Title here

Justin Papagelis *
Department of Mathematics and Statistics, Amherst College

November 17, 2022

Abstract

For my project, I plan to explore different bootstrap methods for creating confidence intervals and then perform comparisons between a couple of the methods. My paper will re-introduce the idea of bootstrap to my peers and give some background on constructing confidence intervals. We will also go deeper into the theory behind the construction of confidence intervals from bootstrapped data. These various methods to create a confidence interval from a bootstrap can include the percentile method, bias-corrected method, accelerated method, and the studentized method, as well as others. I will demonstrate how these bootstrap methods work using "toy examples," which will be datasets in which a specific bootstrap method is appropriate. To demonstrate my understanding of the methods, I will write a simulation to compare a few of them and determine how they perform against each other. I will show my understanding of the different methods of creating confidence intervals from bootstrapped data by communicating the statistical theory in a concise and accessible way to my peers. Writing the simulation and sharing conclusions will demonstrate my ability to implement statistical methods in practice as well as my ability to analyze the results of the simulation.

Keywords: 3 to 6 keywords, that do not appear in the title

*The authors gratefully acknowledge ...

1 Introduction

This template will be used for you to submit your final project. You'll need to install the *rticles* package, make sure you have the agsm.bst file and bibliography.bib file, and the gfx folder and figure file in order to compile. If you make your own bibliography.bib file later, that's fine - change the name above to match your file name. You'll want the setup for the file to look the same in your repo as in the class repo, and then compile to asa_article when you knit. You should change the name of the file to something other than "test" though. Remember that if something goes wrong with the file, you can always look back at the class repo for the original files and their structure.

This template was adapted from the template Prof. Horton provided Stat 495 in Fall 2021, used with permission.

2 Exposition

The portion of the paper where you describe the new technique as though you were teaching it to a classmate (with whatever name you want to give it - Literature, Background, or Exposition) will be based on your annotated bibliography. You can add sources to what you already have in the annotated bibliography (especially if my feedback says you need to!). To be sure that you are on track with the exposition, a full draft of this portion of the paper is due before Thanksgiving break. This will also help with incorporating proper citation early in the writing process, as you'll have a bibliography that we can implement for this portion of the paper at a minimum.

Exposition -> what is a bootstrap –> parametric –> non-parametric –> briefly: what is jackknife nah –> SE and bias? -> what is a bootstrapped confidence interval –> what can it do -> how do we create them -> different types of bootstrapped CI –> standard t interval, approx classical, swapped SE, percentile, Bias-corrected, BCa —> for sample mean or corr (check example sheet)

Bootstrapping is a statistical method of resampling that allows the estimation of a test statistic from an unknown distribution. In particular, bootstrapping is a computational heavy method which is useful for many different situations.

First, we will introduce the non-parametric bootstrap. Suppose we have a random sample $X = (x_1, x_2, \ldots, x_n)$ from our unknown distribution, F and a statistic of interest $\hat{\theta} = \hat{\theta}(X)$. Ideally, the desired test statistic could be found by repeatedly sampling new reproductions of X from F. However F is unknown so this is not possible. The non-parametric bootstrap creates an estimate $\hat{F}$ from F using our sample X without making any parametric assumptions (such as distribution). Therefore, the bootstrap sample could be represented as $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$ where each x_i^* is sampled randomly with equal probability and with replacement from $\{x_1, x_2, \ldots, x_n\}$. From this bootstrap sample, a bootstrap replication of the test statistic can be computed using $\hat{\theta}^* = \hat{\theta}(X^*)$. A large number, B, of bootstrap samples are drawn independently and the corresponding bootstrap replication of the test statistic is calculated.

$$\hat{\theta}^{*b} = \hat{\theta}(X^{*b}) \text{ for } b = 1, 2, \ldots, B$$

. The bootstrap estimate of the test statistic is the empirical value of the test statistic from all of the $\hat{\theta}^{*b}$ replications. As B increases, $\hat{F}$ approaches F which means that the tests statistic of interest approaches its true value as well.

