Data 8
Spring 2022

Lecture 22

Midterm Review

Announcements

(Continuing Last Lecture…)

Discussion Question

- Manufacturers of Super Soda run a taste test
- 91 out of 200 tasters prefer Super Soda over its rival

Question: Do fewer people prefer Super Soda than its rival, or is this just chance?

Null hypothesis: The same proportion of people prefer Super as Rival

Alternative hypothesis: A smaller proportion of people prefer Super

Test statistic: Number of people (out of 200) who prefer Super

p-value: Start at the observed statistic and look which way? Left

(Demo)

Hypothesis Test Concerns

The outcome of a hypothesis test can be affected by:

- The hypotheses you investigate:
  How do you define your null distribution?
- The test statistic you choose:
  How do you measure a difference between samples?
- The empirical distribution of the statistic under the null:
  How many times do you simulate under the null distribution?
- The data you collected:
  Did you happen to collect a sample that is similar to the population?
- The truth:
  If the alternative hypothesis is true, how extreme is the difference?

Hypothesis Test Effects

Number of simulations: Make it as large as possible so that the empirical distribution of the test statistic under the null distribution is good. No new data needs to be collected.

Number of observations: A larger sample will lead you to reject the null more reliably if the alternative is in fact true.

Difference from the null: If the null hypothesis is false, but the truth is similar to the null hypothesis, then even a large sample may not provide enough evidence to reject the null.

Tables & Arrays

Spring 2017 Midterm Q1

A table named **pay** contains one row for each UC Berkeley faculty member and these columns:

- **dept**: a string, the department of the faculty member.
- **name**: a string, the first name of the faculty member.
- **role**: a string, one of: Assistant Professor, Associate Professor, Professor, or Lecturer
- **salary**: an int, last year's salary paid by the university.

dept	name	role	salary
Journalism	Jeremy	Lecturer	111,528
Economics	Christina	Professor	349,727
South & Southeast Asian Studies	Penelope	Associate Professor	127,119

```
... (2056 rows omitted)
```

(a) (2 pt) The total salary amount paid to all faculty.

```
_______________(pay.___________________(_______________________________________))
```

Spring 2017 Midterm Q1

A table named **pay** contains one row for each UC Berkeley faculty member and these columns:

- **dept**: a string, the department of the faculty member.
- **name**: a string, the first name of the faculty member.
- **role**: a string, one of: Assistant Professor, Associate Professor, Professor, or Lecturer
- **salary**: an int, last year's salary paid by the university.

dept	**name**	**role**	**salary**
Journalism	Jeremy	Lecturer	111,528
Economics	Christina	Professor	349,727
South & Southeast Asian Studies	Penelope	Associate Professor	127,119

```
... (2056 rows omitted)
```

(b) **(3 pt)** The name of the third highest paid faculty member. (Assume no two faculty have the same salary.)

```
pay._________(_________ , _____________________________).column(_________).item(_________)
```

Spring 2017 Midterm Q1

A table named **pay** contains one row for each UC Berkeley faculty member and these columns:

- **dept**: a string, the department of the faculty member.
- **name**: a string, the first name of the faculty member.
- **role**: a string, one of: Assistant Professor, Associate Professor, Professor, or Lecturer
- **salary**: an int, last year's salary paid by the university.

dept	name	role	salary
Journalism	Jeremy	Lecturer	111,528
Economics	Christina	Professor	349,727
South & Southeast Asian Studies	Penelope	Associate Professor	127,119

```
... (2056 rows omitted)
```

(c) (3 pt) The number of lecturers in the department that has the most lecturers. (One has more than the rest.)

```
max(pay.________(________ , __________________).________(__________).column('count'))
```

Simulation

Generating Random Samples (10.4)

- From an array of values (evaluates to the sample array):
  - `np.random.choice(array, sample_size)`
- From the rows of a table (evaluates to the sample table):
  - `tbl.sample(sample_size)` (default method: with replacement)
  - `tbl.sample(sample_size, with_replacement=False)`
  - `tbl.sample(with_replacement=False)` shuffles the table by randomly permuting all the rows
- From a categorical distribution (evaluates to the sample distribution):

