COMS BC1016
Introduction to Computational Thinking and Data Science

Lecture 12: Hypothesis Testing

Oct 27, 2025

Reminders/Announcements

- Labs start again this week!
- HW 5 Released, due next Wednesday
- Midterms are being graded
  - Aim to finish grading next week and pass back during next week lab
- Final Project!!
  - No final exam :)
  - https://www.eysalee.com/courses/f25/bc1016_final.html

Lecture Outline

- Final Project Info
- Model Recap
- Hypothesis Testing
- Total Variation Distance

Final Project Info

Overview

- Real world data science!
    - Explore a real dataset and ask questions that the dataset can help answer
        - Info on the report and required components discussed later
- At this point in the semester, we have *not* covered everything needed to complete this project!
    - We’re giving you the data sets early so you can start thinking about groups and which data sets you find interesting

Deadlines

~ 3 weeks away

5 weeks away

~ 6.5 weeks away

1. **Group Declaration**: Deadline TBA
   Please read the "Group Guidelines" section for guidelines on how to form your final project groups and complete the Google Form to indicate your group.
2. **Project Proposal**: Due **Friday, Nov 14** at 11:59pm
   Each group will select a final project notebook and dataset to work on for the final project and complete the introduction section of the report.
3. **Progress Report**: Due **Monday, Dec 1** at 11:59pm
   At this point, groups should be about ~60% done with the final project. For the progress report, groups should list out what analysis remains and how they plan on approaching it. Additionally, groups should share if they are running into any issues with their analysis that they may need assistance with or have questions about.
4. **Final Project Report**: Due **Friday, Dec 12** at 11:59pm
   Groups will submit the completed reports along with a completed peer review.

Note: We will require all students to complete a peer review to share how work was distributed among team members. Any major discrepancies in the distribution of work will be factored into individual grades on this assignment.

Final Project Grading Breakdown:

- Project Proposal - 10%
- Progress Report - 25%
- Final Report - 65%

Final Project Groups

- Groups of 2 (optionally 3 but with extra work)
  - Groups must be able to attend the same lab section for the final lab
    - Students in different labs and groups of 3 must email the teaching staff for approval
- Google Form for registering your group: Link
  - Deadline TBD, expected 1.5-2 weeks

Datasets

- Can choose any one of these four categories:
  - Airbnb
  - NYC Restaurants
  - Seattle Pets
  - Spotify
- *Do not look at or use any external analysis of these datasets. Doing so will result in a 0 for your final project grade.*

Welcome to Stats (Pt II of BC1016)

Programming

Iteration

Manipulating Arrays & Tables

Data Types

Functions

Building Visualizations

Conditionals

Midterm Exam

Statistics

Probabilities

Confidence Intervals

Correlation

Linear Regression

P-value & Statistical Significance

Residuals

Final Project

Stats in Data Science

- How to test hypotheses about the world using data
- How to report on how confident you are in the results of a test
- How to design experiments and consider sample size
- How to communicate your process and findings
- Note:
  - You largely know all the programming you need to do this
  - Not much new programming material this part of the class

Model Recap

Models

A model is a set of assumptions about data

- In data science, many models involve assumptions about the processes that involve randomness

- Question: Does the model fit the data?

Terminology

Parameter: Number associated with the population

- Example: average, max, min, mean

Statistic: A number calculated from the sample, can be used to describe the distribution

- A statistic can be used as an estimate of a parameter

- Example: sample mean, sample max, sample min

Assessing Models

- Suppose we have a statistical model that describes how data should behave (based on certain assumptions)
    - If we can use that model to **generate (simulate)** fake data, then we can see what the model “thinks” the data should look like
    - By simulating data, we can see the kinds of outcomes or patterns the model expects, i.e., its **predictions**
- We can then compare the predictions to the data that were observed (irl data)
- If the data and the model’s predictions are not consistent, that is evidence against the model.

Swain vs. Alabama

- Would an 8% black jury be a realistic outcome if jury section were truly unbiased?

