Estimation, Part 3

The Bootstrap

NASHVILLE
SOFTWARE
SCHOOL

Estimation

Recall: We can construct a confidence interval for a parameter if we understand what the sampling distribution of that parameter looks like.

Eg. Using *t*-distributions for the sampling distribution of the mean

Problems:

- We have to make assumptions about the population of interest to use particular sampling distributions.
- It's not always easy to find the sampling distribution for certain parameters.

Estimation

Big Idea: Say we want to find a 95% confidence interval for some parameter p.

When building a confidence interval, we needed to determine the *margin of error*. This was done by understanding something about the sampling distribution of the mean.

Specifically, we needed to know about the amount of variability in the sampling distribution of the mean. This was where the t-distribution came in handy.

NASHVILLE SOFTWARE SCHOOL

Bootstrap Estimation

Big Idea: We want a 95% confidence interval for the parameter *p*, based on taking a single sample and calculating the statistic *s*.

Say we can can find numbers *a* and *b* so that for 95% of samples,

$$a \leq s - p \leq b$$

Bootstrap Estimation

Big Idea: We want a 95% confidence interval for the parameter *p*, based on taking a single sample and calculating the statistic *s*.

Say we can can find numbers *a* and *b* so that for 95% of samples,

$$a \leq s - p \leq b$$

Example: When estimating the mean, we used the fact that for 95% of samples,

$$-t_{0.025} \cdot \frac{s}{\sqrt{n}} \leq \bar{x} - \mu \leq t_{0.025} \cdot \frac{s}{\sqrt{n}}$$

For 95% of samples: $$a \leq s - p \leq b$$

For 95% of samples: $a \leq s - p \leq b$

For 95% of samples: $-s + a \leq -p \leq -s + b$

For 95% of samples: $$a \leq s - p \leq b$$

For 95% of samples: $$-s + a \leq -p \leq -s + b$$

For 95% of samples: $$s - a \geq p \geq s - b$$

For 95% of samples: $$a \leq s - p \leq b$$

For 95% of samples: $$-s + a \leq -p \leq -s + b$$

For 95% of samples: $$s - a \geq p \geq s - b$$

We get a 95% confidence interval: $$[s - b, s - a]$$

Bootstrap Estimation

Question: How do we find a and b?

Bootstrap Estimation

Question: How do we find a and b?

The Bootstrap Principle: We can understand the variation in the sample statistic s by resampling the original data.

Bootstrap Estimation

Question: How do we find a and b?

The Bootstrap Principle: We can understand the variation in the sample statistic s by resampling the original data.

Specifically, by resampling we can approximate the distribution of $p - s$.

Bootstrap Estimation

Question: How do we find a and b?

The Bootstrap Principle: We can understand the variation in the sample statistic s by resampling the original data.

Specifically, by resampling we can approximate the distribution of $p - s$.

So what do we mean by **resampling**?

Bootstrap Estimation

Given a dataset, a **resample** from that dataset is a random sample drawn with replacement from that dataset of the same size as the original dataset.

Bootstrap Estimation

Given a dataset, a **resample** from that dataset is a random sample drawn with replacement from that dataset of the same size as the original dataset.

With replacement means that each element can be included potentially multiple times.

Resampling

Original Data:

5	15	16	19	21	22

Resampling

Original Data:	5	15	16	19	21	22
Resample # 1:	5	16	16	19	19	21

Resampling

Original Data:	5	15	16	19	21	22
Resample # 1:	5	16	16	19	19	21
Resample # 2:	15	16	19	21	22	22

Resampling

Original Data:	5	15	16	19	21	22
Resample # 1:	5	16	16	19	19	21
Resample # 2:	15	16	19	21	22	22
Resample # 3:	5	5	19	19	21	22

Big Idea:

To build a 95% confidence interval, we need to find *a* and *b* so that for 95% of samples,

$$a \leq s - p \leq b$$

Big Idea:

To build a 95% confidence interval, we need to find *a* and *b* so that for 95% of samples,

$$a \leq s - p \leq b$$

But we don't know the distribution of *s - p*; we're trying to estimate *p*.

Big Idea:

To build a 95% confidence interval, we need to find *a* and *b* so that for 95% of samples,

$$a \leq s - p \leq b$$

But we don't know the distribution of *s* - *p*; we're trying to estimate *p*.

However, we can approximation it using the distribution of s^* - *s*, where s^* is the statistic computed on a resample drawn for our initial sample.

$$s - p \Longleftrightarrow s^* - s$$

Resampling

Original Data:	5	15	16	19	21	22
Resample # 1:	5	16	16	19	19	21
Resample # 2:	15	16	19	21	22	22
Resample # 3:	5	5	19	19	21	22

$$\bar{x} = 16.333$$

Resampling

Original Data:	5	15	16	19	21	22	$\bar{x} = 16.333$
Resample # 1:	5	16	16	19	19	21	$\bar{x}^* = 16$ $\bar{x}^* - \bar{x} = -0.333$
Resample # 2:	15	16	19	21	22	22	
Resample # 3:	5	5	19	19	21	22	

Resampling

Original Data:	5	15	16	19	21	22	$\bar{x} = 16.333$
Resample # 1:	5	16	16	19	19	21	$\bar{x}^* = 16$ $\bar{x}^* - \bar{x} = -0.333$
Resample # 2:	15	16	19	21	22	22	$\bar{x}^* = 19.167$ $\bar{x}^* - \bar{x} = 2.833$
Resample # 3:	5	5	19	19	21	22	

Resampling

Original Data:	5	15	16	19	21	22	$\bar{x} = 16.333$
Resample # 1:	5	16	16	19	19	21	$\bar{x}^* = 16$ $\bar{x}^* - \bar{x} = -0.333$
Resample # 2:	15	16	19	21	22	22	$\bar{x}^* = 19.167$ $\bar{x}^* - \bar{x} = 2.833$
Resample # 3:	5	5	19	19	21	22	$\bar{x}^* = 15.167$ $\bar{x}^* - \bar{x} = -1.167$

Bootstrap Estimation

Procedure for Building a 95% Bootstrap Confidence Interval:

Given a sample, find the sample statistic *s*.

Bootstrap Estimation

Procedure for Building a 95% Bootstrap Confidence Interval:

Given a sample, find the sample statistic s.

1. Draw a large number (10,000 or so) resamples from the original sample and calculate the statistic s^* for each.

Bootstrap Estimation

Procedure for Building a 95% Bootstrap Confidence Interval:

Given a sample, find the sample statistic s.

1. Draw a large number (10,000 or so) resamples from the original sample and calculate the statistic s^* for each.
2. Find the 0.025 and 0.975 quantiles of the set of $s^* - s$, a and b, respectively.

Bootstrap Estimation

Procedure for Building a 95% Bootstrap Confidence Interval:
Given a sample, find the sample statistic *s*.

1. Draw a large number (10,000 or so) resamples from the original sample and calculate the statistic s^* for each.
2. Find the 0.025 and 0.975 quantiles of the set of $s^* - s$, *a* and *b*, respectively.
3. The 95% confidence interval is given by

$$[s - b, s - a]$$