Hypothesis Testing and Confidence Interval

Zeyang (Arthur) Yu

Princeton University

November 13, 2024

Outline

1. Hypothesis Testing and Confidence Interval: Motivation
2. Standard Error
3. Hypothesis Testing
4. Confidence Interval
5. Duality: Confidence Interval and Hypothesis Testing

Outline

1. Hypothesis Testing and Confidence Interval: Motivation

2. Standard Error

3. Hypothesis Testing

4. Confidence Interval

5. Duality: Confidence Interval and Hypothesis Testing

Hypothesis Testing and C.I.: Motivation

Recall: two canonical problems in mathematical statistics

- Sample: $\{X_i\}_{i=1}^n$ distributed according to P (population *dist.*)

- Assume $\{X_i\}_{i=1}^n$ is an (aka, *iid*) sample

- "Learn" some "features" of P (e.g., a *param.* $\theta(P)$) from the data

- $\theta(P)$ can be either a descriptive or a causal *param.*

- Test a hypothesis about $\theta(P)$

- Quantify the strength of evidence from data against a statement

- Give a binary decision: reject or not reject the statement

  Want to control the errors of making such binary decision

- Construct a confidence region for $\theta(P)$

- Construct a random set C_n based on the data

- The random set covers the true *param.* with pre-specified *prob.*

$$P(\theta(P) \in C_n) \geq 1 - \alpha, \alpha \in (0, 1)$$

Outline

1 Hypothesis Testing and Confidence Interval: Motivation

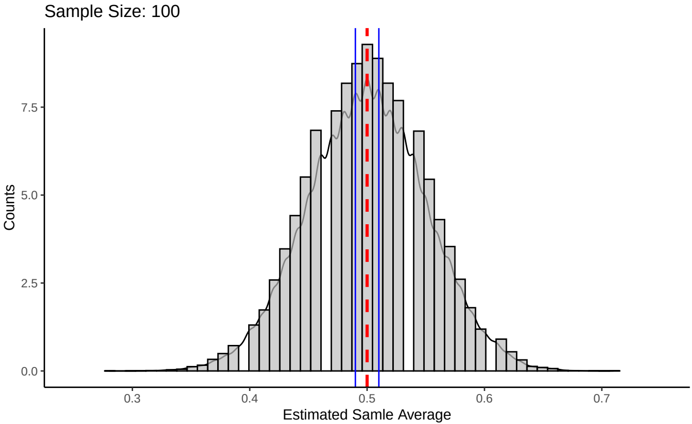
2 Standard Error

3 Hypothesis Testing

4 Confidence Interval

5 Duality: Confidence Interval and Hypothesis Testing

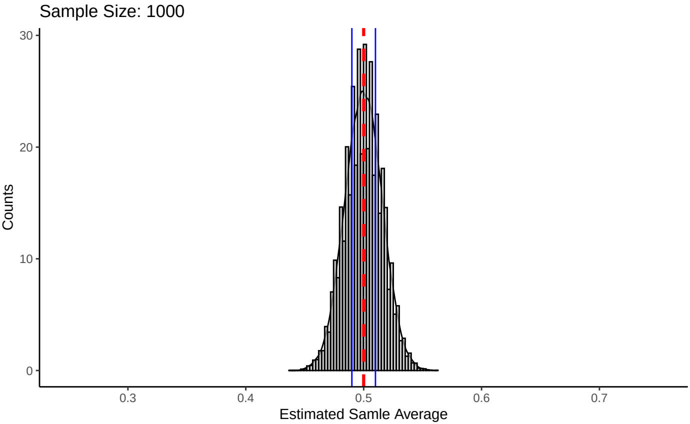
Standard Error: Motivation



Simulation Setup

- $\{X_i\}_{i=1}^{100}$ is *iid* draw from Ber(0.5)
- $\theta(P) = E(X_i)$, $\hat{\theta}_n = \frac{\sum_{i=1}^{n} X_i}{n}$

Standard Error: Motivation (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{1000}$ is *iid* draw from Ber(0.5)
- $\theta(P) = E(X_i)$, $\hat{\theta}_n = \frac{\sum_{i=1}^{n} X_i}{n}$

Standard Error: Definitions

What we've learned from previous simulations

- Given that the observed $\hat{\theta}_n$ being close to true $\theta(P) = 0.5$
  - Need to think the variability of $\hat{\theta}_n$ to gauge strength of evidence
  - $\hat{\theta}_n$ close to $\theta(P)$ but is noisy: NO GOOD!
- Thus, demand for a new "jargon": variability of $\hat{\theta}_n$
  - Naturally, apply the idea of variance to *r.v.* $\hat{\theta}_n$

Standard error: definition

The standard deviation of an estimator $\hat{\theta}_n$ is called the standard error, denoted by se:

$$\text{se} = \sqrt{\text{Var}\left(\hat{\theta}_n\right)}.$$

The estimated standard error is denoted by $\widehat{\text{se}}$: $\widehat{\text{se}} = \sqrt{\widehat{\text{Var}}\left(\hat{\theta}_n\right)}$.

