Gov 50: 18. Estimation: Surveys

Matthew Blackwell

Harvard University

Fall 2018

1/ Today's agenda

Logistics

- HW4 due tonight.

Logistics

- HW4 due tonight.
- Midterm 2 next Thursday.

Logistics

- HW4 due tonight.
- Midterm 2 next Thursday.
    - Same basic structure as last midterm.

Logistics

- HW4 due tonight.
- Midterm 2 next Thursday.
  - Same basic structure as last midterm.
  - Cumulative, but more focused on new material.

Logistics

- HW4 due tonight.
- Midterm 2 next Thursday.
  - Same basic structure as last midterm.
  - Cumulative, but more focused on new material.
  - Review session on Tuesday.

Logistics

- HW4 due tonight.
- Midterm 2 next Thursday.
  - Same basic structure as last midterm.
  - Cumulative, but more focused on new material.
  - Review session on Tuesday.
  - Practice midterm out now.

Where are we? Where are we going?

- Up to now: what kinds of samples should we observe if we know the population distribution?

Where are we? Where are we going?

- Up to now: what kinds of samples should we observe if we know the population distribution?
- Now: what can I learn about the population distribution from my sample.

Where are we? Where are we going?

- Up to now: what kinds of samples should we observe if we know the population distribution?
- Now: what can I learn about the population distribution from my sample.
- Lessons today applicable to most statistical procedures.

How popular is Donald Trump?

- What proportion of the public approves of Trump's job as president?

How popular is Donald Trump?

- What proportion of the public approves of Trump's job as president?
- Latest Gallup poll:

How popular is Donald Trump?

- What proportion of the public approves of Trump’s job as president?
- Latest Gallup poll:
    - Oct. 29th–Nov. 4th

How popular is Donald Trump?

- What proportion of the public approves of Trump's job as president?
- Latest Gallup poll:
  - Oct. 29th–Nov. 4th
  - 1500 adult Americans

How popular is Donald Trump?

- What proportion of the public approves of Trump's job as president?
- Latest Gallup poll:
  - Oct. 29th–Nov. 4th
  - 1500 adult Americans
  - Telephone interviews

How popular is Donald Trump?

- What proportion of the public approves of Trump's job as president?
- Latest Gallup poll:
  - Oct. 29th–Nov. 4th
  - 1500 adult Americans
  - Telephone interviews
  - Approve (40%), Disapprove (54%)

How popular is Donald Trump?

- What proportion of the public approves of Trump's job as president?
- Latest Gallup poll:
  - Oct. 29th–Nov. 4th
  - 1500 adult Americans
  - Telephone interviews
  - Approve (40%), Disapprove (54%)
- What can we learn about Trump approval in the population from this one sample?

2/ Samples and estimators

Samples from the population

- Our focus: simple random sample of size n from some population $Y_1, \ldots, Y_n$

Samples from the population

- Our focus: simple random sample of size n from some population $Y_1, \ldots, Y_n$
  - $\rightsquigarrow$ i.i.d. random variables

Samples from the population

- Our focus: simple random sample of size n from some population $Y_1, \dots, Y_n$
  - $\rightsquigarrow$ i.i.d. random variables
  - e.g.: $Y_i = 1$ if i approves of Trump, $Y_i = 0$ otherwise.

Samples from the population

- Our focus: simple random sample of size n from some population $Y_1, \ldots, Y_n$
  - $\rightsquigarrow$ i.i.d. random variables
  - e.g.: $Y_i = 1$ if i approves of Trump, $Y_i = 0$ otherwise.
- **Statistical inference** is using data to guess something about the population distribution of Y_i.

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution
  - Also called: parameters.

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution
  - Also called: parameters.
- Examples of quantities of interest:

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution
  - Also called: parameters.
- Examples of quantities of interest:
  - $\mu = \mathbb{E}[Y_i]$: the population mean (turnout rate in the population).

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution
  - Also called: parameters.
- Examples of quantities of interest:
  - $\mu = \mathbb{E}[Y_i]$: the population mean (turnout rate in the population).
  - $\sigma^2 = \mathbb{V}[Y_i]$: the population variance.

