in the sampling distribution of the bootstrap samples. In fact, Efron proved that this third method was closely related to using jackknifing (Efron, 1979). The mathematics behind Efron's proof of this are difficult to explain, but it is based on creating a theoretical bootstrap distribution based on the expected number of each observation. Based on the number of observations of each type in the original distribution, the probability for that observation being selected into a bootstrap sample can be calculated. Multiplying the total number of observations that would be in a bootstrap sample (n) by these probabilities, gives the expected number of each observation in any bootstrap sample (Efron, 1979). Using this, Efron expanded the probability distribution in a Taylor Series using concepts about multinomial distributions. The resulting estimate of the standard error and the mean of the sampling distribution closely resemble the standard error and mean estimate given by a specific jackknife procedure known as the infinitesimal jackknife procedure developed by Louis Jaeckel (Efron, 1979; Efron, 1981). A copy of this derivation from his paper is included in Appendix A, and for more information see "Bootstrap Methods: Another Look at the Jackknife" (Efron, 1979) and "Nonparametric Estimates of the Standard Error: The Jackknife, the Bootstrap, and Other Methods" (Efron, 1981).

While these latter two variations are more theoretically based, the method outlined in the above box is the easiest to understand the bootstrap sampling distribution. This method tends to be the most commonly used; although, none of the three methods Efron proposed in his paper immediately took hold. It took many years for bootstrapping to become a commonly used method in statistics. Most people struggled to understand how the methods worked or accept the premise that the methods are based upon (Champkin, 2010). For most statisticians in the early 1980s, it was uncomfortable to simply assume the sample to be representative of the population. Efron's method appeared to many as unfounded so jackknifing remained the predominant method of resampling for many years following the discovery of bootstrapping.

Another reason jackknifing remained the primary method was that bootstrapping developed before software was capable of carrying out the computations bootstrapping required. This meant in order for a statistician to use the method, they would have had to do it by hand, which would require excessive amounts of time and energy. It is hard to imagine bootstrapping being difficult and time consuming given that modern computers can complete the process in a matter of seconds. However, the statistical software most statisticians currently use to do bootstrap resampling methods was not even developed until 1995—over fifteen years after Efron introduced the concept. In fact, when Efron introduced bootstrapping to the statistical world, most homes did not have computers and the computers that did exist had only a portion of the processing power required in bootstrapping to store the large datasets and complete the necessary operations. The method had to be carried out manually. To complete this process even with only fifteen bootstrapped samples, as done above, means the researcher must generate fifteen random number sequences, create samples corresponding to these sequences, find the statistic of interest for each of these fifteen samples, find the mean of all fifteen samples, and calculate the variance in the fifteen samples. If the researcher wanted a much better picture of the bootstrapped sampling distribution, they would want at least fifty bootstrapped samples. In addition to this, the researcher may be interested in a statistic that is also time-consuming to calculate such as the trimmed mean.

The time that bootstrap resampling methods might save when taking the original sample would be minimal compared to the amount of time it would take to do such an analysis by hand. Not only did the software capabilities prevent individuals from using the method, but it prevented many from thoroughly researching the method or testing it empirically. Efron's bootstrap resampling methods were ahead of their time, resulting in jackknifing—a less time-consuming method with fewer opportunities to make errors—being the preferred method despite its inferiority.

As computing ability advanced, bootstrapping methods were more thoroughly examined and expanded upon. In fact, over the next four decades, Efron himself and many other notable statisticians worked to refine bootstrapping and to develop specific subcategories of bootstrapping. In 1981, two years after Efron's original paper, it was shown through comparison of the distribution resulting from the Monte Carlo bootstrapping technique, the one involving using empirical bootstrapped samples to approximate the distribution, to the bootstrap sampling distribution given through the analytical method, that Efron's proposed method closely approximated the bootstrap sampling distribution when working with means (Bickel & Freedman, 1981). In fact, the method was shown to work for a variety of examples and only failed when estimating a statistic from uniformly distributed data (Bickel & Freedman, 1981). This type of research on the bootstrap methods continues today and has resulted in many validations of the method as well as adjustments and corrections. Research on the bootstrap has also led to an expansion of the bootstrapping method.

