

Intermediate Statistics

Hypothesis Testing

Alex Hoagland, PhD

February, 2025

We've now discussed estimation in two ways:

- 1 **Point Estimation**: A single value
- 2 **Confidence Intervals**: A range of values

Typically, we care more about the estimates **in reference to a given value**:

- Is the effect of a policy greater than 0?
- Does something affect women more than men?
- Do people prefer Popeye's chicken sandwiches over Chick-Fil-A?

We've now discussed estimation in two ways:

- 1 **Point Estimation**: A single value
- 2 **Confidence Intervals**: A range of values

Typically, we care more about the estimates **in reference to a given value**:

- Is the effect of a policy greater than 0?
- Does something affect women more than men?
- Do people prefer Popeye's chicken sandwiches over Chick-Fil-A?

All of these questions can be answered under the framework of **hypothesis testing**

Why Hypothesis Test?

In almost every case, our statistical analysis is driven towards **decision-making**:

- Should there be another trial?
- Should we change/renew the policy?
- Should I really get this PhD?

Why Hypothesis Test?

In almost every case, our statistical analysis is driven towards **decision-making**:

- Should there be another trial?
- Should we change/renew the policy?
- Should I really get this PhD?

Hypothesis testing **automates** (at least a little bit) **that process**.

HYPOTHESIS AND TEST PROCEDURES

What is Hypothesis Testing?

Statistical hypothesis testing is an algorithmic way of incorporating the **scientific method** into data analysis

- We have a benchmark hypothesis (**null hypothesis**, \mathcal{H}_0) that we start with

What is Hypothesis Testing?

Statistical hypothesis testing is an algorithmic way of incorporating the **scientific method** into data analysis

- We have a benchmark hypothesis (**null hypothesis**, \mathcal{H}_0) that we start with
- We want to test this against an **alternative hypothesis** (\mathcal{H}_a) using our collected data

What is Hypothesis Testing?

Statistical hypothesis testing is an algorithmic way of incorporating the **scientific method** into data analysis

- We have a benchmark hypothesis (**null hypothesis**, \mathcal{H}_0) that we start with
- We want to test this against an **alternative hypothesis** (\mathcal{H}_a) using our collected data
- If the estimated results are **drastically different** from the world of the null hypothesis, we **reject** it in favor of the alternative

What is Hypothesis Testing?

Statistical hypothesis testing is an algorithmic way of incorporating the **scientific method** into data analysis

- We have a benchmark hypothesis (**null hypothesis**, \mathcal{H}_0) that we start with
- We want to test this against an **alternative hypothesis** (\mathcal{H}_a) using our collected data
- If the estimated results are **drastically different** from the world of the null hypothesis, we **reject** it in favor of the alternative

As per the scientific method, a test has two outcomes: we either **reject \mathcal{H}_0** or **fail to reject** (never **accept** \mathcal{H}_0)

The Null Hypothesis, \mathcal{H}_0

The null hypothesis is representative of the **status quo**

- The test is performed to see if there is evidence in favor of something you (the researcher) believes to be **superior** to \mathcal{H}_0

The Null Hypothesis, \mathcal{H}_0

The null hypothesis is representative of the **status quo**

- The test is performed to see if there is evidence in favor of something you (the researcher) believes to be **superior** to \mathcal{H}_0
- Typically, \mathcal{H}_0 is associated with “null” effects:

The Null Hypothesis, \mathcal{H}_0

The null hypothesis is representative of the **status quo**

- The test is performed to see if there is evidence in favor of something you (the researcher) believes to be **superior** to \mathcal{H}_0
- Typically, \mathcal{H}_0 is associated with “null” effects:
 - ▶ The effect of a new policy is 0
 - ▶ The difference between groups is 0
 - ▶ A logistic regression coefficient (odds ratio) is different from 1

The null hypothesis is representative of the **status quo**

- The test is performed to see if there is evidence in favor of something you (the researcher) believes to be **superior** to \mathcal{H}_0
- Typically, \mathcal{H}_0 is associated with “null” effects:
 - ▶ The effect of a new policy is 0
 - ▶ The difference between groups is 0
 - ▶ A logistic regression coefficient (odds ratio) is different from 1

Typically, we write \mathcal{H}_0 as an **equality**: $\mathcal{H}_0 : \theta = \theta_0$

There are three ways to assert that a null hypothesis is misspecified:

- 1 $\theta > \theta_0$
- 2 $\theta < \theta_0$
- 3 $\theta \neq \theta_0$

There are three ways to assert that a null hypothesis is misspecified:

- 1 $\theta > \theta_0$
- 2 $\theta < \theta_0$
- 3 $\theta \neq \theta_0$

The first two are **one-sided** alternatives, the second is **two-sided**.

- The alternative hypothesis should be **contextually-motivated**
 - ▶ Like any good hypothesis, it needs a story
- When properly motivated, one-sided tests have more **power** because they put less strain on the testing procedure

A Generalized Testing Outline

To perform a hypothesis test properly, each of these elements must be **explicitly stated**:

- 1 The **null** and **alternative** hypotheses, \mathcal{H}_0 and \mathcal{H}_a

A Generalized Testing Outline

To perform a hypothesis test properly, each of these elements must be **explicitly stated**:

- 1 The **null** and **alternative** hypotheses, \mathcal{H}_0 and \mathcal{H}_a
- 2 A **test statistic**: the estimated value $\hat{\theta}$

A Generalized Testing Outline

To perform a hypothesis test properly, each of these elements must be **explicitly stated**:

- 1 The **null** and **alternative** hypotheses, \mathcal{H}_0 and \mathcal{H}_a
- 2 A **test statistic**: the estimated value $\hat{\theta}$
- 3 A **rejection region** that denotes what values of the test statistic will lead to a rejection of \mathcal{H}_0

A Generalized Testing Outline

To perform a hypothesis test properly, each of these elements must be **explicitly stated**:

- 1 The **null** and **alternative** hypotheses, \mathcal{H}_0 and \mathcal{H}_a
- 2 A **test statistic**: the estimated value $\hat{\theta}$
- 3 A **rejection region** that denotes what values of the test statistic will lead to a rejection of \mathcal{H}_0
 - ▶ Typically a **tail** of a distribution
 - ▶ Can therefore be denoted by a single parameter α , which we call the **significance level** (just like in CIs)
 - ▶ Must be **pre-specified**

A Testing Example

Suppose that we are concerned that there is gender pay inequality in an office. We collect data on wages and wish to test the difference $\mu_m - \mu_f$.

A Testing Example

Suppose that we are concerned that there is gender pay inequality in an office. We collect data on wages and wish to test the difference $\mu_m - \mu_f$.

- 1 $\mathcal{H}_0 : \mu_m - \mu_f = 0$ (no discrimination)
- 2 $\mathcal{H}_a : \mu_m - \mu_f \neq 0$ (some discrimination)

A Testing Example

Suppose that we are concerned that there is gender pay inequality in an office. We collect data on wages and wish to test the difference $\mu_m - \mu_f$.

