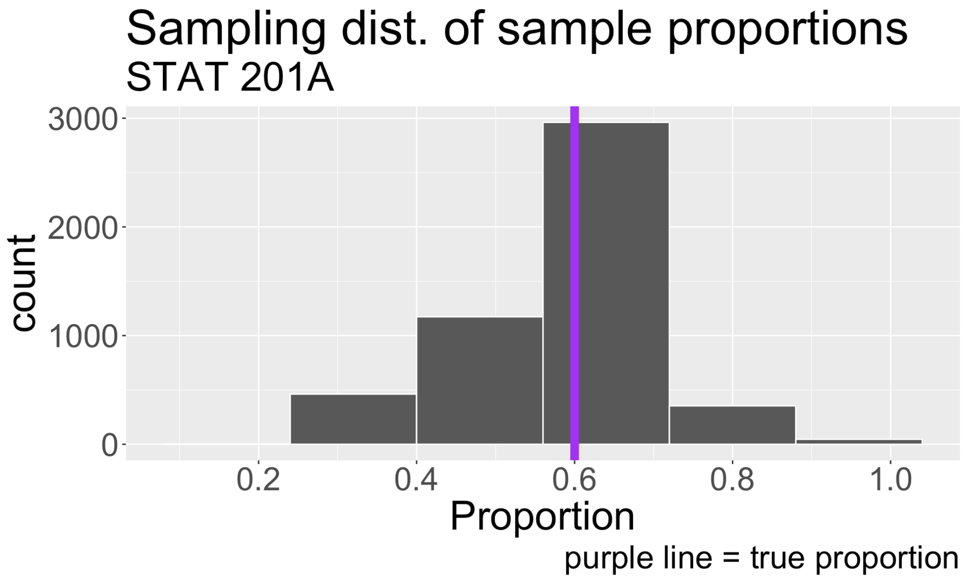
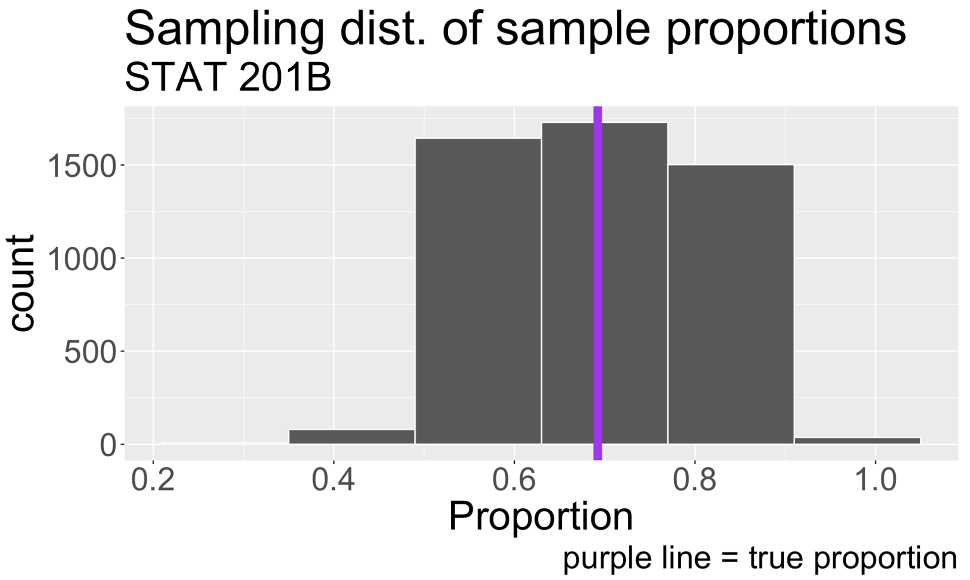
Bootstrap Confidence Intervals

10/9/24

Recap

- Sampling distribution describes how statistic behaves under repeated sampling from population
- Let's return to the data collected from our activity.
  - I will repeatedly take SRS of $n = 10$ values from the population (call this $\vec{x}$) and calculate $\hat{p}$. Sampling distribution of $\hat{p}$ is as follows:





Bootstrap recap

Taking new samples each time is costly! Bootstrap distribution is an *approximation* of the sampling distribution of the statistic!

