STA248 - Tutorial 3 Activity: Bootstrapping v.s. CLT

Due: March 28, 2021 at 10 PM EDT

Submission Instructions

For this tutorial only, you may work with **one other classmate** but it is expected that by the end of the activity, each student will be able to complete the activity independently (i.e. learn the skills covered in this tutorial). If you are working with a partner, you will first create a group and submit one copy as a group on Crowdmark. **Note that once you form a group, it cannot be changed.** Read and follow the instructions below carefully.

- There have been some recent changes to how work is submitted on Crowdmark. Make sure to review the submission process HERE, including image size limitations and how to fix them.
- It is expected and required that all problems requiring the extensive use of R beyond simple calculations be completed in an R markdown file with outputs and text cleanly knit to pdf.
- Questions ((a), (b), . . . , etc.) should be clearly labeled in your text. Make use of headers to create eye-catching titles!
- Use `\newpage` to separate your parts onto different pages for easier upload.
- Your output should include: your code chunks, relevant output values, and written responses using properly displayed LaTeX notation. For problems that involve large simulations, please do not print out the full vector/data frame. Instead, use glimpse() or head() to display the first few rows for your own verification purposes and also for the teaching team to see your results.

Introduction

One core objective in statistics is to accurately understand and infer *uncertainty* of the estimated model. When scientists try to measure the average effect size of a new drug (measure of difference in efficacy between treatments), they would like to know the range of plausible values for this unknown effect size. This is often described using a **confidence interval**. Often times statisticians rely heavily on the asymptotic normality of the sampling distribution known as the **Central Limit Theorem** (CLT) to measure this uncertainty, however, situations arise when a naive implementation of the CLT can be problematic.

RECALL

The assumptions in place for CLT include that we have a random sample of random variables $X_1, ..., X_n$ from the same distribution (identically distributed). CLT tells us that ***as the sample size grows to infinity****, that*

$$\bar{X}_n \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right)$$

That is, as the sample size grows, the distribution of the sample mean converges to a normal distribution.

The convergence rate of the test statistic might be too slow; there might be heterogeneous, non-IID clusters of observations in the dataset; or the asymptotic distribution of the test statistic is simply difficult to derive theoretically. In these cases, **bootstrapping** and the family of resampling based methods serve as viable alternatives.

In this tutorial, we will go through implementing bootstrapping step by step and compare the use of bootstrapping against CLT when constructing confidence intervals.

Part A: Bootstrapping v. Generating New Data (6 points)

Note: This is intended as a preparation for Part B. Make sure you understand what the R code does and complete the preparation work. Answer the questions at the end of Part A before you move onto Part B!

The most important thing to know is that **bootstrapping is not the same as generating a new dataset!** Let's start with bootstrapping first.

Bootstrapping

You are a physician working on a cure to the mysterious disease X. Luckily, you have found a new drug that can potentially help your patients! Now you're interested in knowing the **median** treatment effect of this potential drug. You randomly sampled 200 patients and measure the patients' responses to the drug. Since it is more challenging to derive the asymptotic distribution of a median than the mean, we can use bootstrapping to help us build a confidence interval!

First, load in the dataset containing drug effects for the 200 patients.

```
library(tidyverse)
effect_size <- read_csv('effect_a.csv')
```

The data frame effect_size contains only one column utility which measures the drug response for a patient in standard units. Positive utilities indicate improved health while negative utilities indicate otherwise.

Now, suppose you would like to construct a bootstrapped sample from this dataset. One can simply *sample with replacement* from the observations we have.

```
boot_sample <- sample(effect_size$Utility, size=200, replace=TRUE)
```

The sample function in R will now sample 200 observations with replacement. Note that each observation has an equal probability of being selected in this bootstrapped sample.

Let's now find the median treatment effect (utility) for this bootstrapped sample. **Is it similar to the median in the original dataset?**

```
median(boot_sample)               # Median of the bootstrapped sample
median(effect_size$Utility)       # Median of the original dataset
```

Clearly, one bootstrapped sample is not enough to construct a confidence interval. Let's now build 1000 bootstrapped samples instead. Remember, each bootstrapped sample contains 200 observations - the same number of observations of your original dataset.

```
B = 1000      ## 1000 bootstrapped samples
n = 200       ## 200 observations per bootstrapped sample

boot_sample <- matrix(sample(effect_size$Utility, size = B * n, replace = TRUE), B, n)
```

The matrix boot_sample contains 1000 rows (each represents a bootstrapped sample) and 200 columns (each represents an observation in a particular bootstrapped sample).

We will be using the *centred medians* to construct a CI for the median. That is, we would compute the bootstrapped deviations (boot.median-data.median) and construct the CI using the percentiles of bootstrapped deviations.

To construct such a confidence interval, we need to find the median for each bootstrapped sample and subtract it from the sample median to construct the bootstrapped deviations. The 97.5% percentile and the 2.5% percentile of the 1000 bootstrapped deviations represent the upper and lower bound of the *deviation from the sampled median*.

```
sample_median <- median(effect_size$Utility)
boot_median <- apply(boot_sample, 1, median)
boot_dev <- boot_median - sample_median
boot_ci <- -quantile(boot_dev, c(0.025, 0.975)) + sample_median
```

The true median we used to simulate the data is close to 1.84. Note that we only need **one** dataset to construct the bootstrapped confidence interval! Every single bootstrapped sample is derived from the original dataset at our disposal.

