Unit 3: Monte Carlo Methods in Inference

Chapter 7 in "Statistical Computing with R"

Anna Ly

Department of Mathematical and Computational Sciences
University of Toronto Mississauga

February 10, 2026

Overview

1. Introduction
2. R Progamming DOs and DON'Ts
3. Estimation
4. (Aside) Rant About Statistics
5. Hypothesis Testing

Introduction

Introduction

- Monte Carlo methods are helpful for estimation.
- You will likely employ these types of techniques if you do a thesis within statistics.
- We'll use it to estimate the following:
  - Mean
  - Standard Error
  - MSE (Mean squared error)
  - Alternatives to MSE
  - Confidence Levels
- We can also use Monte Carlo methods for hypothesis testing and generating p-values.

Introduction

Consider these questions which keep me up at night:

- What is the goal of statistics?
- Why do we perform hypothesis tests?
- What is actually going on when we do a hypothesis test?
- Do we actually make reasonable assumptions for the null hypothesis?
- Why do we have the threshold $\alpha = 0.05$?
- Why do the mathematicians mock statisticians?
- Can we actually start changing statistics education?

I have a scheduled rant later. One of them is coming.

R Programming DOs and DON'Ts

R Programming DOs and DON'Ts

- It's odd to establish these now instead of earlier.
- It's hard to juggle multiple things at once:
  - Teaching you how to code.
  - Teaching you best practices.
  - Teaching you how to write tests.
- There's also the pedagogical dilemma: efficient code tends to be harder to understand and read.
- However, what's the point of learning through bad methods? I had to learn efficient methods myself!
- Nonetheless, now we're coding more complex structures. It's time to drill better coding practices!

R Programming DOs and DON'Ts

- I don't care if your code is absolutely optimised; I just want you to shake off obviously inefficient habits.
- Naturally, when coding, you will think of the inefficient solution first, which is natural.
- However, you should recognise that a more efficient solution exists and practice writing them.
- After all, efficiency mostly matters for large-scale projects!
- This is obviously not comprehensive. There is a more detailed guide here. (Click text).
- We will not deduct marks in examinations for inefficient code, UNLESS you do the number one sin of growing `R` objects.
- For the project, there will be marks deducted if I comment on the inefficiencies and you do not fix them later.

Appending Entries Is Inefficient!

Inefficient For Loop and Appending

```
n = 100; x1 <- rnorm(n);
# NEVER DO
est <- c()
for(i in 1:n){
  val <- x1[i]^2
  est <- c(est, val)
}
# INSTEAD
est2 <- x1^2
```

Ask yourself: did you need to write a for-loop for your operation?

...I Saw This in a Textbook Once

Repeating

```
# NEVER DO
v1 <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
# INSTEAD (either or)
v2 <- rep(0, 10)
v3 <- numeric(10) # 0 case only
```

Easy Integer List

```
# NEVER DO
v1 <- c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)
# INSTEAD
v2 <- 1:10
```

Matrices Are Efficient!

You will see in the examples for MC estimation that I will be using a lot of matrices! Some elementary functions below:

Basic Matrix Use

```
n <- 10000
m <- matrix(rnorm(n*n), nrow = n)
rowSums(m)
colSums(m)
rowMeans(m)
colMeans(m)
```

How to Test Efficiency?

system.time()

```
> n = 100; x1 <- rnorm(n);
> # NEVER DO
> system.time({
+   est <- c()
+   for(i in 1:n){
+     val <- x1[i]^2
+     est <- c(est, val)
+   }
+ })
    user  system elapsed
   0.002   0.000   0.002
```

How to Test Efficiency?

system.time()

```
> # INSTEAD
> system.time({
+   est2 <- x1^2
+ })
   user  system elapsed
      0       0       0
```

Monte Carlo Estimation

General Idea for MC Estimation

- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution X.
- An estimator $\hat{\theta}$ for a parameter θ are is a function of the random sample.
- Random variates from the sampling distribution of X can be generated by repeatedly drawing independent random samples $\mathbf{x}^{(j)}$ for $j = 1, 2, \ldots, m$ where m represents the Monte Carlo sample size.
- We can compute $\hat{\theta}$ for each sample j.

