Lecture 7: Confidence intervals

Jacob M. Montgomery

Quantitative Political Methodology

Lecture 7: Confidence intervals

Class business

- We will be working on CIs today and hypothesis tests on Wednesday
- Next Monday we will talk about posters and do some catch up
- PS2 due next class
- PS3 due on 10/9
- Midterm on 10/11

Learning objectives

- Defining: Estimates, confidence intervals, confidence levels
- Calculating confidence intervals
- Place confidence intervals into the larger story of this class

The story so far . . .

- We are hunting population parameters [μ, σ].

The story so far ...

- We are hunting population parameters [μ, σ].
  - What percentage of Americans approve of President Trump?

The story so far . . .

- We are hunting population parameters [μ, σ].
  - What percentage of Americans approve of President Trump?
  - How much carbon is emmitted from cars on I-64?

The story so far . . .

- We are hunting population parameters [μ, σ].
  - What percentage of Americans approve of President Trump?
  - How much carbon is emmitted from cars on I-64?
  - What is the wage gap for women in America?
- We sample the population and calculate sample statistics [$\bar{y}, s$]

The story so far . . .

- We are hunting population parameters $[\mu, \sigma]$.
  - What percentage of Americans approve of President Trump?
  - How much carbon is emmitted from cars on I-64?
  - What is the wage gap for women in America?
- We sample the population and calculate sample statistics $[\bar{y}, s]$
- Today we are going to show how to use *sample statistics* to estimate *population parameters*.

The story so far . . .

- We are hunting population parameters [μ, σ].
  - What percentage of Americans approve of President Trump?
  - How much carbon is emmitted from cars on I-64?
  - What is the wage gap for women in America?
- We sample the population and calculate sample statistics [$\bar{y}, s$]
- Today we are going to show how to use *sample statistics* to estimate *population parameters*.
- How? Probability theory and sampling distributions.

The story so far . . .

- We are hunting population parameters [μ, σ].
  - What percentage of Americans approve of President Trump?
  - How much carbon is emmitted from cars on I-64?
  - What is the wage gap for women in America?
- We sample the population and calculate sample statistics [$\bar{y}, s$]
- Today we are going to show how to use *sample statistics* to estimate *population parameters*.
- How? Probability theory and sampling distributions.
- $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

The story so far . . .

- We are hunting population parameters $[\mu, \sigma]$.
  - What percentage of Americans approve of President Trump?
  - How much carbon is emmitted from cars on I-64?
  - What is the wage gap for women in America?
- We sample the population and calculate sample statistics $[\bar{y}, s]$
- Today we are going to show how to use *sample statistics* to estimate *population parameters*.
- How? Probability theory and sampling distributions.
- $\bar{y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- This will be our first true statistical inference.

Estimation basics

A ***point estimate*** *is a sample statistics that gives a good guess about a population parameter.*

- **Example**: Point estimation for population mean ($\hat{\mu}$)
  - $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
  - $med(y_1, y_2, \ldots, y_n)$

Estimation basics

*A **point estimate** is a sample statistics that gives a good guess about a population parameter.*

- **Example**: Point estimation for population mean ($\hat{\mu}$)
  - $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
  - $med(y_1, y_2, \ldots, y_n)$
- **Example**: Point estimate for population standard deviation ($\hat{\sigma}$)
  - $S = \sqrt{\frac{\sum(y_i - \bar{y})^2}{n-1}}$

Estimation basics

How do we choose among possible estimators?

We want our estimators to be:

- Unbiased (i.e., accurate),

Estimation basics

How do we choose among possible estimators?

We want our estimators to be:

- Unbiased (i.e., accurate), $E(\hat{\mu}) = \mu$ with repeated sampling
- Efficient (i.e, precise), $\sigma_{\hat{\mu}}$ is small(er)
- Consistent (as $n \to \infty$ then $E(\hat{\mu}) \to \mu$)

Estimation basics

How do we choose among possible estimators?

We want our estimators to be:

- Unbiased (i.e., accurate), $E(\hat{\mu}) = \mu$ with repeated sampling
- Efficient (i.e, precise), $\sigma_{\hat{\mu}}$ is small(er)
- Consistent (as $n \to \infty$ then $E(\hat{\mu}) \to \mu$)

Notes:

- Sometimes there are tradeoffs between these (e.g., median)
- **This** is why there is such a funny equation for S.

Summary

The point estimates for populations parameters μ and σ are:

- denote $\hat{\mu}$ and $\hat{\sigma}$

Summary

The point estimates for populations parameters μ and σ are:

- denote $\hat{\mu}$ and $\hat{\sigma}$
- "best" estimated by $\bar{y}$ and S

Summary

The point estimates for populations parameters μ and σ are:

- denote $\hat{\mu}$ and $\hat{\sigma}$
- "best" estimated by $\bar{y}$ and S
- They are "best" in terms of bias and efficiency.

