
DATA 8

Fall 2016

# Review I, December 5

## Inference

Slides created by Ani Adhikari and John DeNero

# Plan for This Week

- **This lecture**: I will review:
  - Inference methods
- **Tue 12-4**: I will have office hours in 413 Evans
- **Wed during lecture**: Theory of Prob/Stat
- GSIs will review **Wed and Thurs during lab times**:
  - First hour: review problems on particular topic
  - Second hour: office hour
  - Topics and review leaders TBA; watch Piazza
- Fri: Go see a dumb movie or relax in some other way

# Final Exam

- **Monday December 12, 8:00 - 11:00**
- **RSF Field House**
- Bring something to write with and something to erase with; but not your breakfast. Water is OK.
- We will provide a couple of reference sheets, with drafts posted on Piazza during RRR week
- 16 questions (six 5-pointers, five 6-pointers, five 8-pointers).
- Covers the whole course

# Big Picture of Course Contents

1. Python
2. Describing data
3. General concepts of inference
4. Theory of probability and statistics
5. Methods of inference

# 1. Python

- Textbook sections
  - General features and Table methods: 3.1 - 8.2, 15.3
  - `np.median`: 9.3
  - `proportions_from_distribution`: 10.1
  - `percentile`: 11.1
  - `np.mean`, `np.std`: 12.1, 12.2
  - `stats.norm.cdf`: 12.3
  - `minimize`: 13.3

# 2. Describing Data

- Tables: Chapter 5
- Classifying and cross-classifying: 7.2, 7.3
- Distributions and visualization: Chapter 6, 7.5
- Center and spread: 12.1-12.3
- Linear trend and non-linear patterns: 7.1, Chapter 13

# 3. General Concepts of Inference

- Study, experiment, treatment, control, confounding, randomization, causation, association: Chapter 2
- Distribution: 6.1, 6.2
- Sampling, probability sample: 8.5
- Probability distribution, empirical distribution, law of averages: 9.1
- Population, sample, parameter, statistic, estimate, bias, variability: 9.3
- Model: 10.2, 14.1, every null and alternative hypothesis

# 5. Methods of Inference

(We'll do Item 4 after Item 5).

- Making conclusions about unknown features of the population or model, based on assumptions of randomness

# Estimating a Parameter

- **Question:** What is the value of the parameter?
- **Terms:** **predict, estimate, construct a confidence interval, confidence level**
- **Answer:** Between x and y, with 95% confidence
- **Method** (11.2, 11.3):
  - Bootstrap the sample; compute estimate
  - Repeat; draw empirical histogram of estimates
  - Confidence interval is “middle 95%” of estimates
- Can replace 95% by other confidence level (not 100%)

# Meaning of “95% Confidence”

- You’ll never get to know whether or not your constructed interval contains the parameter.
- The confidence is in the process that generates the interval.
- The process generates a good interval (one that contains the parameter) about 95% of the time.
- End of 11.2

# Main Uses of Confidence Intervals

- To **estimate** a parameter: 11.3
  - Regression **prediction**, if regression model holds: Predict *y* based on a new *x*: 14.3
- To **test** the null hypothesis that a parameter is equal to a specified value: 11.4
  - In the regression model, used for testing whether the slope of the true line is 0: 14.2
  - In A/B testing, used for testing whether the difference between true means is 0: 16.2, 16.3

# Tests of Hypotheses

- **Null:** A well specified chance model: need to say exactly what is due to chance, and what the hypothesis specifies.
- **Alternative:** The null isn't true; something other than chance is going on; might have a direction
- **Test Statistic:** A statistic that helps you decide between the two hypotheses, based on its empirical distribution under the null

- 10.2

# The P-value

- The chance, **under the null hypothesis**, that the test statistic comes out equal to the one in the sample or more in the direction of the alternative
- If this chance is small, then:
  - If the null is true, something very unlikely has happened.
  - Conclude that the data support the alternative hypothesis better than they support the null.
- 10.3

# An Error Probability

- Even if the null is true, your random sample might indicate the alternative, just by chance
- The **cutoff** for P is the chance that your test makes the wrong conclusion when the null hypothesis is true
- Using a small cutoff limits the probability of this kind of error
- Second half of 10.3

# One Categorical Sample

- **Null:** The sample was drawn at random from a specified distribution.
- Test statistic: TVD between distribution in sample and distribution specified in the null.
- Method:
  - Simulation: Generate samples from the distribution specified in the null.

- 10.1 (juries), 10.2+10.3 (Mendel)

# One Sample, One Parameter

- **Null:** The parameter is equal to a specified value.
- **Alternative:** The parameter is not equal to that value; or parameter is greater than the value; or parameter is less than the value
- Test Statistic: Statistic that estimates the parameter
- Method:
  - Bootstrap: Construct a confidence interval and see if the specified value is in the interval.
- 10.2+10.3 (GSI's defense), 14.2 (slope of true line)