10. Hypothesis Testing

Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- Recently: how to build estimators to estimate parameters.

Where are we? Where are we going?

- Recently: how to build estimators to estimate parameters.
- Also learned properties of these estimators in finite and large samples.

Where are we? Where are we going?

- Recently: how to build estimators to estimate parameters.
- Also learned properties of these estimators in finite and large samples.
- Now: how to use estimates to test a particular hypothesis about a parameter.

Where are we? Where are we going?

- Recently: how to build estimators to estimate parameters.
- Also learned properties of these estimators in finite and large samples.
- Now: how to use estimates to test a particular hypothesis about a parameter.
- We'll draw on our probability knowledge from earlier in the term!

1/ Hypothesis Testing Examples

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.
- R.A. Fisher was skeptical so he devised a test:

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.
- R.A. Fisher was skeptical so he devised a test:
    - Prepare 8 cups of tea, 4 milk-first, 4 tea-first

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.
- R.A. Fisher was skeptical so he devised a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups in a **random** order, asked her to pick which 4 are milk-first

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.
- R.A. Fisher was skeptical so he devised a test:
    - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
    - Present cups in a **random** order, asked her to pick which 4 are milk-first
- She guessed all correctly!

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.
- R.A. Fisher was skeptical so he devised a test:
  - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
  - Present cups in a **random** order, asked her to pick which 4 are milk-first
- She guessed all correctly!
  - This is our data. What can we learn from it?

The lady tasting tea

- Biologist Muriel Bristol claimed she could tell whether tea or milk was added first to a cup.
- R.A. Fisher was skeptical so he devised a test:
    - Prepare 8 cups of tea, 4 milk-first, 4 tea-first
    - Present cups in a **random** order, asked her to pick which 4 are milk-first
- She guessed all correctly!
    - This is our data. What can we learn from it?
    - There is uncertainty: she could have guessed randomly.

Taste tests

- **Statistical thought experiment**: how often would she get all 4 correct **if she were guessing randomly**?

Taste tests

- **Statistical thought experiment**: how often would she get all 4 correct **if she were guessing randomly**?
    - Only one way to choose all 4 correct cups, but 70 ways of choosing 4 cups among 8.

Taste tests

- **Statistical thought experiment**: how often would she get all 4 correct **if she were guessing randomly**?
    - Only one way to choose all 4 correct cups, but 70 ways of choosing 4 cups among 8.
    - Choosing at random $\approx$ picking each of these 70 with equal probability.

Taste tests

- **Statistical thought experiment**: how often would she get all 4 correct **if she were guessing randomly**?
    - Only one way to choose all 4 correct cups, but 70 ways of choosing 4 cups among 8.
    - Choosing at random $\approx$ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.

Taste tests

- **Statistical thought experiment**: how often would she get all 4 correct **if she were guessing randomly**?
    - Only one way to choose all 4 correct cups, but 70 ways of choosing 4 cups among 8.
    - Choosing at random $\approx$ picking each of these 70 with equal probability.
- Chances of guessing all 4 correct is $\frac{1}{70} \approx 0.014$ or 1.4%.
- $\rightsquigarrow$ the guessing at random hypothesis might be implausible.

Social pressure effect

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY—VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

Social pressure effect

```
load("../assets/gerber_green_larimer.RData")
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment == "Neighbors"])
contr.mean <- mean(social$voted[social$treatment == "Civic Duty"])
neigh.mean - contr.mean
```

```
## [1] 0.0634
```

- Treatment effect of 6.3 percentage points.

Social pressure effect

```
load("../assets/gerber_green_larimer.RData")
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment == "Neighbors"])
contr.mean <- mean(social$voted[social$treatment == "Civic Duty"])
neigh.mean - contr.mean
```

```
## [1] 0.0634
```

- Treatment effect of 6.3 percentage points.
- But the estimator varies from sample to sample by random chance.

Social pressure effect

```
load("../assets/gerber_green_larimer.RData")
social$voted <- 1 * (social$voted == "Yes")
neigh.mean <- mean(social$voted[social$treatment == "Neighbors"])
contr.mean <- mean(social$voted[social$treatment == "Civic Duty"])
neigh.mean - contr.mean
```

```
## [1] 0.0634
```

- Treatment effect of 6.3 percentage points.
- But the estimator varies from sample to sample by random chance.
- Could it be this big by random chance if there was no effect at all?

Difference in means

- **Treated group** $Y_1, Y_2, \ldots, Y_{n_y}$ i.i.d. with population mean μ_y and population variance σ_y^2

Difference in means

- **Treated group** $Y_1, Y_2, \ldots, Y_{n_y}$ i.i.d. with population mean μ_y and population variance σ_y^2
- **Control group** $X_1, X_2, \ldots, X_{n_x}$ i.i.d. with population mean μ_x and population variance σ_x^2

Difference in means

- **Treated group** $Y_1, Y_2, \ldots, Y_{n_y}$ i.i.d. with population mean μ_y and population variance σ_y^2
- **Control group** $X_1, X_2, \ldots, X_{n_x}$ i.i.d. with population mean μ_x and population variance σ_x^2
- Quantity of interest: **population differences in average turnout**

$$\tau = \mathbb{E}[Y_i] - \mathbb{E}[X_i]$$

Difference in means

- **Treated group** $Y_1, Y_2, \ldots, Y_{n_y}$ i.i.d. with population mean μ_y and population variance σ_y^2
- **Control group** $X_1, X_2, \ldots, X_{n_x}$ i.i.d. with population mean μ_x and population variance σ_x^2
- Quantity of interest: **population differences in average turnout**

$$\tau = \mathbb{E}[Y_i] - \mathbb{E}[X_i]$$

- Estimator: sample difference in means: $\widehat{\tau}_n = \overline{Y}_{n_y} - \overline{X}_{n_x}$

Difference in means

- **Treated group** $Y_1, Y_2, \dots, Y_{n_y}$ i.i.d. with population mean μ_y and population variance σ_y^2
- **Control group** $X_1, X_2, \dots, X_{n_x}$ i.i.d. with population mean μ_x and population variance σ_x^2
- Quantity of interest: **population differences in average turnout**

$$\tau = \mathbb{E}[Y_i] - \mathbb{E}[X_i]$$

- Estimator: sample difference in means: $\widehat{\tau}_n = \overline{Y}_{n_y} - \overline{X}_{n_x}$
- We estimate the standard error of $\widehat{\tau}_n$ with:

$$\widehat{\text{se}}[\widehat{\tau}_n] = \sqrt{\frac{s_y^2}{n_y} + \frac{s_x^2}{n_x}}$$

2/ Hypothesis Testing Framework

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.
- We might have hypotheses about causal inferences:

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.
- We might have hypotheses about causal inferences:
  - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.
- We might have hypotheses about causal inferences:
    - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)
    - Do treaties constrain countries? (behavior different among treaty signers?)

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.
- We might have hypotheses about causal inferences:
  - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)
  - Do treaties constrain countries? (behavior different among treaty signers?)
- We might also have hypotheses about other parameters:

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.
- We might have hypotheses about causal inferences:
  - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)
  - Do treaties constrain countries? (behavior different among treaty signers?)
- We might also have hypotheses about other parameters:
  - Is the share of Biden supporters more than 50%?

