Week 03: Monte Carlo Estimation

Mark M. Fredrickson (mfredric@umich.edu)

Computational Methods in Statistics and Data Science (Stats 406)

Monte Carlo Estimation

Estimation Review

As with hypothesis testing, we want to **infer** something about a **population** based on a **sample**:

- Specific **parameters** in the distribution function.
- **Moments** of the population (mean, variance, skew, etc).
- Probabilities or quantiles: $P(X > c) = \theta$ or $P(X \leq \theta) = c$

Estimation Review

As with hypothesis testing, we want to **infer** something about a **population** based on a **sample**:

- Specific **parameters** in the distribution function.
- **Moments** of the population (mean, variance, skew, etc).
- Probabilities or quantiles: $P(X > c) = \theta$ or $P(X \leq \theta) = c$

To make our guesses, we'll use an **estimator** (another term for a **statistic**):

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

Estimation Review

As with hypothesis testing, we want to **infer** something about a **population** based on a **sample**:

- Specific **parameters** in the distribution function.
- **Moments** of the population (mean, variance, skew, etc).
- Probabilities or quantiles: $P(X > c) = \theta$ or $P(X \leq \theta) = c$

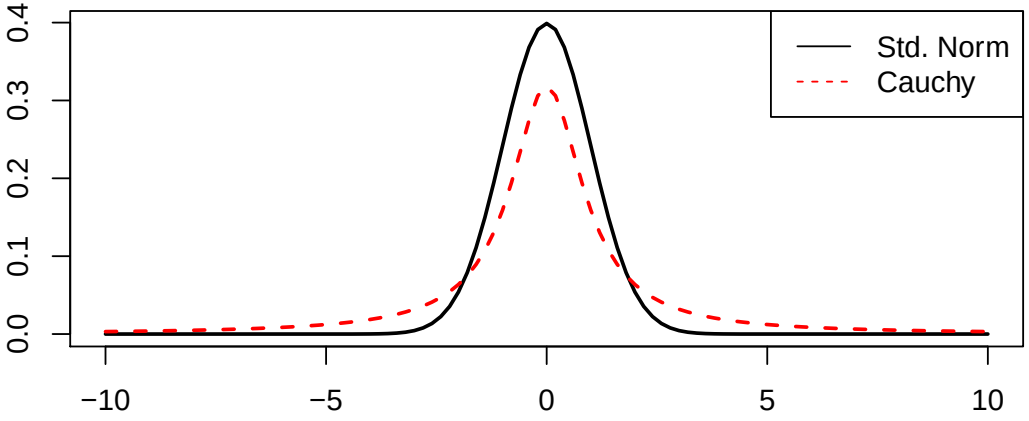
To make our guesses, we'll use an **estimator** (another term for a **statistic**):

$$\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$$

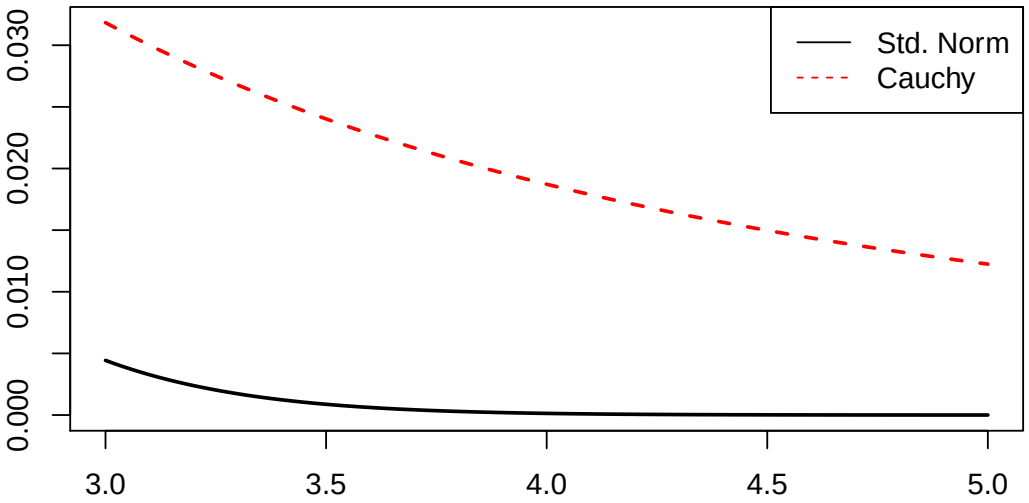
Understanding the distribution of $\hat{\theta}$ allows us to describe uncertainty and compare different estimators.

“Light” and “heavy” tails

The **Cauchy distribution** is known to have **“heavy tails”** (i.e., puts more mass away from the center).



Zooming in on the tails



Some properties of a Cauchy distribution

The PDF of the Cauchy is given by:

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

Some properties of a Cauchy distribution

The PDF of the Cauchy is given by:

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

The parameter θ **is the median** of the Cauchy distribution.

Some properties of a Cauchy distribution

The PDF of the Cauchy is given by:

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

The parameter θ **is the median** of the Cauchy distribution.

Interesting fact: The Cauchy distribution does not have a mean: $E(X) = \infty$!

Some properties of a Cauchy distribution

The PDF of the Cauchy is given by:

$$f(x) = \frac{1}{\pi} \frac{1}{1 + (x - \theta)^2}$$

The parameter θ **is the median** of the Cauchy distribution.

Interesting fact: The Cauchy distribution does not have a mean: $E(X) = \infty$!

