Simulation Studies

BIOS 6611

CU Anschutz

Week 1

1. What is simulation?
2. Why simulate?
3. Example
4. Reproducibility
5. Distribution of the sample vs population

What is simulation?

What is simulation?

- A **simulation** mimics the data generation process
- We "generate" samples from a known distribution
  - **Distribution**: a mathematical depiction of the shape describing the chances a variable realizes certain values (e.g., the normal distribution)
- Since we sample from a known distribution of our choosing, we then know the "truth" for the population.
- Within a single simulation, we will generate a sample size of n observations from our chosen distribution.
- We then repeat the process of our simulation many times to estimate our chosen model, estimator, algorithm, etc.

Why simulate?

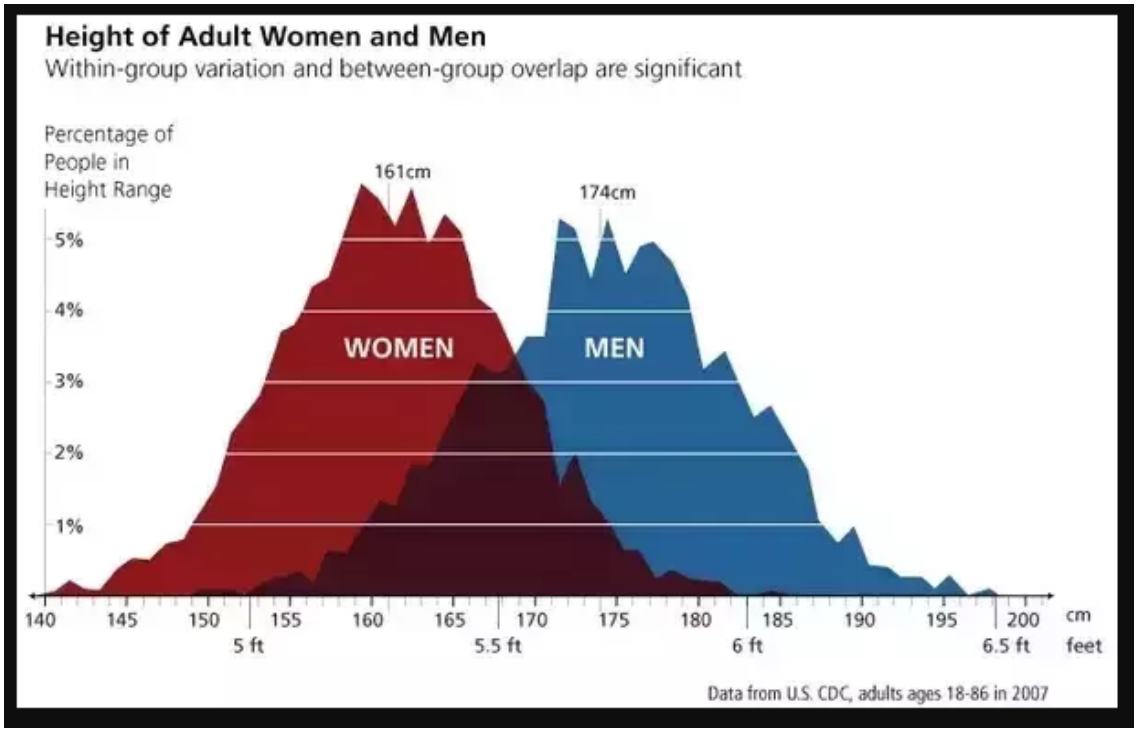
Why simulate?

- Much of statistical theory is based on *asymptotic* properties (i.e., as $n \to \infty$ what would we expect) and are based on assumptions (e.g., normality).
- We usually have finite samples and assumptions may be violated.
- If we know the "truth", then we can evaluate how well different statistical methods work if we were to use them in the given context.
- For example, could assess multiple estimators using criteria discussed in "Properties of Estimators" lecture (e.g., means, medians, and modes to estimate central tendency) under different scenarios (e.g., varying sample sizes, simulating from different distributions, etc.).