Standard Error: Examples

se of sample mean

- se of sample mean:

$$\text{se of } \frac{1}{n}\sum_{i=1}^{n} X_i = \sqrt{\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)} = \sqrt{\frac{\text{Var}(X_i)}{n}} = \frac{\sigma}{\sqrt{n}},$$

  where σ is the standard deviation of X_i.

- Of course, we do not know σ, thus, needs to estimate from data
- $\widehat{\text{se}}$ of sample mean:

$$\widehat{\text{se}} \text{ of } \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{\hat{\sigma}}{\sqrt{n}},$$

  where $\hat{\sigma}$ is: $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$.

Standard Error: Examples (Conti.)

$\widehat{\text{se}}$ of difference in means

- Dif in means: causal vs. descriptive
- n_0 *iid* samples $\{X_i\}_{i=1}^{n_0}$ and n_1 *iid* samples $\{Y_i\}_{i=1}^{n_1}$

- $\{X_i\}_{i=1}^{n_0}$ from treatment group, $\{Y_i\}_{i=1}^{n_1}$ from control group

- $\{X_i\}_{i=1}^{n_0}$ PU students' wage, $\{Y_i\}_{i=1}^{n_1}$ non-PU students' wage

- $\widehat{\text{se}}$ of difference in means:

$$\widehat{\text{se}} \text{ of } \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i - \frac{1}{n_1}\sum_{i=1}^{n_1} X_i\right) = \sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}},$$

where $\hat{\sigma}_{X_i}$ and $\hat{\sigma}_{Y_i}$ are:

$$\hat{\sigma}^2_{X_i} = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i^2 - \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i\right)^2, \quad \hat{\sigma}^2_{Y_i} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^2 - \left(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\right)^2.$$

Outline

1. Hypothesis Testing and Confidence Interval: Motivation

2. Standard Error

3. Hypothesis Testing

4. Confidence Interval

5. Duality: Confidence Interval and Hypothesis Testing

Hypothesis Testing: Motivation

Example 1 for hypothesis testing: setup

- Question: is the coin toss a fair one?
- Consider the simulation we did in previous classes
- Unit of analysis: one coin toss
- Data: results for each coin toss across different tosses

Example 1 for hypothesis testing: how to make decisions?

- H_0: Fair coin toss $\rightsquigarrow P(\text{head}) = \frac{1}{2}$
- How to measure the evidence for testing H_0?
- How to quantify the strength of the evidence against H_0?
- What is the threshold for the strength of evidence to conclude?
- What are the errors we make in this decision-making process?
- How to control the error probability?

Hypothesis Testing: Motivation (Conti.)

Example 2 for hypothesis testing: setup

- Question: is there racial discrimination in the U.S. labor market?
- Consider the RCT by Bertrand and Mullainathan (2004, AER)
  - Unit of analysis: resumes
  - Treatment: African-American- or White-sounding names
  - Outcome: callbacks from employers

Example 2 for hypothesis testing: how to make decisions?

- H_0: No discrimination $\rightsquigarrow$ no difference on callbacks
- How to measure the evidence for testing H_0?
- How to quantify the strength of the evidence against H_0?
- What is the threshold for the strength of evidence to conclude?
- What are the errors we make in this decision-making process?
- How to control the error probability?

Hypothesis Testing: H_0 and H_1

Parameter space

- Logically speaking, need to specify what we want to test first
- Recall: interested in learning some "feature" of the data $\theta(P)$
- Θ: parameter space in which the *param.* of interest resides
  - All possible values for $\theta(P)$
- Partition parameter space into two disjoint sets: Θ_0 and Θ_1
  - Disjoint: $\Theta_0 \cap \Theta_1 = \emptyset$

    Partition all possible values for $\theta(P)$ to two disjoint parts

H_0 and H_1

- Null hypothesis: $H_0 : \theta \in \Theta_0$
  - Will decide whether the evidence is strong enough against H_0
- Alternative hypothesis: $H_1 : \theta \in \Theta_1$

Hypothesis Testing: H_0 and H_1 (Conti.)

