DATA 8

Spring 2020

Lecture 18

Decisions & Uncertainty

Weekly Goals

- Monday
  - Comparing distributions
  - Hypothesis tests and p-values
- **Today**
  - Making decisions with incomplete information
  - Error probabilities
- Friday
  - A/B testing
  - Permutation Test

Announcements

- This Week
  - Ramesh's office hours cancelled today
  - Lab attendance is optional (Lab 6 now due Fri 11:59PM)
  - Homework 6 due tomorrow
- Next Week
  - Tutoring sections via Google Hangouts
  - Midterm Review Lab attendance will be optional
  - Midterm on March 13th, 7PM
    - **Midterm Review Cancelled**
      - Video/Slides uploaded before Wednesday

Decisions and Uncertainty

Incomplete Information

- We are trying to choose between two views of the world, based on data in a sample.
- It is not always clear whether the data are consistent with one view or the other.
- Random samples can turn out quite extreme. It is unlikely, but possible.

Review: Terminology

Testing Hypotheses

- A test chooses between two views of how data were generated
- The views are called **hypotheses**
- The test picks the hypothesis that is better supported by the observed data

Null and Alternative

The method only works if we can simulate data under one of the hypotheses.

- **Null hypothesis**
  - A well defined chance model about how the data were generated
  - We can simulate data under the assumptions of this model – “under the null hypothesis”
- **Alternative hypothesis**
  - A different view about the origin of the data

Test Statistic

- The statistic that we choose to simulate, to decide between the two hypotheses

Questions before choosing the statistic:

- What values of the statistic will make us lean towards the null hypothesis?

- What values will make us lean towards the alternative?
  - Preferably, the answer should be just “high” or just “low”. Try to avoid “both high and low”.

Prediction Under the Null Hypothesis

- Simulate the test statistic under the null hypothesis; draw the histogram of the simulated values
- This displays the **empirical distribution of the statistic under the null hypothesis**
- It is a prediction about the statistic, made by the null hypothesis
  - It shows all the likely values of the statistic
  - Also how likely they are **(if the null hypothesis is true)**
- The probabilities are approximate, because we can't generate all the possible random samples

Conclusion of the Test

Resolve choice between null and alternative hypotheses

- Compare the **observed test statistic** and its empirical distribution under the null hypothesis
- If the observed value is **not consistent** with the distribution, then the test favors the alternative (“data is more consistent with the alternative”)

Whether a value is consistent with a distribution:

- A visualization may be sufficient
- If not, there are conventions about “consistency”

Another Example

The Problem

- Large(-ish) Statistics class divided into 12 discussion sections
- Graduate Student Instructors (GSIs) lead the sections

- After the midterm, students in Section 3 notice that the average score in their section is lower than in others

The GSI's Defense

GSI's position (Null Hypothesis):

- If we had picked my section at random from the whole class, we could have got an average like this one.

Alternative:

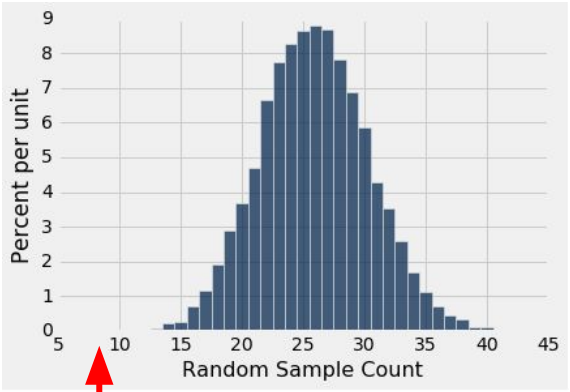
- No, the average score is too low. Randomness is not the only reason for the low scores.

(Demo)

Statistical Significance

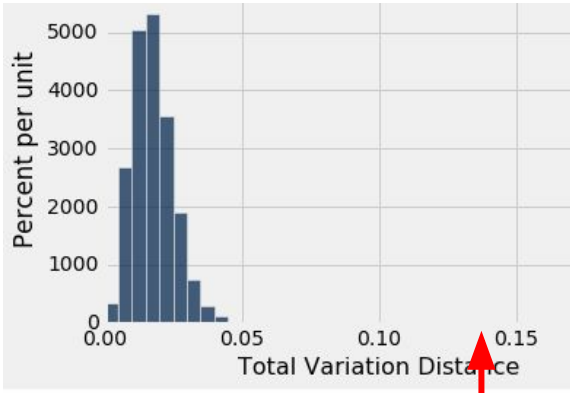
Tail Areas

Alabama Jury



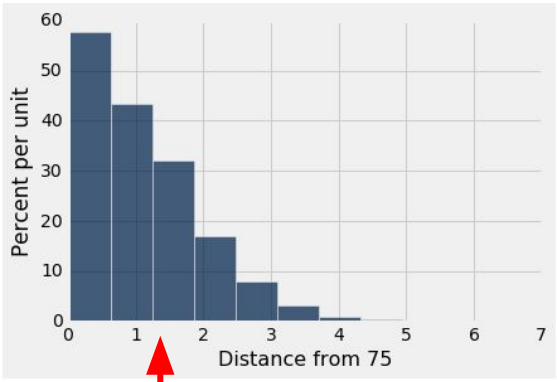
Observed Number (8)

