Example: CD4 Data Standard vs Bootstrap Inference

Bootstrap

Example: cd4 data

Dataset: cd4.csv

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

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

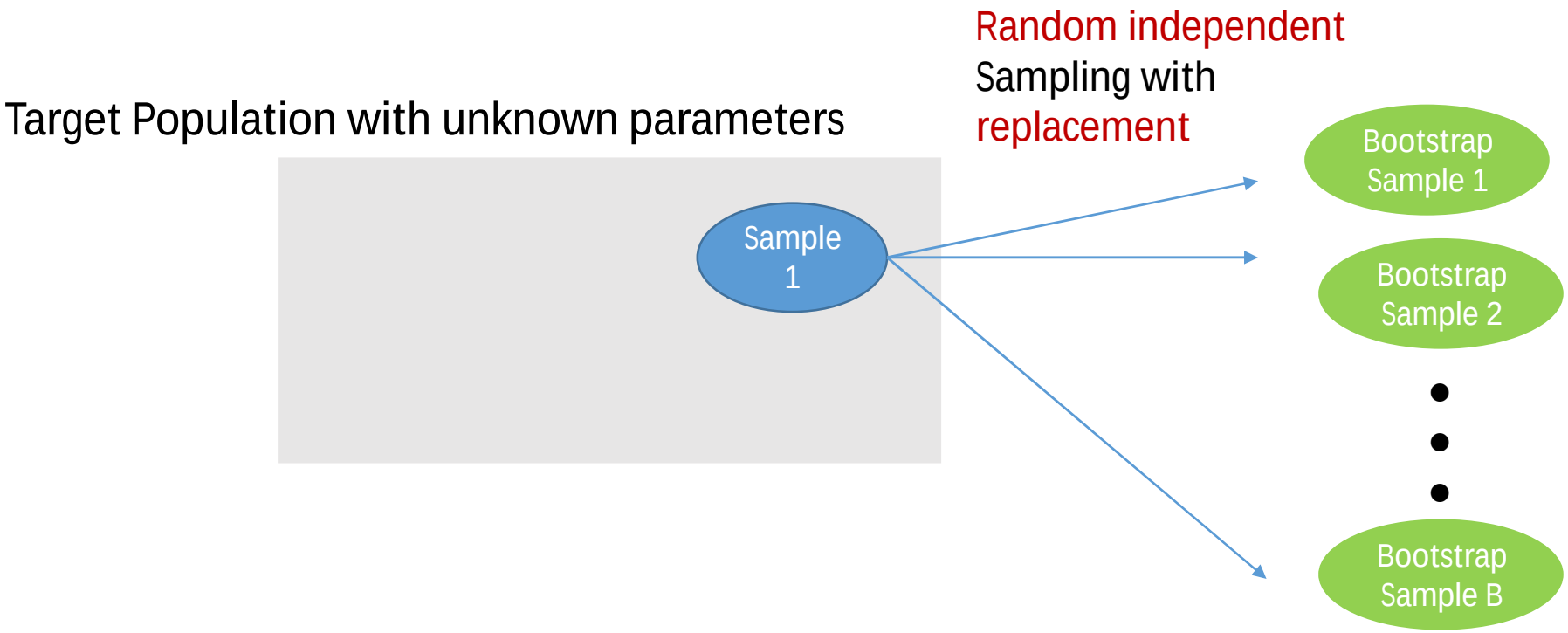
Table 1: Before bootstrap

	Baseline.Standard.Statistic	Baseline.Bootstrap.Statistic	Year1.Standard.Statistic	Year1.Bootstrap.Statistic
Mean	328.8	*NA*	409.3	*NA*
95% CI for Mean	290.86 to 366.74	*NA*	355.01 to 463.59	*NA*
SD	81.06	*NA*	116	*NA*
95% CI for SD	61.64 to 118.39	*NA*	88.21 to 169.42	*NA*
Prop Normal Range	0.05	*NA*	0.2	*NA*
95% CI for Prop N.R.	0.003 to 0.269	*NA*	0.066 to 0.443	*NA*

- Standard Formulas applied and results displayed under "Standard Statistic" column for variables:
  - Baseline Score
  - Year1 Score
- Code available in my RScript cd4table1.R

Source: Chew C.H. (2022) A.I., Analytics & Data Science, Vol. 2.

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.

Source: Chew C.H. (2022) A.I., Analytics & Data Science, Vol. 2.

