An R Implementation of the Bias-Corrected and Accelerated Bootstrap Confidence Interval

Luke Sianchuk

August 13, 2021

Contents

Abstract

The bootstrap is a powerful tool in computational statistics which can be used to construct non parametric confidence intervals around an estimate. Although widely applicable in a variety of problem settings, the bootstrap can still be uniquely tailored such that specific statistical issues are addressed. The following paper provides an introduction and R implementation of the bias-corrected and accelerated bootstrap interval (BC_a), which corrects for bias and skewness in the bootstrap distribution. We begin by surveying the early statistical literature and defining the required terms of the BC_a interval. This is followed by an illustrative example drawn from an ecological context, where we walk through the construction of a confidence interval by defining a collection of versatile and reusable functions in the R programming language.

Introduction

Overview

Throughout history, curiosity has driven some of society's greatest minds towards myriad scientific contributions. For centuries, talented scholars have pondered over the natural world, seeking explanations through patterns and logic and communicating through the language of mathematics. With the invention of computers, enormous leaps have been continuously made in the field of statistics. Here, we dive into a specific instance of the bootstrap, a revolutionary resampling technique used to generate estimates of a population's parameters. Through surveyed literature and an illustrative example, we examine how the accelerated bias-corrected percentile bootstrap has secured a crucial role in statistical methodology.

Problem and Significance

In 1979, Bradley Efron first introduced the bootstrap in his paper, "Bootstrap Methods: Another look at the jacknife" [1]. Named after the age old impossibility of "pulling one's self up by their bootstraps," the method bases itself upon a seemingly counter-intuitive exercise in logic. Bootstrapping seeks to deduce new information from what is already available to the statistician. By taking a sample and sampling from it many times over, with replacement, one is able to produce thousands of simulated datasets which can be analyzed to provide meaningful results.

The bootstrap resampling technique is useful because it allows us to assess the variability of some estimator when the data generation distribution is unknown. This is a very powerful technique, made possible by our technical advancements in computing power. Further, once delineated, the distribution of bootstrap samples can be used to construct confidence intervals around the parameter of interest.

However, as with any statistical methodology, there exists inherent limitations. One of the simplest ways of constructing a bootstrap confidence interval is known as the percentile method [2]. This nonparametric methodology simply takes the desired quantiles from the bootstrap distribution as bounds for the confidence interval. But, this provides an estimate based only on the bootstrap samples and does not adjust for any skewness of the bootstrap distribution. Thus, in 1984, Efron postulated the bias-corrected and accelerated bootstrap interval as a means to address these issues [3].

Organization

The following section of this paper proceeds to detail the methodology and underlying mathematics of the bootstrap method, as well as the formulation of the bias-correction and acceleration factors. The

next section walks the reader through an illustrative example of the bias-corrected and accelerated bootstrap interval, complete with R code and figures. This is proceeded by a discussion and conclusions.

Methodology

Bootstrapping

The bootstrap is a stochastic process based on the resampling of a sample from an unknown distribution. Suppose that we have a sample given by

$$Q(x_1, x_2, \ldots, x_i)$$

where each of the X_i is independent and identically distributed from an unknown distribution F. The bootstrap tells us to sample from Q, with replacement, to generate a number of bootstrap samples $Q_1, Q_2, \ldots, Q_B$, which come from the empirical cumulative distribution function, denoted by $\hat{F}$. In the interest of investigating some parameter $\theta = \theta(F)$, we calculate an estimate, $\hat{\theta}_i$ from each Q_i. The set $(\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B)$ provides an estimate of the sampling distribution.

The number of possible bootstrap samples grows exponentially with increasing sample sizes. It is usually infeasible to compute all possible combinations, and thus, only a small fraction are studied. As a rule of thumb, Wilcox suggests that for general use, one should generate 599 samples [4]. However, this may vary from problem to problem. For best results, one should examine the variability of the bootstrap estimates and continue to resample until settling is observed.

Percentile Confidence Interval

The bootstrap percentile confidence interval was introduced in Efron's 1981 paper, "Nonparametric Standard Errors and Confidence Intervals" [2]. Given B bootstrap samples and their corresponding estimates, $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_B$, the percentile interval is found simply by taking percentiles of the parameter estimates.

