Precept 3: Random Samples

Soc 500: Applied Social Statistics

Simone Zhang[1]

Princeton University

September 2016

[1]This set of slides draws on material from lecture slides, Matt Blackwell, Justin Grimmer and Jens Hainmueller.

Logistics

- Reactions to the problem set?
- Solutions will be posted at 11:00
- New problem set is out. Any questions so far?

Today's Tasks

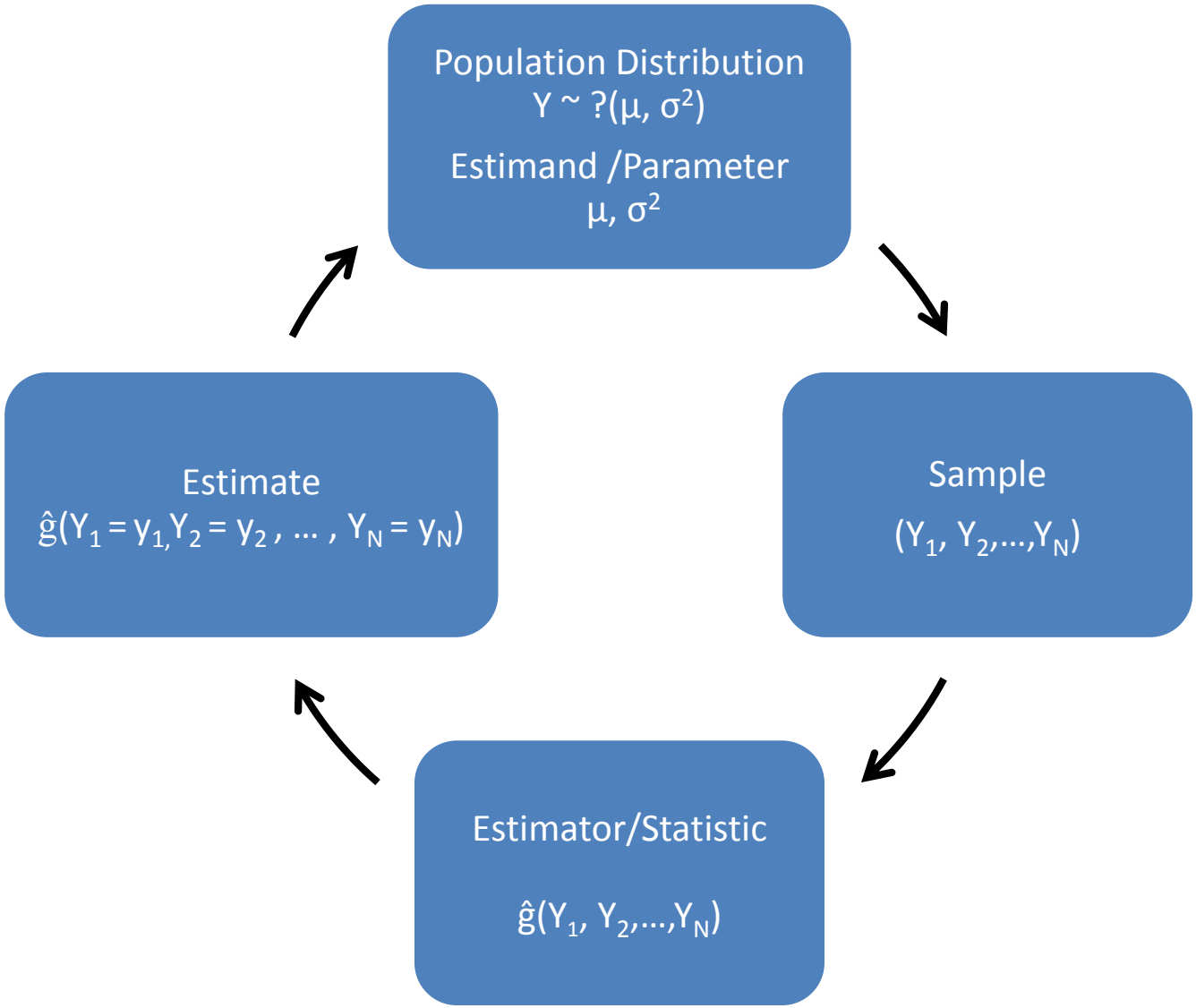
- Review material presented in lecture
  - sampling
  - estimators (and their properties)
  - CLT, t-distribution
  - confidence intervals
- Cover computational examples
  - rnorm(), pnorm(), qnorm()
  - drawing random samples
  - generating CIs