What is a hypothesis?

- A **hypothesis** is just a statement a population parameter, θ.
- We might have hypotheses about causal inferences:
  - Does social pressure induce higher voter turnout? (mean turnout higher in social pressure group compared to Civic Duty group?)
  - Do treaties constrain countries? (behavior different among treaty signers?)
- We might also have hypotheses about other parameters:
  - Is the share of Biden supporters more than 50%?
  - Are traits of treatment and control groups different?

Hypothesis testing procedure

1. Choose null and alternative hypotheses

Hypothesis testing procedure

1. Choose null and alternative hypotheses
2. Choose a test statistic, T_n

Hypothesis testing procedure

1. Choose null and alternative hypotheses
2. Choose a test statistic, T_n
3. Choose a test level, α

Hypothesis testing procedure

1. Choose null and alternative hypotheses
2. Choose a test statistic, T_n
3. Choose a test level, α
4. Determine rejection region

Hypothesis testing procedure

1. Choose null and alternative hypotheses
2. Choose a test statistic, T_n
3. Choose a test level, α
4. Determine rejection region
5. Reject if T_n in rejection region, fail to reject otherwise

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually “no effect/difference/relationship.”

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually “no effect/difference/relationship.”
  - We denote this hypothesis as $H_0 : \theta = \theta_0$.

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
    - This is usually "no effect/difference/relationship."
    - We denote this hypothesis as $H_0 : \theta = \theta_0$.
    - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually "no effect/difference/relationship."
  - We denote this hypothesis as $H_0 : \theta = \theta_0$.
  - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
    - This is usually “no effect/difference/relationship.”
    - We denote this hypothesis as $H_0 : \theta = \theta_0$.
    - H_0: Social pressure doesn’t affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
    - Usually, “there is a relationship/difference/effect.”

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
    - This is usually "no effect/difference/relationship."
    - We denote this hypothesis as $H_0 : \theta = \theta_0$.
    - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
    - Usually, "there is a relationship/difference/effect."
    - We denote this as $H_1 : \theta \neq \theta_0$.

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually "no effect/difference/relationship."
  - We denote this hypothesis as $H_0 : \theta = \theta_0$.
  - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
  - Usually, "there is a relationship/difference/effect."
  - We denote this as $H_1 : \theta \neq \theta_0$.
  - H_1: Social pressure affects turnout ($H_1 : \tau \neq 0$)

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually "no effect/difference/relationship."
  - We denote this hypothesis as $H_0 : \theta = \theta_0$.
  - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
  - Usually, "there is a relationship/difference/effect."
  - We denote this as $H_1 : \theta \neq \theta_0$.
  - H_1: Social pressure affects turnout ($H_1 : \tau \neq 0$)
- One-sided vs. two-sided alternatives:

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
    - This is usually "no effect/difference/relationship."
    - We denote this hypothesis as $H_0 : \theta = \theta_0$.
    - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
    - Usually, "there is a relationship/difference/effect."
    - We denote this as $H_1 : \theta \neq \theta_0$.
    - H_1: Social pressure affects turnout ($H_1 : \tau \neq 0$)
- One-sided vs. two-sided alternatives:
    - One-sided: $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually "no effect/difference/relationship."
  - We denote this hypothesis as $H_0 : \theta = \theta_0$.
  - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
  - Usually, "there is a relationship/difference/effect."
  - We denote this as $H_1 : \theta \neq \theta_0$.
  - H_1: Social pressure affects turnout ($H_1 : \tau \neq 0$)
- One-sided vs. two-sided alternatives:
  - One-sided: $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
  - Two-sided: $H_1 : \theta \neq \theta_0$

Null and alternative hypotheses

- The **null hypothesis** is the hypothesis we want to test.
  - This is usually "no effect/difference/relationship."
  - We denote this hypothesis as $H_0 : \theta = \theta_0$.
  - H_0: Social pressure doesn't affect turnout ($H_0 : \tau = 0$)
- The **alternative hypothesis** is the complement of the null hypothesis
  - Usually, "there is a relationship/difference/effect."
  - We denote this as $H_1 : \theta \neq \theta_0$.
  - H_1: Social pressure affects turnout ($H_1 : \tau \neq 0$)
- One-sided vs. two-sided alternatives:
  - One-sided: $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
  - Two-sided: $H_1 : \theta \neq \theta_0$
  - Two-sided much more common, one-sided involves ignoring evidence in one direction.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.
    - **Statistical thought experiments**: Assume we know (part of) the true DGP.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.
    - **Statistical thought experiments**: Assume we know (part of) the true DGP.
- Rejection based on a **test statistic**, $T_n = T(Y_1, \dots, Y_n)$.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.
    - **Statistical thought experiments**: Assume we know (part of) the true DGP.
- Rejection based on a **test statistic**, $T_n = T(Y_1, \ldots, Y_n)$.
    - Will help us adjudicate between the null and the alternative.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.
    - **Statistical thought experiments**: Assume we know (part of) the true DGP.
- Rejection based on a **test statistic**, $T_n = T(Y_1, \ldots, Y_n)$.
    - Will help us adjudicate between the null and the alternative.
    - Typically: larger values of $T_n \rightsquigarrow$ null less plausible.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.
  - **Statistical thought experiments**: Assume we know (part of) the true DGP.
- Rejection based on a **test statistic**, $T_n = T(Y_1, \dots, Y_n)$.
  - Will help us adjudicate between the null and the alternative.
  - Typically: larger values of $T_n \rightsquigarrow$ null less plausible.
  - A test statistic is a r.v.

General framework

- **Hypothesis tests** choose to reject or not reject the null hypothesis based on the observed data.
  - **Statistical thought experiments**: Assume we know (part of) the true DGP.
- Rejection based on a **test statistic**, $T_n = T(Y_1, \ldots, Y_n)$.
  - Will help us adjudicate between the null and the alternative.
  - Typically: larger values of $T_n \rightsquigarrow$ null less plausible.
  - A test statistic is a r.v.
- Intuitively, reject null of no effect when $|\overline{Y} - \overline{X}|$ is large.

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
    - If our data is in R, we reject H_0

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:
  - One-sided tests: $T_n > c$

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:
  - One-sided tests: $T_n > c$
  - All the samples of size n leading to a T_n greater than c.

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
    - If our data is in R, we reject H_0
    - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:
    - One-sided tests: $T_n > c$
    - All the samples of size n leading to a T_n greater than c.
    - Two-sided tests: $|T_n| > c$

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:
  - One-sided tests: $T_n > c$
  - All the samples of size n leading to a T_n greater than c.
  - Two-sided tests: $|T_n| > c$
- The c here is the **critical value** that defines the rejection region, C:

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:
  - One-sided tests: $T_n > c$
  - All the samples of size n leading to a T_n greater than c.
  - Two-sided tests: $|T_n| > c$
- The c here is the **critical value** that defines the rejection region, C:
  - One-sided $C = \{t : t > c\}$, two-sided: $C = \{t : |t| > c\}$.