Estimating the Mean

Example

Suppose that X_1, X_2 are iid from a standard Gaussian distribution.

(a) Find the exact value of $\mathbb{E}[|X_1 - X_2|]$.

(b) Derive a Monte Carlo estimator of $\mathbb{E}[|X_1 - X_2|]$.

(c) Use part (b) to find an approximate value of $\mathbb{E}[|X_1 - X_2|]$.

Solution.

Estimating the Mean

Estimating the Mean

Critiquing R-code

R Code from Textbook

```
m <- 10000
g <- numeric(m)
for (i in 1:m) {
  x <- rnorm(2)
  g[i] <- abs(x[1] - x[2])
}
val2 <- mean(g)
```

Some glaring issues:

- Unnecessary for-loop
- Running `rnorm()` *m* amount of times; could've done this easily in the beginning.

Estimating Standard Error

Standard Error

Standard error of the mean $\overline{X}$ of a sample of size n is $\sqrt{\mathbb{V}(X)/n}$.

Estimators for Standard Error

An intuitive choice to estimate the variance is:

$$\widehat{\mathbb{V}(X)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

But the **unbiased** estimator is:

$$\widehat{\mathbb{V}(X)} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

Estimating Standard Error

Example

In the previous example, find the estimate of the standard error for $|X_1 - X_2|$ using two methods. Compare it to the true standard error.

Solution.

Estimating MSE

Mean Squared Error (MSE)

$$MSE(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2] = \mathbb{V}[\hat{\theta}] + Bias(\hat{\theta})^2$$

Based on this formula, the Monte Carlo estimate can be found using the following:

1. Generate m random samples $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ from the distribution of interest X, where $x^{(i)}$ are of size n for $i = 1, 2, \ldots, m$.
2. Compute the estimator: $\hat{\theta}^{(j)} = \hat{\theta}(x_1^{(j)}, x_2^{(j)}, \ldots, x_n^{(j)})$ for $j = 1, 2, \ldots, m$.
3. Compute: $\widehat{MSE}(\hat{\theta}) = \frac{1}{m} \sum_{j=1}^{m} (\hat{\theta}^{(j)} - \theta)^2$.

Warning: do not confuse m and n in this case!!

Estimating MSE

Example

Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\theta, \theta^2)$. We want to estimate θ. We have two possible candidates:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad , \hat{\theta}_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Compute the MSE of these two estimators and determine which one is "better".

Solution.

Estimating Alternatives to MSE

- Why do we use the MSE? Karlin (1958) noted advantages of using MSE when comparing estimators:
  - For an unbiased estimator $\hat{\theta}$ of θ, the MSE equals its variance.
  - Squared error emphasizes large deviations, so a smaller MSE implies more consistent accuracy.
- However, Rao (1980) argued that MSE does not indicate how often an estimator lies close to the true parameter.
- Focusing solely on MSE encourages reducing the variance of unbiased estimators, hence the origin of solving for the unique minimum variance unbiased estimators (UMVUE).
- However, many useful estimators, such as the MLE, may be biased.

Estimating Alternatives to MSE

There are some alternatives to the MSE:

Mean Absolute Error

$$\mathbb{E}[|\hat{\theta} - \theta|]$$

Root Mean Squared Error

$$\sqrt{\mathbb{E}[(\hat{\theta} - \theta)^2]}$$

You do not need to memorize these.

Another possible way to compare estimators is to use the Pitman closeness (PC) criterion.

