ECS7024 Statistics for Artificial Intelligence and Data Science

Topic 15: Bootstrap

William Marsh

Outline

- Aim: outline the idea of the bootstrap, a direct method of estimate a confidence interval for a statistic

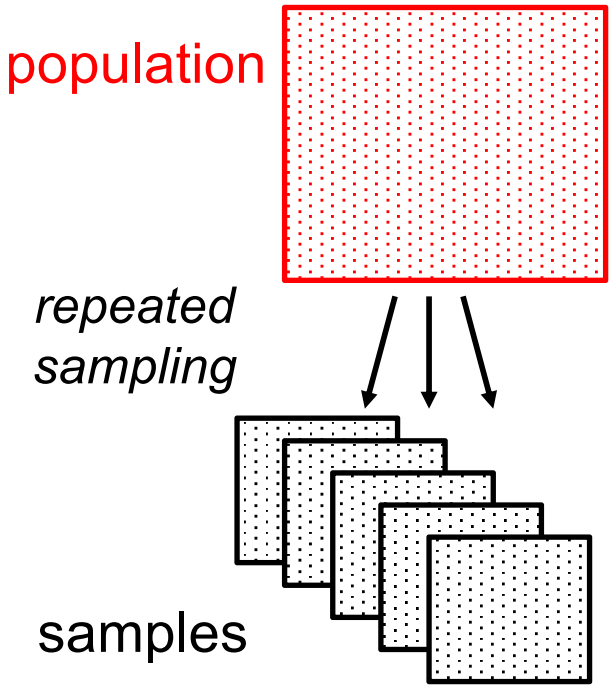
- Principles
- Example
- Discussion

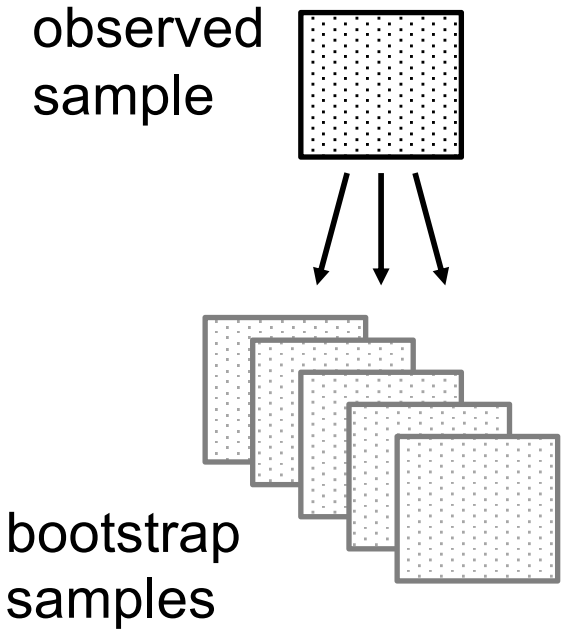
Bootstrap Principles

CI without a Sample Distribution from Theory

Bootstrap

- In a simulation we repeated sample a known population
- In a bootstrap, we resample the sample





Non-parametric Bootstrap

Bootstrap sample based on sample

No distribution assumed

Sampling with Replacement

- Re-sample from the ‘observed sample’ with replacement
- Bootstrap sample
  - Same size as original
  - Some records omitted
  - Some records repeated

Related Terms

- ‘Bootstrap aggregation’ or ‘bagging’
- Resampling (with or w/o replacement)
- Permutation test