Generating A New Dataset

Imagine that you know the *true* probability distribution of the treatment effect. One can directly simulate data from this true probability distribution and construct confidence intervals without bootstrapping at all! (Of course, we almost never know what the true probability distribution is in real life.)

The true distribution belongs to a classic distribution called the Mixture of Gaussians. The drug is likely to benefit eighty percent of the patients ($Utility \sim N(2, 0.5)$) while being more likely to be harmful to the other 20 percent ($Utility \sim N(-1, 0.5)$).

To be precise,

$$\begin{cases} X_i \sim N(2, 0.5), & \text{if } U(0,1) < 0.8 \\ X_i \sim N(-1, 0.5), & \text{if } U(0,1) \geq 0.8 \end{cases}$$

Here is the code to generate one new dataset from the true probability distribution

```
#The number of samples from the true (mixture) distribution
N <- 200

#Sample N datapoints from a Uniform Distribution
U <- runif(N)

#Store the samples from the mixture distribution
samples <- array(NA, dim=N)

#Sampling from the mixture
for(i in 1:N){
    if(U[i] < 0.8){
        samples[i] = rnorm(1,2,0.5)
    }else{
        samples[i] = rnorm(1,-1,0.5)
    }
}
```

One can easily calculate sample medians from the generated datasets. This exercise is left for those who are interested in further comparisons.

Part A Questions (6 points)

a) (1 point) Report the confidence interval for the median you found via bootstrapping.
b) (3 points) Is the bootstrapped confidence interval symmetric around the median estimated from your original sample? Is this surprising? Why or why not?
c) (1 point) Does your bootstrapped confidence interval include the true population median? (Hint: Don't worry too much if it doesn't. Remember that a correctly derived 95% confidence interval will fail 5% of the time.)
d) (1 point) How do your data and resulting median from bootstrapping differ from the those generated from the Mixture of Gaussians? (This is not a trick question!)

Part B: Bootstrapped Confidence Intervals (16 points)

You are a physician working on a cure to another mysterious disease Y, however, you have worse luck since only 30 patients are available to you! Now you're interested in knowing the **mean** treatment effect of this potential drug.

First, load in the dataset containing drug effects for the 30 patients.

```
library(tidyverse)
effect_size <- read_csv('effect_b.csv')
```

The data frame `effect_size` contains only one column `utility` which measures the drug response/treatment effect for a patient in standard units. Positive utilities indicate improved health while negative utilities indicate otherwise.

Question 1 (4 points)

Construct the 95% bootstrapped confidence interval for the average treatment effect (utility) using 1500 bootstrapped samples.

Question 2 (2 points)

Calculate the 95% CLT-based confidence interval for the average treatment effect (utility) using only the original dataset.

Question 3 (6 points)

To investigate the reliability of bootstrapped CIs compared with those constructed with CLT, we will pretend that we are able to draw new samples from the population.

Generate a *new dataset of 30 observations* from this population. The true underlying distribution is the gamma distribution with shape=4 and rate=2 (Expected value: $\frac{4}{2} = 2$). For each new dataset, repeat questions 1 and 2 and save your results.

Do this 999 times each so you have in total (including the initial dataset) 1000 bootstrapped CIs and 1000 CLT-based CIs for the **mean** treatment effect (utility).

Determine the proportion of bootstrapped confidence intervals and CLT-based confidence intervals that contain the true population mean, μ.

Hint: You can do this on your own from scratch, or fill in the missing code blocks in the given code below.

```
boot_ci <- matrix(NA, nrow=1000, ncol=2)
clt_ci <- matrix(NA, nrow=1000, ncol=2)

B =                 ## Define the number of bootstrapped samples needed to
                    ##construct one bootstrapped CI
n =                 ## Define the number of samples needed for each
                    ##bootstrapped sample
```

```
for(i in 1:1000){
  effect <- rgamma(n=100, shape=4, rate=2)

  ## Build a bootstrapped sample and calculate its mean
  ## Code goes in here



  #########################

  ## Construct bootstrapped CI and CLT-based CI for each iteration
  boot_ci[i, ] <-
  clt_ci[i,] <-

  #########################

  ## Determine the proportion of CIs that cover the true population mean, mu=2.
  print(sum(apply(boot_ci, 1, function(x){x[1] <2 & x[2] > 2})))
  print(sum(apply(clt_ci, 1, function(x){x[1] <2 & x[2] > 2})))

}
```

Question 4 (4 points)

Are there any differences in terms of the proportion of CIs that cover the true population mean, μ? Some people think bootstrapping can be a great way to construct confidence intervals for the mean when the sample size is too small for CLT to be considered valid. What do you think? Support your answer by drawing from your findings in the activity along with any theory we have covered. When else would bootstrapped CIs be desirable over CLT?