BIOS 635: Bootstrap

Kevin Donovan

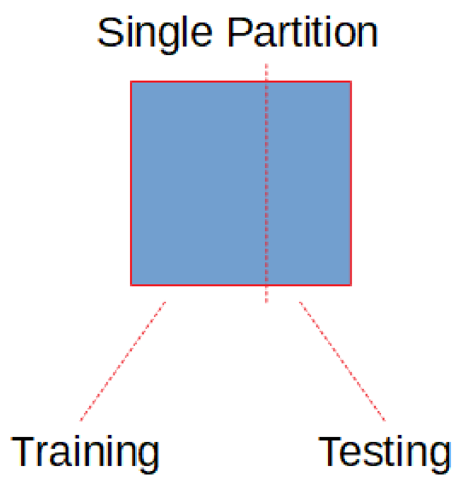
3/2/2021

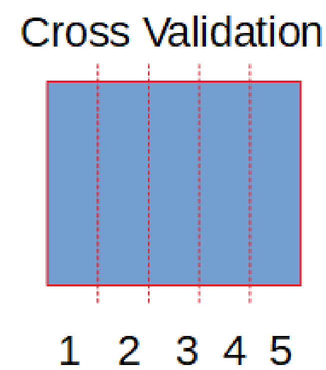
Review

- Homework 5 due on 3/5 at 11PM through GitHub Classroom
- Article Evaluation 2 assigned, due on 3/2 through GitHub Classroom
- Last lecture: cross validation

Data partitioning

- **Recall**: can generate training and testing datasets using
    - *Holdout method*
    - *K-fold cross validation*

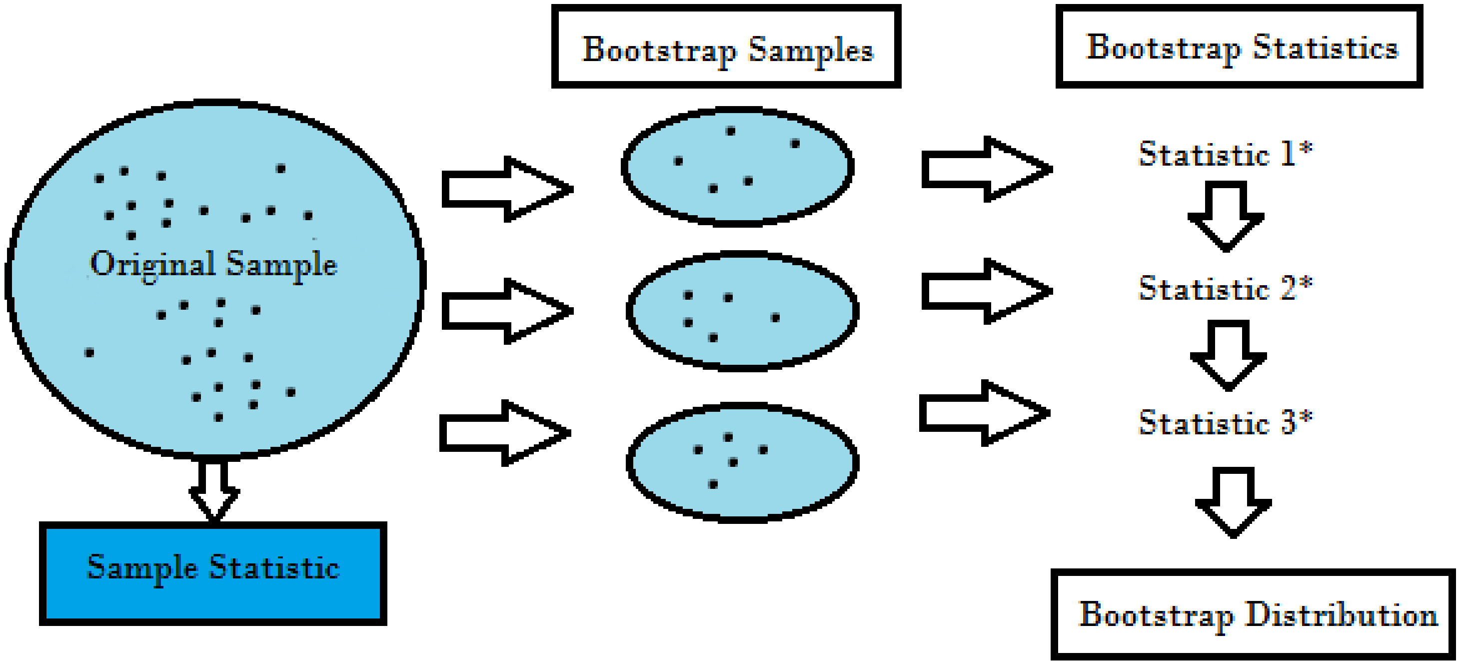




1) Training : 1-4, Testing : 5
2) Training : 1-3+4, Testing : 4
3) Training : 1-2+4-5, Testing : 3
4) Training : 1+3-5, Testing : 2
5) Training : 2-5, Testing : 1