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution
  - Also called: parameters.
- Examples of quantities of interest:
  - $\mu = \mathbb{E}[Y_i]$: the population mean (turnout rate in the population).
  - $\sigma^2 = \mathbb{V}[Y_i]$: the population variance.
  - $\mu_1 - \mu_0 = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$: the population ATE.

Point estimation

- **Point estimation**: providing a single "best guess" as to the value of some fixed, unknown **quantity of interest**, θ.
  - θ is a feature of the population distribution
  - Also called: parameters.
- Examples of quantities of interest:
  - $\mu = \mathbb{E}[Y_i]$: the population mean (turnout rate in the population).
  - $\sigma^2 = \mathbb{V}[Y_i]$: the population variance.
  - $\mu_1 - \mu_0 = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$: the population ATE.
- These are the things we want to learn about.

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.

- An **estimate** is one particular realization of the estimator

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.

- An **estimate** is one particular realization of the estimator
- Ideally we'd like to know the **estimation error**, $\hat{\theta} - \theta$

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.

- An **estimate** is one particular realization of the estimator
- Ideally we'd like to know the **estimation error**, $\hat{\theta} - \theta$
- Problem: θ is unknown.

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.

- An **estimate** is one particular realization of the estimator
- Ideally we'd like to know the **estimation error**, $\hat{\theta} - \theta$
- Problem: θ is unknown.
- Solution: figure out the properties of $\hat{\theta}$ using probability.

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \dots, Y_n)$.

- An **estimate** is one particular realization of the estimator
- Ideally we'd like to know the **estimation error**, $\hat{\theta} - \theta$
- Problem: θ is unknown.
- Solution: figure out the properties of $\hat{\theta}$ using probability.
  - $\hat{\theta}$ is a r.v. because it is a function of r.v.s.

Estimators

Estimator

An estimator, $\hat{\theta}$, of some parameter θ, is some function of the sample: $\hat{\theta} = h(Y_1, \ldots, Y_n)$.

- An **estimate** is one particular realization of the estimator
- Ideally we'd like to know the **estimation error**, $\hat{\theta} - \theta$
- Problem: θ is unknown.
- Solution: figure out the properties of $\hat{\theta}$ using probability.
  - $\hat{\theta}$ is a r.v. because it is a function of r.v.s.
  - $\leadsto$ $\hat{\theta}$ has a distribution.

3/ Properties of estimators

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump
- There are many different possible estimators:

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump
- There are many different possible estimators:
  - $\hat{\theta} = \overline{Y}_n$ the sample proportion of respondents who support Trump.

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump
- There are many different possible estimators:
  - $\hat{\theta} = \overline{Y}_n$ the sample proportion of respondents who support Trump.
  - $\hat{\theta} = Y_1$ just use the first observation

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump
- There are many different possible estimators:
  - $\hat{\theta} = \overline{Y}_n$ the sample proportion of respondents who support Trump.
  - $\hat{\theta} = Y_1$ just use the first observation
  - $\hat{\theta} = \max(Y_1, \ldots, Y_n)$

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump
- There are many different possible estimators:
  - $\hat{\theta} = \overline{Y}_n$ the sample proportion of respondents who support Trump.
  - $\hat{\theta} = Y_1$ just use the first observation
  - $\hat{\theta} = \max(Y_1, \ldots, Y_n)$
  - $\hat{\theta} = 0.5$ always guess 50% support

Estimating Trump's support

- Parameter θ: **population proportion** of adults who support Trump
- There are many different possible estimators:
  - $\hat{\theta} = \overline{Y}_n$ the sample proportion of respondents who support Trump.
  - $\hat{\theta} = Y_1$ just use the first observation
  - $\hat{\theta} = \max(Y_1, \ldots, Y_n)$
  - $\hat{\theta} = 0.5$ always guess 50% support
- How good are these different estimators?