Procedure:

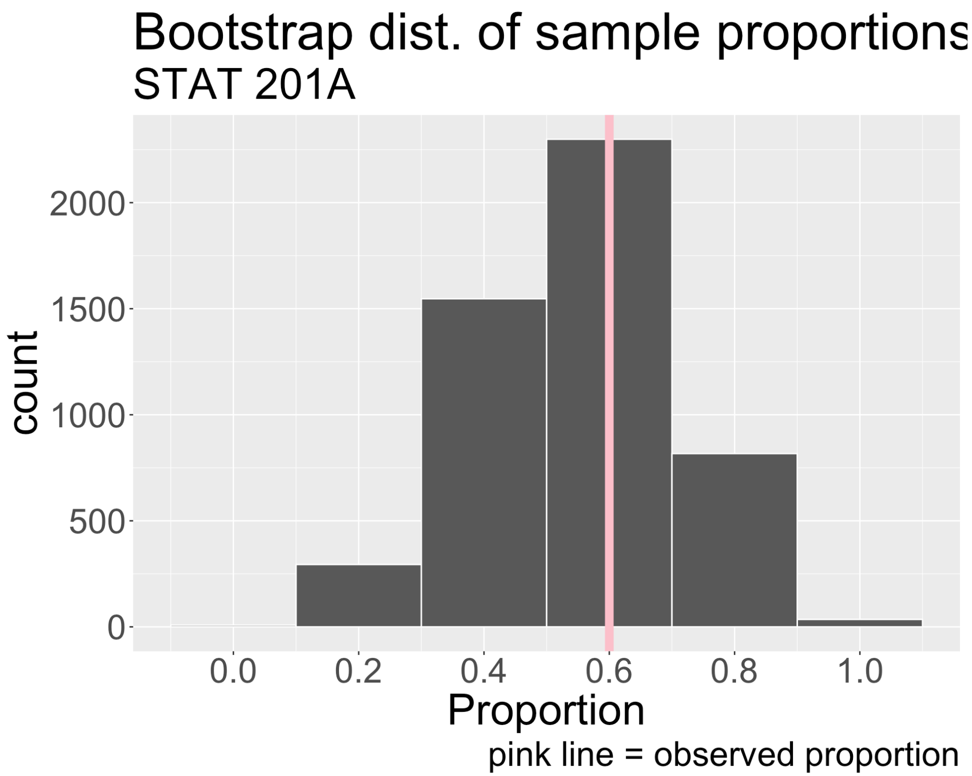
1. Assume we have a sample $x_1, x_2, \ldots, x_n$ from the population. Call this sample $\vec{x}$. Note the sample size is n
2. Choose a large number B. For b in $1, 2, \ldots, B$:
   i. Resample: take a sample of size n with *replacement* from $\vec{x}$. Call this set of resampled data $\vec{x}_b^*$
   ii. Calculate: calculate and record the statistic of interest from $\vec{x}_b^*$