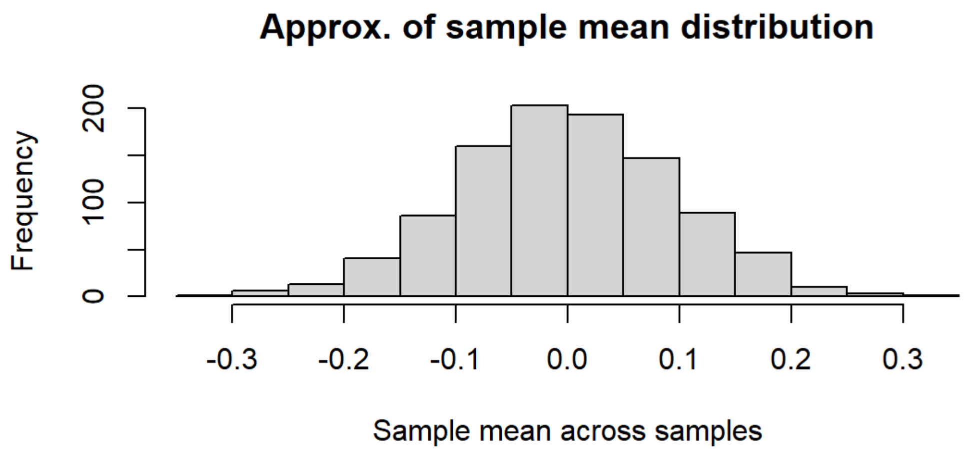
Bootstrap

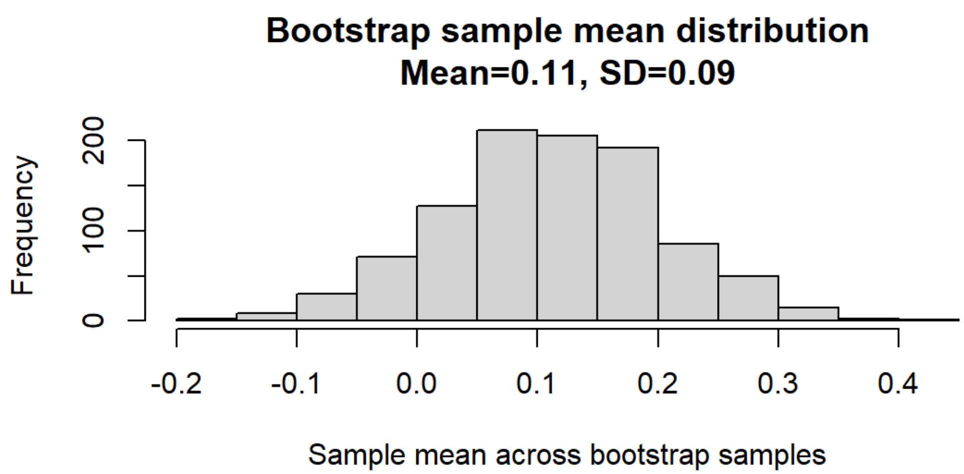
- **Another method**: resampling via *bootstrap*
- **Idea**: generate multiple samples from data by *sampling with replacement* m times
    - *Repeat process* B *times* $\rightarrow$ B *samples of size* m *each created*
    - *Calculate statistic in each* B *samples* $\rightarrow \{\hat{\alpha}_1, \ldots, \hat{\alpha}_B\}$
    - *Use* $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_B\}$ *to assess sample variability of* $\hat{\alpha}$


Bootstrap Samples
Bootstrap Statistics
Original Sample
Statistic 1*
Statistic 2*
Statistic 3*
Sample Statistic
Bootstrap Distribution

