Lecture 8: Hypothesis Testing 1

Jacob M. Montgomery

Quantitative Political Methodology

Lecture 8: Hypothesis Testing 1

Roadmap

Last class:

- We figured out how to calculate a sampling distribution.
- We figured out how create a confidence interval to estimate a parameter.

Roadmap

Last class:

- We figured out how to calculate a sampling distribution.
- We figured out how create a confidence interval to estimate a parameter.

This class:

- What is a hypothesis test?
- The five steps of hypothesis testing.

Roadmap

Last class:

- We figured out how to calculate a sampling distribution.
- We figured out how create a confidence interval to estimate a parameter.

This class:

- What is a hypothesis test?
- The five steps of hypothesis testing.

Next class:

- Hypothesis tests with small samples
- Types of errors

Hypothesis testing: The big picture

The goal of hypothesis testing is to see if the data agree with some prediction we make based on our theory.

> ***Definition 1***: *A hypothesis is just some statement about a characteristic of a variable or a set of variables.*

Hypothesis testing: The big picture

The goal of hypothesis testing is to see if the data agree with some prediction we make based on our theory.

> ***Definition 1***: *A hypothesis is just some statement about a characteristic of a variable or a set of variables.*
>
> ***Definition 2***: *In statistics, a hypothesis is a statement about a population. It is usually a prediction that a parameter describing some characteristic of a variable takes a particular numerical value or falls in a certain range of values.*

Hypothesis testing: The big picture

The goal of hypothesis testing is to see if the data agree with some prediction we make based on our theory.

> ***Definition 1****: A hypothesis is just some statement about a characteristic of a variable or a set of variables.*
>
> ***Definition 2****: In statistics, a hypothesis is a statement about a population. It is usually a prediction that a parameter describing some characteristic of a variable takes a particular numerical value or falls in a certain range of values.*

To test a hypothesis, we take our data and conduct a *significance test*. Does the data support my hypothesis?

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter.

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter. **Everything** we do for the rest of this class will concern hypothesis testing of one kind or another.

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter. **Everything** we do for the rest of this class will concern hypothesis testing of one kind or another.

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter. **Everything** we do for the rest of this class will concern hypothesis testing of one kind or another.

- This is why it is called political **science**.

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter. **Everything** we do for the rest of this class will concern hypothesis testing of one kind or another.

- This is why it is called political **science**.
- Every empirical research question is a hypothesis test.

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter. **Everything** we do for the rest of this class will concern hypothesis testing of one kind or another.

- This is why it is called political **science**.
- Every empirical research question is a hypothesis test.
- This is *exciting*.

The five steps of hypothesis testing

CIs are about estimating a parameter value. Hypothesis testing is about testing scientific claims about a parameter. **Everything** we do for the rest of this class will concern hypothesis testing of one kind or another.

- This is why it is called political **science**.
- Every empirical research question is a hypothesis test.
- This is *exciting*.
- Examples?

Step 1 of 5: Make some assumptions about your data

To gain leverage on the problem, we have to make some assumptions about the data and where it came from. These assumptions determine how (and if) we can test our hypothesis.

Step 1 of 5: Make some assumptions about your data

To gain leverage on the problem, we have to make some assumptions about the data and where it came from. These assumptions determine how (and if) we can test our hypothesis.

Important assumptions include:

- Type of data

Step 1 of 5: Make some assumptions about your data

To gain leverage on the problem, we have to make some assumptions about the data and where it came from. These assumptions determine how (and if) we can test our hypothesis.

Important assumptions include:

- Type of data
- Population distributions

Step 1 of 5: Make some assumptions about your data

To gain leverage on the problem, we have to make some assumptions about the data and where it came from. These assumptions determine how (and if) we can test our hypothesis.

Important assumptions include:

- Type of data
- Population distributions
- Sample size

Step 1 of 5: Make some assumptions about your data

To gain leverage on the problem, we have to make some assumptions about the data and where it came from. These assumptions determine how (and if) we can test our hypothesis.

