

Inference and Estimation

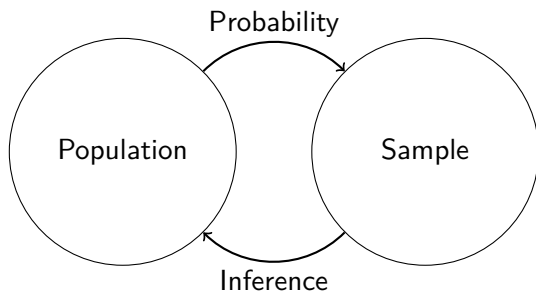
PSC4375: Week 11

Prof. Weldzius

Villanova University

Slides Updated: 2025-04-04

Remember our goal



- We want to learn about the chance process that generated our data.
- Now we switch to **inference**.
 - What can I learn about the population distribution from my sample?

What are random variables?

- What proportion of the public approves of Trump's job as president?
- Latest Gallup poll:
 - March 3-16, 2025
 - 1500 adult Americans
 - Approve (43%), Disapprove (53%)
- What can we learn about Trump approval in the population from this one sample?

Do you approve or disapprove of the way Donald Trump is handling his job as president?

Second-term trend

	Approve %	Disapprove %	No opinion %
2025			
2025 Mar 3-16	43	53	4
2025 Feb 3-16	45	51	5
2025 Jan 21-27	47	48	4

GALLUP

Samples from the population

- 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.
- **Statistical inference:** using data to guess something about the population distribution of Y_i .

Point estimation

- **Quantity of interest:** some feature of the population distribution.
 - Also called: parameters.
 - These are the things we want to learn about.
- **Point estimation:** providing a single “best guess” about this q.o.i.
- 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.

Estimators

- **Estimator**: function of the data that produces estimates of the q.o.i.
 - An **estimate** is one particular realization of the estimator
- Ideally we'd like to know the **estimation error**, estimator - truth
 - Problem: θ is unknown.
- Solution: figure out the properties of estimator using probability.
 - Estimator is a r.v. because it is a function of r.v.s. (the data)
 - \leadsto estimator has a distribution.

Estimating Trump's support

- Parameter p : **population proportion** of adults who support Trump
- There are many different possible estimators:
 - $\hat{p} = \frac{1}{n} \sum_{i=1}^n Y_i$ the sample proportion of respondents who support Trump.
 - $\hat{p} = Y_1$ just use the first observation
 - $\hat{p} = \max(Y_1, \dots, Y_n)$
 - $\hat{p} = 0.5$ always guess 50% support
- How good are these different estimators?

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
- X_i is **Bernoulli** with probability of success p
 - “success” = “selecting a Trump approver”
 - $p = \mathbb{P}(Y_i = 1)$ the population proportion of Trump approvers.
- Sample proportion is the same as the sample mean:

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

Sample mean properties

sample proportion = population proportion + chance error

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

- Remember: the sample mean/proportion is a random variable.
 - Different samples give different sample means.
 - Chance error “bumps” sample mean away from population mean
- $\leadsto \bar{Y}$ has a distribution across repeated samples.

Central tendency of the sample mean

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

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

- **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?
 - This is the standard deviation of the estimator.
- Special rule for sample proportions:

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

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

$$\sqrt{\mathbb{V}(\bar{Y})} = \sqrt{\frac{\bar{Y}(1-\bar{Y})}{n}} \approx 0.012$$

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

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

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

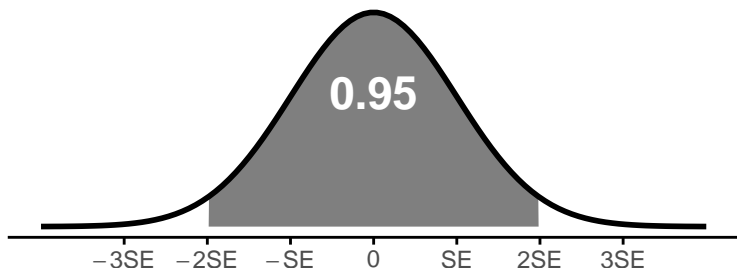
$$\bar{Y} \stackrel{\text{approx}}{\sim} N(\mathbb{E}(\bar{Y}), \frac{\mathbb{V}(Y_i)}{n})$$

- In this case:

$$\bar{Y} \stackrel{\text{approx}}{\sim} N(p, \frac{p(1-p)}{n})$$

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

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

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

- 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 $N(0, 1)$:

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

```
## [1] 1.644854
```

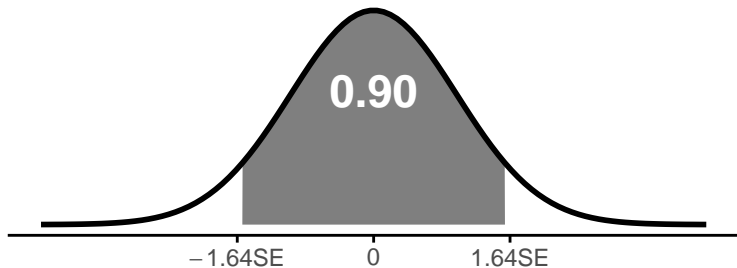
```
qnorm(0.025, lower.tail = FALSE)
```

```
## [1] 1.959964
```

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

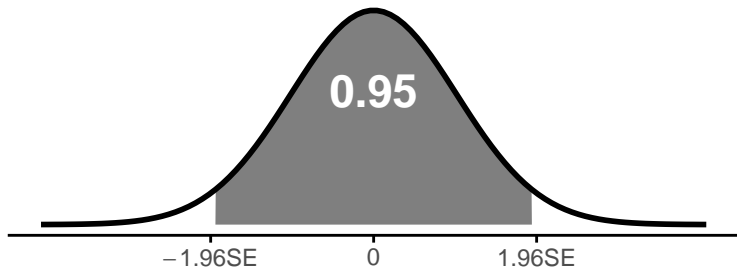
```
## [1] 2.575829
```


Z-values



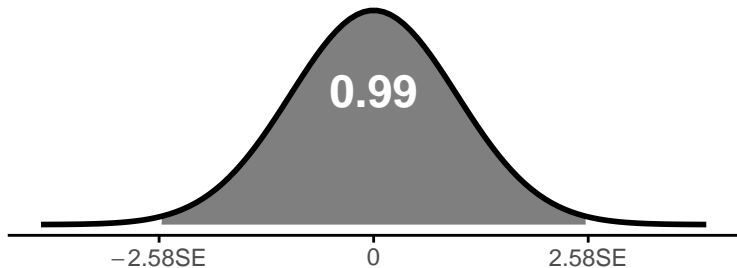
$$CI_{90} = \bar{Y} \pm 1.64 \times SE$$

Z-values



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

Z-values



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

CIs for the Gallup Poll

- Gallup poll: $\bar{Y} = 0.43$ with an SE of 0.012.
- 90% CI:

$$[0.43 - 1.64 \times 0.012, 0.43 + 1.64 \times 0.012] = [0.410, 0.449]$$

- 95% CI:

$$[0.43 - 1.96 \times 0.012, 0.43 + 1.96 \times 0.012] = [0.406, 0.454]$$

- 99% CI:

$$[0.43 - 2.58 \times 0.012, 0.43 + 2.58 \times 0.012] = [0.399, 0.461]$$

- More confidence \rightsquigarrow wider intervals

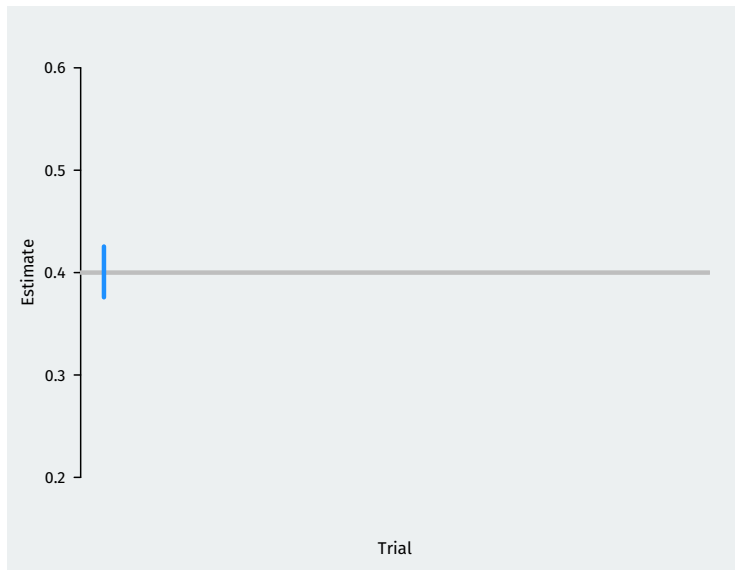
Interpretation and simulation

- 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.

