DATA 8
Spring 2020

Lecture 22

Midterm Review

Announcements

- Midterm is tonight, 7:10 - 9:00 pm **on Gradescope.com**
  - Multiple versions - will receive email at 6:00 pm
- Midterm Review walkthrough posted here
- Midterm concerns?
  - This will make you feel better :)

Testing Hypotheses

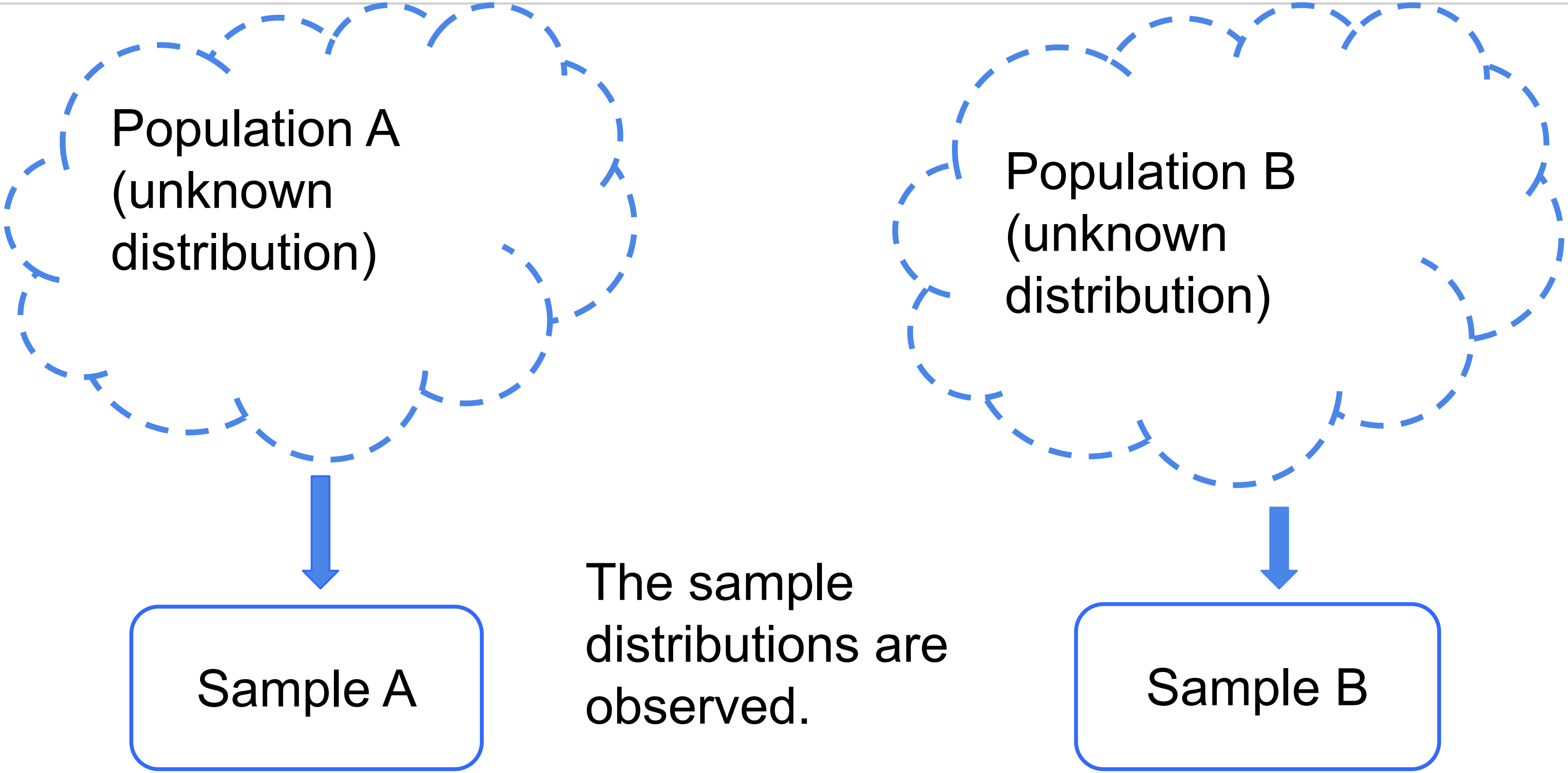
Before You Compute Anything

Figure out the viewpoints the question wants to test.