Efron's original paper introduced the method of nonparametric bootstrapping described above. In addition to this, hidden in remark K of the notes section of this paper, Efron introduced parametric bootstrapping (Efron, 1979). The methods are very similar, primarily distinguished by where the bootstrapped samples are taken from. As seen above, in nonparametric bootstrapping, the bootstrapped samples are generated by resampling with replacement directly from the original sample. In parametric bootstrapping the original sample is theorized to follow some specific model and the resulting samples are generated by sampling from this model (Efron, 1981). For example, a researcher may take a sample of trunk widths for trees in various locations in a forest and assume that this type of data will follow a normal distribution. Using the sample mean and standard deviation as the parameter estimates for this hypothesized normal distribution, random samples are then generated from this hypothesized distribution. Once the samples are created, the remaining steps are carried out exactly as they would be in the nonparametric bootstrap. Even more so than Efron's nonparametric bootstrap, the parametric bootstrap struggled to find popularity. While most of this was likely due to the same reasons nonparametric bootstrapping was largely ignored, part of it may have been due to Efron leaving the introduction of this method for his notes section. Despite the lack of immediate popularity and like its nonparametric counterpart, the parametric bootstrap could reduce the necessary sample size for inference and could help statisticians understand the uncertainty in their inferences. The parametric bootstrap also provides a great advantage over the nonparametric bootstrap in being able to sample any value within the theorized distribution rather than just those from the original sample. This creates a more complete estimate of the sampling distribution. However, the parametric bootstrap also relies on an accurate model being fit to the data at the start, which can be very difficult to do. This may have been an additional reason that parametric bootstrapping struggled even more than its nonparametric counterpart to gain acceptance.

Two years after the introduction of these methods of bootstrapping, Donald Rubin took Efron ‘s nonparametric bootstrap and manipulated it to operate with Bayesian probabilities rather than frequentist probabilities (Rubin, 1981). In doing so, Rubin communicated with Efron and was able to resolve one of the drawbacks of nonparametric bootstrapping that Efron himself acknowledged in his original paper. Rubin's result became known as the Bayesian bootstrap (Efron, 1979; Rubin, 1981). The methods involved in this bootstrap are very similar to those involved in the original nonparametric bootstrap. The primary difference is that the methods apply to posterior probabilities. These are probabilities that are updated based on some other information. For example, the probability of a man making a free throw is much different than the probability of a man who is a professional basketball player making a free throw. The fact that the man is a professional basketball player changes the likelihood of him making a free throw. The posterior probability is the probability of him making it with the knowledge that he is a professional basketball player while the anterior or prior probability is the estimated probability before this information is known.

Bayesian bootstrapping is done by taking $(n-1)$[5] random variates from the uniform distribution [0,1], meaning each random variate is equally likely (Rubin, 1981). The random variates are then placed in ascending order with zero as an additional entry on the low end and one as an additional entry on the high end (Rubin, 1981). The difference between the successive entries are then calculated, and these n difference are placed into a vector g_i. This vector is then applied to the vector of data values such that x_1 is weighted by probability g_1 and x_n is weighted by probability g_n (Rubin, 1981). This creates one Bayesian bootstrapped sample. This process is repeated to get many samples and from this, a distribution is created similarly to how it is created through nonparametric bootstrapping. However, rather than representing the sampling distribution, the distribution resulting from the Bayesian bootstrap represents the posterior distribution of the parameter (Rubin, 1981). This distribution is advantageous because it allows researchers to make

[5] In general, n represents the sample size.

statements on the likelihood of the value of the parameter rather than just the expected frequency of the sample statistic under a hypothesized parameter value (Rubin, 1981). For example, the Bayesian bootstrap method would have allowed the researcher in our oxygen level in the river example to give a likelihood that the median is a particular value whereas the nonparametric bootstrap will only be able to tell us the likelihood of observing the data we did if the true mean oxygen level of that portion of river was some hypothesized value such as 5ppm. With the nonparametric bootstrap, the researcher must compare the estimated statistic to some hypothesized value for the parameter or can create a confidence interval for the parameter, while Bayesian bootstrapping assigns specific probabilities to parameter estimates.

Also in 1981, Efron himself created an adjustment to the confidence intervals created from bootstrap methods. As mentioned previously in this paper, the sampling distribution that results from the bootstrap method often times does not appear normal. This is because in nonparametric bootstrapping only certain numbers in the distribution can be chosen—the ones in the original sample. This results in large gaps in the sampling distribution. If a researcher were to create a normal-based confidence interval using this information, the assumption of normality would be violated, and the confidence interval would not be accurate. In order to combat this, Efron introduced a method for bias-corrected confidence intervals that accounted for the nonnormality seen in the bootstrapped estimate of the sampling distribution. He improved upon these intervals again in 1987 to create BC_a confidence intervals (also known as bias-corrected and adjusted confidence intervals. These adjustments to the original method have been successful and still are the primary methods used today.

