

# Bootstrapping, Confidence Intervals, Sampling, Probability

## Data 100 Final Review

*Suraj Rampure, Neil Shah*

# Sampling Techniques

## Convenience sample

Quite literally, sampling whoever is convenient

## Simple random sample (SRS)

Sampling uniformly at random from the population

- For example, if we have five students, A, B, C, D, and E, and want to select two of them, our sample will look like AB, AC, AD, AE, BC, BD, BE, CD, CE, or DE.
- There are  $\binom{5}{2}$  total possible samples, and each is equally likely (with probability  $\frac{1}{\binom{5}{2}} = \frac{1}{10}$ ).
- Each student appears in exactly 4 of the samples, so the probability that any one specific student appears in our sample is  $\frac{5-1}{\binom{5}{2}} = \frac{4}{10} = \frac{2}{5}$ .
  - (Extra): In general, if we have  $n$  students and want to choose  $k$  of them, the probability that one specific student is chosen in our sample is  $\frac{\binom{n-1}{k-1}}{\binom{n}{k}}$ .

## Cluster sample

In cluster sampling, we first split our population into clusters. Then, we use SRS to select clusters themselves, and sample everyone within a cluster.

- For example, if we wanted to survey some number of CS and EECS majors at Berkeley, we could split our population into 1st years, 2nd years, 3rd years and 4th years. We could choose two of these clusters at random (i.e. 1/2, 1/3, 1/4, 2/3, 2/4, 3/4) and then **sample all students** in the two clusters.
- We could also choose just one cluster.

## Stratified sample

In stratified sampling, we also split our population into groups (now called strata). However, we now use SRS to sample *within every strata*.

- Continuing with the above example, if we were to proceed with stratified sampling, we would use SRS to select **some** number of 1st, 2nd, 3rd, and 4th years

## Multi-stage sample

Multi-stage samples can be thought of as a combination of cluster samples and stratified samples. In a cluster sample, we:

1. Divide our population into clusters
  2. Use SRS to select clusters
  3. Use SRS within each cluster
- In our example, this would mean first using SRS to select some number of class ranks (e.g. 1st years and 3rd years)
  - Then, we would use SRS to sample **some** 1st years, and **some** 3rd years

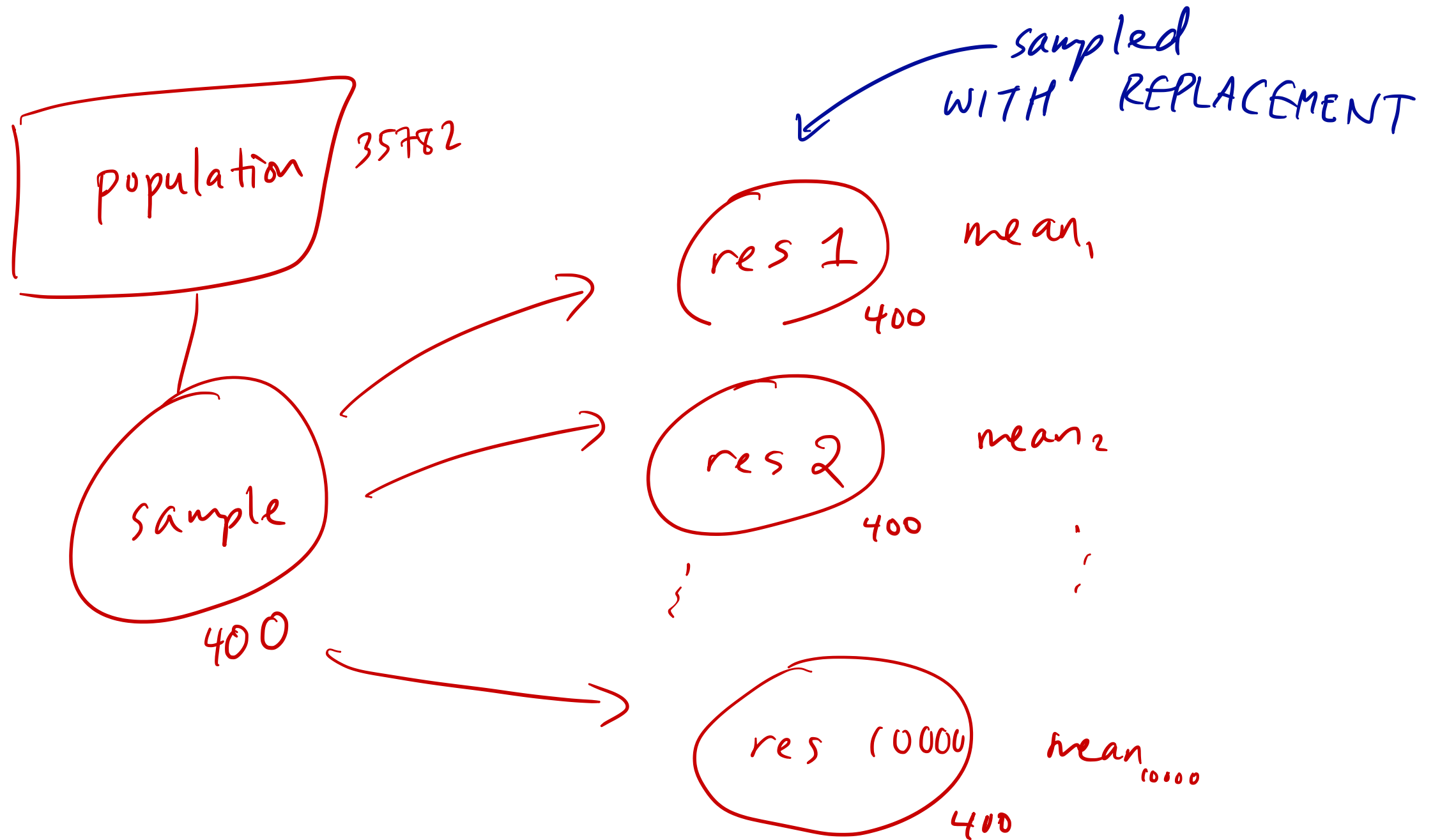
### Recap of the three "grouped" sampling techniques:

- Cluster: Use SRS to select groups, sample everyone within selected groups
- Stratified: Use SRS to sample within every group
- Multi-stage: Use SRS to select groups, use SRS to sample within selected groups

# Bootstrapping

Suppose we want to estimate some parameter of a population (for example, the mean of the heights of everyone at Berkeley), given a (large) sample of the population.

- With bootstrapping, we treat our acquired sample as the new population, and repeatedly resample from it (with replacement). In each of these resamples, we calculate the statistic of interest (e.g. the mean).
- Suppose we resampled 10000 times. We now have 10000 statistics, and we can use these 10000 to create a **confidence interval**
- A  $p\%$  confidence interval is acquired by looking at the inner  $p\%$  of values



to create a  $p\%$  CI:  
look at middle  $p\%$  of values

## Interpreting Confidence Intervals

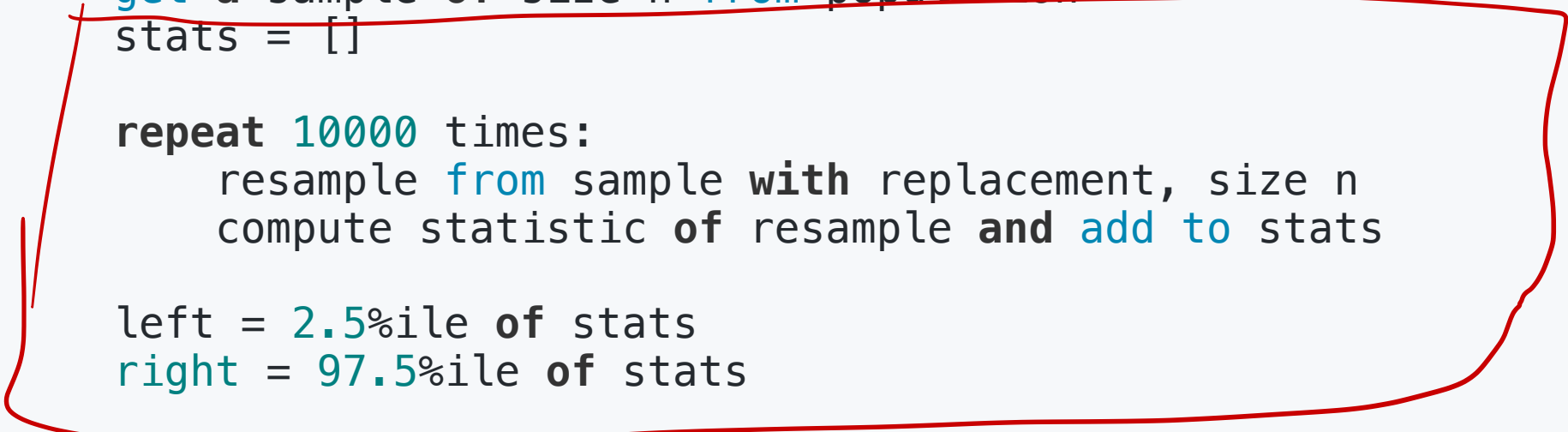
It is **not true** that a 95% confidence interval means that there is a 95% chance that the true population parameter is within the interval. The true parameter is a constant; there is no randomness associated with it, and either it is in the interval, or it is not.

What a confidence interval makes a statement about is the **process**:

- Suppose we repeated this process many times – the process of creating an original sample and bootstrap resampling from it – and each time we created a confidence interval. We would expect 95% of these 95% confidence intervals to contain the true population parameter.

In other words:

```
population = [...]  
CIs = []  
repeat 100 times:  
    get a sample of size n from population  
    stats = []  
    repeat 10000 times:  
        resample from sample with replacement, size n  
        compute statistic of resample and add to stats  
    left = 2.5%ile of stats  
    right = 97.5%ile of stats  
    add [left, right] to CIs
```



We are making the claim that roughly ~95 of the confidence intervals stored in CIs will contain the population parameter.