Suppose $X_i \overset{\text{iid}}{\sim} \text{Cauchy}(\theta)$ for $i = 1, \ldots, n$. How do we **estimate** θ?

Estimating θ

As θ is the median, the **sample median** is a natural choice.

Estimating θ

As θ is the median, the **sample median** is a natural choice.

While the Cauchy distribution does not have a mean, **the sample mean will exist** for any finite n. Perhaps this would be a good estimator?

Estimating θ

As θ is the median, the **sample median** is a natural choice.

While the Cauchy distribution does not have a mean, **the sample mean will exist** for any finite n. Perhaps this would be a good estimator?

Which of these estimators is "better?" What does it mean for an estimator to be better?

Estimating θ

As θ is the median, the **sample median** is a natural choice.

While the Cauchy distribution does not have a mean, **the sample mean will exist** for any finite n. Perhaps this would be a good estimator?

Which of these estimators is "better?" What does it mean for an estimator to be better?

Let's investigate the **operating characteristics** of estimators: bias, variance, and mean squared error.

Sampling Distributions

Recall that our **estimators are statistics** and **have a distribution**.

Sampling Distributions

Recall that our **estimators are statistics** and **have a distribution**.

Since this distribution comes from sampling n values, we call this the **sampling distribution**. We will **estimate** the sampling distribution using **Monte Carlo techniques**.

```
> n <- 10
> true_theta <- 10
> samples <- rerun(10000, rcauchy(10, location = true_theta))
```

Sampling Distributions

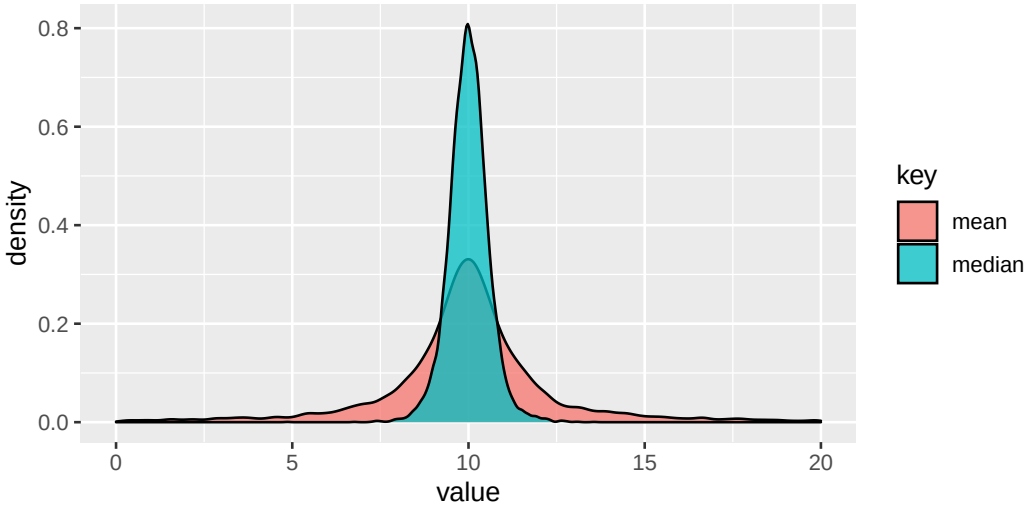
Recall that our **estimators are statistics** and **have a distribution**.

Since this distribution comes from sampling n values, we call this the **sampling distribution**. We will **estimate** the sampling distribution using **Monte Carlo techniques**.

```
> n <- 10
> true_theta <- 10
> samples <- rerun(10000, rcauchy(10, location = true_theta))

> median_sampling_dist <- map_dbl(samples, median)
> mean_sampling_dist <- map_dbl(samples, mean)
```

Sampling Distributions



Bias

When we performed hypothesis tests, we asked what **kinds of errors we could make**.

Bias

When we performed hypothesis tests, we asked what **kinds of errors we could make**.

Similar to rejecting a **true null hypothesis**, we can **under/over estimate the truth on average**:

$$|\mathrm{E}\left(\hat{\theta}\right) - \theta| > 0$$

Bias

When we performed hypothesis tests, we asked what **kinds of errors we could make**.

Similar to rejecting a **true null hypothesis**, we can **under/over estimate the truth on average**:

$$|\mathrm{E}\left(\hat{\theta}\right) - \theta| > 0$$

We call $\mathrm{E}\left(\hat{\theta}\right) - \theta$ the **bias** of the estimator. An estimator is **unbiased** if $E(\hat{\theta}) = \theta$.

Bias

When we performed hypothesis tests, we asked what **kinds of errors we could make**.

Similar to rejecting a **true null hypothesis**, we can **under/over estimate the truth on average**:

$$|\mathrm{E}\left(\hat{\theta}\right) - \theta| > 0$$

We call $\mathrm{E}\left(\hat{\theta}\right) - \theta$ the **bias** of the estimator. An estimator is **unbiased** if $E(\hat{\theta}) = \theta$.

Estimated bias of the two estimators:

```
> mean(median_sampling_dist) - true_theta
[1] 0.001323
> mean(mean_sampling_dist) - true_theta
[1] 11.77
```

More sampling distribution details

```
> summary(median_sampling_dist)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.87    9.66   10.00   10.00   10.33   13.57

> summary(mean_sampling_dist)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  -9051       9      10      22      11  119723
```

Comparing variance

We also asked how often we could **reject false null hypotheses** (power).

Comparing variance

We also asked how often we could **reject false null hypotheses** (power).