Estimating Alternatives to MSE

Pitman Closeness Criterion

Suppose we want to compare two estimators, $\hat{\theta}_1$ and $\hat{\theta}_2$, compared to a common parameter θ. Then, the PC of $\hat{\theta}_1$ relative to $\hat{\theta}_2$ is:

$$\pi_{\hat{\theta}_1,\hat{\theta}_2}(\theta) := \mathbb{P}(|\hat{\theta}_1 - \theta| \leq |\hat{\theta}_2 - \theta|).$$

If $\pi_{\hat{\theta}_1,\hat{\theta}_2}(\theta) > \frac{1}{2}$ then $\hat{\theta}_1$ is said to be **Pitman closer** to θ, implying it is a more desirable estimator.

The PC also has an intuitive appeal in applied settings. For example, Keating and Mason (1985) noted that in elections, voters typically support the candidate whose position is *probably* the closest to their own, not the one minimising squared differences. Similarly, customers choose the convenience store that *probably* has the shortest distance.

Common misconception! PC does not evaluate the magnitude of how close an estimator is to the parameter.

Estimating Alternatives to MSE

- Despite its flaws, the MSE criterion has dominated statistics education and research.
- It is also genuinely the easiest to compute.
- Others, especially the PC criterion, are more complicated to derive exact expressions for.
- With Monte Carlo methods, we may simulate the PC criterion instead of deriving an exact expression. :3

Estimating Alternatives to MSE

Example

Suppose $X_1, X_2, \ldots, X_n$ are iid $N(\theta, \theta^2)$. We want to estimate θ. We have two possible candidates:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad , \hat{\theta}_2 = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Compute the PC probability of these two estimators and use the PC criterion to determine which one is "better".

Solution.

Estimating Confidence Levels

- Let $\hat{\theta}$ be a statistic that is Gaussian distributed with mean θ and standard error $\sigma_{\hat{\theta}}$.
- As $n \to \infty$, $Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}}$ has a standard Gaussian distribution as well.
- The endpoints of a $100(1-\alpha)\%$ confidence interval for θ is given by:

$$\hat{\theta}_L = \hat{\theta} - z_{\alpha/2}\sigma_{\hat{\theta}}, \quad \hat{\theta}_U = \hat{\theta} + z_{\alpha/2}\sigma_{\hat{\theta}},$$

  where $z_{\alpha/2}$ is defined as the value where $\mathbb{P}(Z > z_{\alpha/2}) = \alpha/2$ is true.
- In most cases, we won't know $\sigma_{\hat{\theta}}$ and we are generating from a sample of size n; thus, we may use the t-distribution and $t_{\alpha/2,n-1}$ instead (where $n-1$ represents the degree of freedom.)

Estimating Confidence Levels

To generate a confidence level from the previous scenario, employ the following:

1. Generate m random samples $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ from a Gaussian distribution where $x^{(j)}$ are of size n for $j = 1, 2, \ldots, m$.
2. For each replicate $x^{(j)}$:
   - 2.1 Compute the statistic $\hat{\theta}^{(j)}$
   - 2.2 Compute the sample standard error $\widehat{\sigma_{\hat{\theta}}}^{(j)}$.
   - 2.3 Compute the upper and lower bounds:

$$\hat{\theta}_L^{(j)} = \hat{\theta}^{(j)} - t_{\alpha/2,n-1}\widehat{\sigma_{\hat{\theta}}}^{(j)}, \quad \hat{\theta}_U^{(j)} = \hat{\theta}^{(j)} + t_{\alpha/2,n-1}\widehat{\sigma_{\hat{\theta}}}^{(j)},$$

   - 2.4 Compute $y_i = \mathbf{I}(\hat{\theta}_L^{(j)} \leq \theta \leq \hat{\theta}_U^{(j)})$
3. Compute the empirical confidence level $\overline{y} = \frac{1}{m}\sum_{j=1}^{m} y_j$

Estimating Confidence Levels

Example

Generate a Gaussian sample of size 30 with mean 5 and variance 2. Using 10000 replicates, compute the empirical confidence level for $\hat{\mu} = \sum_{i=1}^{n} x_i$ given $\alpha = 0.05$.

Solution.