The Big Picture

In studying the world, we usually run into the following challenge:

- There's some quantity of interest we want to know about a population, the **estimand**, which we consider to have a "true" value
  e.g. what's the average height of a penguin?
- Ideally, we'd like to collect information on every member of the population. But usually, that's not possible. Instead we collect data on a random sample drawn from the population.
  e.g. measure the heights of a random sample of penguins
- This week is about understanding how to infer the "true" population-level distribution from the data we do have in a sample
  e.g. by calculating the mean height of penguins in our sample

An Overview


Population Distribution
$Y \sim ?(\mu, \sigma^2)$
Estimand /Parameter
μ, σ^2
Sample
$(Y_1, Y_2, \ldots, Y_N)$
Estimator/Statistic
$\hat{g}(Y_1, Y_2, \ldots, Y_N)$
Estimate
$\hat{g}(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N)$

Estimands, Estimators, and Estimates

Estimands, Estimators, and Estimates

- The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

Estimands, Estimators, and Estimates

- The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N} \sum_{i=1}^{N} y_i$

Estimands, Estimators, and Estimates

- The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N} \sum_{i=1}^{N} y_i$

- **Estimators** are functions of sample data (i.e. **statistics**) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)

Estimands, Estimators, and Estimates

- The goal of statistical inference is to learn about the unobserved population distribution, which can be characterized by **parameters**.

- **Estimands** are the parameters that we aim to estimate. Often written with greek letters (e.g. μ, θ, population mean) : $\frac{1}{N} \sum_{i=1}^{N} y_i$

- **Estimators** are functions of sample data (i.e. **statistics**) which we use to learn about the estimands. Often denoted with a "hat" (e.g. $\hat{\mu}, \hat{\theta}$)

- **Estimates** are particular values of estimators that are realized in a given sample (e.g. sample mean)

Clarifying Notation and Terms You'll Encounter

- Estimand / Population Parameter (Theoretical)
  - Population mean: $\mu = E[X] = \frac{1}{N}\sum_{i=1}^{N} X_i$
  - Population variance: $\sigma^2 = E[(X - E(X))^2] = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$
- Estimator (Links Data to Estimand)
  - Estimator for population mean: $\hat{\mu}$
  - Estimator for population variance: $\hat{\sigma}^2$
- Estimate (Calculated from a Given Sample), e.g.
  - Sample mean: $\overline{X}_n = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$
  - Sample variance: $s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X}_n)^2$