H_0 and H_1: example 1

- $\Theta = [0, 1]$
  - $\theta = P(\text{Head}) \in [0, 1]$: *prob.* of getting head is between 0 and 1
- $\Theta_0 = \{0.5\}$: fair toss space

  $\Theta_1 = [0, 0.5) \cup (0.5, 1]$: unfair (unfair in both directions) toss space
- $H_0 : \theta = 0.5$ (fair) versus $H_1 : \theta \neq 0.5$ (not fair)

H_0 and H_1: example 2

- $\Theta = [-1, 1]$
  - $\theta = E(\text{Callback} \mid \text{Black}) - E(\text{Callback} \mid \text{White}) \in [-1, 1]$
- $\Theta_0 = \{0\}$: no racial discrimination in terms of callback

  $\Theta_1 = [-1, 0) \cup (0, 1]$: discrimination (in both directions) space

  $[-1, 0)$: discriminate Black; $(0, 1]$: discriminate White
- $H_0 : \theta = 0$ (no discrimination) versus $H_1 : \theta \neq 0$ (discrimination)

Hypothesis Testing: Type I and Type II Error

Heuristically, we make the following errors

- Recall that we are testing the following hypothesis:

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta_1$$

- Logically, can make either H_0 is true or H_1 is true

- If H_0 is true, but data tells us to reject H_0
- If H_1 is true, but data tells us not to reject (loosely, "accept") H_0

Type I and type II error: definition

	Not reject H_0	Reject H_0
H_0 true	✓	type I error (false *pos.*)
H_1 true	type II error (false *neg.*)	✓

- *neg.*: negative, *pos.*: positive

Hypothesis Testing: Type I and Type II Error (Conti.)

Type I and type II error: example 1

	Not reject H_0	Reject H_0
It's a fair coin toss (H_0✓)	✓	type I error
It's not a fair coin toss (H_1✓)	type II error	✓

- $\theta = P(\text{Head})$
- $H_0 : \theta = 0.5$ (fair coin toss) v.s. $H_1 : \theta \neq 0.5$ (unfair coin toss)

Type I and type II error: example 2

	Not reject H_0	Reject H_0
No racial discrimination (H_0✓)	✓	type I error
Racial discrimination exists (H_1✓)	type II error	✓

- $\theta = E(\text{Callback} \mid \text{Black}) - E(\text{Callback} \mid \text{White})$
- $H_0 : \theta = 0$ (no discrimination) v.s. $H_1 : \theta \neq 0$ (discrimination)

Hypothesis Testing: Test Statistic

Test statistic: definition

A test statistic T is a function that maps from data to a number.

Test statistic: heuristics

- Intuitively, the test statistic should use the following information:

- $\theta(P)$: we want to learn about $\theta(P)$, this should appear somewhere
- $\hat{\theta}_n$: this our guess of $\theta(P)$ from data, should use this too
- $\widehat{\text{se}} = \sqrt{\widehat{\text{Var}}\left(\hat{\theta}_n\right)}$: variability of the guess matters, should use this

Test statistic: it needs help from CLT

Hopefully, under suitable assumptions, we can get:

$$\frac{\left(\hat{\theta}_n - \theta(P)\right)}{\widehat{\text{se}}} \xrightarrow{\mathcal{D}} N(0,1),$$

Hypothesis Testing: Test Statistic (Conti.)

Test statistic: example 1

Let $\{X_i\}_{i=1}^n$ be a sequence of *iid* random variables on $\mathbb{R}$ with distribution P. Suppose that $\sigma^2(P)$ exists, then:

$$\frac{\left(\overline{X}_n - E(X_i)\right)}{\frac{\hat{\sigma}(P)}{\sqrt{n}}} \xrightarrow{\mathcal{D}} N\left(0, 1\right),$$

where $\hat{\sigma}^2(P) = \hat{E}(X_i^2) - \hat{E}(X_i)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 - \left(\frac{1}{n}\sum_{i=1}^n X_i\right)^2$.

Wait, the denominator looks familiar...

The denominator in this version's CLT is an estimated standard error:

$$\widehat{\text{se}} = \frac{\hat{\sigma}(P)}{\sqrt{n}}.$$

Hypothesis Testing: Test Statistic (Conti.)