Bootstrap Steps

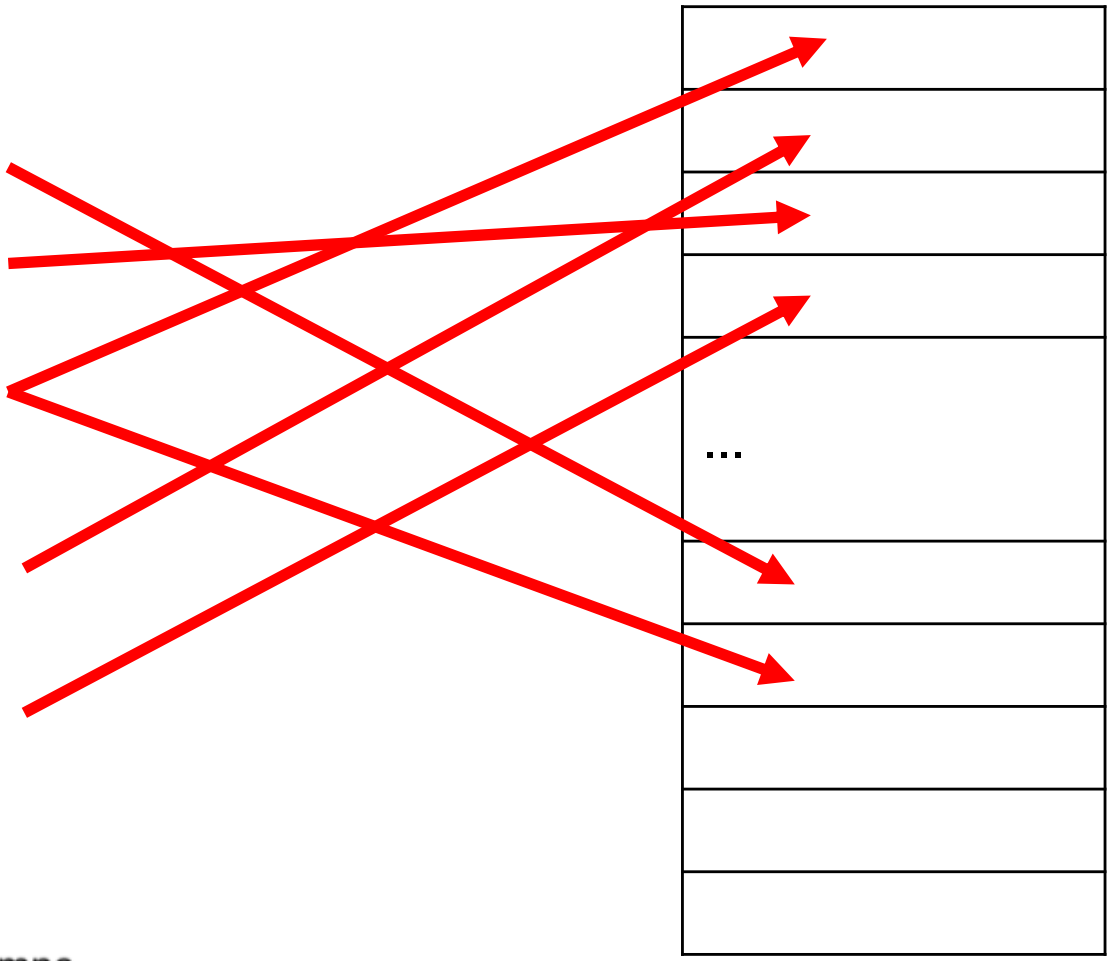
1. Resample from the sample
2. Calculate the statistic (e.g. mean) of interest for each new sample
3. Consider (i.e. plot) the distribution of the statistic
   - Use the quantiles to create a CI on the statistic

Step 1: Resample with Replacement

- Data has 200 values

	value
0	0.614272
1	1.487629
2	0.732561
3	1.937260
4	0.300582
...	...
195	4.381353
196	-0.464248
197	1.706641
198	2.866323
199	2.090060

200 rows × 1 columns



Step 2: Calculate Means

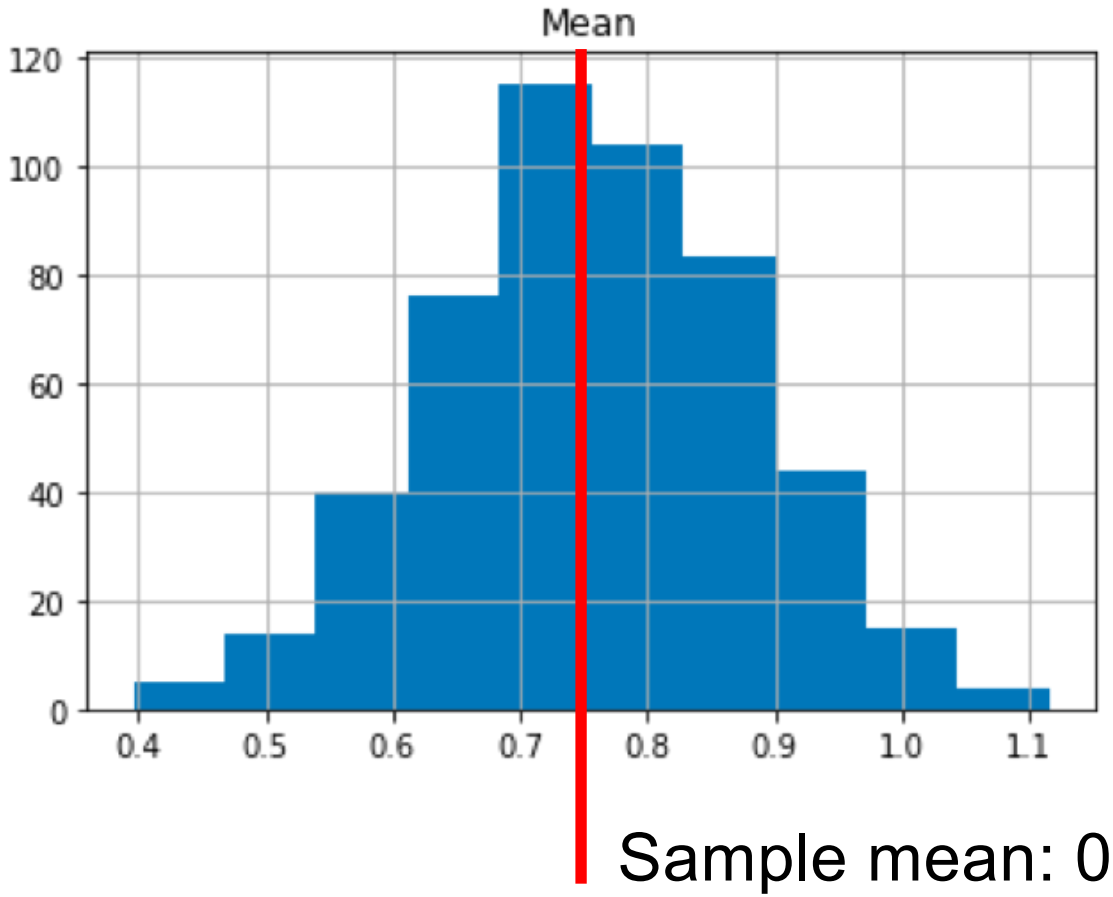
- For each bootstrap sample
- Run 500 bootstraps estimates

