NYU Center for Data Science

DS-UA 111
Data Science for Everyone

Week 12: Lecture 1
Resampling

Center for
Data Science

NYU

Could sampling from a sample teach us anything about a population?

DS-UA 111
Data Science for Everyone

Week 12: Lecture 1
Resampling

Adapted from Adhikari, DeNero, Wagner, Milner, Foulkes

Announcements

- Please check Week 12 agenda on NYU Classes
  - Homework 3/4
  - Lab 7
  - Project Milestone
- Refer to the Calendar linked to NYU Classes

Review

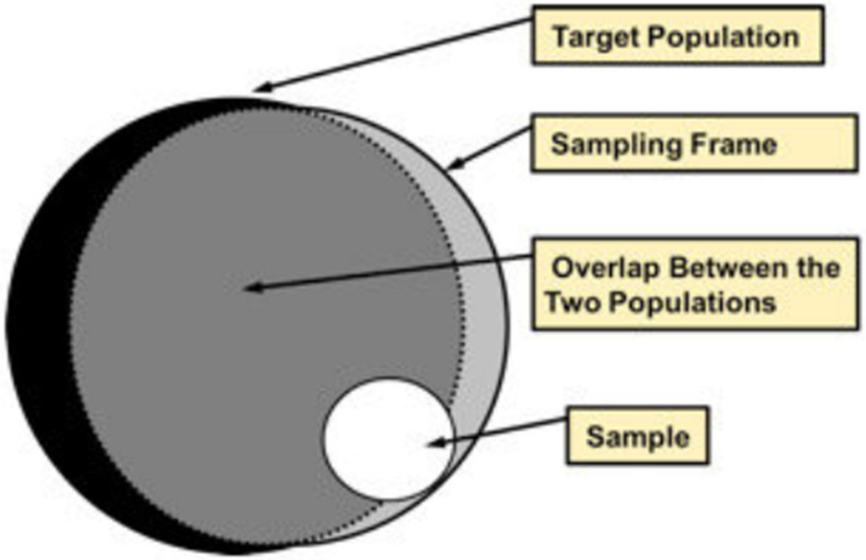
- Permutation Testing
  - Does the distribution of some feature match between two groups?
- Designing Experiments

References

- Comparing Samples:
  - Chapter 12.3

Review

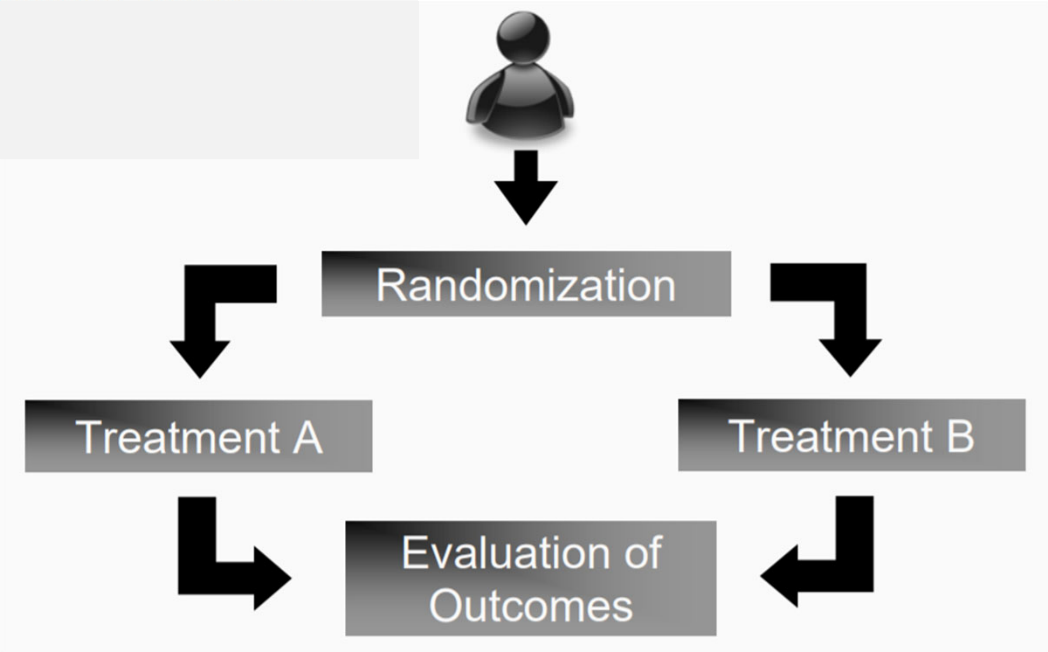
- Researchers perform studies on a target population
- The sampling frame is the subset of the population eligible for inclusion in the sample
- The sample contains randomly or deterministically selected participants or observations from the sampling frame



