Bootstrap

A Key Technique used in Random Forest

In-class session slides

Intended Learning Outcomes

Identify aspects of business problems that cause standard analytics models to become useless or less effective.

Apply advanced techniques to overcome or mitigate the weaknesses of standard analytics models.

Evaluate performance of the advanced predictive techniques.

Explain the workings and results of the advanced predictive techniques in the context of the business problem to client/employer.

Propose business solutions/recommendations based on the advanced predictive techniques.

Quiz

Ungraded. Check your understanding of this Session Content.
Use your real name (not nickname) in the quiz.

Example: CD4 Data Standard vs Bootstrap Inference

Bootstrap

Example: cd4 data

Subject	Baseline	One year	Subject	Baseline	One year
1	2.12	2.47	11	4.15	4.74
2	4.35	4.61	12	3.56	3.29
3	3.39	5.26	13	3.39	5.55
4	2.51	3.02	14	1.88	2.82
5	4.04	6.36	15	2.56	4.23
6	5.10	5.93	16	2.96	3.23
7	3.77	3.93	17	2.49	2.56
8	3.35	4.09	18	3.03	4.31
9	4.10	4.88	19	2.66	4.37
10	3.35	3.81	20	3.00	2.40

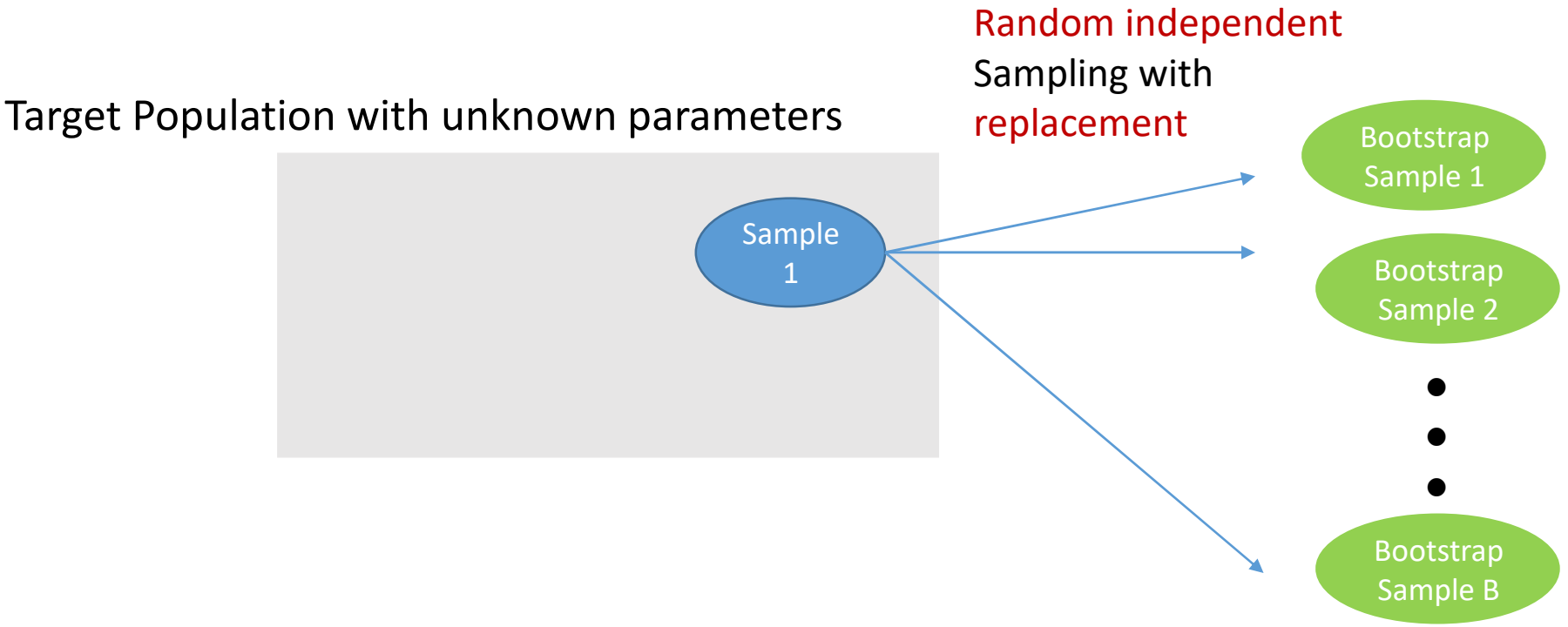
- cd4 counts (in hundreds) of HIV patients before and one year after experimental drug trial.
- Source: DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals (with Discussion). Statistical Science 11: 189-228
- Data available in cd4.csv
- Construct Table1 to compare univariate Standard Statistics vs Bootstrap Statistics for Mean, SD, and Proportion of cd4 count in normal range.

Example: cd4

"The CD4 count is like a snapshot of how well your immune system is functioning. CD4 cells (also known as CD4+ T cells) are white blood cells that fight infection. The more you have, the better. These are the cells that the HIV virus kills. As HIV infection progresses, the number of these cells declines. When the CD4 count drops below 200 due to advanced HIV disease, a person is diagnosed with AIDS. A normal range for CD4 cells is about 500-1,500. Usually, the CD4 cell count increases as the HIV virus is controlled with effective HIV treatment."

Source: https://www.hiv.va.gov/patient/diagnosis/labs-CD4-count.asp

Inference from Sample to Population (with bootstrap)



Each bootstrap sample has the same size (n) as the original sample (sized n).
B is chosen to be a large number e.g. 2000, 10,000, … etc.
Inference about the population based on the B bootstrap samples.

Exercise 1: Bootstrap vs standard statistics in Table 2

Part of Pre-class learning activities.

- Run cd4table1.R

a. Correlation between Baseline and Year1 cd4.

b. Linear Regression with Y = Year1 cd4 count and X = Baseline cd4 count.

Ref: https://www.statmethods.net/advstats/bootstrapping.html

c. Analysis of Difference in medical outcome (D):

   i. Define D = Year 1 cd4 – Baseline cd4.

   ii. Is the Difference statistically significant? What is the business conclusion?

d. Create and save your answers in Table 2.

Class Activity: Table 2 to be completed

	Standard.Statistic	Bootstrap.Statistic
Correlation	NA	NA
CI for Correlation	NA	NA
b0	NA	NA
CI for beta0	NA	NA
b1	NA	NA
CI for beta1	NA	NA
D	NA	NA
CI for D	NA	NA

Answer in cd4 Table 2. Rscript solution: cd4.r

Note: Due to random selection of bootstrap samples, it is fine to have a different but close answer to the Bootstrap Statistic column.

Exercise 2: Bootstrap Reflection

Est. Duration: 30 mins.

Using Microsoft Excel:

1. Create a column of ten case IDs: 1, 2, …10
2. Create 5 bootstrap samples.
3. What are the cases not selected in each of the bootstrap sample?

Reflection:

4. Mathematically, what is the probability that case ID i in the original sample of size n is not in a bootstrap sample?
5. Each bootstrap sample would have some cases from the original sample that is not selected into the bootstrap sample. If you use the bootstrap sample instead of the original sample, would this be a waste of data? Explain.

Answers for Exercise 2

Reminder

Please complete the Pre-Class Learning Activities before next class.

Reflection on your Learning

Go	NTULearn Class Site > Journal
Post	Read the instructions and post entry on this week's learning. • Reply on the 3 questions as stated in the Journal Instructions.