Bootstrap

BC2407 Seminar 7

References

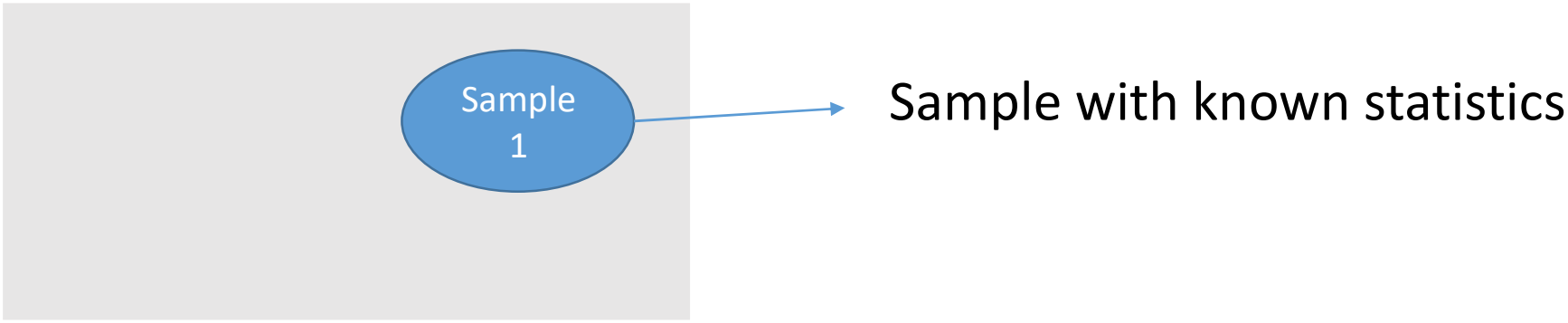
- Gareth James et. Al. (2017) An Introduction to Statistical Learning.
  - Section 5.2 The Bootstrap, pp. 187 – 190.
  - Textbook download: http://faculty.marshall.usc.edu/gareth-james/ISL/
- Chew C.H. (2022) Artificial Intelligence, Analytics and Data Science, Vol. 2.
  - Est. Q3 2022.

Bootstrap as an advanced statistical technique for Machine Learning

- Bootstrap is a general statistical technique, not a model.
- An extremely creative way to use the given data sample.
- Allows one to infer from sample to unknown population without those assumptions.
- The basis for some advanced Machine Learning methods.
  - Essential for Random Forest (next topic).

Inference from Sample to Population (without bootstrap)

Target Population with unknown parameters



Example: Election

- Population: All eligible voters as listed in the voter registry
- Population Parameter: Proportion of all voters who voted for party A. Unknown constant.
- Sample: 100 eligible voters
- Sample Statistic: Proportion of all voters who voted for party A in the sample. Known, but varies from sample to sample.

Inference from Sample to Population (without bootstrap)

Population Parameter	Sample Statistic	Inference via Confidence Interval	Key Inference Assumption
Mean: μ	$\bar{x}$	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$, or $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$	CLT applies (with sufficient sample size), or x has normal dist.
Standard Dev: σ	s	$\sqrt{\frac{(n-1)s^2}{\chi^{2*}_{df,\frac{\alpha}{2}}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^{2*}_{df,1-\frac{\alpha}{2}}}}$	x has normal dist.

For 95% CI: $z^* = 1.96$, $t^* = 2.776$, $\chi^{2*}_{df=n-1=39,0.975} = 23.65$, $\chi^{2*}_{df=39,0.025} = 58.12$

Inference from Sample to Population (without bootstrap)

Population Parameter	Sample Statistic	Inference via Confidence Interval	Key Inference Assumption
Proportion: π	p	$p \pm z^* \sqrt{\frac{p(1-p)}{n}}$	π is a binomial dist. parameter & np ≥ 5, n(1−p) ≥ 5
Correlation: ρ	r	Fisher Transform: $\mathrm{F} = \frac{1}{2}\frac{1+r}{1-r}$ $F_l = F - \frac{z^*}{\sqrt{n-3}}$ $F_u = F + \frac{z^*}{\sqrt{n-3}}$ $\frac{e^{2F_l} - 1}{e^{2F_l} + 1} < \rho < \frac{e^{2F_u} - 1}{e^{2F_u} + 1}$	x_A and x_B have iid Bivariate Normal Dist.