Test statistic: example 2

Let n_0 *iid* samples $\{X_i\}_{i=1}^{n_0}$ and n_1 *iid* samples $\{Y_i\}_{i=1}^{n_1}$ with distribution $P_{X_i} \times P_{Y_i}$. Suppose that $\sigma^2_{X_i}$ and $\sigma^2_{Y_i}$ exists and $\frac{n_0}{n_1} \to c < +\infty$ as $n_0, n_1 \to +\infty$, then:

$$\frac{\left(\left(\overline{X}_{n_0} - \overline{Y}_{n_1}\right) - (E(X_i) - E(Y_i))\right)}{\sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}}} \xrightarrow{\mathcal{D}} N(0,1),$$

where $\hat{\sigma}_{X_i}$ and $\hat{\sigma}_{Y_i}$ are:

$$\hat{\sigma}^2_{X_i} = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i^2 - \left(\frac{1}{n_0}\sum_{i=1}^{n_0} X_i\right)^2, \quad \hat{\sigma}^2_{Y_i} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i^2 - \left(\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i\right)^2.$$

Hypothesis Testing: Test Statistic (Conti.)

Test statistic: heuristics of its appearance

- By CLT, can choose test statistic in the following way:

$$T = \left| \frac{\left(\hat{\theta}_n - \theta(P)\right)}{\widehat{\mathrm{se}}} \right| = \frac{\text{deviation from } \hat{\theta}_n \text{ to } \theta(P)}{\text{rescale by variability of } \hat{\theta}_n}$$

- Wait a sec, we do not know $\theta(P)$...
- But, we know we want to test $\theta(P)$:

$$\text{for example:} \quad H_0 : \theta(P) = 0 \quad \text{v.s.} \quad H_1 : \theta(P) \neq 0$$

Test statistic: what it tells us

- Under H_0, T quantifies the deviation of $\hat{\theta}_n$ from $\theta(P) \in H_0$:
- If H_0 is true, such deviation should be small
- Therefore, large $T \rightsquigarrow$ evidence against H_0

Hypothesis Testing: Test Statistic (Conti.)

Test statistic: example 1 (conti.)

A test statistic for testing a sample mean is:

$$T = \left| \frac{\left(\bar{X}_n - E(X_i)\right)}{\frac{\hat{\sigma}(P)}{\sqrt{n}}} \right|, \quad \text{where } E(X_i) \in H_0 = \{0.5\}.$$

Test statistic: example 2 (conti.)

A test statistic for testing difference in mean is:

$$T = \left| \frac{\left(\left(\bar{X}_{n_0} - \bar{Y}_{n_1}\right) - (E(X_i) - E(Y_i))\right)}{\sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}}} \right|, \quad \text{where } E(X_i) - E(Y_i) \in H_0 = \{0\}.$$

Hypothesis Testing: Rejection Region

Recall the 2nd canonical problem in mathematical statistics

- Test a null hypothesis (H_0) about $\theta(P)$: reject or not reject H_0

Now, make a binary decision based on test statistic

- Given we now have test statistic, can have a binary rule:

$$\begin{aligned} T > c \text{ (large deviation from } H_0) &\quad \rightsquigarrow \text{ reject } H_0 \\ T \leq c \text{ (not a large deviation from } H_0) &\quad \rightsquigarrow \text{ fail to reject } H_0 \end{aligned}$$

- $T > c$: is called a rejection region
- c: is called a critical value

Rejection region: remarks

- $T = \left| \frac{(\hat{\theta}_n - \theta(P)))}{\widehat{\text{se}}} \right| > c \Longleftrightarrow \frac{(\hat{\theta}_n - \theta(P)))}{\widehat{\text{se}}} > c$ or $\frac{(\hat{\theta}_n - \theta(P)))}{\widehat{\text{se}}} < -c$

Hypothesis Testing: Rejection Region (Conti.)

Test statistic: example 1 (conti.)

- Rejection region for testing a sample mean is:

$$\frac{\left(\overline{X}_n - E(X_i)\right)}{\frac{\hat{\sigma}(P)}{\sqrt{n}}} \underbrace{\geq c}_{\text{evidence deviates to: toss favors head}}$$

$$\frac{\left(\overline{X}_n - E(X_i)\right)}{\frac{\hat{\sigma}(P)}{\sqrt{n}}} \underbrace{\leq -c}_{\text{evidence deviates to: toss favors tail}}$$

where under the null, $E(X_i) = 0.5$

- Thus, reject the null when evidence deviates to: toss is unfair

- Unfair toss can be either favor head or favor tail

Hypothesis Testing: Rejection Region (Conti.)

Test statistic: example 2 (conti.)