Bootstrap

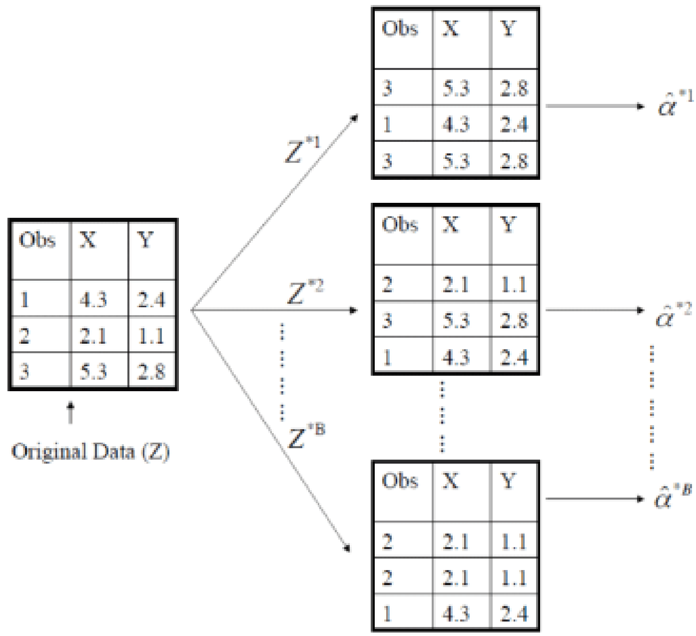
- Simple example: computing sample mean
- Suppose sample of variable X observed: $X_1, \ldots, X_n$ for $n = 100$
- We know sample mean $\bar{X} \sim \mathrm{Normal}(\mu, \sigma^2/n)$ by *Central Limit Theorem*
- Suppose $\mu = 0$ and $\sigma^2 = 1$. Let's look at the distribution of $\bar{X}$ via bootstrap


Approx. of sample mean distribution
Frequency
0
100
200
-0.3
-0.2
-0.1
0.0
0.1
0.2
0.3
Sample mean across samples


Bootstrap sample mean distribution
Mean=0.11, SD=0.09
Frequency
0
100
200
-0.2
-0.1
0.0
0.1
0.2
0.3
0.4
Sample mean across bootstrap samples

Bootstrap

- Can see variance of bootstrap sample means $\approx$ sample mean variance $1/\sqrt{100}$
  - *Recall: also called **standard error** of statistic*
- Can use to create confidence interval or do hypothesis testing as well
- Sampling with replacement $\rightarrow$ row can be included more then once
  - *Idea: mimics independent random sampling*
  - *Ex. three observations, computing statistic $\hat{\alpha}$*


Obs X Y
1 4.3 2.4
2 2.1 1.1
3 5.3 2.8
Original Data (Z)
Z^{*1}
Z^{*2}
Z^{*B}
Obs X Y
3 5.3 2.8
1 4.3 2.4
3 5.3 2.8
Obs X Y
2 2.1 1.1
3 5.3 2.8
1 4.3 2.4
Obs X Y
2 2.1 1.1
2 2.1 1.1
1 4.3 2.4
$\hat{\alpha}^{*1}$
$\hat{\alpha}^{*2}$
$\hat{\alpha}^{*B}$

Bootstrap algorithm

Suppose Z denotes the dataset with n rows (obs) and p columns (variables)

