

COMS BC1016

Introduction to Computational Thinking and Data Science

Lecture 14: Hypothesis Testing

BARNARD COLLEGE OF COLUMBIA UNIVERSITY

Sept 30, 2025

Copyright © 2026 Barnard College

March 23, 2026

Reminders/Announcements

- **Labs start again this week!**
 - Lab schedule has been updated on the main course website
- **HWs 5-9 are being pushed back a week**
 - HW 5 will release next week Monday, etc
 - **HW 9 is now optional**
 - To keep due dates within the semester, there's slightly less time to work on HW 9
 - **HW 9 replaces your second lowest homework grade if you do it**

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 early so you can start thinking about groups and which data you find interesting

Final Project Website

Webpage with details is now up: https://www.eysalee.com/courses/s26/bc1016_final.html

COMS BC1016 Spring '26 [Course Home](#) [Syllabus](#) [Final Project](#)

COMS BC1016 Final Project

Spring 2026
Barnard College

The final project for BC1016 provides an opportunity to bring together, apply, and communicate your knowledge of data science and statistics from this course. You will work in groups of 2 to choose one of the provided datasets to analyze and submit a writeup of your analysis and conclusions.

AI and Usage of Outside Resources

As a reminder, the syllabus states:

For your final project, AI generated text is **not permitted** as part of your written descriptions in your final report. Your report must include your own original writing and reflections. Violations can result in a failing grade for the assignment and/or the course.

Please note this policy includes using generative AI to produce analysis. **You are not allowed to submit AI generated code.**

If you like to do anything more advanced than what we have covered in class, you must include a brief explanation of where you learned these concepts. For instance, if you have taken prior statistics classes and would like to perform a more complicated analysis or if you have experience with (and would like to use) Python libraries we have not covered in class.

Project Milestones and Deadlines

1. **Group Declaration:** Deadline **Wednesday, April 1**





Please complete one of the two Google Forms to indicate your group preference: [Group Declaration](#) (if you know what group you want to be in) or [Group Matching](#) (if you do not have a group and would like us to form one for you)

2. **Project Proposal:** Due **Friday, April 17** at 11:59pm

Each group will select a final project notebook and dataset to work on for the final project and complete the Exploratory Data Analysis section.

Based on your exploratory data analysis, you will state the hypothesis you are planning to test.

Project Milestones and Deadlines

- 1.5 weeks away  1. **Group Declaration:** Deadline **Wednesday, April 1**
Please complete one of the two Google Forms to indicate your group preference: [Group Declaration](#) (if you know what group you want to be in) or [Group Matching](#) (if you do not have a group and would like us to form one for you)
- ~3.5 weeks away  2. **Project Proposal:** Due **Friday, April 17** at 11:59pm
Each group will select a final project notebook and dataset to work on for the final project and complete the Exploratory Data Analysis section. Based on your exploratory data analysis, you will state the hypothesis you are planning to test.
- 5 weeks away  3. **Progress Report:** Due **Monday, April 27** 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.
- ~ 6.5 weeks away  4. **Final Project Report:** Due **Friday, May 8** 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:

- Group Declaration - 1%
- Project Proposal - 9%
- Progress Report - 25%
- Final Report - 65%

Final Project Groups

- This project is designed to be worked in groups of 2, but you are allowed to work in groups of 1-3
 - Groups of 3 will have slightly more to do to account for the extra teammate
 - Working alone will require the same amount of work as a group of 2
- Fill out a Google Form letting us know who you want to work with:
 - If you already know what group you want to be in (including alone): [Group Declaration](#)
 - If you don't have a group and want us to put you into one: [Group Matching](#)
- Complete at least one of these forms by **Wednesday, April 1**
 - This is worth **1%** of your final project grade

Datasets

- Can choose any one of these four categories :
 - NYC Restaurant Inspections
 - Seattle Pet Licenses
 - Taylor Swift Songs
 - Allrecipes
- If your group wants to analyze something aligned with your interests outside of these options, you must get my approval
- *Do not look at or use any external analysis of these datasets. Doing so will result in a 0 for your final project grade.*

Programming

Data Types

Iteration

Manipulating Arrays
& Tables

Functions

Conditionals

Building Visualizations

Statistics

Probabilities

Confidence Intervals

Midterm Exam

Correlation

Linear Regression

P-value & Statistical
Significance

Residuals

Final Project

Midterm Recap (Notebook Demo)

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 selection were truly unbiased?