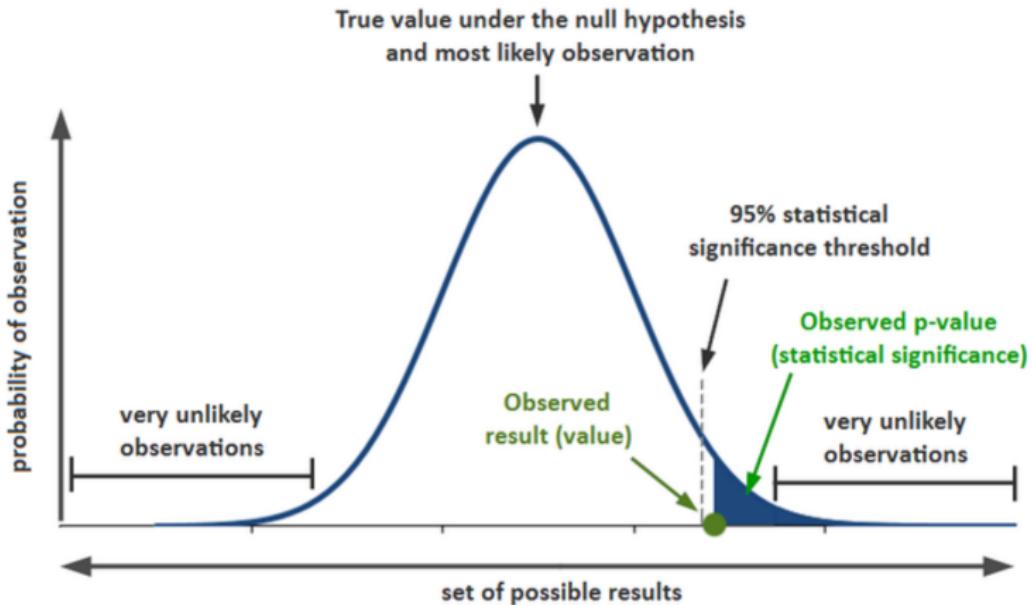
- 1 $\mathcal{H}_0 : \mu_m - \mu_f = 0$ (no discrimination)
- 2 $\mathcal{H}_a : \mu_m - \mu_f \neq 0$ (some discrimination)
- 3 Test statistic: $\bar{x}_m - \bar{x}_f$, an unbiased and consistent estimator of $\mu_m - \mu_f$

A Testing Example

Suppose that we are concerned that there is gender pay inequality in an office. We collect data on wages and wish to test the difference $\mu_m - \mu_f$.

- 1 $\mathcal{H}_0 : \mu_m - \mu_f = 0$ (no discrimination)
- 2 $\mathcal{H}_a : \mu_m - \mu_f \neq 0$ (some discrimination)
- 3 Test statistic: $\bar{x}_m - \bar{x}_f$, an unbiased and consistent estimator of $\mu_m - \mu_f$
- 4 Rejection region: $\alpha = 0.05$, so that we reject if the test statistic is in the 95th percentile or higher of its distribution *under the null*

A Testing Example



Source: [Steemit post](#)

Errors in Hypothesis Testing

The testing procedure should correctly identify when to reject/fail to reject most of the time. However, there is always potential for errors, to which we give specific names:

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Errors in Hypothesis Testing

Q: What do these errors mean in the context of our example?

Q: What do these errors mean in the context of our example?

- Generally, we can't demand error-free tests (too idealist!)
- Instead, we try to minimize $P(\text{Type 1 Error}) = \alpha$ and $P(\text{Type 2 Error}) = \beta$.

Q: What do these errors mean in the context of our example?

- Generally, we can't demand error-free tests (too idealist!)
- Instead, we try to minimize $P(\text{Type 1 Error}) = \alpha$ and $P(\text{Type 2 Error}) = \beta$.
- Since we've fixed a θ_0 under the null, this α is fixed (and is the same as α in previous slides!)
- β varies depending on the true parameter value θ (which we can't know!)

Q: What do these errors mean in the context of our example?

- Generally, we can't demand error-free tests (too idealist!)
- Instead, we try to minimize $P(\text{Type 1 Error}) = \alpha$ and $P(\text{Type 2 Error}) = \beta$.
- Since we've fixed a θ_0 under the null, this α is fixed (and is the same as α in previous slides!)
- β varies depending on the true parameter value θ (which we can't know!)
- There is a **tradeoff** in minimizing these errors—decreasing α usually requires increasing β for all θ

Which Error is Worse?

Q: Which error should we be more worried about?

- I get this question a lot
- The answer: **both!**

Which Error is Worse?

Q: Which error should we be more worried about?

- I get this question a lot
- The answer: **both!**

Consider a (grim) example of investigating a report for child maltreatment:

- What's the (binary) hypothesis we are testing?
- What's the Type 1 Error? Why is it bad?
- What's the Type 2 Error? Why is it bad?
- In what ways might policy be set up to prioritize one of these errors? Research? What should we do about it?

Which Error is Worse?

Q: Which error should we be more worried about?

- I get this question a lot
- The answer: **both!**

Consider a (grim) example of investigating a report for child maltreatment:

- What's the (binary) hypothesis we are testing?
- What's the Type 1 Error? Why is it bad?
- What's the Type 2 Error? Why is it bad?
- In what ways might policy be set up to prioritize one of these errors? Research? What should we do about it?

Unfortunately, research tends to bias us **towards Type II errors** (we'll revisit this when we get to power)

TESTS FOR MEANS AND PROPORTIONS

Tests for a Population Mean

So let's do our first **hypothesis tests!**

$$\mathcal{H}_0 : \mu = \theta_0$$

$$\mathcal{H}_a : \mu \neq \theta_0$$

Test Statistic : a function of \bar{x}

$$\alpha := 0.05$$

So let's do our first **hypothesis tests!**

$$\mathcal{H}_0 : \mu = \theta_0$$

$$\mathcal{H}_a : \mu \neq \theta_0$$

Test Statistic : a function of \bar{x}

$$\alpha := 0.05$$

We will cover **three cases** for this test

- 1 A normal population, σ known
- 2 Large sample tests (non-normal popln)
- 3 A normal population, σ unknown

Motivating Example

Throughout, we will use the example of [Card and Kreuger \(1994\)](#)

- Interested in a 1992 policy in NJ that raised minimum wage
- Our goal: measure its effects on employment

Motivating Example

Throughout, we will use the example of [Card and Kreuger \(1994\)](#)

- Interested in a 1992 policy in NJ that raised minimum wage
- Our goal: measure its effects on employment
- θ_0 = Avg employment rate **before** policy change
- \bar{x} = Avg employment rate **after** change

Motivating Example

Throughout, we will use the example of [Card and Kreuger \(1994\)](#)

- Interested in a 1992 policy in NJ that raised minimum wage
- Our goal: measure its effects on employment
- θ_0 = Avg employment rate **before** policy change
- \bar{x} = Avg employment rate **after** change
 - ▶ **Q:** If we reject \mathcal{H}_0 , what are we concluding?

Case 1: Normal Population, Known σ

This case is rarely true, but helpful to ground intuition.

Case 1: Normal Population, Known σ

This case is rarely true, but helpful to ground intuition.