Rejection

- **Rejection region** R is the region of the sample space in which we reject the null.
  - If our data is in R, we reject H_0
  - If our data is not in R, we retain/fail to reject H_0.
- Regions often based on the test statistic. For scalar hypotheses:
  - One-sided tests: $T_n > c$
  - All the samples of size n leading to a T_n greater than c.
  - Two-sided tests: $|T_n| > c$
- The c here is the **critical value** that defines the rejection region, C:
  - One-sided $C = \{t : t > c\}$, two-sided: $C = \{t : |t| > c\}$.
  - Reject when $T_n \in C$.

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.
  - No treatment effect, but we reject the null

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.
    - No treatment effect, but we reject the null
- **Type II error** not rejecting the null hypothesis when it is false.

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.
  - No treatment effect, but we reject the null
- **Type II error** not rejecting the null hypothesis when it is false.
  - Treatment effect is nonzero, but we cannot reject the null

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.
    - No treatment effect, but we reject the null
- **Type II error** not rejecting the null hypothesis when it is false.
    - Treatment effect is nonzero, but we cannot reject the null
- Consequences depend the context:

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.
    - No treatment effect, but we reject the null
- **Type II error** not rejecting the null hypothesis when it is false.
    - Treatment effect is nonzero, but we cannot reject the null
- Consequences depend the context:
    - Treatment effects: false discovery (type I) vs undetected finding (type II).

Type I and Type II errors

	H_0 True	H_0 False
Retain H_0	Awesome!	Type II error
Reject H_0	Type I error	Good stuff!

- **Type I error**: rejecting the null hypothesis when it is in fact true.
  - No treatment effect, but we reject the null
- **Type II error** not rejecting the null hypothesis when it is false.
  - Treatment effect is nonzero, but we cannot reject the null
- Consequences depend the context:
  - Treatment effects: false discovery (type I) vs undetected finding (type II).
  - Medical diagnosis: false positive (type I) vs false negative (type II).

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?
  - The **power** of a test against an alternative $\theta_1 \in H_1$ is $\pi(\theta_1)$

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?
  - The **power** of a test against an alternative $\theta_1 \in H_1$ is $\pi(\theta_1)$
  - We want to maximize power against alternative

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?
  - The **power** of a test against an alternative $\theta_1 \in H_1$ is $\pi(\theta_1)$
  - We want to maximize power against alternative

- **Size** of a test is the probability of a Type I error:

$$\pi(\theta_0) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta_0\right)$$

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?
  - The **power** of a test against an alternative $\theta_1 \in H_1$ is $\pi(\theta_1)$
  - We want to maximize power against alternative

- **Size** of a test is the probability of a Type I error:

$$\pi(\theta_0) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta_0\right)$$

  - Size of two-sided test: $\mathbb{P}(|T_n| > c \mid \theta_0)$

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?
  - The **power** of a test against an alternative $\theta_1 \in H_1$ is $\pi(\theta_1)$
  - We want to maximize power against alternative

- **Size** of a test is the probability of a Type I error:

$$\pi(\theta_0) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta_0\right)$$

  - Size of two-sided test: $\mathbb{P}(|T_n| > c \mid \theta_0)$
  - Size of one-sided test: $\mathbb{P}(T_n > c \mid \theta_0)$

Features of a test

- Good tests: reject null when they should, retain when they shouldn't.
- **Power function** of a test: probability of rejection as a function of θ:

$$\pi(\theta) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta\right) = \mathbb{P}\left(T_n \in C \mid \theta\right)$$

  - **Hypotheticals!** if we knew θ, what is the probability of rejecting the null?
  - The **power** of a test against an alternative $\theta_1 \in H_1$ is $\pi(\theta_1)$
  - We want to maximize power against alternative

- **Size** of a test is the probability of a Type I error:

$$\pi(\theta_0) = \mathbb{P}\left(\text{Reject } H_0 \mid \theta_0\right)$$

  - Size of two-sided test: $\mathbb{P}(|T_n| > c \mid \theta_0)$
  - Size of one-sided test: $\mathbb{P}(T_n > c \mid \theta_0)$
  - We want to minimize the size of a test.

Test statistic example

- What is an example of a test statistic and how we know its distribution?

Test statistic example

- What is an example of a test statistic and how we know its distribution?
- By the CLT, the difference in means is asymptotically normal:

$$\frac{\widehat{\tau}_n - \tau}{\widehat{\mathsf{se}}[\widehat{\tau}_n]} \xrightarrow{d} \mathcal{N}(0, 1)$$

Test statistic example

- What is an example of a test statistic and how we know its distribution?
- By the CLT, the difference in means is asymptotically normal:

$$\frac{\widehat{\tau}_n - \tau}{\widehat{\mathsf{se}}[\widehat{\tau}_n]} \overset{d}{\to} \mathcal{N}(0,1)$$

- Under the null of $H_0 : \tau = \mathbb{E}[Y_i] - \mathbb{E}[X_i] = 0$, then **asymptotically**:

$$T_n = \frac{\widehat{\tau}_n}{\widehat{\mathsf{se}}[\widehat{\tau}_n]} \overset{d}{\to} \mathcal{N}(0,1)$$

Test statistic example

- What is an example of a test statistic and how we know its distribution?
- By the CLT, the difference in means is asymptotically normal:

$$\frac{\widehat{\tau}_n - \tau}{\widehat{\text{se}}[\widehat{\tau}_n]} \xrightarrow{d} \mathcal{N}(0, 1)$$

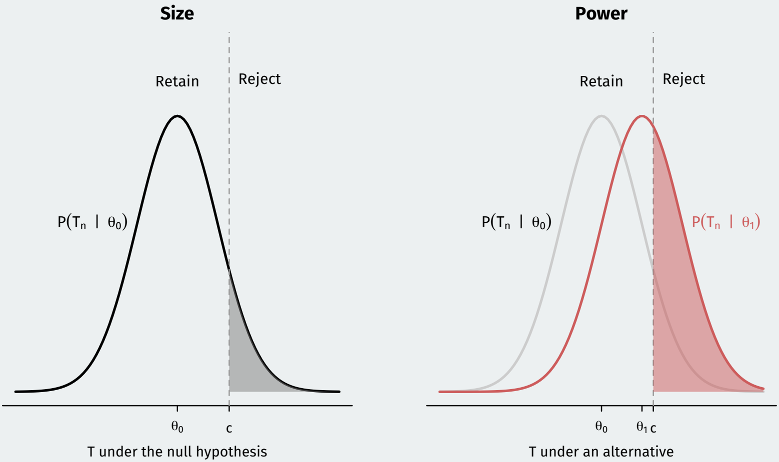
- Under the null of $H_0 : \tau = \mathbb{E}[Y_i] - \mathbb{E}[X_i] = 0$, then **asymptotically**:

$$T_n = \frac{\widehat{\tau}_n}{\widehat{\text{se}}[\widehat{\tau}_n]} \xrightarrow{d} \mathcal{N}(0, 1)$$