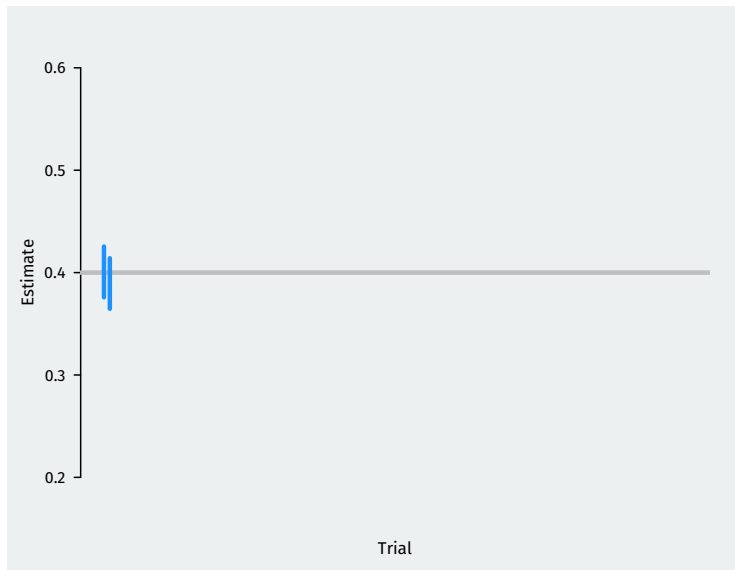
Plotting the CIs



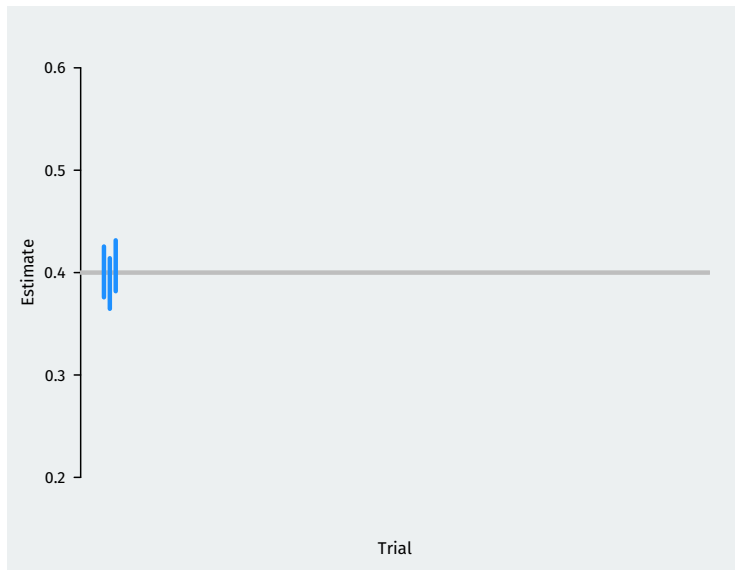
Plotting the CIs



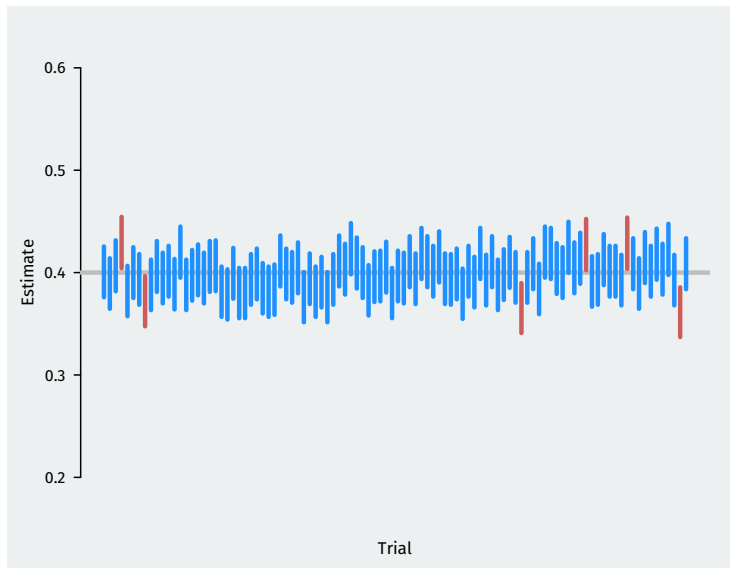
Plotting the CIs



Plotting the CIs



Plotting the CIs



Inference for experiments

- More interesting to compare across groups.
 - Differences in public opinion across groups
 - Difference between treatment and control groups.
- Bedrock of causal inference!

