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

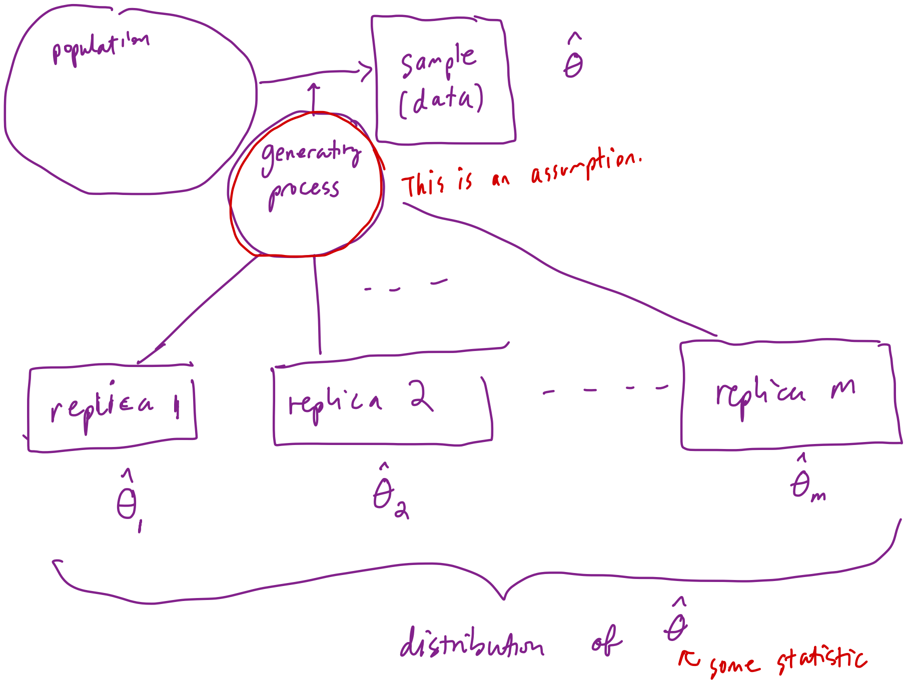
② Estimating $\hat{Var}\,\hat{\theta} = \frac{Var\, g(X)}{m}$ → sample $X_1, \dots, X_m \sim f$ $\hat{Var}(\hat{\theta}) = \frac{1}{m} \cdot \frac{1}{m} \sum_{i=1}^{m} (g(x_i) - \hat{\theta})^2$

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 similar to the "parametric bootstrap" where we simulate from a process that (is assumed to) generate the data.

↳ repeatedly sampling 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 $X_1, \dots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$) is of the form

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

Interpretation:

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

Comments:

1. (L, U) are derived from stat theory.
2. (L, U) are statistics (computed from data). If I collect new data I will get new (L, U).

Mathematical interpretation:

$$P\left(\bar{X} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96 \frac{\sigma}{\sqrt{n}} \right) = \overset{\text{confidence level.}}{0.95}$$

$$\iff P\left(-1.96 < \underbrace{\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}}_{N(0,1)} < 1.96 \right) = 0.95$$

where by assumptions of data generation, $\dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$.

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

This holds when we have full data from $N(\mu, \sigma^2)$, but with real data these assumptions may not hold exactly

=> need to estimate confidence

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) denote a CI for parameter θ, then

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

So, if we have formulas for $\underbrace{L \text{ and } U}_{\leftarrow \text{ from stat theory.}}$, 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 $\underbrace{\text{practice}}_{\text{are our assumptions about the data reasonable?}}$.

↖ from asymptotic theory

1.2 Vocabulary

We say $P(L < \theta < U) = P(\underset{\uparrow \text{ statistic}}{\text{CI}} \text{ contains } \underset{\uparrow \text{ true parameter value.}}{\theta}) = 1 - \alpha$.

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

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

= simulation-based estimate of the proportion of CIs that contain 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 the parameter of interest.

$X \sim \text{Bern}(p)$

p is the parameter of interest.

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

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 $C_j = [L_j, U_j]$                              } estimating integral.
    ③ $y_j = \mathbb{I}(\theta \in C_j) = \mathbb{I}(L_j \leq \theta \leq U_j)$

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

1.4 Motivation

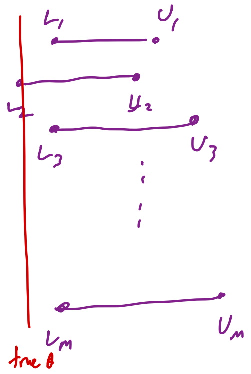
Why do we want empirical and nominal coverage to match?

Because it suggests our state α is accurate.

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

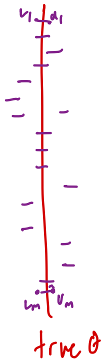
$\Rightarrow$ coverage will be low.

I thought this method was giving me 95% confidence but actually it is 5% confidence.



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

$\Rightarrow$ low coverage.

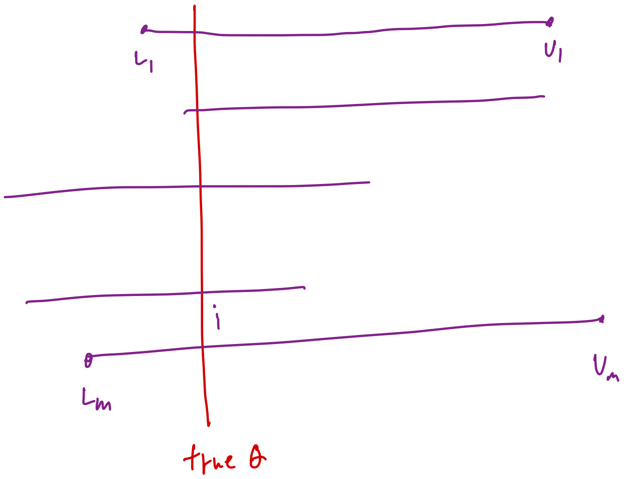


Example 1.4 Estimates of $[L, U]$ have variance that is larger than it should be.

$\Rightarrow$ high coverage.

A little bit high is ok, but if have 100% coverage the CI's 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 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 part a estimate of coverage 100 times. Plot the distribution histogram of the results. This is the Monte Carlo estimate of the distribution of the coverage.

robustness to broken assumptions.

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 $\text{Unif}[-1, 1]$ and variance unknown. What happens to the coverage? What can we do to improve coverage in this case and why?