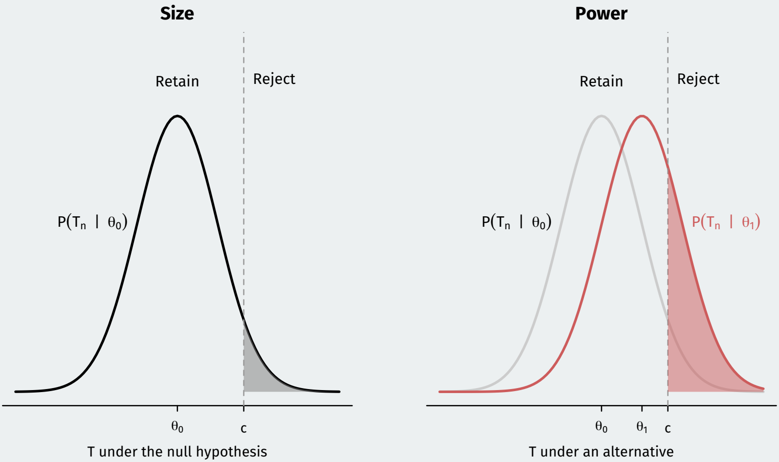
- Under an alternative $H_1 : \tau = \tau_1$:

$$T_n \xrightarrow{d} \mathcal{N}\left(\frac{\tau_1}{\text{se}(\widehat{\tau})}, 1\right)$$

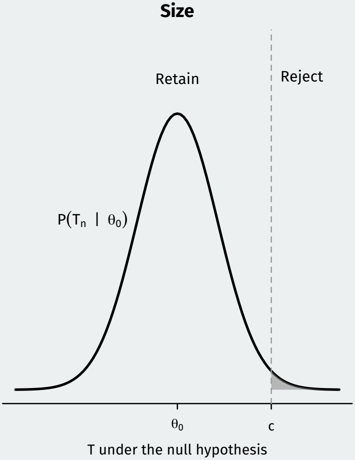
Size-power trade-off

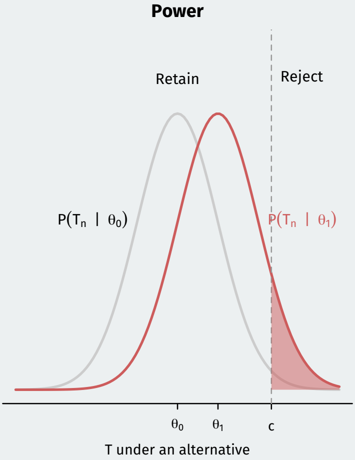


Size-power trade-off



Size-power trade-off





Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.

Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.
  - Classical Neyman-Pearson approach: fix the rate of Type I errors.

Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.
    - Classical Neyman-Pearson approach: fix the rate of Type I errors.
- **Significance level** α is researcher-selected maximum size of a test.

Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.
  - Classical Neyman-Pearson approach: fix the rate of Type I errors.
- **Significance level** α is researcher-selected maximum size of a test.
  - Convention in social sciences is $\alpha = 0.05$, but nothing magical there

Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.
    - Classical Neyman-Pearson approach: fix the rate of Type I errors.
- **Significance level** α is researcher-selected maximum size of a test.
    - Convention in social sciences is $\alpha = 0.05$, but nothing magical there
    - Particle physicists at CERN use $\alpha \approx \frac{1}{1{,}750{,}000}$

Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.
  - Classical Neyman-Pearson approach: fix the rate of Type I errors.
- **Significance level** α is researcher-selected maximum size of a test.
  - Convention in social sciences is $\alpha = 0.05$, but nothing magical there
  - Particle physicists at CERN use $\alpha \approx \frac{1}{1{,}750{,}000}$
- Frequentist justification: in the long run, at most $100 \times \alpha\%$ decisions will be Type I errors

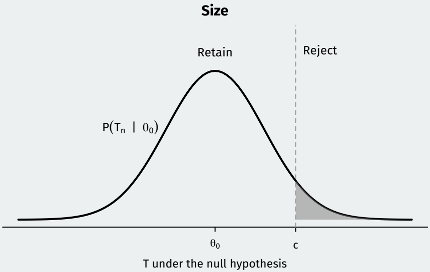
Controlling the size of a test

- Generally cannot simultaneously reduce both types of errors.
  - Classical Neyman-Pearson approach: fix the rate of Type I errors.
- **Significance level** α is researcher-selected maximum size of a test.
  - Convention in social sciences is $\alpha = 0.05$, but nothing magical there
  - Particle physicists at CERN use $\alpha \approx \frac{1}{1{,}750{,}000}$
- Frequentist justification: in the long run, at most $100 \times \alpha\%$ decisions will be Type I errors
  - Fisher (and Bayesians) didn't like this: relied on repeated sampling.

Controlling the size of a test

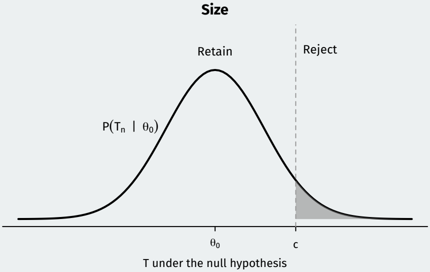
- Generally cannot simultaneously reduce both types of errors.
  - Classical Neyman-Pearson approach: fix the rate of Type I errors.
- **Significance level** α is researcher-selected maximum size of a test.
  - Convention in social sciences is $\alpha = 0.05$, but nothing magical there
  - Particle physicists at CERN use $\alpha \approx \frac{1}{1{,}750{,}000}$
- Frequentist justification: in the long run, at most $100 \times \alpha\%$ decisions will be Type I errors
  - Fisher (and Bayesians) didn't like this: relied on repeated sampling.
  - Still the dominant approach in the social sciences.

One-sided tests



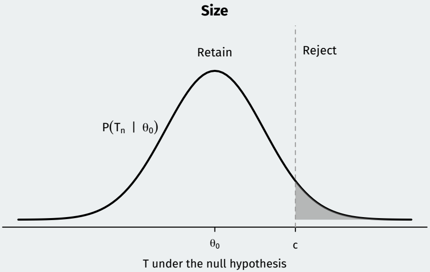
- How to select c to make $\alpha = 0.05$?

One-sided tests



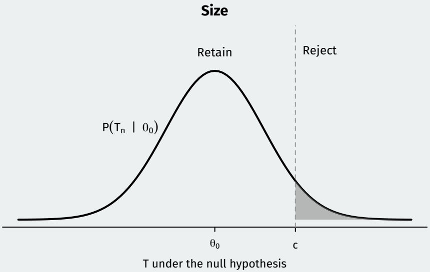
- How to select c to make $\alpha = 0.05$?
  - Let $G_0(t) = \mathbb{P}(T_n \leq t \mid \theta_0)$ be the c.d.f. under the null.

One-sided tests



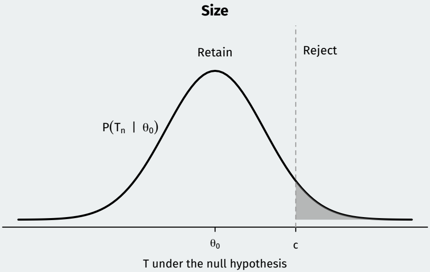
- How to select c to make $\alpha = 0.05$?
    - Let $G_0(t) = \mathbb{P}(T_n \leq t \mid \theta_0)$ be the c.d.f. under the null.
    - We want to find c that puts α probability in the tail: $1 - G_0(c) = \alpha$.