1. Randomly select n obs from Z, creating **bootstrap dataset** Z_1

- Selection done **with replacement**

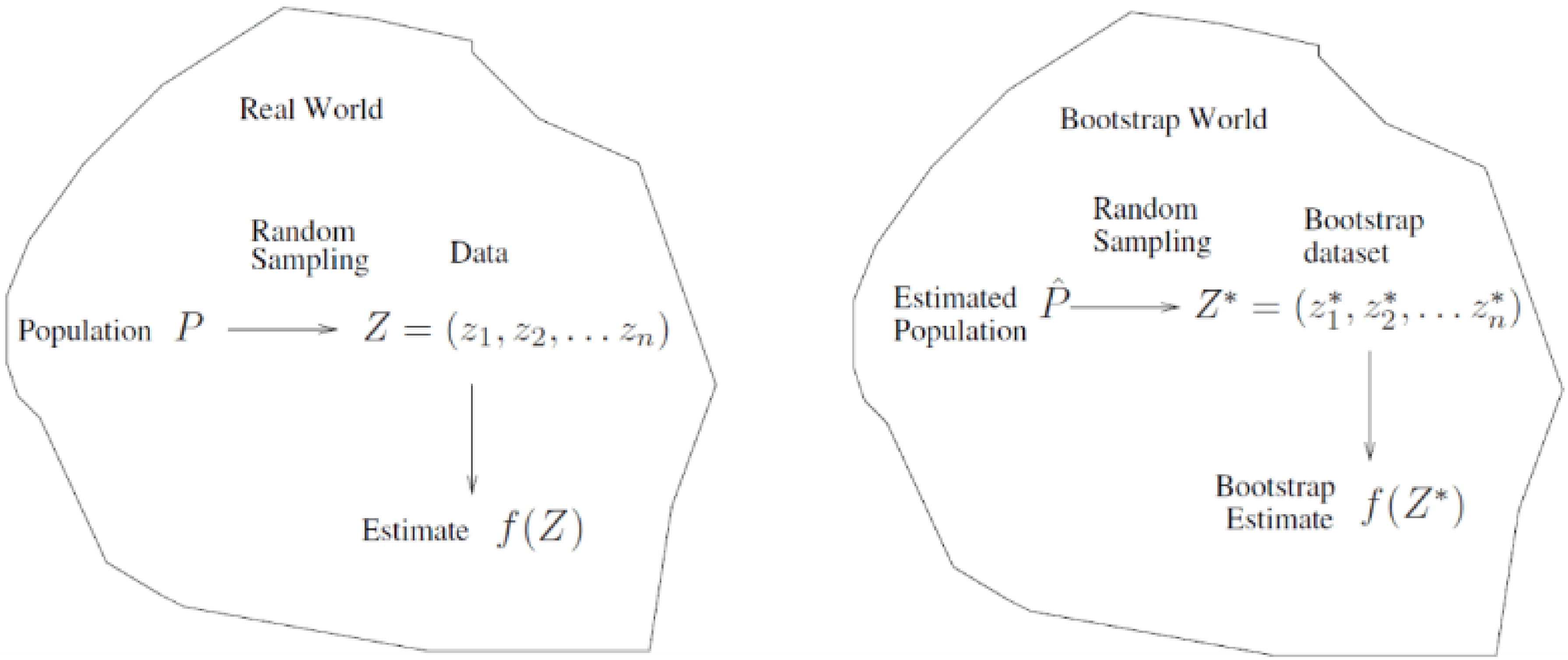
2. Using Z_1 calculate statistic of interest, denoted $\hat{\alpha}_1$
3. Repeat 1 and 2 B times, creating set of estimates: $\{\hat{\alpha}_1, \ldots, \hat{\alpha}_B\}$
4. Can estimate SE of statistic $\hat{\alpha}$ using bootstrap sample SE

$$\hat{SE}_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}(\hat{\alpha}_r - \bar{\hat{\alpha}})^2}$$

where $\bar{\hat{\alpha}} = \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}_r$ denotes bootstrap sample mean

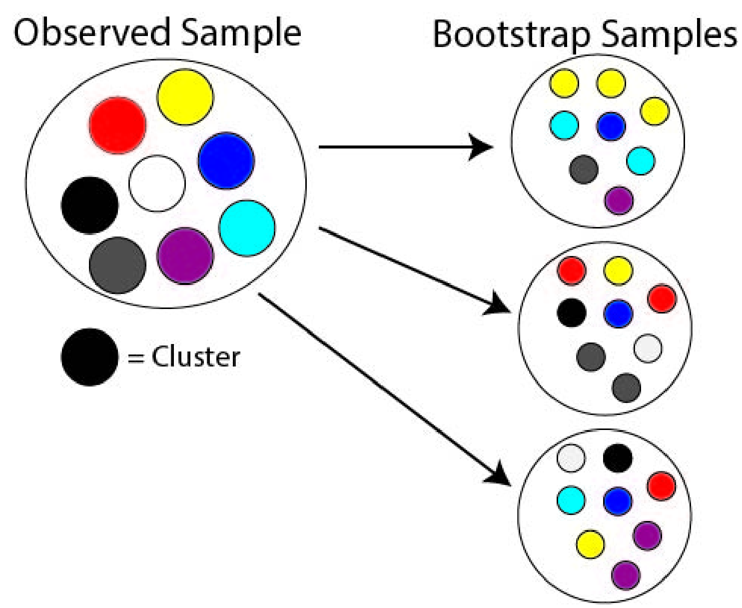
- Can show $\hat{SE}_B(\hat{\alpha}) \approx SE(\hat{\alpha})$

Bootstrap visual



Bootstrap with clustered data

- Suppose some obs in sample are correlated
  - *Denoted as clusters*
- How does this change the bootstrap sampling?



Bootstrap and prediction

- Using bootstrap for data partitioning
  - *Use bootstrap as testing and original as training? (or vice versa)*
  - ***Issue**: Bootstrap as significant overlap with sample* ($\approx \frac{2}{3}$)
  - $\rightarrow$ *bootstrap error estimate **biased downward***
  - *What about for tuning? Sometimes used (`train` in `caret` uses by default)*
- K-fold CV forces separate training and testing sets at each iteration
  - $\rightarrow$ ***always use CV instead***
  - *Possible solution with bootstrap: use **out-of-bag** (OOB) sample*
  - *OOB discussed with random forests later on*