DATA SCIENCE FOR EVERYONE

Prof. Spirling (Prof Jones-Rooy)

Sept 30, 2019

4: Statistics 2, part 1

ANNOUNCEMENTS

1. HW1 back out soon. Lenient on lateness if we got it by Friday.
2. HW2 out tonight. One week till deadline. Academic honesty policy applies. Format correctly. TAs will reject non-pdfs and anything they cannot read (late penalty). Take advantage of example code!
3. Bring laptops to labs because TAs will teach you how to upload data and code.
4. "Engagement" (again).

FROM LAST WEEK
PERMUTATION INFERENCE

- This is called **"permutation inference"**: we will permute (shuffle) the labels and each time record the difference between smoking and non-smoking weights.

- We will do this many, many times. The ***distribution*** of all those differences will be the distribution of our statistic under the null hypothesis.

in *proportion* to relative frequency of *actual* labels

- We can then ***compare*** the observed statistic with that distribution to see whether the null could have generated what we saw.

OUTLINE FOR TODAY

1. Estimates vary between samples: want to be systematic about this uncertainty
2. Percentiles
3. Bootstrap (re) sampling for confidence intervals

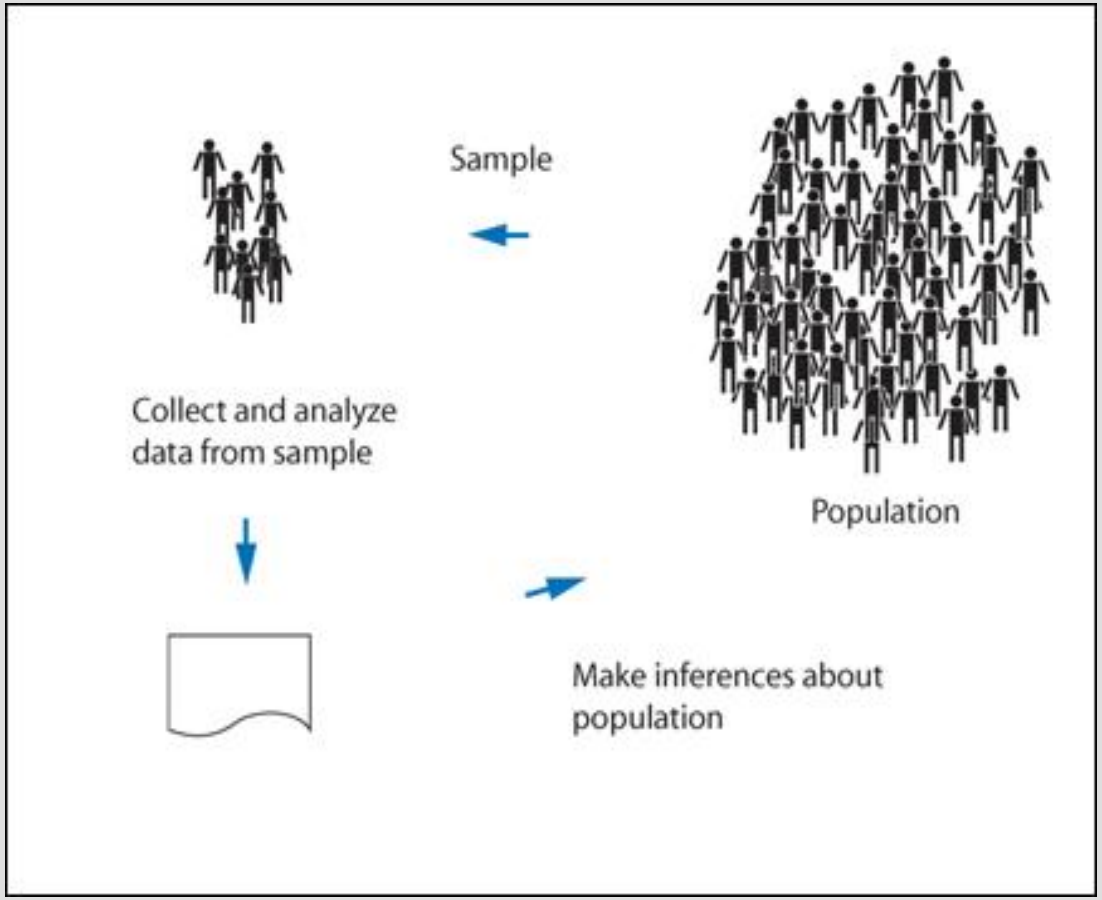
POPULATION

- The population is ***the universe of cases we want to describe***. We could ask:
- How will the United States (every voter) vote in the 2020 Presidential election?
- What is the average income in Britain (every household or every person)?

- We call the characteristic we care about (vote intention, income) the ***population parameter***.
- Unfortunately, outside of a ***census***, we cannot study the population directly

SAMPLES

- So, instead we will use ***samples*** from the population, and use those to ***estimate*** the parameters of the population.
- We'll be able to talk about our ***uncertainty*** around those estimates.
- Focus on the properties of ***large***, ***random*** samples.



RANDOM SAMPLES

- A ***statistic*** from a random sample from the population can give us a reasonable "point" estimate.
- But we want to know how much "***error***" we have around that estimate. How certain or ***uncertain*** are we about our estimate?

- We have to systematically answer the following question:

"how different ***could*** this estimate have been, if we had drawn a ***different*** random sample?"

PERCENTILES

the p^{th} percentile is the point such that p% of the observations fall below, (100-p)% fall above.

- the 50^{th} percentile is the median
- 25^{th} percentile is the first (lower) quartile
- 75^{th} percentile is the third (upper) quartile
- NB: definitions differ slightly across different software

HOW THIS IS CALCULATED

- take the sample (n=8 here) 1,3,9,7,5,3,11,3
- rank the observations 1,3,3,3,5,7,9,11

- Find p% of n: (p/100)×n. Call that ***k***.
- If k is an ***integer***, take the kth element of the sorted collection.
- If k is ***not an integer***, round it up to the next integer, and take that element of the sorted collection.

- 25^{th} percentile is the 2^{nd} entry (0.25x8 = 2) : 3
- 75^{th} percentile is the 6^{th} entry (0.75x8=6): 7

INTERQUARTILE RANGE

- take the sample (n=8 here) 1,3,9,7,5,3,11,3
- rank the observations 1,3,3,3,5,7,9,11

- The ***interquartile range*** (IQR) is the **difference between the first and third quartiles** in the sample.

- Here, this IQR = 7 – 3 = 4.

PARTNER EXERCISE

- Suppose all voters care about is where candidates from two parites are on a left-right spectrum, and they vote for the candidate nearest to them. In this world, the ***median voter*** in the electorate is very important, and candidates need to win them over. Why?

- The child percentile growth charts in US Doctors's offices are based on large amount of data from the ***1960s and 1970s***. Some people want to update them, some don't. Give an argument for both sides.

HOW ESTIMATES VARY

- We know how to generate the median (or any statistics) from a ***random sample***.
- But that was just ***one*** random sample from the population: we want to know how much the statistic would ***vary*** if we took different random samples.
- We could work this out ***analytically***, but it's difficult, esp for certain statistics.
- So, instead, we'll use the "***bootstrap***."