Once we have collected data, we can calculate \bar{x}

- If \mathcal{H}_0 is true, then $\bar{x} \approx \theta_0$

This case is rarely true, but helpful to ground intuition.

Once we have collected data, we can calculate \bar{x}

- If \mathcal{H}_0 is true, then $\bar{x} \approx \theta_0$
- Hence, the standardized mean:

$$Z = \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}$$

is a good measure of **how far** estimator is from truth
when the null hypothesis is true

This case is rarely true, but helpful to ground intuition.

Once we have collected data, we can calculate \bar{x}

- If \mathcal{H}_0 is true, then $\bar{x} \approx \theta_0$
- Hence, the standardized mean:

$$Z = \frac{\bar{X} - \theta_0}{\sigma/\sqrt{n}}$$

is a good measure of **how far** estimator is from truth
when the null hypothesis is true

- Hence, if $|Z|$ is too large, we can reject \mathcal{H}_0

Calculating the Test Statistic

Suppose that $\theta_0 = 0.20$ with $\sigma = .10$.¹ We randomly sample 100 stores after the change and obtain a mean of 0.16 as the employment rate. We are interested in performing this test with 95% confidence, so $\alpha = 0.05$.

Q: What is the test statistic?

¹Card and Krueger looked at fast food employment as their case study, hence this low value for employment.

Calculating the Test Statistic

Suppose that $\theta_0 = 0.20$ with $\sigma = .10$.¹ We randomly sample 100 stores after the change and obtain a mean of 0.16 as the employment rate. We are interested in performing this test with 95% confidence, so $\alpha = 0.05$.

Q: What is the test statistic?

$$\begin{aligned} Z &= \frac{0.16 - 0.20}{0.10/10} \\ &= \frac{-0.04}{0.01} \\ &= -4 \end{aligned}$$

¹Card and Krueger looked at fast food employment as their case study, hence this low value for employment.

Calculating the Test Statistic

Suppose that $\theta_0 = 0.20$ with $\sigma = .10$.¹ We randomly sample 100 stores after the change and obtain a mean of 0.16 as the employment rate. We are interested in performing this test with 95% confidence, so $\alpha = 0.05$.

Q: What is the test statistic?

$$\begin{aligned} Z &= \frac{0.16 - 0.20}{0.10/10} \\ &= \frac{-0.04}{0.01} \\ &= -4 \end{aligned}$$

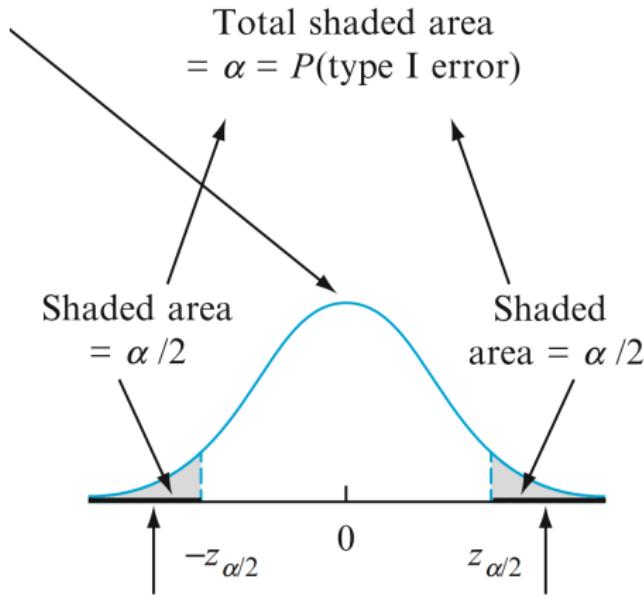
Okay, but **what does this mean?**

- This is where we need our **rejection region**

¹Card and Krueger looked at fast food employment as their case study, hence this low value for employment.

Rejection Regions

Our rejection region gives us **cutoff values** for the test stat:



Rejection region: either

$$z \geq z_{\alpha/2} \text{ or } z \leq -z_{\alpha/2}$$

Our rejection region gives us **cutoff values** for the test stat:

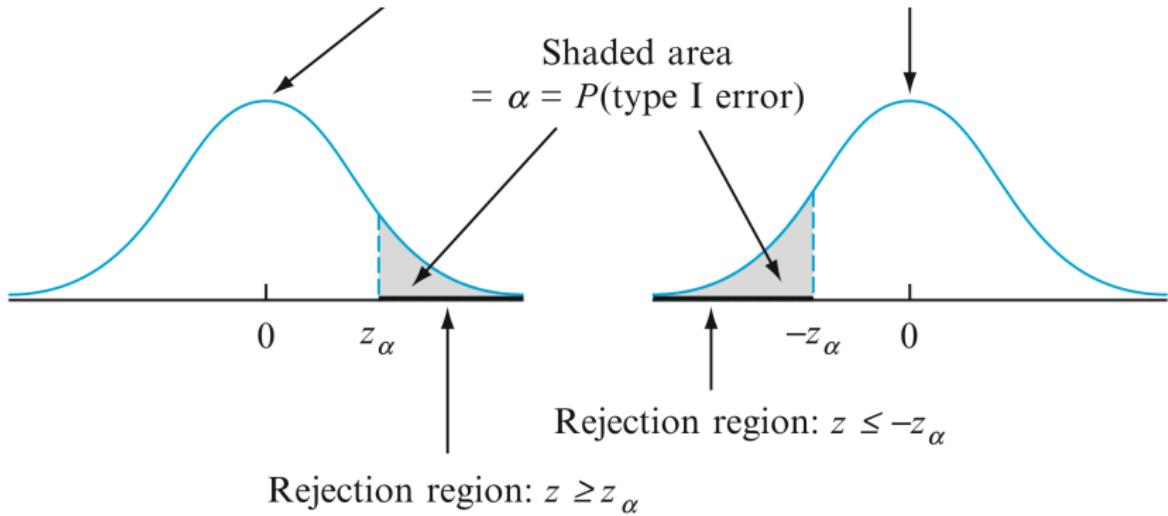
In this case, if $\alpha = 0.05$, then $Z_{0.05/2} = Z_{0.025} = \pm 1.96$.

- Since the test stat is **larger in absolute value** than the cutoff, we **reject the null hypothesis**
- In this case², what is the interpretation?

²Which is not exactly what Card and Krueger found.

One-Sided vs. Two-Sided Rejection Regions

Earlier, we performed a **two-sided** test. But what if we only want to test for a **decrease** in employment?



One-Sided vs. Two-Sided Rejection Regions

Earlier, we performed a **two-sided** test. But what if we only want to test for a **decrease** in employment?

A one-sided test has a **larger rejection region** in the area of interest

One-Sided vs. Two-Sided Rejection Regions

Earlier, we performed a **two-sided** test. But what if we only want to test for a **decrease** in employment?

A one-sided test has a **larger rejection region** in the area of interest

- Greater probability of rejecting the null hypothesis
- Hence, must be context-motivated: you *know* difference should be 0 or one direction
- If your results suggested the opposite direction, you would attribute that to random sampling!

One-Sided vs. Two-Sided Rejection Regions

Earlier, we performed a **two-sided** test. But what if we only want to test for a **decrease** in employment?