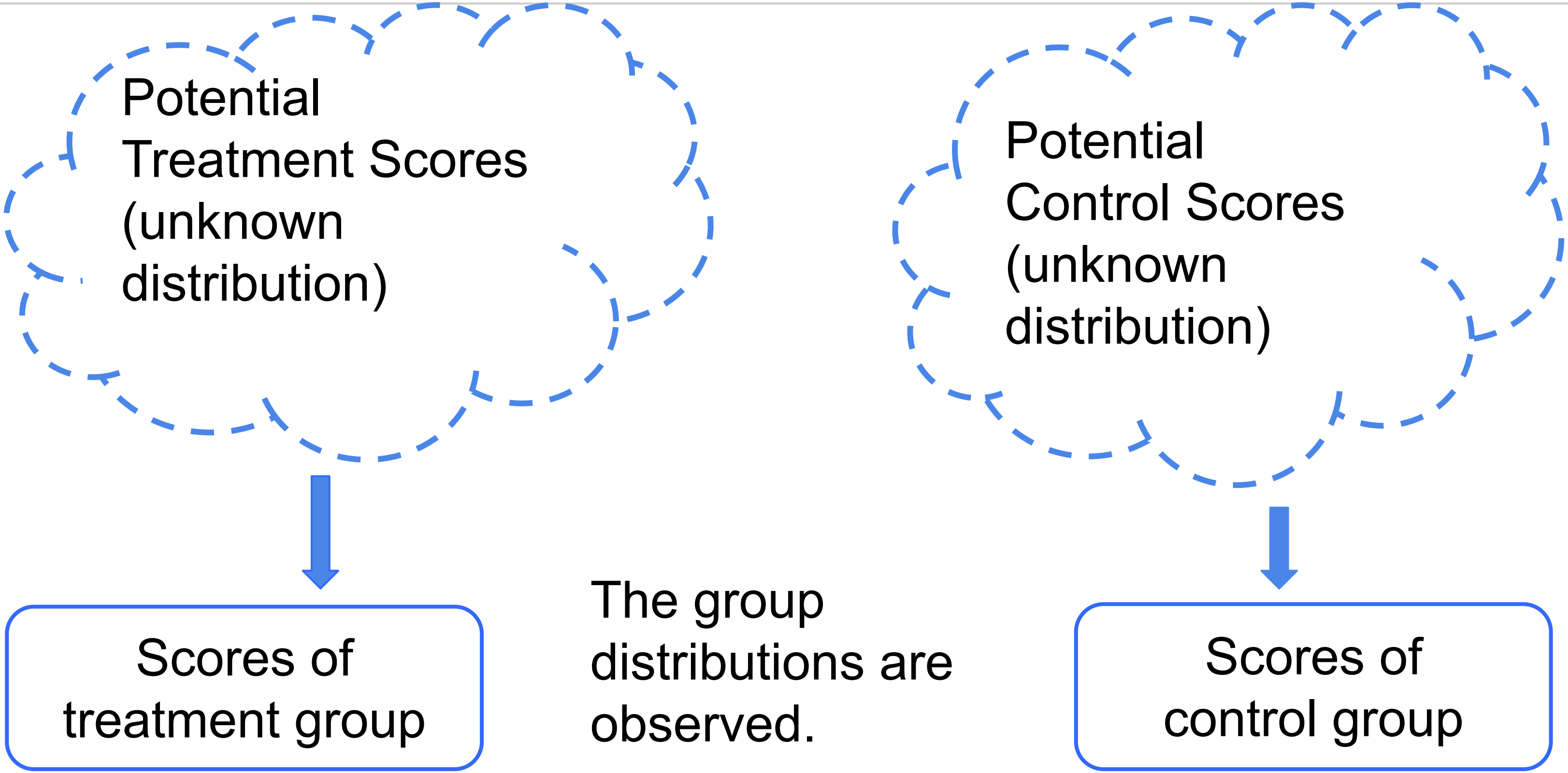
- **Null hypothesis:** Completely specified chance model under which you can simulate data
- **Alternative hypothesis:** The opposing viewpoint in the question
- **Test statistic:** Should help you decide which of the two hypotheses is better supported by the data
    - For the P-value calculation: What kinds of values of this statistic make you lean towards the alternative?