The equivalent for estimation is the **variance of the sampling distribution**.

```
> var(median_sampling_dist)

[1] 0.3292

> var(mean_sampling_dist)

[1] 1449215
```

A better estimator

The **order statistics** of a sample are the **sorted values**:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

I.e., $X_{(1)}$ is the sample minimum; $X_{(n)}$ is the sample maximum.

A better estimator

The **order statistics** of a sample are the **sorted values**:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

I.e., $X_{(1)}$ is the sample minimum; $X_{(n)}$ is the sample maximum.

Another estimator takes the average of some middle portion of the order statistics, we call this a "trimmed mean".

```
> trimmed_mean <- function(x, k) {
+   n <- length(x)
+   orderstats <- sort(x)
+   mean(orderstats[k:(n - k)])
+ }
```

How should we pick k?

```
> ks <- 1:4
> map(ks, function(k) {
+   ts <- map_dbl(samples, ~ trimmed_mean(.x,  k = k))
+   t_bias <- mean(ts) - true_theta
+   t_var  <- var(ts)
+   c(t_bias, t_var)
+ }) %>% bind_cols

# A tibble: 2 x 4
        V1     V2     V3     V4
     <dbl>  <dbl>  <dbl>  <dbl>
1    -4.22 -0.418 -0.251 -0.202
2 11534.    1.27   0.460  0.358
```

Bigger sample size n

```
> n <- 500
> k <- round(0.38 * n / 2)
> samples_500 <- data.frame(replicate(1000, rcauchy(n, true_theta)))
> median_500 <- map_dbl(samples_500, median)
> trimmed_500 <- map_dbl(samples_500, trimmed_mean, k = k)
```

Bigger sample size n

```
> n <- 500
> k <- round(0.38 * n / 2)
> samples_500 <- data.frame(replicate(1000, rcauchy(n, true_theta)))
> median_500 <- map_dbl(samples_500, median)
> trimmed_500 <- map_dbl(samples_500, trimmed_mean, k = k)

> c(mean(median_500), var(median_500))

[1] 9.995974 0.005108

> c(mean(trimmed_500), var(trimmed_500))

[1] 9.990400 0.006029
```

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean squared error:

$$\text{MSE}(\hat{\theta}) = E((\hat{\theta} - \theta)^2)$$

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean squared error:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\ &= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 \end{aligned}$$

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean squared error:

$$\begin{aligned}\text{MSE}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\ &= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 \\ &= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta})]^2 - 2E(\hat{\theta})\theta + \theta^2\end{aligned}$$

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean squared error:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\
&= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 \\
&= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta})]^2 - 2E(\hat{\theta})\theta + \theta^2 \\
&= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2
\end{aligned}$$

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean squared error:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\
&= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 \\
&= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta})]^2 - 2E(\hat{\theta})\theta + \theta^2 \\
&= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\
&= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
\end{aligned}$$

Mean Squared Error

How do we combine bias and variance into a single measure of "goodness"?

Mean squared error:

$$\begin{aligned}
\text{MSE}(\hat{\theta}) &= E((\hat{\theta} - \theta)^2) \\
&= E(\hat{\theta}^2) - 2E(\hat{\theta})\theta + \theta^2 \\
&= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta})]^2 - 2E(\hat{\theta})\theta + \theta^2 \\
&= E(\hat{\theta}^2) - [E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\
&= \text{Var}(\hat{\theta}) + \text{Bias}(\hat{\theta})^2
\end{aligned}$$

For **unbiased estimators**, then we simply prefer those with **low variance**.

Estimating MSE

When we can estimate MSE using the same tricks as before: **estimating the distribution of $\hat{\theta}$** and then use MC integration.

Estimating MSE

When we can estimate MSE using the same tricks as before: **estimating the distribution of $\hat{\theta}$** and then use MC integration.

```
> mean((median_500 - true_theta)^2)

[1] 0.005119

> mean((trimmed_500 - true_theta)^2)

[1] 0.006115
```

Estimating MSE

When we can estimate MSE using the same tricks as before: **estimating the distribution of $\hat{\theta}$** and then use MC integration.

```
> mean((median_500 - true_theta)^2)

[1] 0.005119

> mean((trimmed_500 - true_theta)^2)

[1] 0.006115
```

Since both were nearly unbiased, this is basically equal to variance.

Confidence Intervals for MSE

It is useful to give statements of how precise our estimates of the MSE are:

```
> t.test((median_500 - true_theta)^2)$conf.int

[1] 0.004648 0.005590
attr(,"conf.level")
[1] 0.95

> t.test((trimmed_500 - true_theta)^2)$conf.int

[1] 0.005594 0.006636
attr(,"conf.level")
[1] 0.95
```

Example: estimating upper bound of uniform distribution

Consider a distribution with an **unknown upper bound** θ given by:

$$f(x) = I(x \in [0, \theta])\frac{1}{\theta}$$

Example: estimating upper bound of uniform distribution

Consider a distribution with an **unknown upper bound** θ given by:

$$f(x) = I(x \in [0, \theta])\frac{1}{\theta}$$

Suppose we have a sample of 20 IID draws from f. How can we **estimate** θ?

- Method of moments estimation: find θ as a function of the moments of the distribution.
- Maximum likelihood estimation: find the θ that maximizes the joint density (likelihood function).

In general, for any uniform distribution on $[a, b]$, the mean is given by $(b - a)/2$ (the midpoint of the range).

In general, for any uniform distribution on $[a, b]$, the mean is given by $(b - a)/2$ (the midpoint of the range).