A one-sided test has a **larger rejection region** in the area of interest

- Greater probability of rejecting the null hypothesis
- Hence, must be context-motivated: you *know* difference should be 0 or one direction
- If your results suggested the opposite direction, you would attribute that to random sampling!
- **Q:** Does a one-sided test change our results for our example? (now, $Z_\alpha = -1.645$)

HYPOTHESIS TEST PROCEDURES

- 1 State the null and alternative hypotheses (context \Rightarrow math)
- 2 Choose the correct test statistic and give its formula under \mathcal{H}_0
- 3 State the significance level α and the associated rejection region (a picture may help)
- 4 Compute the test statistic under \mathcal{H}_0 and determine the results of the test (reject or fail to reject)
- 5 Interpret the test results (math \Rightarrow context)

Case 2: Large Sample Tests

When n is large enough, we can invoke the Central Limit Theorem to deal with potentially non-normal populations

- Same logic as for confidence intervals
- Since $s \approx \sigma$,

$$Z = \frac{\bar{X} - \theta_0}{S/\sqrt{n}}$$

has an **approximately** normal distribution. We can then use **the same** test procedure!

- Recall that $n > 30$ is our rule of thumb for "large samples"

What if we **can't** rely on the CLT?

What if we **can't** rely on the CLT?

- Just as with confidence intervals, we can assume population is **close to normal**
- In that case,

$$T = \frac{\bar{X} - \theta_0}{S/\sqrt{n}}$$

has a t_{n-1} distribution.

What if we **can't** rely on the CLT?

- Just as with confidence intervals, we can assume population is **close to normal**
- In that case,

$$T = \frac{\bar{X} - \theta_0}{S/\sqrt{n}}$$

has a t_{n-1} distribution.

We can use the same **test procedure**, but our **rejection regions** will change slightly.

Example: Tipping

One restaurant manager is very angry that he isn't receiving a lot of tips. He wants to test whether his customers are systematically leaving tips below the customary 15%. He doesn't want to get too angry without cause, so he sets $\alpha = 0.01$.

Q: What is the testing procedure?

Example: Tipping

One restaurant manager is very angry that he isn't receiving a lot of tips. He wants to test whether his customers are systematically leaving tips below the customary 15%. He doesn't want to get too angry without cause, so he sets $\alpha = 0.01$.

Q: What is the testing procedure?

$$\mathcal{H}_0 : \mu = 15$$

$$\mathcal{H}_a : \mu < 15$$

Test Stat :
$$\frac{\bar{X} - 15}{S/\sqrt{n}}$$

$$\alpha = 0.01$$

Example: Tipping

One restaurant manager is very angry that he isn't receiving a lot of tips. He wants to test whether his customers are systematically leaving tips below the customary 15%. He doesn't want to get too angry without cause, so he sets $\alpha = 0.01$.

Suppose that our **data** comes from $N = 16$ bills, and that it has moments $(\bar{x}, s^2) = (12, 23)$.

Q: What is the test statistic? The confidence region? The results of the test?

Example: Tipping

One restaurant manager is very angry that he isn't receiving a lot of tips. He wants to test whether his customers are systematically leaving tips below the customary 15%. He doesn't want to get too angry without cause, so he sets $\alpha = 0.01$.

Suppose that our **data** comes from $N = 16$ bills, and that it has moments $(\bar{x}, s^2) = (12, 23)$.

Q: What is the test statistic? The confidence region? The results of the test?

$$T = \frac{-3}{4.8/4} = \frac{-12}{4.8} = -2.5$$

$$t_{0.01, 15} = 2.602$$

We therefore **fail to reject** in this small sample.

Example: Tipping

One restaurant manager is very angry that he isn't receiving a lot of tips. He wants to test whether his customers are systematically leaving tips below the customary 15%. He doesn't want to get too angry without cause, so he sets $\alpha = 0.01$.

The owner isn't satisfied, so he collects a **larger sample** of $N = 100$ bills. The mean is still 12, but the variance grows to 144.

Q: What are the results of the test now?

Example: Tipping

One restaurant manager is very angry that he isn't receiving a lot of tips. He wants to test whether his customers are systematically leaving tips below the customary 15%. He doesn't want to get too angry without cause, so he sets $\alpha = 0.01$.

The owner isn't satisfied, so he collects a **larger sample** of $N = 100$ bills. The mean is still 12, but the variance grows to 144.

Q: What are the results of the test now?

The test statistic becomes:

$$Z = \frac{-3}{12/10} = \frac{-30}{12} = -2.5$$

$$Z_{0.01} = 2.326$$

We therefore **reject the null hypothesis** given the additional data.

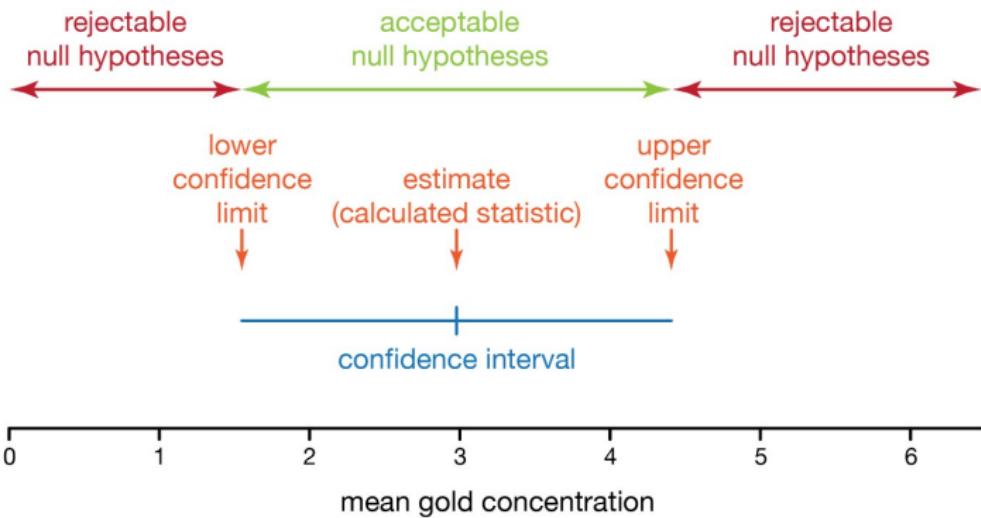
CONCLUDING THOUGHTS FOR TODAY

Relationship between \mathcal{H}_0 and confidence intervals

Do you notice a relationship between test statistics and confidence intervals?

- The margin of error in a CI is the same as the test statistic!
- That is, a test's critical values *define* a CI around \mathcal{H}_0

How does α relate to the level of confidence?

