

Engaging Students Using Simulation-Based Introductions to Computational Statistical Methods

Njesa Totty, PhD

Assistant Professor
Department of Mathematics
Framingham State University

eCOTS 2024

Reasons to Teach Statistical Computing

“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

- Curriculum incorporating simulations enabled students to apply material in new contexts (Son et al. 2021)
- Randomization-based curriculum increased retention (Tintle et al. 2012)
- Strong presence and support in discussions on statistics in the undergraduate curriculum (e.g. Nolan and Temple Lang 2010; Nolan and Lang 2015; Leman et al. 2015; Horton and Hardin 2015)

Teaching the Simple Bootstrap

- Statistical computing technique by Efron (1979) for estimating the standard deviation of a sample statistic
 - Supports student understanding of sampling distributions and confidence intervals (R. H. Lock and P. F. Lock 2008)
 - Equips students to make statistical inference in a variety of scenarios (Howington 2017)
- Does have underlying assumptions which could be misinterpreted if not carefully communicated (e.g. Hesterberg 2015; Hayden 2019)
- Using simulations to teach simple bootstrap could help students understand:
 - Assumptions behind methods
 - Consequences of inappropriate use

Percentile Interval Assumption

Let $\hat{\theta}(\mathbf{X})$ be an unobserved sample estimate for the population parameter θ , $\hat{\theta}(\mathbf{x})$ the observed statistic, $\hat{\theta}(\mathbf{x}^*)$ a bootstrap estimate, and $\alpha \in (0, 1)$ the significance level.

Assumption: There exists some monotone transformation $\hat{\phi} = m(\hat{\theta}(\mathbf{X}))$ such that $\hat{\phi} \sim \text{Normal}(\phi, c^2)$ for all population distributions F , including the case $F = \hat{F}$, where $\phi = m(\theta)$, for some standard deviation c .

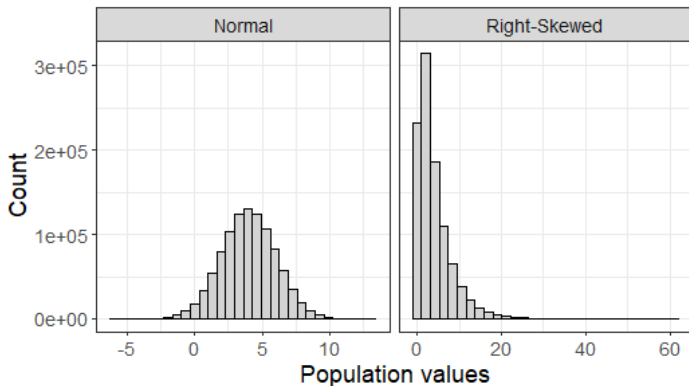
The $(1 - \alpha) * 100\%$ *percentile interval* given by (Efron and Tibshirani 1993) is

$$(r_{\alpha/2}^*, r_{1-\alpha/2}^*),$$

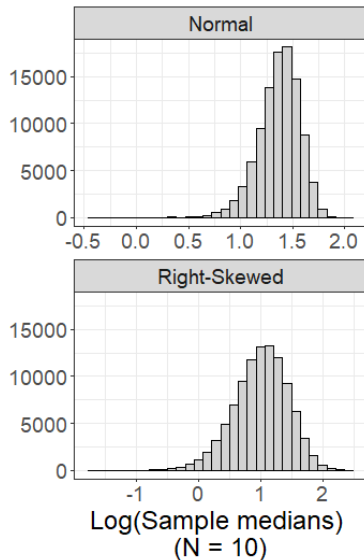
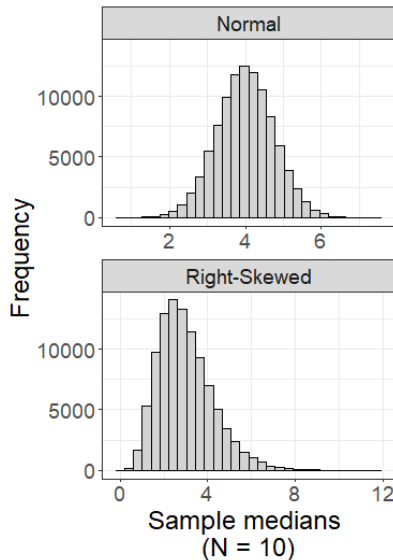
where the p quantile of $\hat{\theta}(\mathbf{x}^*)$ is r_p^* .

