DATA 8

Spring 2021

Discussion 7

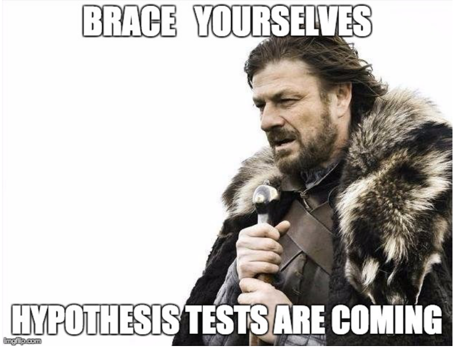
Assessing Models

Materials: tinyurl.com/d8-disc07
or access through kevin-miao.com under teaching

DIS W8AM | Kevin Miao

Today

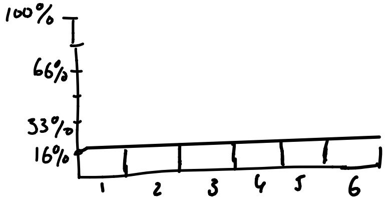
- Announcements
- Review: Testing Models
- Worksheet
    - Link: www.tinyurl.com/d8-disc07



Announcements

- **Discussion Attendance Points**
- All office hours have been converted to group settings
- Assignments:
  - **Vitamin 6** is due tonight
  - **Homework 6** is due Thursday
- **Regrades** for homework 4, vitamin 5 and lab 5 **due Friday**
  - Gradescope: Submit regrade via button
  - OkPy: Email me
- *Informal OH:* Feel free to stay after discussion until 9:30am, if you have homework/project/course related questions.

Theoretical Distribution

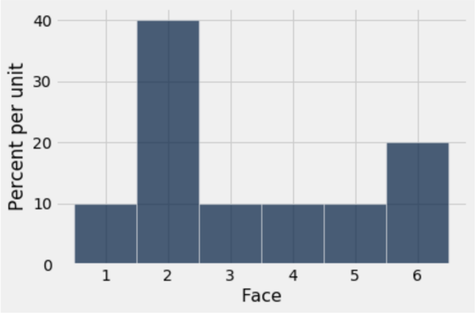


1/6 ↔

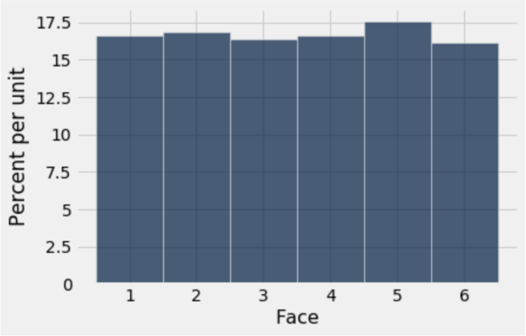
🎲 Law of Large Numbers

Empirical Distributions

actually perform the experiment and just count.



number of rolls : small



number of rolls : large

Testing Models

Scenario:
We know a fair coin has ½, ½ probability landing tails and head. Given a coin and flip it 10 times (we don’t know if it is fair or not), we see it lands heads 7 times. Is this possible with a fair coin? Is this possible with a biased coin?

Hypothesis testing allows us to simulate a fair coin and formalize this problem into a procedure.

Hypothesis Testing

- **Objective:** We want to see whether it is likely that a coin is fair, given that we know it lands heads 7 times, for instance.
- **Concretely**: We want to use simulations. We can only simulate a fair coin (i.e. we know that a fair coin lands heads/tails ½, ½ of the time).
- **Procedure**
  - Simulation Model (Null Model)
  - Alternative Model
  - Test Statistic
    - We need to (arbitrarily) create a statistic (formula) that is
      - **High** if it is close to the **alternative**
      - **Low** if it is close to the **simulation**
  - Simulate under the simulation
  - 🧑‍⚖️ Decide whether the coin is fair or biased (Later in the class)

But when do we decide the coin is fair or unfair?

To the worksheet! ✍️

tinyurl.com/d8-disc07

Vitamin Question

Question 1c

Make sure to just enter the number. Do not add in any words, whitespaces or invalid numbers since the auto-grader will mark your answer wrong and we will not allow regrades.

Data 8 Spring 2021
Discussion: Testing Models (Disc 07)

When we observe something different from what we expect in real life (i.e. four 3's in six rolls of a fair die), a natural question to ask is "Was this observed difference what we expect due to random chance? Or was it due to something other than random chance?"

Hypothesis testing allows us to answer that question in a scientific and consistent manner, using the power of computation and statistics to conduct simulations and draw conclusions from our data.