Survey

- Assume a simple random sample of n voters: $n = 1500$

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump
- Y_i is **Bernoulli** with probability of success p

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump
- Y_i is **Bernoulli** with probability of success p
  - "probability of success" = "probability of randomly selecting a Trump approver"

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump
- Y_i is **Bernoulli** with probability of success p
  - "probability of success" = "probability of randomly selecting a Trump approver"
  - Remember that p is the expectation of Y_i

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump
- Y_i is **Bernoulli** with probability of success p
  - "probability of success" = "probability of randomly selecting a Trump approver"
  - Remember that p is the expectation of Y_i
  - That is, $p = \mathbb{P}(Y_i = 1) = \mathbb{E}(Y_i)$

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump
- Y_i is **Bernoulli** with probability of success p
  - "probability of success" = "probability of randomly selecting a Trump approver"
  - Remember that p is the expectation of Y_i
  - That is, $p = \mathbb{P}(Y_i = 1) = \mathbb{E}(Y_i)$
- Sample proportion is the same as the sample mean:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i = \frac{\text{number who support Trump}}{n}$$

Survey

- Assume a simple random sample of n voters: $n = 1500$
- Define r.v. Y_i for Trump approval:
  - $Y_i = 1 \rightsquigarrow$ respondent i approves of Trump
  - $Y_i = 0 \rightsquigarrow$ respondent i disapproves of Trump
- Y_i is **Bernoulli** with probability of success p
  - "probability of success" = "probability of randomly selecting a Trump approver"
  - Remember that p is the expectation of Y_i
  - That is, $p = \mathbb{P}(Y_i = 1) = \mathbb{E}(Y_i)$
- Sample proportion is the same as the sample mean:

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i = \frac{\text{number who support Trump}}{n}$$

- $\theta = p$ and $\hat{\theta} = \overline{Y}$

Sample mean properties

sample mean = population mean + chance error

Sample mean properties

$$\text{sample mean} = \text{population mean} + \text{chance error}$$
$$\overline{Y} = p + \text{chance error}$$

Sample mean properties

$$\text{sample mean} = \text{population mean} + \text{chance error}$$

$$\overline{Y} = p + \text{chance error}$$

- Remember: the sample mean is a random variable.

Sample mean properties

$$\text{sample mean} = \text{population mean} + \text{chance error}$$

$$\overline{Y} = p + \text{chance error}$$

- Remember: the sample mean is a random variable.
  - Different samples give different sample means.

Sample mean properties

$$\text{sample mean} = \text{population mean} + \text{chance error}$$

$$\overline{Y} = p + \text{chance error}$$

- Remember: the sample mean is a random variable.
  - Different samples give different sample means.
  - Chance error “bumps” sample mean away from population mean

Sample mean properties

$$\text{sample mean} = \text{population mean} + \text{chance error}$$

$$\overline{Y} = p + \text{chance error}$$

- Remember: the sample mean is a random variable.
  - Different samples give different sample means.
  - Chance error "bumps" sample mean away from population mean
- $\rightsquigarrow \overline{Y}$ has a distribution across repeated samples.

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.
  - From last week, $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_i] = p$

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.
  - From last week, $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_i] = p$
  - $\rightsquigarrow$ chance error is 0 on average:

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.
  - From last week, $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_i] = p$
  - $\rightsquigarrow$ chance error is 0 on average:

$$\mathbb{E}[\overline{Y} - p]$$

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.
  - From last week, $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_i] = p$
  - $\rightsquigarrow$ chance error is 0 on average:

$$\mathbb{E}[\overline{Y} - p] = \mathbb{E}[\overline{Y}] - p$$

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.
  - From last week, $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_i] = p$
  - $\rightsquigarrow$ chance error is 0 on average:

$$\mathbb{E}[\overline{Y} - p] = \mathbb{E}[\overline{Y}] - p = 0$$

Central tendency of the sample mean

- Expectation: average of the estimates across repeated samples.
  - From last week, $\mathbb{E}[\overline{Y}] = \mathbb{E}[Y_i] = p$
  - $\rightsquigarrow$ chance error is 0 on average:

$$\mathbb{E}[\overline{Y} - p] = \mathbb{E}[\overline{Y}] - p = 0$$

- **Unbiasedness**: Sample proportion is on average equal to the population proportion.

Spread of the sample mean

- **Standard error**: how big is the chance error on average?

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\overline{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\overline{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

- Problem: we don't know p!