- Rejection region for testing difference in means is:

$$\frac{\left(\left(\bar{X}_{n_0} - \bar{Y}_{n_1}\right) - (E(X_i) - E(Y_i))\right)}{\sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}}} \underbrace{\geq c}_{\text{evidence deviates to: discriminates White}}$$

$$\frac{\left(\left(\bar{X}_{n_0} - \bar{Y}_{n_1}\right) - (E(X_i) - E(Y_i))\right)}{\sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}}} \underbrace{\leq -c}_{\text{evidence deviates to: discriminates Black}}$$

where under the null, $E(X_i) - E(Y_i) = 0$

- Thus, reject the null when evidence deviates to: discrimination

Hypothesis Testing: Size and Level

Size of a test: motivation

- Might make a type I error when using test *stats.* and *rej.* region
  - If toss is fair, coincidentally, I get 65% of tails in multiple draws
  - If there is no discrimination, coincidentally, the data *rej.* this
- Given we're making errors, need to control the *prob.* of error
- Naturally, need to define a jargon for the *prob.* of such an error

Size of a test: definition

When testing the following hypothesis:

$$H_0 : \theta(P) = \theta_0 \quad \text{v.s.} \quad H_1 : \theta(P) \neq \theta_0,$$

the size of a test is:

$$\alpha = P_{\theta(P) \in H_0}(T > c).$$

Hypothesis Testing: Size and Level (Conti.)

Size of a test: remarks

- Now, let us unpack what α means

$$\underbrace{\alpha}_{\text{probability of Type I error}} = P_{\underbrace{\theta(P) \in H_0}_{\text{given that the truth is null}}} \Big(\underbrace{T > c}_{\text{evidence tells me to reject null}} \Big)$$

- Sum up, size of a test in this test: probability of Type I error

Size of a test: example 1 & 2

1. Example 1: test whether a coin toss is fair
   - Given it's fair toss, probability of evidence showing us it's unfair
2. Example 2: test whether there is racial discrimination
   - Given no racial *disc.*, *prob.* of evidence showing there is *disc.*

? Can you give an example of the size of a medical test?

Hypothesis Testing: Size and Level (Conti.)

Level of a test: motivation

- Recall: want to control Type I error with hypothesis testing

Level of a test: definition

A test is to have level α if its size is less or equal to α.

Level of a test: example 1 & 2

1. Example 1: test whether a coin toss is fair
   - Given it's fair toss, the Type I error of the test $\leq \alpha$
2. Example 2: test whether there is racial discrimination
   - Given no racial *disc.*, the Type I error of the test $\leq \alpha$

Level of a test: remarks

- Choose the level of the test before you start your analysis!
- Critical value c is partly determined by the level of the test

Hypothesis Testing: P-Value

P-Value: motivation

- So far, data tells us evidence about the null hypothesis H_0
- Naturally, want to measure the strength of evidence against H_0

P-Value: definition

P-value is the smallest level at which we can reject null hypothesis

P-Value: strength of evidence against H_0

p-value	evidence	*stats.* significance
< 0.01	very strong evidence against H_0	*sig*. at 1% level
$0.01 - 0.05$	strong evidence against H_0	*sig*. at 5% level
$0.05 - 0.1$	weak evidence against H_0	*sig*. at 10% level
> 0.1	little evidence against H_0	not *sig*.

Hypothesis Testing: P-Value (Conti.)

P-Value: two "theorems"

1. P-Value: how "unlikely" the observed test *stats.* is if H_0 is true

$$\text{p-value} = P_{\underbrace{\theta_0}_{\text{if } H_0 \text{ is true}}}\Big(\underbrace{T}_{\text{test statistics}} > \underbrace{T_n}_{\text{observed test statistics in data}} \Big)$$

2. Recall the test *stats.* we've used: $T = \frac{(\hat{\theta}_n - \theta(P))}{\widehat{\text{se}}}$, then, by CLT:

$$P_{\theta_0}(|T| > c) = P_{\theta_0}\left(\left| \frac{\left(\hat{\theta}_n - \theta(P)\right)}{\widehat{\text{se}}} \right| > c \right) \approx 2\Phi(-c),$$

where Φ is the CDF of the standard normal distribution ($N(0,1)$).

- Thus, to have strong evidence against H_0, need:

$$P_{\theta_0}(|T| > c) \leq 0.05 \quad \text{which leads to } c = 1.96$$

Hypothesis Testing: P-Value (Conti.)

P-Value: remarks

- Large p-value: not strong evidence in favor of H_0

  - Maybe: H_0 true

  - Maybe: H_0 false, but test doesn't do a good job of detecting signal

  With intimidating jargon: test might be underpowered

- P-Value IS NOT: $P(H_0 \mid \text{Data})$!

  - Loosely speaking, P-Value = $P(\text{Data} \mid H_0)$

  In general, $P(H_0 \mid \text{Data}) \neq P(\text{Data} \mid H_0)$

- Subtlety of using p-value: reject or not when p-value= 0.1001?

P-Value: a repetition of an important point

- P-Value measures the strength of evidence against H_0

  - $H_0 \checkmark \rightsquigarrow T$ not extreme/large

  - T extreme/large $\rightsquigarrow H_0$ false

Hypothesis Testing: P-Value (Conti.)

