

# Chapter 7: Monte Carlo Methods in Inference

Monte Carlo methods may refer to any method in statistical inference or numerical analysis where simulation is used.

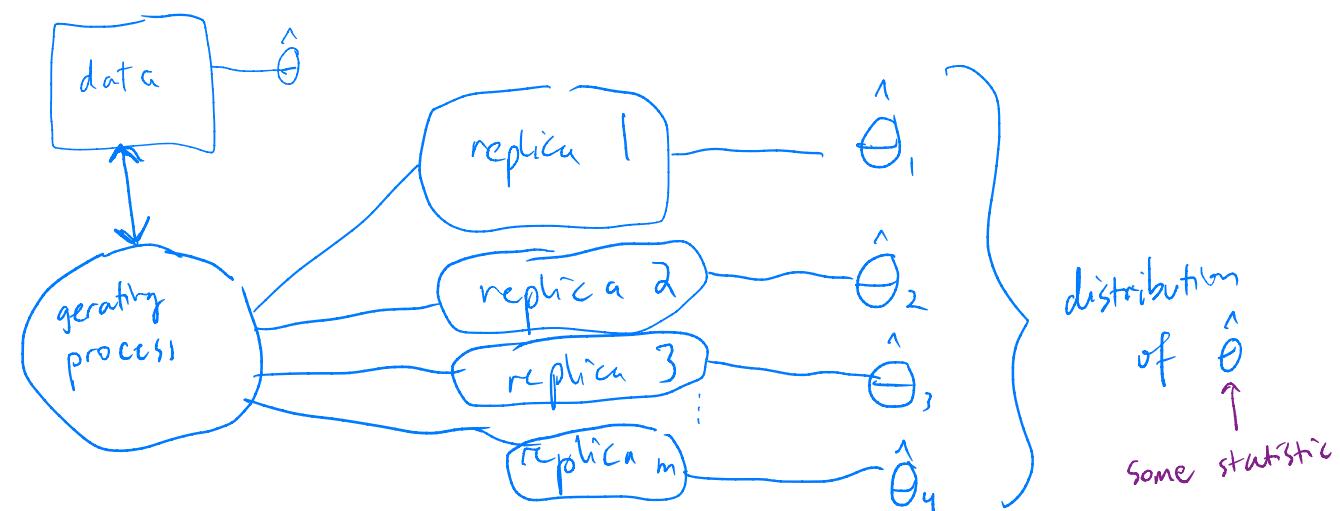
We have so far learned about Monte Carlo methods for estimation.

- ① Estimating  $\hat{\theta} = \int h(x) dx$  via rewriting  $\theta = E g(x)$ ,  $X \sim f$  and sampling  $X_1, \dots, X_m \sim f$ ,  $\hat{\theta} = \sum_{i=1}^m g(X_i)$
- ② Estimating  $\text{Var } \hat{\theta} = \frac{\text{Var } g(X)}{m}$ , sample  $X_1, \dots, X_n \sim f$ ,  $\text{Var } \hat{\theta} = \frac{1}{n} \sum_{i=1}^m (g(X_i) - \hat{\theta})^2$ .

We will now look at Monte Carlo methods to estimate coverage probability for confidence intervals, Type I error of a test procedure, and power of a test. Inference!

In statistical inference there is uncertainty in an estimate. We will use repeated sampling (Monte Carlo methods) from a given probability model to investigate this uncertainty.

This is also called "parametric bootstrap," where we simulate from the process that generated the data — repeatedly sampling under identical conditions — to have a close replica of the process reflected in the sample.



# 1 Monte Carlo Estimate of Coverage

## 1.1 Confidence Intervals

Recall from your intro stats class that a 95% confidence interval for  $\mu$  (when  $\sigma$  is known and  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ) is of the form

$$\left( \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}} \right)$$

Interpretation:

If I repeated the study 100 times and completed a CI for each using the formula above, I expect about 95 of the CI's to include the true mean  $\mu$ .

Comments:

1.  $(L, U)$  are derived from statistical theory

2.  $(L, U)$  are statistics (computed from data). If I collect new data, I get new  $(L, U)$ .

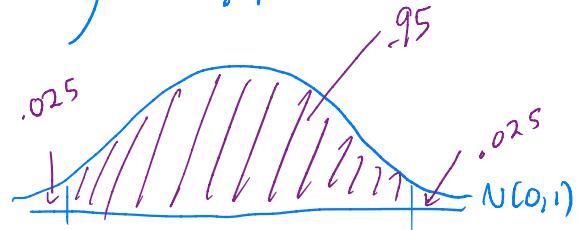
Mathematical interpretation:

$$P\left(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}\right) = .95$$

confidence level

$$\Leftrightarrow P\left(-1.96 < \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

where by CLT,  $\frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$



$$\text{i.e. } \int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 0.95$$

This holds when we have full data from  $N(\mu, \sigma^2)$ , but with real data may not be exact  $\Rightarrow$  need to estimate!

**Definition 1.1** For  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ ,  $\sigma$  known, the  $(1 - \alpha)100\%$  confidence interval for  $\mu$  is

$$\left( \bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right),$$

where

$$z_{1-\frac{\alpha}{2}} = 1 - \frac{\alpha}{2} \text{ quantile of } N(0, 1). \quad \hat{=} qnorm(1 - \alpha/2)$$

In general,

Let  $[L, U]$  be a CI for parameter  $\theta$ , then

$$P(L < \theta < U) = 1 - \alpha \quad (\text{an integral!})$$

So, if we have formulas for  $L$  and  $U$ , we can use Monte Carlo integration to estimate  $\alpha$ .