Two Random Samples: A/B Testing

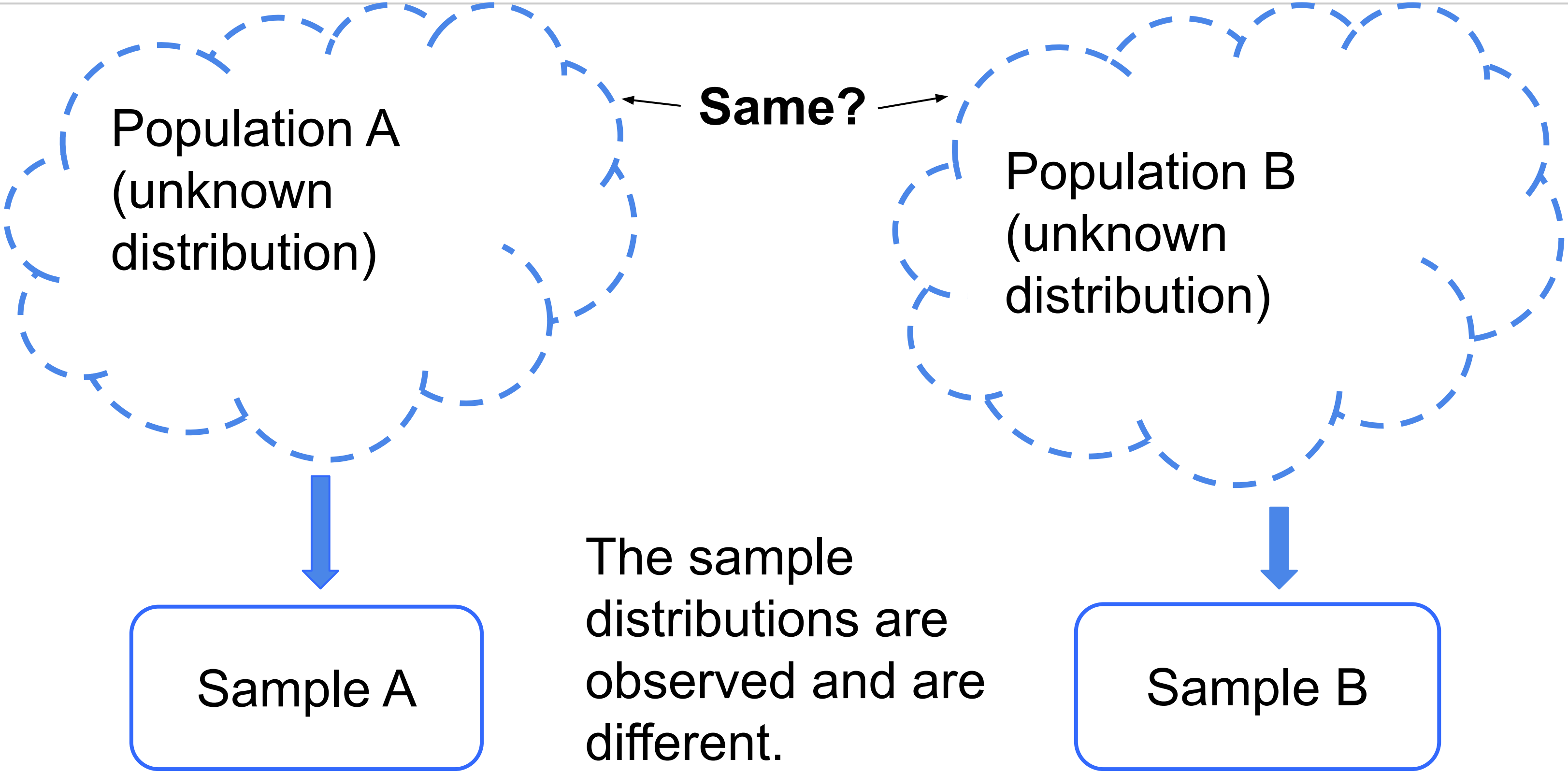
Populations and Samples



Example: RCT


Potential Treatment Scores (unknown distribution)
Potential Control Scores (unknown distribution)
Scores of treatment group
The group distributions are observed.
Scores of control group