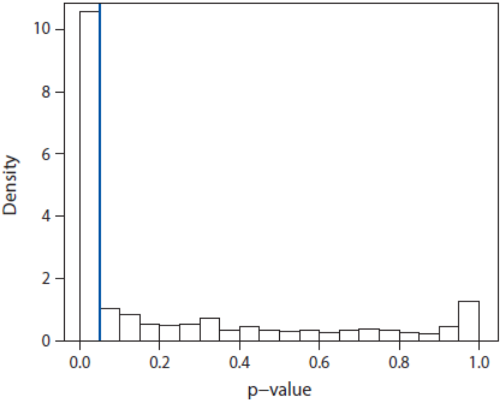


Figure 7.4. The Distribution of p-Values for Hypothesis Tests Published in Two Leading Political Science Journals.

Hypothesis Testing: A Cookbook

A Cookbook for Testing Statistical Hypothesis

1. Select a significance level α

   Select a test statistic T with some distribution under H_0
2. Calculate the critical value c such that Type I error $\leq \alpha$

   Calculate the p-value
3. Binary statistical decision rule

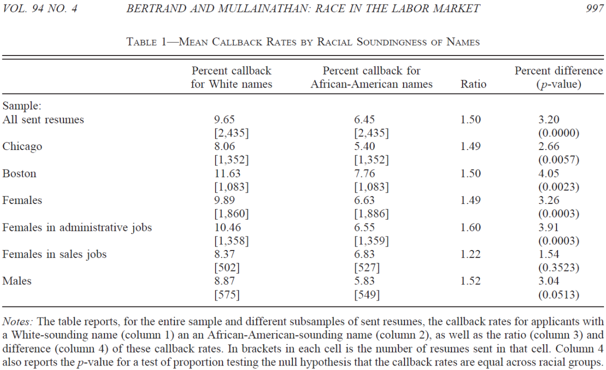
   Reject H_0 if $T_n > c$, equivalently $p < \alpha$

   Fail to reject H_0 if $T_n \leq c$, equivalently $p \geq \alpha$

Warning!

- Choose significance level and test statistic before seeing data
- The procedure is valid for testing one H_0
- Don't be dogmatic about p-value: many ways to torture data

Hypothesis Testing: An Empirical Example



VOL. 94 NO. 4 BERTRAND AND MULLAINATHAN: RACE IN THE LABOR MARKET 997

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65	6.45	1.50	3.20
	[2,435]	[2,435]		(0.0000)
Chicago	8.06	5.40	1.49	2.66
	[1,352]	[1,352]		(0.0057)
Boston	11.63	7.76	1.50	4.05
	[1,083]	[1,083]		(0.0023)
Females	9.89	6.63	1.49	3.26
	[1,860]	[1,886]		(0.0003)
Females in administrative jobs	10.46	6.55	1.60	3.91
	[1,358]	[1,359]		(0.0003)
Females in sales jobs	8.37	6.83	1.22	1.54
	[502]	[527]		(0.3523)
Males	8.87	5.83	1.52	3.04
	[575]	[549]		(0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the p-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

P-Value for the racial differences in all resumes sample

- Test statistic is: 4.116
- P-Value is: 0.00004 $\rightsquigarrow$ very strong evidence against H_0: no *disc.*
- Codes on Canvas "lecture" folder

Outline

1. Hypothesis Testing and Confidence Interval: Motivation

2. Standard Error

3. Hypothesis Testing

4. **Confidence Interval**

5. Duality: Confidence Interval and Hypothesis Testing

Confidence Interval: Motivation

Recall: last canonical problems in mathematical statistics

- Sample: $\{X_i\}_{i=1}^n$ distributed according to P (population *dist.*)
- Assume $\{X_i\}_{i=1}^n$ is an (aka, *iid*) sample
- "Learn" some "features" of P (e.g., a *param.* $\theta(P)$) from the data
- $\theta(P)$ can be either a descriptive or a causal *param.*
- Construct a confidence region for $\theta(P)$
- Construct a random set C_n based on the data
- This random set covers the true *param.* with pre-specified *prob.*

$$P(\underbrace{\theta(P)}_{\text{true parameter}} \in \underbrace{C_n}_{\text{random set: sample maps to this set}}) \geq 1 - \underbrace{\alpha}_{\text{pre-specified}},$$

where $\alpha \in (0,1)$

- Question: relationship between hypothesis testing and C.I.?

Confidence Interval: Motivation (Conti.)