An estimate of  $1 - \alpha$  tells us about the behavior of our estimator  $[L, U]$  in practice.

## 1.2 Vocabulary

We say  $P(L < \theta < U) = P(\text{CI contains } \theta) = 1 - \alpha$ .

$\uparrow$  statistic       $\nwarrow$  the value

are our assumptions  
about our data  
reasonable?

$1 - \alpha =$   
nominal (named) coverage

$1 - \hat{\alpha} =$  empirical coverage or empirical confidence level

= simulation based estimate of the proportion of time  
that CI contains  $\theta$ .

### 1.3 Algorithm

Let  $X \sim F_X$  and  $\theta$  is the parameter of interest.

**Example 1.1**

$N(\mu, 1)$  is a function of  $\mu$ .

Consider a confidence interval for  $\theta$ ,  $C = [L, U]$ . (from stat theory).

Then, a Monte Carlo Estimator of Coverage could be obtained with the following algorithm.

a) For  $j = 1, \dots, m$ ,

① sample  $X_1^{(j)}, \dots, X_n^{(j)} \sim F$

② Compute  $C_j = [L_j, U_j]$

③  $y_j = I[\theta \in C_j] = I[L_j < \theta < U_j]$

b)  $\hat{\alpha} = \frac{1}{m} \sum_{i=1}^m y_i = \text{empirical coverage.}$

## 1.4 Motivation

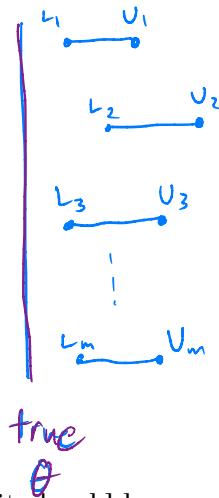
Why do we want empirical and nominal coverage to match?

Because it suggests our stated  $\alpha$  is accurate.

**Example 1.2** Estimates of  $[L, U]$  are biased.

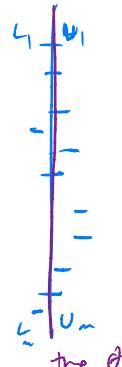
$\Rightarrow$  coverage will be low.

("I thought my method is 95% accurate, but it was 0% accurate!")



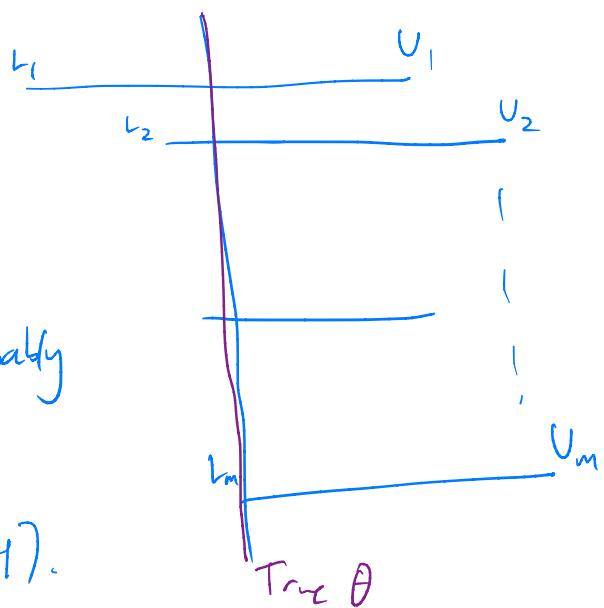
**Example 1.3** Estimates of  $[L, U]$  have variance that is smaller than it should be.

$\Rightarrow$  Low coverage



**Example 1.4** Estimates of  $[L, U]$  have variance that is larger than it should be.

$\Rightarrow$  high coverage



A bit too high is OK, but

if you have 100% coverage  
then CIs based on the method probably  
aren't useful.

(ex. 100% of GPAs are between 0 and 4).

## Your Turn

We want to examine empirical coverage for confidence intervals of the mean.

1. Coverage for CI for  $\mu$  when  $\sigma$  is known,  $\left(\bar{x} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right)$ .
  - a. Simulate  $X_1, \dots, X_n \stackrel{iid}{\sim} N(0, 1)$ . Compute the empirical coverage for a 95 confidence interval for  $n = 5$  using  $m = 1000$  MC samples.
  - b. Plot 100 confidence intervals using `geom_segment()` and add a line indicating the true value for  $\mu = 0$ . Color your intervals by if they contain  $\mu$  or not.
  - c. Repeat the Monte Carlo estimate of coverage 100 times. Plot the distribution of the results. This is the Monte Carlo estimate of the distribution of the coverage.
2. Repeat part 1 but without  $\sigma$  known. Now you will plug in an estimate for  $\sigma$  (using `sd()`) when you estimate the CI using the same formula that assumes  $\sigma$  known. What happens to the empirical coverage? What can we do to improve the coverage? Now increase  $n$ . What happens to coverage?
3. Repeat 2a. when the data are distributed  $\text{Unif}[-1, 1]$  and variance unknown. What happens to the coverage? What can we do to improve coverage in this case and why?