NYU | Center for Data Science

DS-UA 111
Data Science for Everyone

Week 12: Lecture 2
Confidence Intervals

Center for
Data Science

NYU

How can we use resampling for hypothesis testing?

DS-UA 111
Data Science for Everyone

Week 12: Lecture 2
Confidence Intervals

Adapted from Adhikari, DeNero, Wagner, Milner

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

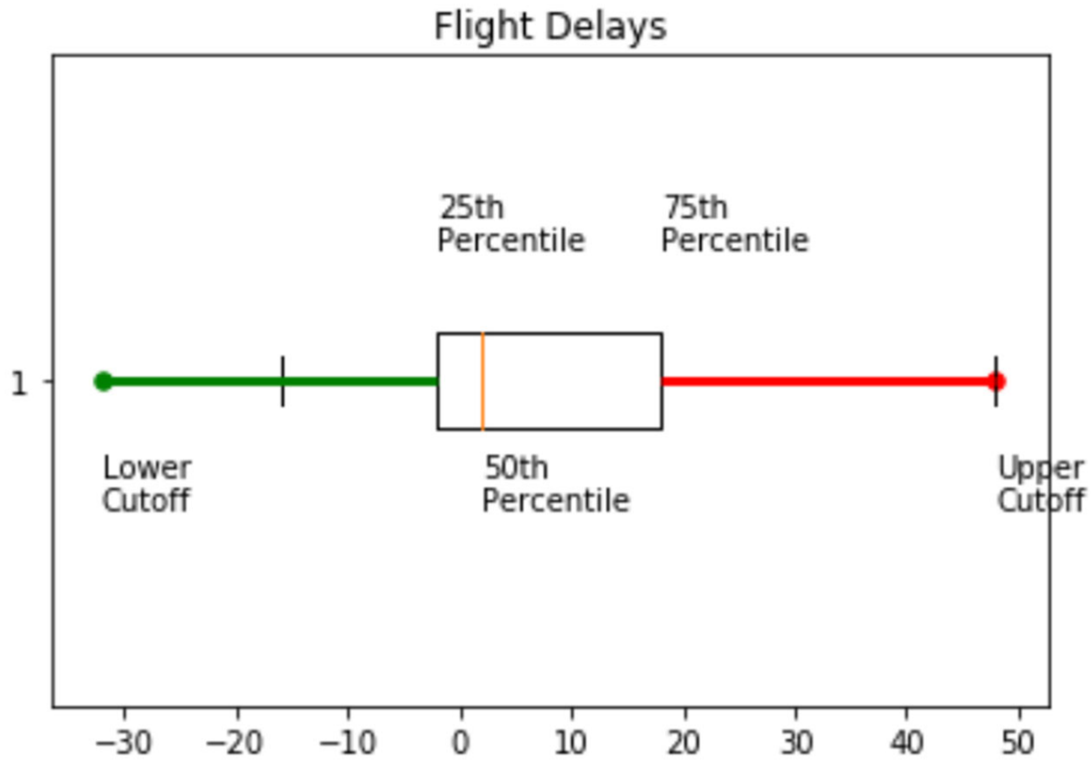
- Understanding Quantiles
  - Percentiles
  - Box-Plot
- Resampling
  - Bootstrap Method