For a significance level of α, the lower and upper limits of the interval are given by the $100 \times \frac{\alpha}{2}$ and $100 \times (1 - \frac{\alpha}{2})$ percentiles, respectively. These can be obtained by sorting the data and picking out the lower and upper estimates based on the desired significance level. The intended coverage, given by $1 - 2\alpha$, is given directly from the percentiles. For instance, given $B = 1000$ bootstrap samples and signficance level $\alpha = 0.05$, the percentile confidence interval spans the 50^{th} and 950^{th} ordered values.

These percentiles are denoted:

$$(\hat{\theta}_{lower}, \hat{\theta}_{upper}) = (\hat{\theta}^{(\alpha)}, \hat{\theta}^{(1-\alpha)})$$

Bias-Correction

The first component to be discussed in the bias-corrected and accelerated bootstrap confidence interval is the bias correction factor. Statistical bias refers to the difference between the expectation of an estimator $\hat{\theta}$, and the quantity θ being estimated. By estimating the bias of $\hat{\theta}$, one can correct for it simply by computing $\hat{\theta} - \hat{bias}$.

Efron an Tibshirani present an excellent depiction of the BC_a bootstrap interval in chapter 14.3 of their book, "An Introduction to the Bootstrap" [5]. Following their notation, we denote the value of the bias-correction factor by $\hat{z}_0$. This value is obtained from the proportion of bootstrap

estimates which are less than the observed statistic. This is calculated by:

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

where # denotes counting the number of observations in the braces,
$\hat{\theta}_b$ are individual bootstrap estimates,
$\hat{\theta}$ is the original estimate,
and Φ^{-1} is the inverse function of a standard normal's cumulative distribution function.

The value of $\hat{z}_0$ provides a measure, in normal units, of the discrepency between the median of $\hat{\theta}_b$ and $\hat{\theta}$. Notice that a bias-correction factor of 0 is obtained when exactly half of the bootstrap estimates are less than or equal to the value of $\hat{\theta}$.

Acceleration

The acceleration factor, denoted $\hat{\alpha}$, is most easily defined by using the jackknife values of a statistic. The jackknife is a resampling technique which predates the bootstrap, first appearing in the scientific literature from Quenouille in 1949 [6], and named by Tukey in 1958 [7]. To estimate a parameter using the jackknife, a single observation is omitted from the sample prior to computing the estimate. This is repeated for each of the observations, generating from n observations, n samples of size $n-1$. The jackknife estimate is then found through the aggregation of each of the generated samples.

Here, we denote $x_{(i)}$ as the original sample with the i^{th} observation omitted. Then, we let $\hat{\theta}_{(i)} = s(x_{(i)})$ denote the jackknife estimate for that sample, and further define $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(i)}/n$. Utilizing these definitions, the acceleration is given by:

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6[\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2]^{3/2}}$$

The acceleration provides a measure, on a normalized scale, of the rate of change of the standard error of $\hat{\theta}$ with respect to the true value θ. It is proportional to the skewness of the bootstrap distribution, and thus helps to construct a better confidence interval when faced with skewed estimates.

The BCa Interval Definition

With the pieces described above, we can now construct the BC_a confidence interval. This interval is also described in terms of percentiles, but with appropriate adjustment based on the bias-correction and acceleration parameters. The interval, with an intended coverage of $1 - 2\alpha$ is defined:

$$(\hat{\theta}_{lower}, \hat{\theta}_{upper}) = (\hat{\theta}^{(\alpha_1)}, \hat{\theta}^{(\alpha_2)})$$

where

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

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

in which Φ denotes the cumulative distribution function of the standard normal, and $z^{(\alpha)}$ is the 100αth percentile of the standard normal. Notice that this formulation deteriorates into that of the percentile interval when both $\hat{z}_0$ and $\hat{\alpha}$ are equal to 0.

Properties of the BCa Interval

An important property of the BC_a interval is that it is transformation respecting [3]. This means that if we change the estimated parameter θ to some function $f(\theta)$, then the interval endpoints transform correctly by the same function. Further, the BC_a interval has been shown to be second order accurate, whereas the percentile interval is only of the first order [3]. The errors of the BC_a interval go to zero at a rate much quicker than that of the more classical interval.