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\bar{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

- Problem: we don't know p!
- Solution: **estimate** the SE:

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\overline{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

- Problem: we don't know p!
- Solution: **estimate** the SE:

$$\sqrt{\widehat{\mathbb{V}}(\overline{Y})} =$$

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\overline{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

- Problem: we don't know p!
- Solution: **estimate** the SE:

$$\sqrt{\widehat{\mathbb{V}}(\overline{Y})} = \sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}}$$

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\overline{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

- Problem: we don't know p!
- Solution: **estimate** the SE:

$$\sqrt{\widehat{\mathbb{V}}(\overline{Y})} = \sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}} = \sqrt{\frac{0.37 \times (1-0.37)}{1500}}$$

Spread of the sample mean

- **Standard error**: how big is the chance error on average?
- We can use a special rule to binary r.v.s:

$$\sqrt{\mathbb{V}(\overline{Y})} = \sqrt{\frac{p(1-p)}{n}}$$

- Problem: we don't know p!
- Solution: **estimate** the SE:

$$\sqrt{\widehat{\mathbb{V}}(\overline{Y})} = \sqrt{\frac{\overline{Y}(1-\overline{Y})}{n}} = \sqrt{\frac{0.37 \times (1-0.37)}{1500}} \approx 0.012$$

4/ Confidence intervals

Confidence intervals

- Awesome: sample proportion is correct on average.

Confidence intervals

- Awesome: sample proportion is correct on average.
- Awesomer: get an range of plausible values.

Confidence intervals

- Awesome: sample proportion is correct on average.
- Awesomer: get an range of plausible values.
- **Confidence interval**: way to construct an interval that will contain the true value in some fixed proportion of repeated samples.

CLT

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?

CLT

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
  - Find a range of plausible chance errors and add them to $\overline{Y}$

CLT

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
  - Find a range of plausible chance errors and add them to $\overline{Y}$
- Central limit theorem:

$$\overline{Y} \overset{\text{approx}}{\sim} N\left(\mathbb{E}(Y_i), \frac{\mathbb{V}(Y_i)}{n}\right)$$

CLT

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
  - Find a range of plausible chance errors and add them to $\overline{Y}$
- Central limit theorem:

$$\overline{Y} \overset{\text{approx}}{\sim} N\left(\mathbb{E}(Y_i), \frac{\mathbb{V}(Y_i)}{n}\right)$$

- In this case:

$$\overline{Y} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

CLT

$$\overline{Y} - p = \text{chance error}$$

- How can we figure out a range of plausible chance errors?
  - Find a range of plausible chance errors and add them to $\overline{Y}$
- Central limit theorem:

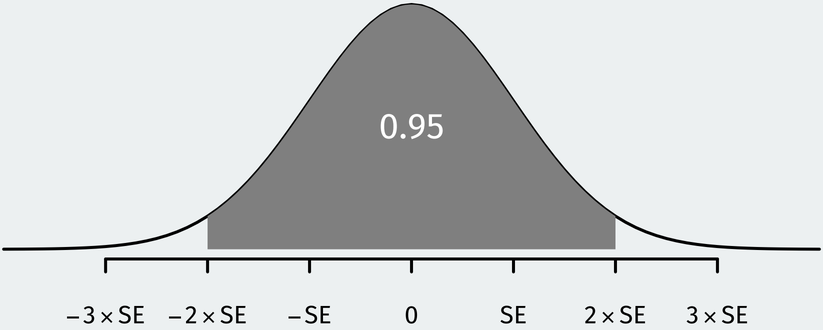
$$\overline{Y} \overset{\text{approx}}{\sim} N\left(\mathbb{E}(Y_i), \frac{\mathbb{V}(Y_i)}{n}\right)$$

