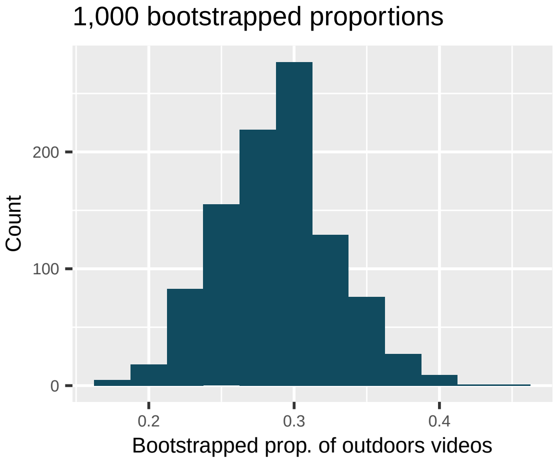
Problem Set 9

Confidence Intervals with the Bootstrap

1. **Outside YouTube videos.** Let's say that you want to estimate the proportion of YouTube videos which take place outside (define "outside" to be if any part of the video takes place outdoors). You take a random sample of 128 YouTube videos[1] and determine that 37 of them take place outside. You'd like to estimate the proportion of all YouTube videos which take place outside, so you decide to create a bootstrap interval from the original sample of 128 videos.



   a. Describe in words the relevant statistic and parameter for this problem. If you know the numerical value for either one, provide it. If you don't know the numerical value, explain why the value is unknown.

   b. What notation is used to describe, respectively, the statistic and the parameter?

   c. If using software to bootstrap the original dataset, what is the statistic calculated on each bootstrap sample?

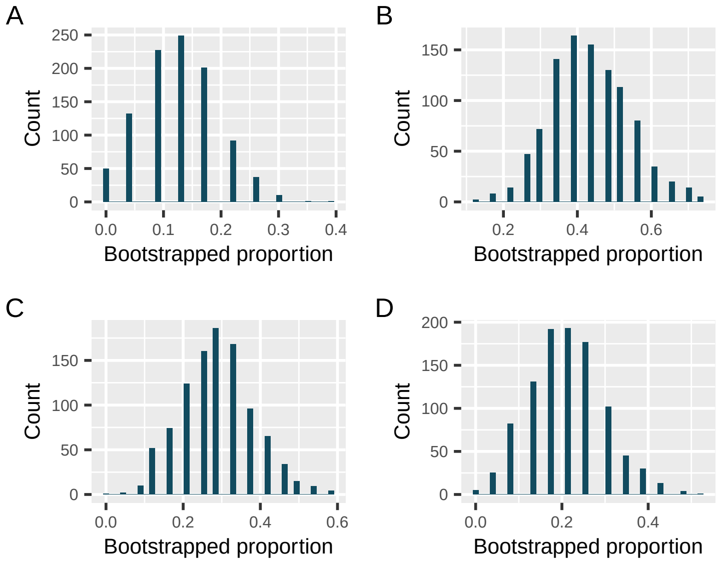
   d. When creating a bootstrap sampling distribution (histogram) of the bootstrapped sample proportions, where should the center of the histogram lie?

   e. The histogram provides a bootstrap sampling distribution for the sample proportion (with 1000 bootstrap repetitions). Using the histogram, estimate a 90% confidence interval for the proportion of YouTube videos which take place outdoors.

   f. In words of the problem, interpret the confidence interval which was estimated in the previous part.

2. **Bootstrap distributions of $\hat{p}$** Each of the following four distributions was created using a different dataset. Each dataset was based on $n = 23$ observations.

[1] There are many choices for implementing a random selection of YouTube videos, but it isn't clear how "random" they are.



Consider each of the following values for the true popluation p (proportion of success). Datasets A, B, C, D were bootstrapped 1000 times, with bootstrap proportions as given in the histograms provided. For each parameter value, list the datasets which could plausibly have come from that population. (Hint: there may be more than one dataset for each parameter value.)

a. $p = 0.05$

b. $p = 0.25$

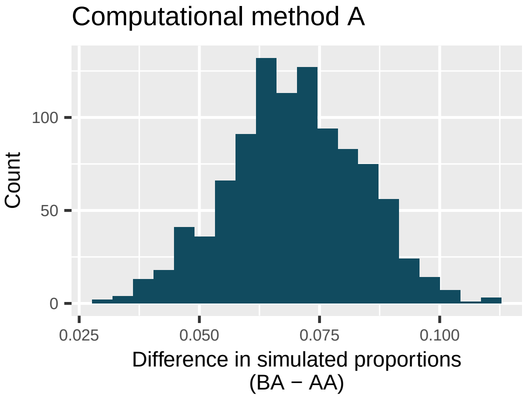
c. $p = 0.45$

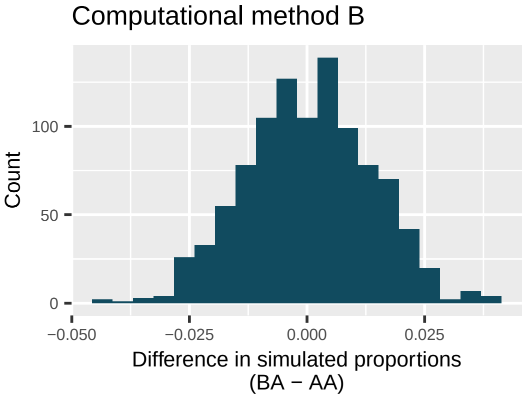
d. $p = 0.55$

e. $p = 0.75$

3. **COVID-19 and degree completion.** A 2021 Gallup poll surveyed 3,941 students pursuing a bachelor's degree and 2,064 students pursuing an associate degree (students were not randomly selected but were weighted so as to represent a random selection of currently enrolled US college students). The poll found that 51% of the bachelor's degree students and 44% of associate degree students said that the COVID-19 pandemic will negatively impact their ability to complete the degree. (Gallup 2021)

   Below are two histograms which represent different computational approaches (both use 1,000 repetitions) to research questions which could be asked from the Gallup data which was provided. One of the histograms can be used to do a randomization test on whether the proportions of bachelor's and associate students who think the COVID-19 pandemic will negatively impact their ability to complete the degree. The other histogram is a bootstrap distribution used to quantify the difference in the proportions of bachelor's and associate's students who feel this way.





a. Are the center and standard error of the two graphs approximately the same? Explain.

b. Write a research question which could be addressed using this histogram with computational method A.

c. Write a research question which could be addressed using this histogram with computational method B.

Gallup. 2021. "Half of College Students Say COVID-19 May Impact Completion." https://www.openintro.org/go?id=textbook-gallup-2021-covid-college-impact.