DATA 8

Fall 2020

Tutoring Section 7

Sampling and Hypothesis Testing

Logistics

- *Woohoo! Project 1 done.* 🎉
- Midterm next week! How are we feeling?
- **Next Week**: Midterm Review
  - Check your email this week so I know what you want to work on

Per usual:

- Feedback Form:
    - Form: https://tinyurl.com/feedbackD8Kevin

All resources can be found on kevin-miao.com

Today

- Sampling
    - *Population* vs *Sample*
    - `sample`
    - `np.random.choice`
  - Worksheet
- Hypothesis Testing
    - General Outline
    - `sample_proportions`
  - Worksheet

Sampling

- **Goal:** We want to know/generalize something about a **population**, i.e. "*How much time do Berkeley students work spend their homework?*"
- **Constraint:** It's extremely difficult to ask **all Berkeley students** (population)
- **Solution:** ***Sampling*** where we take samples, representative parts of the population and ask them this question *"how much time do you spend on your homework?"*
- Numbers associated with the **population** are called **parameters**
- Numbers that say something about **samples** are called **statistics**

How many people < 20 y/o live in the U.S.

Parameter : Census information

statistic : survey

Sampling *functions*

- **Sampling from a table:**

```
tbl.sample(sample_size, with_replacement=True/False)
```

- Returns a *table*
- *If you are sampling, **with_replacement = True***
- *We will discuss later when you use **with_replacement=False*** → A/B testing

- **Sampling from an array:**

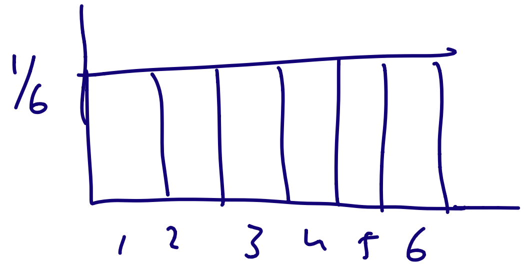
```
np.random.choice(array, sample_size)
```

- Returns an *array*

Worksheet

Link: **https://tinyurl.com/d8tutweek7**

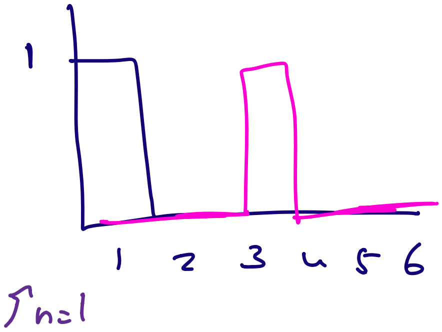
Law of Averages



Theoretical Probability

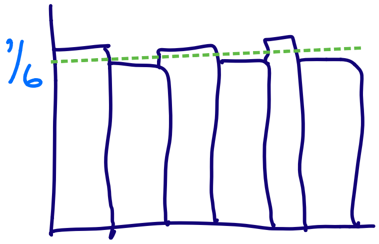
kind of the same

Emperical



n = # of rolls

n = large



law of averages

Q1a-b

Relationship between theoretical & empirical:

There's! From the empirical probability we can see what the theoretical probability is.

Law of averages If you conduct a chance experiment a large number of times, theoretical & empirical become close.

Practice Problems

1.1 Let's use the example of rolling a fair die. Remember: rolling a die is always sampling "with replacement".

a) What is the probability that you will roll a 5? Is this an empirical or a theoretical probability? Is there a relationship between the two?

1/6 chance that we roll a 5; empirical observed; Theoretical ; (real probability).

b) Complete the function `roll_die`, which takes in no arguments and uses the `dice` table to the right to roll a dice a single time and returns the value that is randomly picked.

item

Side
1
2
3
4
5
6

```
def roll_die():
    x = np.random.choice(dice.column('Side'))
    return x          ↳      ... , 1).item(0)
      - OR -
return dice.sample(1, with_replacement=True).column('Side').item(0)
```

Q1c-d

c) Simulate rolling a die 10 times and store the results in an array called `simulated_rolls`.

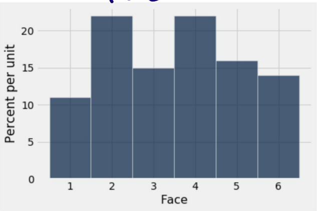
```
simulated_rolls = make_array()

for i in np.arange(10)__:
    face = roll_die()
    simulated_rolls = np.append(sim_rolls, face)
```

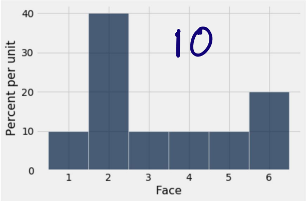
d) We've generated histograms of dice roll results for samples of size 10, 100, 1000, and 10,000 below. Which histograms correspond to which sample sizes, and why? Do these histograms show a probability or empirical distribution?

empirical b/c observing results!

