Bootstrap Confidence Intervals

3/26/25

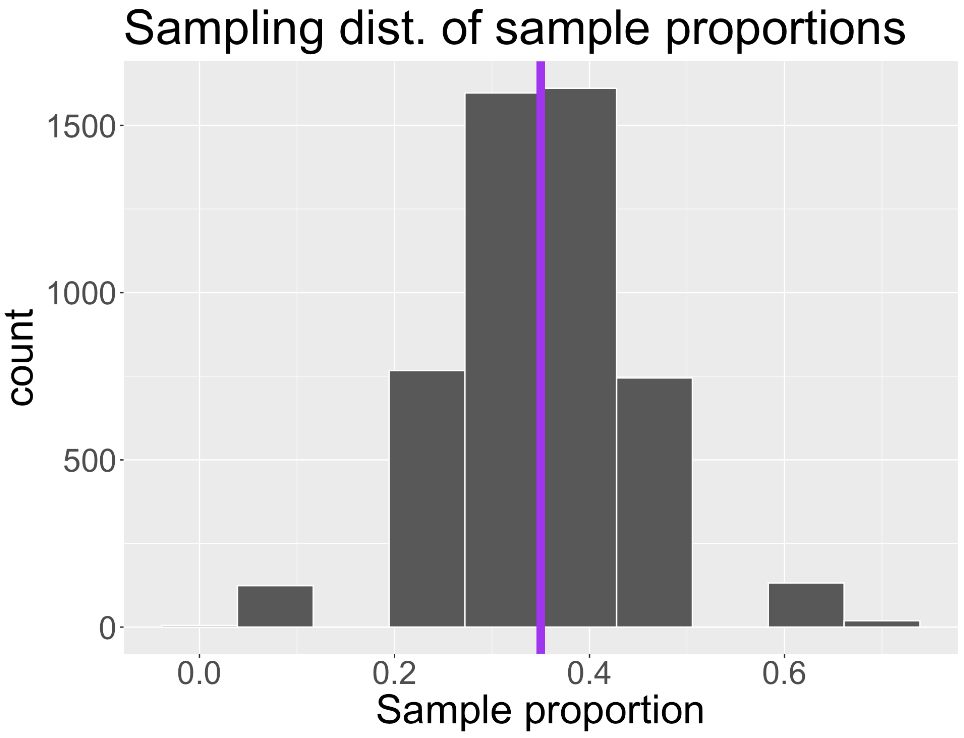
Housekeeping

- DataFest groups!

Samplint distribution recap

- Sampling distribution describes how statistic behaves under repeated sampling from population
- Recall research question from last class: what proportion of STAT 201A students drink coffee regularly?

- Since I took a census, I actually do have access to true sampling distribution of the sample proportion!
- I will repeatedly take SRS (i.e. without replacement) of $n = 10$ values from the population and calculate



purple line = true proportion

Bootstrap recap

If instead I could not repeatedly sample from population, we could obtain bootstrap distribution as an *approximation* of the sampling distribution of the statistic!

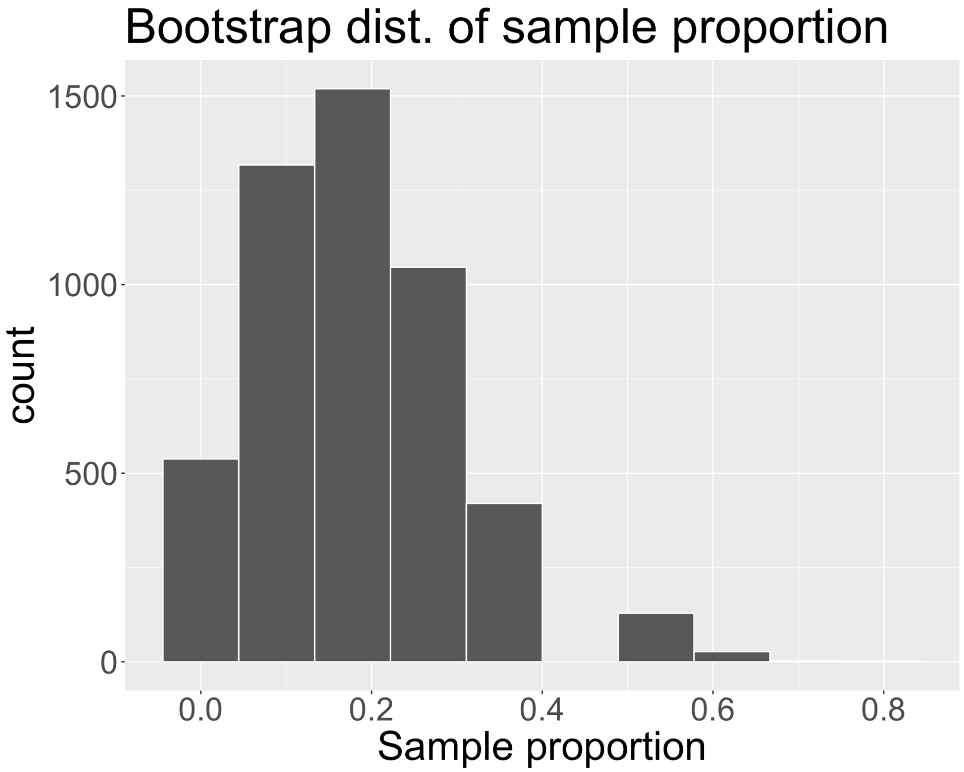
Procedure:

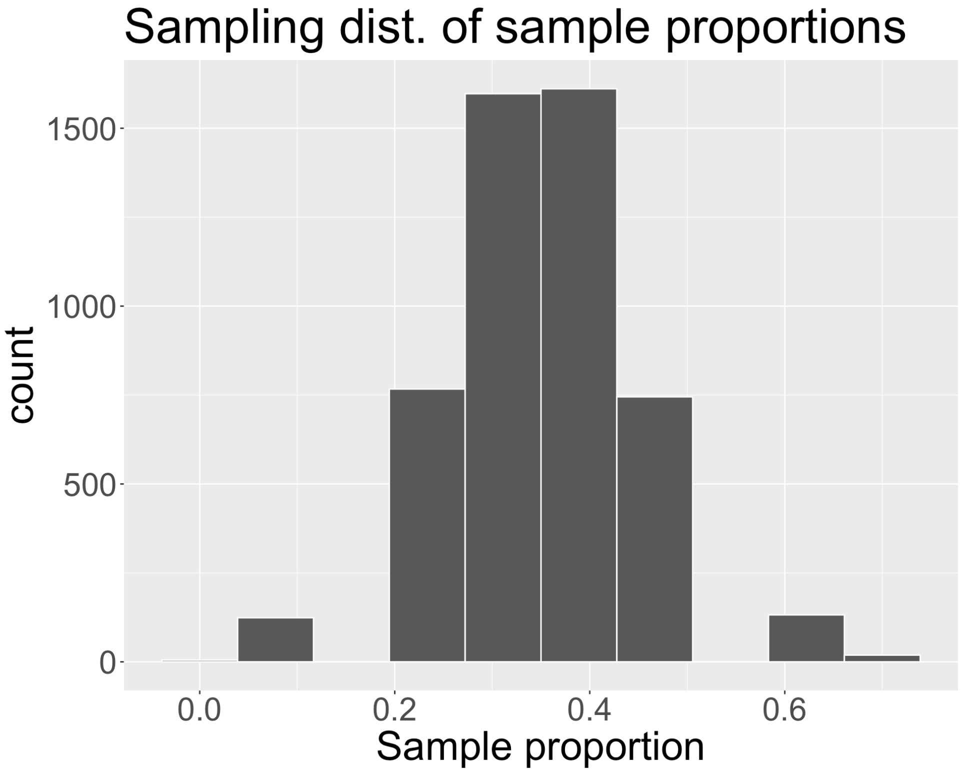
1. Assume we have a sample from the population. Call this sample . Note the sample size is
2. Choose a large number . For in :
    i. Resample: take a sample of size with *replacement* from . Call this set of resampled data
    ii. Calculate: calculate and record the statistic of interest from

At the end of this procedure, we will have a distribution of **resample or bootstrap statistics**

Bootstrap distribution from activity

In our original sample of , we had . We have the following bootstrap distribution of sample proportions, obtained from 5000 iterations:





- Notice that our bootstrap distribution isn’t a great approximation (maybe did not yield a representative sample)

Answering estimation question

- Great…but what do we do with the bootstrap distribution?
- Recall our research question: What proportion of STAT 201A drink coffee regularly?
  - Could respond using our single point estimate:
  - But due to variability, we recognize that the point estimate will rarely (if ever) equal population parameter
- Rather than report a single number, why not report a range of values?
  - This is possible only if we have a sampling distribution to work with!!