Model-Based Inference

- Linear Regression
  - P-values of betas
  - Confidence Interval of betas
- Logistic Regression
  - P-values of betas
  - Confidence Interval of betas
  - P-values of Odds Ratios
  - Confidence Interval of Odds Ratios
- Inference assumptions typically involve z, t or χ^2
- Other models have inference results too

Non-standard Inference

- What if:
  - You are not sure whether assumptions are valid and do not want to assume.
  - Sample or domain knowledge suggests assumptions are not valid. E.g. not independent, not Normal dist.,…
  - There are no (or you do not know) standard inference formulas e.g. median, 99 percentile, 10% trimmed mean, parameters of newly invented model, etc.
- Then, how do you do inference e.g. confidence interval?
  - Ans: 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

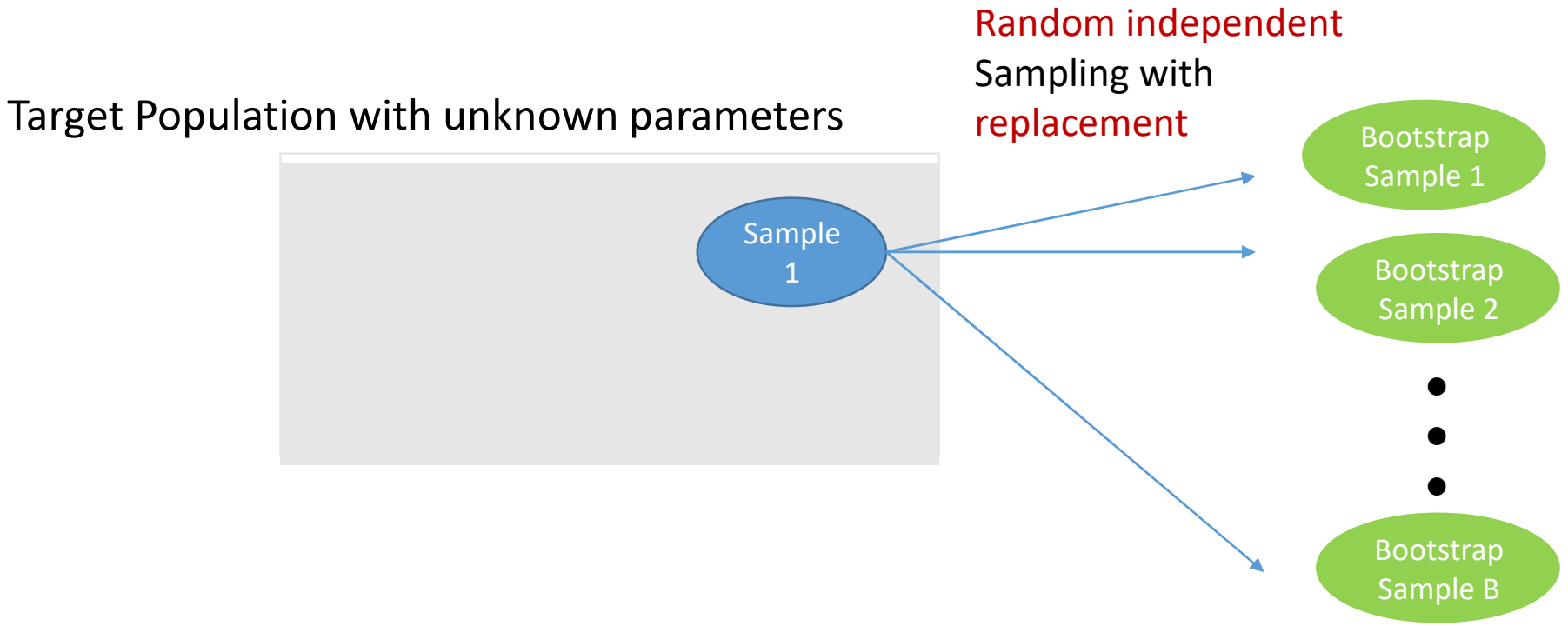
Source: https://www.hiv.va.gov/patient/diagnosis/labs-CD4-count.asp

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

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.

Q&A: What is the probability that case i in the original sample is not in a bootstrap sample?

- This result will be useful in Random Forest [next topic].
- Explain your answer to your classmates at the table (5 mins).
- *Hint: Refer to diagram in previous slide.*

P(case i not in a bootstrap sample) = $(1 - 1/n)^n$

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?

- Ans: We need an efficient way to generate the mean statistic from B bootstrap samples. Each bootstrap sample is a different 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 supplies 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).