Alameda Jury



Observed TVD (0.14)

Pea Plants



Observed Distance (1.32)

Conventions About Inconsistency

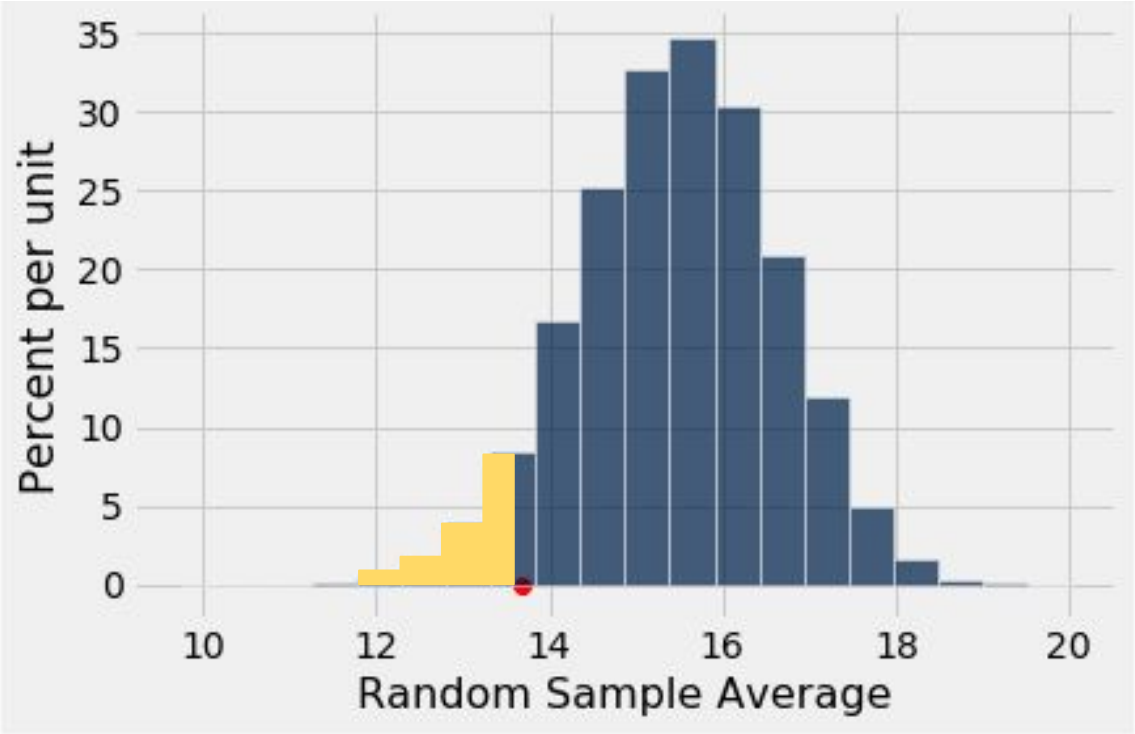
- **"Inconsistent with the null":** The test statistic is in the tail of the empirical distribution under the null hypothesis
- **"In the tail," first convention:**
  - The area in the tail is less than 5%
  - The result is "statistically significant"
- **"In the tail," second convention:**
  - The area in the tail is less than 1%
  - The result is "highly statistically significant"

(Demo)

The P-Value as an Area

Empirical distribution of the test statistic under the null hypothesis

The red dot is the observed statistic.



Definition of the *P*-value

Formal name: **observed significance level**

The *P*-value is the chance,

- under the null hypothesis,
- 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.

How We've Tested Thus Far

Hypothesis Testing Review

- **One Category** *(ex: percent of flowers that are purple)*
  - Test Statistic (1): `empirical_percentage`
  - Test Statistic (1): `abs(empirical_percentage - null_percentage)`
  - How to Simulate: `sample_proportions(n, null_dist)`
- **Multiple Categories** *(ex: ethnicity distribution of jury panel)*
  - Test Statistic: `tvd(empirical_dist, null_dist)`
  - How to Simulate: `sample_proportions(n, null_dist)`
- **Numerical Data** *(ex: scores in a lab section)*
  - Test Statistic: `empirical_mean`
  - How to Simulate: `population_data.sample(n, with_replacement=False)`