Important assumptions include:

- Type of data
- Population distributions
- Sample size
- Sample method (i.e., randomization)

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a the null hypothesis (determined sceptic)

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a the null hypothesis (determined sceptic)
- Second, we use a sample statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a the null hypothesis (determined sceptic)
- Second, we use a sample statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.
- Third, we (kind of) conclude that our research hypothesis is true. This is usually called the alternative hypothesis.

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a the null hypothesis (determined sceptic)
- Second, we use a sample statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.
- Third, we (kind of) conclude that our research hypothesis is true. This is usually called the alternative hypothesis.
- To do this, we first need to specify our hypotheses.

Step 3 of 5: Calculate a Test Statistic

We calculate a statistic summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*.

Step 3 of 5: Calculate a Test Statistic

We calculate a statistic summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*. Usually this is something equivalent to a Z-statistic.

Step 4 of 5: Calculate a P-Value

*A **P-Value** is a measure of surprise.*

Step 4 of 5: Calculate a P-Value

A ***P-Value*** *is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme* ***or more****?"*

Step 4 of 5: Calculate a P-Value

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- P-Values are very difficult for many people to understand.
- Smaller P-Values more strongly contradict the null

Step 5 of 5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$

Step 5 of 5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$ (Yay! More greek letters!)

Step 5 of 5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$ (Yay! More greek letters!)
- We "reject the null" and conclude that the evidence supports the alternative hypothesis.

Large sample significance testing for means

Step 1: Assumptions

- $n \geq 30$
- Quantitative data (i.e., not a proportion)

Large sample significance testing for means

Step 1: Assumptions

- $n \geq 30$
- Quantitative data (i.e., not a proportion)

Step 2: State hypotheses

- $H_0 : \mu = \mu_0$ (e.g., $\mu_0 = 0$)
- This is a `two-sided test'', but it may be aone-sided.”

Step 3: Calculate a test statistic

- $TS = \dfrac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}}$
- TS is our "Test statistic"
- Just as before, this comes from the sampling distribution of $\bar{Y}$

Step 4: P-Value

- $p = Pr(Z \geq |\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}|) + Pr(Z \leq -|\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}|)$
- $= 2 \times Pr(Z \geq |\frac{\bar{Y}-\mu_0}{\sigma_{\bar{Y}}}|)$
- We use both tails, because we want to find the probability of error in both directions.

Step 5: Draw a conclusion

- If $p \leq \alpha$ we conclude that the evidence supports H_a
- If $p > \alpha$ we say that "we cannot reject the null hypothesis."

Example: Females “like” the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you “like” the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is “neutral.”

Example: Females “like” the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you “like” the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is “neutral.”

- $\bar{y} = 60.5$,

Example: Females "like" the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you "like" the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is "neutral."

- $\bar{y} = 60.5, n = 626,$

Example: Females “like” the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you “like” the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is “neutral.”

- $\bar{y} = 60.5, n = 626, S = 24.9$

Example: Females "like" the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you "like" the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is "neutral."

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$

Example: Females "like" the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you "like" the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is "neutral."

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$
- Assumptions? Continuous data, large sample, random sample.
- $H_0 : \mu = 58, H_a : \mu \neq 58$

Example: Females "like" the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you "like" the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is "neutral."

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$
- Assumptions? Continuous data, large sample, random sample.
- $H_0 : \mu = 58, H_a : \mu \neq 58$
- $Z - score = \frac{60.5 - 58}{1} = 2.5$

Example: Females "like" the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you "like" the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is "neutral."

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$
- Assumptions? Continuous data, large sample, random sample.
- $H_0 : \mu = 58, H_a : \mu \neq 58$
- $Z - score = \frac{60.5-58}{1} = 2.5$
- $P - value = 2 \times 0.006 = 0.012$

Example: Females “like” the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you “like” the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is “neutral.”