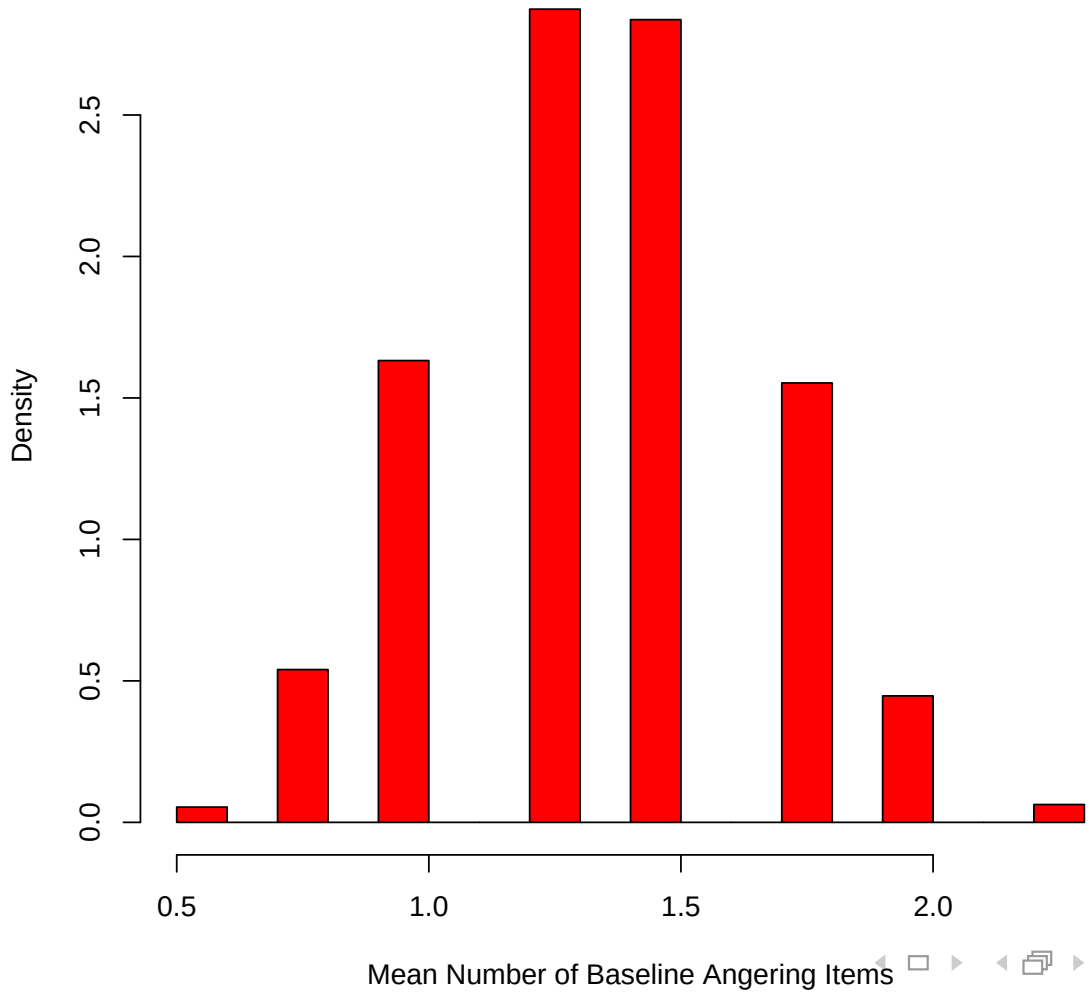
Sampling Distribution

Consider using the sample mean as an estimator for the "true" mean: $\hat{\mu} = \overline{X}_n$:

- Usually we only ever observe one sample of size n - so we get one value of $\overline{X}_n$
- But consider the hypothetical case that we got 10,000 random samples of size n. By random chance, the samples may look different from each other. Each sample would have its own $\overline{X}_n$
- The sampling distribution of $\overline{X}_n$ gives the probability density of he possible values of $\overline{X}_n$

Sampling Distribution of the Sample Mean

Example:



Other estimators (e.g. sample variance) also have sampling distributions.
We can describe sampling distributions in terms of their center (i.e. mean) and spread (i.e. standard error).

The Central Limit Theorem

The Central Limit Theorem tells us something cool about sample means ($\overline{X}_n$).
From the lecture slides, as n increases, the sampling distribution of $\overline{X}_n$ becomes more bell-shaped. This is the basic implication of the **Central Limit Theorem**:

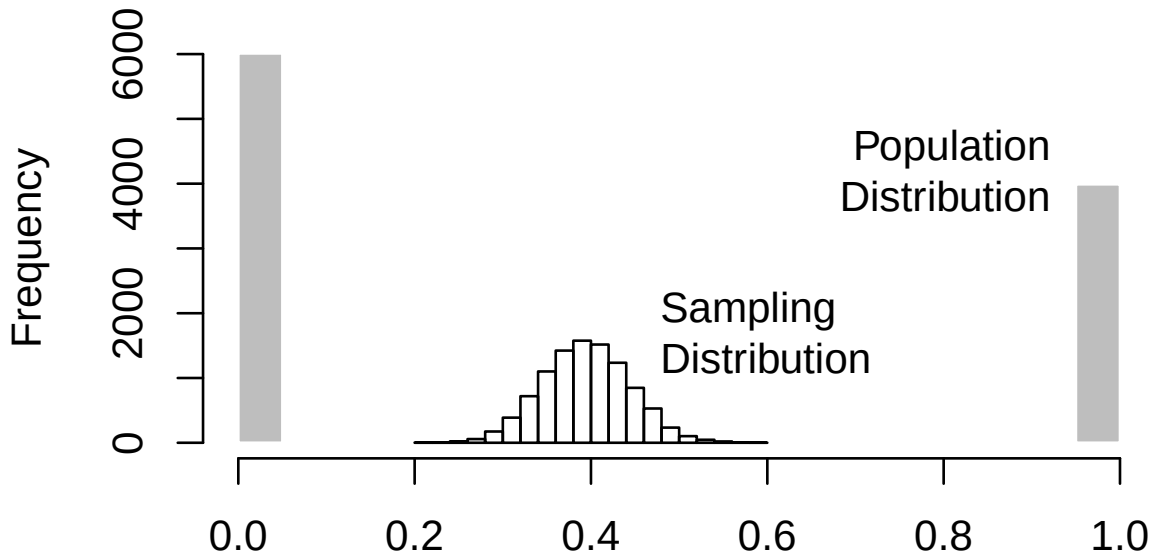
If $X_1, \ldots, X_n \sim_{i.i.d.} ?(\mu, \sigma^2)$ and n is large, then

$$\overline{X}_n \sim_{approx} N(\mu, \frac{\sigma^2}{n})$$

so

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

Population vs. Sampling Distribution



t-distribution

When does the t-distribution come in handy?

- Note the the CLT kicks in asymptotically (as n $\rightarrow \infty$)
- t-distribution useful when we have smaller sample sizes

t-distribution

When does the t-distribution come in handy?

- Note the the CLT kicks in asymptotically (as n $\to \infty$)
- t-distribution useful when we have smaller sample sizes

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim$$

t-distribution

When does the t-distribution come in handy?

- Note the the CLT kicks in asymptotically (as n $\to \infty$)
- t-distribution useful when we have smaller sample sizes

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim ??$$

t-distribution

When does the t-distribution come in handy?

- Note the the CLT kicks in asymptotically (as n $\to \infty$)
- t-distribution useful when we have smaller sample sizes

$$\frac{\overline{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$$

t-distribution

Since we have to estimate σ, the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is still bell-shaped but has fatter tails.

t-distribution

Since we have to estimate σ, the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is still bell-shaped but has fatter tails.

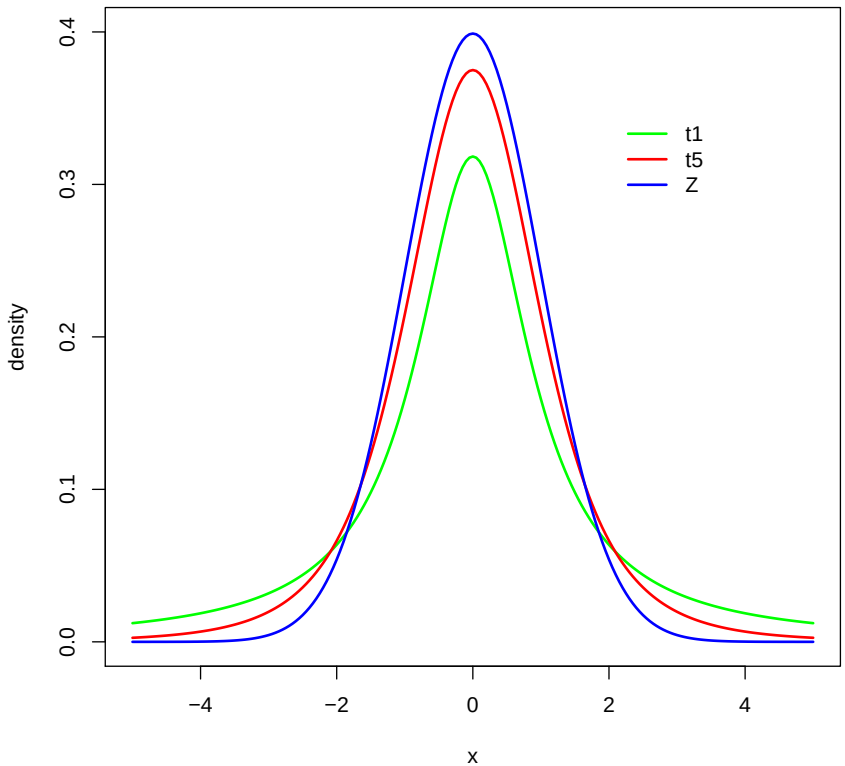
As the sample size increases, our estimates of σ improve and extreme values of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ become less likely.

t-distribution

Since we have to estimate σ, the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is still bell-shaped but has fatter tails.