References

- Comparing Samples:
  - Chapters 13.1, 13.2

Review

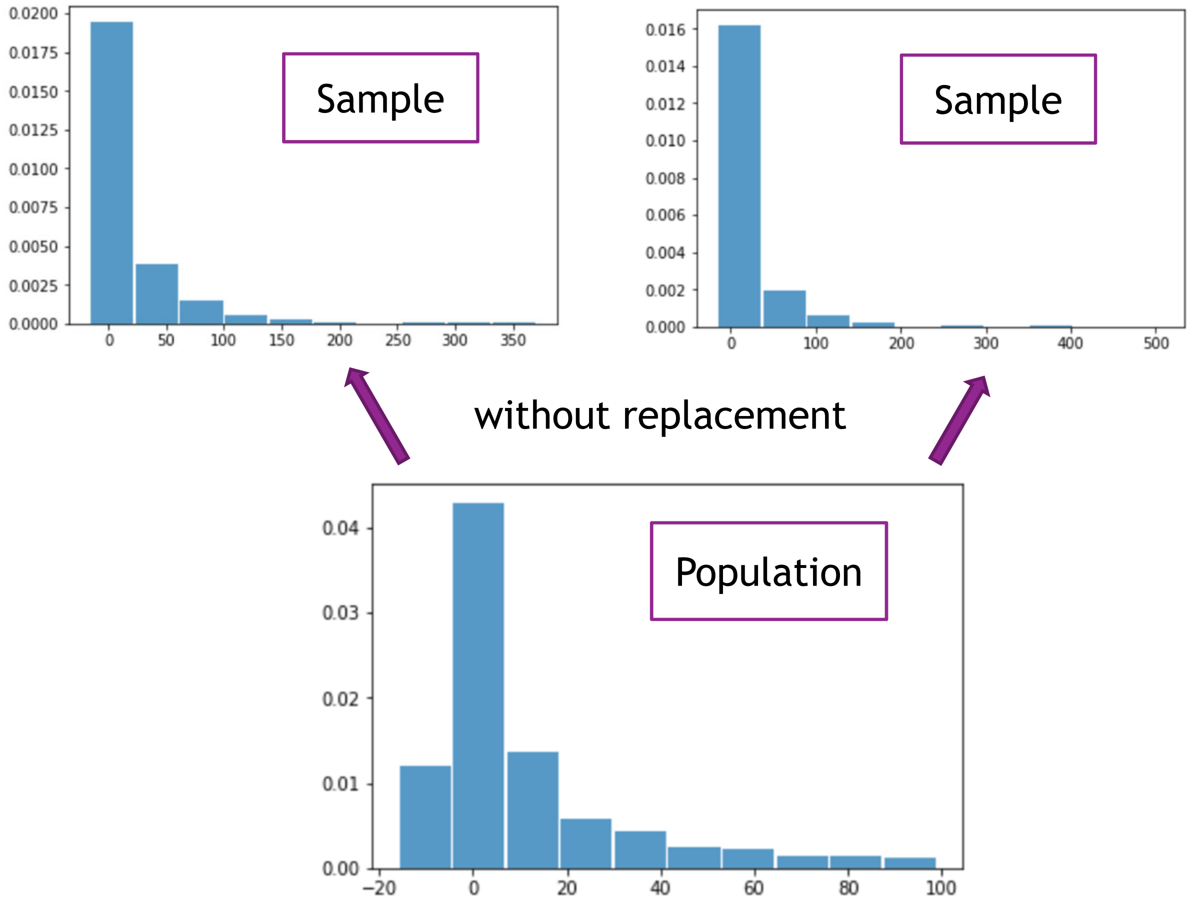
- Boxplot
  - The left end of the box indicates 25th percentile
  - The right end of the box indicates 75th percentile
  - The middle of the box indicates 50th percentile
  - The whiskers indicate the values closest to the upper bound and lower bound for outliers determined by the Inter-Quartile Range



