Chapter 7: Monte Carlo Methods in Inference

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

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

1) Estimating $\theta = \int_{\mathcal{X}} h(x)dx$ via rewriting $\theta = E[g(X)]$, $X \sim f$ and sampling $X_1, \dots, X_m \sim f$, $\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m} g(X_i)$.

2) Estimating $Var\,\hat{\theta} = \frac{Var\, g(X)}{m}$, sample $X_1, \dots, X_m \sim f$, $\hat{Var}(\hat{\theta}) = \frac{\frac{1}{m}\sum_{i=1}^{m}[(g(x_i) - \hat{\theta})^2]}{m}$

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!

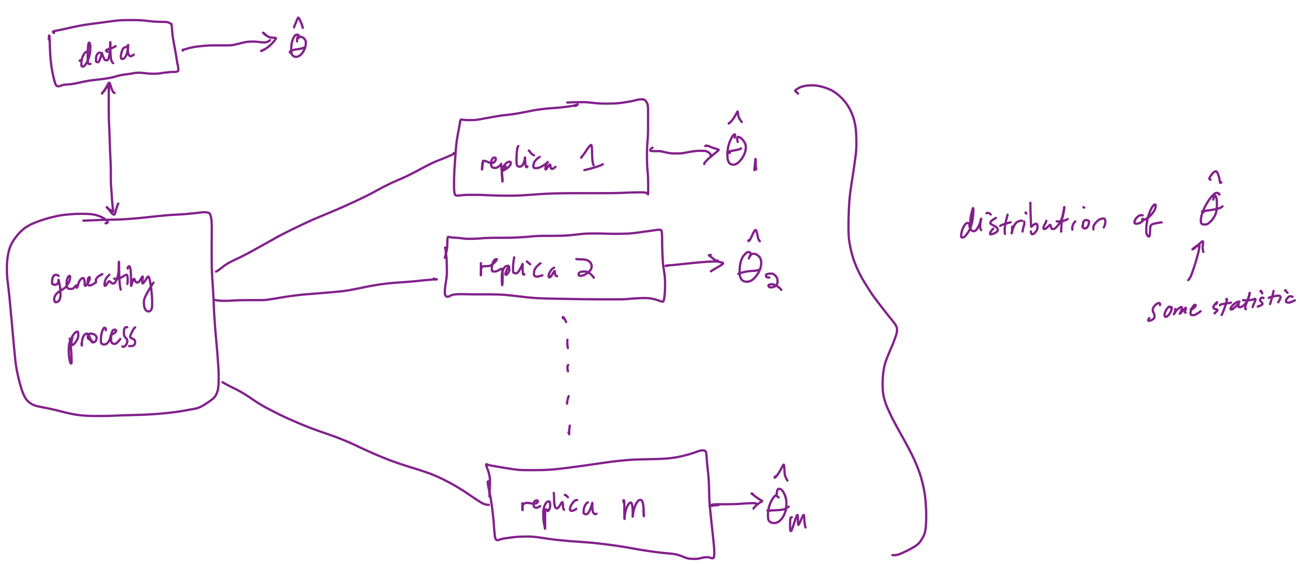
In statistical inference there is uncertainty in an estimate. We will use repeated sampling (Monte Carlo methods) from a given probability model to investigate this uncertainty.

This is also called a "parametric bootstrap"

Idea: we will simulate from process that generated our data

↳ repeatedly sample under identical conditions

in order to have a close replica of process reflected in our sample.



1 Monte Carlo Estimate of Coverage

1.1 Confidence Intervals

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

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

> Interpretation:

If I repeated my study 100 times and computed a CI for each study using the formula above, I expect 95 of the CI's to include the true mean μ.

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:

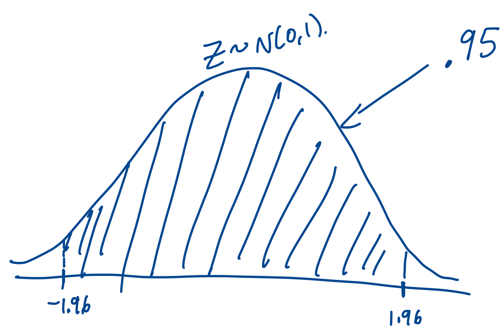
$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95$$

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

because assumed $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

(similar statement for approx. of $\bar{X}$ without normality from CLT).



$$\text{i.e.} \underbrace{\int_{-1.96}^{1.96} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx}_{\text{could estimate in real scenario!}} = 0.95$$

This holds when we have full data from $N(\mu, \sigma^2)$

BUT w/ real data this may not be true!

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

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

where

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

In general,

Let $[L, U]$ be a confidence interval for parameter θ, then

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

So, if we have formulas for L and U, we can use Monte Carlo integration to estimate α. $1-\alpha$

from stat theory.

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

$1-\alpha$ is from asymptotic theory

are our assumptions about our data reasonable?

1.2 Vocabulary

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

statistic

$1 - \alpha =$ nominal coverage

$1 - \hat{\alpha} =$ empirical coverage

= simulation based estimate of the proportion of CI's that contain θ.

1.3 Algorithm

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

Example 1.1

$N(\mu, 1)$, μ is a parameter that fully specifies this distribution (interested in estimating it).

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

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

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

1. Sample $X_1^{(j)}, \ldots, X_n^{(j)} \sim F_X$
2. Compute $C_j = [\underbrace{L(X_1^{(j)}, \ldots, X_n^{(j)})}_{= L_j}, \underbrace{U(X_1^{(j)}, \ldots, X_n^{(j)})}_{= U_j}]$
3. $y_j = \mathbb{I}[\theta \in C_j] = \mathbb{I}[L_j < \theta < U_j]$

b) Compute $1 - \hat{\alpha} = \frac{1}{m} \sum_{j=1}^{m} y_j$ = empirical coverage.

1.4 Motivation

Why do we want empirical and nominal coverage to match?

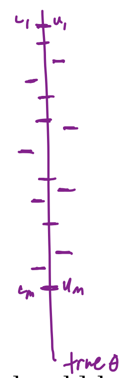
Because it suggests out stated coverage is accurate.

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

$\Rightarrow$ coverage will be low

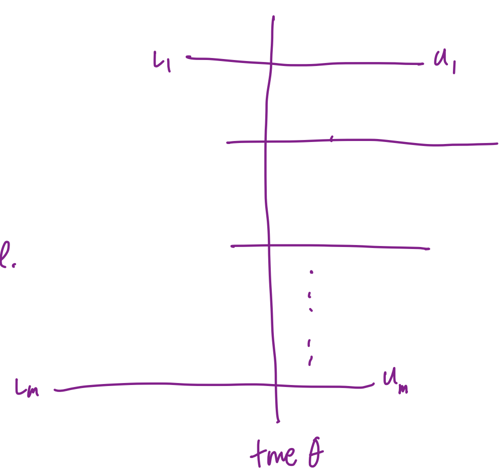
"I thought my method contained the true value 95%
but it actually contained the truth 0%!"

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
the CI's based on 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 μ when σ is known, $\left(\overline{x} - z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$.

    a. Simulate $X_1, \dots, X_n \overset{iid}{\sim} N(0, 1)$. 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 μ 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 σ known. Now you will plug in an estimage for σ (using `sd()`) when you estimate the CI using the same formula that assumes σ known. What happens to the empirical coverage? What can we do to improve the coverage? Now increase n. What happens to coverage?

3. Repeat 1a. 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?