- In this case:

$$\overline{Y} \overset{\text{approx}}{\sim} N\left(p, \frac{p(1-p)}{n}\right)$$

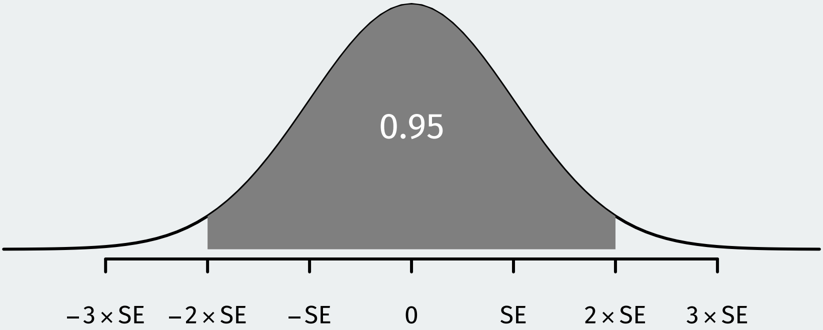
- Chance error: $\overline{Y} - p$ is approximately normal with mean 0 and SE equal to $\sqrt{\frac{p(1-p)}{n}}$

Chance errors



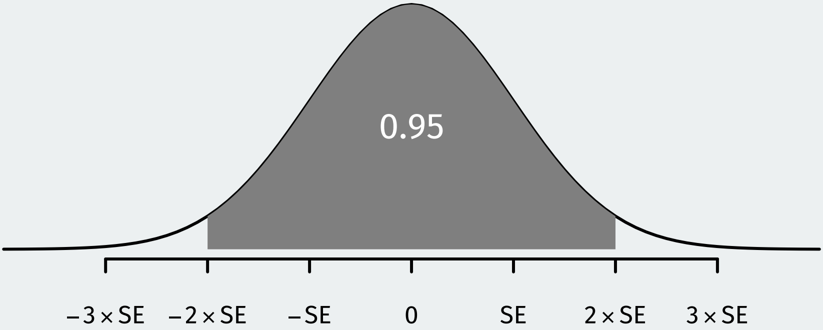
- We know 95% of chance errors will be within $\approx 2 \times SE$

Chance errors



- We know 95% of chance errors will be within $\approx 2 \times SE$
  - (actually it's $1.96 \times SE$)

Chance errors



- We know 95% of chance errors will be within $\approx 2 \times SE$
  - (actually it's $1.96 \times SE$)
- $\rightsquigarrow$ range of plausible chance errors is $\pm 1.96 \times SE$

Confidence interval

- First, choose a **confidence level**.

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as “plausible”?

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as “plausible”?
  - Convention is 95%.

Confidence interval

- First, choose a **confidence level**.
    - What percent of chance errors do you want to count as "plausible"?
    - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ of the probability density above it.

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ of the probability density above it.
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as “plausible”?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ of the probability density above it.
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$
  - 90% CI $\rightsquigarrow \alpha = 0.1 \rightsquigarrow z_{\alpha/2} = 1.64$

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ of the probability density above it.
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$
  - 90% CI $\rightsquigarrow \alpha = 0.1 \rightsquigarrow z_{\alpha/2} = 1.64$
  - 95% CI $\rightsquigarrow \alpha = 0.05 \rightsquigarrow z_{\alpha/2} = 1.96$

Confidence interval

- First, choose a **confidence level**.
  - What percent of chance errors do you want to count as "plausible"?
  - Convention is 95%.
- $100 \times (1 - \alpha)\%$ confidence interval:

$$CI = \overline{Y} \pm z_{\alpha/2} \times SE$$

  - In polling, $\pm z_{\alpha/2} \times SE$ is called the **margin of error**
- $z_{\alpha/2}$ is the $N(0, 1)$ z-score that would put $\alpha/2$ of the probability density above it.
  - $\mathbb{P}(-z_{\alpha/2} < Z < z_{\alpha/2}) = \alpha$
  - 90% CI $\rightsquigarrow \alpha = 0.1 \rightsquigarrow z_{\alpha/2} = 1.64$
  - 95% CI $\rightsquigarrow \alpha = 0.05 \rightsquigarrow z_{\alpha/2} = 1.96$
  - 99% CI $\rightsquigarrow \alpha = 0.01 \rightsquigarrow z_{\alpha/2} = 2.58$

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

```
qnorm(0.05, lower.tail = FALSE)
```

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

```
qnorm(0.05, lower.tail = FALSE)
```

```
## [1] 1.64
```

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

```
qnorm(0.05, lower.tail = FALSE)
```

```
## [1] 1.64
```

```
qnorm(0.025, lower.tail = FALSE)
```