The Question



The Hypotheses

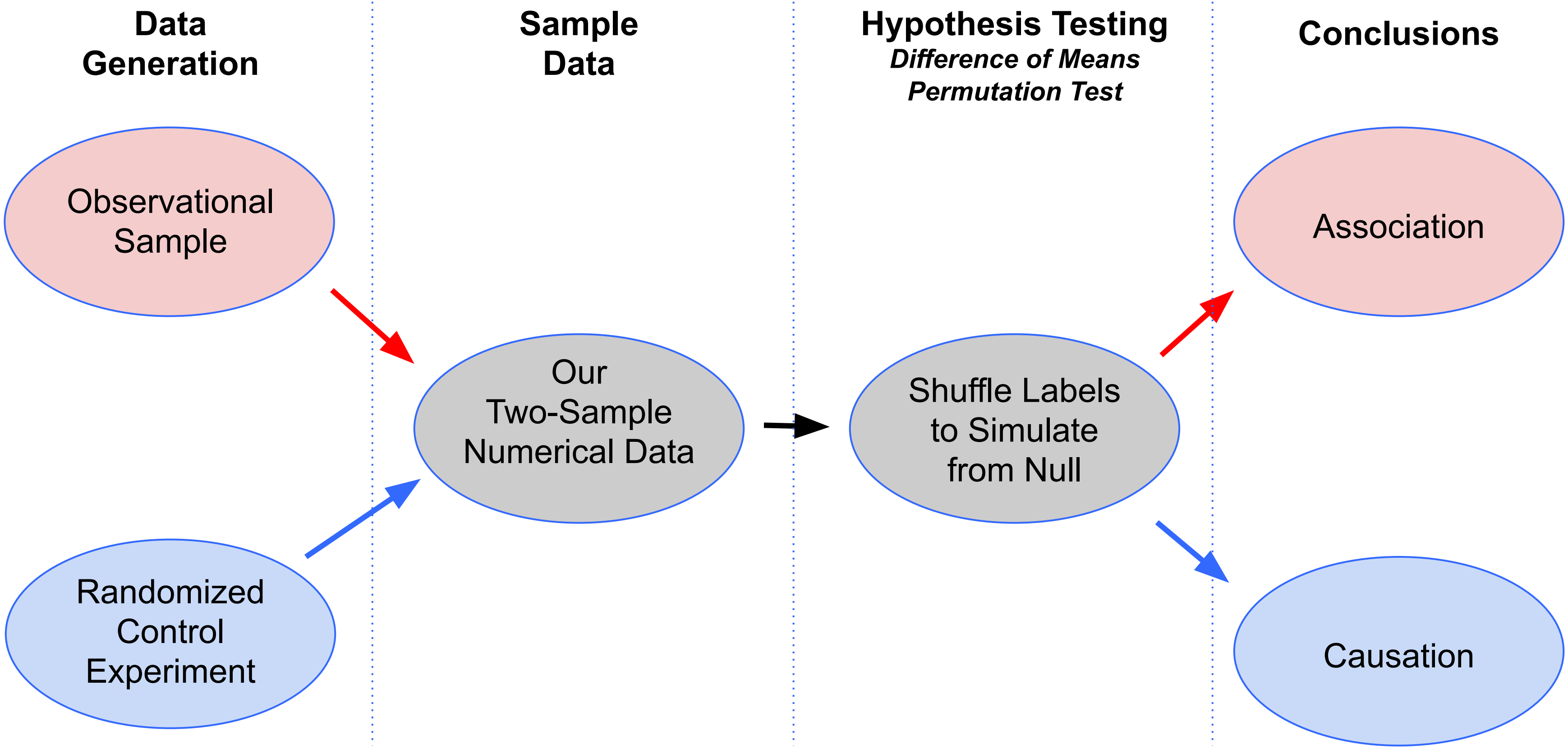
- **Null:** The distributions in the two populations are the same. (The distributions in the samples are different due to chance.)
- The alternative depends on the question. For example:
  - The values in Group A are on average smaller than the values in Group B.
  - … larger than ...
  - ... different from ...

Simulating Under the Null

If the two population distributions are the same, then:

- It doesn’t matter which sampled individual is labeled A and which is labeled B
- So you can label at individuals random, provided you ensure that the two randomly labeled groups have the same sizes as the original ones
- This ensures comparability of the simulated statistics and the observed one

Random Assignment & Shuffling



The P-Value of a Test

Definition of the *P*-value

The *P*-value is the chance,

- if the null hypothesis is true,
- that the test statistic
- is equal to the value that was observed in the data
- or is even further in the direction of the alternative.