Bootstrap the Mean Statistic

```
library(boot)
# Bootstrap the mean
samplemean <- function(data, indices) {
  return(mean(data[indices], na.rm = T))
}
boot.Baseline <- boot(data=data1$Baseline, statistic=samplemean, R=10000)
```

- First step is to define the statistic that will be bootstrapped via a function.
- Indices is a vector of numbers that determines the random set of records selected within data.
- The function (that you write) require data and a vector of indices. This function will be called B times, one for each bootstrap replication. Every time, the dataframe will be the same, but the bootstrap sample will be different, depending on the [random] choice of indices.

Q: Function to generate sample mean?

- Why do we need to write a function to generate sample mean, when there is already a standard in-built function mean() in Base R?

PAUSE and REFLECT

Q: Function to generate sample mean?

- Why do we need to write a function to generate sample mean, when there is already a standard in-built function mean() in Base R?

Ans:

- We need an efficient way to generate the mean statistic from 10,000 bootstrap samples.
- Each bootstrap sample is a different data sample.
- We can choose to write 10,000 lines (or a for loop) ; each line just compute the mean of a specific but different bootstrap sample 10,000 times; or
- Write a function (2 lines) that will be run 10,000 times by the boot() function. At each execution, boot() function auto-generates a fresh set of random indices to select a new bootstrap sample.

Results of 10,000 bootstrap of Baseline cd4 mean

```
# view results of bootstrap
boot.Baseline
plot(boot.Baseline)
# 95% BCA confidence interval from Bootstrap of Mean
boot.ci(boot.Baseline, type="bca", conf = 0.95)
```

```
> boot.ci(boot.Baseline, type="bca", conf = 0.95)
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 10000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.Baseline, conf = 0.95, type = "bca")

Intervals :
Level       BCa
95%   (296.3, 365.2 )
Calculations and Intervals on Original Scale
```

- There are several variant of bootstrap C.I.
- Best to use BCA (Bias Corrected and Accelerated).

Run the Rscript: cd4table1.R

- Bootstrap the Mean of Year1 cd4 count.
- Create functions and conduct bootstrap of SD and Proportion.
- Compare the difference between standard statistics and bootstrap statistics.
- Code provided in cd4table1.R

Cd4 Table 1 Results

	Baseline.Standard.Statistic	Baseline.Bootstrap.Statistic	Year1.Standard.Statistic	Year1.Bootstrap.Statistic
Mean	328.8	328.45351	409.3	409.38417
95% CI for Mean	290.86 to 366.74	296.3 to 365.25	355.01 to 463.59	362.5 to 461.37
SD	81.06	78.2252656011766	116	112.279775830505
95% CI for SD	61.64 to 118.39	63.16 to 110.32	88.21 to 169.42	92.29 to 149.39
Prop Normal Range	0.05	0.04977	0.2	0.20084
95% CI for Prop N.R.	0.003 to 0.269	0 to 0.15	0.066 to 0.443	0.05 to 0.35

- Bootstrap point estimates of the parameter are based on the average of the B bootstrap samples.
- Observe that Bootstrap BCA confidence intervals are all better than the standard confidence intervals.

Bootstrap using Python

- Scikit-learn resample()
  - https://machinelearningmastery.com/a-gentle-introduction-to-the-bootstrap-method/
- Bootstrap Confidence Interval via Percentile Method
  - https://machinelearningmastery.com/calculate-bootstrap-confidence-intervals-machine-learning-results-python/
  - Python Library Bootstrapped from Facebook
    - https://pypi.org/project/bootstrapped/

- *Opinion: For bootstrap, R boot package is far more flexible and simpler to use [assuming you know how to write R functions], and includes an advanced bootstrap confidence interval computation - bca.*

Summary

- Bootstrap
  - Random Sampling with Replacement
  - B bootstrap samples. Nowadays, typically set B = 10,000.
- Rpackage boot
  - Require user to write their own function for the statistic(s) to be bootstrap.
  - “Cost” for flexibility - You need to write the function that generates the statistic.
  - Benefit is the ability conduct bootstrap for any statistic to get its distribution for inference purposes.
    - i.e. as long as you can define the function in R, you can bootstrap it!
  - Several types of Bootstrap Confidence Interval are available. Best to use Bias Corrected and Accelerated (bca) version.
- Random Forest used bootstrap to boost stability in CART.