Review

Randomized Controlled Trial

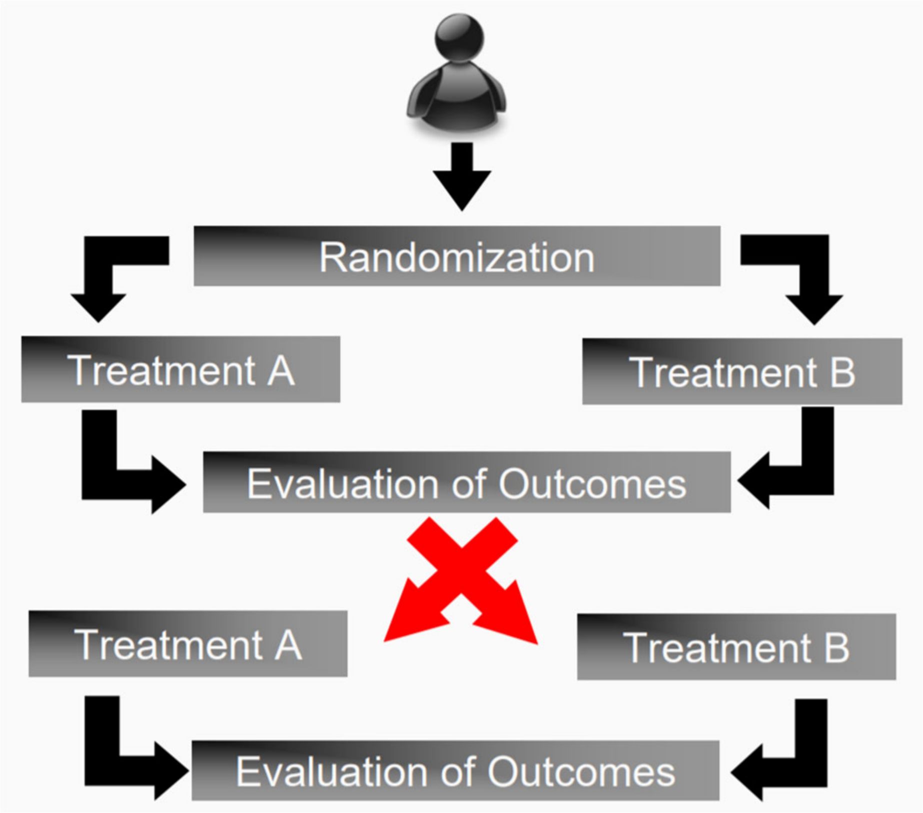
- Researchers randomly split the participants between two groups receiving either treatment A or treatment B
- Here neither the researchers nor the participants should know the division into groups