SHOULD I ADD ABOUT PARAMETRIC BOOTSTRAP

Confidence intervals are tools that are used to estimate a parameter. Specifically, a confidence interval gives a range in which the true value of the parameter may lie. An α-level standard confidence interval is given by

$$\hat{\theta} \pm z_\alpha \hat{\sigma}$$

where $\hat{\theta}$ is a point estimate of the parameter of interest θ, $\hat{\sigma}$ is the estimate of the standard deviation of $\hat{\theta}$ and z_α is the $(100 * \alpha)$th percentile of the normal deviation. We say that the confidence interval constructed in this manner has a chance of capturing the true parameter with a probability of α.

The standard confidence interval is built based on the assumption that the distribution from which we are sampling is Normal. This means that for an unknown distribution, the standard confidence interval could present an incorrect range. However, the same process can be used with bootstrap sampling to form the bootstrap percentile method. This means that an approximate bootstrap confidence interval will be created in the same automatic

way that the standard confidence interval was created. For bootstrapped confidence intervals, the number of bootstrap replications B must be large (around 2000) due to the nature of confidence intervals requiring greater accuracy.

The percentile method interval is defined as the interval between the $100 * \alpha$ and the $100(1-\alpha)$ percentiles of the bootstrap distribution of $\hat{\theta}$. That is, the $(1-2\alpha)$ coverage interval can be defined as $[\hat{\theta}^*_\alpha, \hat{\theta}^*_{1-\alpha}]$. To go further, we can say define $\hat{G}(t)$ as the bootstrap cdf, or the proportion of bootstrap samples less than t.

$$\hat{G}(t) = \frac{\#\{\hat{\theta^{*b}} \leq t\}}{B}$$

Thus the αth percentile point of the distribution is given by

$$\hat{\theta}_p[\alpha] = \hat{\theta^*_\alpha} = \hat{G}^{-1}(\alpha).$$

It follows that the percentile interval can be represented as

$$\left[\hat{G}^{-1}(\alpha), \hat{G}^{-1}(1-\alpha)\right].$$

In the case that the bootstrap distribution of $\hat{\theta}^* \sim N(\hat{\theta}, \hat{\sigma}^2)$, the corresponding percentile interval would be equivalent to the standard interval. However, this is not usually the case. When the bootstrap distribution is non-normal, we can suppose that there exists, for all θ,

$$\hat{\phi} \sim N(\phi, \tau^2),$$

for some monotone transformation $\hat{\phi} = g(\hat{\theta}), \phi = g(\theta)$, and τ is a constant. In other words, this transformation perfectly normalizes the distribution of $\hat{\theta}$. This transformation invariant can be applied to the bootstrap replications such that

$$\hat{\phi}^{*b} = g\left(\hat{\theta}^{*b}\right) \text{ for } b = 1, 2, \ldots, B.$$

The corresponding percentiles of the distribution transform similarly, $\hat{\phi}^*_\alpha = g\left(\hat{\theta}^*_\alpha\right)$. Or we can say that the $(1-2\alpha)$ percentile interval is $\hat{\phi} \pm \tau z_\alpha$ which can also be represented as $[\hat{\phi}^*_\alpha, \hat{\phi}^*_{1-\alpha}]$. This means that the interval on the θ scale can be defined as

$$\hat{\theta}^*_\alpha = g^{-1}(\hat{\phi} \pm \tau z_\alpha).$$

Or $\left[g^{-1}(\hat{\phi} \pm \tau z_{1-\alpha}), g^{-1}(\hat{\phi} \pm \tau z_{\alpha})\right]$. Therefore, the percentile method produces a correct interval for ϕ and due to the transformation invariance, also produces a correct percentile interval for θ. This method assumes the existence of some monotone normalizing mapping $\hat{\phi} = g(\hat{\theta}), \phi = g(\theta)$ and relies on that to create a correct interval. Since the process is automatic, we do not need to know the transformation itself, only that it exists. However, in some cases, no monotone normalizing mapping will exist.