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

Demo: cd4table1.R

- Bootstrap the Mean of Year1 cd4 count.
- Write functions and conduct bootstrap of SD and Proportion.
- Compare the difference between standard statistics and bootstrap statistics.
- Code avail 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.

Learning Activity 1: Bootstrap vs standard statistics

Est. Duration: 30 mins

1. Run cd4table1.R
2. Using R, conduct Inference on:
   a. Correlation between Baseline and Year1 cd4.
   b. Linear Regression with Y = Year1 cd4 and X = Baseline cd4.

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

   c. Analysis of Difference in medical outcome (D):
      - D = Year 1 cd4 – Baseline cd4.
      - Is the difference statistically significant?

3. 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

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

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

	Standard.Statistic	Bootstrap.Statistic
Correlation	0.7232	0.7157
CI for Correlation	0.4127 to 0.8831	0.4921 to 0.8604
b0	69.0379	67.989
CI for beta0	-96.4676 to 234.5434	-56.6604 to 185.3545
b1	1.0349	1.0393
CI for beta1	0.5454 to 1.5243	0.7174 to 1.4576
D	80.5	80.3154
CI for D	42.9812 to 118.0188	48.8 to 117.6419

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

Python Bootstrap

- 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”: https://pypi.org/project/bootstrapped/

- *Coordinator Opinion: For bootstrap, R boot package is far more flexible and simpler to use, and includes the most advanced bootstrap confidence interval - bca. But you are free to use any other R or Python package and have different opinion.*

Creating your own functions in R

Optional for those who do not know how to create own functions.

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 1, Appendix B.

Ability to do bootstrap for any statistics you define

- If you can create your own functions in R, your ability to do bootstrap is only limited by your imagination.
- Otherwise, you are limited to only using
  - Functions created by someone else, and
  - That you know of, and
  - Suitable for the problem you are solving.
- Your new function may consists of other functions previously defined by you or others.
- Extremely useful in order to use the full power of boot().

Importance of Functions in R

"To understand computations in R, two slogans are helpful:

Everything that exists is an object.

Everything that happens is a function call."

John Chambers

Creator of the S programming language,

Core member of the R programming language project.

Adjunct Professor of Statistics, Stanford University.

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 1, Appendix B.

Example: mean() function

- At R console, ?mean to view documentation from mean function creator. Shows definition of the function and examples on usage.

Arithmetic Mean

Description

Generic function for the (trimmed) arithmetic mean.

Usage

```
mean(x, ...)

## Default S3 method:
mean(x, trim = 0, na.rm = FALSE, ...)
```

Arguments

`x`	An R object. Currently there are methods for numeric/logical vectors and date, date-time and time interval objects. Complex vectors are allowed for `trim = 0`, only.
`trim`	the fraction (0 to 0.5) of observations to be trimmed from each end of `x` before the mean is computed. Values of trim outside that range are taken as the nearest endpoint.
`na.rm`	a logical value indicating whether NA values should be stripped before the computation proceeds.
`...`	further arguments passed to or from other methods.

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 1, Appendix B.

User Defined Functions

- Anyone can create functions in R.
- This is my sum3() function defined mathematically:

  sum3(x, y, z = 1) = x + 2y + z

- sum3(1, 2) = 1 + 2(2) + 1 = 6
- sum3(2, 1) = 2 + 2(1) + 1 = 5
- sum3(y = 2, x = 1) = 1 + 2(2) + 1 = 6
- sum3(1, 2, −1) = $1 + 2(2) - 1 = 4$
- sum3(1) = error!

Note: X and Y are mandatory arguments, Z is optional with a default value.

Learning Activity 2: Create your R function

Est. Duration: 10 mins

- Create the sum3() function in R.
- *Hint: https://www.statmethods.net/management/userfunctions.html*
- Verify your answers using the numerical examples in previous slide.

Solution: my sum3() function created in R

Console Terminal

D:/Dropbox/Datasets/ADA1/2_Fundamentals/

```
> sum3 <- function(x, y, z = 1) {
      ans = x + 2*y + z
      return(ans)
  }
```

Console Terminal

D:/Dropbox/Datasets/ADA1/2_Fundamentals/

```
> sum3(1, 2)
[1] 6
> sum3(2, 1)
[1] 5
> sum3(y = 2, x = 1)
[1] 6
> sum3(1, 2, -1)
[1] 4
> sum3(1)
Error in sum3(1) : argument "y" is missing, with no default
```

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 1, Appendix B.

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.