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

```
qnorm(0.05, lower.tail = FALSE)
```

```
## [1] 1.64
```

```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.96
```

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

```
qnorm(0.05, lower.tail = FALSE)
```

```
## [1] 1.64
```

```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.96
```

```
qnorm(0.005, lower.tail = FALSE)
```

Standard normal z-scores in R

- `qnorm(x, lower.tail = FALSE)` will find the value of z so that $\mathbb{P}(Z < z)$ is equal to `x`, where Z is $N(0, 1)$:

```
qnorm(0.05, lower.tail = FALSE)
```

```
## [1] 1.64
```

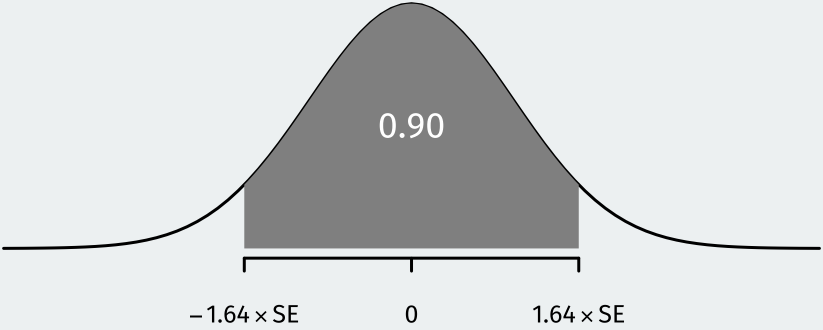
```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.96
```

```
qnorm(0.005, lower.tail = FALSE)
```

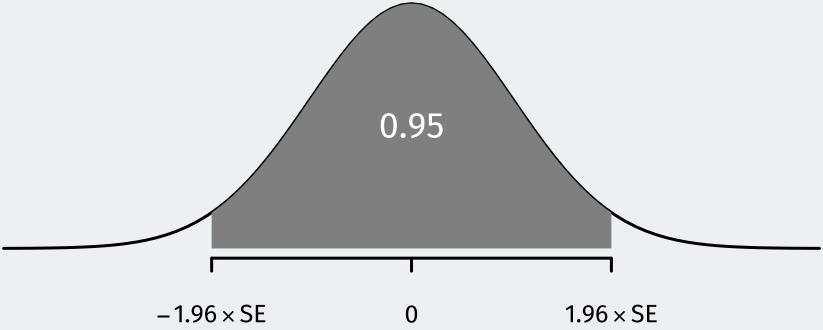
```
## [1] 2.58
```

Z-values



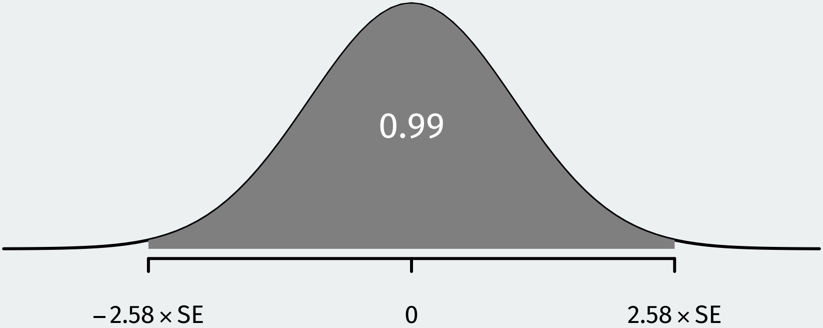
$$CI_{90} = \overline{Y} \pm 1.64 \times SE$$

Z-values



$$CI_{95} = \overline{Y} \pm 1.96 \times SE$$

Z-values



$$CI_{99} = \overline{Y} \pm 2.58 \times SE$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012]$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

$$[0.37 - 1.96 \times 0.012,\ 0.37 + 1.96 \times 0.012]$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

$$[0.37 - 1.96 \times 0.012,\ 0.37 + 1.96 \times 0.012] = [0.346, 0.394]$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

$$[0.37 - 1.96 \times 0.012,\ 0.37 + 1.96 \times 0.012] = [0.346, 0.394]$$

- 99% CI:

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

$$[0.37 - 1.96 \times 0.012,\ 0.37 + 1.96 \times 0.012] = [0.346, 0.394]$$

- 99% CI:

$$[0.37 - 2.58 \times 0.012,\ 0.37 + 2.58 \times 0.012]$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

$$[0.37 - 1.96 \times 0.012,\ 0.37 + 1.96 \times 0.012] = [0.346, 0.394]$$

- 99% CI:

$$[0.37 - 2.58 \times 0.012,\ 0.37 + 2.58 \times 0.012] = [0.339, 0.401]$$

CIs for the Gallup poll

- Gallup poll: $\overline{Y} = 0.37$ with an SE of 0.012.
- 90% CI:

$$[0.37 - 1.64 \times 0.012,\ 0.37 + 1.64 \times 0.012] = [0.350, 0.389]$$

- 95% CI:

$$[0.37 - 1.96 \times 0.012,\ 0.37 + 1.96 \times 0.012] = [0.346, 0.394]$$

- 99% CI:

$$[0.37 - 2.58 \times 0.012,\ 0.37 + 2.58 \times 0.012] = [0.339, 0.401]$$

- More confidence $\rightsquigarrow$ wider intervals

Interpretation and simulation

- Be careful about interpretation:

Interpretation and simulation

- Be careful about interpretation:
  - A 95% confidence interval will contain the true value in 95% of repeated samples.

Interpretation and simulation

- Be careful about interpretation:
  - A 95% confidence interval will contain the true value in 95% of repeated samples.
  - For a particular calculated confidence interval, truth is either in it or not.

Interpretation and simulation

- Be careful about interpretation:
  - A 95% confidence interval will contain the true value in 95% of repeated samples.
  - For a particular calculated confidence interval, truth is either in it or not.
- A simulation can help our understanding:

Interpretation and simulation

- Be careful about interpretation:
  - A 95% confidence interval will contain the true value in 95% of repeated samples.
  - For a particular calculated confidence interval, truth is either in it or not.
- A simulation can help our understanding:
  - Draw samples of size 1500 assuming population approval for Trump of $p = 0.4$.

Interpretation and simulation

- Be careful about interpretation:
  - A 95% confidence interval will contain the true value in 95% of repeated samples.
  - For a particular calculated confidence interval, truth is either in it or not.
- A simulation can help our understanding:
  - Draw samples of size 1500 assuming population approval for Trump of $p = 0.4$.
  - Calculate 95% confidence intervals in each sample.

Interpretation and simulation

- Be careful about interpretation:
  - A 95% confidence interval will contain the true value in 95% of repeated samples.
  - For a particular calculated confidence interval, truth is either in it or not.
- A simulation can help our understanding:
  - Draw samples of size 1500 assuming population approval for Trump of $p = 0.4$.
  - Calculate 95% confidence intervals in each sample.
  - See how many overlap with the true population approval.

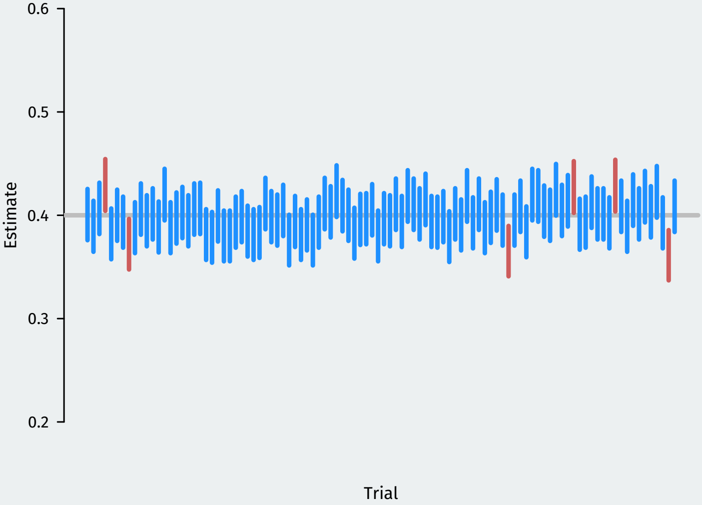
Plotting the CIs

Plotting the CIs

Plotting the CIs

Plotting the CIs

Plotting the CIs



5/ How big of a sample do I need?