`sample_proportions(sample_size, population_distribution)`

Process (10.3.3)

To simulate a statistic many times:

- Define a function that returns one simulated value
- Make an empty collection array
- Run a `for` loop:
    - Each time, call the function
    - Append the newly simulated value to the collection array

Function Returns Gain on One Bet

```
def bet_on_one_roll():
    """Returns net gain on one bet"""
    # roll a die once and record the number of spots
    x = np.random.choice(np.arange(1, 7))
    if x <= 2:
        return -1
    elif x <= 4:
        return 0
    elif x <= 6:
        return 1
```

Iteration

Simulate the net gains on each of 5 bets:

```
outcomes = make_array()

for i in np.arange(5):
    outcome_of_bet = bet_on_one_roll()
    outcomes = np.append(outcomes, outcome_of_bet)
```

After the `for` loop is complete, `outcomes` will be an array of length 5.

Histograms

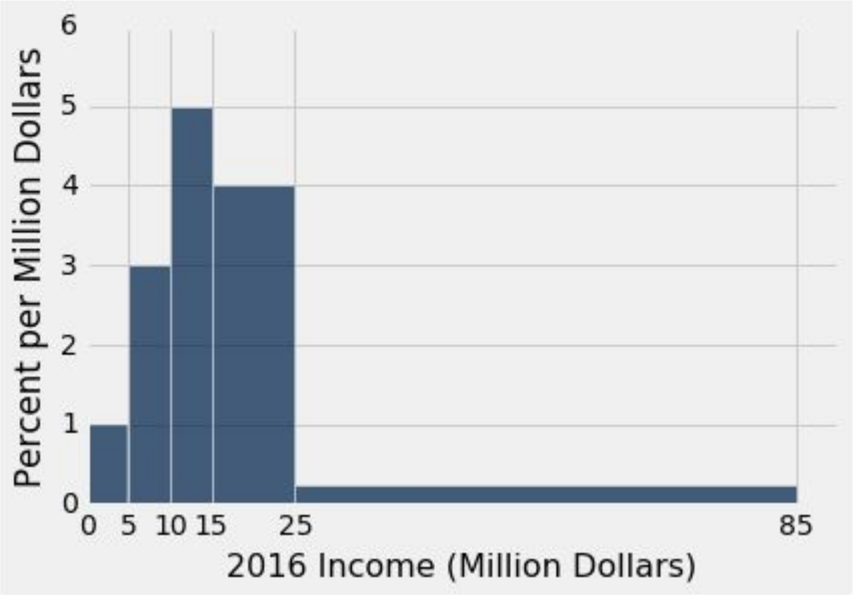
Using the Density Scale (7.2.5)

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

(b) What percent of incomes are in the [15, 25) bin?

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

(d) What percent make 75 million or more?



Answers

(a) [15, 25)

(b) 40%

(c) 4.33 percent per million dollars

(d) Unknown; maybe 0, but not more than 60 * 0.2 = 12%

Calculating Chances

Problem-Solving Method (9.5)

Most problems involve multiple trials. Here's a method that works widely.

- **Ask yourself what the first trial has to be.** If there's a clear answer (e.g. "not a six") whose probability you know, almost certainly you can continue the process with the multiplication rule.
- If there's no clear answer (e.g. "could be R, could be B, but then the next one would have to be B, or R …"), **list all the distinct ways** your event could occur and add up their chances.
- If the list above is long and complicated, **look at the complement**. If the complement is simpler (e.g. the complement of "at least one" is "none"), you can find its chance and subtract that from 1.

Exercise 1

Marbles: G, G, G, G, R, R, R, R, R, R. Draw 3 at random **with** replacement.

P(all G) = ?

P(all R) = ?

P(all same color) = ?

P(at least one G) = ?

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

P(all R) = (6/10)*(6/10)*(6/10)

P(all same color) = P(all G) + P(all R)

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

Exercise 2

Marbles: G, G, G, G, R, R, R, R, R, R. Draw 3 at random **without** replacement.

P(all G) = ?

P(all R) = ?

P(all same color) = ?

P(at least one G) = ?