P-value is high → more evidence for the null

P-value is low → more evidence for the alternative

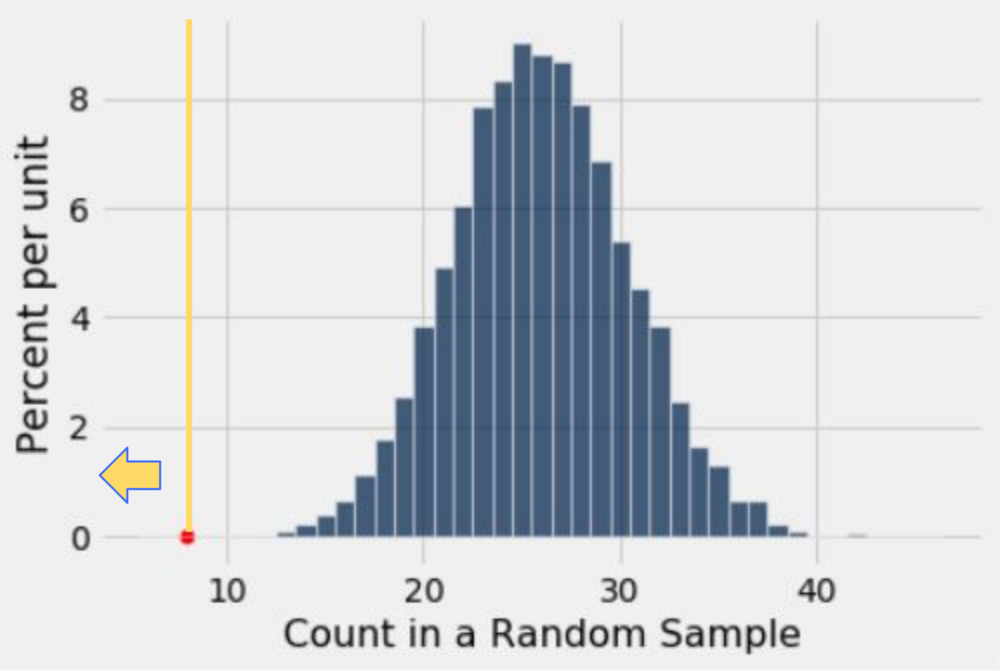
Swain v. Alabama

- **Null:** The jury panel was drawn at random from a population that had 26% black men.
- **Alternative:** There were too few black men on the panel for it to look like a random sample.
- **Test statistic:**

  Number of black men in panel
- **Small values of the statistic support the alternative.**

Statistic Simulated Under the Null

- The P-value is the area at or **to the left** of the observed value (red dot)
- Very close to 0%
- Test favors the alternative



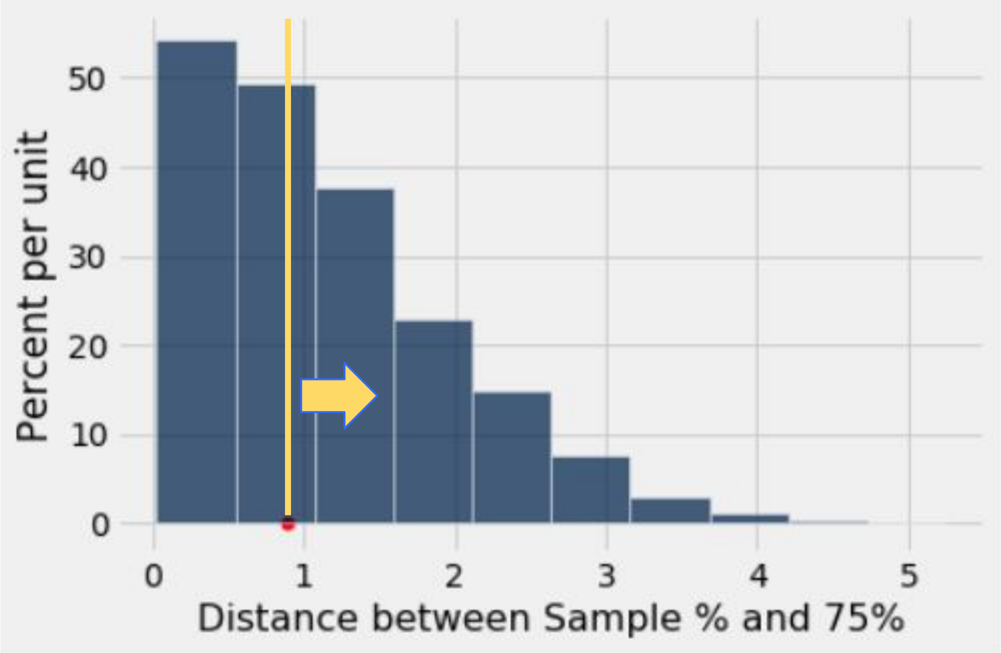
Mendel's Model

- **Null:** Each pea plant has 75% chance of being purple flowering, independently of all other plants.
- **Alternative:** The model isn't good.
- **Test statistic:**