Eligible Jurors	Empaneled Jurors	Final Jury
26% Black	8% Black	0 Black

Assessing *Swain v Alabama*

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

Model: Panelists were selected at random and the small number of Black panelists is by chance

Alternative view: too few Black panelists for it to have been a random sample

Statistic: Number (count) of Black panelists

Sampling from a Categorical Distribution

- Sample at random from a categorical distribution

  – `sample_proportions(sample_size, pop_distribution)`

- `pop_distribution` is a list or array that adds up to 1

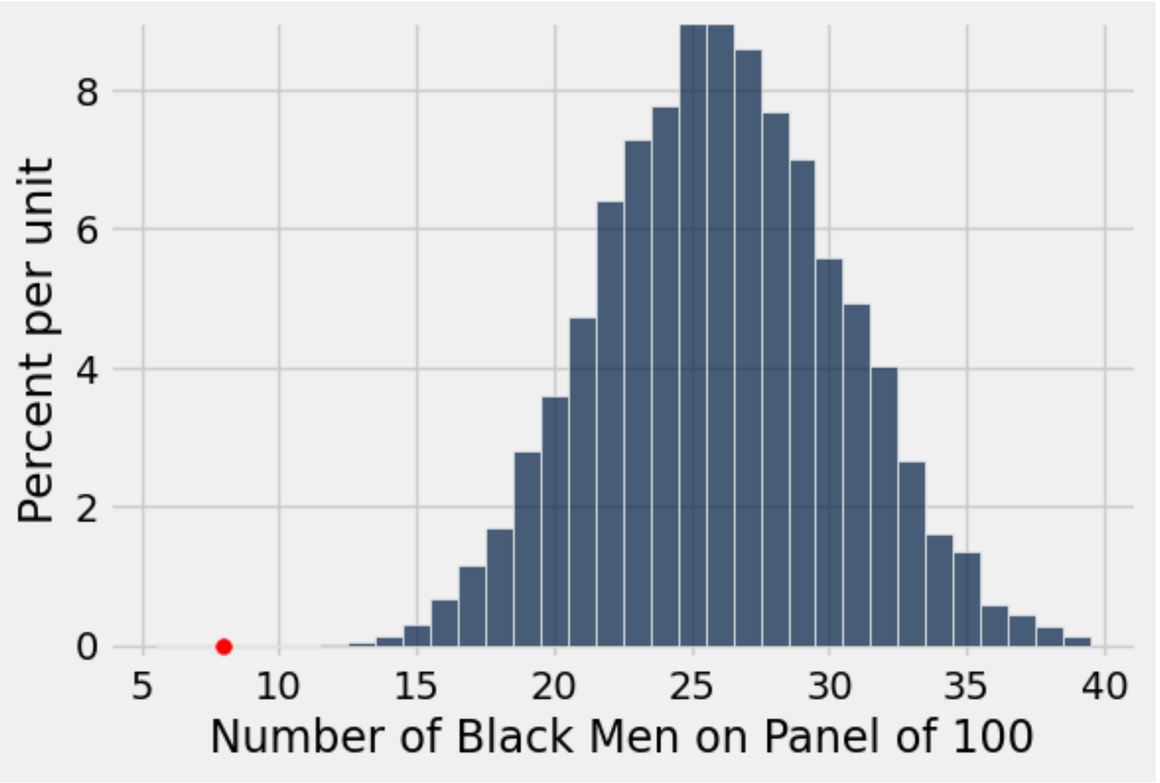
- Function returns an array containing the empirical distribution of the categories in the sample

Swain v Alabama Example

- Used `sample_proportions` to simulate sampling 10,000 jury panels

- Plotted experiment and percent of Black panelists (8%) was **highly unlikely** under random sampling

- We observed **statistical bias**, when differences between the parameters and the statistics are systematically in one direction