Review

Estimation

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



Sample

Sample

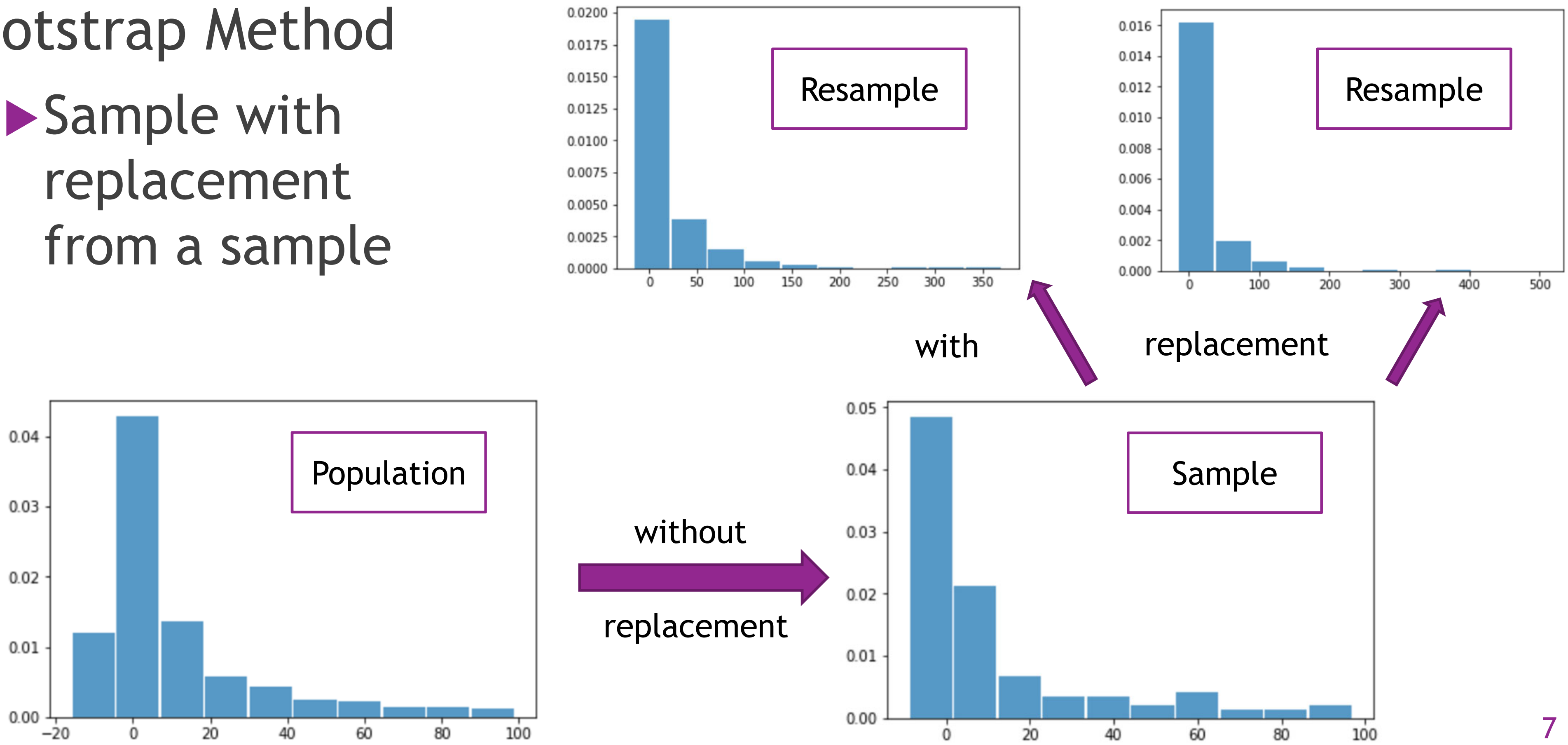
without replacement

Population

Review

Bootstrap Method

- Sample with replacement from a sample



with replacement

without replacement

Agenda

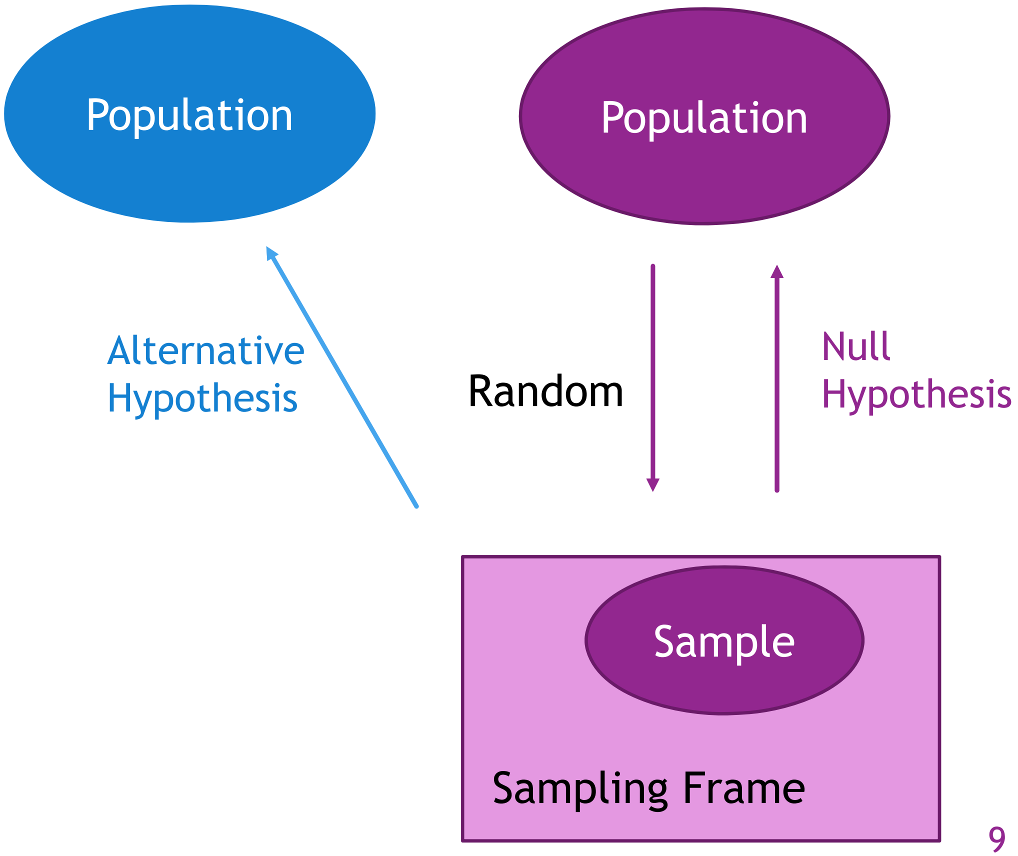
- Confidence Intervals
  - Resampling for hypothesis testing
- Averages
