DATA252/DATA551: Modeling and Simulation

Lecture 7: Bootstrapping[1]

March 23, 2020

[1] Materials based on *The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application*, Lauer and Grantz, et. al., published at Annals.org on March 10, 2020. Data and code available at https://github.com/HopkinsIDD/ncov_incubation.

General idea

- Bootstrapping is a method that relies on *resampling* of the observed data

Data : $\{1, 1, 2, 3\}$

Bootstrap sample : $\{2, 2, 3, 1\}$

Another bootstrap sample : $\{1, 1, 1, 1\}$

bootstrap ?

" pull yourself up by the bootstrap "

- Often computationally intensive
- Developed by Bradley Efron in 1979

Motivation

Incubation period of 10 patients:

```
##  [1]  9.1  9.5  4.0  0.3  2.5  5.0 11.0  4.0  1.5  2.0
```

A **statistic** is a function (i.e., a numeric summary) of data:

mean incubation period = 4.89

A statistic is not the truth. . . we need to perform **statistical inference**

(true average inc. period ≠ 4.89)

examples : CI: true average incubation period is 4.89 ± 2 days

hypothesis testing: is the true average incubation period < 5 days?

Motivation

sample mean (4.89)

To make inference on the average incubation period based on $\bar{X}$, we need to understand the **randomness of $\bar{X}$** (i.e., the "sampling distribution" of a statistic). Imagine two scenarios:

1. Not much variation in $\bar{X}$: 4.89, 4.82, 4.87, 4.51

   another 10 patients

   CI: $\bar{X}$ ± small error (4.89 ± 0.5 days)

   Hypothesis testing: true mean < 5 days? Yes

2. A LOT of variation in $\bar{X}$: 4.89, 7.5, 1.2, 5.5 ...

   another 10 patients

   CI: $\bar{X}$ ± large error (4.89 ± 5 days)

   Hypothesis testing: true mean < 5 days? No.

Motivation

How do we know the sampling distribution of $\bar{X}$?

Variation in $\bar{X}$:

mean of $\bar{X}$: true average incubation period (μ)

SD of $\bar{X}$: $\frac{\sigma}{\sqrt{n}}$ (σ is the population SD, n is the sample size)

estimated by $\frac{s}{\sqrt{n}}$ (s is the sample SD)

Shape of distr. of $\bar{X}$: normal (CLT)

However:

1. sample size $n=10$ might be too small
2. What if we don't use $\bar{X}$, but use some other statistic, like median?

Solution:

Simulation via bootstrapping!

↳ enable you to study the sampling distribution of a statistic w/out relying on CLT.

Bootstrap samples

To get a bootstrap sample: resample the original data with replacement; sample size stays the same.

```
##  [1]  9.1  9.5  4.0  0.3  2.5  5.0 11.0  4.0  1.5  2.0
```

(n=10)

#1 bootstrap sample: 0.3, 2.5, 11.0, 4.0, 4.0, 0.3, 0.3, 11.0, 9.1, 9.5 → n=10

↳ mean $\bar{X}_1 = 4.0$ 5.2

#2 bootstrap sample: 9.5, 2.5, 2.5, 5.0, 9.1, 9.5, 11.0, 4.0, 1.2, 2.0 → n=10

↳ mean $\bar{X}_2 = 5.51$

⋮

repeat for a large # of times (say 1000 times): get $\underbrace{\bar{X}_1, \bar{X}_2, \cdots, \bar{X}_{1000}}_{\text{SD, histogram, 95\% CI}}$

Bootstrapping with R

Original data:

```
mydata=round(ncov_simple[1:10],1)
mydata
```

```
##  [1]  9.1  9.5  4.0  0.3  2.5  5.0 11.0  4.0  1.5  2.0
```

To get one bootstrap sample:

```
sample(mydata,replace=T)
```

```
##  [1]  1.5  0.3  9.1  9.1  9.5  1.5  4.0  9.5  9.1 11.0
```

Bootstrapping with R

Repeat via a for loop

```
bootsize=30
bootresult=rep(NA,bootsize)
for(i in 1:bootsize){
  bootresult[i]=mean(sample(mydata,replace=T))
}
bootresult
```

