Investigating coverage of 1-sample proportion theoretical and bootstrap methods

Connor Demorest

4/16/2021

Introduction:

In introductory statistics classes, we learn all about all these theory-based, parametric methods of testing hypotheses and creating confidence intervals that *generally* work in many situations. When we make these parametric models of our data, we then learn about the assumptions that we are using; typically, it's something to do with having normal data or close enough, or using the Central Limit Theorem to our advantage to assume it's close enough, as well as independence of observations.

In this project, I wanted to examine the long-run probability of coverage for a 1-sample proportion percentile bootstrap confidence interval (called the "bootstrap method" from now on) as described in the Montana State Introductory Statistics with R textbook compared to the normal approximation to the binomial distribution and using normal distribution quantiles (which will be called the "theory-based method" from now on). I also wanted to compare the coverage probabilities of both the bootstrap method and the theory-based methods to the rule of thumb that $n * p > 5\ (or\ 10)$ should be sufficient to have good coverage properties.

These inequalities $n * p > 5$ and $n * (1 - p) > 5$ seem to come out of nowhere. From a search on the internet, the only mention I can find of the origin of this rule of thumb comes from a paper by Cochran (1952), where he discusses the χ^2 goodness of fit test developed by Neyman and Pearson. In that test of the goodness of fit, there is a similar assumption that the expected value for any one cell should be greater than 5 or 10, which Cochran describes as "…appear to have been arbitrarily chosen" by Neyman and Pearson.

Figure 1 shows the sampling distributions for many values of $n * p$. As the values of $n * p$ decrease, there is "bunching up" near 0 that would be an issue if a normal approximation was used to model the probability of success. This plot makes it appear that the assumption that $n * p > 10$ may be too strong of an assumption, and $n * p \approx 7.5$ may be enough to assume that theoretical methods are reasonable to assume are accurate.