So for a uniform $X \sim U[0, \theta]$:

$$\mathsf{E}(X) = \frac{\theta}{2}$$

In general, for any uniform distribution on $[a, b]$, the mean is given by $(b - a)/2$ (the midpoint of the range).

So for a uniform $X \sim U[0, \theta]$:

$$\mathrm{E}(X) = \frac{\theta}{2}$$

Solving for θ, gives

$$\theta = 2\mathrm{E}(X)$$

In general, for any uniform distribution on $[a, b]$, the mean is given by $(b - a)/2$ (the midpoint of the range).

So for a uniform $X \sim U[0, \theta]$:

$$\mathrm{E}(X) = \frac{\theta}{2}$$

Solving for θ, gives

$$\theta = 2\mathrm{E}(X) \Rightarrow \tilde{\theta} = 2\bar{X}$$

For IID data, the likelihood in this case will be

$$L(\theta) = \prod_{i=1}^{n} I(x_i \in [0, \theta]) \frac{1}{\theta} = \frac{1}{\theta^n} \prod_{i=1}^{n} I(x_i \in [0, \theta])$$

Maximum likelihood

For IID data, the likelihood in this case will be

$$L(\theta) = \prod_{i=1}^{n} I(x_i \in [0, \theta]) \frac{1}{\theta} = \frac{1}{\theta^n} \prod_{i=1}^{n} I(x_i \in [0, \theta])$$

Notice:

- $\prod_{i=1}^{n} I(x_i \in [0, \theta]) = 0$ **if any** $x_i > \theta$.
- $1/\theta^n$ is **decreasing in** θ.

Maximum likelihood

For IID data, the likelihood in this case will be

$$L(\theta) = \prod_{i=1}^{n} I(x_i \in [0, \theta]) \frac{1}{\theta} = \frac{1}{\theta^n} \prod_{i=1}^{n} I(x_i \in [0, \theta])$$

Notice:

- $\prod_{i=1}^{n} I(x_i \in [0, \theta]) = 0$ **if any** $x_i > \theta$.
- $1/\theta^n$ is **decreasing in** θ.

These two facts imply, the **MLE is the sample maximum**: $\hat{\theta} = \max_i X_i$.

Sampling distribution for $n = 20, \theta = 2$

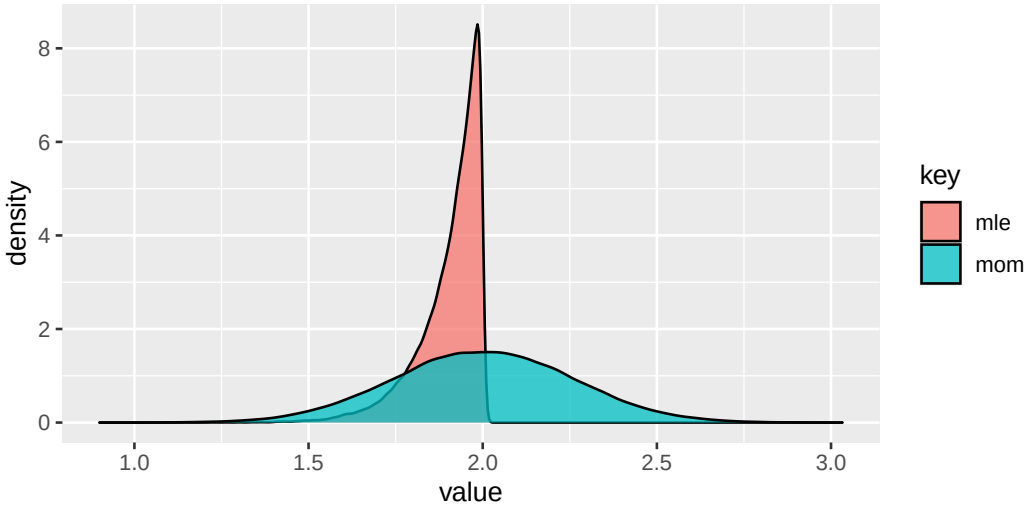
```
> k <- 100000
> n <- 20
> theta <- 2
```

Sampling distribution for $n = 20, \theta = 2$

```
> k <- 100000
> n <- 20
> theta <- 2

> samples <- rerun(k, runif(n, min = 0, max = theta))
> mle_sampling_dist <- map_dbl(samples, max)
> mom_sampling_dist <- map_dbl(samples, ~ 2 * mean(.x))
```

Sampling Distributions



Operating characteristics

Bias:

```
> t.test(mle_sampling_dist - theta)$conf.int[1:2] # 1:2 silence attr
[1] -0.09571 -0.09458
> t.test(mom_sampling_dist - theta)$conf.int[1:2]
[1] -0.001261  0.001942
```

Operating characteristics

Bias:

```
> t.test(mle_sampling_dist - theta)$conf.int[1:2] # 1:2 silence attr
[1] -0.09571 -0.09458
> t.test(mom_sampling_dist - theta)$conf.int[1:2]
[1] -0.001261  0.001942
```

MSE:

```
> t.test((mle_sampling_dist - theta)^2)$conf.int[1:2] # 1:2 silence attr
[1] 0.01718 0.01762
> t.test((mom_sampling_dist - theta)^2)$conf.int[1:2]
[1] 0.06617 0.06733
```

Estimation: Summary

- As with hypothesis tests, our goal is **understand the operating characteristics** of different estimators.

Estimation: Summary

- As with hypothesis tests, our goal is **understand the operating characteristics** of different estimators.
- We can evaluate for **bias** and **variance** of estimators, and combine these with **mean squared error**.