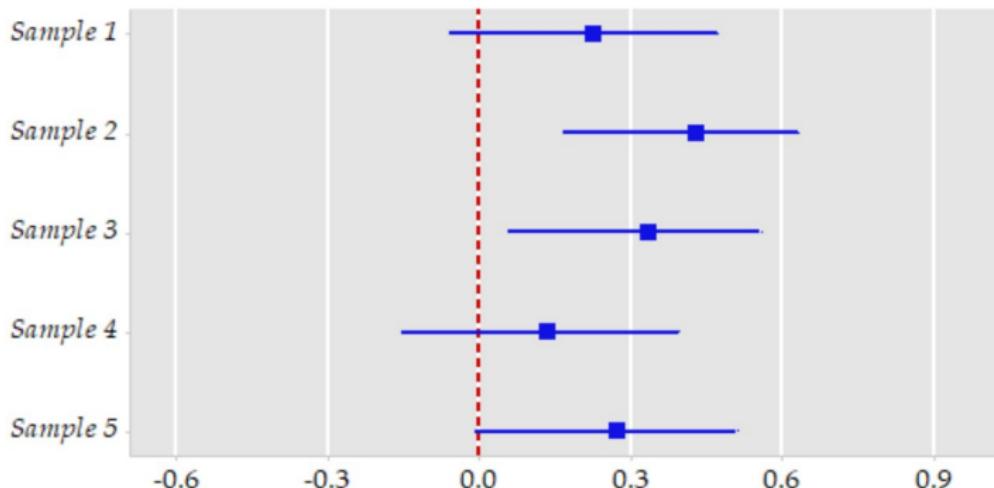


Relationship between \mathcal{H}_0 and confidence intervals

Do you notice a relationship between **test statistics** and **confidence intervals**?

- The margin of error in a CI is the same as the test statistic!
- That is, a test's critical values *define* a CI around \mathcal{H}_0

What is significant here? What is significantly different from other samples?



A General Version of the Test Statistic

For large samples, we can write the test statistic for any parameter θ against a null $\mathcal{H}_0 : \theta = \theta_0$ in a general way:

$$\text{Test Statistic} : \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

A General Version of the Test Statistic

For large samples, we can write the test statistic for any parameter θ against a null $\mathcal{H}_0 : \theta = \theta_0$ in a general way:

$$\text{Test Statistic} : \frac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$$

When $\sigma_{\hat{\theta}}$ involves unknown parameters, we replace it with its **estimator** $\hat{\sigma}_{\hat{\theta}}$

Example: Large Sample Tests for a Proportion

Suppose that we are interested in testing **proportions**, rather than means

- We test whether p is different from some p_0 .
- Based on the assumption that $\hat{p} = X/n$ is approximately normally distributed when n is large enough

Example: Large Sample Tests for a Proportion

Suppose that we are interested in testing **proportions**, rather than means

- We test whether p is different from some p_0 .
- Based on the assumption that $\hat{p} = X/n$ is approximately normally distributed when n is large enough

Given the general test for large samples, we can write the test statistic as:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

Example: Large Sample Tests for a Proportion

Suppose that we are interested in testing **proportions**, rather than means

- We test whether p is different from some p_0 .
- Based on the assumption that $\hat{p} = X/n$ is approximately normally distributed when n is large enough

Given the general test for large samples, we can write the test statistic as:

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

- In this special case, we **do not** need to estimate $\sigma_{\hat{p}}$
- For small samples, would need to use the binomial distribution directly \Rightarrow can be quite tedious

TWO-SAMPLE TESTS

- Two-sample hypothesis testing is used to compare two population means or proportions.
- Applications include medical trials, A/B testing, demographic differences, balances, etc.
- Remember that **so much of statistics is some version of a two-sample test!**
- Assumptions depend on whether samples are **independent** or **paired**.

Null and Alternative Hypotheses:

- $H_0 : \mu_1 = \mu_2$ (no difference in population means)
- $H_A : \mu_1 \neq \mu_2, \mu_1 > \mu_2$, or $\mu_1 < \mu_2$ (difference in means)

Test Statistics:

- For independent samples (assuming equal variances):

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Pooled standard deviation:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Equal Variance Assumption

- When conducting a two-sample t -test, assuming equal variances across samples may get you narrower confidence intervals and simpler math
- But R can handle anything!
- Note: recall that if sample sizes are equal, the test is fairly robust to variance differences.
- Otherwise, want to default to Welch's t -test, which does not assume equal variance:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Levene's test or F-test can be used to check for equality of variances before choosing the appropriate test.

Matched Pairs Testing

- Used when the samples are dependent (paired observations, such as pre-test and post-test data).
- Hypotheses:

$$H_0 : \mu_D = 0, \quad H_A : \mu_D \neq 0$$

where μ_D is the mean difference between paired observations.

- Test statistic:

$$t = \frac{\bar{D}}{s_D / \sqrt{n}}$$

where \bar{D} is the mean difference, s_D is the standard deviation of differences, and n is the number of pairs.

- Often more powerful than independent sample tests due to reduced variability.

HANDLING CATEGORICAL DATA

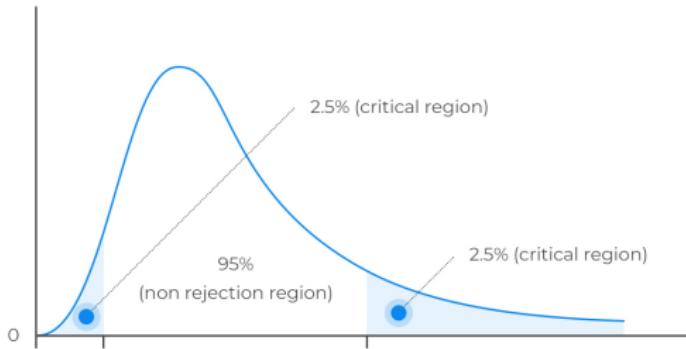
Introduction to Chi-Square Tests

What if our data isn't (quasi)-continuous? How can we adapt testing?

- Chi-square tests are used to analyze categorical data.
- Sort of a hybrid between t -tests and proportion testing (looking at frequencies across discrete buckets)



Two-tailed Chi-Square test (5% significance)



Example:

- Used to determine if a categorical variable follows a specified distribution.
- Null Hypothesis (H_0): The observed distribution follows the expected distribution.
- Test Statistic:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed count and E_i is the expected count.

- The test statistic follows a chi-square distribution with $k - 1$ degrees of freedom.

- Used to determine if two categorical variables are independent.
- Null Hypothesis (H_0): The two categorical variables are independent.
- Test Statistic:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed count in cell (i, j) and E_{ij} is the expected count computed as:

$$E_{ij} = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

- The test statistic follows a chi-square distribution with $(r - 1)(c - 1)$ degrees of freedom.

- Use `chisq.test()` for both goodness of fit (distributional) and independence tests.
- Need to have a clear rationale for the test—why do we need to generalize beyond a t -test?
- If your research question can be simplified, this is a good thing!

P-VALUES

How Strong is Our Evidence?

Our testing procedure so far has been:

- 1 Choose a significance level α

How Strong is Our Evidence?

Our testing procedure so far has been:

- 1 Choose a significance level α
- 2 This specifies the rejection region for \mathcal{H}_0

How Strong is Our Evidence?

Our testing procedure so far has been:

- 1 Choose a significance level α
- 2 This specifies the rejection region for \mathcal{H}_0

Is there a more **intuitive approach** to hypothesis testing?

How Strong is Our Evidence?