One-sided tests



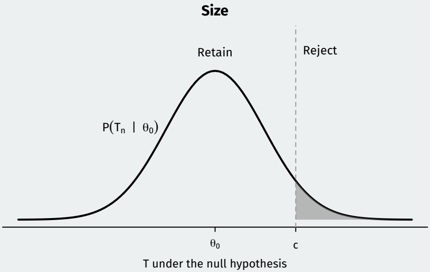
- How to select c to make $\alpha = 0.05$?
  - Let $G_0(t) = \mathbb{P}(T_n \leq t \mid \theta_0)$ be the c.d.f. under the null.
  - We want to find c that puts α probability in the tail: $1 - G_0(c) = \alpha$.
  - Use the **quantile function**: $c = G_0^{-1}(1 - \alpha)$

One-sided tests



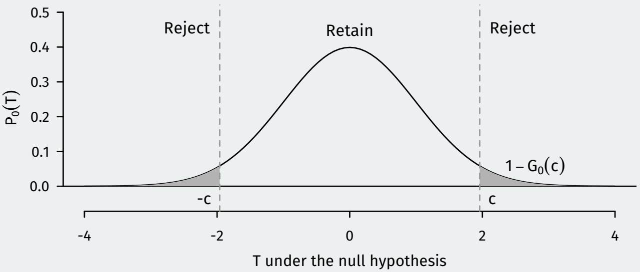
- How to select c to make $\alpha = 0.05$?
    - Let $G_0(t) = \mathbb{P}(T_n \leq t \mid \theta_0)$ be the c.d.f. under the null.
    - We want to find c that puts α probability in the tail: $1 - G_0(c) = \alpha$.
    - Use the **quantile function**: $c = G_0^{-1}(1 - \alpha)$
- If $G_0 \sim N(0, 1)$ and $\alpha = 0.05$, then $c = \Phi^{-1}(0.95) = 1.645$

One-sided tests

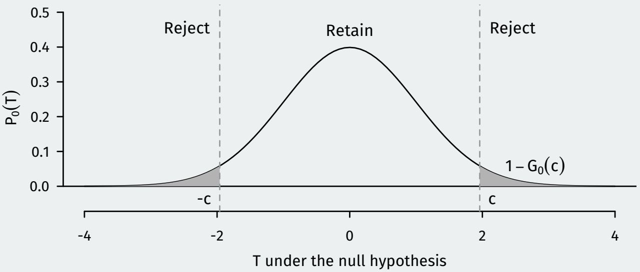


- How to select c to make $\alpha = 0.05$?
  - Let $G_0(t) = \mathbb{P}(T_n \leq t \mid \theta_0)$ be the c.d.f. under the null.
  - We want to find c that puts α probability in the tail: $1 - G_0(c) = \alpha$.
  - Use the **quantile function**: $c = G_0^{-1}(1 - \alpha)$
- If $G_0 \sim N(0, 1)$ and $\alpha = 0.05$, then $c = \Phi^{-1}(0.95) = 1.645$
  - Reject null if $T_n > 1.645$, fail to reject if $T_n \leq 1.645$

Two-sided rejection region

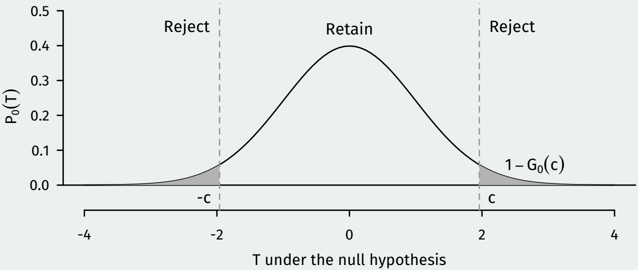


Two-sided rejection region



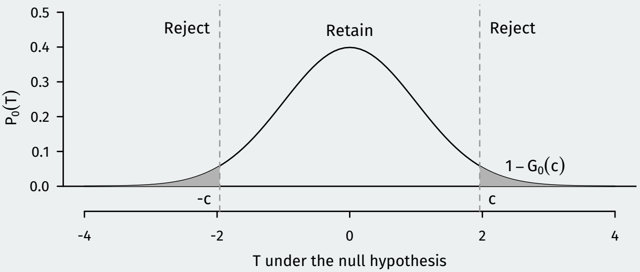
- What's the rejection region $|T_n| > c$ if $\alpha = 0.05$?

Two-sided rejection region



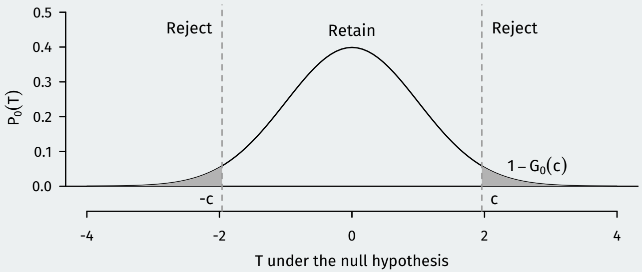
- What's the rejection region $|T_n| > c$ if $\alpha = 0.05$?
- For symmetric G_0 and given c, we have test size $\pi(\theta_0) = 2(1 - G_0(c))$

Two-sided rejection region



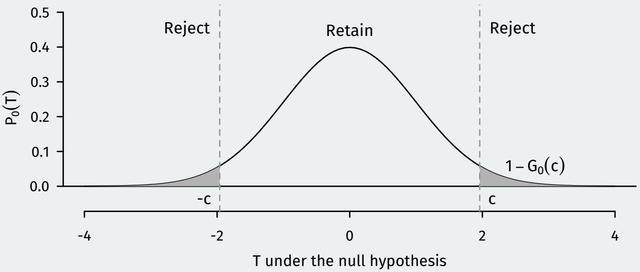
- What's the rejection region $|T_n| > c$ if $\alpha = 0.05$?
- For symmetric G_0 and given c, we have test size $\pi(\theta_0) = 2(1 - G_0(c))$
- Critical values: $c = G_0^{-1}(1 - \alpha/2)$

Two-sided rejection region



- What's the rejection region $|T_n| > c$ if $\alpha = 0.05$?
- For symmetric G_0 and given c, we have test size $\pi(\theta_0) = 2(1 - G_0(c))$
- Critical values: $c = G_0^{-1}(1 - \alpha/2)$
  - Find c such that $\alpha/2$ is in each tail

Two-sided rejection region



- What's the rejection region $|T_n| > c$ if $\alpha = 0.05$?
- For symmetric G_0 and given c, we have test size $\pi(\theta_0) = 2(1 - G_0(c))$
- Critical values: $c = G_0^{-1}(1 - \alpha/2)$
    - Find c such that $\alpha/2$ is in each tail
    - For $G_0 \sim \mathcal{N}(0, 1)$ and $\alpha = 0.05$, then $c = 1.96$

Final hypothesis test

1. Hypotheses: $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$

Final hypothesis test

1. Hypotheses: $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$
2. Test statistic: $T_n = \widehat{\tau}_n / \widehat{\text{se}}[\widehat{\tau}_n]$

Final hypothesis test

1. Hypotheses: $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$
2. Test statistic: $T_n = \widehat{\tau}_n / \widehat{\text{se}}[\widehat{\tau}_n]$
3. Use $\alpha = 0.05$

Final hypothesis test

1. Hypotheses: $H_0 : \tau = 0$ vs. $H_1 : \tau \neq 0$
2. Test statistic: $T_n = \widehat{\tau}_n / \widehat{\text{se}}[\widehat{\tau}_n]$
3. Use $\alpha = 0.05$
4. Rejection region is $|T_n| > 1.96$.