Estimation: Summary

- As with hypothesis tests, our goal is **understand the operating characteristics** of different estimators.
- We can evaluate for **bias** and **variance** of estimators, and combine these with **mean squared error**.
- As with generating **null distributions**, we pick a distribution for the X_i, and the **estimate the distribution of** $\hat{\theta}$ (**sampling distribution**).

Estimation: Summary

- As with hypothesis tests, our goal is **understand the operating characteristics** of different estimators.
- We can evaluate for **bias** and **variance** of estimators, and combine these with **mean squared error**.
- As with generating **null distributions**, we pick a distribution for the X_i, and the **estimate the distribution of** $\hat{\theta}$ (**sampling distribution**).
- Usually we wish to compare estimators or change aspects of the problem to see how the performance changes (ample size, parameter values).

Confidence Intervals

Beyond point estimates

So far, we have looked at the **operating characteristics** of **point estimates**.

Beyond point estimates

So far, we have looked at the **operating characteristics** of **point estimates**.

Point estimates are useful, but **they do not communicate uncertainty**.

Beyond point estimates

So far, we have looked at the **operating characteristics** of **point estimates**.

Point estimates are useful, but **they do not communicate uncertainty**.

When performing hypothesis tests, we started by either **accepting or rejecting the null hypothesis** and then extended this idea to computing **p-values**.

Beyond point estimates

So far, we have looked at the **operating characteristics** of **point estimates**.

Point estimates are useful, but **they do not communicate uncertainty**.

When performing hypothesis tests, we started by either **accepting or rejecting the null hypothesis** and then extended this idea to computing **p-values**.

A similar concept in estimation is to construct a **confidence interval** for θ (target of inference).

Capturing Uncertainty: Confidence Intervals

A $(1-\alpha)\times 100\%$ **confidence interval** (CI) is a pair of **random variables** A and B such that

$$\Pr(A \leq \theta, B \geq \theta) > 1-\alpha \quad \text{and} \quad \Pr(A \leq B) = 1$$

Capturing Uncertainty: Confidence Intervals

A $(1-\alpha)\times 100\%$ **confidence interval** (CI) is a pair of **random variables** A and B such that

$$\Pr(A \leq \theta, B \geq \theta) > 1-\alpha \quad \text{and} \quad \Pr(A \leq B) = 1$$

Equivalent notation:

$$\Pr(\theta \in [A, B]) > 1-\alpha$$

Capturing Uncertainty: Confidence Intervals

A $(1-\alpha)\times 100\%$ **confidence interval** (CI) is a pair of **random variables** A and B such that

$$\Pr(A \leq \theta, B \geq \theta) > 1-\alpha \quad \text{and} \quad \Pr(A \leq B) = 1$$

Equivalent notation:

$$\Pr(\theta \in [A, B]) > 1-\alpha$$

Confidence intervals are sometimes called **interval estimators** because A and B are typically functions of the data, $X_1, \ldots, X_n$:

$$A = A(X_1, \ldots, X_n) \quad B = B(X_1, \ldots, X_n)$$

Confidence interval construction

The notation $A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)$ has a natural connection to **test statistics** and highlights one way of construction confidence intervals: **find the set of null hypotheses not rejected at the α-level**.

Confidence interval construction

The notation $A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)$ has a natural connection to **test statistics** and highlights one way of construction confidence intervals: **find the set of null hypotheses not rejected at the α-level**.

$$A = \inf\{\theta_0 : T(X_1, \ldots, X_n) \notin \mathcal{R}(\theta_0, \alpha)\}$$
$$B = \sup\{\theta_0 : T(X_1, \ldots, X_n) \notin \mathcal{R}(\theta_0, \alpha)\}$$

where $\mathcal{R}(\theta_0, \alpha)$ is the rejection region for T when $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

Confidence interval construction

The notation $A(X_1, \ldots, X_n), B(X_1, \ldots, X_n)$ has a natural connection to **test statistics** and highlights one way of construction confidence intervals: **find the set of null hypotheses not rejected at the α-level**.

$$A = \inf\{\theta_0 : T(X_1, \ldots, X_n) \notin \mathcal{R}(\theta_0, \alpha)\}$$
$$B = \sup\{\theta_0 : T(X_1, \ldots, X_n) \notin \mathcal{R}(\theta_0, \alpha)\}$$

where $\mathcal{R}(\theta_0, \alpha)$ is the rejection region for T when $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

This is called the "test inversion" method of CI construction.

Connections to hypothesis tests

Many of the concepts from hypothesis tests have direct analogs to CIs (no matter how they are constructed):

- Type I error: The **confidence coefficient/coverage** is $\Pr(\theta \in [A, B])$, which is greater than $1 - \alpha$.

Connections to hypothesis tests

Many of the concepts from hypothesis tests have direct analogs to CIs (no matter how they are constructed):

- Type I error: The **confidence coefficient/coverage** is $\Pr(\theta \in [A, B])$, which is greater than $1 - \alpha$.
  In other words, the probability of $[A, B]$ **excluding** θ is no more than α.

Connections to hypothesis tests

Many of the concepts from hypothesis tests have direct analogs to CIs (no matter how they are constructed):

- Type I error: The **confidence coefficient/coverage** is $\Pr(\theta \in [A, B])$, which is greater than $1 - \alpha$.
  In other words, the probability of $[A, B]$ **excluding** θ is no more than α.
- Power: No matter what θ is, we want $[A, B]$ to cover it.