(Aside) Rant About Statistics

If I could only pick one thing for everyone to learn from this course, it would be this rant.

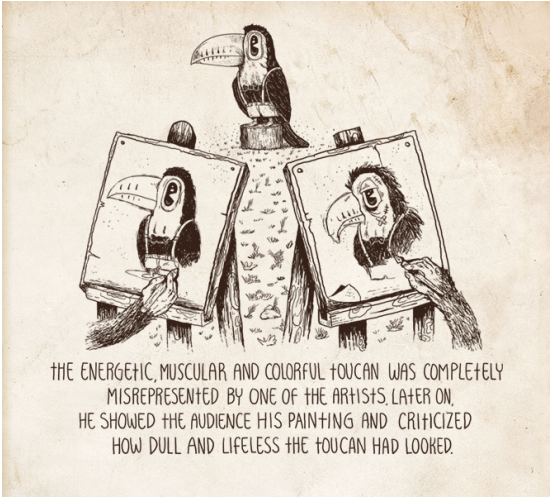
I don't believe we should be doing hypothesis testing.

This take is not unique.

Hypothesis testing amounts to disproving a strawman.

*Under specific certain conditions hypothesis tests are fine. But I've never seen anyone do a hypothesis test without using a strawman.

What is a "Strawman?"



From "An Illustrative Book of Bad Arguments" (highly recommended.)

Null Hypothesis Is Never True

T-Test Hypothesis Testing Set-Up

$H_0 : \mu_1 - \mu_2 = 0, \quad H_1 : \mu_1 - \mu_2 \neq 0$

Key questions to think about:

- Why do we say we "fail to reject" rather than "accept" the null hypothesis?
- Is the null hypothesis ever literally true in real-world data?
- Why is 0 the benchmark difference? What does "no effect" even mean?
- Suppose we estimate $\mu_1 - \mu_2 = 0.000001$. Does that make H_0 true?
- If I don't get the results I want, can't I choose a very extreme or unrealistic null to obtain statistical significance?

If you do not assume a simple null hypothesis then the null hypothesis can be true. However, all statistical tests you have learned assume a simple null hypothesis...

What is a *p*-value?

Pick one (or none at all):

(i) The probability that the null hypothesis is true.

(ii) The probability that the alternative hypothesis is true.

(iii) One minus the probability that the alternative hypothesis is true.

(iv) One minus the probability of replication.

Some people who teach statistics can't even answer this question properly (source).

Bob on **March 20, 2019 7:30 PM at 7:30 pm** said:

This times 1 billion. I'm sick of hearing about students 'not understanding p values'. Who is teaching them? How do you pass a stats course and not understand p values? Why don't these students fail? Lazy and incompetent lecturers, that's the problem.

What is a p-value?

p-value

Given that the null hypothesis is true, the p-value is the probability of observing a test statistic at least as extreme as the one obtained from the data.

If the null hypothesis is never true, what is the point of computing this?

If you do not assume a simple null hypothesis then the null hypothesis can be true. However, all statistical tests you have learned assume a simple null hypothesis...

What is a *p*-value?

David J. Marcus on **March 21, 2019 8:18 PM at 8:18 pm** said:

You can't teach p-values correctly because they don't make sense. Andrew has the right approach: Spend a little time on them to explain why you shouldn't use them. Then teach things that do make sense and work.

- Is there a point in teaching *p*-values?
- Educators and students don't understand it...
- The definition of a *p*-value is confusing and flawed...
- Some statisticians are begging to abandon the concept of statistical significance...

What Can We Do?

- Traditional statistics education heavily emphasises hypothesis testing and p-values.
- Courses for *non-statistics* majors often teach hypothesis testing as the primary tool for making research conclusions.
- Courses for *statistics* majors also dedicate significant time to hypothesis testing and p-values.
- Many applied fields, especially in the medical and social sciences, rely on hypothesis testing and p-values to draw conclusions.
- I even wrote a hypothesis testing "how-to" guide for STA304 students. (My biggest regret!)
- For the record: I began questioning hypothesis testing halfway through my master's degree.
- So... can we undo the damage?