  - Understanding differences from the average

References

- Estimation
  - Chapters 13.3, 13.2
  - Chapters 14.1, 14.2

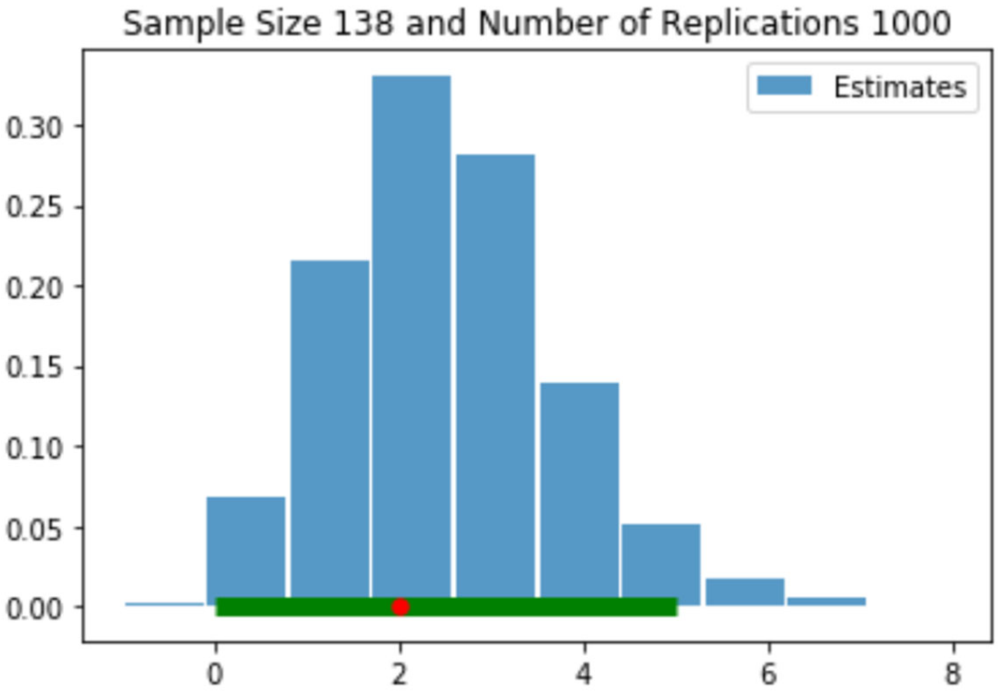
Hypothesis Testing

- We use hypothesis testing to generate evidence about two understandings of a population
  - Null Hypothesis
  - Alternative Hypothesis
- For the evidence we can compute statistics from
  - Samples
  - Simulation
  - Resamples