Social pressure experiment

- Back to the Social Pressure Mailer GOTV example.
 - Primary election in MI 2006
- Treatment group: postcards showing their own and their neighbors' voting records.
 - Sample size of treated group, $n_T = 360$
- Control group: received nothing.
 - Sample size of the control group, $n_C = 1890$

Outcomes

- Outcome: $X_i = 1$ if i vote, 0 otherwise.
- Turnout rate (sample mean) in treated group, $\bar{X}_T = 0.37$
- Turnout rate (sample mean) in control group, $\bar{X}_C = 0.30$
- Estimated **average treatment effect**

$$\widehat{ATE} = \bar{X}_T - \bar{X}_C = 0.07$$

Inference for the difference

- Parameter: **population ATE** $\mu_T - \mu_C$
 - μ_T : Turnout rate in the population if everyone received treatment.
 - μ_C : Turnout rate in the population if everyone received control.
- Estimator: $\widehat{ATE} = \bar{X}_T - \bar{X}_C$
- \bar{X}_T is a r.v. with mean $\mathbb{E}[\bar{X}_T] = \mu_T$
- \bar{X}_C is a r.v. with mean $\mathbb{E}[\bar{X}_C] = \mu_C$
- $\rightsquigarrow \bar{X}_T - \bar{X}_C$ is a r.v. with mean $\mu_T - \mu_C$
 - Sample difference in means is on average equal to the population difference in means.

Simulation

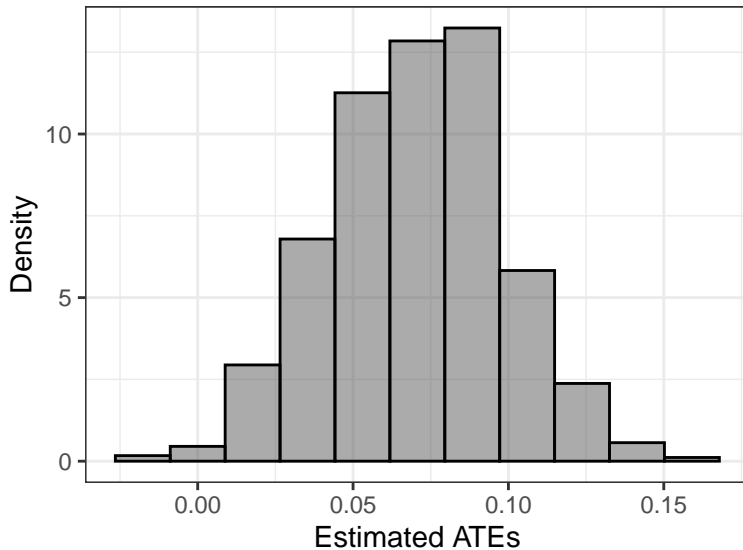
- What if these were the true population means? We would still expect some **variation** in our estimates:

```
xt.sims <- rbinom(1000, size = 360, prob = 0.37) / 360
xc.sims <- rbinom(1000, size = 1890, prob = 0.30) / 1890

diff.sims <- as.tibble(xt.sims - xc.sims)

diff.sims %>%
  ggplot(aes(x = value, y = ..density..)) +
    geom_histogram(bins = 11, alpha = 0.5, color = "black") +
    labs(x = "Estimated ATEs", y = "Density") +
    theme_bw()
```

Simulations



Standard error

- Is an $\widehat{ATE} = 0.07$ big?
- How much variation would we expect in the difference in means across repeated samples?
- **Variance** of our estimates:

$$\begin{aligned}\mathbb{V}(\widehat{ATE}) &= \mathbb{V}(\bar{X}_T - \bar{X}_C) = \mathbb{V}(\bar{X}_T) + \mathbb{V}(\bar{X}_C) \\ &= \frac{\mu_T(1 - \mu_T)}{n_T} + \frac{\mu_C(1 - \mu_C)}{n_C}\end{aligned}$$

- **Standard error** is the square root of this variance:

$$\widehat{SE}_{\widehat{ATE}} = \sqrt{\frac{\bar{X}_T(1 - \bar{X}_T)}{n_T} + \frac{\bar{X}_C(1 - \bar{X}_C)}{n_C}} = 0.028$$

- SE represents how far, on average, $\bar{X}_T - \bar{X}_C$ will be from $\mu_T - \mu_C$

Confidence intervals

- We can construct confidence intervals based on the CLT like last time.

$$\begin{aligned} CI_{95} &= \widehat{ATE} \pm 1.96 \times \widehat{SE}_{\widehat{ATE}} \\ &= 0.07 \pm 1.96 \times 0.028 \\ &= 0.07 \pm 0.054 \\ &= [0.016, 0.124] \end{aligned}$$

- Range of possibilities taking into account plausible chance errors.
- 0 not included in this CI \rightsquigarrow chance error as big as the estimated effect unlikely