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$
- Assumptions? Continuous data, large sample, random sample.
- $H_0 : \mu = 58, H_a : \mu \neq 58$
- $Z - score = \frac{60.5-58}{1} = 2.5$
- $P - value = 2 \times 0.006 = 0.012$
- Is that good enough?

Example: Females "like" the Democratic party

2004 ANES asked all respondents to rate the Democratic party on a feeling thermometer (0-100). We are just going to use data from female respondents. We are going to assume, for the purposes of this class, that any thermometer rating above 58 means you "like" the Democratic party. Test the research hypothesis that women are not neutral towards the Democratic party (they either like or dislike them). A score of exactly 58 means the respondent is "neutral."

- $\bar{y} = 60.5, n = 626, S = 24.9 \Rightarrow \sigma_{\bar{y}} \approx 1.0$
- Assumptions? Continuous data, large sample, random sample.
- $H_0 : \mu = 58, H_a : \mu \neq 58$
- $Z - score = \frac{60.5-58}{1} = 2.5$
- $P - value = 2 \times 0.006 = 0.012$
- Is that good enough?
- Why are we using a two-sided test?

Large sample test of proportions

Castenedat v. Partida

- The true number of Mexican-Americans was 79.1% of the population.
- Individuals were selected for jury participation using the "key man" system.
  - Key men in the area provide a list of possible jurors
  - Jurors are selected at random from the list
- 45.5% of the members of a grand jury (assume n=60) were Mexican-American.

Large sample test of proportions

Castenedat v. Partida

- The true number of Mexican-Americans was 79.1% of the population.
- Individuals were selected for jury participation using the "key man" system.
  - Key men in the area provide a list of possible jurors
  - Jurors are selected at random from the list
- 45.5% of the members of a grand jury (assume n=60) were Mexican-American.

Research hypothesis: The "'key man" system produces lists that significantly under-represent Mexican-Americans.

Step 1 of 5: Make some assumptions about your data

To get some leverage on the problem, we have to make some assumptions about the data and where it came from. The assumptions will determine how (and if) we can test our hypothesis.

Step 1 of 5: Make some assumptions about your data

To get some leverage on the problem, we have to make some assumptions about the data and where it came from. The assumptions will determine how (and if) we can test our hypothesis.

Important assumptions include:

- Type of data
- Population distributions
- Sample size
- Sample method (i.e., randomization)

Step 1 of 5: Make some assumptions about your data

To get some leverage on the problem, we have to make some assumptions about the data and where it came from. The assumptions will determine how (and if) we can test our hypothesis.

Important assumptions include:

- Type of data
- Population distributions
- Sample size
- Sample method (i.e., randomization)

Jury Example:

- *Type of data*: Nominal data
- *Population distributions*: No assumptions needed
- *Sample size*: Large enough for the Central Limit Theorem
- *Sampling method*: Jury members are selected at random

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a the null hypothesis (determined sceptic)
- Second, we use a statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.
- Third, we (kind of) conclude that our research hypothesis is true. This is usually called the alternative hypothesis.
- To do this, we first need to specify our hypotheses.

Step 2 of 5: Formulate null and alternative hypotheses

We are going to try and support our research hypothesis using a technique called *proof by contradiction*.

- First, we set up a the null hypothesis (determined sceptic)
- Second, we use a statistic to show that the data we have observed would be very unlikely if the null hypothesis were true.
- Third, we (kind of) conclude that our research hypothesis is true. This is usually called the alternative hypothesis.
- To do this, we first need to specify our hypotheses.

Jury example:

- *Null hypothesis*: The proportion of Mexican-Americans provided by key men is the same as the proportion of Mexican-Americans in the district.
- $H_0 : \pi > 0.791$
- *Alternative hypothesis:* The proportion is less than that. $H_a : \pi \leq 0.791$

Step 3 of 5: Calculate a Test Statistic

We calculate a statistic summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*.