Our testing procedure so far has been:

- 1 Choose a significance level α
- 2 This specifies the rejection region for \mathcal{H}_0

Is there a more **intuitive approach** to hypothesis testing?

- Motulsky starts with intuition of a ***p-value*** (Ch. 15)
- We compare a **test statistic** T to a **hypothetical distribution**
 - ▶ If the test stat is “unlikely” given distribution for \mathcal{H}_0 , we reject

How Strong is Our Evidence?

Our testing procedure so far has been:

- 1 Choose a significance level α
- 2 This specifies the rejection region for \mathcal{H}_0

Is there a more **intuitive approach** to hypothesis testing?

- Motulsky starts with intuition of a ***p-value*** (Ch. 15)
- We compare a **test statistic** T to a **hypothetical distribution**
 - ▶ If the test stat is “unlikely” given distribution for \mathcal{H}_0 , we reject
- We can therefore define the **probability** $\Pr(T|\mathcal{H}_0)$.
 - ▶ What is probability we see results as extreme as T given \mathcal{H}_0 ?

Definition. The *p-value* is the probability **conditional on the null being true** of observing a test statistic at least as contradictory as T .

Definition. The *p-value* is the probability **conditional on the null being true** of observing a test statistic at least as contradictory as T .

- The smaller the *p-value*, the more likely we should **reject** \mathcal{H}_0 .
- Now, α defines a “**sufficiently small**” probability for rejection

Example: Tipping

Last class, we had an angry manager concerned about tips. We tested a stat of $Z = -2.5$ against a null hypothesis of $\mu = 15\%$. The p -value is thus

$$p\text{-value} = \mathbb{P}(Z \leq -2.5) \text{ when } \mu = 15.$$

Example: Tipping

Last class, we had an angry manager concerned about tips. We tested a stat of $Z = -2.5$ against a null hypothesis of $\mu = 15\%$. The p -value is thus

$$p\text{-value} = \mathbb{P}(Z \leq -2.5) \text{ when } \mu = 15.$$

When we can use the CLT, $Z \sim \mathcal{N}(0, 1)$ (**Q:** What if we can't?), so

$$\begin{aligned} p\text{-value} &\approx \text{area under the } z \text{ curve to the left of } -2.5 \\ &= \Phi(-2.5) = 0.006. \end{aligned}$$

Example: Tipping

Last class, we had an angry manager concerned about tips. We tested a stat of $Z = -2.5$ against a null hypothesis of $\mu = 15\%$. The p -value is thus

$$p\text{-value} = \mathbb{P}(Z \leq -2.5) \text{ when } \mu = 15.$$

When we can use the CLT, $Z \sim \mathcal{N}(0, 1)$ (**Q:** What if we can't?), so

$$\begin{aligned} p\text{-value} &\approx \text{area under the } z \text{ curve to the left of } -2.5 \\ &= \Phi(-2.5) = 0.006. \end{aligned}$$

There is a **0.6% chance** of getting Z under \mathcal{H}_0 !

- Hence, for any $\alpha \geq 0.006$, we would **reject** \mathcal{H}_0 .
- Since p defines the smallest α for which we reject, we occasionally call it the **observed significance level** of the data

When $p < \alpha$ (so that we **reject** the null), we generally refer to p as **significant**

- Otherwise, p is **insignificant**

When $p < \alpha$ (so that we **reject** the null), we generally refer to p as **significant**

- Otherwise, p is **insignificant**

R can perform all the steps of a hypothesis test for you!

- It even calculates p -values for you!
- Let's check this out in R.

Like with confidence intervals, **precision of language** is important here:

Like with confidence intervals, **precision of language** is important here: The p -value **is not**:

- The probability that \mathcal{H}_0 is true (we assumed \mathcal{H}_0 !)
- The probability that we reject \mathcal{H}_0 (high $p \not\rightarrow \mathcal{H}_0$ is true)

Like with confidence intervals, **precision of language** is important here: The p -value **is not**:

- The probability that \mathcal{H}_0 is true (we assumed \mathcal{H}_0 !)
- The probability that we reject \mathcal{H}_0 (high $p \not\rightarrow \mathcal{H}_0$ is true)

Instead, the p -value **is**:

- The conditional probability that we obtain T if \mathcal{H}_0 is true
- It is therefore **hypothesis dependent**

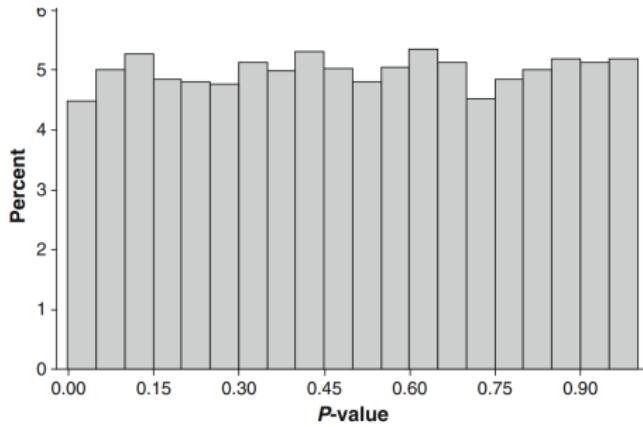
Like with confidence intervals, **precision of language** is important here: The p -value **is not**:

- The probability that \mathcal{H}_0 is true (we assumed \mathcal{H}_0 !)
- The probability that we reject \mathcal{H}_0 (high $p \not\rightarrow \mathcal{H}_0$ is true)

Instead, the p -value **is**:

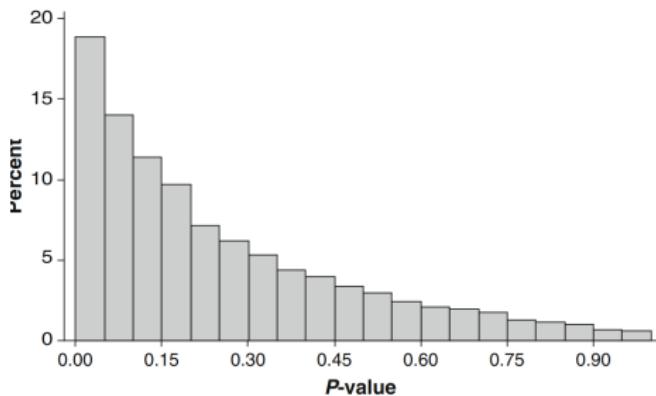
- The conditional probability that we obtain T if \mathcal{H}_0 is true
- It is therefore **hypothesis dependent**
- The p -value is also **sample dependent**, so it is a random variable
 - ▶ We talk about its **distribution** under two cases:
 - ▶ \mathcal{H}_0 is **true**
 - ▶ \mathcal{H}_0 is **false**

The Distribution of p -values



When \mathcal{H}_0 is true, the p -value has a **uniform** distribution

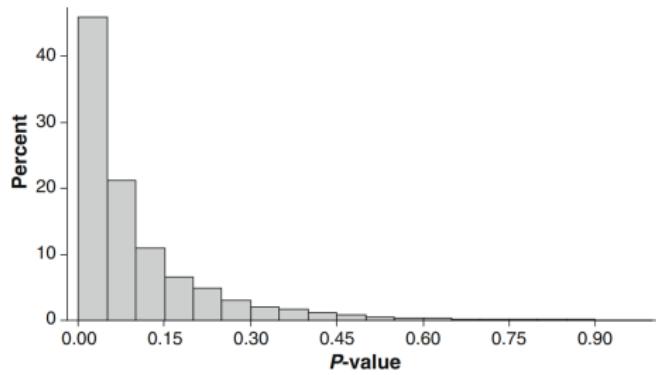
The Distribution of p -values



When \mathcal{H}_0 is “slightly” false, the p -value is skewed towards lower values

Q: How likely are we to reject the null hypothesis here (for $\alpha = 0.05$)? What type of error are we likely to commit?