  | percent purple in sample − 75 |
- **Large values of the statistic support the alternative.**

Statistic Simulated Under the Null

- The P-value is the area at or **to the right** of the observed value (red dot)
- Bigger than 50%
- Test favors the null



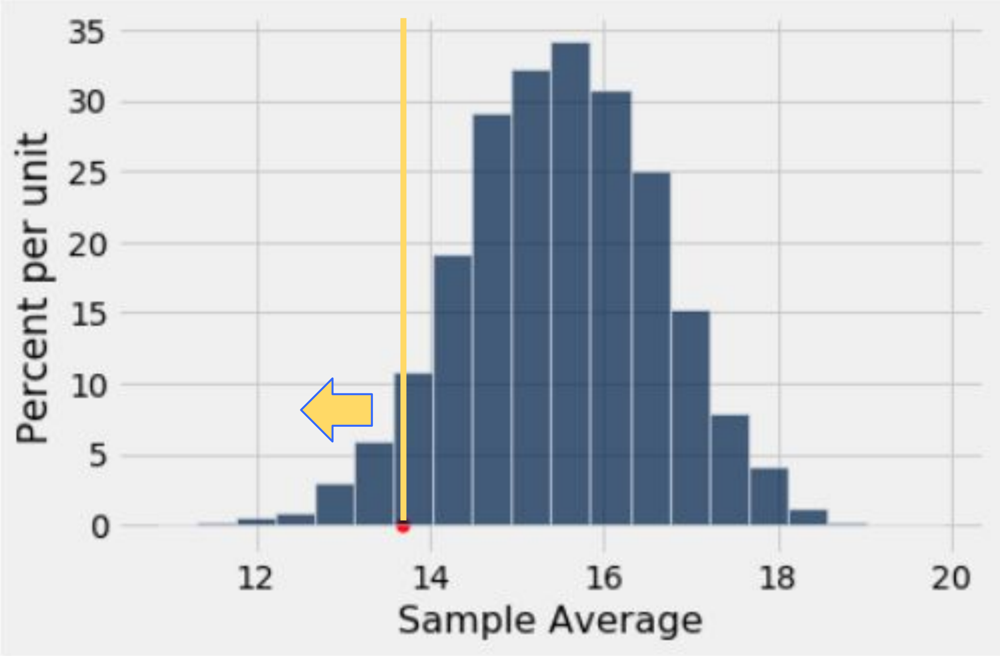
GSI’s Defense

- **Null:** Section 3 scores are are like a sample drawn at random without replacement from the whole class.
- **Alternative:** The Section 3 average is too low for the section to be a random sample from the class.
- **Test statistic:**

  Section 3 average
- **Small values of the statistic support the alternative.**

Statistic Simulated Under the Null

- The P-value is the area at or **to the left** of the observed value (red dot)
- About 5.6%
- Test favors the null if you are strict about the 5% cutoff



Hypothesis Testing Review

- **1 Sample: One Category** *(e.g. percent of flowers that are purple)*
  - Test Statistic: `empirical_percent`, `abs(empirical_percent - null_percent)`
  - How to Simulate: `sample_proportions(n, null_dist)`
- **1 Sample: Multiple Categories** *(e.g. ethnicity distribution of jury panel)*
  - Test Statistic: `tvd(empirical_dist, null_dist)`
  - How to Simulate: `sample_proportions(n, null_dist)`
- **1 Sample: Numerical Data** *(e.g. scores in a lab section)*
  - Test Statistic: `empirical_mean`, `abs(empirical_mean - null_mean)`
  - How to Simulate: `population_data.sample(n, with_replacement=False)`
- **2 Samples: Numerical Data** *(e.g. birth weights of smokers vs. non-smokers)*
  - Test Statistic: `group_a_mean - group_b_mean`, `group_b_mean - group_a_mean`, `abs(group_a_mean - group_b_mean)`
  - How to Simulate: `empirical_data.sample(with_replacement=False)`

Cutoffs vs. P-values

The Cutoff

- It is your threshold for deciding whether or not you think the P-value is small.
- It is an *error probability*: approximately the chance that the test concludes the alternative when the null is true
  - You get to choose the cutoff. So you get to control this error probability.
- The cutoff does not depend on the data. It is often chosen before the data are collected.