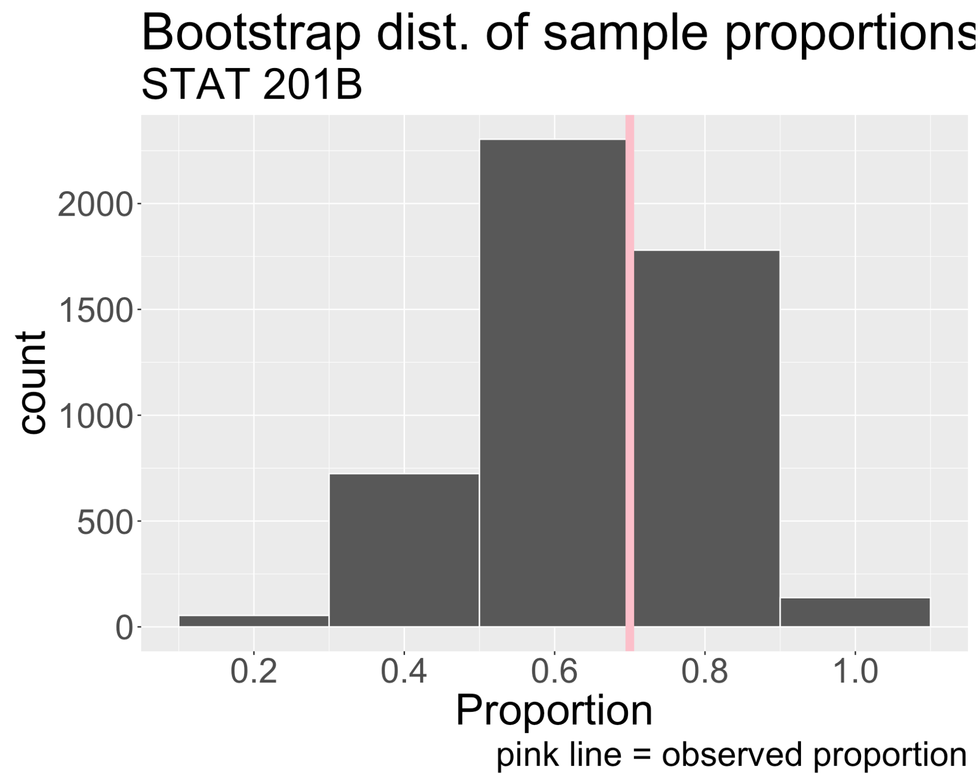
At the end of this procedure, we will have a distribution of **resample or bootstrap statistics**

Bootstrap distribution from activity

We have the following bootstrap distribution of sample proportions, obtained from $B = 5000$ iterations:

Bootstrap dist. continued





- Notice where the bootstrap distribution is centered
- What do we do with the bootstrap distribution?

Answering estimation question

Recall our research question: What proportion of STAT 201A/STAT 201B students get at least 7 hours of sleep a night?

- Could respond using our single point estimate: $\hat{p}_A = 0.6$ or $\hat{p}_B = 0.7$
- 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 distribution to work with!!