Review

Cross-Over Design

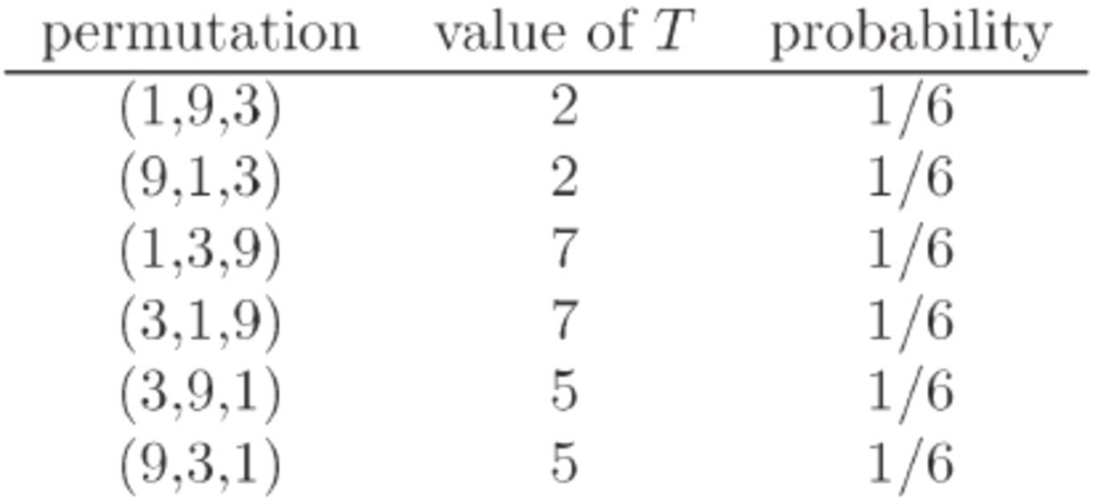
- Researchers repeat the experiment switching the participants between the groups.
- Between the two rounds, each patient has both treatment A and treatment B



Review

Steps for Permutation Testing

1. Fix a null hypothesis and alternative hypothesis
2. Determine a test statistic
3. Calculate the observed test statistic for the sample
4. Simulate test statistics under the null hypothesis with many trials
5. Calculate p-value for the observed test statistic with the empirical distribution



permutation	value of T	probability
(1,9,3)	2	1/6
(9,1,3)	2	1/6
(1,3,9)	7	1/6
(3,1,9)	7	1/6
(3,9,1)	5	1/6
(9,3,1)	5	1/6

Agenda

- Understanding Quantiles
  - Percentiles
  - Box-Plot
- Resampling
  - Bootstrap Method

References

- Estimation
  - Chapters 13.1, 13.2

Percentiles

- Remember that quantiles are cut-points that split a dataset consisting of numbers into subsets of equal size
- Percentiles split the dataset into subsets of size 1%.
- Assume that we arrange the data in increasing order
- The pth percentile of the data is the smallest number in the dataset that is at least as large as p% of all the values.

Percentiles

Steps for Computing Percentiles

- Arrange the data in increasing order.
- Calculate (p/100)×n where n is the size of the dataset. Call this number k.
- If k is an integer, then take the k^{th} number in the dataset.
- If k is not an integer, then round k up to the next integer and take that number in the dataset.

