bootEd: An R Package for Introducing Students to Bootstrap Intervals With Assumption Checking

Njesa Totty

PhD Student
Department of Statistics
Oregon State University

USCOTS 2021

Introduction

- Using R to teach simple bootstrap intervals
- Percentile, basic, and studentized bootstrap intervals
- Understand assumptions and perform simulations with R

Reasons to Teach Bootstrapping

- Textbooks: Tibshirani and Efron 1993; Field et al. 2012; Ismay and Kim 2019; Lock et al. 2020
- Retaining more concepts (Tintle et al. 2012)
- Active learning and relevance (Wood 2005; Mills 2002)

Reasons to Teach Bootstrapping

'Students should demonstrate an understanding of, and ability to use, basic ideas of statistical inference, both hypothesis tests and interval estimation, in a variety of settings.'

- GAISE College Report ASA Revision Committee 2016

Issues with Teaching Bootstrapping

Some misconceptions about the simple bootstrap (Hayden 2019):

- More accurate for small samples
- More accurate for non-normal populations
- Works for any statistic
- **Fewer or no assumptions**
- Easier for students to understand

Simple Bootstrap Intervals

Use $\hat{\theta}$, an estimate of θ, to construct a 95% interval estimate:

1. Denote 0.025 and 0.975 quantiles of $\hat{\theta} - \theta$ as $a_{0.025}$ and $a_{0.975}$
2. 95% equi-tailed interval for θ is $(\hat{\theta} - a_{0.975}, \hat{\theta} - a_{0.025})$
3. Use bootstrapping to estimate $a_{0.025}$ and $a_{0.975}$:
   - If $t_p = p$-th quantile of sampling distribution, then $a_p = t_p - \theta$
   - Bootstrap analogue: $a_p^* = t_p^* - \hat{\theta}$
   - Use ordered values $t^*_{(B+1)p}$, where $B =$ number of bootstrap samples, for better estimates

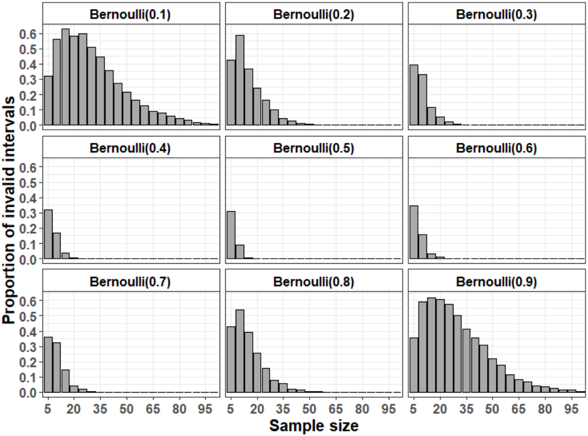
A. C. Davison and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Vol. 1. Cambridge University Press

Bootstrap Intervals Overview

	Interval	Assumptions
Basic	$(2\hat{\theta} - t^*_{(B+1)0.975}, 2\hat{\theta} - t^*_{(B+1)0.025})$	Bootstrap distribution is good estimate of sampling distribution
Percentile	$(t^*_{(B+1)(0.025)}, t^*_{(B+1)(0.975)})$	Symmetry in sampling distribution
Studentized	$\left(\hat{\theta} - z^*_{(B+1)(1-\alpha/2)}\hat{\sigma}, \quad \hat{\theta} - z^*_{(B+1)(\alpha/2)}\hat{\sigma}\right)$	Normal approximation reasonable starting point

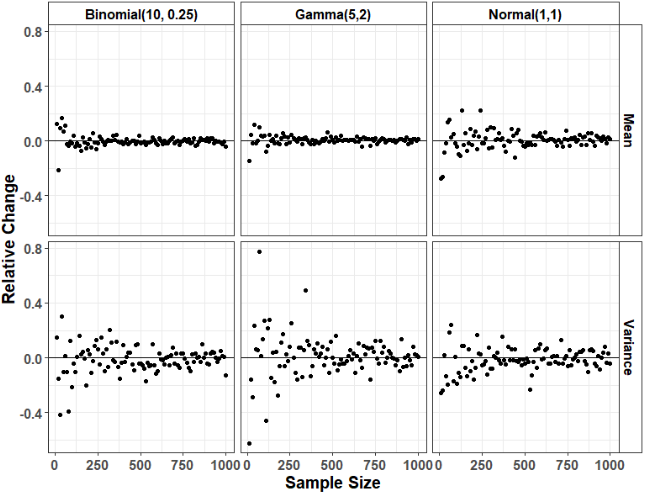
Simulation Results

Bootstrap distribution inconsistent as estimate of sampling distribution in certain scenarios (Andrews 2000)



Simulation Results

Relative change between 0.95 quantile of sampling distribution and bootstrap distribution