Confidence intervals

- Analogy: would you rather go fishing with a single pole or a large net?
  - A range of values gives us a better chance at capturing the true value
- A **confidence interval** provides such a range of plausible values for the parameter (more rigorous definition coming soon)
  - “Interval”: specify a lower bound and an upper bound
  - Confidence intervals are not unique! Depending on the method you use, you might get different intervals

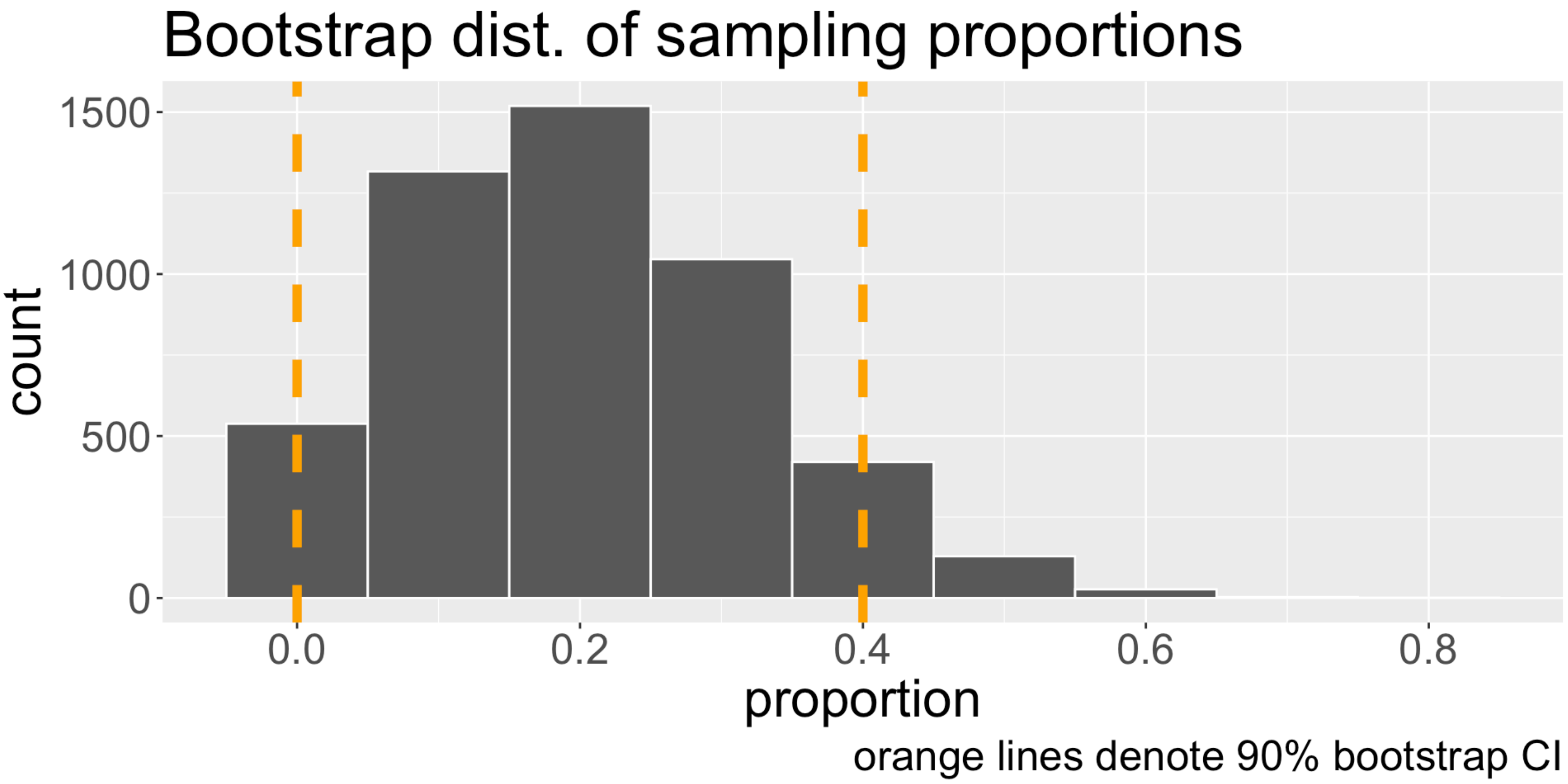
Bootstrap percentile interval

- The % **bootstrap percentile interval** is obtained by finding the bounds of the middle % of the bootstrap distribution
- Called "percentile interval" because the bounds are the and percentiles of the bootstrap distribution

- If , then the bounds would be at which percentiles?

- For our purposes, "bootstrap confidence interval" will be equivalent to "bootstrap percentile interval"
- `quantile()` function in R gives us easy way to obtain percentiles: `quantile(x, p)` gives us -th percentile of `x`

Visualizing bootstrap confidence interval



- Our 90% bootstrap CI for : (0, 0.4)

Interpreting a confidence interval

- Our 90% bootstrap CI for : (0, 0.4). Does this mean there is a 90% chance/probability that the true proportion lies in the interval?
  - Answer: NO
- Remember: bootstrap distribution is based on our original sample
  - If we started with a different original sample , then our estimated 90% confidence interval would also be different
- What a confidence interval (CI) represents: if we take many independent repeated samples from this population using the same method and calculate a % CI for the parameter in the exact same way, then in theory, % of these intervals should capture/contain the parameter
  - represents the long-run proportion of CIs that theoretically contain the true parameter
  - However, we never know if any particular interval(s) actually do!