Question 1. Francie is flipping a coin. She thinks it is fair, but is not sure. She flips it 10 times, and gets heads 9 times.

She wishes to determine whether the coin was actually unfair, or whether the coin was fair and her result of 9 heads in 10 flips was by random chance.

a. What is a possible model that you can simulate under?

The coin is fair. Any deviation is due to random chance. coin lands 1/2, 1/2 heads and tails.

b. What is an alternative model for Francie's coin? You don't necessarily have to be able to simulate under this model.

The coin is biased.

c. What is a good statistic that you could simulate? Calculate that statistic for your observed data.
Hint: If the coin was unfair, it could be biased towards heads or biased towards tails.

$|\text{The number of heads} - 5| \xrightarrow[\text{in}]{\text{plug}} |9-5| = \boxed{4}$

d. Complete the function `flip_coin_10_times`, which takes no arguments and returns the absolute difference between the observed number of heads in 10 flips of a fair coin and the expected number of heads in 10 flips of a fair coin.

heads tails → arrays

```
def flip_coin_10_times():
    probabilities = make_array(0.5, 0.5)
    proportions = sample_proportions(__10, probabilities__)
    num_heads = __proportions.item(0) * 10__
    return __abs(num_heads - 5)__
```

e.g. [3/10, 7/10]
[5/10, 5/10]

e. How would you change `flip_coin_10_times` to use np.random.choice instead of sample_proportions?

↳ arrays

```
def flip_coin_10_times():
    choices = make_array("Heads", "Tails")
    flips = np.random.choice(choices, 10)
    heads = np.count_nonzero(flips == "Heads")
    return abs(heads - 5)
```

↳ $\underbrace{[H, T, H \ldots]}_{10}$

f. How would you change `flip_coin_10_times` to use .sample instead of sample_proportions?

↳ table

face
H
T

```
def flip_coin_10_times():
    table = Table().with_column("face", make_array("H","T
    flips = table.sample(10, with_replacement = True)
    num_heads = flips.where("face", "H").num_rows
    return abs(num_heads - 5)
```

faces
H
T
H

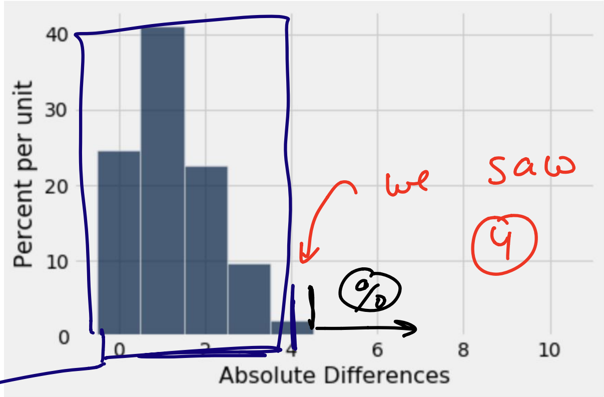
→ $\boxed{10}$

g. Complete the code below to simulate the experiment 10000 times and record the statistic in each of those trials in an array called `abs_differences`.

```
trials = 10 000
abs_differences = make_array()

for i in np.arange(trials):
    abs_diff_one_trial =
    flip_coin_10_times()
    abs_differences =
    np.append(abs_differences, abs_diff_one_trial)
```

h. Suppose we performed the simulation and plotted a histogram of `abs_differences`. The histogram is shown below.



fair

Is the observed statistic described in the question consistent with the model we simulated under?

argue either way:

fair → 4 appears on graph

unfair → 4's bar height is really short

→ more on this next time!

Question 2. As a student fed up with waiting times at office hours, you scout out the number of people in office hours (OH) from 11-12, 12-1, and 1-2 in B6 Evans. The Head GSI claims that the distribution of students is even across the three times, but you do not believe so. You observe the following data:

categorical

OH Time	Number of Students
11-12	250
12-1	300
1-2	200

Being a cunning Data 8 student, you would like to test the Head GSI's claim. Before you design your test, consider: are office hour times numerical data or categorical data?

a. What is the Head GSI's hypothesis?

no difference in which OH students prefer. The distribution is equal (1/3, 1/3, 1/3). Any random deviation is due to chance.

b. What is the student’s hypothesis?

There is a difference.

c. Which hypothesis (Head GSI or student) can you simulate under?

null hypothesis (Head GSI)

d. What is a good statistic to use? *Hint: What is a good statistic for measuring the distance between two categorical distributions?*

TVD → Categorical Distribution

End of Section
How did I do?

https://tinyurl.com/kevind8feedback