Example

Simulation Example with the Normal Distribution

- We think the height of all women in America follows a normal distribution, with an average height of 161 cm and a standard deviation of 5 cm.
  - Note: to simulate, you need to know or assume the distribution you want to sample from and its parameters.



Simulating in R

Let's simulate one sample of size $n = 6$ from this distribution using `rnorm` in R:

```
# For help documentation, type ?rnorm and hit

# Simulate data
sample <- rnorm(n = 6, mean = 161, sd = 5)
sample
```

```
## [1] 157.5625 158.6203 167.2073 167.3912 159.7194 168.7052
```

However, we still need to add one *very* important component...

Reproducibility

Setting the seed

- But wait! Let's say, 3 months from now, we want to replicate this exact same sample
- To do this, need to set the **seed**
- The seed is the initial value used in the random number generation. Different seed = different sample.
- If don't set the seed, R generates its own using current time
- For *reproducibility*, it is important to set the seed so others (including future you) can replicate results

Back to R code example

- Let's generate 10 simulations with sample size $n = 6$.
- Further, save the mean of each sample in a vector called `thetas`.
- Note: set the seed OUTSIDE the for loop. If done inside, we get same set of numbers in every simulation (bad).

```
# set seed
set.seed(812)

# number of simulations
reps <- 10

# initialize vector of thetas (NA = missing)
thetas <- rep(NA, reps)
thetas
```

```
##  [1] NA NA NA NA NA NA NA NA NA NA
```

```r
# Loop through simulations

for (i in 1:reps){
  sample <- rnorm(n = 6, mean = 161, sd = 5)
  sample_mean <- mean(sample)
  thetas[i] <- sample_mean
}

thetas
```

```
## [1] 159.6370 160.4007 159.2373 158.4635 157.1915
```

```
## [1] 163.1119 162.8384 158.1156 162.8484 164.1790
```

Note: there may be computationally faster ways to do this (e.g., the `apply` functions). But, `for` loops may be more intuitive and useful for conceptual understanding.

Distribution of the sample vs population

Distribution of sample vs population

- None of the 10 samples had the sample mean exactly equal the population mean
- Samples usually look different from the population they are being sampled from, unless the sample covers almost the entire population
- Each (random) sample will be different from any other (random) sample

Sampling Distributions

The collection of sample means (`thetas`) has its own distribution, known as the **sampling distribution** of the statistic we estimate (the sample mean in our case).

The sampling distribution of any statistic relies on the underlying population distribution, the statistic being computed, and the sample size.

For example, consider a normal population with mean μ and variance σ^2. Assume we repeatedly take samples of size n from this population and calculate the sample mean, $\bar{x}$, for each sample. We can show that the sampling distribution of the sample mean is