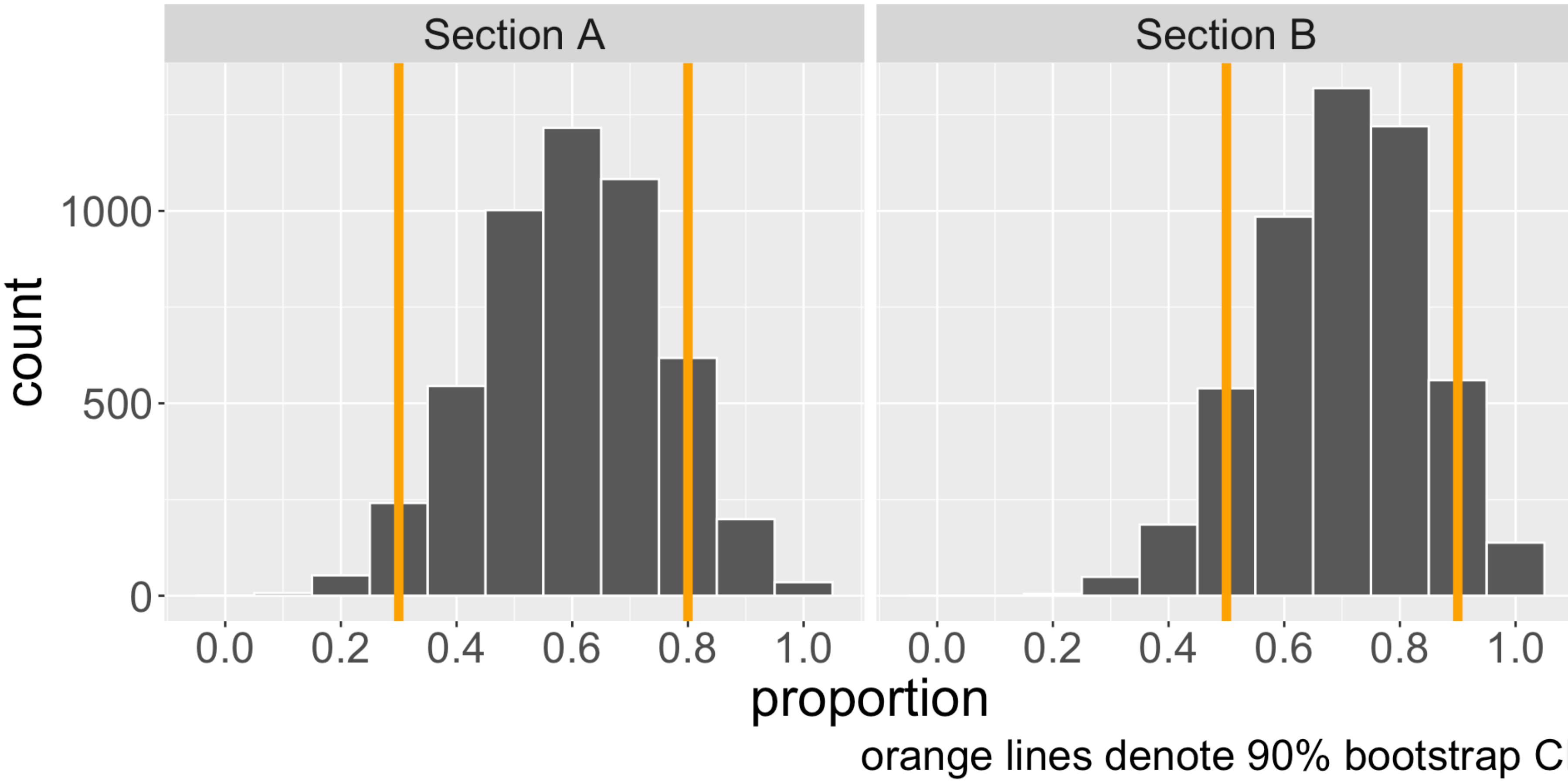
Confidence intervals

- Analogy: would you rather go fishing with a single pole or a large net?
  - A range of plausible values gives us a better chance at capturing the parameter
- A confidence interval provides such a range of values (more rigorous definition coming soon)
  - “Interval” = we 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 $\gamma \times 100\%$ **bootstrap percentile interval** is obtained by finding the bounds of the middle $\gamma \times 100\%$ of the bootstrap distribution
  - e.g. If I want a 90% bootstrap percentile interval, where would the bounds be?
- Called "percentile interval" because the bounds are the $(1-\gamma)/2$ and $(1+\gamma)/2$ percentiles of the bootstrap distribution
  - e.g. if $\gamma = 0.80$, then the bounds would be $(1-0.80)/2 = 0.10$ and $(1+0.80)/2 = 0.90$ percentiles
- For our purposes, "bootstrap confidence interval" will be equivalent to "bootstrap percentile interval"

Obtaining bootstrap confidence interval



- Section A 90% confidence interval for p_A: (0.3, 0.8)
- Section B 90% confidence interval for p_B: (0.5, 0.9)

Interpreting a confidence interval

- Our 90% confidence interval is: (0.3, 0.8) or (0.5, 0.9). 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 $\vec{x}$, 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 $\gamma \times 100$ % CI for the parameter in the exact same way, then in theory, $\gamma \times 100$ % of these intervals should capture/contain the parameter
  - γ 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): We are $\gamma \times 100$ % confident that the population parameter is between the lower bound and upper bound.
  - 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.”