Confidence interval: example 1

- Recall Example 1: test whether a coin toss is fair
- We have n sample points (i.e., n tosses)
- Based on n tosses, want to construct a C.I. such that

$$P(\underbrace{\theta(P)}_{\text{true } P(\text{head})} \in \underbrace{C_n}_{\text{random set: sample maps to this set}}) \geq 1 - \underbrace{\alpha}_{\text{pre-specified}}$$

Confidence interval: example 2

- Recall Example 1: test whether there is racial discrimination
- Based on $n_0 + n_1$ applicants, want to construct a C.I. such that

$$P(\underbrace{\theta(P)}_{\text{true } \textit{discr. param.}} \in \underbrace{C_n}_{\text{random set: sample maps to this set}}) \geq 1 - \underbrace{\alpha}_{\text{pre-specified}}$$

Confidence Interval: Definition

Confidence interval: definition

A $1-\alpha$ confidence interval for a parameter $\theta(P)$ is an interval $C_n = (a_n, b_n)$, where a_n and b_n are functions of the sample such that

$$P_{\theta(P)}(\theta(P) \in C_n) \geq 1-\alpha, \quad \text{for all } \theta \in \Theta.$$

Confidence interval: it needs help from CLT

- Recall that under suitable assumptions, we can get:

$$\frac{\left(\hat{\theta}_n - \theta(P)\right)}{\widehat{\text{se}}} \xrightarrow{\mathcal{D}} N(0,1),$$

- Can use this to construct C.I. since we know the *dist.* of $\frac{\left(\hat{\theta}_n - \theta(P)\right)}{\widehat{\text{se}}}$

- Note that: $\theta(P)$ is unknown, $\hat{\theta}_n$ and $\widehat{\text{se}}$ known from the sample

Confidence Interval: Definition (Conti.)

Use CLT to construct C.I.

- By CLT, we know:

$$P_{\theta(P)}\left(\left|\frac{\left(\hat{\theta}_n - \theta(P)\right)}{\widehat{\text{se}}}\right| \leq c\right) \approx 1 - 2\Phi(-c)$$

$$\Leftrightarrow P_{\theta(P)}\left(\hat{\theta}_n - c \times \widehat{\text{se}} \leq \theta(P) \leq \hat{\theta}_n + c \times \widehat{\text{se}}\right) \approx 1 - 2\Phi(-c)$$

Three choices of c for 90%, 95%, and 99% C.I.

$$P_{\theta(P)}\left(\hat{\theta}_n - 1.64 \times \widehat{\text{se}} \leq \theta(P) \leq \hat{\theta}_n + 1.64 \times \widehat{\text{se}}\right) \approx 1 - 2\Phi(-1.64) = 90\%$$

$$P_{\theta(P)}\left(\hat{\theta}_n - 1.96 \times \widehat{\text{se}} \leq \theta(P) \leq \hat{\theta}_n + 1.96 \times \widehat{\text{se}}\right) \approx 1 - 2\Phi(-1.96) = 95\%$$

$$P_{\theta(P)}\left(\hat{\theta}_n - 2.58 \times \widehat{\text{se}} \leq \theta(P) \leq \hat{\theta}_n + 2.58 \times \widehat{\text{se}}\right) \approx 1 - 2\Phi(-2.58) = 99\%$$

Confidence Interval: Examples

95% **confidence interval: example 1 (conti.)**

A confidence interval for an expectation is:

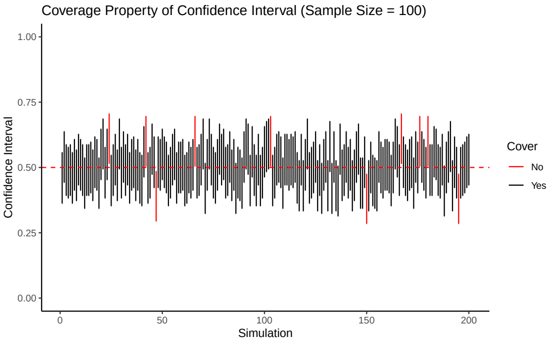
$$\overline{X}_n - 1.96 \times \frac{\hat{\sigma}(P)}{\sqrt{n}} \leq E(X_i) \leq \overline{X}_n + 1.96 \times \frac{\hat{\sigma}(P)}{\sqrt{n}}$$

95% **confidence interval: example 2 (conti.)**

A confidence interval for difference in mean is:

$$\begin{aligned} \left(\overline{X}_{n_0} - \overline{Y}_{n_1}\right) - 1.96 \times \sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}} &\leq E(X_i) - E(Y_i) \\ &\leq \left(\overline{X}_{n_0} - \overline{Y}_{n_1}\right) - 1.96 \times \sqrt{\frac{\hat{\sigma}^2_{X_i}}{n_0} + \frac{\hat{\sigma}^2_{Y_i}}{n_1}} \end{aligned}$$