Eligible Jurors

26% Black

Empaneled Jurors

8% Black

Final Jury

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

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

2. Simulate the statistic under the assumptions of the model

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

3. Draw a histogram of the simulated values

4. Compute the statistic from the sample in the study

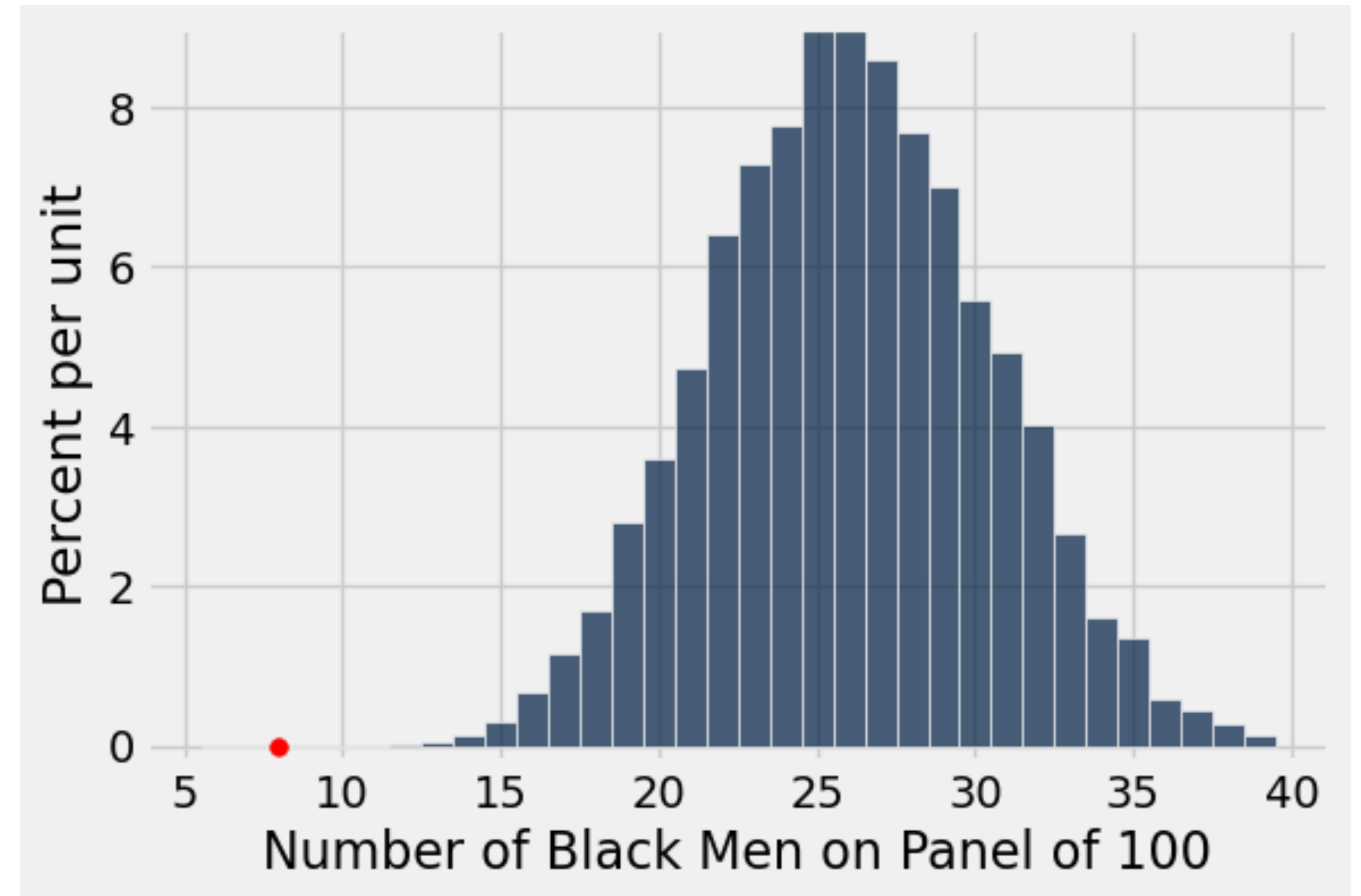
Statistic: Number (count) of Black panelists