Box-Plots

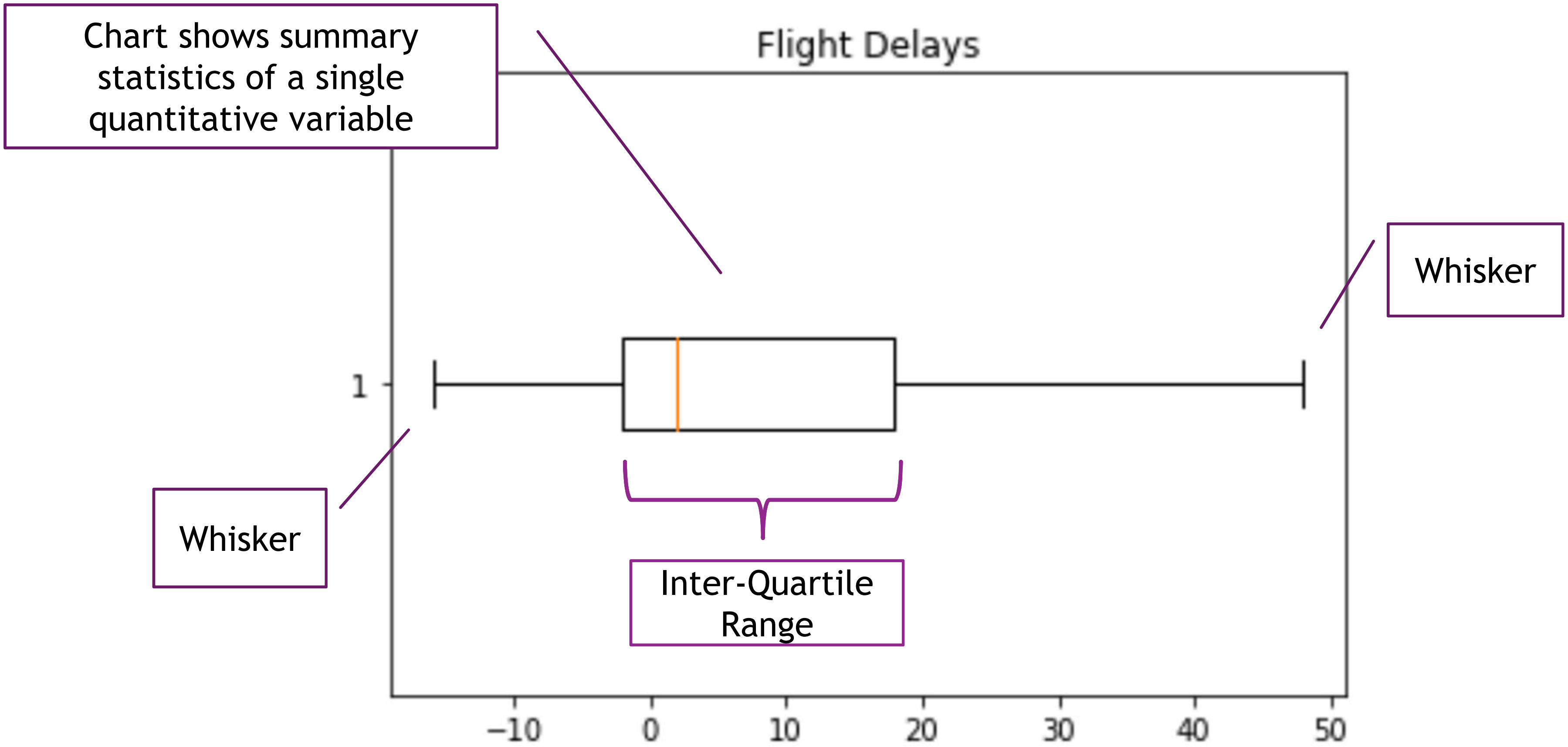


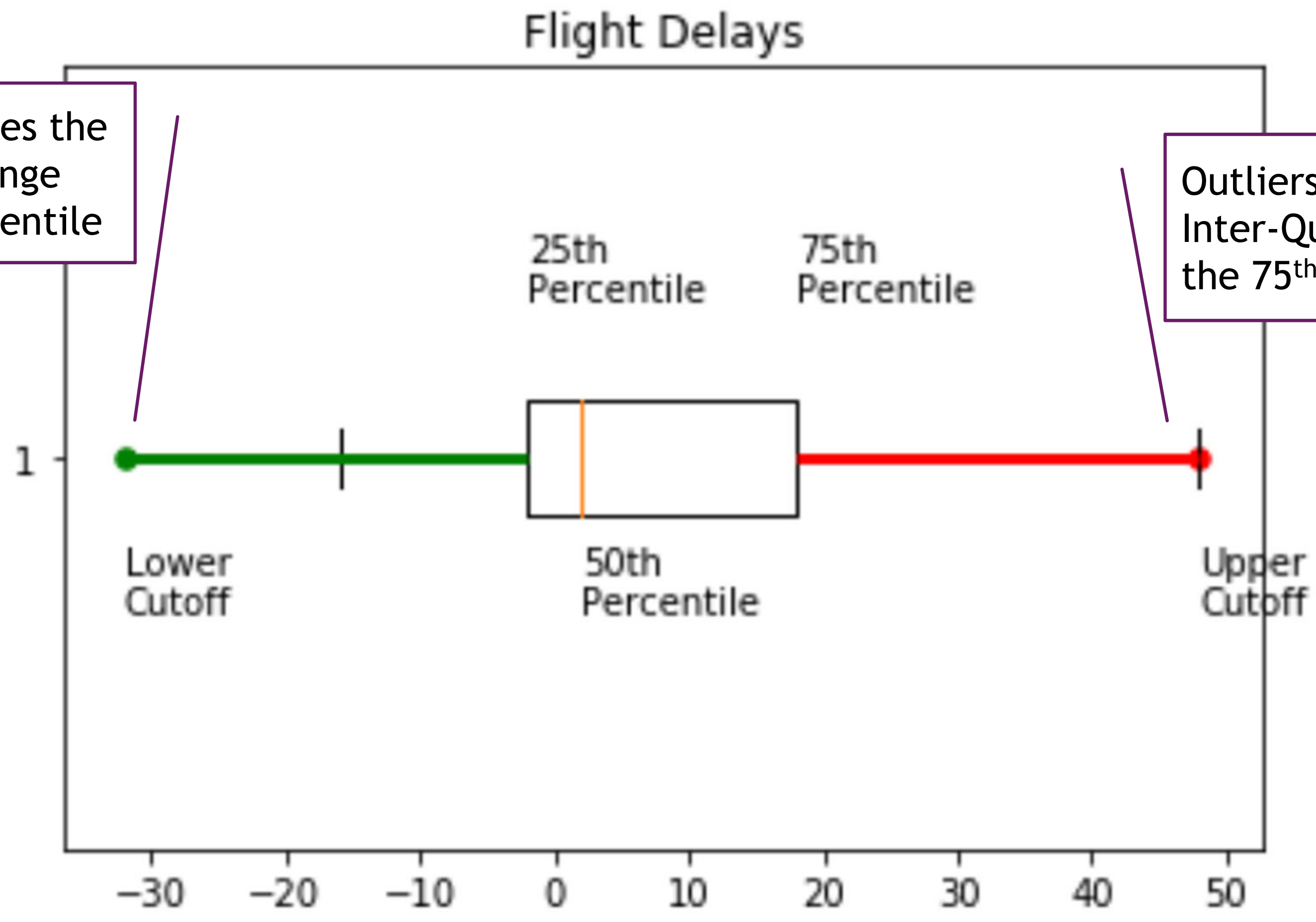
Chart shows summary statistics of a single quantitative variable