Summary

The point estimates for populations parameters μ and σ are:

- denote $\hat{\mu}$ and $\hat{\sigma}$
- "best" estimated by $\bar{y}$ and S
- They are "best" in terms of bias and efficiency.

Note: We are not assuming that the population is normal. We are just assuming that our real goal is to find a good estimate of μ and that n is large.

Class discussion

You are the campaign manager for a candidate who is deciding whether or not to publish a new deficit reduction proposal. You commission a poll of voters in the district to find out whether they approve or disapprove of this proposal. Which of the following statements would you find most useful from your pollster?

Class discussion

You are the campaign manager for a candidate who is deciding whether or not to publish a new deficit reduction proposal. You commission a poll of voters in the district to find out whether they approve or disapprove of this proposal. Which of the following statements would you find most useful from your pollster?

1. We can be 25% confident that between 54 and 55 percent of voters approve of the plan.
2. We can be 95% confident that between 48.5 and 59.5 percent of voters approve of the plan.
3. We can be 99% confident that between 45.75 and and 62.25 percent of voters approve of the plan.
4. We can be 100% confident that between 0 and 100 percent of voters approve of the plan.

Confidence intervals: Some terms

A point estimate is OK, but it is not very useful without knowing how much confidence to have it.

Confidence intervals: Some terms

A point estimate is OK, but it is not very useful without knowing how much confidence to have it. Solution – interval estimation.

Confidence intervals: Some terms

A point estimate is OK, but it is not very useful without knowing how much confidence to have it. Solution – interval estimation.

> *A **confidence interval** for a population parameter is a range of numbers within which a parameter is believed to fall.*
>
> *The **confidence coefficient** is the probability that an interval would contain the parameter with repeated sampling.*

Confidence intervals: Some terms

A point estimate is OK, but it is not very useful without knowing how much confidence to have it. Solution – interval estimation.

> *A* ***confidence interval*** *for a population parameter is a range of numbers within which a parameter is believed to fall.*
>
> *The* ***confidence coefficient*** *is the probability that an interval would contain the parameter with repeated sampling.*

- 0.95 $\rightarrow$ 95% confidence interval

Confidence intervals: Some terms

A point estimate is OK, but it is not very useful without knowing how much confidence to have it. Solution – interval estimation.

> *A **confidence interval** for a population parameter is a range of numbers within which a parameter is believed to fall.*
>
> *The **confidence coefficient** is the probability that an interval would contain the parameter with repeated sampling.*

- 0.95 $\rightarrow$ 95% confidence interval
- 0.70 $\rightarrow$ 70% confidence interval

Confidence interval for population means (large samples)

We can use the sampling distribution of $\bar{y}$ (assuming a large sample) to calculate a confidence interval for the population mean.

- Parameter: μ
- Estimator: $\hat{\mu} =$

Confidence interval for population means (large samples)

We can use the sampling distribution of $\bar{y}$ (assuming a large sample) to calculate a confidence interval for the population mean.

- Parameter: μ
- Estimator: $\hat{\mu} = \bar{y} \sim$

Confidence interval for population means (large samples)

We can use the sampling distribution of $\bar{y}$ (assuming a large sample) to calculate a confidence interval for the population mean.

- Parameter: μ
- Estimator: $\hat{\mu} = \bar{y} \sim N(\mu_{\bar{y}}, \sigma_{\bar{y}})$
- Remember that $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}}$ and

Confidence interval for population means (large samples)

We can use the sampling distribution of $\bar{y}$ (assuming a large sample) to calculate a confidence interval for the population mean.

- Parameter: μ
- Estimator: $\hat{\mu} = \bar{y} \sim N(\mu_{\bar{y}}, \sigma_{\bar{y}})$
- Remember that $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}}$ and $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$

Basic idea:

- We plug in the **estimated** value of σ to get $\hat{\sigma}_{\bar{y}}$.

Basic idea:

- We plug in the **estimated** value of σ to get $\hat{\sigma}_{\bar{y}}$.
- We use $\bar{y}$ to **estimate** μ, which is sometimes denoted $\hat{\mu}$

Basic idea:

- We plug in the **estimated** value of σ to get $\hat{\sigma}_{\bar{y}}$.
- We use $\bar{y}$ to **estimate** μ, which is sometimes denoted $\hat{\mu}$
- Now we have an **estimated** sampling distribution, $N(\bar{y}, \hat{\sigma}_{\bar{y}})$

Basic idea:

- We plug in the **estimated** value of σ to get $\hat{\sigma}_{\bar{y}}$.
- We use $\bar{y}$ to **estimate** μ, which is sometimes denoted $\hat{\mu}$
- Now we have an **estimated** sampling distribution, $N(\bar{y}, \hat{\sigma}_{\bar{y}})$
  - We use our knowledge of the normal distribution to find a CI
  - E.g., we want 2.5% of the probability to be outside of our interval on each side.