Social pressure test

- Calculate test statistic for social pressure mailers:

Social pressure test

- Calculate test statistic for social pressure mailers:

```
neigh_var <-  var(social$voted[social$treatment == "Neighbors"])
neigh_n <- sum(social$treatment == "Neighbors")
civic_var <- var(social$voted[social$treatment == "Civic Duty"])
civic_n <- sum(social$treatment == "Civic Duty")
se_diff <- sqrt(neigh_var/neigh_n + civic_var/civic_n)

## Calcuate test statistic
(0.378-0.315)/se_diff
```

```
## [1] 18.3
```

Social pressure test

- Calculate test statistic for social pressure mailers:

```
neigh_var <-  var(social$voted[social$treatment == "Neighbors"])
neigh_n <- sum(social$treatment == "Neighbors")
civic_var <- var(social$voted[social$treatment == "Civic Duty"])
civic_n <- sum(social$treatment == "Civic Duty")
se_diff <- sqrt(neigh_var/neigh_n + civic_var/civic_n)

## Calcuate test statistic
(0.378-0.315)/se_diff
```

```
## [1] 18.3
```

- $|T_n| = 18.343 > 1.96 \rightsquigarrow$ REJECT!

t-test/Wald test

- Consider **any** asymptotically normal estimator $\hat{\theta}$ for parameter θ.

t-test/Wald test

- Consider **any** asymptotically normal estimator $\hat{\theta}$ for parameter θ.
- Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

t-test/Wald test

- Consider **any** asymptotically normal estimator $\hat{\theta}$ for parameter θ.
- Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- A size-α **t-test** (or **Wald test**) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathrm{se}}[\hat{\theta}]}$$

t-test/Wald test

- Consider **any** asymptotically normal estimator $\hat{\theta}$ for parameter θ.
- Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- A size-α **t-test** (or **Wald test**) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\mathrm{se}}[\hat{\theta}]}$$

- Critical value c calculated in the exact same way as above.

t-test/Wald test

- Consider **any** asymptotically normal estimator $\hat{\theta}$ for parameter θ.
- Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- A size-α **t-test** (or **Wald test**) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{se}}[\hat{\theta}]}$$

- Critical value c calculated in the exact same way as above.
  - For $Z \sim \mathcal{N}(0, 1)$, let $c = z_{\alpha/2}$ such that $\mathbb{P}(Z \leq z_{\alpha/2}) = 1 - \alpha/2$.

t-test/Wald test

- Consider **any** asymptotically normal estimator $\hat{\theta}$ for parameter θ.
- Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.
- A size-α **t-test** (or **Wald test**) rejects H_0 when $|T_n| > c$ where

$$T_n = \frac{\hat{\theta} - \theta_0}{\widehat{\text{se}}[\hat{\theta}]}$$

- Critical value c calculated in the exact same way as above.
    - For $Z \sim \mathcal{N}(0, 1)$, let $c = z_{\alpha/2}$ such that $\mathbb{P}(Z \leq z_{\alpha/2}) = 1 - \alpha/2$.
- Size of the test converges to the **nominal size** as $n \to \infty$:

$$\mathbb{P}(|T_n| > z_{\alpha/2} \mid \theta_0) \to \alpha$$

3/ p-values

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.
  - We rejected null of no population diff-in-means ($H_0 : \tau = 0$) at $\alpha = 0.05$.

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.
  - We rejected null of no population diff-in-means ($H_0 : \tau = 0$) at $\alpha = 0.05$.
  - What about all the other levels like $\alpha = 0.01$?

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.
  - We rejected null of no population diff-in-means ($H_0 : \tau = 0$) at $\alpha = 0.05$.
  - What about all the other levels like $\alpha = 0.01$?
- Alternative: **p-values** are the probability of observing T_n or more extreme under H_0:

$$p = \begin{cases} 1 - G_0(T_n) & \text{if one-sided} \\ 2(1 - G_0(|T_n|)) & \text{if two-sided} \end{cases}$$

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.
    - We rejected null of no population diff-in-means ($H_0 : \tau = 0$) at $\alpha = 0.05$.
    - What about all the other levels like $\alpha = 0.01$?
- Alternative: **p-values** are the probability of observing T_n or more extreme under H_0:

$$p = \begin{cases} 1 - G_0(T_n) & \text{if one-sided} \\ 2(1 - G_0(|T_n|)) & \text{if two-sided} \end{cases}$$

- Interpretation: smallest size α at which a test would reject the null.

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.
  - We rejected null of no population diff-in-means ($H_0 : \tau = 0$) at $\alpha = 0.05$.
  - What about all the other levels like $\alpha = 0.01$?
- Alternative: **p-values** are the probability of observing T_n or more extreme under H_0:

$$p = \begin{cases} 1 - G_0(T_n) & \text{if one-sided} \\ 2(1 - G_0(|T_n|)) & \text{if two-sided} \end{cases}$$

- Interpretation: smallest size α at which a test would reject the null.
  - Can immediately assess tests of all sizes, no need for strict cutoffs.

Why p-values?

- Just rejecting or not rejecting the null hypothesis is not too informative.
  - We rejected null of no population diff-in-means ($H_0 : \tau = 0$) at $\alpha = 0.05$.
  - What about all the other levels like $\alpha = 0.01$?
- Alternative: **p-values** are the probability of observing T_n or more extreme under H_0:

$$p = \begin{cases} 1 - G_0(T_n) & \text{if one-sided} \\ 2(1 - G_0(|T_n|)) & \text{if two-sided} \end{cases}$$

- Interpretation: smallest size α at which a test would reject the null.
  - Can immediately assess tests of all sizes, no need for strict cutoffs.
  - A **continuous** measure of evidence against the null.

Calculate the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$.

Calculate the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$.
- How likely would it be to get a test statistic this extreme or more extreme if there were no treatment effect?