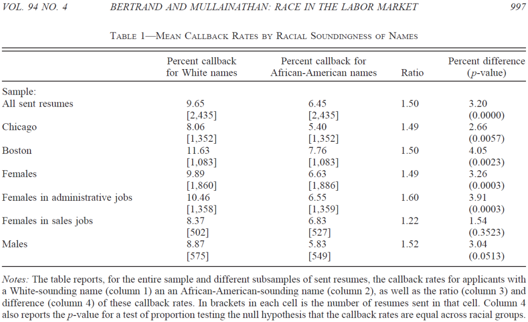
Confidence Interval: Examples (Conti.)



Simulation Setup

- $\{X_i\}_{i=1}^{100}$ is *iid* draw from Ber(0.5)
- $\theta(P) = E(X_i) = 0.5$, $\hat{\theta}_n = \frac{\sum_{i=1}^{n} X_i}{n}$
- 95% of C.I.: $\overline{X}_n - 1.96 \times \frac{\hat{\sigma}(P)}{\sqrt{n}} \leq E(X_i) \leq \overline{X}_n + 1.96 \times \frac{\hat{\sigma}(P)}{\sqrt{n}}$

Confidence Interval: Examples (Conti.)



VOL. 94 NO. 4 *BERTRAND AND MULLAINATHAN: RACE IN THE LABOR MARKET* *997*

TABLE 1—MEAN CALLBACK RATES BY RACIAL SOUNDINGNESS OF NAMES

	Percent callback for White names	Percent callback for African-American names	Ratio	Percent difference (p-value)
Sample:				
All sent resumes	9.65 [2,435]	6.45 [2,435]	1.50	3.20 (0.0000)
Chicago	8.06 [1,352]	5.40 [1,352]	1.49	2.66 (0.0057)
Boston	11.63 [1,083]	7.76 [1,083]	1.50	4.05 (0.0023)
Females	9.89 [1,860]	6.63 [1,886]	1.49	3.26 (0.0003)
Females in administrative jobs	10.46 [1,358]	6.55 [1,359]	1.60	3.91 (0.0003)
Females in sales jobs	8.37 [502]	6.83 [527]	1.22	1.54 (0.3523)
Males	8.87 [575]	5.83 [549]	1.52	3.04 (0.0513)

Notes: The table reports, for the entire sample and different subsamples of sent resumes, the callback rates for applicants with a White-sounding name (column 1) an an African-American-sounding name (column 2), as well as the ratio (column 3) and difference (column 4) of these callback rates. In brackets in each cell is the number of resumes sent in that cell. Column 4 also reports the p-value for a test of proportion testing the null hypothesis that the callback rates are equal across racial groups.

C.I. for the racial differences in all resumes sample

- The 95% C.I. is:

$$[3.203 - 1.96 \times 0.778, 3.203 + 1.96 \times 0.778] = [1.678, 4.729]$$

- Codes on Canvas ”lecture” folder

Outline

1. Hypothesis Testing and Confidence Interval: Motivation
2. Standard Error
3. Hypothesis Testing
4. Confidence Interval
5. Duality: Confidence Interval and Hypothesis Testing

Confidence Interval v.s. Hypothesis Testing

Confidence interval v.s. hypothesis testing: connections?

- What's the connection between C.I. and hypothesis testing?
- Conjecture: there should be some connections between the two
  - Hypothesis testing: reject the guesses that are too ”extreme”
  - Confidence interval: covers the ”reasonable” guesses

Duality: confidence interval and hypothesis testing

A $100(1-\alpha)\%$ confidence interval for $\theta(P)$ consists of all those values of θ_0 for which the null hypothesis that $\theta(P) = \theta_0$ will not be rejected at level α.

Applying duality: two equivalent statements

- A 95% C.I. for $\theta(P)$ does not contain 0
- $H_0 : \theta(P) = 0$ is rejected at 5% level

Learning Goals: Testing Statistical Hypothesis

Students will be able to:

- Understand the definition of standard error
- - Know that s.e. measures the variability of an estimator
- Understand the definition of consistency
- Understand the definition of null and alternative hypothesis
- Understand the definition of Type 1 and Type 2 Error
- Understand the definition of a test statistic
- - If the null hypothesis is true, we can get the *dist.* of test *stats.*
- - Know the two examples for a test *stats.*
- Understand the size and the level of a test
- Understand the definition of p-value
- Conduct hypothesis testing using the cookbook
- Understand the definition of confidence interval
- - Know the duality between confidence interval and testing