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

P(all R) = (6/10)*(5/9)*(4/8)

P(all same color) = P(all G) + P(all R)

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

Testing Hypotheses

Before You Compute … (11.3)

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

Before You Compute … (11.1-3)

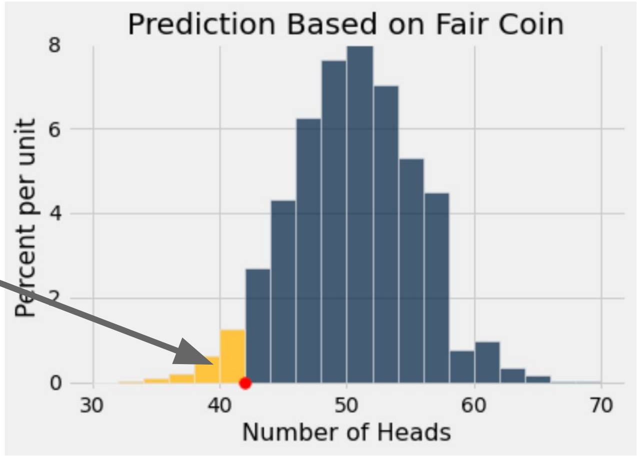
- To choose a test statistic, look at the alternative.
  - If the alternative is “the null is wrong” then use a distance
  - If the alternative specifies a direction (e.g. “too low to be due to chance), don’t use a distance
    - Instead, use a count, or average, or difference from an expected value under the null
- For the p-value: What kinds of values of the test statistic make you lean towards the alternative?
  - If the answer is “large”, the p-value is a right-hand tail
  - If the answer is “small”, the p-value is a left-hand tail

p-Value

Definition of the *p*-value (11.3)

Fair, or biased towards tails?

- The gold area approximates the p-value



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.

p-value is high → evidence of consistency with the null

p-value is low → more evidence for the alternative

Cutoff (11.3, 11.4)

- It is your threshold for deciding whether or not you think the p-value is small. Conventional values: 5%, 1%
- 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.

Factors Affecting the p-Value

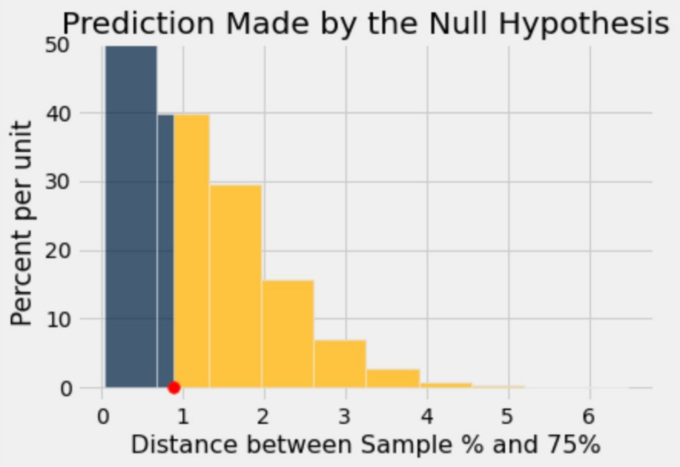
Which of the following does the p-value depend on?

- 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

Based on Tail Area

- Start at the observed value of the test statistic
- **Look in the direction that favors the alternative hypothesis**
  - If that tail is small, the data are not consistent with the null
  - Otherwise, the data are consistent with the null

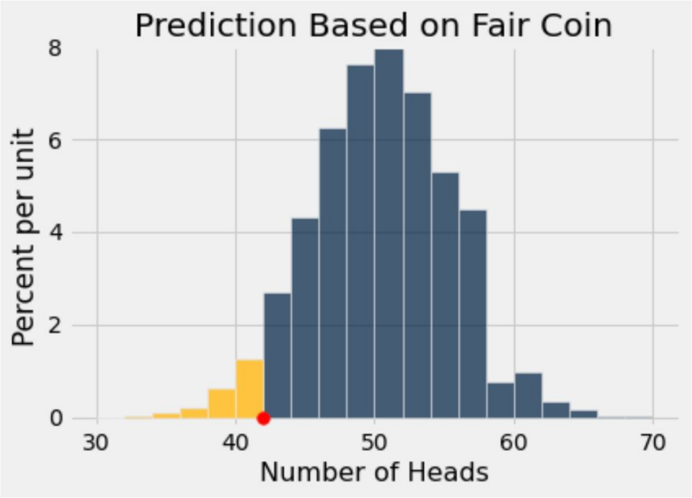


Testing whether or not Mendel's model is good:

- Large values of the distance favor the alternative
- So start at the observed distance and look to the right

Biased Towards Tails?