BOOTSTRAP

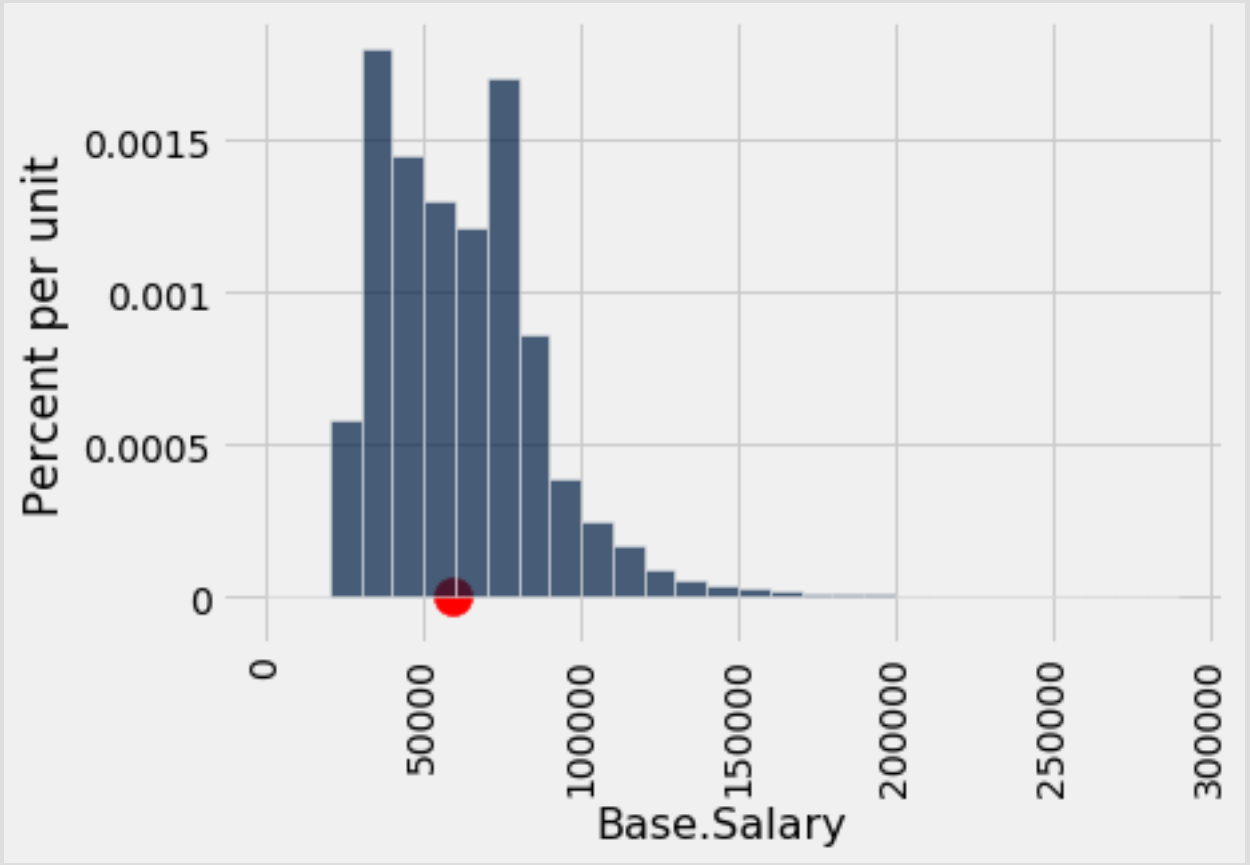
- We could work out the variability of our statistic by taking lots and lots of samples, calculating the statistic each time, and plotting it. Problem is we typically only get ***one sample***.

- The "***bootstrap***" is a procedure for ***simulating*** this process. It will give us the ***sampling distribution***, without much math.

- The bootstrap says "treat the sample ***as if*** it's the population. Draw large numbers of random samples from the sample, and compute the statistic each time. Then study the ***distribution*** of those statistics"

NYC SALARY DATA

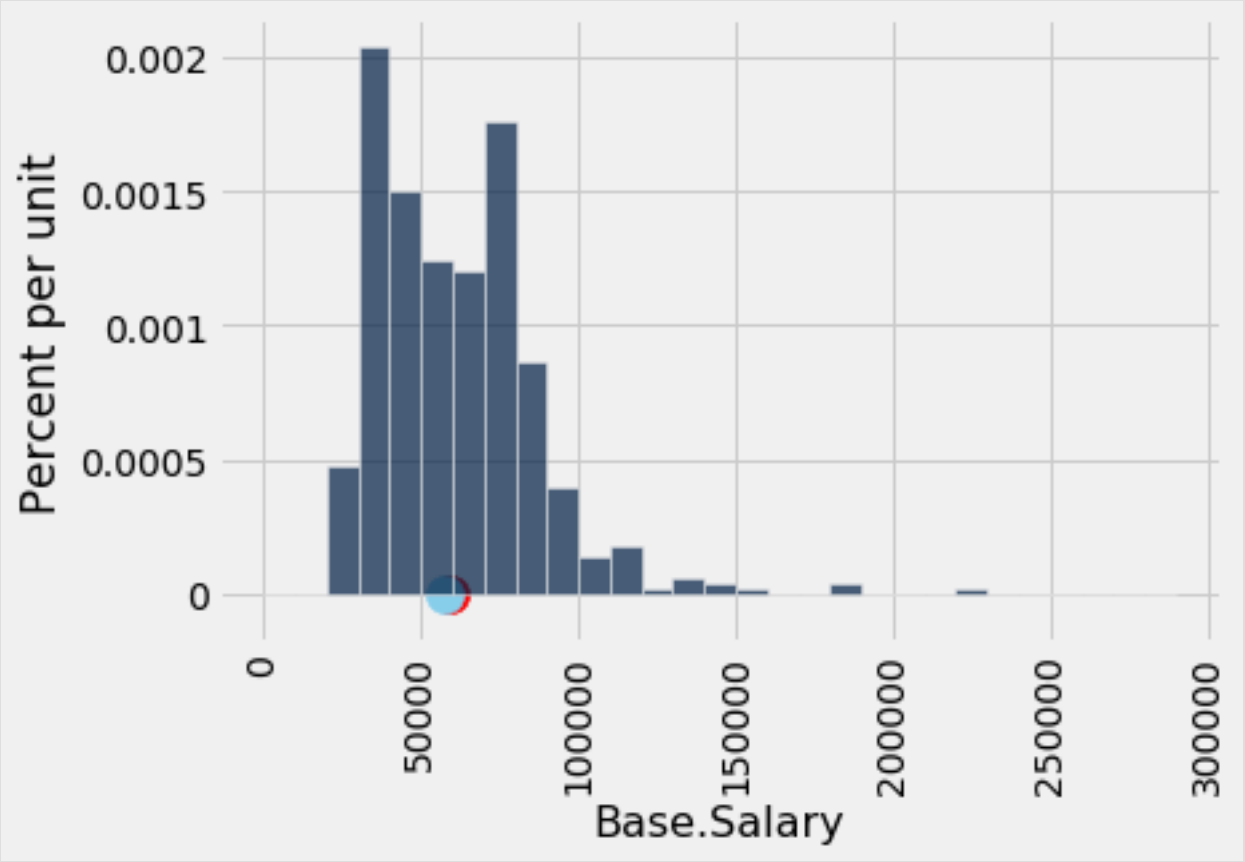
- About 290k people work for NYC as salaried employees.
- Let's call that the ***population***. The median is ***$58,850*** (the max is $300000)
- These are base salaries, so some people get more than this in total.



OUR SAMPLE

- Suppose we didn't have the population, but we had a large random sample of ***500*** rows of the data.

- In our sample, the median was ***$57842***. Close to our population median, but not the same.



BOOTSTRAPPING THE MEDIAN

- Treat the sample ***as if*** it were the population.
- Draw a ***sample of size n each time*** (here, 500) with ***replacement***

- We draw a sample of size n because we want to ***compare our sample*** with other samples of that size ***we could have*** taken.

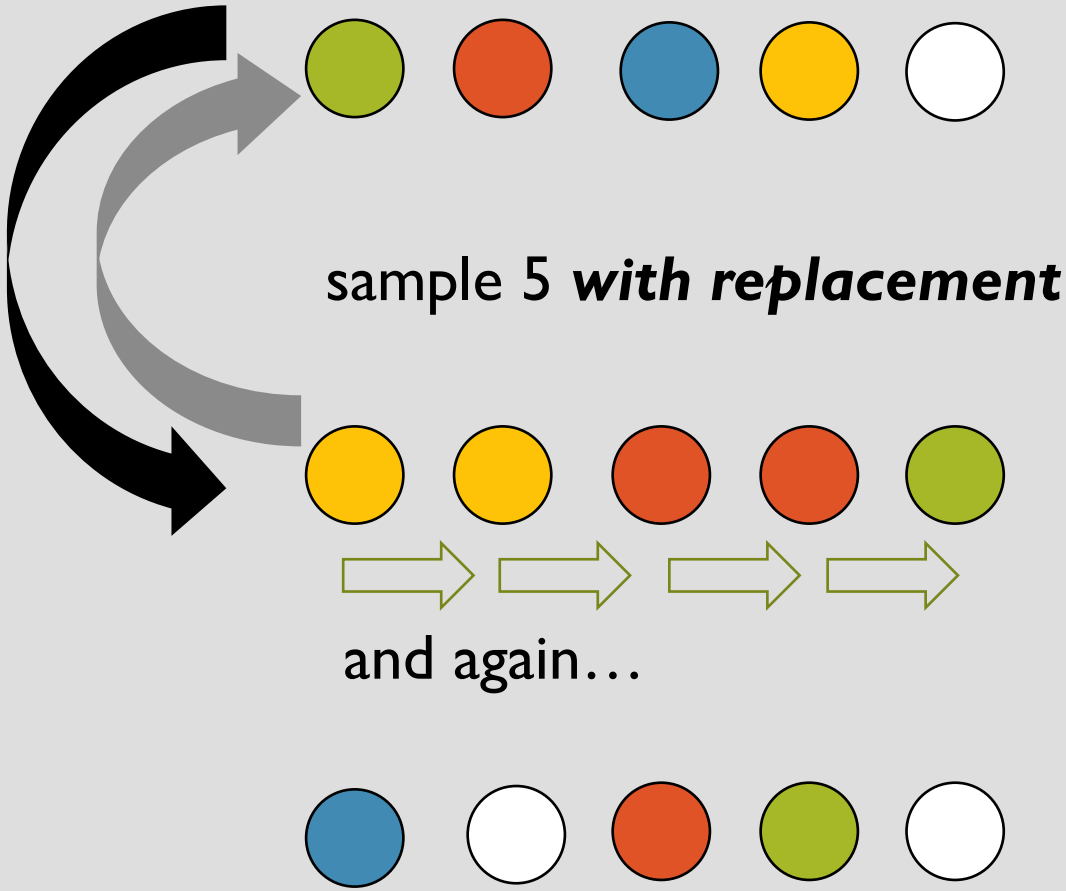
- Sampling with ***replacement*** means we can draw the same row of our data multiple times. This means our new samples can ***differ*** from our actual sample.