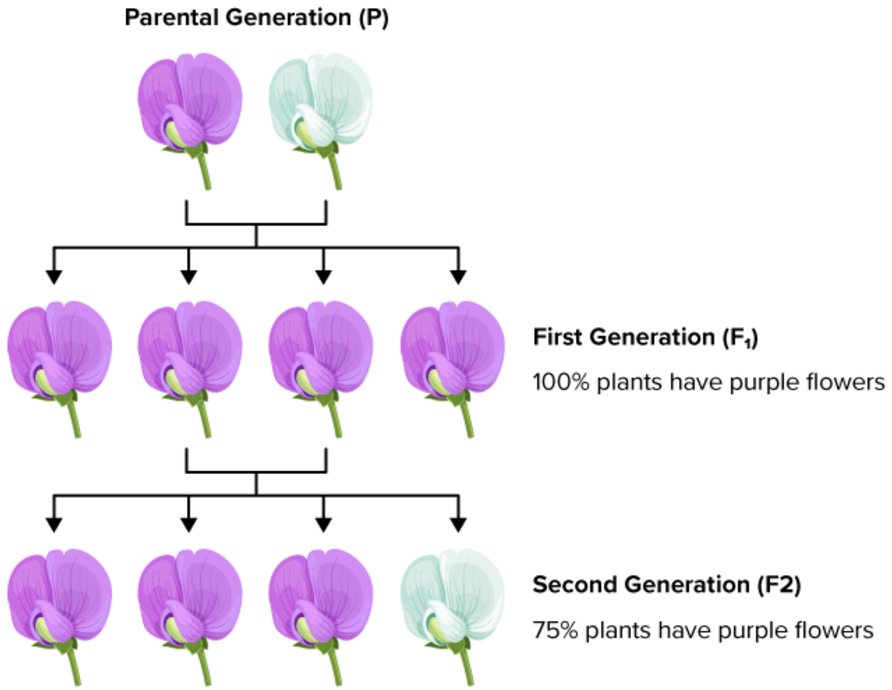
Models and Hypothesis Testing

Gregor Mendel

- Austrian monk interested in botany
- In the 1850s, experimented with pea plants
  - Self-pollinated plants and collected seeds until plants had similar characteristics each generation (e.g., flower color, plant height)
  - He crossed plants with different characteristics and had the offspring self-pollinate

Mendel’s Genetic Model

- Pea plants can either have purple flowers or white flowers (but not a mix)
    - Of 929 plants, 709 had purple flowers
    - Mendel theorized a pattern in the frequency of purple in the second generation
- Model:
    - Crossed plants have a 75% chance of having purple flowers
- Question: Is this model accurate?



Assessing Mendel's Model

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

Model: Plants have a 75% chance of having purple flowers

Assessing Mendel's Model

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

Model: Plants have a 75% chance of having purple flowers

Alternative: The percentage is much larger or much smaller than 75%

Assessing Mendel's Model

1. Choose a statistic that will help you decide whether the data supports the model or an alternative view of the world

2. Simulate the statistic under the assumptions of the model

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

Model: Plants have a 75% chance of having purple flowers

Alternative: The percentage is much larger or much smaller than 75%

Statistic: Distance from 75%

$$\text{distance}(x, y) = |x - y|$$

Mendel Notebook Demo

Hypothesis Testing

- **Hypothesis Testing:** A statistical test in which we choose between two potential views
  - **Null Hypothesis:** Clearly defined model based on chance. Data is generated randomly and under clearly specified assumptions
    - This is the one we can simulate and test!
  - **Alternative Hypothesis**: The observed data differs from the null hypothesis in some way other than chance
    - Doesn’t say how or why the model isn’t good, just that it isn’t good

Hypothesis Testing: Examples

- Swain v Alabama
  - Null Hypothesis: The jury selection represented the larger population
  - Alternative Hypothesis:

Hypothesis Testing: Examples

- Swain v Alabama
  - Null Hypothesis: The jury selection represented the larger population
  - Alternative Hypothesis: Jury selection was biased

Hypothesis Testing: Examples

- Swain v Alabama
  - Null Hypothesis: The jury selection represented the larger population
  - Alternative Hypothesis: Jury selection was biased
- Mendel's Purple Flowers
  - Null Hypothesis: Each plant has a 75% change of having purple flowers
  - Alternative Hypothesis:

Hypothesis Testing: Examples

- Swain v Alabama
  - Null Hypothesis: The jury selection represented the larger population
  - Alternative Hypothesis: Jury selection was biased
- Mendel's Purple Flowers
  - Null Hypothesis: Each plant has a 75% change of having purple flowers
  - Alternative Hypothesis: The chance of having purple flowers is not 75%

Jury Selection in Alameda County

- In 2010, ACLU of Northern California reported that racial and ethnic groups were not properly represented in jury panels in Alameda County, CA

  - 11 felony trials over 2 years (2009 and 2010)

  - Collected demographic data on the 1453 panelists and compared to eligible jurors in the county

Comparing Distributions

- Mendel example we computed the distance of random samples from the model

$$\text{distance}(x, y) = |x - y|$$

- For this, we can compute a generalized version of distance
  - **Total Variation Distance:** Measures the distance between two categorical distributions

Ethnicity	% in Population
Asian	15
Black	18
Latino	12
White	54
Other	1

Computing Total Variation Distance (TVD)

- For each category, compute the difference in proportion between two distributions (under null hypothesis and empirical / observed)

- Take the absolute value of each difference

- Sum for all categories and then divide the sum by 2

```
def tvd(dist1, dist2):
    return sum(abs(dist1 - dist2))/2
```

Summary of Process of Applying TVD

- To assess whether a sample was drawn randomly from a known categorical distribution using TVD:
    - Sample at random from the population and compute the TVD from the random sample
    - Repeat many times
    - Compare the TVD empirical distribution of simulated to the actual TVD from the sample

Jury Selection Notebook Demo

Next time

- Tests with numerical (non-categorical) data
  - P-value!