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.

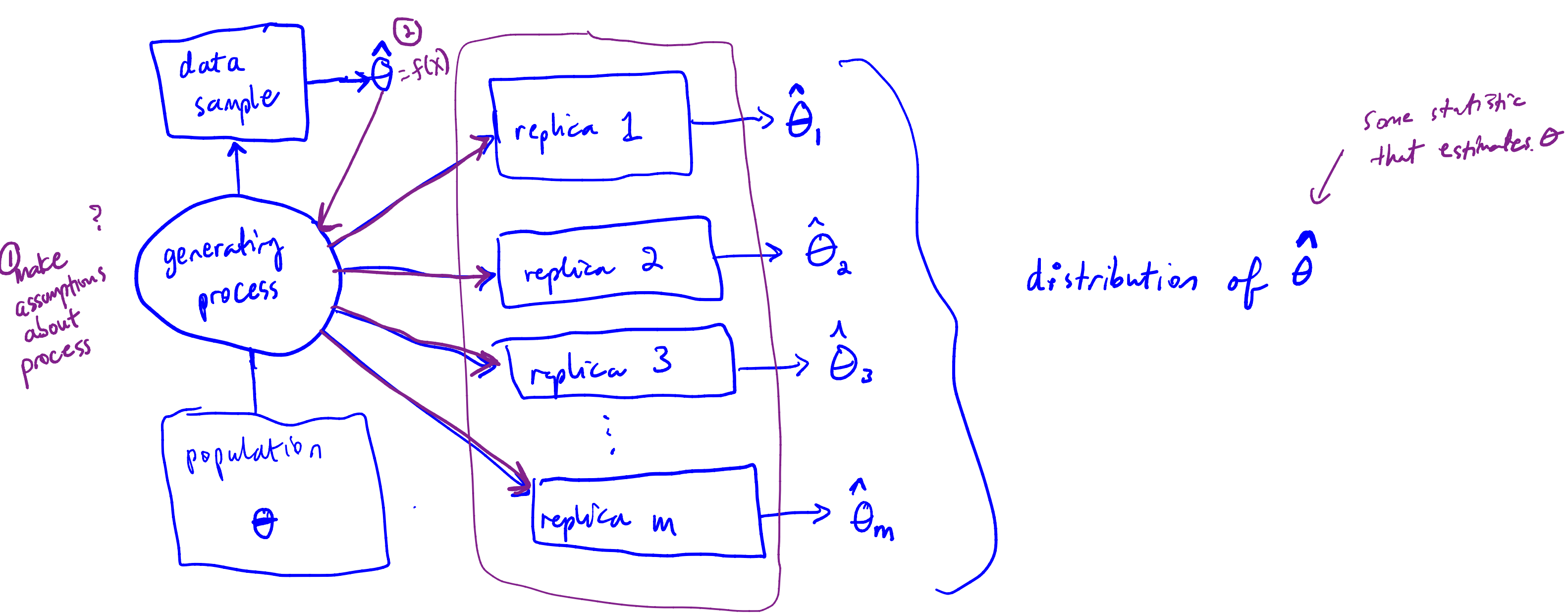
① Estimating $\theta = \int_{\mathcal{X}} h(x)\,dx$ via rewriting $\theta = E[g(X)]$, $X \sim f$ and Sampling $X_1, \dots, X_m \sim f$, $\hat{\theta} = \sum_{i=1}^{m} g(X_i)$

② Estimating $Var\,\hat{\theta} = \frac{Var\, g(X)}{m}$, Sample $X_1, \dots, X_m \sim f$, $\widehat{Var}\,\hat{\theta} = \frac{1}{m}\sum_{i=1}^{m}(g(X_i) - \hat{\theta})^2$

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 (our sample) – repeatedly sample under identical conditions – to have a close replica of the 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 $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$) is of the form

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

Interpretation:

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

Comments:

1. (L, U) are derived from stat. theory.

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

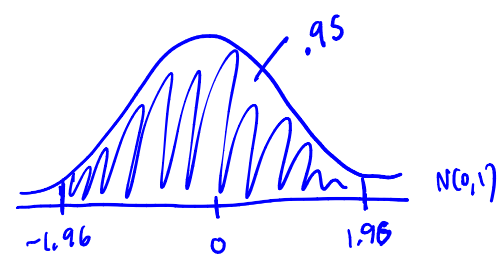
Mathematical interpretation:

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

95% confidence interval

"confidence level"

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



Because we have assumed $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$.

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

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

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

But with real data may not be exact

$\Rightarrow$ need to estimate!

Definition 1.1 For $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, 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 CI for parameter θ, then

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

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

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

true unknown parameter.

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

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

= simulation based estimate of the proportion of time CI contains θ.

1.3 Algorithm

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

Example 1.1

$X \sim N(\mu, 1)$

μ is parameter of interest.

Consider a confidence interval for θ, $C = [L, U]$. ← form determined stat theory.

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

make assumptions posit true value of θ.

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

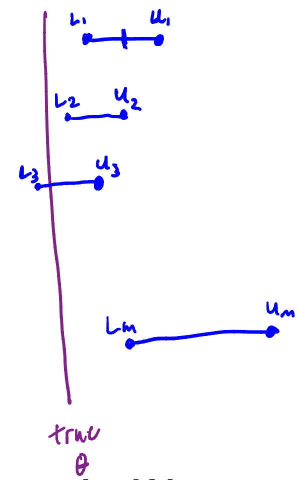
1. Sample $X_1^{(j)}, \dots, X_n^{(j)} \sim \boxed{F}$
2. Compute $C_j = [L_j, U_j]$
3. $y_j = \mathbb{I}[\theta \in C_j] = \mathbb{I}[L_j < \theta < U_j] = \begin{cases} 1 & \text{if } \theta \in C_j \\ 0 & \text{o.w.} \end{cases}$

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

1.4 Motivation

Why do we want empirical and nominal coverage to match?

Because it suggests our stated α is accurate.



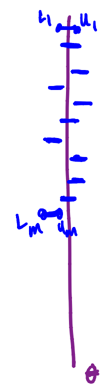
Example 1.2 Estimates of $[L, U]$ are biased.

$\Rightarrow$ coverage being low

I thought my method is right 95% of repeated experiments

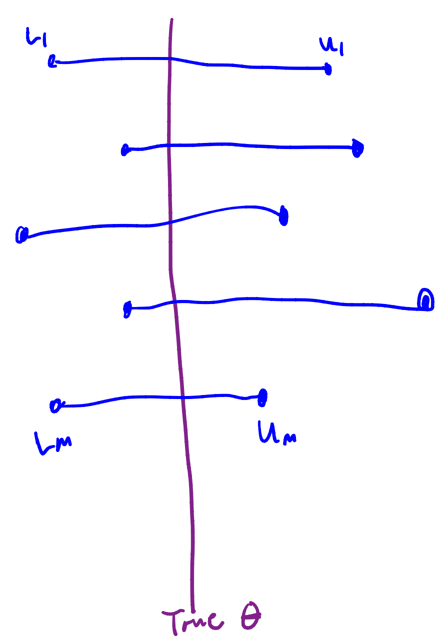
but it was 1% accurate.

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% empirical coverage
then the CI based on your method
probably aren't useful.

(e.g. CI for GPA is $(0, 4)$.
will have 100% coverage)

Your Turn

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

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

   $1-\alpha$

   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 2a. when the data are distributed Unif$[-1, 1]$ and variance unknown. What happens to the coverage? What can we do to improve coverage in this case and why?