SAMPLING

with replacement

- once we sample a unit, it goes *back into* the "population" and can be sampled again (though not necessarily)

- here, sampling 5 ***with replacement*** means getting a "new" sample, different to our original one

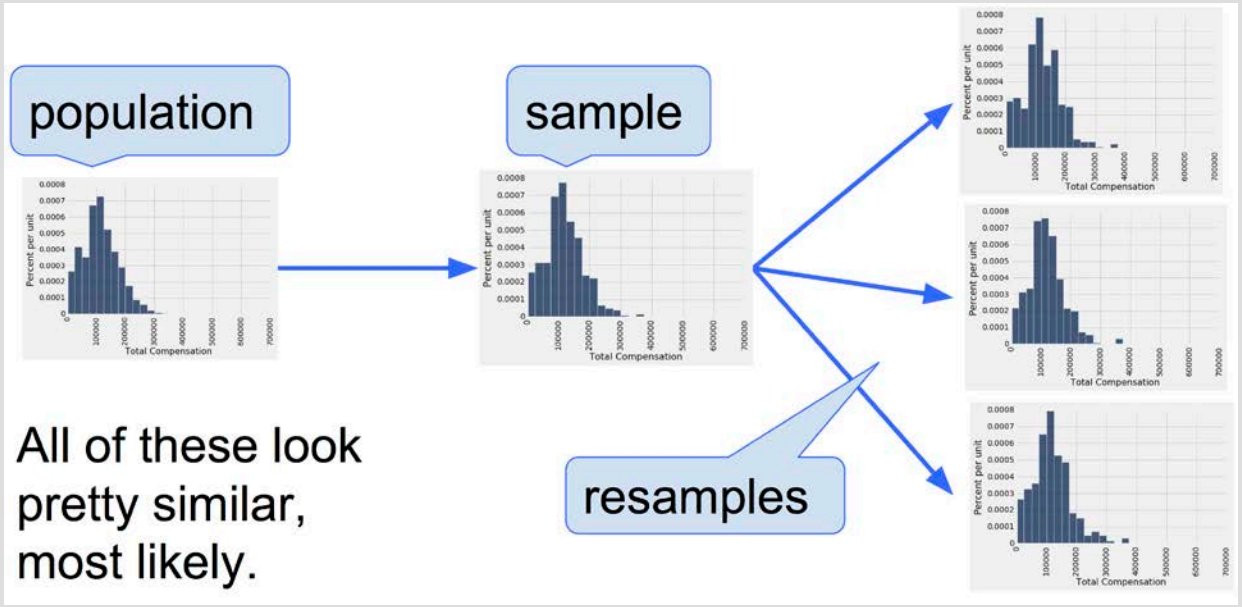


WHY BOOTSTRAPPING WORKS

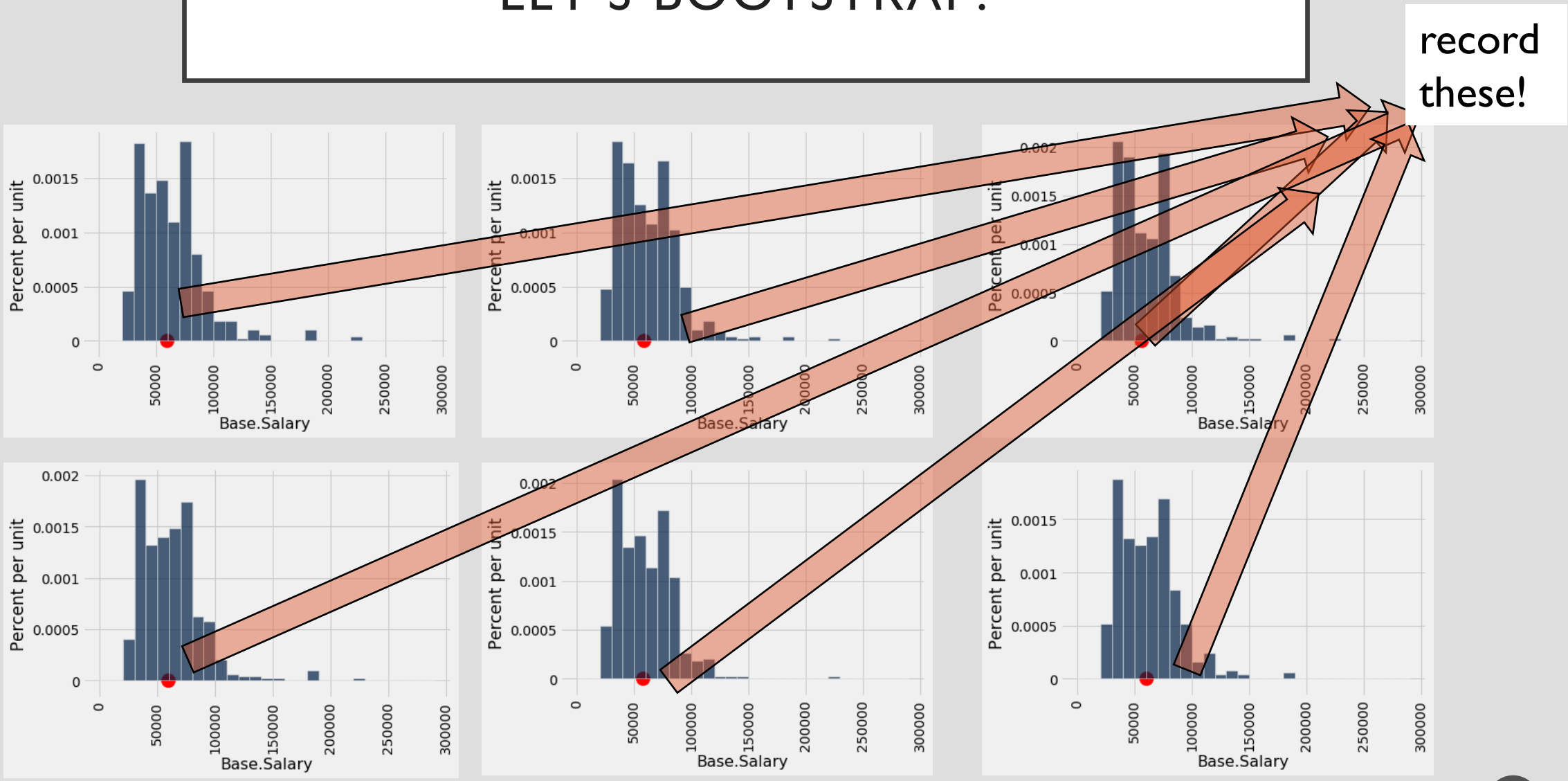
- ***Law of Large Numbers*** (LLN) says our original sample will look like the population

- And we know the ***resamples*** from the sample will look like our original sample.

- If we put all the resamples ***together***, we'll have something that resembles the ***population***.