The Distribution of p -values



When \mathcal{H}_0 is “more” false, the p -value becomes more skewed

Q: What is the probability of a Type II error now?

HOW TO CHOOSE A TEST

Now that we've covered the algorithm for testing, let's talk **interpretation**

- What does it mean for my project when $p < \alpha$?
- Can I rig the system to get what I want?
- What happens to my tests if its assumptions are violated?

Now that we've covered the algorithm for testing, let's talk **interpretation**

- What does it mean for my project when $p < \alpha$?
- Can I rig the system to get what I want?
- What happens to my tests if its assumptions are violated?

These are very important questions in academic research. Today:

- 1 Statistical vs. economic significance
- 2 p -hacking and publication bias
- 3 Likelihood ratio tests and general test statistics

What does “significance” mean?

So you've rejected your first null hypothesis. Does this make you a big shot?

- **Strong** rejection of \mathcal{H}_0 is associated with a **small** p -value
- However, p -values decrease with sample size!

What does “significance” mean?

So you’ve rejected your first null hypothesis. Does this make you a big shot?

- **Strong** rejection of \mathcal{H}_0 is associated with a **small** p -value
- However, p -values decrease with sample size!

Suppose we test $\mathcal{H}_0 : \bar{x} = 100$ against $\mathcal{H}_a : \bar{x} > 100$

n	P-value when $\bar{x} = 101$
25	.3085
100	.1587
400	.0228
900	.0013
1600	.0000335
2500	.000000297
10,000	7.69×10^{-24}

With $n > 400$, we’ve rejected, but the rejection looks **stronger** as $n \rightarrow \infty$

Significant, **Very** Significant, or **Extremely** Significant?

How much stock do we put in the p -value?

Symbol	Phrase	p
+	Approaching Significance	$p < .1$
*	Significant	$p < 0.05$
**	Very Significant	$p < .01$
***	Extremely Significant	$p < .001$

Significant, *Very* Significant, or *Extremely* Significant?

How much stock do we put in the p -value?

Symbol	Phrase	p
+	Approaching Significance	$p < .1$
*	Significant	$p < 0.05$
**	Very Significant	$p < .01$
***	Extremely Significant	$p < .001$

- Remember, hypothesis testing is a *binary* decision—can't make it continuous now!
- What does this mean in the context of big data?
- What does it mean in the context of your project?

So the difference may be (even extremely) statistically significant as $n \rightarrow \infty$. But is going from 100 to 101 a lot?

Economic Significance & Big Data

So the difference may be (even extremely) statistically significant as $n \rightarrow \infty$. But is going from 100 to 101 **a lot**?



So the difference may be (even extremely) statistically significant as $n \rightarrow \infty$. But is going from 100 to 101 a lot?

The context here determines the economic significance of a test:

- We care not only if the data detect a difference, but also what that difference means in the real world
- In the world of big data ($n > 10^6$), this is an especially large concern!

Example ideas:

- Women's labor force participation increases their probability of divorce...by 0.33% ([Dong, 2018](#))

POWER AND P -HACKING

What does “insignificant” mean?

We already know failure to reject $\not\rightarrow$ accepting \mathcal{H}_0

- So **no evidence of a difference** $\not\rightarrow$ **no difference**

What does “insignificant” mean?

We already know failure to reject $\not\rightarrow$ accepting \mathcal{H}_0

- So no evidence of a difference $\not\rightarrow$ no difference

For example:

The other day I shot baskets with Michael Jordan. He shot 7 straight free throws; I hit 3 and missed 4 and then rushed to the sideline, grabbed my laptop, and calculated a P value of .07.

Does this mean there is no difference? **Of course not!**

- What is one big statistical issue here?

What if the two had thrown many more free throws?

- $\mathcal{H}_{\text{Alex}}$: Differences would have started to emerge
- How big does my sample need to be?

What if the two had thrown many more free throws?

- $\mathcal{H}_{\text{Alex}}$: Differences would have started to emerge
- How big does my sample need to be?

This is the concept of **statistical power**: how big do you need N to observe the difference you're after?

- Depends on your question (1-sided or 2? What type of estimate? How large is economically meaningful?)
- Rule of thumb: if $N \uparrow 4x$, CIs $\downarrow 2x$

What if the two had thrown many more free throws?

- $\mathcal{H}_{\text{Alex}}$: Differences would have started to emerge
- How big does my sample need to be?

This is the concept of **statistical power**: how big do you need N to observe the difference you're after?

- Depends on your question (1-sided or 2? What type of estimate? How large is economically meaningful?)
- Rule of thumb: if $N \uparrow 4x$, CIs $\downarrow 2x$
- Formally, the **power** of a test, $1 - \beta$, is
 $\Pr(\text{reject } \mathcal{H}_0 | \mathcal{H}_0 \text{ is false}, N)$

Thinking about Sample Size

In a lot of cases, we can't choose N . But when can we?

Thinking about Sample Size

In a lot of cases, we can't choose N . But when can we?

- RCTs
- Surveys
- How many data requests we send to ICES, etc.

How do we know how big our sample needs to be?

Thinking about Sample Size

In a lot of cases, we can't choose N . But when can we?

- RCTs
- Surveys
- How many data requests we send to ICES, etc.

How do we know how big our sample needs to be?

For a simple hypothesis test (one mean), look at the **margin of error** of a CI:

$$ME = \frac{CV * SD}{\sqrt{N}}$$
$$\Rightarrow N = 4 \left(\frac{SD}{CV} \right)^2$$

Power Calculations

β (remember what that is?) depends on sample size (N) but also:

- randomness/dispersion in data
- actual size of hypothesized difference

Power Calculations

β (remember what that is?) depends on sample size (N) but also:

- randomness/dispersion in data
- actual size of hypothesized difference

Suppose we run a lot of experiments. How do we calculate power given these?

	Reject \mathcal{H}_0	Don't Reject \mathcal{H}_0	Total
\mathcal{H}_0 is true	A	B	A + B
\mathcal{H}_0 is false	C	D	C + D

Power Calculations

β (remember what that is?) depends on sample size (N) but also:

- randomness/dispersion in data
- actual size of hypothesized difference

Suppose we run a lot of experiments. How do we calculate power given these?

	Reject \mathcal{H}_0	Don't Reject \mathcal{H}_0	Total
\mathcal{H}_0 is true	A	B	A + B
\mathcal{H}_0 is false	C	D	C + D

If you specify N , SDs, and α , you can calculate power yourself!