Calculate the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$.
- How likely would it be to get a test statistic this extreme or more extreme if there were no treatment effect?

$$\mathbb{P}(|T_n| > 18.5 \mid \tau_0) = \mathbb{P}(T_n > 18.5 \mid \tau_0) + \mathbb{P}(T_n < -18.5 \mid \tau_0)$$

Calculate the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$.
- How likely would it be to get a test statistic this extreme or more extreme if there were no treatment effect?

$$\begin{aligned}\mathbb{P}(|T_n| > 18.5 \mid \tau_0) &= \mathbb{P}(T_n > 18.5 \mid \tau_0) + \mathbb{P}(T_n < -18.5 \mid \tau_0) \\ &= 2 \times \mathbb{P}(T_n < -18.5 \mid \tau_0)\end{aligned}$$

Calculate the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$.
- How likely would it be to get a test statistic this extreme or more extreme if there were no treatment effect?

$$\begin{aligned}\mathbb{P}(|T_n| > 18.5 \mid \tau_0) &= \mathbb{P}(T_n > 18.5 \mid \tau_0) + \mathbb{P}(T_n < -18.5 \mid \tau_0) \\ &= 2 \times \mathbb{P}(T_n < -18.5 \mid \tau_0)\end{aligned}$$

- Use the `pnorm()` function:

Calculate the p-value

- Social pressure test statistic, $t_{\text{obs}} = 18.5$.
- How likely would it be to get a test statistic this extreme or more extreme if there were no treatment effect?

$$\begin{aligned} \mathbb{P}(|T_n| > 18.5 \mid \tau_0) &= \mathbb{P}(T_n > 18.5 \mid \tau_0) + \mathbb{P}(T_n < -18.5 \mid \tau_0) \\ &= 2 \times \mathbb{P}(T_n < -18.5 \mid \tau_0) \end{aligned}$$

- Use the `pnorm()` function:

```
2 * pnorm(-18.5)
```

```
## [1] 2.06e-76
```

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.
- p-values are **not**:

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.
- p-values are **not**:
    - An indication of a large substantive effect

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.
- p-values are **not**:
    - An indication of a large substantive effect
    - The probability that the null hypothesis is false

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.
- p-values are **not**:
    - An indication of a large substantive effect
    - The probability that the null hypothesis is false
    - The probability that the alternative hypothesis is true

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.
- p-values are **not**:
  - An indication of a large substantive effect
  - The probability that the null hypothesis is false
  - The probability that the alternative hypothesis is true
- p-values are just a transformation of the test statistic to the $[0, 1]$ scale.

Be careful with p-values

- Low p-value $\rightsquigarrow$ data unlikely given the null $\rightsquigarrow$ evidence against the null.
- p-values are **not**:
  - An indication of a large substantive effect
  - The probability that the null hypothesis is false
  - The probability that the alternative hypothesis is true
- p-values are just a transformation of the test statistic to the $[0, 1]$ scale.
- p-hacking controversy: not about p-values per se, but about significance cutoffs

4/ Power Analyses

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?
- Choose the *n* to ensure you can reject the null under a hypothesized effect size.

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?
- Choose the n to ensure you can reject the null under a hypothesized effect size.
    - Small effect sizes (half percentage point) will require huge n

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?
- Choose the n to ensure you can reject the null under a hypothesized effect size.
    - Small effect sizes (half percentage point) will require huge n
    - Large effect sizes (10 percentage points) will require smaller n

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?
- Choose the n to ensure you can reject the null under a hypothesized effect size.
    - Small effect sizes (half percentage point) will require huge n
    - Large effect sizes (10 percentage points) will require smaller n
- If we fail to reject a null hypothesis, two possible states of the world:

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?
- Choose the *n* to ensure you can reject the null under a hypothesized effect size.
    - Small effect sizes (half percentage point) will require huge *n*
    - Large effect sizes (10 percentage points) will require smaller *n*
- If we fail to reject a null hypothesis, two possible states of the world:
    - Null is true (no treatment effect)

Effect sizes

TABLE 2. Effects of Four Mail Treatments on Voter Turnout in the August 2006 Primary Election

	Experimental Group				
	Control	Civic Duty	Hawthorne	Self	Neighbors
Percentage Voting	29.7%	31.5%	32.2%	34.5%	37.8%
N of Individuals	191,243	38,218	38,204	38,218	38,201

- Why use sample sizes of 38,000 for each treatment condition?
- Choose the *n* to ensure you can reject the null under a hypothesized effect size.
  - Small effect sizes (half percentage point) will require huge *n*
  - Large effect sizes (10 percentage points) will require smaller *n*
- If we fail to reject a null hypothesis, two possible states of the world:
  - Null is true (no treatment effect)
  - Null is false (there is a treatment effect), but test had low power.

Power analysis

- **Power analysis**: evaluate the power function for various sample sizes.

Power analysis

- **Power analysis**: evaluate the power function for various sample sizes.
    - Prob. of rejecting different possible effects at different sample sizes.

Power analysis

- **Power analysis**: evaluate the power function for various sample sizes.
    - Prob. of rejecting different possible effects at different sample sizes.
    - Can be done before the experiment to plan for sample sizes

Power analysis

- **Power analysis**: evaluate the power function for various sample sizes.
  - Prob. of rejecting different possible effects at different sample sizes.
  - Can be done before the experiment to plan for sample sizes
- Easiest to see in a one-sided test of $H_0 : \theta = 0$.

Power analysis

- **Power analysis**: evaluate the power function for various sample sizes.
    - Prob. of rejecting different possible effects at different sample sizes.
    - Can be done before the experiment to plan for sample sizes
- Easiest to see in a one-sided test of $H_0 : \theta = 0$.
- Let $T_n = \hat{\theta}/\widehat{\text{se}}[\hat{\theta}]$ be the test statistic and the power function is:

$$\pi_n(\theta) = \mathbb{P}[T_n > c \mid \theta]$$

Power analysis

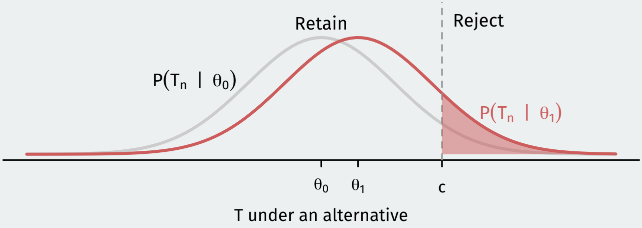
- **Power analysis**: evaluate the power function for various sample sizes.
  - Prob. of rejecting different possible effects at different sample sizes.
  - Can be done before the experiment to plan for sample sizes
- Easiest to see in a one-sided test of $H_0 : \theta = 0$.
- Let $T_n = \hat{\theta}/\widehat{\text{se}}[\hat{\theta}]$ be the test statistic and the power function is:

$$\pi_n(\theta) = \mathbb{P}[T_n > c \mid \theta]$$

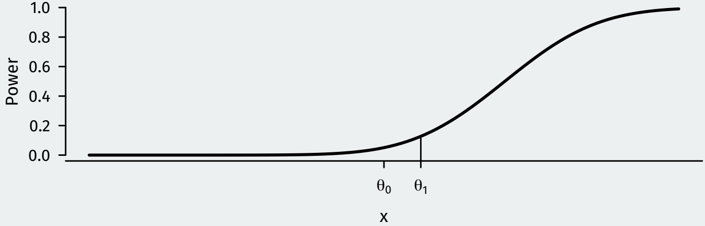
- If T_n is approximately $\mathcal{N}(0, 1)$ under the null, then under $H_1 : \theta = \theta_1$,

$$T_n \overset{a}{\sim} \mathcal{N}\left(\frac{\theta_1}{\widehat{\text{se}}[\hat{\theta}]}, 1\right) \qquad \rightsquigarrow \qquad \pi_n(\theta_1) = 1 - \Phi\left(c - \frac{\theta_1}{\widehat{\text{se}}[\hat{\theta}]}\right)$$