Margin of error

- Margin of error in a close race:

MoE

Margin of error

- Margin of error in a close race:

$$\text{MoE} = \pm 1.96 \times SE$$

Margin of error

- Margin of error in a close race:

$$\text{MoE} = \pm 1.96 \times SE = \pm 1.96\sqrt{\frac{0.5 \times 0.5}{n}}$$

Margin of error

- Margin of error in a close race:

$$\text{MoE} = \pm 1.96 \times SE = \pm 1.96 \sqrt{\frac{0.5 \times 0.5}{n}} \approx \pm \frac{1}{\sqrt{n}}$$

Margin of error

- Margin of error in a close race:

$$\text{MoE} = \pm 1.96 \times SE = \pm 1.96\sqrt{\frac{0.5 \times 0.5}{n}} \approx \pm\frac{1}{\sqrt{n}}$$

- Gallup polls have $n = 1500$ which implies $\text{MoE} = \pm 2.9$ percentage points.

Margin of error

- Margin of error in a close race:

$$\text{MoE} = \pm 1.96 \times SE = \pm 1.96 \sqrt{\frac{0.5 \times 0.5}{n}} \approx \pm \frac{1}{\sqrt{n}}$$

- Gallup polls have $n = 1500$ which implies MoE $= \pm 2.9$ percentage points.

> Gallup tracks daily the percentage of Americans who approve or disapprove of the job Donald Trump is doing as president. Daily results are based on telephone interviews with approximately 1,500 national adults; Margin of error is ±3 percentage points.

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n}$$

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n} \rightsquigarrow n = \frac{1.96^2 p(1-p)}{\text{MoE}^2}$$

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n} \rightsquigarrow n = \frac{1.96^2 p(1-p)}{\text{MoE}^2}$$

- Say you wanted an MoE of 0.03 for a true proportion of $p = 0.3$:

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n} \rightsquigarrow n = \frac{1.96^2 p(1-p)}{\text{MoE}^2}$$

- Say you wanted an MoE of 0.03 for a true proportion of $p = 0.3$:

$$n = \frac{1.96^2 \times 0.3 \times 0.7}{0.03^2}$$

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n} \rightsquigarrow n = \frac{1.96^2 p(1-p)}{\text{MoE}^2}$$

- Say you wanted an MoE of 0.03 for a true proportion of $p = 0.3$:

$$n = \frac{1.96^2 \times 0.3 \times 0.7}{0.03^2} = \frac{0.81}{0.0009}$$

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n} \rightsquigarrow n = \frac{1.96^2 p(1-p)}{\text{MoE}^2}$$

- Say you wanted an MoE of 0.03 for a true proportion of $p = 0.3$:

$$n = \frac{1.96^2 \times 0.3 \times 0.7}{0.03^2} = \frac{0.81}{0.0009} = 900$$

How big does my survey need to be?

- If we know the margin of error that we'd like, we can figure out what sample size we would need.
- Sample size calculation:

$$\text{MoE}^2 = \frac{1.96^2 p(1-p)}{n} \rightsquigarrow n = \frac{1.96^2 p(1-p)}{\text{MoE}^2}$$

- Say you wanted an MoE of 0.03 for a true proportion of $p = 0.3$:

$$n = \frac{1.96^2 \times 0.3 \times 0.7}{0.03^2} = \frac{0.81}{0.0009} = 900$$

- But we don't know p! $\rightsquigarrow$ use $p = 0.5$ since this require the biggest n.

Next steps

- Today: how to assess uncertainty in our survey estimates.

Next steps

- Today: how to assess uncertainty in our survey estimates.
- After midterm:

Next steps

- Today: how to assess uncertainty in our survey estimates.
- After midterm:
  - uncertainty in estimating treatment effects.

Next steps

- Today: how to assess uncertainty in our survey estimates.
- After midterm:
  - uncertainty in estimating treatment effects.
  - hypothesis tests.

Next steps

- Today: how to assess uncertainty in our survey estimates.
- After midterm:
  - uncertainty in estimating treatment effects.
  - hypothesis tests.
  - uncertainty in linear regression.