Example

Background

Presented here is an implementation of the bias-corrected and accelerated bootstrap with usage in an ecological context. The following example is based on an analysis done by Fader and Juliano in their 2013 paper, "An empirical test of the aggregation model of coexistence and consequences for competing container-dwelling mosquitoes" [8].

A main goal of this paper was to test whether two species of mosquitoes exhibited intraspecific aggregation, meaning that they exist in clumped patches of the available habitat. The J-index was created as a means to quantify this phenomenon [9]. This is defined:

$$J = \frac{\sum x_i(x_i - 1)}{m^2 N} - 1$$

where x_i is the number of individuals in container i,
m is the average of the x_i,
and N is the total number of individuals.

The above index measures the proportional increase in the average amount of individuals of the same species present in a sample, relative to the expected amount. This expected value is based on a Poisson distribution, since the samples are assumed to be independent and randomly dispersed. This means that for $J = 0$, individuals are randomly distributed, whereas for $J = 1$, there exists a 100% increase in the mean number of members of the same species with each individual in the same sample. The authors used SAS to implement the BC_a bootstrap on this estimate since it corrects for the observed bias and skewness of the bootstrap distribution. I proceed by using R further investigate bootstrapping of the J-index.

Standard Bootstrapping

To begin the example, we will generate a sample of random data. This will be a vector of length 100, with integer values from 0 to 100.

```
# Set seed for reproducibility
set.seed(20560770)
# Length of vector
N = 100
# Generate data
mydata = round(runif(n=N,min=0,max=100))
```

I will now create a function which calculates the J-index of a given sample.

```
J.index = function(x){
  return((sum(x*(x-1))/(mean(x)^2*length(x)))-1)
}
```

Notice that from the given sample, we have that $J = \hat{\theta}$ is equal to:

```
theta.hat = J.index(mydata)
theta.hat
```

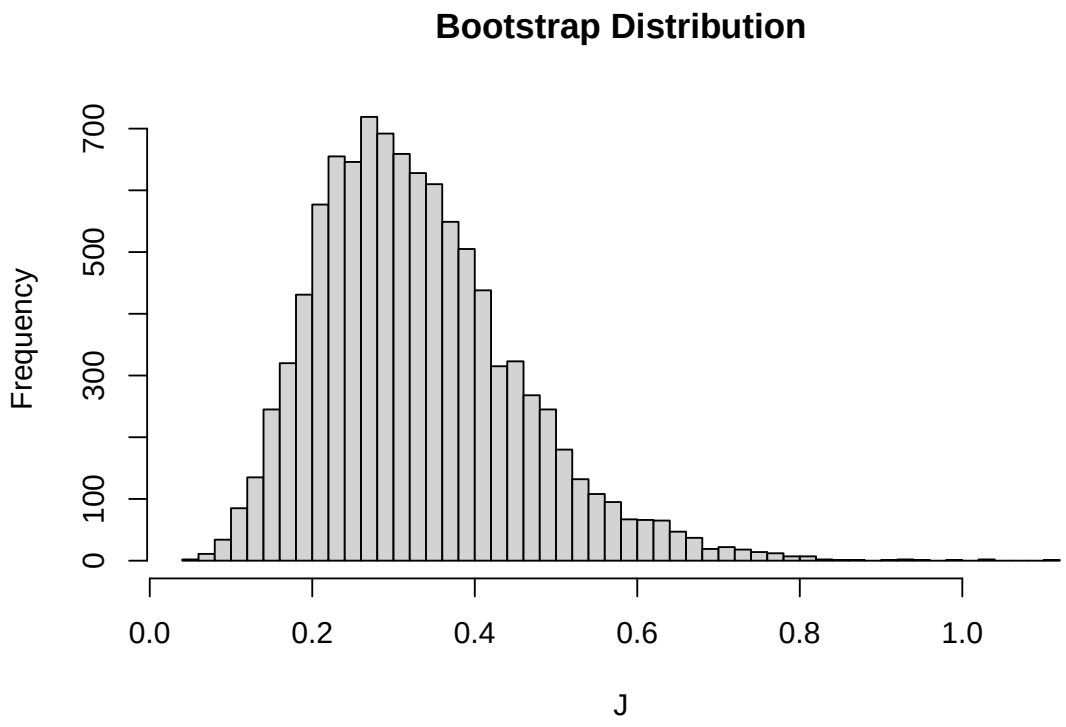
```
## [1] 0.3401116
```

We proceed now by implementing the bootstrap. We will fill a matrix Q with $B = 10000$ bootstrap samples, each of length 20.

```
# Set bootstrap parameters
B = 10000
n=20
# initialize matrix
Q = matrix(0, nrow = B, ncol = n)
# Compute the resampling
for (i in 1:B){
  Q[i, ] = mydata[sample(n, replace = TRUE)]
}
```