— repeating for 30 times

one bootstrap sample

vector of 30

```
##  [1] 2.79 4.59 4.93 2.28 6.22 5.51 4.45 5.98 4.11 4.25 6
## [15] 4.75 7.11 5.32 4.61 6.51 5.34 5.22 3.76 4.52 4.51 5
## [29] 5.76 3.66
```

$\bar{X}_1$ $\bar{X}_2$ $\bar{X}_3$

Exercise: Use the ncov_simple dataset (incubation periods of 50 patients) to obtain $B = 1000$ bootstrap samples; record the mean of each bootstrap sample

Inference using the bootstrap samples

vector of length 1000

```
bootresult[1:30]
```

```
##  [1] 4.2700 4.2352 4.6716 3.8338 4.2082 4.1408 4.0656 4.
## [11] 3.9812 4.0412 4.3816 4.0514 4.2444 3.8868 4.0136 4.
## [21] 3.9426 3.9026 4.4288 3.9806 4.6898 3.5940 3.5314 3.
```

Using the 1000 bootstrap samples, we can estimate the SD of $\bar{X}$:

```
sd(bootresult)
```

```
## [1] 0.3192128
```

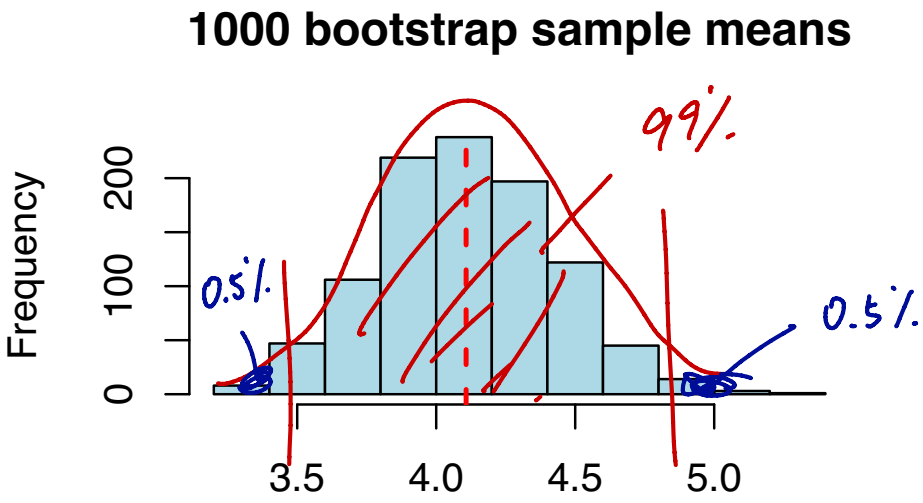
In comparison, without using bootstrap, we can estimate the SD of $\bar{X}$ as usual: true SD $= \sigma/\sqrt{n}$, estimate $= s/\sqrt{n}$

```
sd(ncov_simple)/sqrt(50)
```

s $\sqrt{n}$

```
## [1] 0.3309482
```

Inference using the bootstrap samples



A 95% confidence interval of the mean incubation period

```
quantile(bootresult,probs=c(0.025,0.975))
```

1000 numbers

```
##     2.5%    97.5%
## 3.50512 4.73448
```

Try: obtain a 99% confidence interval:

quantile(bootresult, probs = c(0.005 , 0.995))

Exercise: median

The **median** incubation period is often more of interest (why?) Compared to the mean, properties of the median are harder to calculate mathematically. Instead, we can use bootstrap to study these properties.

Exercise: use the ncov_simple dataset

1. Get the sample median of the dataset. median( )
2. Obtain $B = 5000$ bootstrap samples; record the median of each bootstrap sample.
3. Estimate the SD of the median. Does it have more variation or less variation compared to the mean? 0.39 median : more variation
4. Obtain a 95% confidence interval for the median incubation period. (3, 4.5) median = 4

Limitations

The basic bootstrap method **does not work well** in the following situations:

(4, 3, 4, 5, 7)

- when the original data are not independent

  patient 1 and patient 2 are from the same household

  bootstrap samples ignore the underlying correlation.

- when the original data are not identically distributed

incubation period ~ age

- inc. period 4 3 4 5 7
- age 48 55 32 70 51

- when studying the properties of maximum and minimum

(4, 3, 4, 5, 7) → max = 7

bootstrap samples can never give a max > 7

- when the underlying true distribution has a heavy tail

(4, 3, 4, 5, 23)

outlier

bootstrap tends to over sample the outlier.