The original nonparametric bootstrap method proposed by Bradley Efron in 1979 has greatly changed the field of statistics. As mentioned earlier, it allows researchers to work with smaller samples and statistics with unknown sampling distributions. Similarly, it allows researchers to more accurately measure the uncertainty in their estimates and inferences and allows researchers to check certain assumptions that might need to be met to do hypothesis testing on their data. The invention of bootstrap resampling methods has expanded the scope of research, and it continues to do so as statisticians work on expanding the method. 1979 marked the introduction of nonparametric and parametric bootstrapping. This was followed in 1981 with the development of the Bayesian bootstrap and bias adjusted confidence intervals. In the years since, there have been many more developments including the smooth bootstrap, the semiparametric bootstrap (which has many variations within it), and the block bootstrap. Developments continue to be made to extend the bootstrap method to various types of data. Using similar ideas as those in the Quenouille-Tukey jackknife method, Efron developed and launched a whole new branch of resampling methods that use randomization principles. Efron named these methods "bootstrapping" in order to emphasize that resampling one's own data to create the sampling distribution resembled the way Baron Munchausen pulled himself up by the bootstraps in the tall tale written by R.E. Raspe (Graham, 2018). Bootstrapping is now commonly used and has drastically shaped the field of statistics, allowing researchers to pull themselves up by the bootstraps to undertake studies and make inferences on the data that would not otherwise have been possible.

Acknowledgement

This paper was completed as part of the course requirements for M429: History of Mathematics, taught by Prof. Bharath Sriraman in Spring 2019.

References

Bickel, P., & Freedman, D. (1981). Some Asymptotic Theory for the Bootstrap. *The Annals of Statistics*, 9(6), 1196-1217. Retrieved from http://www.jstor.org/stable/2240410.

Champkin, J. (2010, December). Bradley Efron. *Significance.* 178-181. https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1740-9713.2010.00460.x

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

Efron, B. (1979). Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7(1), 1-26. Retrieved from http://www.jstor.org/stable/2958830.

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

Efron, B., & Stein, C. (1981). The Jackknife Estimate of Variance. *The Annals of Statistics*, 9(3), 586-596. Retrieved from http://www.jstor.org.weblib.lib.umt.edu:8080/stable/2240822.

Graham, J. (2018) Class Notes for STAT 451: Methods in Bootstrapping 5.8.

Holmes, S & Morris, C. & Tibshirani, R. (2003) Bradley Efron: A Conversation with Good Friends. *Statistical Science*, 18(2), 268-281. doi: 10.1214/ss/1063994981.

Martin, M. (1990). On Using the Jackknife to Estimate Quantile Variance. *The Canadian Journal of Statistics / La Revue Canadienne De Statistique*, 18(2), 149-153. Retrieved from http://www.jstor.org.weblib.lib.umt.edu:8080/stable/3315563.

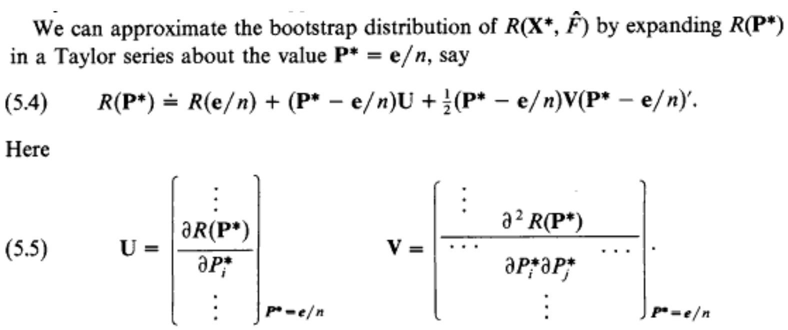
Quenouille, M. (1949). Problems in Plane Sampling. The Annals of Mathematical Statistics, 20(3), 355-375. Retrieved from http://www.jstor.org/stable/2236533.

Rubin, D. (1981). The Bayesian Bootstrap. *The Annals of Statistics*, 9(1), 130-134. Retrieved from http://www.jstor.org/stable/2240875.

Shao, J., & Wu, C. (1989). A General Theory for Jackknife Variance Estimation. *The Annals of Statistics,* 17(3), 1176-1197. Retrieved from http://www.jstor.org.weblib.lib.umt.edu:8080/stable/2241717.

Appendix A

1.) The derivation of the Taylor series as shown by Bradley Efron in his paper. (Efron, 1979)



We can approximate the bootstrap distribution of $R(\mathbf{X}^*, \hat{F})$ by expanding $R(\mathbf{P}^*)$ in a Taylor series about the value $\mathbf{P}^* = \mathbf{e}/n$, say

$$R(\mathbf{P}^*) \doteq R(\mathbf{e}/n) + (\mathbf{P}^* - \mathbf{e}/n)\mathbf{U} + \tfrac{1}{2}(\mathbf{P}^* - \mathbf{e}/n)\mathbf{V}(\mathbf{P}^* - \mathbf{e}/n)'. \tag{5.4}$$

Here

$$\mathbf{U} = \left[\begin{matrix} \vdots \\ \dfrac{\partial R(\mathbf{P}^*)}{\partial P_i^*} \\ \vdots \end{matrix}\right]_{P^* = e/n} \qquad \mathbf{V} = \left[\begin{matrix} & \vdots & \\ \cdots & \dfrac{\partial^2 R(\mathbf{P}^*)}{\partial P_i^* \partial P_j^*} & \cdots \\ & \vdots & \end{matrix}\right]_{P^* = e/n}. \tag{5.5}$$