The bootstrap distribution is obtained by calculating the J-index for each of these samples. Computing this and plotting the histogram:

```
# Apply the previous J.index function
J.boot = apply(X=Q,MARGIN=1,FUN=J.index)
# Generate histogram
hist(J.boot,breaks=50,main="Bootstrap Distribution",xlab="J")
```

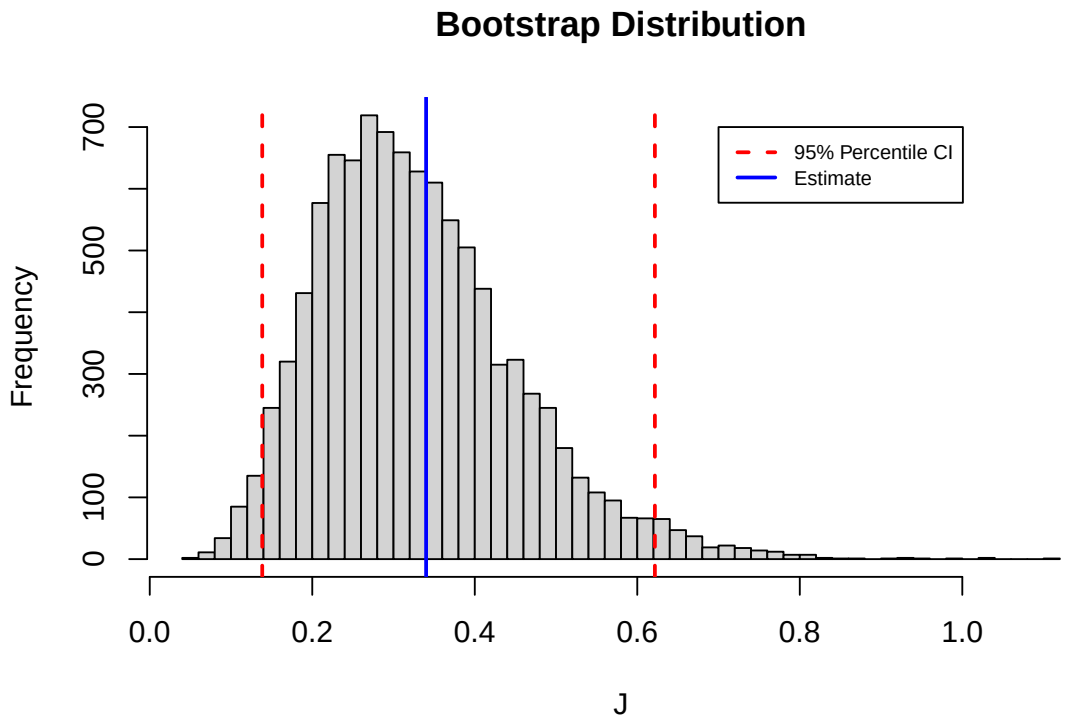


As observed by Fader and Juliano, the bootstrap distribution of the J-index is positively skewed. This

suggests that the percentile confidence interval may not be the best tool in constructing a confidence interval around $\hat{\theta}$. For illustration purposes, plotting the 95% percentile confidence interval as well as the location of $\hat{\theta}$:

```
# Get lower and upper percentiles
alpha=0.05
lower = quantile(x=J.boot,alpha/2)
upper = quantile(x=J.boot,1-alpha/2)

# Redraw histogram
hist(J.boot,breaks=50,main="Bootstrap Distribution",xlab="J")
abline(v=c(lower,upper),lty=2,col="red",lwd=2)
abline(v=theta.hat,col="blue",lwd=2)
legend(0.7, 700, legend=c("95% Percentile CI", "Estimate"),
       col=c("red", "blue"), lty=c(2,1),lwd=2, cex=0.7)
```