Connections to hypothesis tests

Many of the concepts from hypothesis tests have direct analogs to CIs (no matter how they are constructed):

- Type I error: The **confidence coefficient/coverage** is $\Pr(\theta \in [A, B])$, which is greater than $1 - \alpha$.
  In other words, the probability of $[A, B]$ **excluding** θ is no more than α.
- Power: No matter what θ is, we want $[A, B]$ to cover it.
  Short intervals will exclude more false null hypotheses, so we want $E(B - A)$ (**expected length**) to be small.

Connections to "estimator $\pm c$"

For $X_i \sim N(\mu, \sigma^2)$, independent with σ^2 known, let's test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

at the α level using $\bar{X}$ as the statistic.

Connections to "estimator $\pm c$"

For $X_i \sim N(\mu, \sigma^2)$, independent with σ^2 known, let's test

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0$$

at the α level using $\bar{X}$ as the statistic.

We accept the null if

$$-\Phi(1-\alpha/2) \leq \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \Phi(1-\alpha/2)$$

Solving for μ_0

Solving for μ_0 gives us

$$\bar{X} - \Phi(1-\alpha/2)(\sigma/\sqrt{n}) \leq \mu_0 \leq \bar{X} + \Phi(1-\alpha/2)(\sigma/\sqrt{n})$$

or

$$\mu_0 \in \bar{X} \pm \Phi(1-\alpha/2)(\sigma/\sqrt{n})$$

Solving for μ_0

Solving for μ_0 gives us

$$\bar{X} - \Phi(1-\alpha/2)(\sigma/\sqrt{n}) \leq \mu_0 \leq \bar{X} + \Phi(1-\alpha/2)(\sigma/\sqrt{n})$$

or

$$\mu_0 \in \bar{X} \pm \Phi(1-\alpha/2)(\sigma/\sqrt{n})$$

The "plus-minus" type intervals show up for **shift parameters** that change the center of the sampling distribution but not the variance or other properties.

Example: Coverage and Expected Length $X_i \sim N(\mu, \sigma^2)$

Suppose we believe we know σ^2 and

$$X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

we want to use $\bar{X}$ to estimate μ.

Example: Coverage and Expected Length $X_i \sim N(\mu, \sigma^2)$

Suppose we believe we know σ^2 and

$$X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

we want to use $\bar{X}$ to estimate μ.

We will use two facts (without proof):

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

$$\frac{\bar{X} - \mu}{\hat{s}/\sqrt{n}} \sim t(n-1), \quad \hat{s} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

Example: Coverage and Expected Length $X_i \sim N(\mu, \sigma^2)$

Suppose we believe we know σ^2 and

$$X_i \overset{\text{iid}}{\sim} N(\mu, \sigma^2)$$

we want to use $\bar{X}$ to estimate μ.

We will use two facts (without proof):

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

$$\frac{\bar{X} - \mu}{\hat{s}/\sqrt{n}} \sim t(n-1), \quad \hat{s} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}$$

This leads to two kinds of $(1-\alpha) \times 100\%$ confidence intervals:

$$\bar{X} \pm z_{\alpha/2}(\sigma/\sqrt{n}), \quad \bar{X} \pm t_{\alpha/2}(n-1)(\hat{s}/\sqrt{n})$$

Simulation to assess coverage and average width

```
> alpha <- 0.05
> k <- 10000
> n <- 20
> mu <- 2 # unknown truth
> sigma2 <- 10 ; sigma <- sqrt(sigma2) # known variance
```

Simulation to assess coverage and average width

```
> alpha <- 0.05
> k <- 10000
> n <- 20
> mu <- 2 # unknown truth
> sigma2 <- 10 ; sigma <- sqrt(sigma2) # known variance

> samples_norm <- rerun(k, rnorm(n, mean = mu, sd = sigma))
> cis_norm <- map(samples_norm,
+                 ~ mean(.x) + c(-1, 1) * qnorm(1 - alpha/2) * (sigma / sq
> cis_t <- map(samples_norm,
+              ~ mean(.x) + c(-1, 1) * qt(1 - alpha/2, df = n - 1) * (sd(.x
```

Estimating coverage

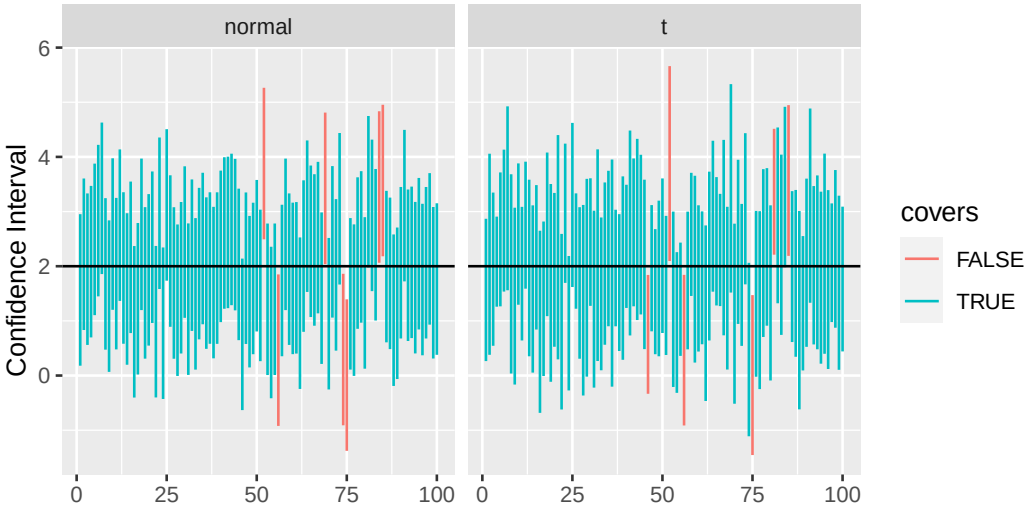
Coverage is $P(A \leq \theta \leq B)$:

```
> map_dbl(cis_norm, ~ .x[1] <= mu && mu <= .x[2]) %>% mean
[1] 0.9516
> map_dbl(cis_t, ~ .x[1] <= mu && mu <= .x[2]) %>% mean
[1] 0.9512
```