As the sample size increases, our estimates of σ improve and extreme values of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ become less likely.

Eventually the t distribution converges to the standard normal.

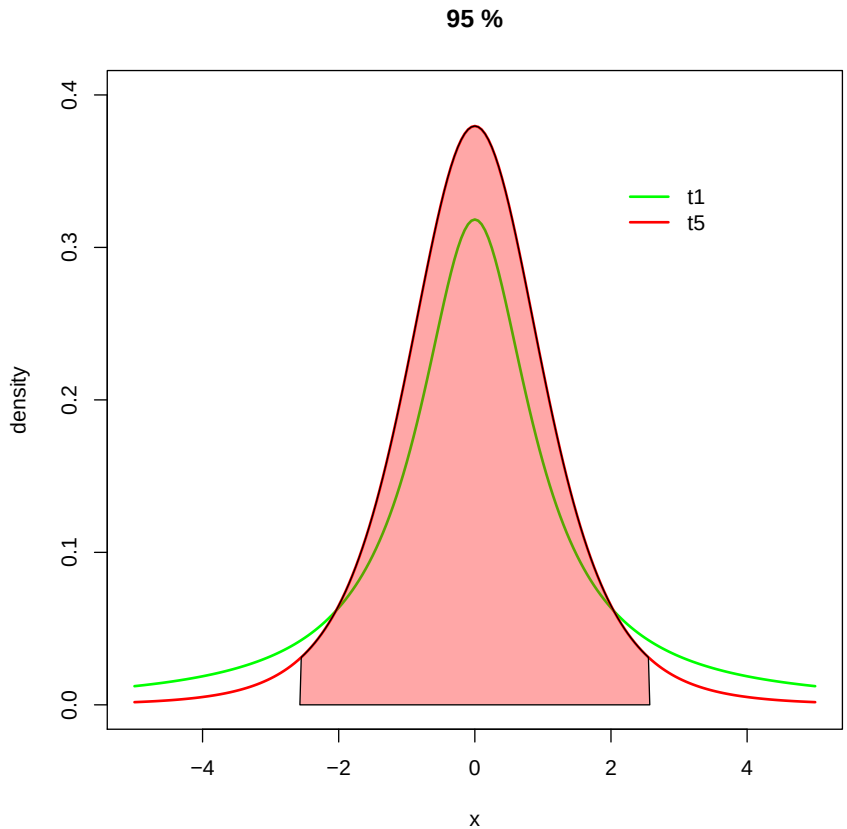


t-distribution

Since we have to estimate σ, the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is still bell-shaped but has fatter tails.

As the sample size increases, our estimates of σ improve and extreme values of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ become less likely.

Eventually the t distribution converges to the standard normal.