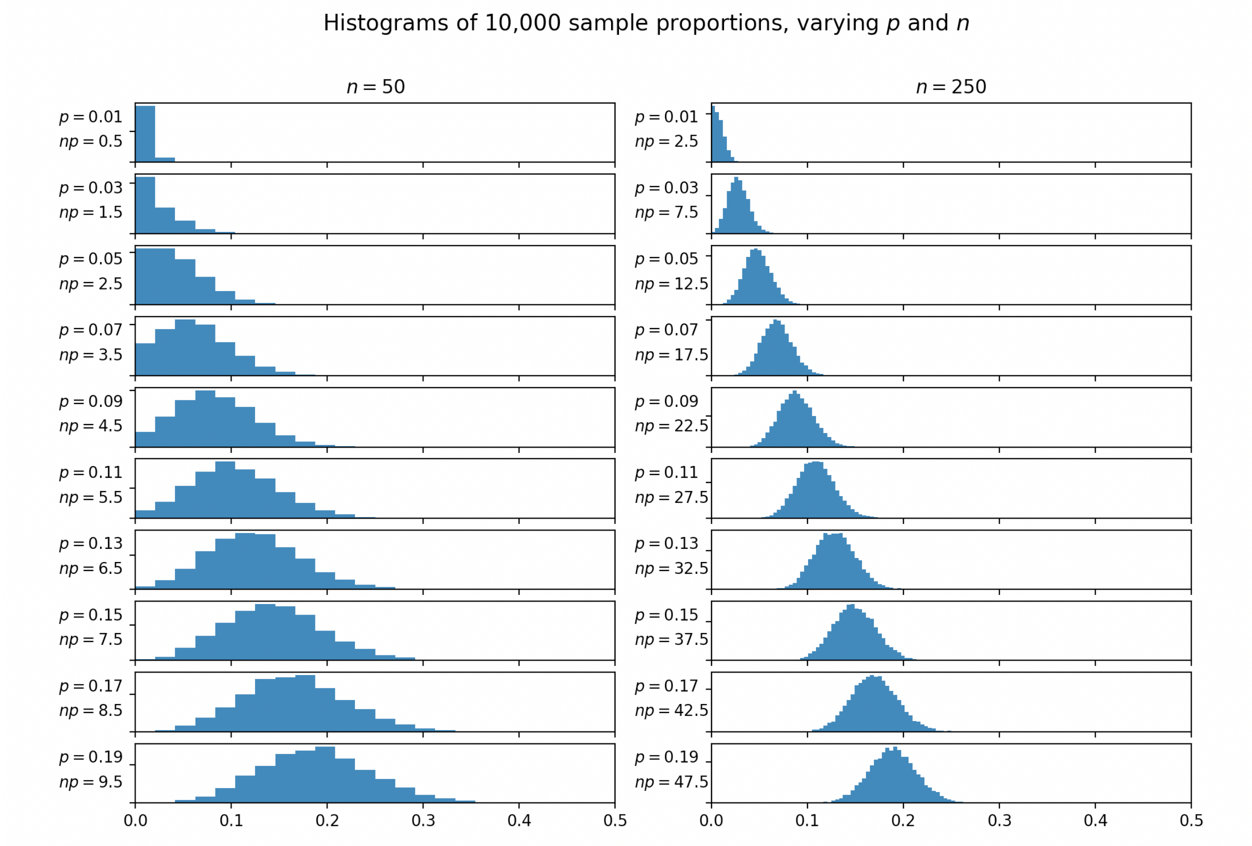


Figure 1: Sampling distribution of the binomial for different n and p

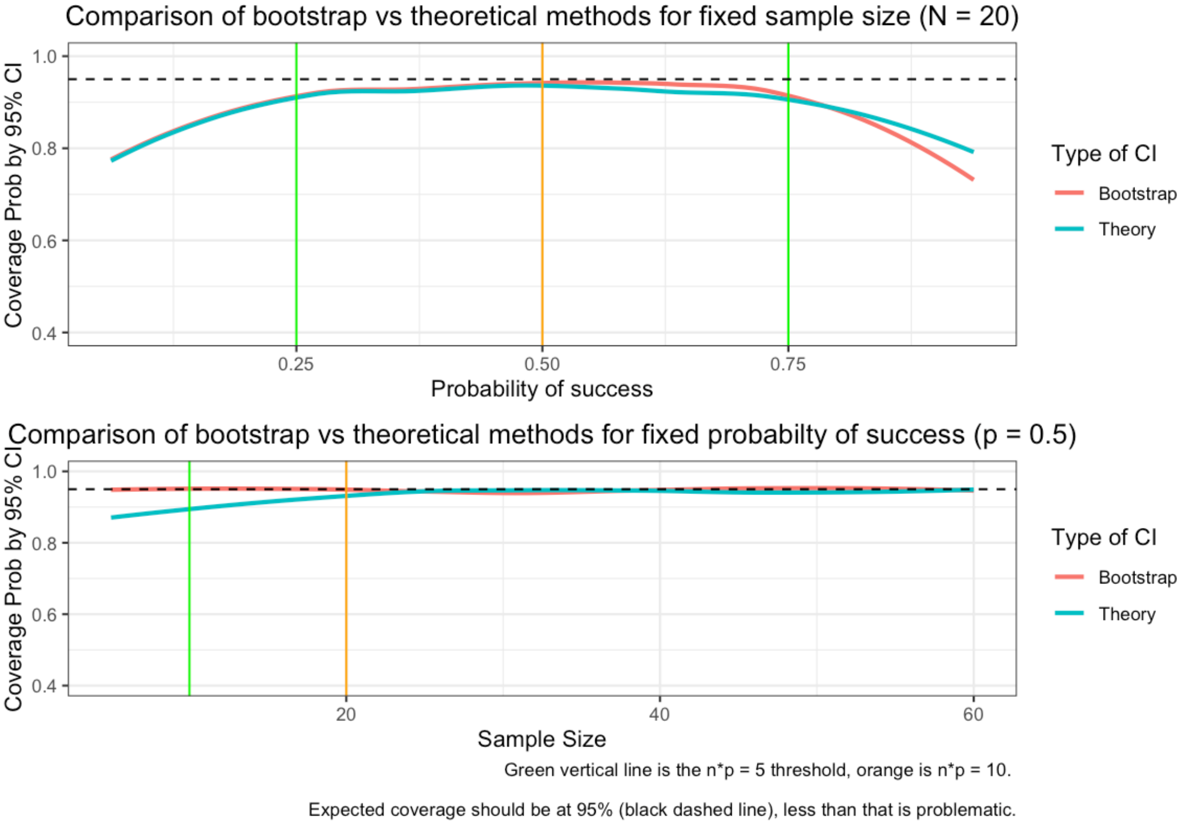
Another method of creating confidence intervals and testing hypotheses is using bootstrap resamples of the empirical distribution to create the "bootstrap distribution" that should approximate the sampling distribution after many resamples. To create a 95% confidence interval, we could then use the 2.5th and 97.5th percentile of the bootstrap distribution to get the approximate 95% confidence interval. This is a method used in the Stat 216 textbook (Carnegie, Hancock, et al, 2021). This bootstrap resampling percentile method of creating confidence intervals has been shown to be less accurate than theory-based methods for small sample sizes, but more accurate for larger sample sizes. (Hesterberg, 2015). However, in that article, Hesterberg argues that there are pedagogical advantages to using these bootstrap methods compared to theory based methods.