Let's Take a Breath...

Before throwing out hypothesis testing entirely, let's pause and think:

- What is the ultimate goal of statistics?
- If p-values are flawed, can they still serve a useful purpose?
- Are there practical, accessible alternatives to p-values?
- Why are we so drawn to simple, black-and-white answers?

Let's Calm Down...

- Yoav Benjamini: "In some sense it offers a first line of defense against being fooled by randomness, separating signal from noise, because the models it requires are simpler than any other statistical tool needs."
- Ben Recht: "Statistical tests constrain outcomes in participatory systems. Engineers want to push features to get promoted; data science teams insist on AB tests to ensure these features don't harm key metrics. Drug companies want to make a ton of money; clinical trials ensure drugs aren't harmful and have a chance of being beneficial. Academics want to publish as many papers as possible to get their h-index to the moon; journals insist on some NHSTs to placate editors. The purpose of statistical tests is regulation."

What Do I Recommend?

- An analysis based on confidence intervals is actually quite reasonable.
- You have likely noticed a similarity between the confidence interval not including 0 and a p-value being less than the α threshold.
- To clarify my position: I don't have an actual problem with hypothesis testing, rather I have a problem with statistical tests that were created with a strawman null hypothesis. Also I despise an analysis strictly on p-values.
- If you want to do hypothesis testing properly, you should construct a reasonable non-point null hypothesis based on the context. (I.e., $H_0 : \mu_1 - \mu_2 \in (-1, 1)$).
- This will require you to construct a hypothesis test by hand, which I'll go over a simple example at the end.

Why I Still Teach Hypothesis Testing

- My scepticism about hypothesis testing isn't new—but it's still not mainstream.
- In many professional settings, you'll still be expected to use hypothesis testing methods.
- Let's be honest: part of pursuing a university degree is preparing for a white-collar career.
- Interviewers still ask, "What is a p-value?"
- Alternatively, even if you don't personally follow a particular religion (Christianity, Islam, Judaism, Buddhism, Hinduism, etc.), you might still learn about it because many others do, and understanding their beliefs helps you engage with them more thoughtfully.
- In the same way, even if I don't "believe in" hypothesis testing, it's important to understand it because many statisticians do.

Is This Rant Testable?

YES.

Monte Carlo Hypothesis Testing

Hypothesis Testing

Type I Error / False Negative

$$\mathbb{P}(\text{reject } H_0 | H_0 \text{ is true})$$

Type II Error / False Positive

$$\mathbb{P}(\text{fail to reject } H_0 | H_0 \text{ is false})$$

Power

$$\textbf{Power}(\theta) = \mathbb{P}(\text{reject } H_0 | H_0 \text{ is false})$$

We can simulate these values.

Hypothesis Testing

To calculate the **type-I error rate**, we assume that the null hypothesis is true. Therefore, we simulate:

1. Generate m random samples $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ from the null distribution where $x^{(j)}$ are of size n for $j = 1, 2, \ldots, m$.
2. For each replicate $x^{(j)}$:
   2.1 Compute the test statistic T_j.
   2.2 Record the following:

$$I_j = \begin{cases} 1 & H_0 \text{ is rejected at significance level } \alpha, \\ 0 & \text{otherwise.} \end{cases}$$

3. Compute the type-I error rate by looking at the proportion of tests where H_0 was rejected: $\frac{1}{m} \sum_{j=1}^{m} I_j$.

Example

Suppose that $X_1, \ldots, X_{20}$ is a random sample from a $N(\mu, \sigma^2 = 100^2)$ distribution. Test $H_0 : \mu = 500$, $H_1 : \mu \neq 500$ at $\alpha = 0.05$ where our estimator of μ is $\overline{X}$. Compute type-I error.

Solution.