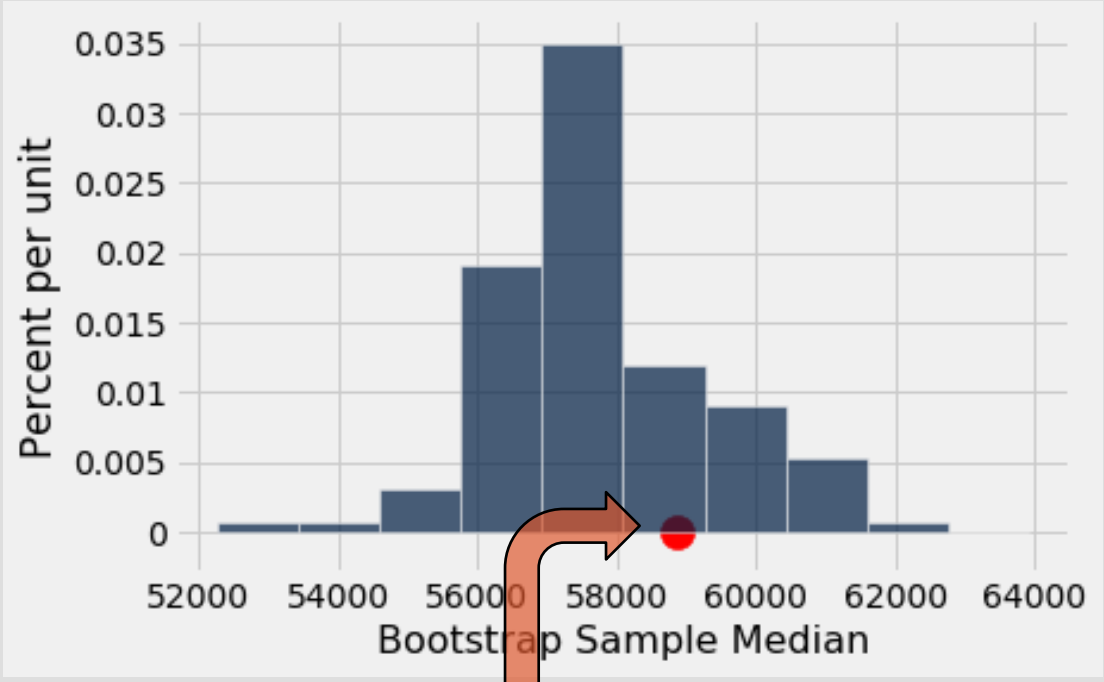


LET'S BOOTSTRAP!



BOOTSTRAP DISTRIBUTION OF THE SAMPLE MEDIANS

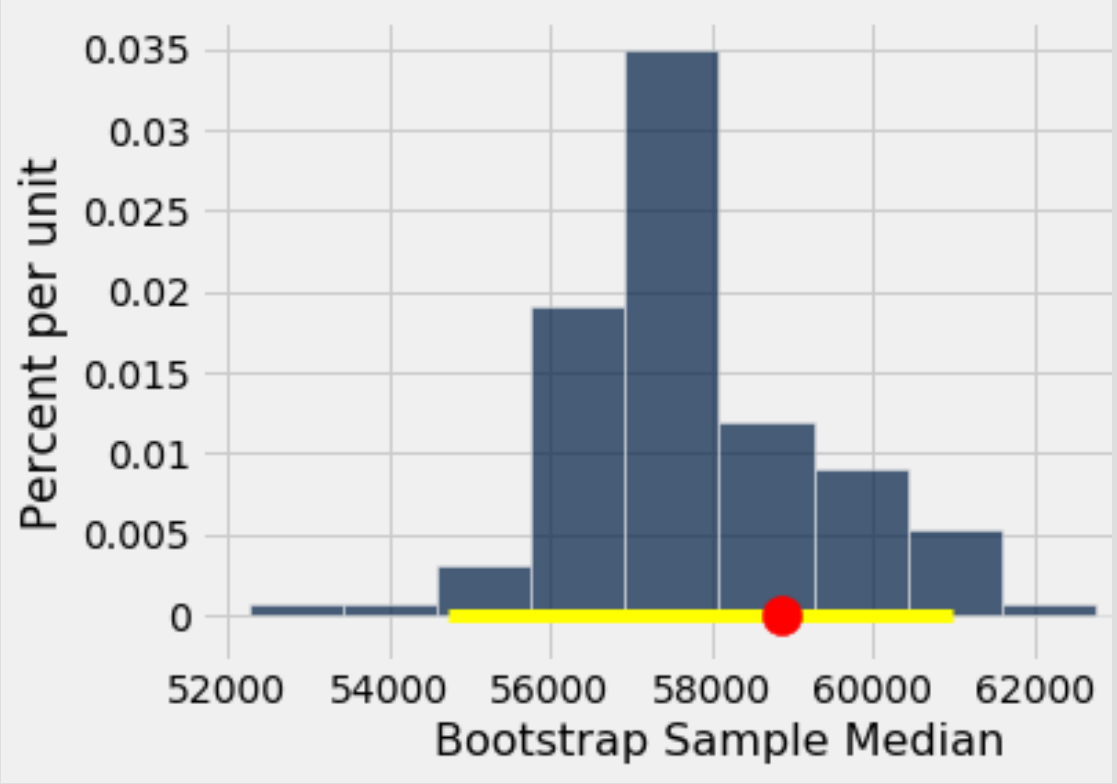
- Each time (5000) we generate a new (re) sample, we ***record the (new) median***.
- Then, we plot all these medians.
- This is the ***bootstrap empirical distribution of the sample medians***.



"true" ***population*** median

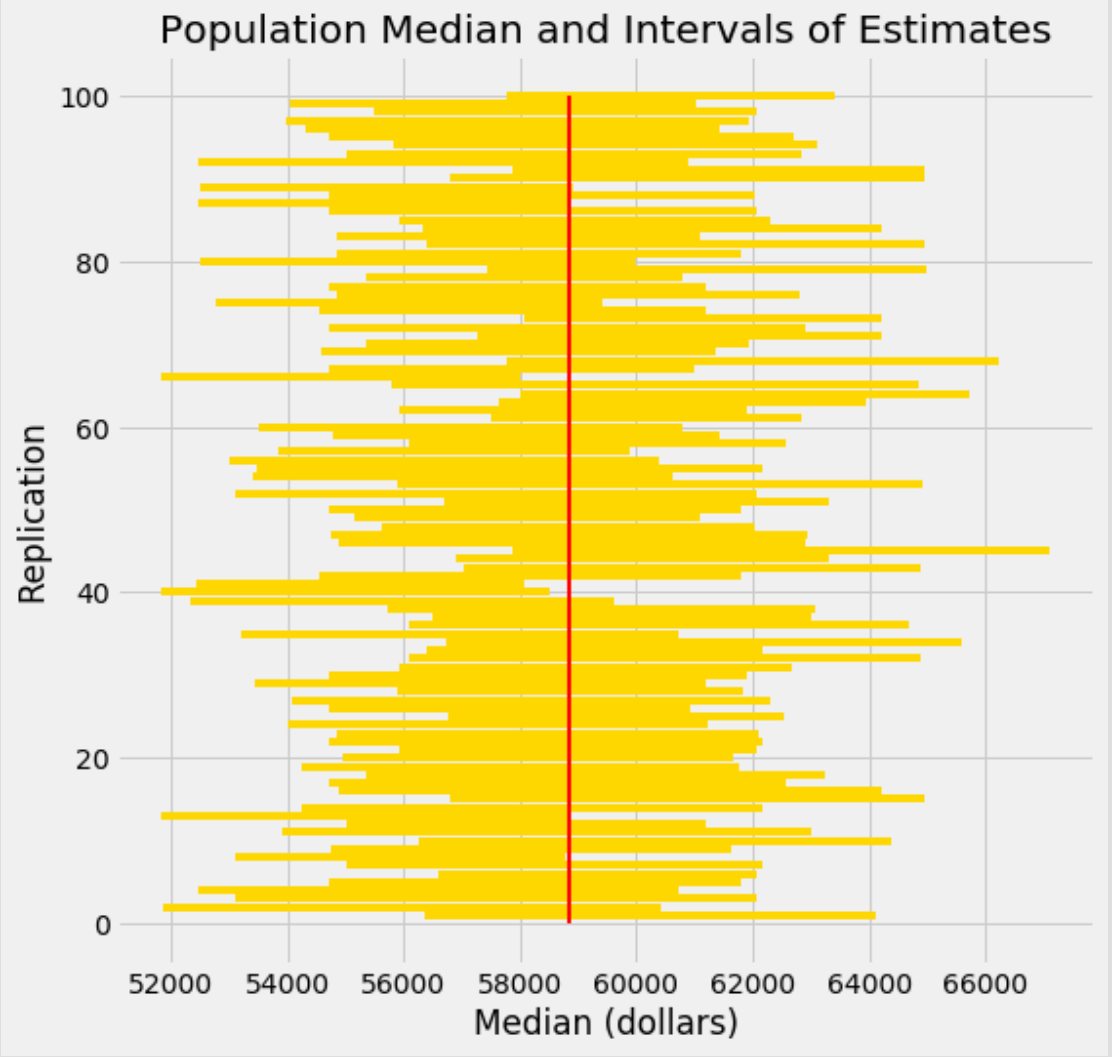
DID WE CAPTURE THE TRUTH?