The next method we will be looking at is the bias-corrected percentile method (BC method) which is an improvement upon the previous percentile method because we now take into account the possibility of bias. It can be shown (SHOULD I SHOW LATER?) that $\hat{\theta}$ is biased upwards relative to θ which means that the confidence intervals should be adjusted downwards. From our simulated bootstrap replications $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$, define $p_0 == \frac{\#\{\hat{\theta^{*b}} \leq \theta\}}{B}$, and define the bias-correction value $z_0 = \Phi^{-1}(p_0)$ where Φ^{-1} is the inverse function of the standard normal cdf. Thus we define a transformation $\hat{\phi} = g(\hat{\theta}), \phi = g(\hat{\theta})$ such that for any θ, $\hat{\phi} \sim N(\phi - z_0\tau, \tau^2)$, with z_0 and τ constants. This means that we can say the bias corrected method has an α-level endpoint can be represented as

$$\hat{\theta}_{BC}[\alpha] = \hat{G}^{-1}\left[\Phi\left(2z_0 + z_\alpha\right)\right].$$

If $\hat{G} = 0.50$, then half of the bootstrap distribution is less than $\hat{\theta}$ and our bias-correction value $z_0 = 0$. In this case, the confidence interval produced by BC would be the same interval produced by the percentile method.

A further modification upon the BC interval is the bias corrected and accelerated method (BCa). For this method, we do assume the the standard error, τ to be constant as we did in the BC interval. Rather, we assume the existence of a monotone transformation $\hat{\phi} = g(\hat{\theta}), \phi = g(\hat{\theta})$ such that for any θ, $\hat{\phi} \sim N(\phi - z_0\tau_\phi, \tau_\phi^2)$ where $\tau_\phi = 1 + a\phi$. The a in known as the acceleration value and is a constant that describes how the standard deviation of $\hat{\phi}$ varies with ϕ. In other words, a is proportional to the skewness of the bootstrap distribution. For example, for one-parameter exponential families, $a = z_0$, however, there are many different algorithms to compute and estimate a. Now, our α-level endpoint from BCa is

$$\hat{\theta}_{BCa}[\alpha] = \hat{G}^{-1}\left[\Phi\left(z_0 + \frac{z_0 + z_\alpha}{1 - a(z_0 + z_a)}\right)\right].$$

If $a = 0$, then $\hat{\theta}_{BCa}[\alpha] = \hat{\theta}_{BC}[\alpha]$. When calculating a BCa interval, the acceleration value a is not a function of the bootstrap distribution and must be calculated separate, however the process is algorithmic and can be calculated without too much work. Each of the three previous methods (percentile, BC, and BCa) all build upon each other and have less restrictive assumptions, however computation increases as we loosen assumptions.

3 More on the template

This template demonstrates some of the basic LaTeX commands and syntax you'll need to know to use the `rticles` package to generate a readable report using R Markdown. Markdown allows various formatting and you can find formatting cheatsheets or guides online to assist as well. Here's an example with bullets and some advice:

- I would encourage you to look closely at this file and explore the various parts and pieces.
- I would suggest that you format your Rmd file so that you have only one sentence per line.
- It makes it *much* easier to see changes in your GitHub commits.

4 Verifications

This section will be just long enough to illustrate what a full page of text looks like, for margins and spacing.

Note that we can refer to sections (e.g., this is section 4, while the previous section was section 3).

Note that we should refer to work using the BibTeX system. Here we can reference papers by Campbell & Austin (2002) and Schubert et al. (2013) through inline citations (see the `bibliography.bib` file for the reference database).

More work that is relevant can also be cited in a traditional fashion (Chi et al. 1981, Galyardt (2014),Galyardt (2012)).

Note that you can capitalize proper nouns in citations (Campbell & Austin 2002) (again, see `bibliography.bib`).

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. **With this spacing we have 30 lines per page.**

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick

brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

5 Examples

Lots of things can be done in this template.