Simulation Results

Coverage proportion of 95% bootstrap intervals for the population mean

Method	N	B	Normal(1, 1)	Gamma(5, 2)	Binomial(10, 0.25)
Percentile	40	99	0.941	0.949	0.939
		999	0.934	0.952	0.957
	100	99	0.950	0.948	0.950
		999	0.934	0.952	0.957
Basic	40	99	0.938	0.937	0.945
		999	0.935	0.930	0.943
	100	99	0.950	0.942	0.949
		999	0.944	0.941	0.942
Student-ized	10	99	0.959	0.960	0.955
		999	0.951	0.953	0.960
	40	99	0.964	0.959	0.957
		999	0.953	0.955	0.958

Simulation Conclusions and Other Notes

- Underlying assumptions exist
- Ignoring assumptions can be problematic
- Population, sample size, number of bootstrap samples impact severity

The bootEd Package

```
install.packages("devtools")
devtools::install_github("tottyn/bootEd")
library(bootEd)
```

- Communicates assumptions to users
- Lightweight: depends, imports, suggests
- Simple documentation and function syntax
- Starter vignette

Implementing bootEd

Code:

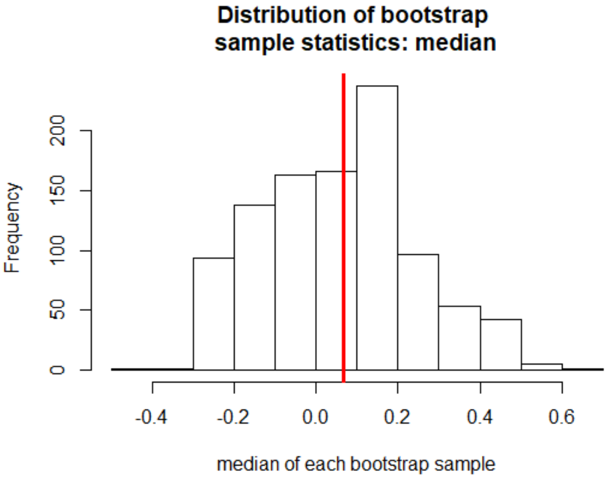
```
percentileMBI(sample = rnorm(100), parameter = "median", B
= 999, siglevel = 0.05, onlyint = FALSE)
```

Output:

```
The percentile bootstrap interval for the median is:
(-0.2199535, 0.4150477).  If it is not reasonable to assume
that the sampling distribution of the statistic of interest
is symmetric this method should not be used.
```

Implementing bootEd

Output of percentileMBI function - bootstrap distribution



Other Package Options

Simplicity of use for bootstrapping:

R Package	Installation	Syntax	Implementation
`boot`	✓		✓
`wboot`	✓	✓	✓
`simpleboot`	✓	✓	✓
`bootstrap`	✓	✓	✓
`bootEd`	✓	✓	✓
`mosaic`		✓	✓
`resample`	✓	✓	✓

Learn More and Contribute

Package: github.com/tottyn/bootEd

Contact: tottyn@oregonstate.edu

Acknowledgments: Advisors Claudio Fuentes and James Molyneux and Senior Instructors, Julie Moore and Katie Jager

References I

Andrews, D. W. K. (2000). 'Inconsistency of the Bootstrap When a Parameter is On the Boundary of the Parameter Space'. In: *Econometrica*, pp. 399–405.

Davison, A. C. and D. V. Hinkley (1997). *Bootstrap Methods and Their Application*. Vol. 1. Cambridge University Press.

Field, A., J. Miles, and Z. Field (2012). *Discovering Statistics Using R*. Sage Publications.

GAISE College Report ASA Revision Committee (2016). *Guidelines for Assessment and Instruction in Statistics Education College Report 2016*. URL: http://www.amstat.org/education/gaise.

Hayden, R. W. (2019). 'Questionable Claims for Simple Versions of the Bootstrap'. In: *Journal of Statistics Education* 27.3, pp. 208–215.

Ismay, C. and A. Y. Kim (2019). *Statistical Inference via Data Science: A Modern Dive into R and the Tidyverse*. CRC Press. URL: https://moderndive.com.

References II

Lock, R. H, P. F. Lock, K. L. Morgan, E. F. Lock, and D. F. Lock (2020). *Statistics: Unlocking the Power of Data*. John Wiley & Sons.

Mills, J. D. (2002). 'Using Computer Simulation Methods to Teach Statistics: A Review of the Literature'. In: *Journal of Statistics Education* 10.1.

Tibshirani, R. and B. Efron (1993). 'An Introduction to The Bootstrap'. In: *Monographs on Statistics and Applied Probability* 57, pp. 1–436.

Tintle, N. L., K. Topliff, J. VanderStoep, V. Holmes, and T. Swanson (2012). 'Retention of Statistical Concepts in a Preliminary Randomization-Based Introductory Statistics Curriculum'. In: *Statistics Education Research Journal* 11.1, p. 21.

Wood, M. (2005). 'The Role of Simulation Approaches in Statistics'. In: *Journal of Statistics Education* 13.3.