Hypothesis Testing

We can simulate the **type-II error rate** and **power** if we are assuming a simple vs. simple test, i.e., the alternative hypothesis is well defined. Assume that the null hypothesis is false. Therefore, we simulate:

1. Generate m random samples $x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ from the alternative distribution where $x^{(j)}$ are of size n for $j = 1, 2, \ldots, m$.
2. For each replicate $x^{(j)}$:
   2.1 Compute the test statistic T_j.
   2.2 Record the following:

$$I_j = \begin{cases} 0 & H_0 \text{ is rejected at significance level } \alpha, \\ 1 & \text{otherwise.} \end{cases}$$

3. Compute the type-II error rate by looking at the proportion of tests where H_0 was not rejected: $\frac{1}{m} \sum_{j=1}^{m} I_j$.

Remember: $\textbf{Power}(\theta) = 1 - \beta$.

Example

Using the same example as the previous example, assume $H_1 : \mu = 600$. Now compute type-I error and power. Also, test for different values for the alternative hypothesis to see how we can inflate our power!!

Solution.

More Formal/General) Hypothesis Testing

Let θ denote a population parameter and let Θ represent the parameter space, i.e., $\theta \in \Theta$. We say the null hypothesis is $H_0 : \theta \in \Theta_0$ and the alternative hypothesis is $H_1 : \theta \in \Theta_0^c$.

α

Let $\pi(\theta)$ denote the probability of rejecting H_0. Then,

$$\alpha = \sup_{\theta \in \Theta_0} \pi(\theta).$$

Likelihood Ratio Test (LRT)

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ represent a vector of realisations of the data. The likelihood ratio test *statistic* for testing $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$ is:

$$\lambda(\mathbf{x}) = \frac{\sup_{\Theta_0} L(\theta|\mathbf{x})}{\sup_{\Theta} L(\theta|\mathbf{x})}$$

A *likelihood ratio test* is any test that has a rejection region of the form $\{\mathbf{x} : \lambda(\mathbf{x}) \leq c\}$, where c is any number satisfying $0 \leq c \leq 1$.

Aside: Supremum/Infimum

Let S represent a subset of a partially ordered set P.

Supremum (Least Upper Bound)

An upper bound b of S is called a supremum of S if for all upper bounds z of S in P, $z \geq b$.

Infimum (Greatest Lower Bound)

A lower bound a of S is called an infimum of S if for all lower bounds y of S in P, $y \leq a$.

These concepts are similar to max/mins but sometimes the max or min doesn't exist. Consider the following:

Example

For $f(x) = e^x$ the infimum is 0 and for $g(x) = -e^x$ the supremum is 0.

Hypothesis Testing

- Earlier, my main criticisms of hypothesis testing were that you're arguing against a strawman. Most methods we use assume a simple null hypothesis.
- However, my criticism is less valid if we instead assume an interval... (However, most people DO NOT construct a hypothesis test by hand.) I'm going to show how we can do a more reasonable hypothesis test.

Simulating a More Complicated Null Hypothesis

Suppose that we had $n = 30$ patients who were sick and we give them some **drugs**. We assume that the amount of time it takes for them to recover in days is Gaussian distributed with $H_0 : \mu \geq 7$. Hence, our alternative hypothesis is $H_1 : \mu < 7$ (that is, if they take less than 7 days to recover, we consider the medicine to be effective.)

Hypothesis Testing

Hypothesis Testing

Disclaimer! This is NOT a Monte Carlo simulation!

```
n <- 30; sig <- 2.5 # assuming variance is known...
# This is simulated but a placeholder to put real data
dat <- rnorm(30, mean = 4, sd = sig)
if(mean(dat) >= 7){
  print("fail to reject")
} else{
  c <- qchisq(0.05, df = 1, lower.tail = FALSE)
  crit.region <- ((mean(dat)-7)/(sig/sqrt(n)))^2
  if(crit.region > c){
    print("reject")
  } else {
    print("fail to reject")
  }
}
```