Estimating expected width

```
> map_dbl(cis_norm, ~ .x[2] - .x[1]) %>% mean
[1] 2.772
> map_dbl(cis_t, ~ .x[2] - .x[1]) %>% mean
[1] 2.918
```

Visual interpretation



Binomial proportion CI

We've already seen that if we are estimating

$$\theta = P(X \leq t)$$

using

$$\hat{\theta} = \frac{1}{n} \sum_{i=1} I(X_i \leq t)$$

the variables $Y_i = I(X_i \leq t)$ are **Bernoulli with** $P(Y_i = 1) = \theta$.

Binomial proportion CI

We've already seen that if we are estimating

$$\theta = P(X \leq t)$$

using

$$\hat{\theta} = \frac{1}{n} \sum_{i=1} I(X_i \leq t)$$

the variables $Y_i = I(X_i \leq t)$ are **Bernoulli with** $P(Y_i = 1) = \theta$.

Using the central limit theorem, if we have a large sample, then

$$\bar{Y} \approx N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

Binomial proportion continued

Again,

$$\bar{Y} \approx N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

so by our previous analysis the interval:

$$\bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}$$

would be a $100 \times (1-\alpha)$ **confidence interval for** θ.

Binomial proportion continued

Again,

$$\bar{Y} \approx N\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

so by our previous analysis the interval:

$$\bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\theta(1-\theta)}{n}}$$

would be a $100 \times (1-\alpha)$ **confidence interval for** θ.

Of course, we don't know θ, but we can stick in an **estimate**:

$$\bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}}$$

This is the standard confidence interval for a proportion.

95% CI for $P(X > 1)$, $X \sim Cauchy(0)$

```
> n <- 10000
> x <- rcauchy(n)
> y <- x > 1
> ybar <- mean(y)
```

95% CI for $P(X > 1)$, $X \sim Cauchy(0)$

```
> n <- 10000
> x <- rcauchy(n)
> y <- x > 1
> ybar <- mean(y)

> (norm_ci <- ybar + c(-1,1) * qnorm(0.975) * sqrt(ybar * (1 - ybar) / n))

[1] 0.2504 0.2676
```

Test inversion method

Let's use the duality of hypothesis tests to create another interval.

Test inversion method

Let's use the duality of hypothesis tests to create another interval.

If we were testing $H_0 : P(X > 1) = \theta_0$, the null hypothesis tells us:

$$\sum_{i=1}^{n} I(X_i > 1) = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, \theta_0)$$

Test inversion method

Let's use the duality of hypothesis tests to create another interval.

If we were testing $H_0 : P(X > 1) = \theta_0$, the null hypothesis tells us:

$$\sum_{i=1}^{n} I(X_i > 1) = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, \theta_0)$$

Let a_0 be the $\alpha/2$ quantile for $\sum_{i=1}^{n} Y_i$ and b_0 be the smallest value such that $P(\sum Y_i \geq b_0) \leq \alpha/2$.

Test inversion method

Let's use the duality of hypothesis tests to create another interval.

If we were testing $H_0 : P(X > 1) = \theta_0$, the null hypothesis tells us:

$$\sum_{i=1}^{n} I(X_i > 1) = \sum_{i=1}^{n} Y_i \sim \text{Binomial}(n, \theta_0)$$

Let a_0 be the $\alpha/2$ quantile for $\sum_{i=1}^{n} Y_i$ and b_0 be the smallest value such that $P(\sum Y_i \geq b_0) \leq \alpha/2$.

We would accept θ_0 if either

$$a_0 < \sum_{i=1}^{n} Y_i < b_0$$

Finding a_0s and b_0s

We will search over a range of possible θ_0 values:

```
> theta_0 <- seq(0, 1, length.out = 1000)
```

Finding a_0s and b_0s

We will search over a range of possible θ_0 values:

```
> theta_0 <- seq(0, 1, length.out = 1000)
```

For each, compute the bounds:

```
> a0 <- qbinom(0.025, size = n, prob = theta_0)
> b0 <- qbinom(0.025, size = n, prob = theta_0,
+               lower.tail = FALSE) - 1
```

Reject or Accept θ_0

Now see for which θ_0 we would accept:

```
> w <- sum(y)
> range(theta_0[a0 < w & w < b0])
```

```
[1] 0.2513 0.2673
```

Reject or Accept θ_0

Now see for which θ_0 we would accept:

```
> w <- sum(y)
> range(theta_0[a0 < w & w < b0])
```

```
[1] 0.2513 0.2673
```

Which shows that the Normal approximation is quite good a $n = 10000$:

```
[1] 0.2504 0.2676
```

Binomial proportions intervals

R provides three methods for computing binomial confidence intervals:

- `t.test`: Using a standard normal approximation
- `binom.test`: Pearson-Clapper interval (similar to exact interval earlier)
- `prop.test`: Wilson "score" interval

Binomial proportions intervals

R provides three methods for computing binomial confidence intervals:

- `t.test`: Using a standard normal approximation
- `binom.test`: Pearson-Clapper interval (similar to exact interval earlier)
- `prop.test`: Wilson "score" interval

Which has proper coverage? Smallest size?