- The true median wasn't quite in the middle of our resamples. But do we "generally" ***capture*** it?
- does the ***middle 95%*** of the resample distribution capture the ***truth***?
- Look at the range of our sampling distribution from ***2.5 percentile to 97.5 percentile***…

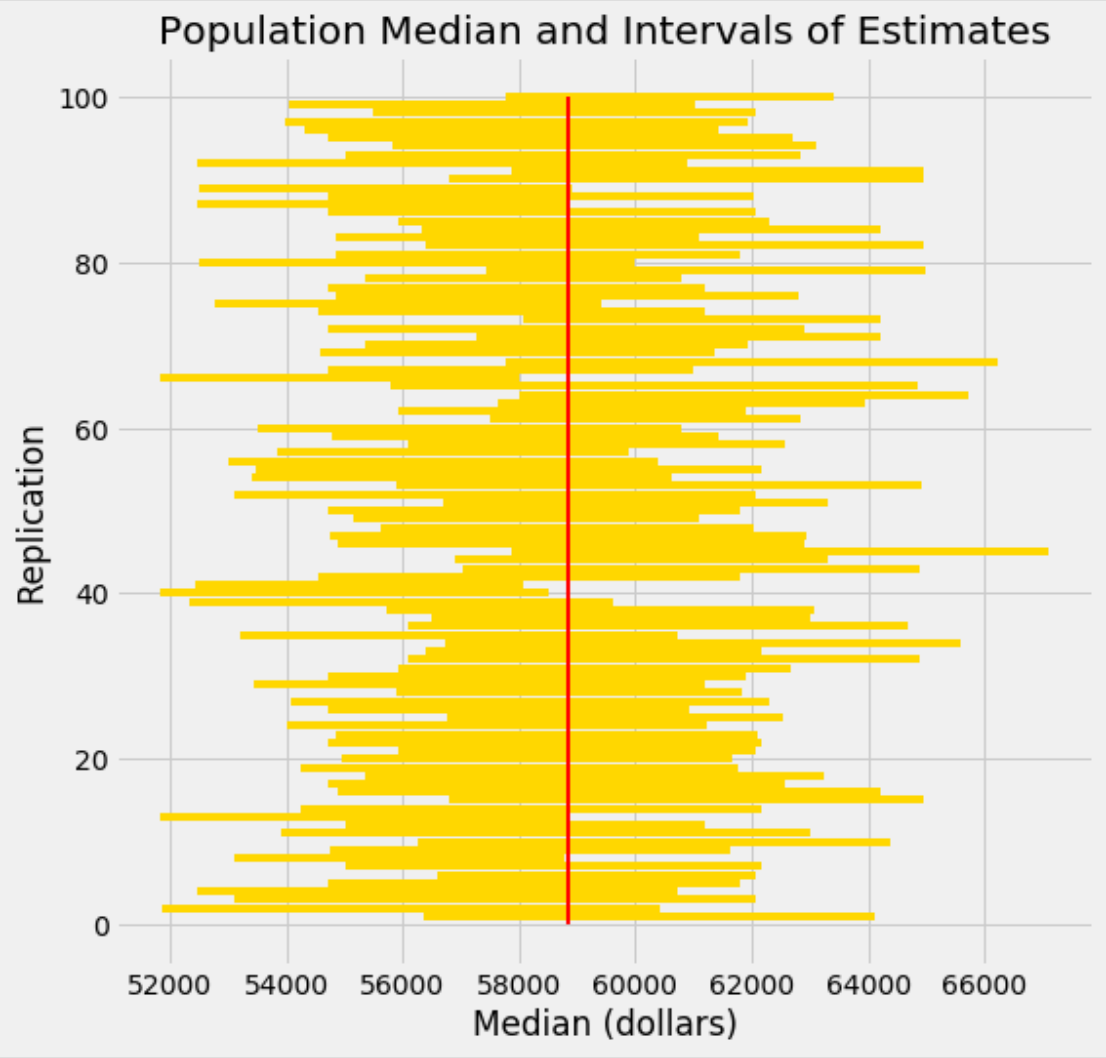


HOW OFTEN WILL THIS WORK?

- The 95% interval captured the truth in ***this*** case. Great!
- But how ***general*** is that result?
- We can repeat this whole set up many times (say, 100 times) and see ***how often*** that ***95% interval*** would capture the truth.



HOW OFTEN WILL THIS WORK?



- This works about ***95% of the time***. 95% of the (95%) intervals captured the truth, 5% didn't.

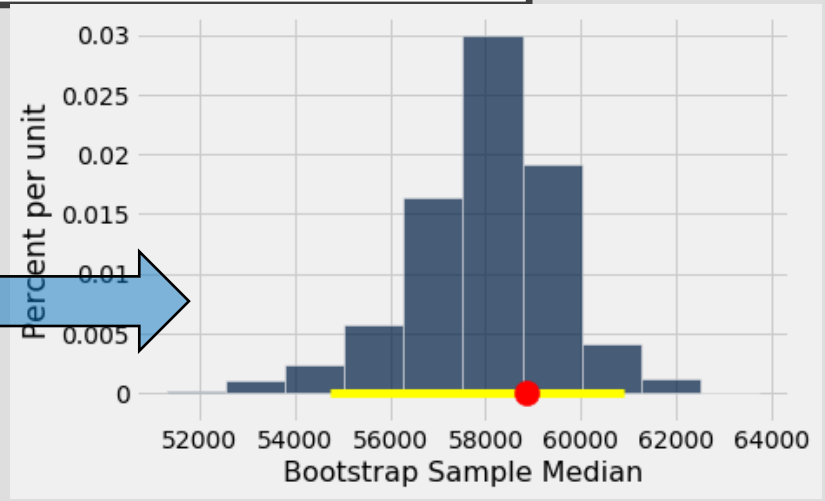
- We could make the interval ***smaller***, but we'd have to accept that it captures the truth less often!

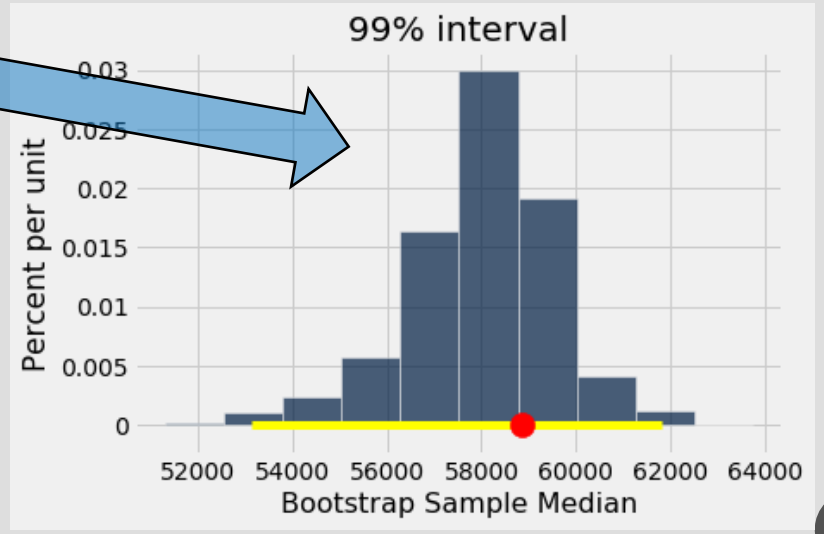
CONFIDENCE LEVEL

- The bootstrapping procedure gives us an ***interval*** of estimates.
- This interval takes into account the ***variability*** from random sampling.
- For the 95% interval, we are "***95% confident***" that it will capture the truth.

CONFIDENCE INTERVAL

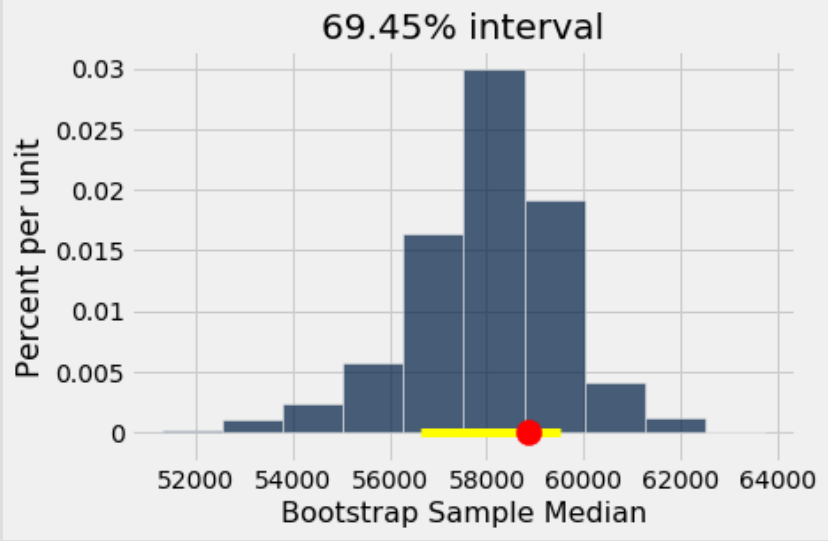
- The interval that captures the truth 95% of the time is the ***95% confidence interval***.
- The interval that captures the truth 99% of the time is the ***99% confidence interval***. (wider!)
- If I want **more** confidence, I need a **wider** interval (why?)