Q: Does 90% power mean 90% of observations in treated group experience difference?

Q: Does 90% power mean 90% of observations in treated group experience difference?

- No! It means 90% of experiments would have significant estimates at α .

Q: Does 90% power mean 90% of observations in treated group experience difference?

- No! It means 90% of experiments would have significant estimates at α .

Q: How much power do I need?

Q: Does 90% power mean 90% of observations in treated group experience difference?

- No! It means 90% of experiments would have significant estimates at α .

Q: How much power do I need?

- In the age of big data, probably more like 90% (but think about this in the context of Type I and Type II errors — what are we saying about their relative value?)

Q: Does 90% power mean 90% of observations in treated group experience difference?

- No! It means 90% of experiments would have significant estimates at α .

Q: How much power do I need?

- In the age of big data, probably more like 90% (but think about this in the context of Type I and Type II errors — what are we saying about their relative value?)

Q: Why does power decrease with α and increase with N ? (Can you answer this one?)

Another issue related to hypothesis testing is *p-hacking*: pushing your research design until you get a significant result.

There are **two** main types of *p*-hacking:

- 1 Performing too many different kinds of tests and zeroing in on the significant ones
- 2 Performing the same test too many times

Another issue related to hypothesis testing is *p-hacking*: pushing your research design until you get a significant result.

There are **two** main types of *p*-hacking:

- 1 Performing too many different kinds of tests and zeroing in on the significant ones
- 2 Performing the same test too many times

Both types forget that p has a uniform distribution under \mathcal{H}_0 ! We are more likely to get **false positives** if we repeatedly sample

Example 1: The Sad Grad Student

Suppose that you are a grad student who desperately wants a good research project (to get a good job). You run an experiment to look at an effect of X on Y :

- 1 First, you test if the correlation of X and Y are positive.
Nothing

Example 1: The Sad Grad Student

Suppose that you are a grad student who desperately wants a good research project (to get a good job). You run an experiment to look at an effect of X on Y :

- 1 First, you test if the correlation of X and Y are positive.
Nothing
- 2 Then, you control for a bunch of other factors. **Still nothing**

Example 1: The Sad Grad Student

Suppose that you are a grad student who desperately wants a good research project (to get a good job). You run an experiment to look at an effect of X on Y :

- 1 First, you test if the correlation of X and Y are positive.
Nothing
- 2 Then, you control for a bunch of other factors. **Still nothing**
- 3 You go out and gather more data. You drop some data points that look clearly wrong. **A little closer**

Example 1: The Sad Grad Student

Suppose that you are a grad student who desperately wants a good research project (to get a good job). You run an experiment to look at an effect of X on Y :

- 1 First, you test if the correlation of X and Y are positive.
Nothing
- 2 Then, you control for a bunch of other factors. **Still nothing**
- 3 You go out and gather more data. You drop some data points that look clearly wrong. **A little closer**
- 4 You notice that a subset of your data seems to exhibit the effect, so you re-test on that sample. **Eureka!**

So **what have you proven?**

Example 1: The Sad Grad Student

Suppose that you are a grad student who desperately wants a good research project (to get a good job). You run an experiment to look at an effect of X on Y :

- 1 First, you test if the correlation of X and Y are positive.
Nothing
- 2 Then, you control for a bunch of other factors. **Still nothing**
- 3 You go out and gather more data. You drop some data points that look clearly wrong. **A little closer**
- 4 You notice that a subset of your data seems to exhibit the effect, so you re-test on that sample. **Eureka!**

So **what have you proven?**

- Likely **nothing**.
- You changed the test 4 times, so you drew from the distribution of p 4 times. Hence, p is no longer interpretable.

Example 2: Wild Heterogeneity Search

You've recognized the error of your ways and moved on to a new research question: how job training affects employees.

- You'll only perform the test once
- But you don't want to miss any results:

Example 2: Wild Heterogeneity Search

You've recognized the error of your ways and moved on to a new research question: how job training affects employees.

- You'll only perform the test once
- But you don't want to miss any results:
 - ▶ What if job training only affects women? Racial minorities? Executives? Construction workers?

Example 2: Wild Heterogeneity Search

You've recognized the error of your ways and moved on to a new research question: how job training affects employees.

- You'll only perform the test once
- But you don't want to miss any results:
 - ▶ What if job training only affects women? Racial minorities? Executives? Construction workers?
 - ▶ To make sure you don't miss anything, you'll divide the population into m subgroups and test the effect of the program on each of them
- **What is the problem?**

Example 2: Wild Heterogeneity Search

You've recognized the error of your ways and moved on to a new research question: how job training affects employees.

- You'll only perform the test once
- But you don't want to miss any results:
 - ▶ What if job training only affects women? Racial minorities? Executives? Construction workers?
 - ▶ To make sure you don't miss anything, you'll divide the population into m subgroups and test the effect of the program on each of them
- **What is the problem?**

Lots of research does this!

- Let's check out this example

Publication Bias and Scrutinizing Research

Unfortunately, we like to publish significant results

(1) Full sample

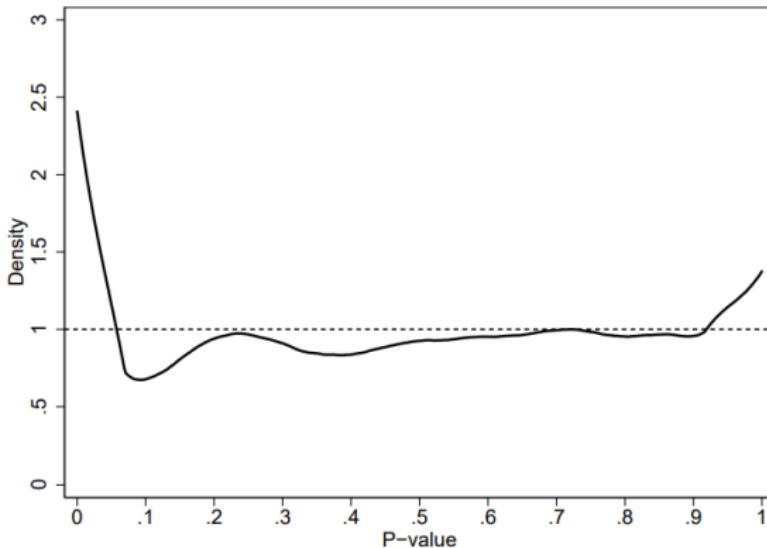


Figure: Source: [Snyder & Zhuo, 2018](#)

There have been lots of discussions lately about **publication bias** and **replication crises** in academic research.

There have been lots of discussions lately about **publication bias** and **replication crises** in academic research.

So **what** should you do about it?

- 1 Play around with [this interactive tool](#) to practice your own *p*-hacking!
- 2 Use your testing power wisely!
- 3 Read research carefully!

TESTING TAKEAWAYS

- 1 There is a pre-specified way to do a test
- 2 Need to know what you're assuming to interpret test correctly
- 3 Need to be careful of performing too many tests
- 4 Need to know what you're actually finding with your results.