MC Setup

```
> theta <- 0.25
> n <- 20
> k <- 1000
> xs <- rbinom(k, size = n, prob =  theta)
```

```
> tints <- map(xs, function(x) {
+     if (x == 0 || x == n) {
+         return(c(0, 1)) # can't estimate
+     } else {
+        return(t.test(c(rep(1, x), rep(0, n - x)))$conf.int)
+     }
+ })
```

```
> tints <- map(xs, function(x) {
+     if (x == 0 || x == n) {
+         return(c(0, 1)) # can't estimate
+     } else {
+        return(t.test(c(rep(1, x), rep(0, n - x)))$conf.int)
+     }
+ })

> bints <- map(xs, ~ binom.test(.x, n)$conf.int)
> pints <- map(xs, ~ prop.test(.x, n)$conf.int)
```

Confidence coefficient

Recall, the **confidence coefficient** is defined as

$$P(A \leq \theta, \theta \leq B)$$

```
> cover <- function(x) { x[1] <= theta && theta <= x[2] }
> (tcover <- map_dbl(tints, cover) %>% mean)

[1] 0.895

> (bcover <- map_dbl(bints, cover) %>% mean)

[1] 0.961

> (pcover <- map_dbl(pints, cover) %>% mean)

[1] 0.981
```

Expected Width

```
> (twidth <- map_dbl(tints, diff) %>% mean)
[1] 0.4053
> (bwidth <- map_dbl(bints, diff) %>% mean)
[1] 0.3944
> (pwidth <- map_dbl(pints, diff) %>% mean)
[1] 0.3893
```

Binomial confidence interval routines

- The t-test method tends to **undercover** and has the **largest intervals**. Has computational issues when $X = 0$.

Binomial confidence interval routines

- The t-test method tends to **undercover** and has the **largest intervals**. Has computational issues when $X = 0$.
- The Pearson-Clopper method is **slightly conservative** (i.e., overcovers), which is reflected by having slightly higher average width.

Binomial confidence interval routines

- The t-test method tends to **undercover** and has the **largest intervals**. Has computational issues when $X = 0$.
- The Pearson-Clopper method is **slightly conservative** (i.e., overcovers), which is reflected by having slightly higher average width.
- The score test inversion method has the **smallest intervals**, but also has **good coverage**.

Binomial confidence interval routines

- The t-test method tends to **undercover** and has the **largest intervals**. Has computational issues when $X = 0$.
- The Pearson-Clopper method is **slightly conservative** (i.e., overcovers), which is reflected by having slightly higher average width.
- The score test inversion method has the **smallest intervals**, but also has **good coverage**.
- Important: these conclusions are for $n = 20$ and $\theta = 0.25$; other values might have other conclusions (e.g., when $\theta \approx 1$)

Confidence Intervals Summary

- A confidence interval is a **pair of statistics** A and B with

$$P(A \leq \theta, B \geq \theta) \geq 1 - \alpha$$

Confidence Intervals Summary

- A confidence interval is a **pair of statistics** A and B with

$$P(A \leq \theta, B \geq \theta) \geq 1 - \alpha$$

- There is a **duality** between hypothesis tests and confidence intervals. One way to create intervals is to **invert a set of hypothesis tests**.

Confidence Intervals Summary

- A confidence interval is a **pair of statistics** A and B with

$$P(A \leq \theta, B \geq \theta) \geq 1 - \alpha$$

- There is a **duality** between hypothesis tests and confidence intervals. One way to create intervals is to **invert a set of hypothesis tests**.
- Confidence interval interpretation: set of hypotheses **not rejected** at the α level.

Confidence Intervals Summary

- A confidence interval is a **pair of statistics** A and B with

$$P(A \leq \theta, B \geq \theta) \geq 1 - \alpha$$

- There is a **duality** between hypothesis tests and confidence intervals. One way to create intervals is to **invert a set of hypothesis tests**.
- Confidence interval interpretation: set of hypotheses **not rejected** at the α level.
- As with Type I error and power, we can **investigate the operating characteristics** of confidence intervals.

Confidence Intervals Summary

- A confidence interval is a **pair of statistics** A and B with

$$P(A \leq \theta, B \geq \theta) \geq 1 - \alpha$$

- There is a **duality** between hypothesis tests and confidence intervals. One way to create intervals is to **invert a set of hypothesis tests**.
- Confidence interval interpretation: set of hypotheses **not rejected** at the α level.
- As with Type I error and power, we can **investigate the operating characteristics** of confidence intervals.
- **Confidence coefficient** (actual probability of including θ, analogous to Type I error)

Confidence Intervals Summary

- A confidence interval is a **pair of statistics** A and B with

$$P(A \leq \theta, B \geq \theta) \geq 1 - \alpha$$

- There is a **duality** between hypothesis tests and confidence intervals. One way to create intervals is to **invert a set of hypothesis tests**.
- Confidence interval interpretation: set of hypotheses **not rejected** at the α level.
- As with Type I error and power, we can **investigate the operating characteristics** of confidence intervals.
- **Confidence coefficient** (actual probability of including θ, analogous to Type I error)
- **Expected width** (ability to exclude incorrect θ, analogous to power)