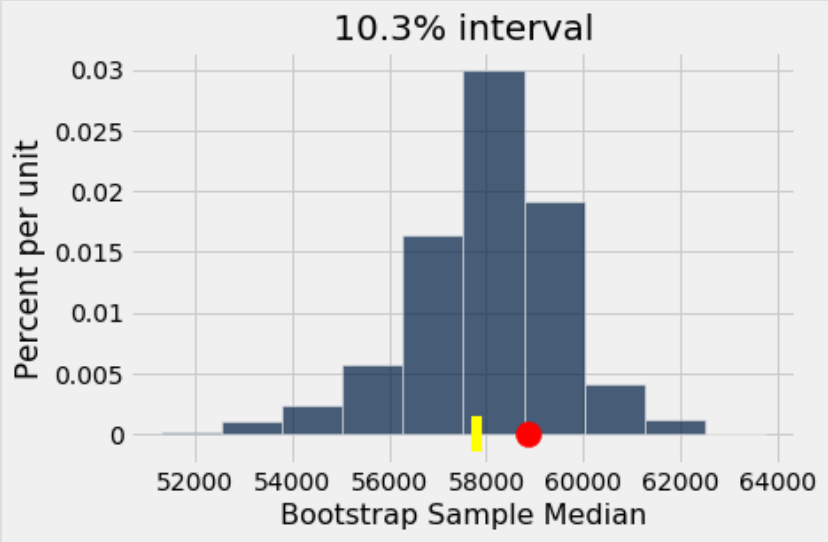


CONFIDENCE INTERVAL

- The interval that captures the truth 69.45% of the time is the ***69.45% confidence interval***



- The interval that captures the truth 10.3% of the time is the ***10.3% confidence interval***



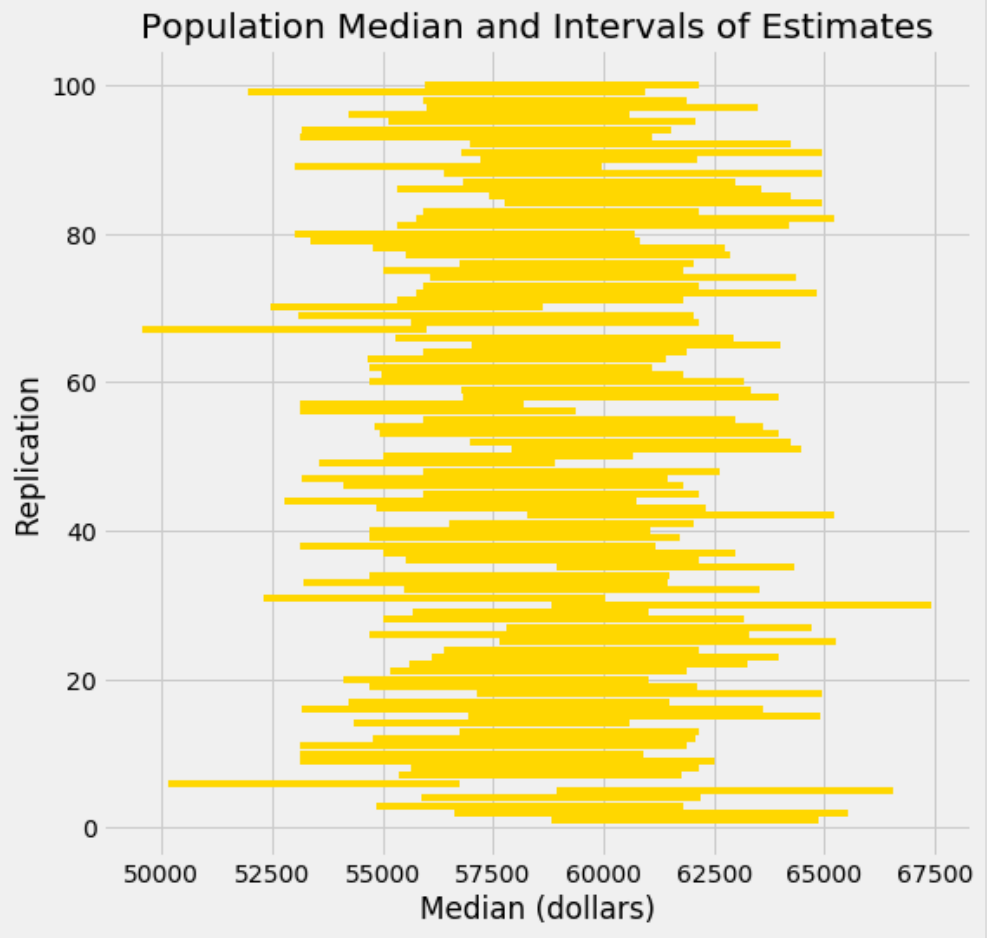
PARTNER EXERCISE

Are the following statements about confidence intervals ***true or false***?

- A given 95% confidence interval tells us where 95% of the variable's values lie.
- A given 95% confidence interval captures the true value of the parameter with probability 0.95.
- A given 95% confidence interval captures the true value of the parameter or it does not.
- A given 95% confidence depends on the random sample it was created from.

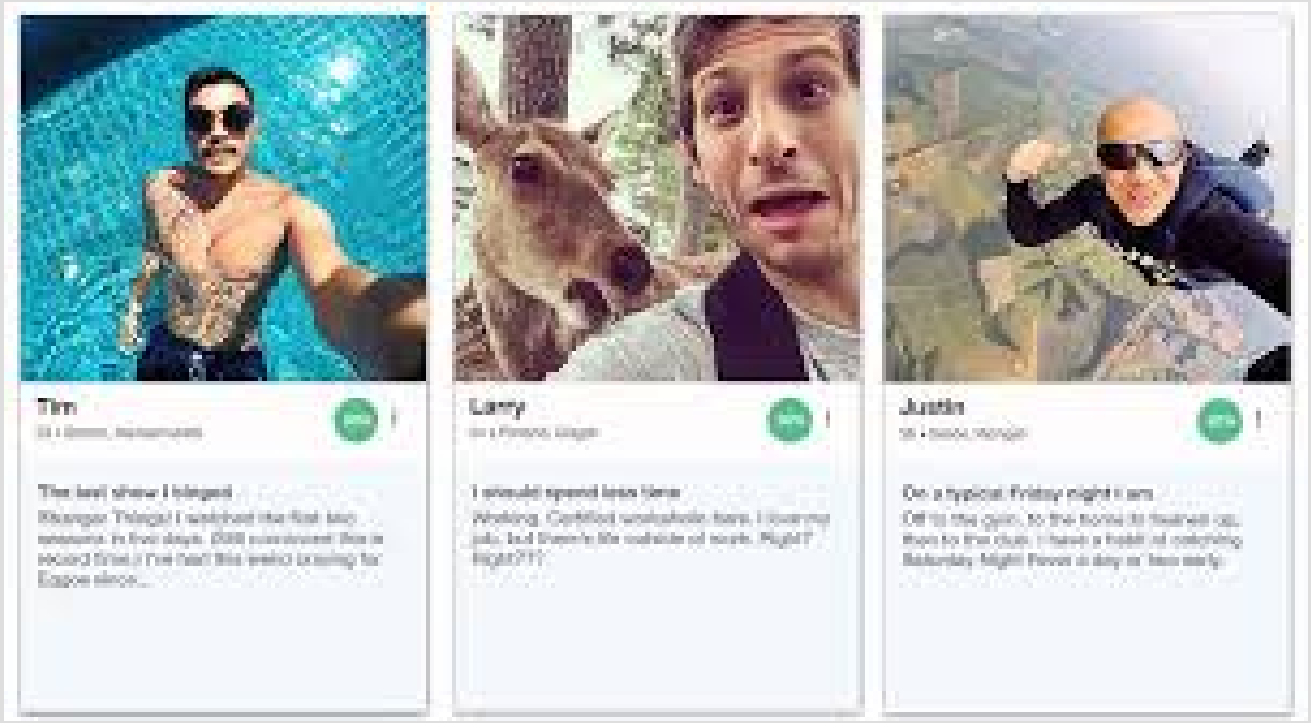
CONFIDENCE LEVEL

- The bootstrapping procedure gives us an ***interval*** of estimates.
- This interval takes into account the ***variability*** from random sampling.
- For the 95% interval, we are "***95% confident***" that it will capture the truth.
- But actually we ***won't*** usually "know" the truth!