Steps:

1. Calculate $\bar{y}$
2. Calculate S

Steps:

1. Calculate $\bar{y}$
2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$

Steps:

1. Calculate $\bar{y}$
2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$
3. How much area do we need under the curve to the right?

Steps:

1. Calculate $\bar{y}$
2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$
3. How much area do we need under the curve to the right? (1-Confidence Coefficient)/2.

Steps:

1. Calculate $\bar{y}$
2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$
3. How much area do we need under the curve to the right? (1-Confidence Coefficient)/2. How much area do we need under the curve to the left?

Steps:

1. Calculate $\bar{y}$
2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \dfrac{S}{\sqrt{n}}$
3. How much area do we need under the curve to the right? (1-Confidence Coefficient)/2. How much area do we need under the curve to the left? (Confidence Coefficient)/2.
4. Find the z-score associated with that number.

Steps:

1. Calculate $\bar{y}$
2. Calculate S and then $\hat{\sigma}_{\bar{y}} = \frac{S}{\sqrt{n}}$
3. How much area do we need under the curve to the right? (1-Confidence Coefficient)/2. How much area do we need under the curve to the left? (Confidence Coefficient)/2.
4. Find the z-score associated with that number.
5. Use these values to calculate $\bar{y} \pm Z \times \hat{\sigma}_{\bar{y}}$

Exercise: If $\bar{y} = 9.6$, $n = 100$, and $S = 4$, what is the 99% confidence interval for μ?

1. Find values for L and R on the *standard normal distribution* such that:

$$Pr(L \leq \mu \leq R) = 0.95$$

1. Find values for L and R on the *standard normal distribution* such that:

$$Pr(L \leq \mu \leq R) = 0.95$$

2. Plug in our estimates, and see that $\bar{y} \sim N(\mu, \sigma_{\bar{y}}) \approx N(\bar{y}, \frac{S}{\sqrt{n}})$
3. $L = \bar{y} - (Z \times \hat{\sigma}_{\bar{y}}), R = \bar{y} + (Z \times \hat{\sigma}_{\bar{y}})$

1. Find values for L and R on the *standard normal distribution* such that:

$$Pr(L \leq \mu \leq R) = 0.95$$

2. Plug in our estimates, and see that $\bar{y} \sim N(\mu, \sigma_{\bar{y}}) \approx N(\bar{y}, \frac{S}{\sqrt{n}})$
3. $L = \bar{y} - (Z \times \hat{\sigma}_{\bar{y}}), R = \bar{y} + (Z \times \hat{\sigma}_{\bar{y}})$
4. Look for $(0.95)/2 = .025$ on the z-table

1. Find values for L and R on the *standard normal distribution* such that:

$$Pr(L \leq \mu \leq R) = 0.95$$

2. Plug in our estimates, and see that $\bar{y} \sim N(\mu, \sigma_{\bar{y}}) \approx N(\bar{y}, \frac{S}{\sqrt{n}})$
3. $L = \bar{y} - (Z \times \hat{\sigma}_{\bar{y}}), R = \bar{y} + (Z \times \hat{\sigma}_{\bar{y}})$
4. Look for $(0.95)/2 = .025$ on the z-table 1.96

Answer: $\bar{y} \pm 1.96 \times \hat{\sigma}_{\bar{y}}$

1. Find values for L and R on the *standard normal distribution* such that:
$$Pr(L \leq \mu \leq R) = 0.95$$

2. Plug in our estimates, and see that $\bar{y} \sim N(\mu, \sigma_{\bar{y}}) \approx N(\bar{y}, \frac{S}{\sqrt{n}})$
3. $L = \bar{y} - (Z \times \hat{\sigma}_{\bar{y}}), R = \bar{y} + (Z \times \hat{\sigma}_{\bar{y}})$
4. Look for $(0.95)/2 = .025$ on the z-table 1.96

Answer: $\bar{y} \pm 1.96 \times \hat{\sigma}_{\bar{y}} = 9.6 \pm 1.96 \times \frac{4}{10} = [8.816, 10.384]$

Reprise:

How to calculate a confidence interval:

1. Calculate $\bar{y}$ and $\sigma_{\bar{y}} = \frac{S}{\sqrt{n}}$
2. How much area do we need under the curve to the left?
   - Example: For a 95% confidence interval we need .025 under the curve.
   - (confidence coefficient)/2
3. Find the z-score associated with that number
4. Use these values to calculate $\bar{y} \pm Z \times \sigma_{\bar{y}}$

Reprise:

How to calculate a confidence interval:

1. Calculate $\bar{y}$ and $\sigma_{\bar{y}} = \frac{S}{\sqrt{n}}$
2. How much area do we need under the curve to the left?
   - Example: For a 95% confidence interval we need .025 under the curve.
   - (confidence coefficient)/2
3. Find the z-score associated with that number
4. Use these values to calculate $\bar{y} \pm Z \times \sigma_{\bar{y}}$

Team time!