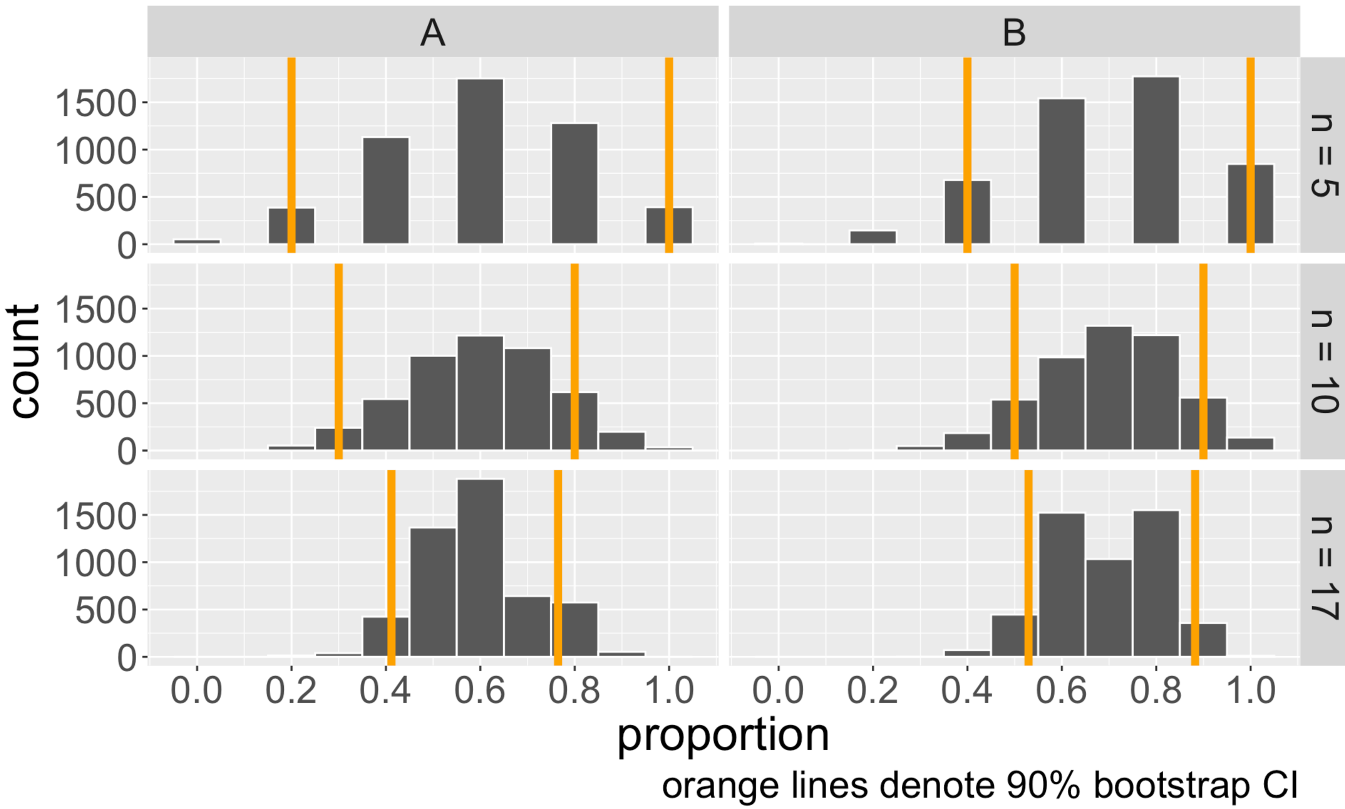
Remarks

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

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

Comparing confidence intervals

Comparing changes in $\gamma \times 100$ % CI for sample sizes: $n = 5, n = 10$, and $n = 17$:



Section A

n	interval
n = 5	(0.2, 1)
n = 10	(0.3, 0.8)
n = 17	(0.41, 0.76)

Section B

n	interval
n = 5	(0.4, 1)
n = 10	(0.5, 0.9)
n = 17	(0.53, 0.88)

What do you notice?

Live code + your turn!

- Live code:
    - in-line code
- You will investigate what happens as we move γ between 0 to 1!