OKCUPID

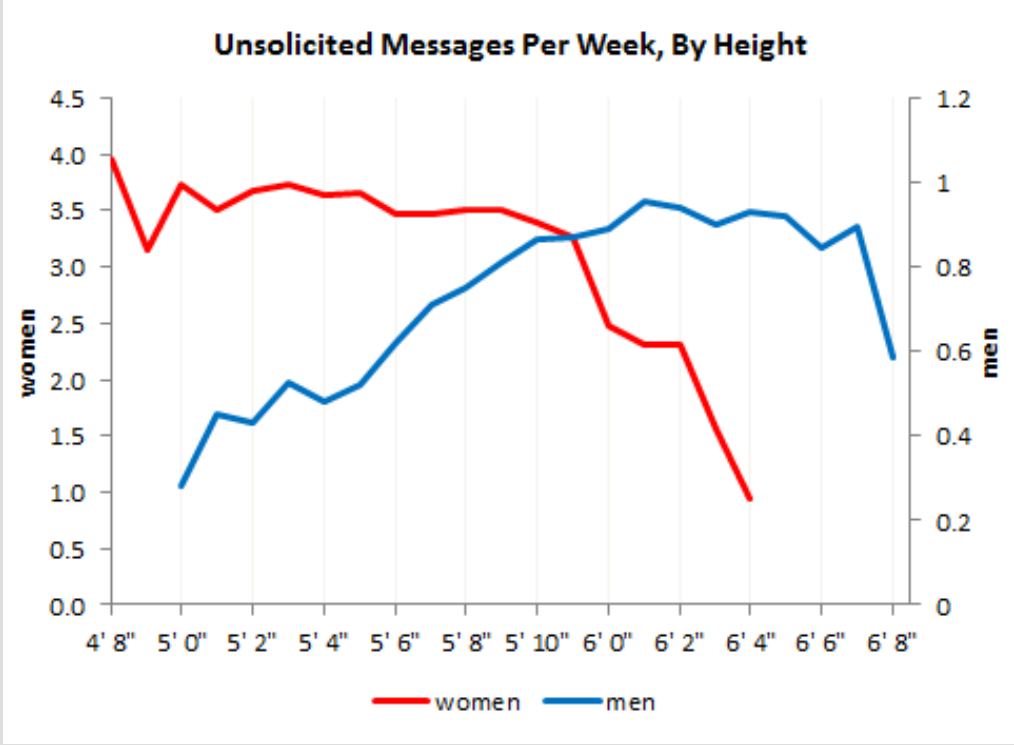
- OkCupid is an online ***dating*** website.
- People post profiles, and others match with them.
- We have an anonymous random sample of ***1112*** ones from men.



CONFIDENCE INTERVAL FOR POPULATION PARAMETER

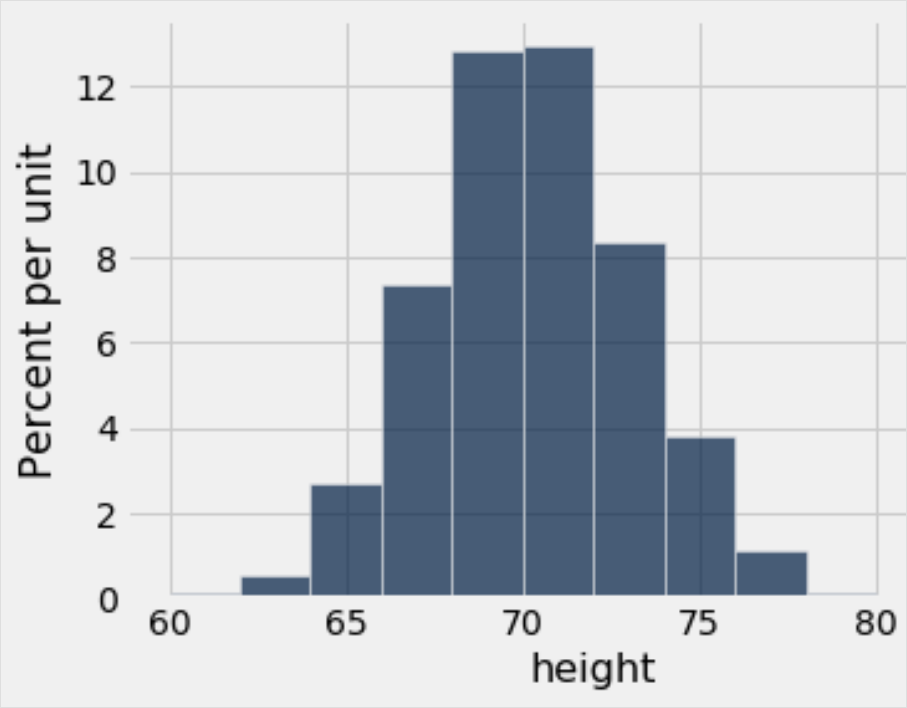
- Evidence that ***height*** is considered important for men seeking partners!

- Q: what's the ***average*** height of men on OkCupid?
- Note: we ***don't have*** the population value this time, and want to make an inference about it.



SAMPLE DISTRIBUTION

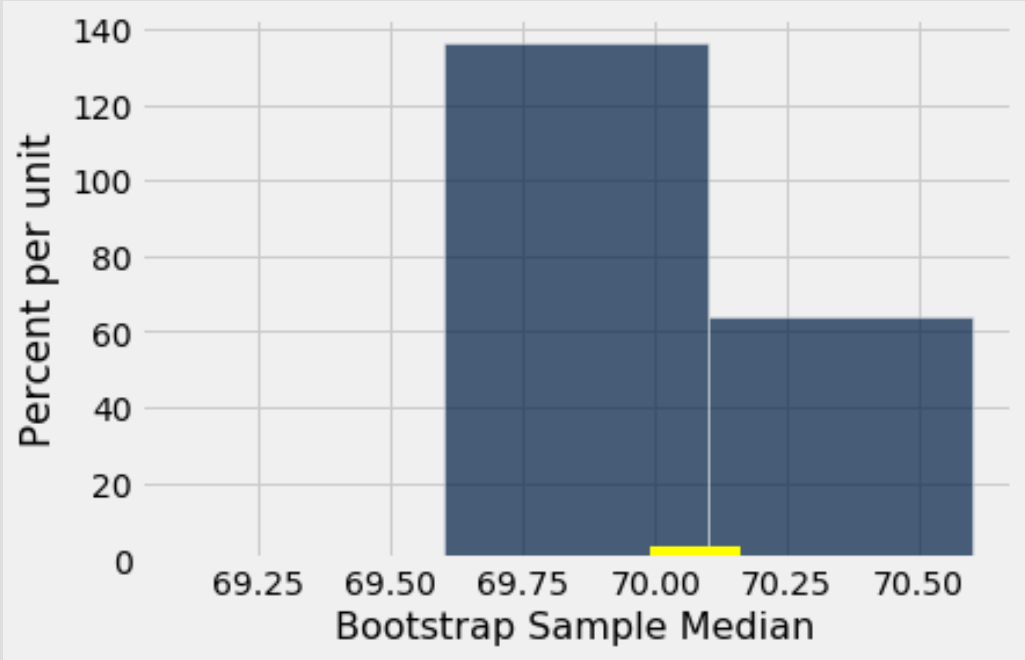
- In our sample, the ***median*** height was ***70.13 inches***, or ~5'10".
- *btw, that's higher than the US male population as whole – why might that be?*



- We can graph the sample distribution, but we ***don't have*** the OkCupid population to compare to.

SAMPLING DISTRIBUTION AND CONFIDENCE INTERVAL

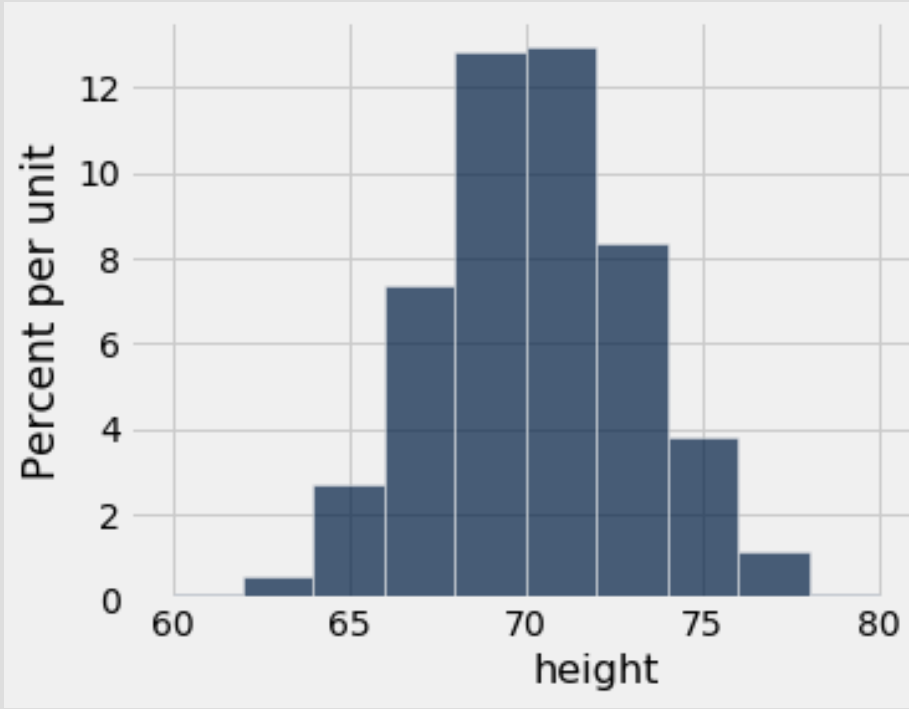
- We'll follow the procedure from before (NYC salaries): we sample with ***replacement***, ***5000*** times, recording the median for each simulation.