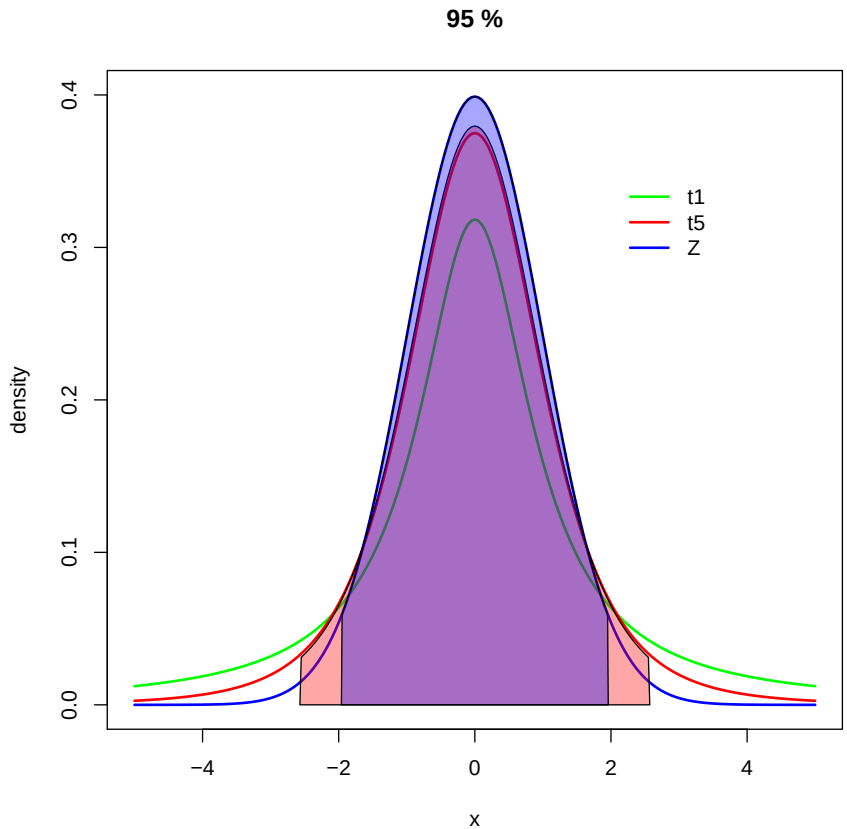


t-distribution

Since we have to estimate σ, the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is still bell-shaped but has fatter tails.

As the sample size increases, our estimates of σ improve and extreme values of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ become less likely.

Eventually the t distribution converges to the standard normal.



Summary of Properties of Estimators

Concept	Criteria	Intuition
Unbiasedness	$E[\hat{\mu}] = \mu$	Right on average
Efficiency	$V[\hat{\mu}_1] < V[\hat{\mu}_2]$	Low variance
Consistency	$\hat{\mu}_n \xrightarrow{p} \mu$	Converge to estimand as $n \to \infty$
Asymptotic Normality	$\hat{\mu}_n \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n})$	Approximately normal in large n

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1-\alpha)\%$$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1-\alpha)\%$$

What is the formula for two-sided confidence intervals?

$$[\overline{X}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1-\alpha)\%$$

What is the formula for two-sided confidence intervals?

$$[\overline{X}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

How do we find $z_{a/2}$? qnorm() : $F^{-1}(p)$
Returns z value at which CDF of Standard Normal equals p

Confidence Intervals

Recall from CLT that $\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$$P\left(-z \leq \frac{\hat{\mu} - \mu}{\hat{SE}[\hat{\mu}]} \leq z \right) = (1-\alpha)\%$$

What is the formula for two-sided confidence intervals?

$$[\overline{X}_n - z_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X}_n + z_{\alpha/2}\frac{s}{\sqrt{n}}]$$

How do we find $z_{a/2}$? qnorm() : $F^{-1}(p)$
Returns z value at which CDF of Standard Normal equals p
What is the width of the confidence interval? $2 * z_{\alpha/2}\frac{s}{\sqrt{n}}$

Confidence Intervals

What about one-sided confidence intervals?

Confidence Intervals

What about one-sided confidence intervals?
An 100(1-α)% upper (one-sided) confidence bound

$$\overline{X}_n + z_\alpha \frac{s}{\sqrt{n}}$$

An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Confidence Intervals

What about one-sided confidence intervals?
An 100(1-α)% upper (one-sided) confidence bound

$$\overline{X}_n + z_\alpha \frac{s}{\sqrt{n}}$$

An 100(1-α)% lower (one-sided) confidence bound

$$\overline{X}_n - z_\alpha \frac{s}{\sqrt{n}}$$

Fulton Data

- Election data from Fulton County, Georgia, aggregated to the precinct level

Table: Fulton Election Data

Variable	Description
precint	precint id
turnout	voter turnout rate
black	percent Black
sex	percent Female
age	mean age
dem	turnout in democratic primary
rep	turnout in republican primary
urban	is the precinct in Atlanta
school	school polling location

Questions?