Answers to Bootstrap

BC2407 Seminar 7

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 2.

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].
- Think (5 mins) and explain to your classmate next to you (5 mins).
- *Hint: Refer to diagram in previous slide.*

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

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 2.

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.

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 2.

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.

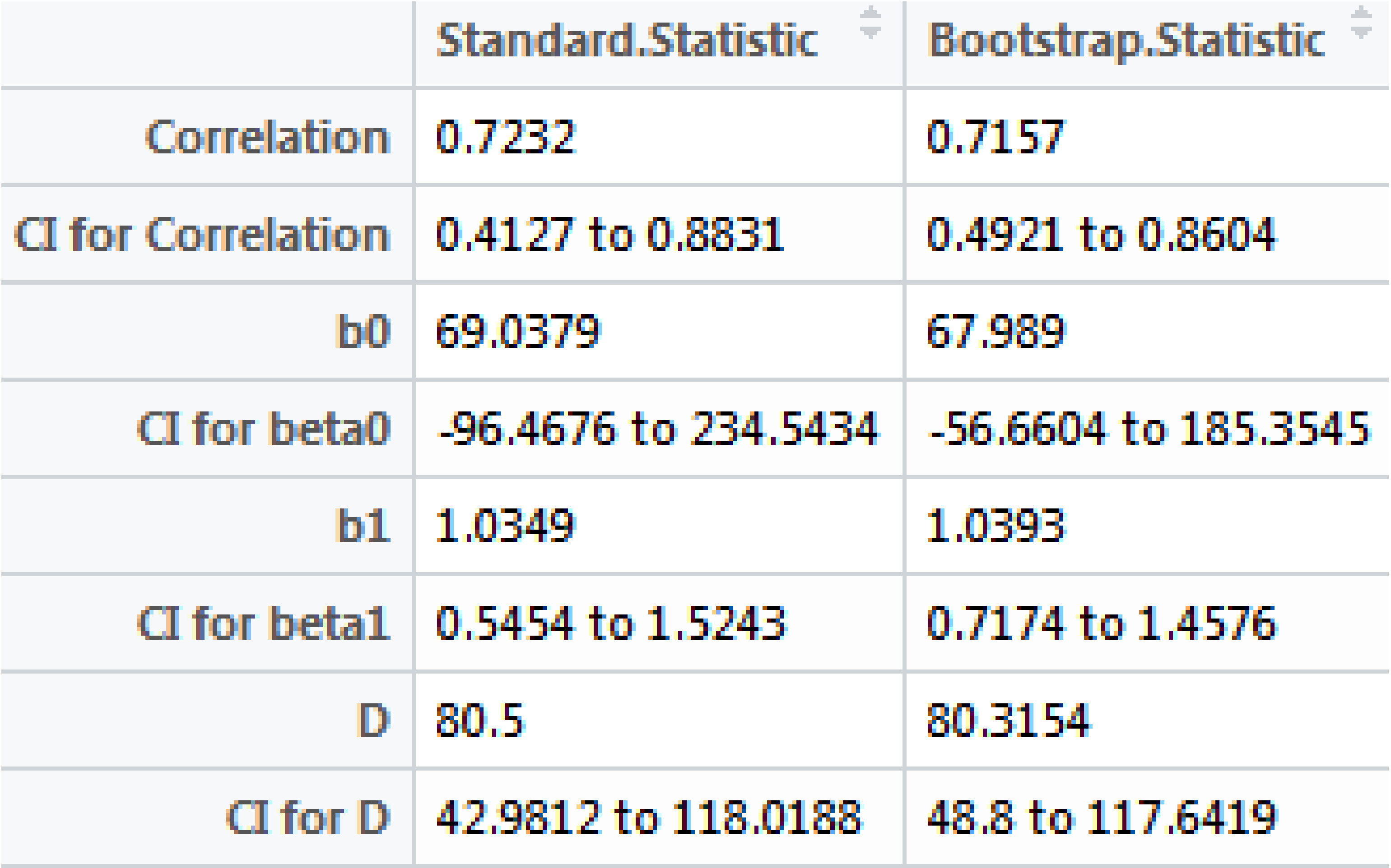
   Hint: https://www.statmethods.net/advstats/bootstrapping.html

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

3. Create and save your answers in Table 2.

Source: Chew C.H. (2020) 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.

Source: Chew C.H. (2020) A.I., Analytics & Data Science, Vol. 2.

Creating your own functions in R

Optional for those who do not know how.

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

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

- 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!