Calculating Bias-Correction

Now, we will implement the BC_a bootstrap. We begin by defining a function which calculates the bias-correction value, given the bootstrap estimates.

```
# Function to calculate bias correction
bias.corr = function(boots,estimate){
  # boots are the estimates from the bootstrap distribution
  # estimate is the value theta hat
  z0 = qnorm(sum(boots < estimate)/B)
```

```
}
```

Testing this function to find $\hat{z}_0$ for this example:

```
z0.hat = bias.corr(boots=J.boot,estimate=theta.hat)
z0.hat
```

```
## [1] 0.2131628
```

Calculating Acceleration

Now, we present a funcion to calculate the acceleration paramater, $\hat{\alpha}$. Recall the expression,

$$\hat{\alpha} = \frac{\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^3}{6[\sum_{i=1}^{n}(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)})^2]^{3/2}}$$

which utilizes the jackknife estimates of a parameter. The following function computes those jackknife values given a sample, performs the required computations, and returns the acceleration value $\hat{\alpha}$.

```
accel = function(values){
  # values is the vector of sample values
  n = length(values)
  # initialize list for theta_i
  theta.i = rep(0,n)
  # Use for loop to generate jackknife estimates
  for (i in 1:n){
    # remove point i for jackknife estimate
    theta.i[i] = J.index(values[-i])
  }
  # compute theta_dot
  theta.dot = sum(theta.i)/n
  # putting pieces together
  numerator = sum((theta.dot-theta.i)^3)
  denominator = 6*(sum((theta.dot-theta.i)^2))^1.5
  a = numerator/denominator
  return(a)
}
```

Testing this function on our data, we see that the acceleration parameter is:

```
a.hat = accel(mydata)
a.hat
```

```
## [1] 0.01778024
```

Constructing the Interval

Now that we have functions for the bias-correction and acceleration values, we can piece together the BC_a interval. We will use a significance level $\alpha = 0.05$.

```
# Defining bias correction and acceleration values
z0.hat = bias.corr(boots=J.boot,estimate=theta.hat)
a.hat = accel(mydata)
# Define alpha
alpha = 0.05
```

```r
# Function to find alpha1 and alpha2 of bca interval
bca = function(z0,a,alpha){
  alpha.1 = pnorm(z0+(z0+qnorm(alpha/2))/(1-a*(z0+qnorm(alpha/2))))
  alpha.2 = pnorm(z0+(z0+qnorm(1-alpha/2))/(1-a*(z0+qnorm(1-alpha/2))))
  return(c(alpha.1,alpha.2))
}
```

We see that the bounds of the BC_a confidence interval are given by:

```r
bca.percentiles = bca(z0=z0.hat,a=a.hat,alpha=alpha)
lower.bca = quantile(x=J.boot,bca.percentiles[1])
upper.bca = quantile(x=J.boot,bca.percentiles[2])
lower.bca
```