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

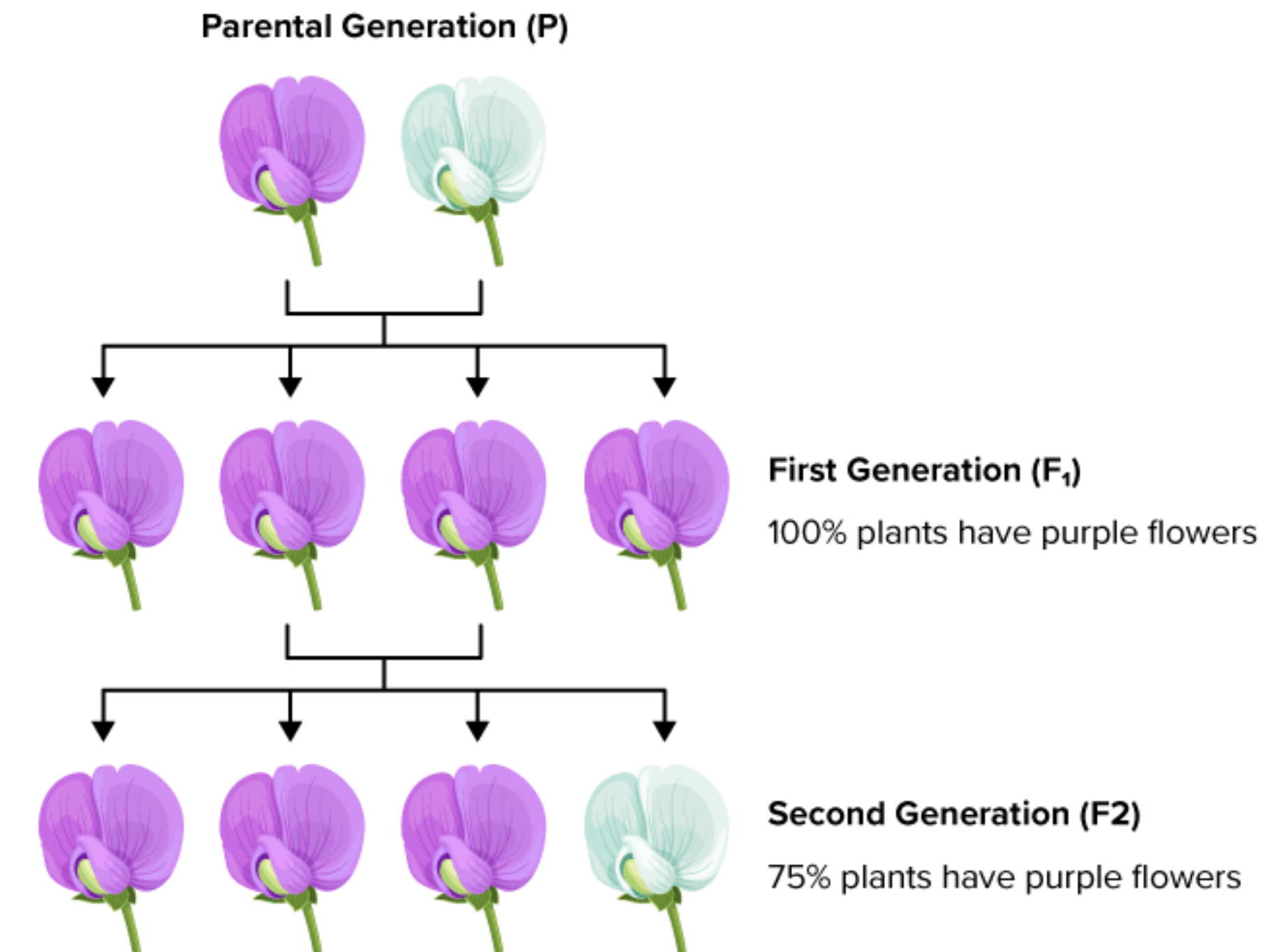
Gregor Mendel

- 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% chance 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% chance of having purple flowers
 - Alternative Hypothesis: The chance of having purple flowers is not 75%

Our Examples So Far

- Swain v Alabama
 - Model: Jury representation matched the overall population
 - Distribution:
 - Black and Non-Black
- Mendel
 - Model: Plants a 75% chance of having purple flowers
 - Distribution:
 - Purple or White