Whisker

Whisker

Inter-Quartile Range

Box-Plots

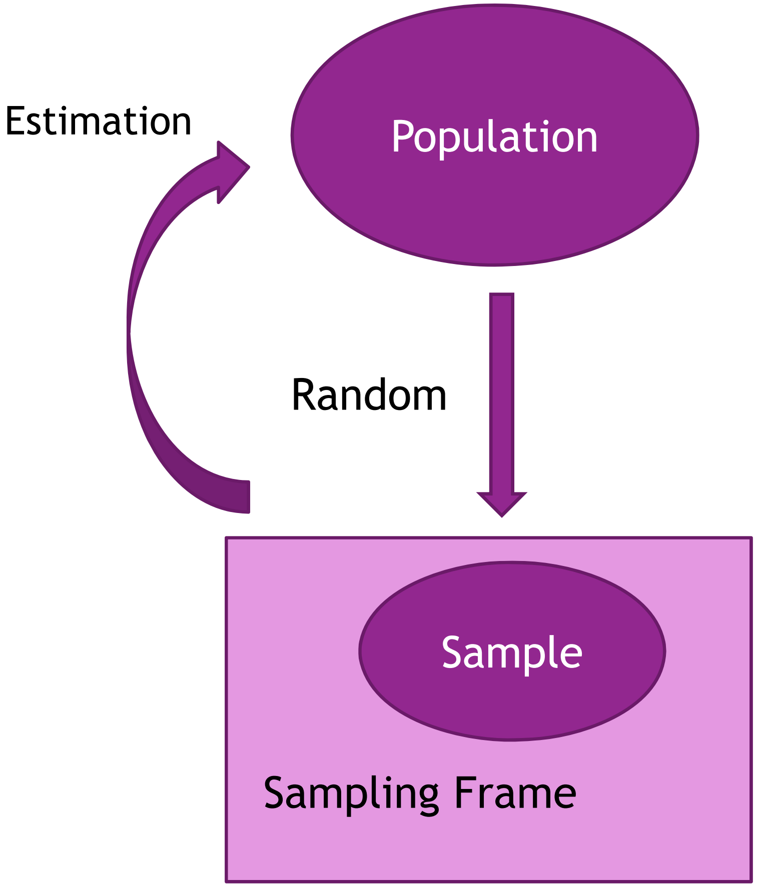


Outliers are 1.5 times the Inter-Quartile Range below the 25th Percentile

Outliers are 1.5 times the Inter-Quartile Range above the 75th Percentile

Estimation

- We want to study populations. We call unknown quantities related to the population parameters
- Hypothesis testing allowed us to compare
  - sample and population
  - two samples



Estimation

- If we have a census containing all data about a population, then we do not need hypothesis testing
- If we lack a census or the census is too large for calculation, then we need samples

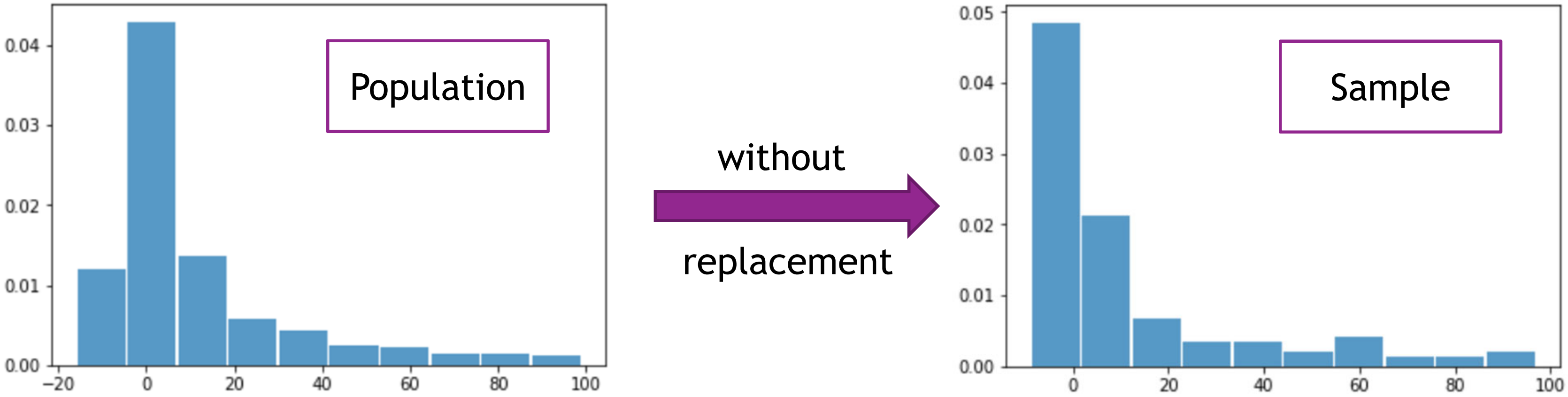
1. Take a sample at random from the population
2. Compute a statistic to estimate the parameter
3. Repeat to understand variability in the estimate

Resampling

- One sample yields one estimate of the parameter. So another estimate means another samples at random from the population.
- However, we may not be able to conduct another observational study or experiment
- If we can make some assumptions about the population, then we could try to simulate samples.
- However, if we lack enough information about the population to make assumptions, then we have to resample.