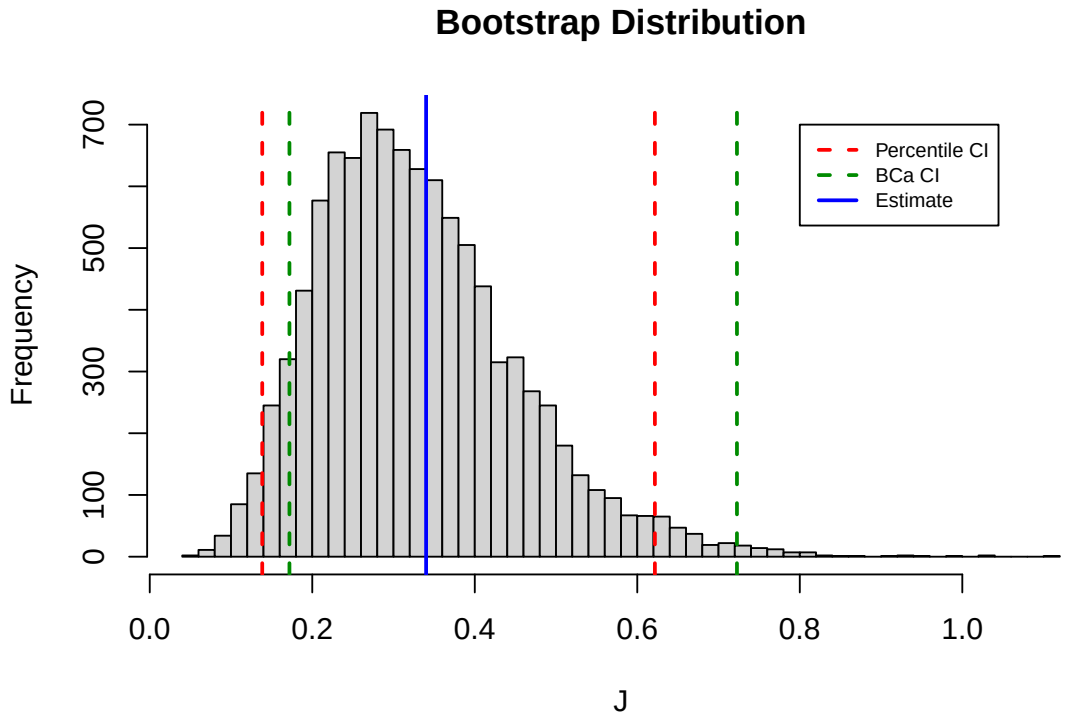
```
## 6.930068%
## 0.1719305
```

```r
upper.bca
```

```
## 99.33126%
## 0.7226531
```

Plotting the previous histogram, but with the BC_a interval included:

```r
hist(J.boot,breaks=50,main="Bootstrap Distribution",xlab="J")
abline(v=c(lower,upper),lty=2,col="red",lwd=2)
abline(v=theta.hat,col="blue",lwd=2)
abline(v=c(lower.bca,upper.bca),lty=2,col="green4",lwd=2)
legend(0.8, 700, legend=c("Percentile CI", "BCa CI","Estimate"),
       col=c("red", "green4","blue"), lty=c(2,2,1),lwd=2, cex=0.7)
```



We see that values from the bootstrap distribution tend to underestimate the parameter value and have a positive skew. Thus, when apply the bias correction and acceleration, the confidence interval has been shifted slightly in the positive direction and widened.

Conclusions

The bootstrap percentile interval provides statisticians with a simple way to construct a non parametric confidence interval around an estimate. However, this can be improved upon through the application of bias-correction and acceleration values. In certain instances, the bootstrap distribution may be skewed and provide biased estimates of a parameter. We have seen a working example of how this can be addressed, using the definitions of the R functions *accel, bias.corr*, and *bca* to construct the BC_a interval.

Since Efron's initial introduction of the bootstrap, he and collaborators have since gone on to produce a plethora of literature surrounding the refinement of the procedure for certain instances. Efron's 1984 paper "Better Bootstrap Confidence Intervals" [3] set the stage for the further development of the resampling method. Through groundbreaking statistical methodology based on techniques such as transformations and bias correction, statisticians have furthered the already remarkable bootstrap to tackle more complex problems. As technological challenges arise, scholars continually advance the bounds of mathematical and statistical knowledge.

References

[1] Bradley Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics, Vol. 7, No. 1, 1-26*, 1979.

[2] Bradley Efron. Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics 9.2: 139-158*, 1981.

[3] Bradley Efron. Better bootstrap confidence intervals. *Journal of the American statistical Association 82.397: 171-185*, 1984.

[4] Rand Wilcox. *Fundamentals of Modern Statistical Methods*. Springer, New York, 2010.

[5] Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, Philadelphia, PA, 1994.

[6] Maurice Quenouille. Problems in plane sampling. *The Annals of Mathematical Statistics: 355-375*, 1949.

[7] John Tukey. Bias and confidence in not quite large samples. *The Annals of Mathematical Statistics. 29: 614*, 1958.

[8] Joseph Fader and Steven Juliano. An empirical test of the aggregation model of coexistence and consequences for competing container-dwelling mosquitoes. *Ecology 94.2: 478-488*, 2013.

[9] Anthony Ives. Aggregation and coexistence in a carrion fly community. *Ecological Monographs 61.1: 75-94*, 1991.