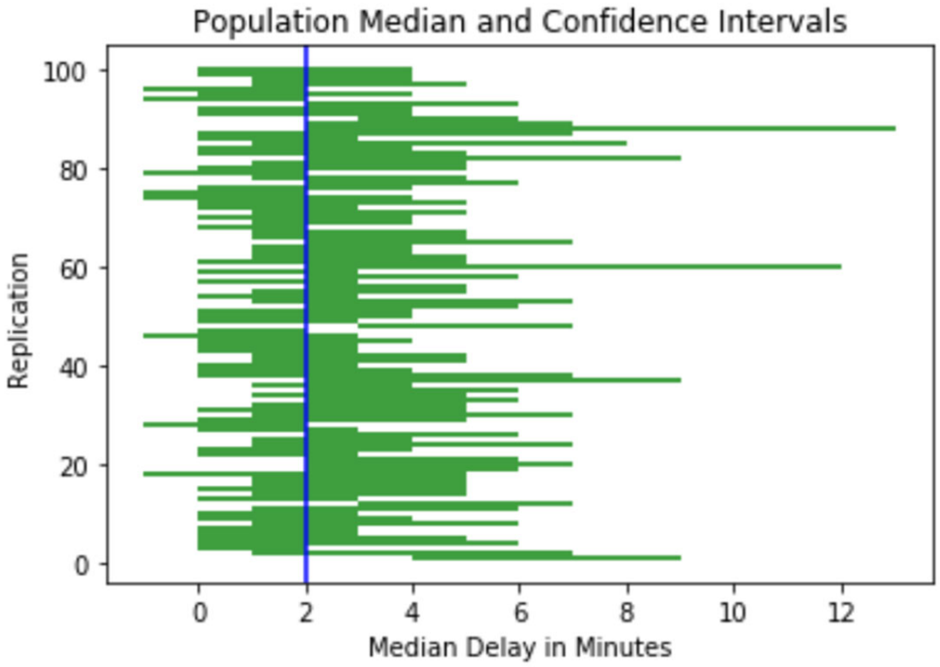
Confidence Intervals

- With test statistics calculated from resamples, we can generate confidence interval.
- These intervals provide ranges of estimates of a parameter. We have confidence in the process that generated the interval
- We could look at any percent between 0 and 100. However we tend to take 90% or 95%



Confidence Intervals

- For constructing a confidence interval for an unknown parameter
  1. Sample without replacement from the population to determine a sample. Larger samples are preferable to smaller sample.
  2. Sample with replacement from the sample to get a resample. Calculate the test statistic on the resample.
  3. Repeat Step 2 many times. Each replication generates another number.
  4. For an approximate 80% confidence interval, take the 10th and 90th percentiles of all the resample estimates.