these are both categorical, and there are just two options

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

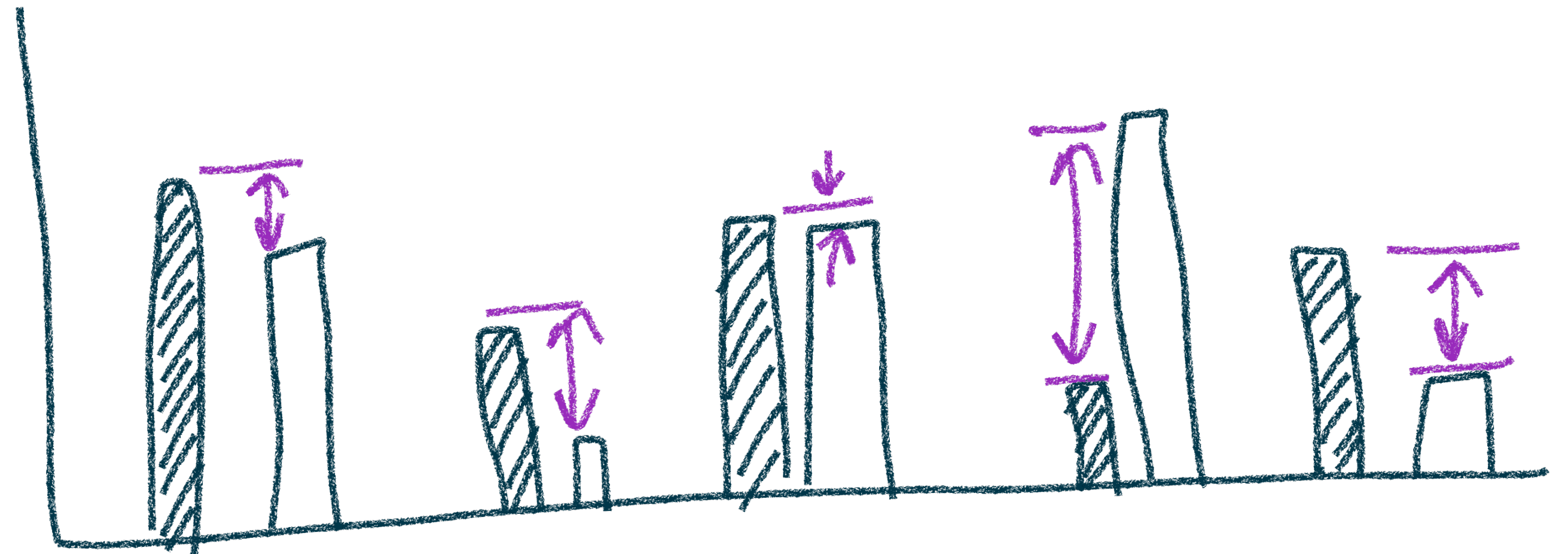
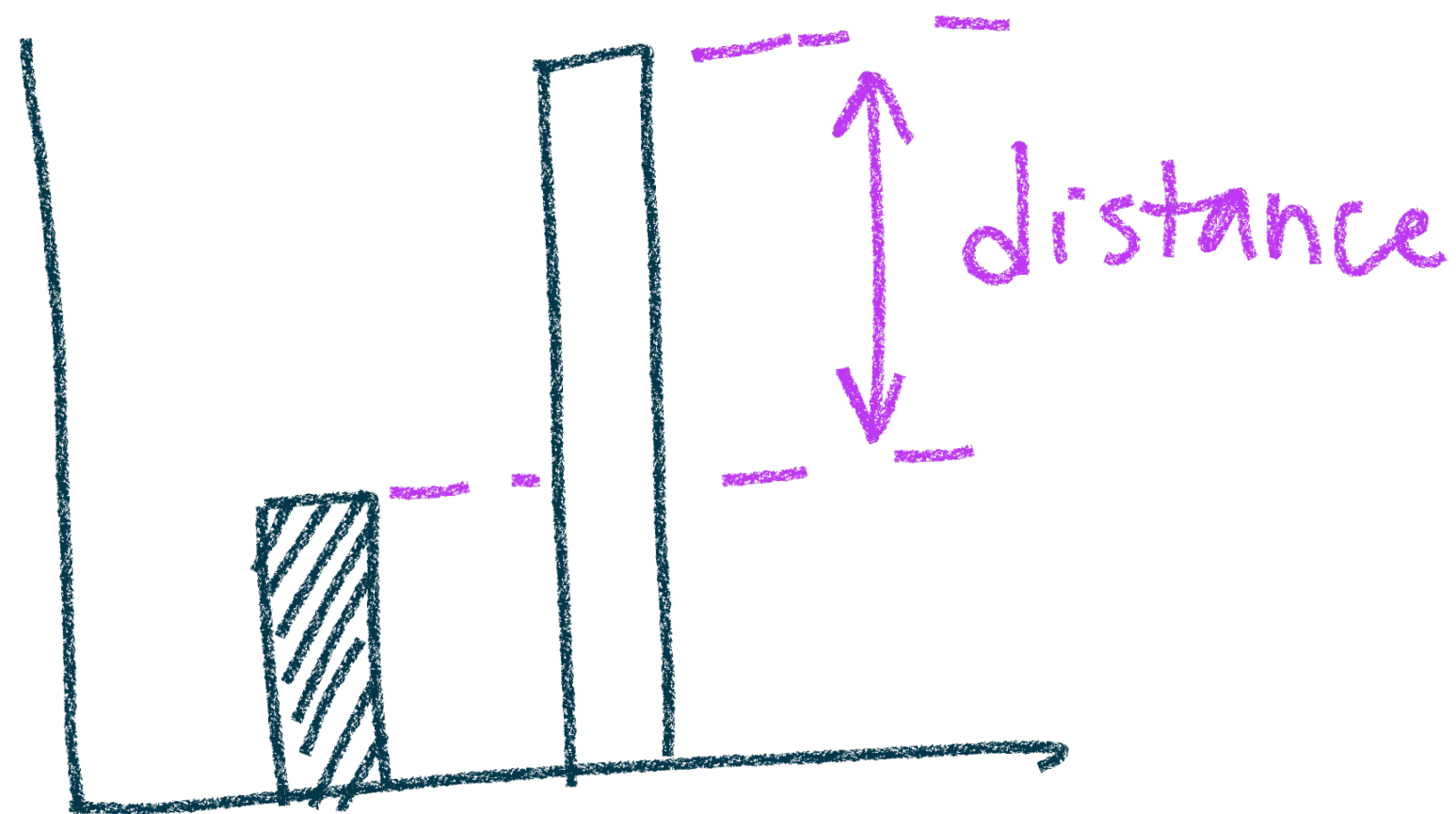
$$\text{distance}(x, y) = |x - y|$$

- 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

Total Variation Distance

▨ distribution A
□ distribution B



sum all these differences
and divide by 2

+/- double counts distances

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!