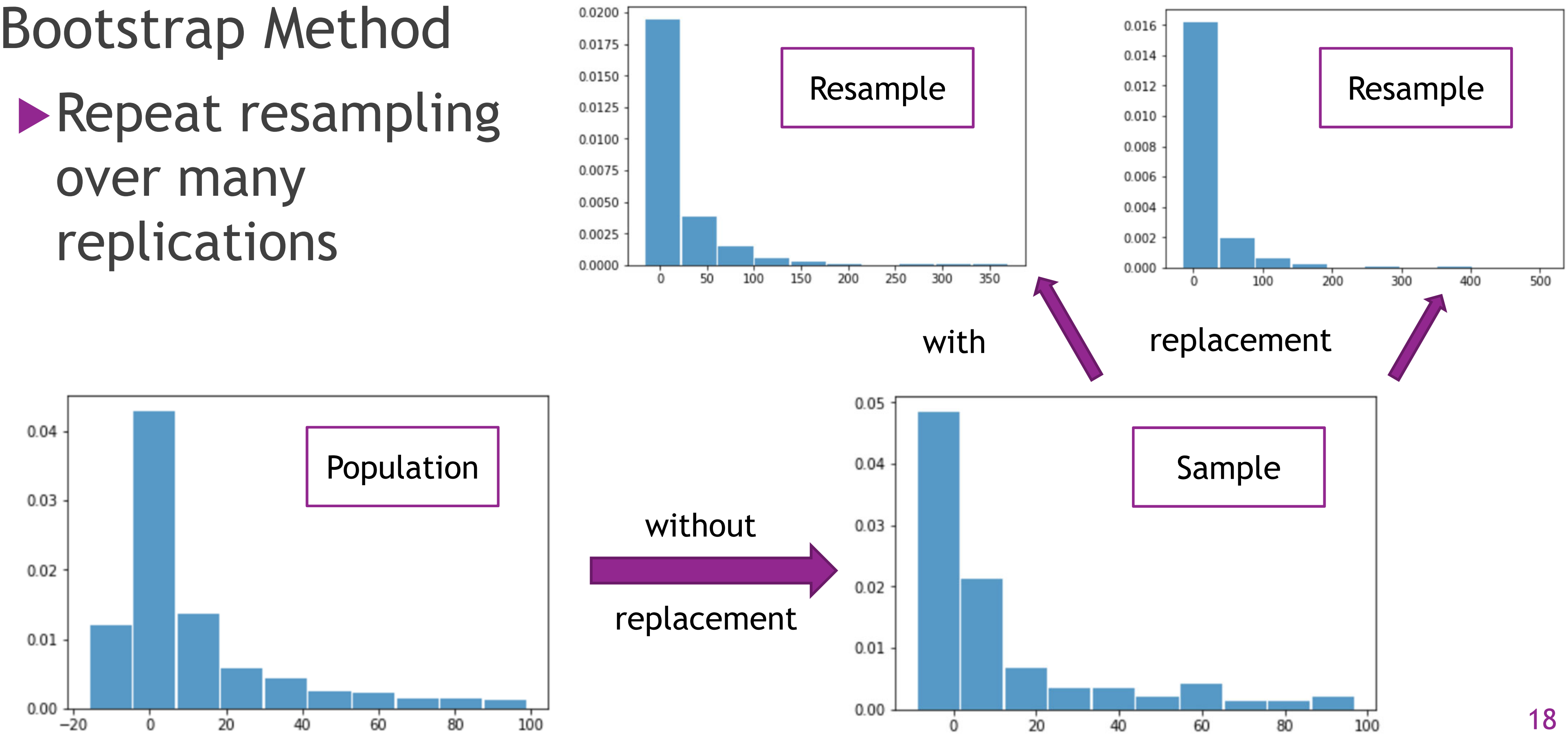
Bootstrap

- Bootstrap Method
  - Draw at random with replacement from the sample
  - The size of the resample should equal the size of the sample