	Mean
0	0.719434
1	0.741623
2	0.572063
3	0.753170
4	0.998408
...	...
495	0.712862
496	0.824634
497	0.803627
498	0.641065
499	0.727630

500 rows × 1 columns

Step 3: Distribution

- Confidence Intervals from the distribution



Step 3: Distribution

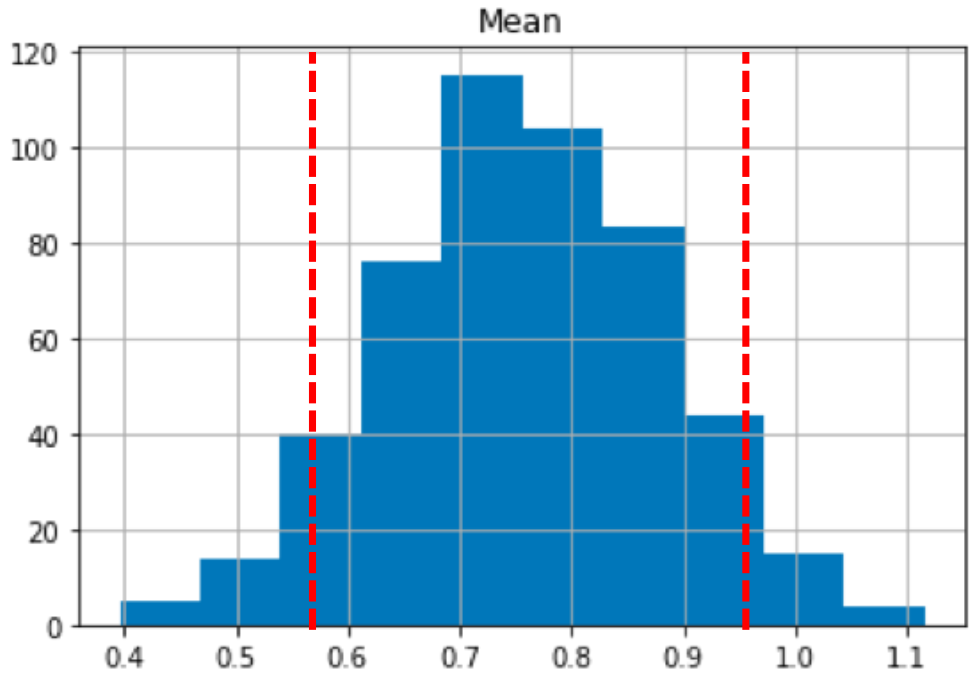
- Confidence Intervals from the

The CI is given by the quantiles of the result data.

```
print('90%% range is %4.3f to %4.3f' %
      (result.Mean.quantile(0.05), result.Mean.quantile(0.95)))
print('95%% range is %4.3f to %4.3f' %
      (result.Mean.quantile(0.025), result.Mean.quantile(0.975)))
```

```
90% range is 0.560 to 0.963
95% range is 0.513 to 0.991
```

Student's t: 90% confidence interval for mean is 0.556 to 0.963



Parametric Boostrap

Concept and Example Application

- Instead of resample data, sample from distribution
  - Use parameters estimated from the data

- Application
  - Chi-square test alternative
  - Use observed parameters
  - How often is data as extreme generated?
- See notebook (in preparation)

Breakout Discussion

Menti Code **3434 4113**

Summary

Advantage

Advantages

- Does not require knowledge of sampling distributions, test statistics
- Can be applied directly to a quantity of interest

Disadvantages

- Not as well recognised

Related Techniques

- Permutation tests
  - Are two datasets from the same population?
- Jackknife
  - Sequentially delete one data value
  - Deterministic
  - May work better for small samples

Summary

- Never mind all that statistical theory
- ... just use a bootstrap