- The ***95% confidence interval*** is ***[69.99, 70.16]***. We don't have the real median to compare it to!

SAMPLING DISTRIBUTION AND CONFIDENCE INTERVAL

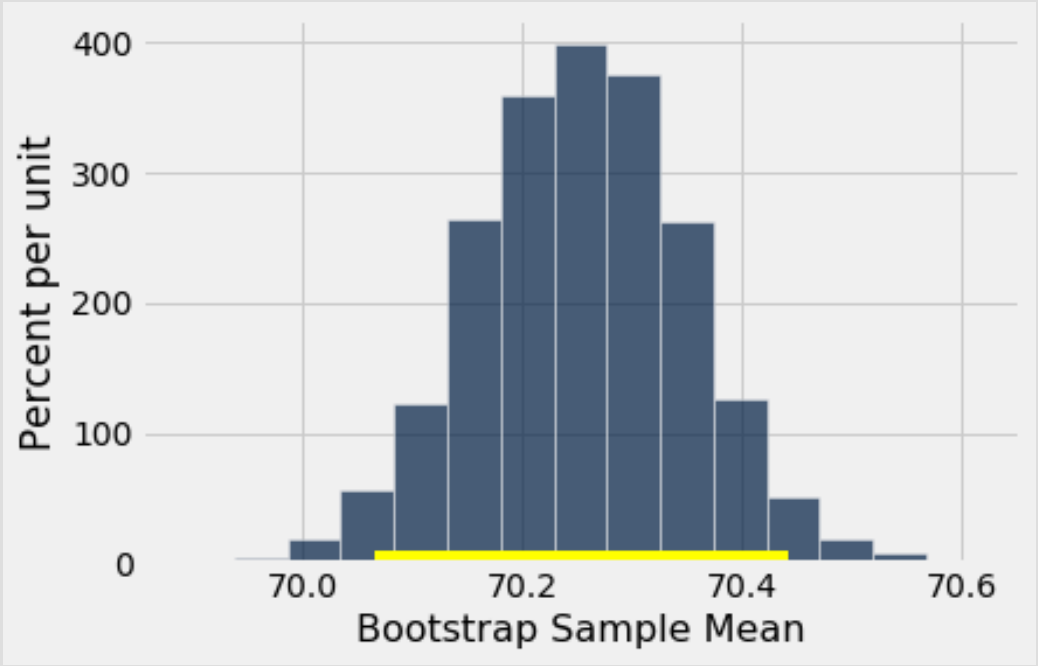
- The median is "***robust***" to outliers, meaning it doesn't vary very much as we resample.



- The ***mean*** is not robust in this way. Let's produce a 95% confidence interval for the mean.
- The mean height here is 70.25

SAMPLING DISTRIBUTION AND CONFIDENCE INTERVAL

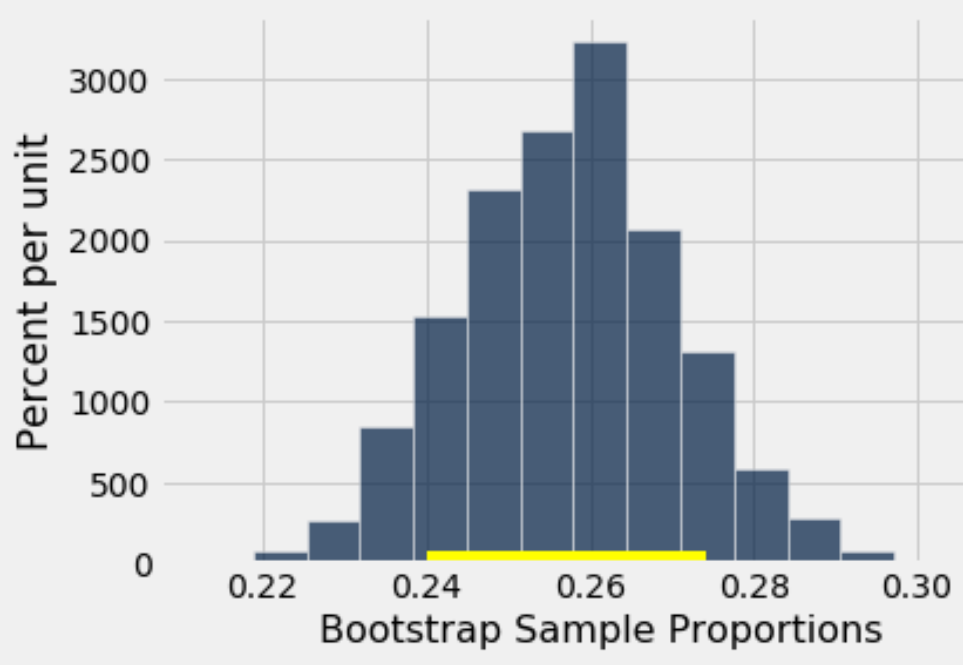
- We'll follow the procedure from before (for the median): we sample with ***replacement***, ***5000*** times, recording the ***mean*** for each simulation.



- The ***95% confidence interval*** is ***[70.06, 70.44]***. We don't have the real mean to compare it to!

SAMPLING DISTRIBUTION AND CONFIDENCE INTERVAL

- Generally, this method of producing confidence intervals works for ***any*** percentile, proportion, ratio etc of the data.



- And it will work for ***different*** percentage confidence intervals too.
- e.g. 80% confidence interval on proportion who say they are "athletic": (0.24, 0.27)

WORKING WITH BOOTSTRAPS

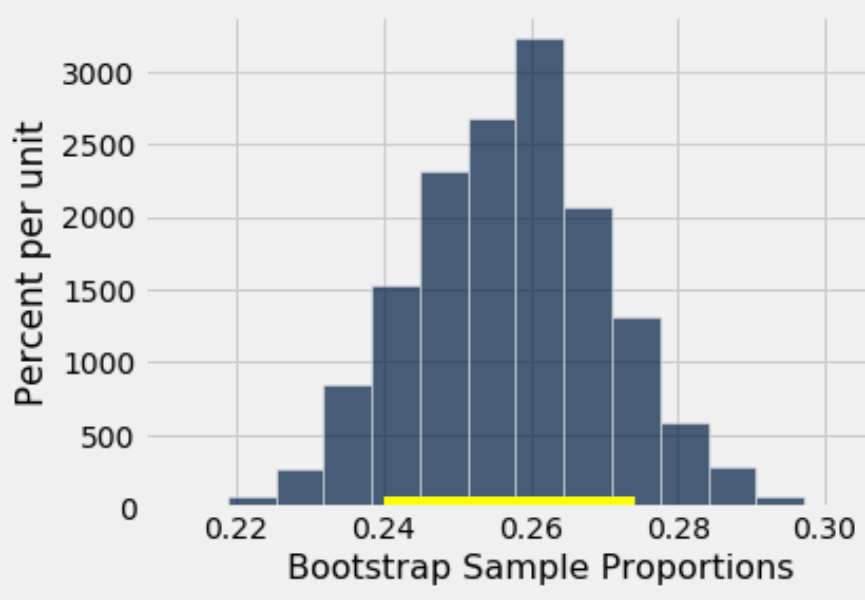
This method for constructing CIs works well if you…

- have a "***large***" random sample, so LLN can kick in
- have ***computational resources*** to do lots (5k, 10k) simulations
- the sampling distribution of the statistic is roughly "***normal***" shaped
- are interested in "***central***" stats like median, mean: can fail for "extreme" statistics like minimum or maximum

WAIT, WHY ARE WE DOING THIS?

The ***p% confidence interval*** gives us a set of numbers: if we follow our procedure many, many times, *p times out of 100* the interval will capture the ***true, population value of the parameter*** we care about.

But why is that ***useful***?



TESTING HYPOTHESES WITH CONFIDENCE INTERVALS

Our 95% confidence interval for mean heights was (***70.06, 70.44***). Suppose…

- **Null hypothesis**: mean height of men on OkCupid is ***same*** as US population (5'9"or 69")

- **Alternative hypothesis**: no, it isn't

TESTING HYPOTHESES WITH CONFIDENCE INTERVALS

- Here, our confidence interval (***70.06, 70.44***). does ***not*** contain 69, so we can ***reject*** the null hypothesis at the ***5%*** level.

- Notice that the ***p-percent*** confidence interval corresponds to ***100-p level of a test***.

- So, if we have a ***99% interval***, we can reject the null at the ***1%*** level (if it doesn't contain the claimed value)

FIN