There seems to be a gap in the literature regarding the coverage properties of the bootstrap methods compared to a for a 1-proportion test, all of the literature I could find discussed the situation of a t-test compared to the bootstrap methods for the difference in two sample means. I want to investigate if my assumption that the same sorts of trends that Hesterberg describes for the two sample means follow for the 1-sample proportion. My presumption is that since both differences in means and sample proportions use the Central Limit Theorem to use approximations for non-normal data to follow normality, the trends should continue.

Methods:

To determine the coverage probabilities of the theory-based method compared to the bootstrap method of creating a confidence interval, I wrote an RShiny app that allows the user to input different values of the probability of success, sample size, confidence level, and number of replications and compares the coverage of a confidence interval of different sizes for the theory based and bootstrap methods. I used the bootstrap method as described in the Montana State Introductory Statistics textbook from the "catstats" R package.

Figures 2 and 3 show examples of the plot comparing the probability of coverage of the unknown probability of success for different values of the probability of success (p) and sample sizes (n). Figure 2 shows the plots with n = 20 and p = 0.5, and Figure 3 shows the plots with n = 50 and p = 0.2. The vertical lines show the cutoff at $n * p = 5$ and $n * p = 10$ to compare the coverage probability at and below those two rules of thumb.

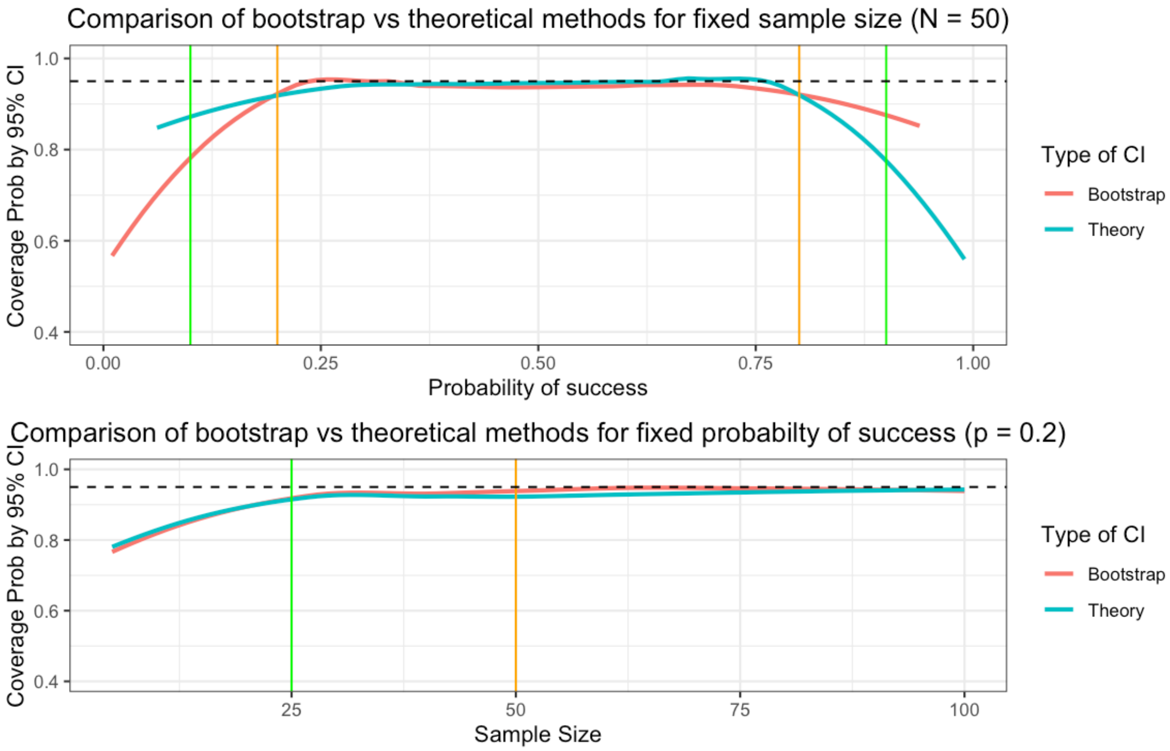

Comparison of bootstrap vs theoretical methods for fixed sample size (N = 20)

Comparison of bootstrap vs theoretical methods for fixed probabilty of success (p = 0.5)

Green vertical line is the n*p = 5 threshold, orange is n*p = 10.

Expected coverage should be at 95% (black dashed line), less than that is problematic.

Figure 2: Coverage of theory-based and bootstrap methods for n = 20 and p = 0.5


Comparison of bootstrap vs theoretical methods for fixed sample size (N = 50)

Comparison of bootstrap vs theoretical methods for fixed probabilty of success (p = 0.2)

Green vertical line is the n*p = 5 threshold, orange is n*p = 10.

Expected coverage should be at 95% (black dashed line), less than that is problematic.

Figure 3: Coverage of theory-based and bootstrap methods for n = 50 and p = 0.2

Results & Discussion:

From the Figures 2 and 3, we can see that the coverage probability for the bootstrap method and the theory-based method are very similar. At values where $n * p < 5$ to the outside of the green vertical lines, there is very low probability that a 95% confidence interval will actually cover a true unknown parameter value of the probability of success π, approaching as low as 75% for proportions near 0 or 1 with a sample size of 20, and hardly improving for sample sizes of 50. There is inconclusive graphical evidence for which of the bootstrap or theory-based methods have better coverage at the small sample sizes and extreme probabilities of success. If there was more time and computing power, I could run hundreds of thousands of replications to get more accurate estimates of the bootstrap and theory-based confidence interval coverage for different combinations of sample sizes and probability of success.

Examining the coverage properties, the of $n * p$ that are less than 5 generally have a large drop off in the coverage probability. The probability of coverage for a 95% confidence interval for values where $n * p > 10$ are nearly all right at the 95% with a small amount of sampling variability. For values where $n * p > 5$ & $n * p < 10$, it seems that there can be a slight loss of coverage power, but it's not clear exactly how much, or whether a theory-based or bootstrap method would be better. The pedagogical advantages that Hesterberg describes are certainly worth the, at most, minute difference in power between the two methods of creating a confidence interval.

It seems that the bare minimum rule of thumb is that $n * p > 5$ leads to around 90% coverage on a 95% confidence interval, so there's not a huge loss of power before that point, and that it's very similar for both bootstrap and theory-based methods. At $n * p > 10$ there is very good coverage by both the bootstrap and theory-based methods. The answer to the question of 'what should the rule of thumb be?' is, as is typical in statistics, it depends. The $n * p > 5$ should be considered an absolute lower bound, and $n * p > 10$ is probably too strong of an assumption. Maybe, as I proposed in the introduction, $n * p > 7.5$ is a reasonable middle ground.

References:

Journal articles

Cochran, W. (1952). The χ^2 Test of Goodness of Fit. The Annals of Mathematical Statistics, 23(3), 315-345. Retrieved April 16, 2021, from http://www.jstor.org/stable/2236678

Tim C. Hesterberg (2015) What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum, The American Statistician, 69:4, 371-386, DOI: 10.1080/00031305.2015.1089789

Textbooks

Carnegie, N., Hancock, S., Meyer, E., Schmidt, J., and Yager, M. (2021). Montana State Introductory Statistics with R. Montana State University.

R packages

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2020). shiny: Web Application Framework for R. R package version 1.5.0. https://CRAN.R-project.org/package=shiny

Baptiste Auguie (2017). gridExtra: Miscellaneous Functions for “Grid” Graphics. R package version 2.3. https://CRAN.R-project.org/package=gridExtra