$$\bar{x} \sim N(\mu, \sigma^2/n) \tag{1}$$

Final Example: Exponential Distribution Mean

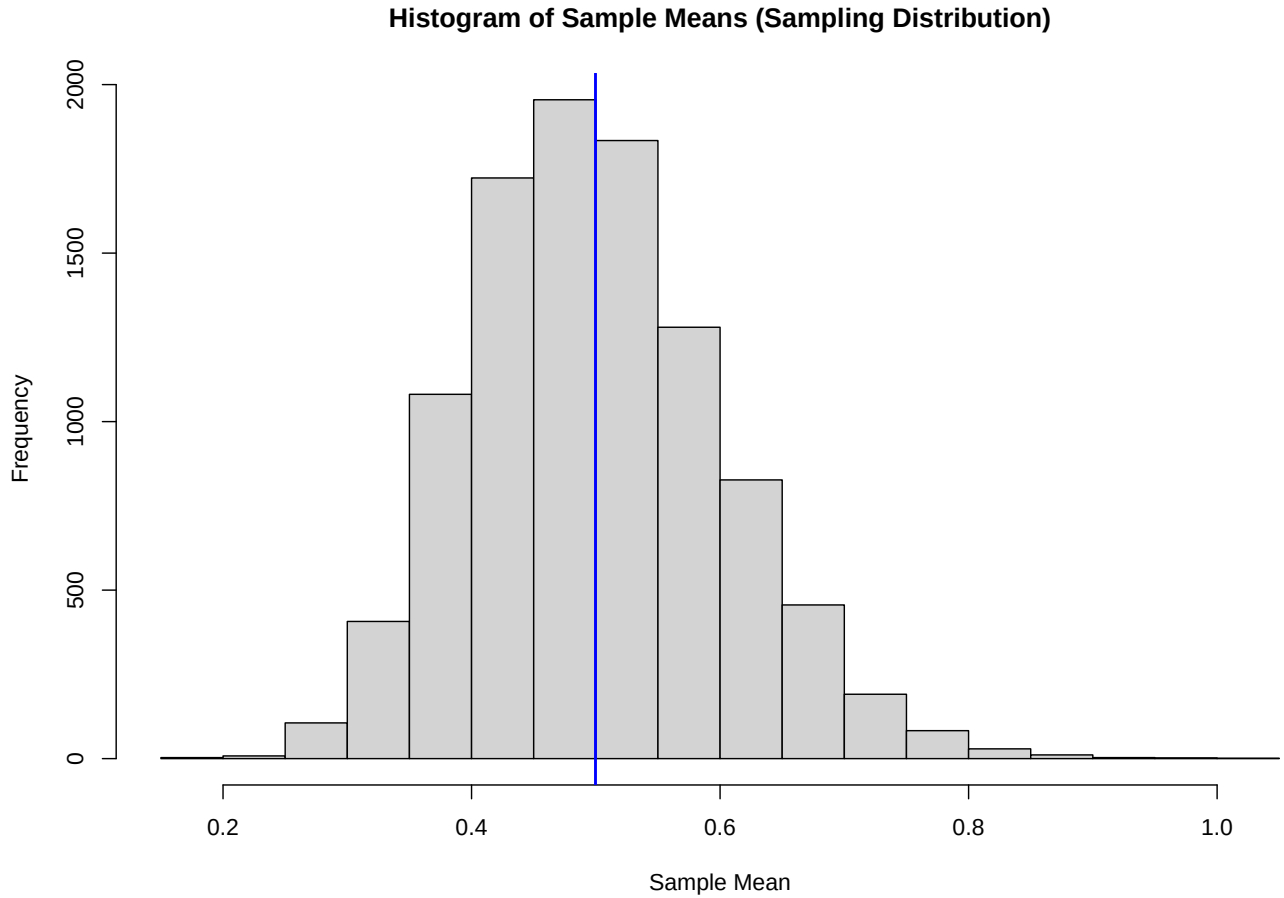
If our population distribution is exponential with rate 2 (i.e., $Exp(\lambda = 2)$), we know the mean is equal to $\frac{1}{\lambda} = \frac{1}{2} = 0.5$. We can also plot its sampling distribution from our simulations. Let's simulate 10,000 data sets with $n = 25$ using `sapply` instead of `for` loops:

```
# set seed for reproducibility
set.seed(6611)

# simulate 10,000 random samples of n=25 from exp(rate=2) dist. and save me
simres <- sapply(1:10000, function(x) mean( rexp(n=25,rate=2) ) )

# mean of 10,000 estimated sample means
mean(simres)
```

```
## [1] 0.4997321
```

Final Example: Exponential Distribution Mean

```
hist(simres, main='Histogram of Sample Means (Sampling Distribution)',
     xlab='Sample Mean') # sampling distribution of sample means
abline(v=mean(simres), col='blue', lwd=2) # add vertical line for mean
```



In Summary

- Simulation generates data where we know the "truth" ("truth" = assumed population distribution)
- Allows us to evaluate how well methods work on the "truth"
- Don't forget to **set the seed!!!**
- Sampling distribution is the distribution of a given (sample-based) statistic