Table 1: This is a sample table caption that should interpret the table!

boro	count
BRONX	1903
BROOKLYN	5294
MANHATTAN	8340
Missing	6
QUEENS	4789
STATEN ISLAND	802

```
Violations %>%
  select(dba, boro) %>%
  unique() %>%
  group_by(boro) %>%
  summarize(count = n(), .groups = "drop") %>%
  knitr::kable(
    caption = "This is a sample table caption that should interpret the table!")
```

Table 1 displays the number of dba's per borough.

```
Violations %>%
  select(dba, boro, inspection_date, score) %>%
  unique() %>%
  group_by(boro) %>%
  summarize(
    count = n(),
    averagescore = mean(score, na.rm = TRUE),
    .groups = "drop"
  ) %>%
  ggplot(aes(x = boro, y = averagescore)) +
  geom_boxplot()
```

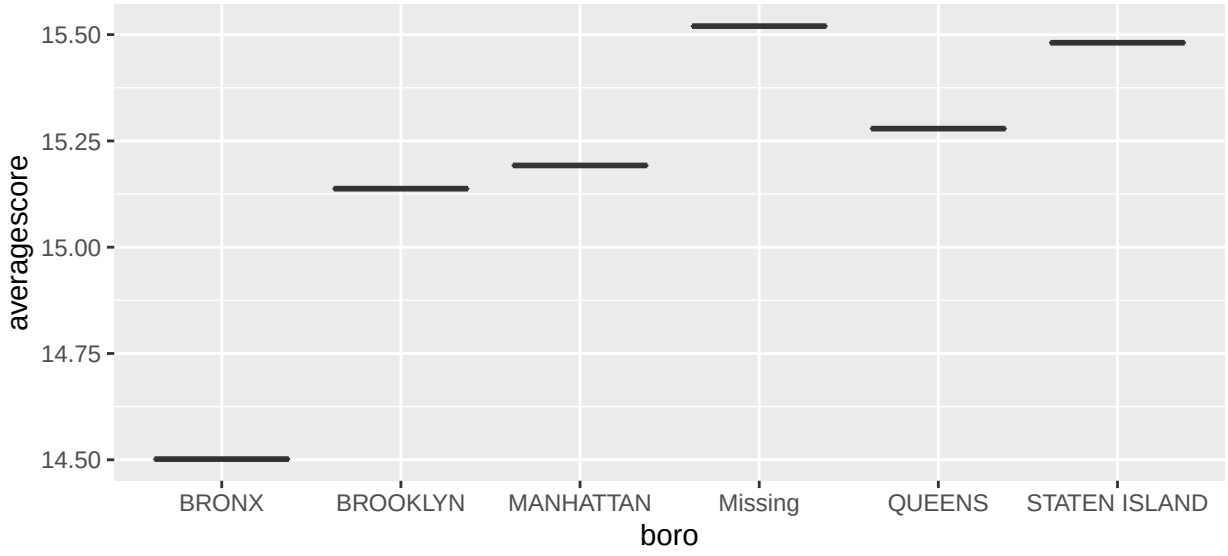

Figure 1: This is a sample figure caption that should interpret the figure!

Figure 1 displays the average inspection score by borough.

Figure 2 displays a campus map.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog. The quick brown fox jumped over the lazy dog.

6 Your Choice of Structure

You are responsible for choosing the structure you want for the report, i.e. what sections and their names. Here are some examples that you could adapt, as appropriate. These are just examples - the first has more sections, some of which may make more sense as