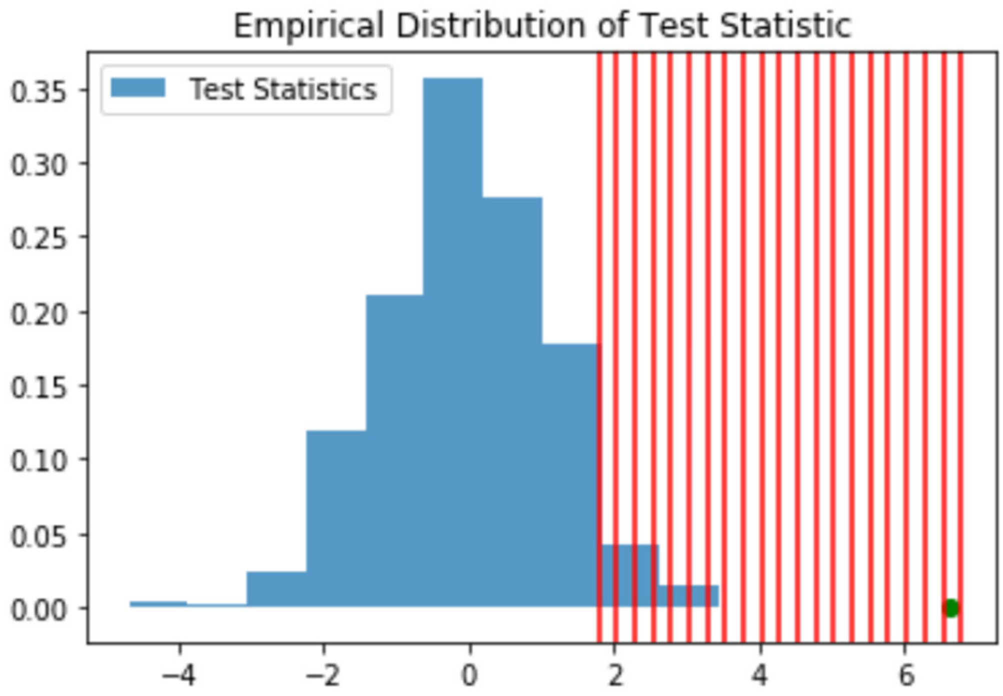


Confidence Intervals

- We can use confidence intervals from resamples for hypothesis testing in the same way that we use p-values for samples or simulations.
  1. Define null hypothesis and alternative hypothesis.
  2. Determine a test statistic for the sample and resamples that estimates the parameter in the population
  3. Construct a confidence interval for the parameter from the test statistics calculated on the resamples.
  4. If the confidence interval does not contain the sample test statistic, then reject the null hypothesis. If the confidence interval does contain the sample test statistic, then fail to reject the null hypothesis.
- If a 95% confidence interval does not contain the number in question, that’s like a p-value being less than a 0.05 cutoff value.

Confidence Intervals

- We could think of confidence intervals like acceptance regions complementary to the rejection regions for p-values.
- We can relate a cut-off for p-value p% and a (100-p)% confidence interval.



Limitations

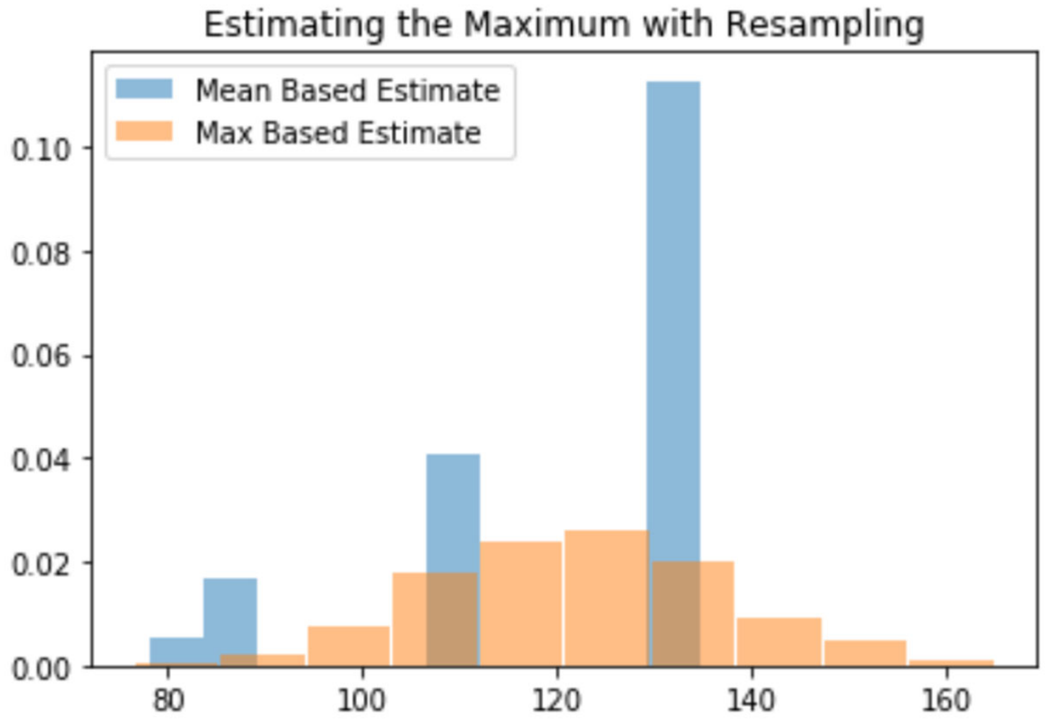
- If we take a p% confidence interval for hypothesis testing then p% of the time we expect the interval to contain the population parameter
  - So we have a false reject about (100-p)% of the time
- However if the goal is to estimate any parameters related to rare elements of the population then the confidence intervals from resamples might be inaccurate. Parameters like
  - Maximum Value
  - Minimum Value