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

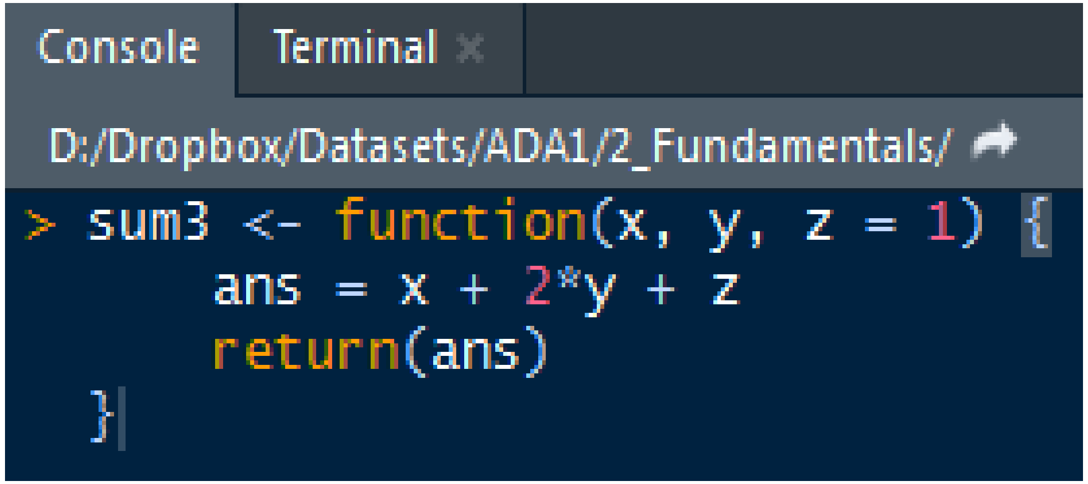
Learning Activity 2: Create your R function

Est. Duration: 10 mins

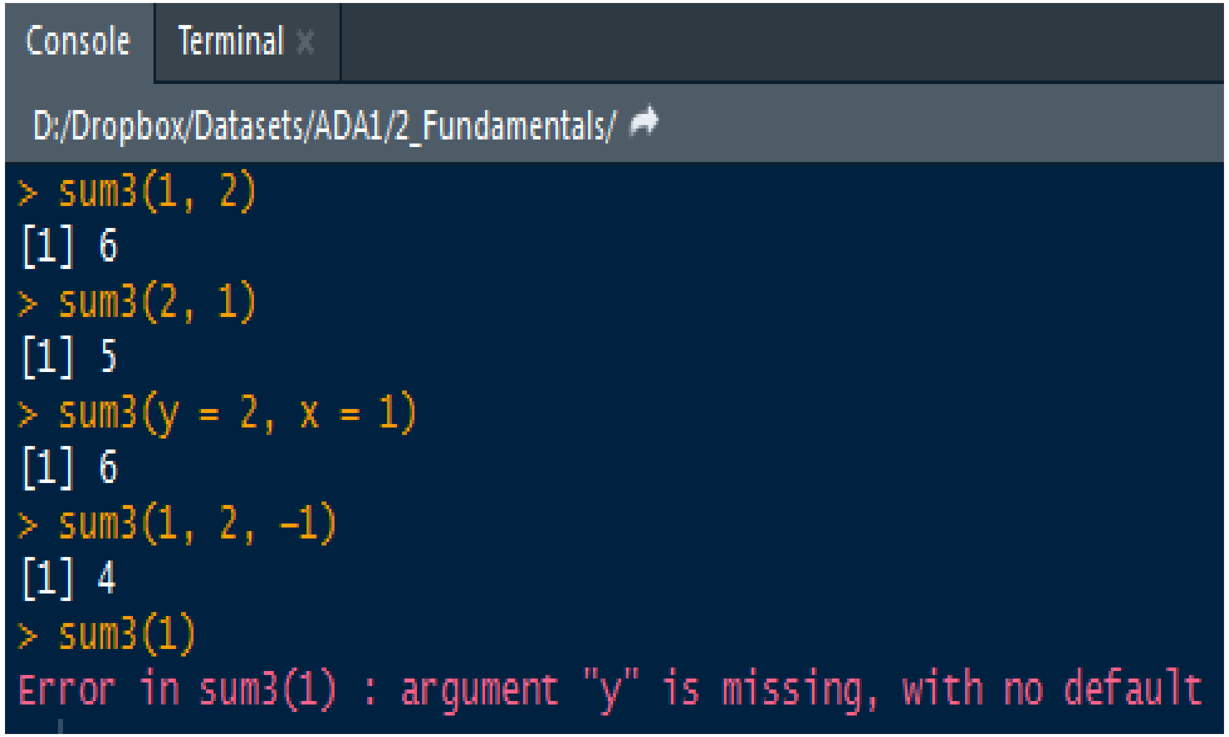
- Create the sum3() function in R.
- *Hint:* *https://www.statmethods.net/management/userfunctions.html*
- Verify your answers using the numerical examples in previous slide.

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

Solution: my sum3() function created in R



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



```
> 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.