Power graph

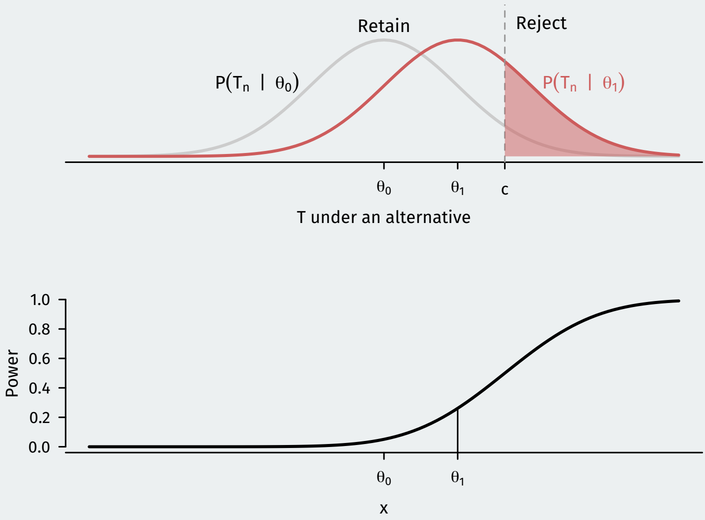


Retain
Reject
$P(T_n \mid \theta_0)$
$P(T_n \mid \theta_1)$
θ_0
θ_1
c
T under an alternative

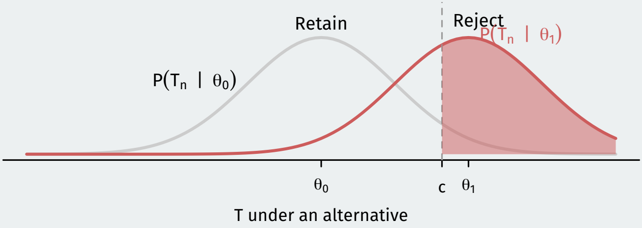


1.0
0.8
0.6
0.4
0.2
0.0
Power
θ_0
θ_1
x

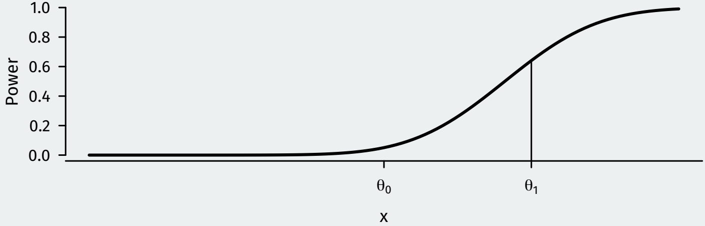
Power graph


Retain
Reject
P(T_n | θ_0)
P(T_n | θ_1)
θ_0
θ_1
c
T under an alternative
1.0
0.8
0.6
0.4
0.2
0.0
Power
θ_0
θ_1
x

Power graph

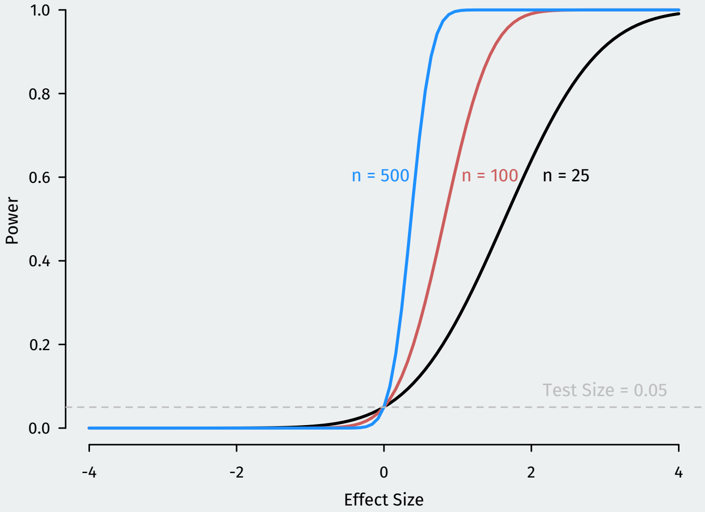


Retain
Reject
$P(T_n \mid \theta_1)$
$P(T_n \mid \theta_0)$
θ_0
c
θ_1
T under an alternative



1.0
0.8
0.6
0.4
0.2
0.0
Power
θ_0
θ_1
x

Power analysis



Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \ldots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \dots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
- Under null of $H_0 : \mu = \mu_0$, we have

$$T_n = \frac{\overline{X}_n - \mu_0}{s_n/\sqrt{n}} \sim t_{n-1}$$

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \ldots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
- Under null of $H_0 : \mu = \mu_0$, we have

$$T_n = \frac{\overline{X}_n - \mu_0}{s_n / \sqrt{n}} \sim t_{n-1}$$

  - **Student's t-distribution** with $n-1$ degrees of freedom.

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \dots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
- Under null of $H_0 : \mu = \mu_0$, we have

$$T_n = \frac{\overline{X}_n - \mu_0}{s_n / \sqrt{n}} \sim t_{n-1}$$

  - **Student's t-distribution** with $n - 1$ degrees of freedom.
- Null distribution is t so we use quantiles of t for critical values.

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \ldots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
- Under null of $H_0 : \mu = \mu_0$, we have

$$T_n = \frac{\overline{X}_n - \mu_0}{s_n / \sqrt{n}} \sim t_{n-1}$$

  - **Student's t-distribution** with $n-1$ degrees of freedom.
- Null distribution is t so we use quantiles of t for critical values.
  - For one-sided test $c = G_0^{-1}(1-\alpha)$ but now G_0 is t with $n-1$ df.

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \ldots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
- Under null of $H_0 : \mu = \mu_0$, we have

$$T_n = \frac{\overline{X}_n - \mu_0}{s_n/\sqrt{n}} \sim t_{n-1}$$

  - **Student's t-distribution** with $n - 1$ degrees of freedom.
- Null distribution is t so we use quantiles of t for critical values.
  - For one-sided test $c = G_0^{-1}(1 - \alpha)$ but now G_0 is t with $n - 1$ df.
  - Basically: use `qt()` instead of `qnorm()` for critical values.

Exact tests under the normal model

- Asymptotics are approximations. Can we ever get **exact** inferences at any sample size?
- Assume **parametric model**: $X_1, \dots, X_n$ are i.i.d. samples from $N(\mu, \sigma^2)$
- Under null of $H_0 : \mu = \mu_0$, we have

$$T_n = \frac{\overline{X}_n - \mu_0}{s_n/\sqrt{n}} \sim t_{n-1}$$

  - **Student's t-distribution** with $n-1$ degrees of freedom.
- Null distribution is t so we use quantiles of t for critical values.
  - For one-sided test $c = G_0^{-1}(1-\alpha)$ but now G_0 is t with $n-1$ df.
  - Basically: use `qt()` instead of `qnorm()` for critical values.
  - Asymptotically equivalent to using the normal, but more conservative

The shape of the t