P-value cutoff vs P-value

- P-value cutoff
  - Does not depend on observed data or simulation
  - Decide on it before seeing the results
  - Conventional values at 5% and 1%
  - Probability of hypothesis testing making an error
- P-value
  - Depends on the observed data and simulation
  - Probability under the null hypothesis that the test statistic is the observed value or further towards the alternative

The P-Value

Which of the following does the P-value depend on?

- Null hypothesis
- Alternative hypothesis
- The choice of test statistic
- The data in the sample
- The cut-off (e.g. 5%)

Answer: All except the cutoff

Probability

Exercise 1

Marbles: G, G, G, G, R, R, R, B, B, Y. Draw 4 at random **with** replacement.

P(all G) = ?

P(no G) = ?

P(at least one G) = ?

P(all G) = (4/10)*(4/10)*(4/10)*(4/10)

P(no G) = (6/10)*(6/10)*(6/10)*(6/10)

P (at least one G) = 1 - P(no G)

Exercise 2

Marbles: G, G, G, G, R, R, R, B, B, Y. Draw 4 at random **without** replacement.

P(all G) = ?

P(no G) = ?

P(at least one G) = ?

P(all G) = (4/10)*(3/9)*(2/8)*(1/7)

P(no G) = (6/10)*(5/9)*(4/8)*(3/7)

P (at least one G) = 1 - P(no G)

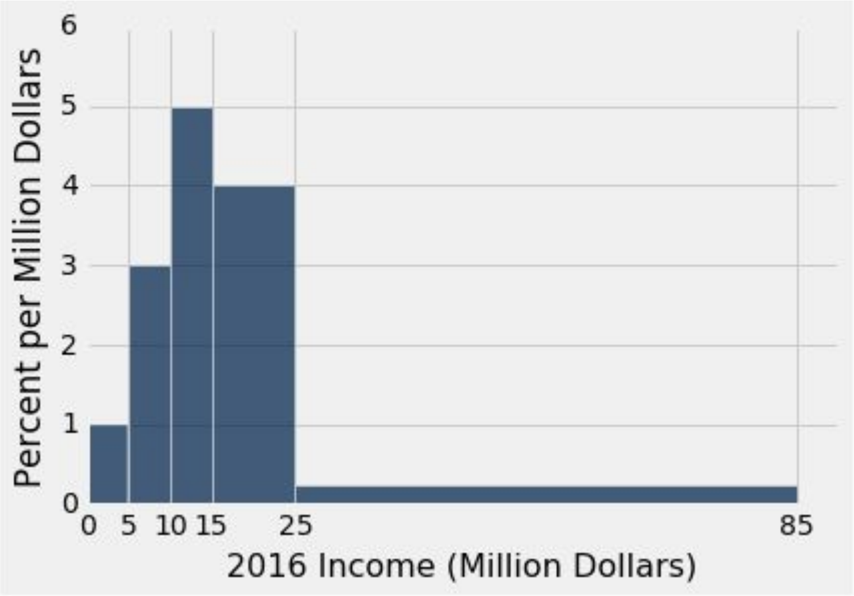
Histograms

Using the Density Scale

(a) Which bin has more people: [10, 15) or [15, 25)?

(b) What percent of incomes are in the [25, 85) bin?

(c) If you draw one bar over [10, 25), how tall will it be?



Answers

(a) [15, 25)

(b) 15%

(c) 4.33 percent per million dollars

Arrays

Arrays

When you want to do the same thing to each of many things => use array operations

Add to end of array => np.append

Count number that aren't zero/False => np.count_nonzero

Tables

Table Operations

Keep some of the columns => select, drop

Keep some of the rows => where, take

Add a column => with_column

Find smallest/biggest => sort, then take first

Table Operations

Combine information from two tables => join

Compute an aggregate, broken down by 1 attribute => group

Compute an aggregate, broken down by 2 attributes => pivot