- **Null:** The coin is fair
- **Alternative:** The coin is biased towards tails
- Statistic: Number of heads



- Small values of the number of heads favor the alternative
- So start at the observed number of heads and look to the left

Review: A/B Tests & Causality

Randomized Controlled Experiment

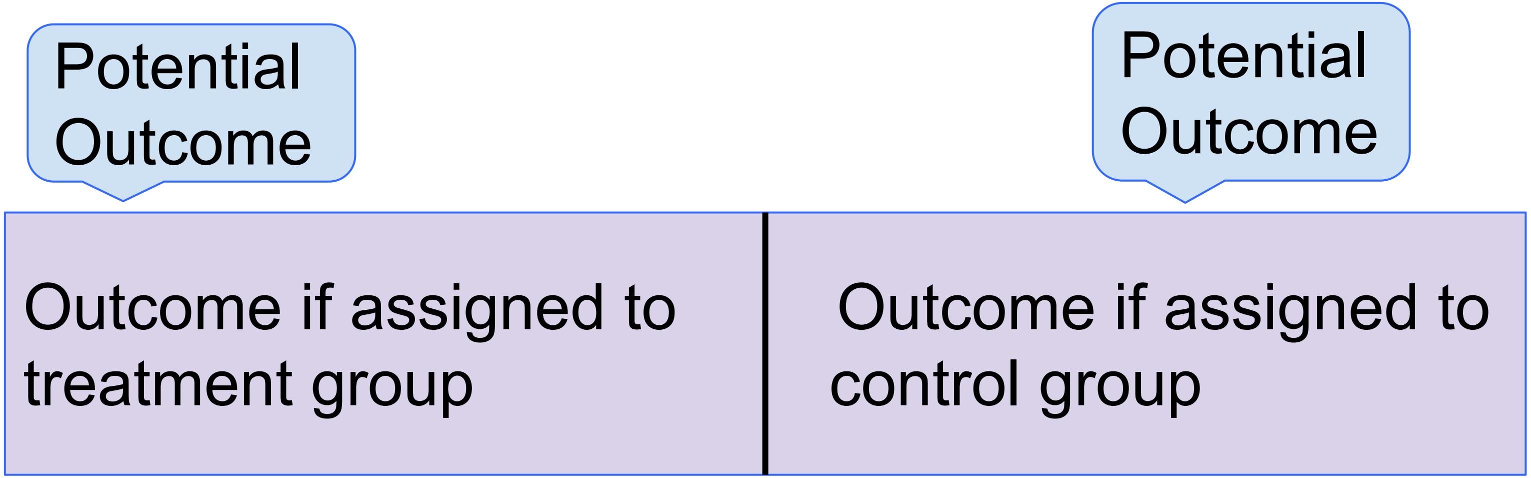
- Sample A: **control group**
- Sample B: **treatment group**

- **If the treatment and control groups are selected at random, then you can make causal conclusions.**

- Any difference in outcomes between the two groups could be due to
  - chance
  - the treatment

Before the Randomization

- In the population there is one imaginary ticket for each of the 31 participants in the experiment.
- Each participant’s ticket looks like this:



The Data

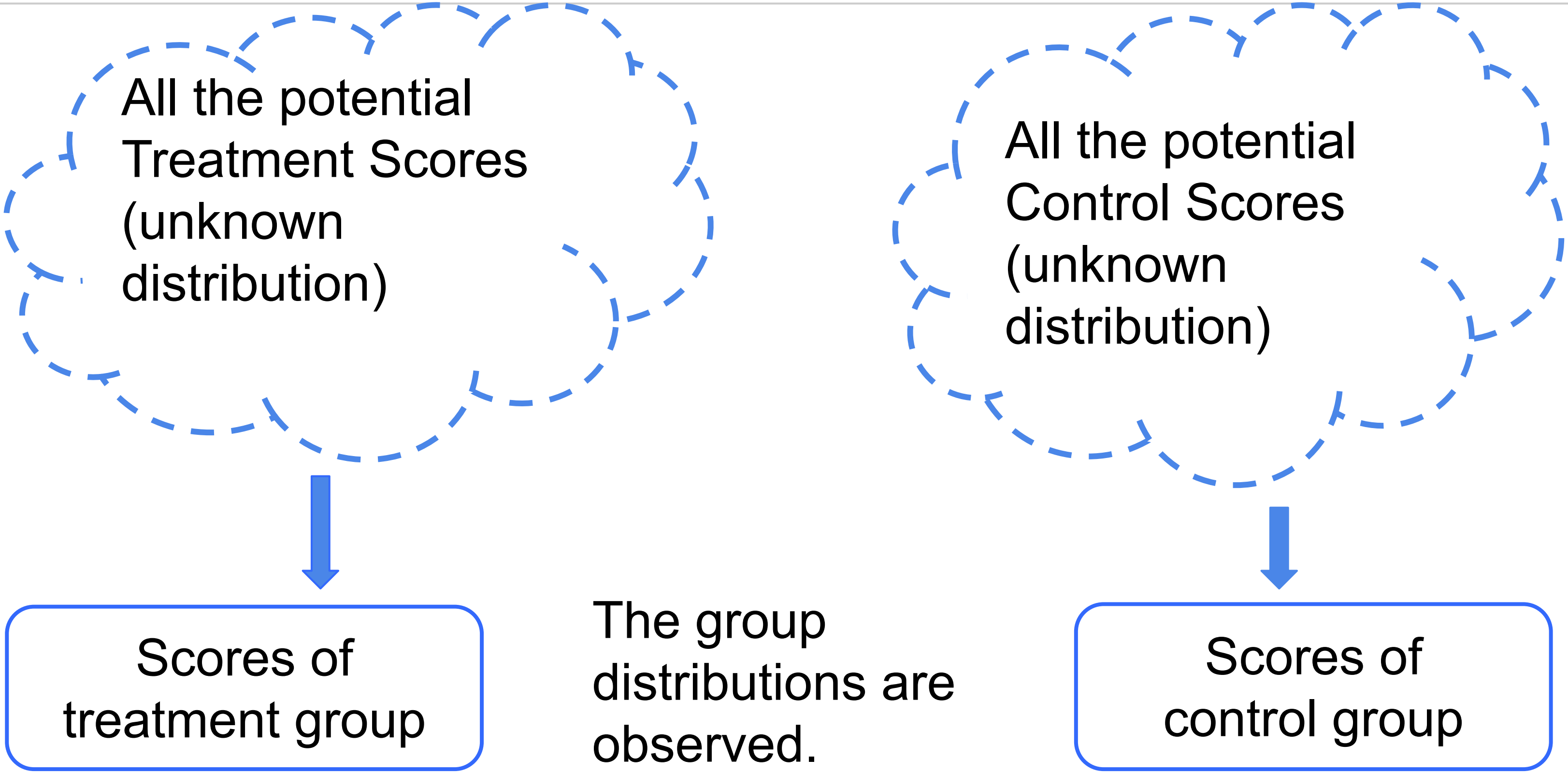
16 randomly picked tickets show:

	Outcome if assigned to control group

The remaining 15 tickets show:

Outcome if assigned to treatment group	

The Question in the RCT



Our Hypotheses

- **Null:**
  - The distribution of all 31 potential control scores is the same as the distribution of all 31 potential treatment scores.
- **Alternative:**
  - Among the 31 potential treatment scores, there is a higher percent for which the patient improves than among the 31 potential control scores.

Conclusion

- If the test rejects the null hypothesis:
  - The data favor the conclusion that the treatment helped.
  - This is a causal conclusion, due to the random assignment of patients to treatment and control.
  - But it is only a conclusion about the 31 patients in the study.
  - To make conclusions in greater generality, more and larger studies are needed.