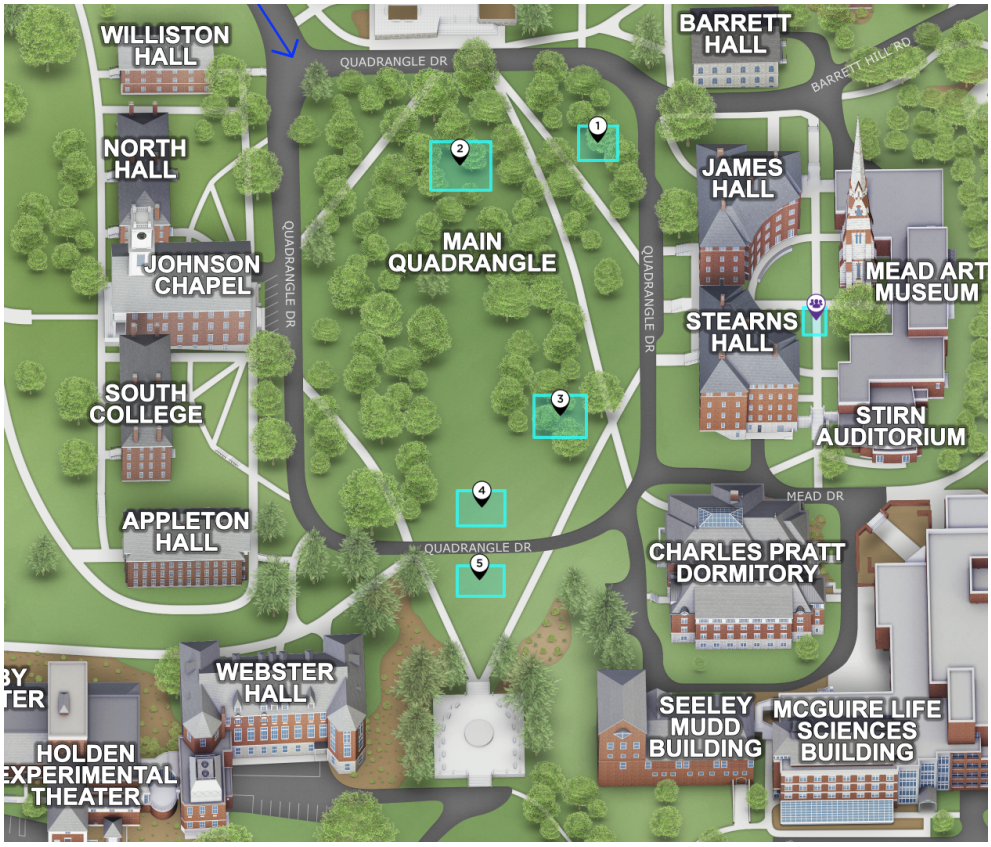

Figure 2: XX another sample figure caption

sub-sections of others. I'm just trying to give you examples. You don't have to have a conclusion or an introduction, etc. You'll need to find what works for you.

6.1 Example 1

- Introduction
- Background
- Topic X (Main exposition section)
- Applications in Literature
- Application to Data (your data)
- Conclusion

6.2 Example 2

- Introduction (including any relevant Background)
- Topic X (exposition and examples)
- Simulation

References

Campbell, J. I. & Austin, S. (2002), 'Use of Python and R to assess effects of response time deadlines on adults' strategy choices for simple addition', *Memory & Cognition* **30**(6), 988–994.

Chi, M. T., Feltovich, P. J. & Glaser, R. (1981), 'Categorization and representation of physics problems by experts and novices', *Cognitive Science* **5**(2), 121–152.

Galyardt, A. (2012), Mixed Membership Distributions with Applications to Modeling Multiple Strategy Usage, PhD thesis, Carnegie Mellon University, Pittsburgh, PA 15213. **URL:** *https://nas.edu/envisioningds*

Galyardt, A. (2014), Interpreting mixed membership models: Implications of erosheva's representation theorem, *in* E. M. Airoldi, D. Blei, E. Erosheva & S. E. Fienberg, eds, 'Handbook of Mixed Membership Models in R', Chapman and Hall.

Schubert, C. C., Denmark, T. K., Crandall, B., Grome, A. & Pappas, J. (2013), 'Characterizing novice-expert differences in macrocognition: an exploratory study of cognitive work in the emergency department', *Annals of Emergency Medicine* **61**(1), 96–109.