Bootstrap

- Bootstrap Method
  - Repeat resampling over many replications



Resample

Resample

with replacement

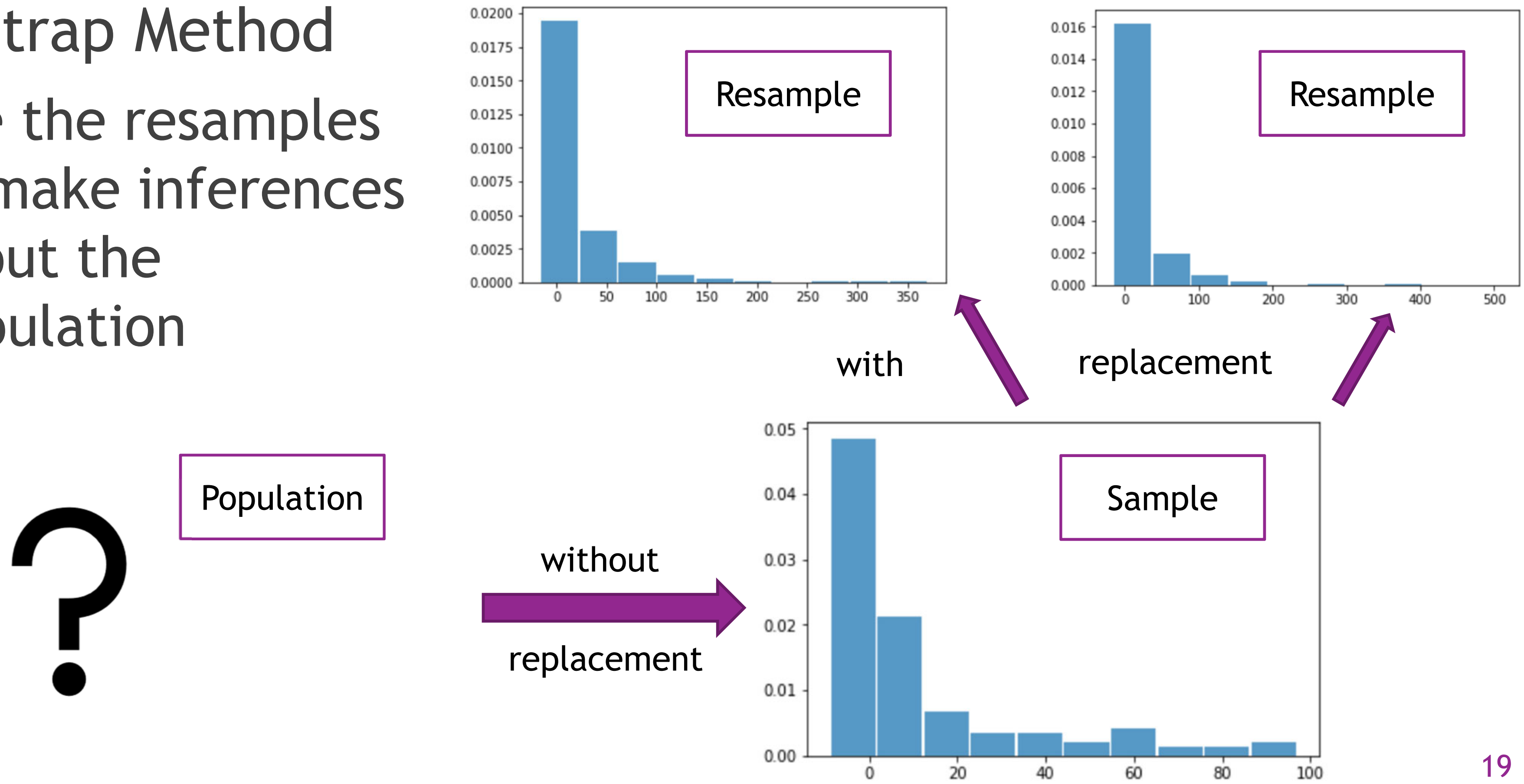
Population

without replacement

Sample

Bootstrap

- Bootstrap Method
  - Use the resamples to make inferences about the population



Population

without replacement

Sample

with replacement

Resample

Resample

Summary

- Understanding Quantiles
  - Percentiles
  - Box-Plot
- Resampling
  - Bootstrap Method

Goals

- Understand the calculation of percentiles
- Visualize different percentiles through a box-plot
- Resample data to assess differences between estimates