Generating Population Data

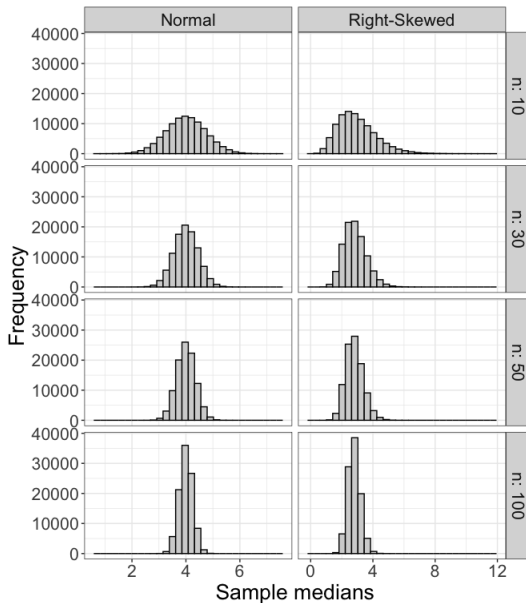
```
# set seed for reproducibility
set.seed(9853)
# set number of draws
ndraw <- 1e+06
# draw from various populations
popdat <- data.frame(values = c(rnorm(ndraw, mean = 4, sd = 2),
                                rexp(ndraw, rate = 0.25)),
                     pop = rep(c("Normal", "Right-Skewed"), each = ndraw))
```



Transforming Sampling Distribution of Median



Varying Sample Size



Coverage Proportions

	Normal	Right-skewed
N = 10	0.946	0.954
N = 30	0.939	0.975
N = 50	0.948	0.934
N = 100	0.94	0.944

Table: Coverage proportions of 1000 two-sided 95% percentile intervals for the median. Samples of size N were drawn from the underlying populations. The number of bootstrap samples was 999. Students can perform simulations such as these to learn how to assess the performance of the percentile interval in different scenarios.

Variability in Coverage

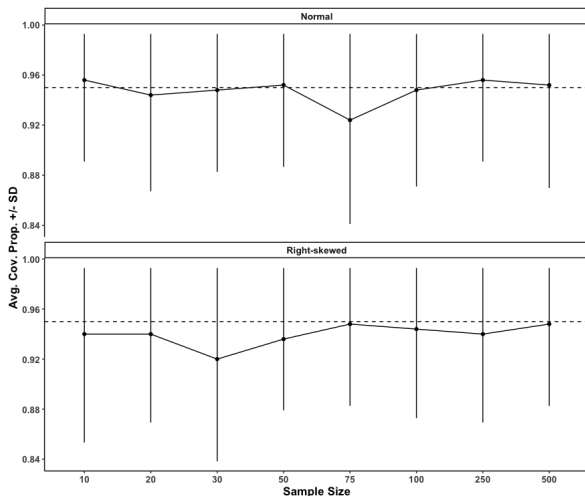


Figure: Average and standard deviation of 25 coverage proportions each coming from 10 95% two-sided percentile intervals for the median using $B = 999$ bootstrap samples.

Main Takeaways

- Teaching statistical computing methods has pedagogical benefits
- Assumptions impact performance
- Enhance teaching by communicating assumptions

Thank you!

More info and references available in pre-print:

- <https://arxiv.org/abs/2112.07737>

Connect or provide feedback:

- ntotty@framingham.edu
- <https://www.linkedin.com/in/njesatotty/>