Remember that the resamples cannot contain data outside of the sample. So if the sample is small, then the confidence intervals from resamples might be inaccurate

Limitations

- Suppose we want to estimate the average age of mothers in a population.
- We use bootstrap resampling to generate approximate 95% confidence interval for the average age of the mothers in the population

  26.9 years to 27.6 years
- True or False
  - About 95% of the mothers in the population were between 26.9 years and 27.6 years old.
  - False: We're estimating that their average age is in this interval.



Mean

- Among different statistics for estimate of parameters, the mean lends itself to resampling.
- We can make some observations about the mean that hold for any population.
  - The mean of numbers might not be contained among the numbers
    - For example the mean of integers could be a fraction
  - The mean is greater than the minimum and less than the maximum
  - The mean aggregates many numbers into one representative number

- Suppose we have data

  {2,3,9,9}

- The mean is

  4.25 = (2 + 3 + 9 + 9) / 4

- We can think of the mean as a weighted sum of the numbers. Here the weights reflect the frequency of the number

$$4.25 = 2*(¼) + 3*(¼) + 3*(¼) + 9*(¼)$$

$$= 2*(¼) + 3*(2/4) + 9*(¼)$$

$$= 2*0.25 + 3*0.5 + 9*0.25$$

Standard Deviation

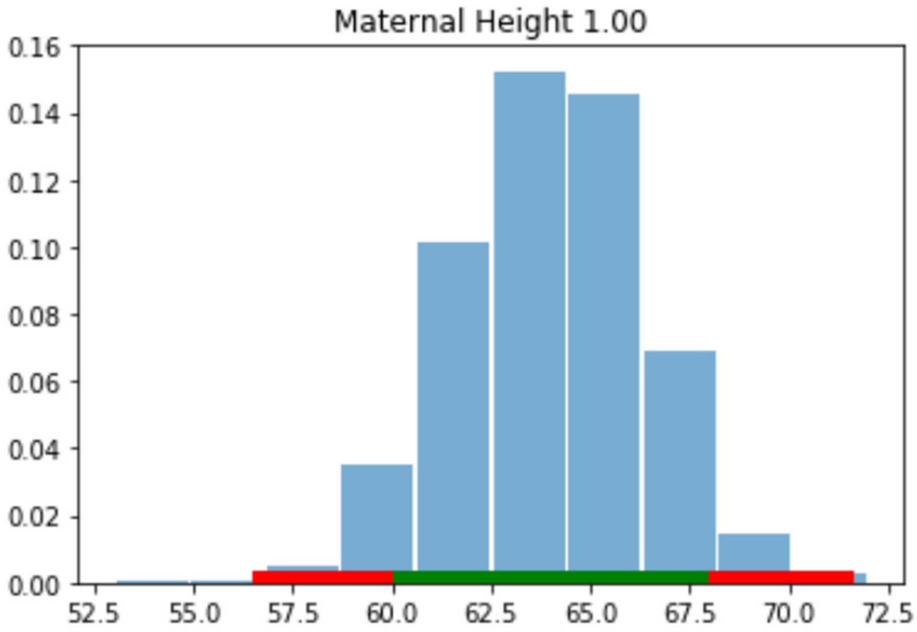
- The standard deviation measures the difference between numbers and the mean of the numbers.
- Regardless of the population we know that a certain amount of the data has to lie near the mean.
- The standard deviation bound tells use the fraction of data greater than

Mean – z * (Standard Deviation)

and less than

Mean + z * (Standard Deviation)

is a least 1 – (1 / z^2)



Summary

- Confidence Intervals
  - Resampling for hypothesis testing
- Averages
  - Understanding differences from the average

Goals

- Use bootstrap method for hypothesis testing
- Understand some limitations of resampling
- Bound the difference between numbers and their average