Interpreting a confidence interval (cont.)

- Correct interpretation (generic) of our interval : We are % confident that the population parameter is between and .
  - Interpret our bootstrap CI in context
- Again: why is this interpretation **incorrect**? “There is a 90% chance/probability that the true parameter value lies in the interval.”

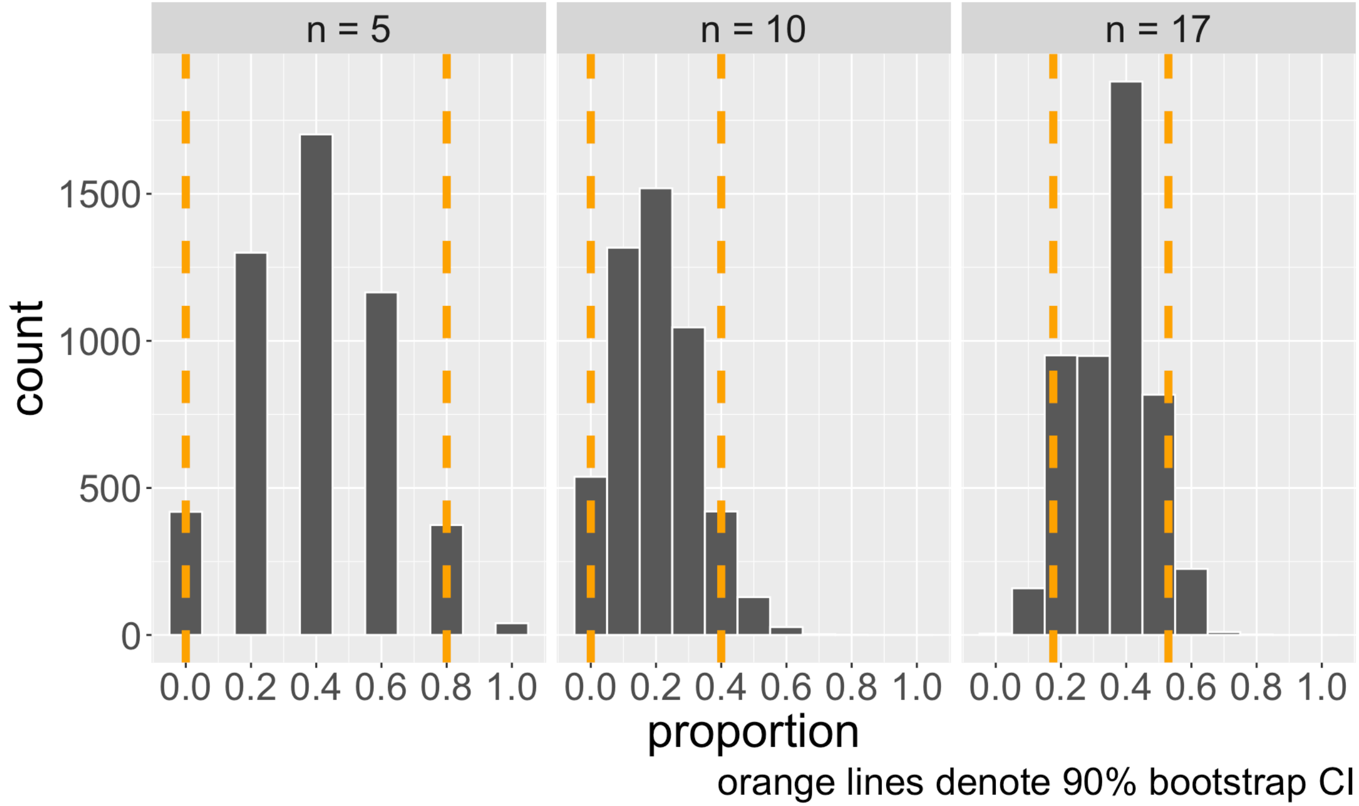
Remarks

- What is a virtue of a “good” confidence interval?
- How do you expect the interval to change as the original sample size changes?

  How do you expect the interval to change as level of confidence changes?
- Once again, a good interval relies on a representative original sample!

Comparing confidence intervals

Comparing changes in 90% bootstrap CI for sample sizes .



n	interval
n = 5	(0, 0.8)
n = 10	(0, 0.4)
n = 17	(0.18, 0.53)

What do you notice about the bootstrap distributions and CIs as increases?

Live code + Coding practice!

- Live code:
  - in-line code
  - setting a seed
- You will investigate what happens as we move between to !