Step 3 of 5: Calculate a Test Statistic

We calculate a statistic summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*. Usually this is something equivalent to a Z-statistic.

Step 3 of 5: Calculate a Test Statistic

We calculate a statistic summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*. Usually this is something equivalent to a Z-statistic.

Jury example :

- $Z - score = \frac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}} = \frac{0.455 - 0.791}{0.052} = -6.46$

Step 3 of 5: Calculate a Test Statistic

We calculate a statistic summarizes how much our data differs from what we would have expected to observe *if the null hypothesis were true*. Usually this is something equivalent to a Z-statistic.

Jury example :

- $Z - score = \frac{\hat{\pi} - \pi_0}{\sigma_{\pi_0}} = \frac{0.455 - 0.791}{0.052} = -6.46$

In this case, Z is measuring the number of standard deviations the observed data is from the population mean and standard deviation assumed by the null hypothesis.

Step 4 of 5: Calculate a P-Value

A ***P-Value*** *is a measure of surprise.*

Step 4 of 5: Calculate a P-Value

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

Step 4 of 5: Calculate a P-Value

> *A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- P-Values are very difficult for many people to understand.
- Smaller P-Values more strongly contradict the null

Step 4 of 5: Calculate a P-Value

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- P-Values are very difficult for many people to understand.
- Smaller P-Values more strongly contradict the null

Jury example:

- $Pr(Z \leq -6.46) = Pr(Z \geq 6.46) = p \approx 0$

Step 4 of 5: Calculate a P-Value

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- P-Values are very difficult for many people to understand.
- Smaller P-Values more strongly contradict the null

Jury example:

- $Pr(Z \leq -6.46) = Pr(Z \geq 6.46) = p \approx 0$
- We would be very surprised to observe a jury with this few Mexican-Americans or less if if $\pi = 0.791$

Step 4 of 5: Calculate a P-Value

*A **P-Value** is a measure of surprise. We ask, "If the null hypothesis is true, how likely is it that we would observe a test-statistic this extreme **or more**?"*

- P-Values are very difficult for many people to understand.
- Smaller P-Values more strongly contradict the null

Jury example:

- $Pr(Z \leq -6.46) = Pr(Z \geq 6.46) = p \approx 0$
- We would be very surprised to observe a jury with this few Mexican-Americans or less if if $\pi = 0.791$
- Why don't we ask $Pr(Z = -6.46)$?

Step 5 of 5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$

Step 5 of 5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$ (Yay! More greek letters!)

Step 5 of 5: Draw a Conclusion

How surprised would you have to be in order to conclude that the *null hypothesis* is false?

- Usually, $p \leq 0.05 \Rightarrow$ statistically significant result
- We would observe a test-statistic this extreme or more $1/20$ times if the null hypothesis was true.
- More generically, we want $p \leq \alpha$ (Yay! More greek letters!)
- We “reject the null” and conclude that the evidence supports the alternative hypothesis.

Jury Example:

- $p \approx 0 \Rightarrow$ we reject the null hypothesis.

Jury Example:

- $p \approx 0 \Rightarrow$ we reject the null hypothesis.
- $H_0 : \pi > 0.791, \ H_a : \pi \leq 0.791$

Jury Example:

- $p \approx 0 \Rightarrow$ we reject the null hypothesis.
- $H_0 : \pi > 0.791, \ H_a : \pi \leq 0.791$
- The evidence supports the hypothesis that the “key man”" system results in a list that includes too few Mexican-Americans.

Jury Example:

- $p \approx 0 \Rightarrow$ we reject the null hypothesis.
- $H_0 : \pi > 0.791,\ \ H_a : \pi \leq 0.791$
- The evidence supports the hypothesis that the “key man”" system results in a list that includes too few Mexican-Americans.
- Science!