YData: An Introduction to Data Science

Lecture 24: Interpreting Confidence

Jessi Cisewski-Kehe and John Lafferty
Statistics & Data Science, Yale University

Spring 2019

Credit: data8.org

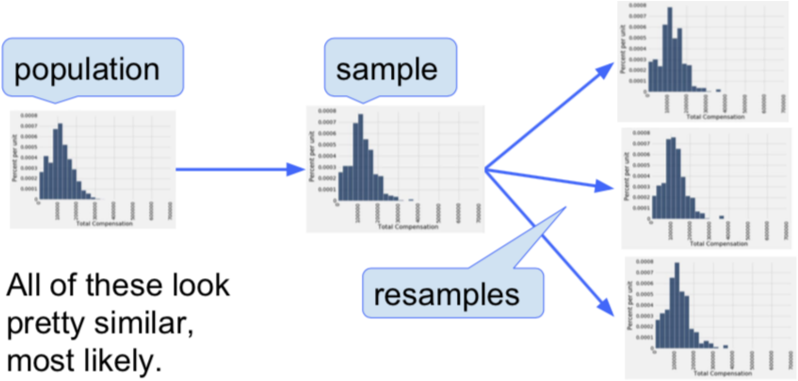
Announcements

The Bootstrap

Key to Resampling

- From the original sample,
  - draw at random
  - with replacement
  - as many values as the original sample contained
- The size of the new sample has to be the same as the original one, so that the two estimates are comparable

Why the Bootstrap Works



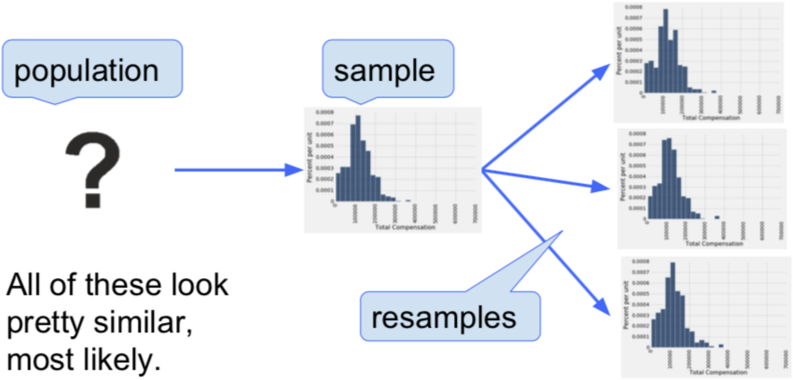
population

sample

resamples

All of these look pretty similar, most likely.

Inference Using the Bootstrap



population

?

sample

resamples

All of these look pretty similar, most likely.

95% Confidence Interval

- Interval of estimates of a parameter

- Based on random sampling

- 95% is called the confidence level
  - Could be any percent between 0 and 100
  - Higher level means wider intervals

- The confidence is in the process that generated the interval:
  - It generates a “good” interval about 95% of the time.

(DEMO)

Use Methods Appropriately

Can You Use a CI Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

Can You Use a CI Like This?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- About 95% of the mothers in the population were between 26.9 years and 27.6 years old.

Answer: **False**. We're estimating that their **average age** is in this interval.

(DEMO)

Is This What a CI Means?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- There is a 0.95 probability that the average age of mothers in the population is in the range 26.9 to 27.6 years.

Is This What a CI Means?

By our calculation, an approximate 95% confidence interval for the average age of the mothers in the population is (26.9, 27.6) years.

True or False:

- There is a 0.95 probability that the average age of mothers in the population is in the range 26.9 to 27.6 years.

Answer: **False**. The average age of the mothers in the population is unknown but it's a constant. It's not random. No chances involved.

When Not to Use The Bootstrap

- If you're trying to estimate very high or very low percentiles, or min and max
- If you're trying to estimate any parameter that's greatly affected by rare elements of the population
- If the probability distribution of your statistic is not roughly bell shaped (the shape of the empirical distribution will be a clue)
- If the original sample is very small

(DEMO)

Confidence Intervals For Testing

Using a CI for Testing

- Null hypothesis: **Population average = x**
- Alternative hypothesis: **Population average $\neq$ x**
- Cutoff for P-value: p%
- Method:
  - Construct a (100-p)% confidence interval for the population average
  - If x is not in the interval, reject the null
  - If x is in the interval, can't reject the null

(DEMO)