100

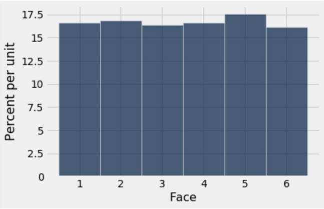


A



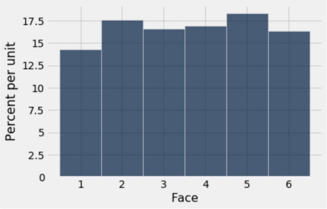
10

B



C

10 000



D

1000

Hypothesis Testing

- **Observation:** When I flip a fair coin 20 times, it is unlikely to land heads 18 times **but** it's **possible**.
- **Problem:** When we observe a certain scenario, how likely is it that that coin is fair or biased?
- **Solution:** Hypothesis testing! We basically simulate a fair **coin** 100, 1000, 10000, … times.

sample_proportions

- **Modeling proportions**

```
Sample_proportions(sample_size, model_proportions)
```

  - Returns an array of proportions the same size as *model_proportions*

- What does it do?
  - You give it an array of model proportions and a sample size. We are going to sample *sample_size* times from this model and calculate the proportions.

`sample_proportions` *example*

- Imagine we are rolling a die.
- Intuitively, we know that the chance to roll a **1, 2, 3, 4, 5, 6** is **1/6**.
- `model_proportions` = [1/6, 1/6, 1/6, 1/6, 1/6, 1/6] (this needs to add up to 1)
- The proportions add up to one, and each entry corresponds to the proportion of rolling its corresponding number of eyes.
- Imagine when we roll **1 time**, how often (in proportions) do we see **1,2,3,4,5,6** eyes.
  - ○ `sample_proportions(1, model_proportions)` → [1,0,0,0,0,0] [0,0,1,0,0,0] → etc
- What about for **100 times**?
  - ○ `sample_proportions(100, model_proportions)`

↳ [0.10, 0.20, ...]

When to use what? 🤷‍♂️

- **tbl.sample(size, with_replacement = True)**
  - Using it if we are sampling and the data is given in the form of a table.
  - *By default size is the size of the full table*
- **tbl.sample(size, with_replacement = False)**
  - This is basically shuffling the rows in a table. → A/B testing
  - Only used in A/B testing.
  - *By default size is the size of the full table*
- **sample_proportions**
  - Using if we are given an array of model_proportions that add up to 1.
- **np.random.choice**
  - If we are given an array, and we want to sample from it!

Q2.1

Practice Problems

2.1 Suppose you are flipping thumbtacks, and thumbtacks always either land pointing up or pointing down. You flip a thumbtack 60 times, and observe the thumbtack land pointing down 45 times. Your friend tells you that a thumbtack lands down with a ⅔ chance, and lands up with a ⅓ chance.

a. Does the thumbtack that you are flipping seem consistent with your friend's model? Why or why not?

b. Complete the function `flip_thumbtack`, which takes in no arguments and randomly flips a thumbtack that lands down with ⅔ probability and lands pointing up with ⅓ probability 60 times, and then returns the number of pointing down results out of 60 tosses.

```
def flip_thumbtack():
    probabilities = ___________________________________________
    proportions = sample_proportions(________________________)
    proportion_down = _________________________________________
    return ____________________________________________________
```

Q2.1

2.1 Suppose you are flipping thumbtacks, and thumbtacks always either land pointing up or pointing down. You flip a thumbtack 60 times, and observe the thumbtack land pointing down 45 times. Your friend tells you that a thumbtack lands down with a ⅔ chance, and lands up with a ⅓ chance.

a. Does the thumbtack that you are flipping seem consistent with your friend's model? Why or why not?

no, 75% vs 67% → not the same.
↳ It's possible || some people might say it's fair!

b. Complete the function `flip_thumbtack`, which takes in no arguments and randomly flips a thumbtack that lands down with ⅔ probability and lands pointing up with ⅓ probability 60 times, and then returns the number of pointing down results out of 60 tosses.

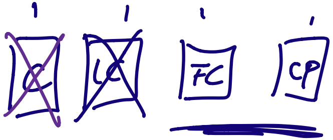
```
def flip_thumbtack():
    probabilities = make_array(2/3, 1/3)
    proportions = sample_proportions(60, probabilities)
    proportion_down = empirical_proportions.item(0)
    return proportion_down * 60
```

EMP down ↑ EMP up ↑

~> [0.60, 0.40]

Q2.2

2.2 Suppose you want to leave your breakfast choices up to chance! You have a cabinet of 4 different cereal brands: Cheerios, Lucky Charms, Fruit Loops, and Cocoa Puffs. Suppose you randomly pick 4 cereal boxes *with replacement*.



a. What is the probability that you pick four unique brands of cereal?

$$1 \times \frac{3}{4} \times \frac{2}{4} \times \frac{1}{4}$$

b. What is the probability that you don't pick Cheerios? in 4 times !

$$\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} \times \frac{3}{4} = \left(\frac{3}{4}\right)^4$$

Q2.3

Not true lol

2.3 In the Netherlands, all men take a military preinduction exam at age 18. The exam includes an intelligence test known as "Raven's progressive matrices" and includes questions about demographic variables like family size. A study was done in 1968, relating the test scores of 18-year-old men to the number of their brothers and sisters. The records of all exams taken in 1968 were used.[1]

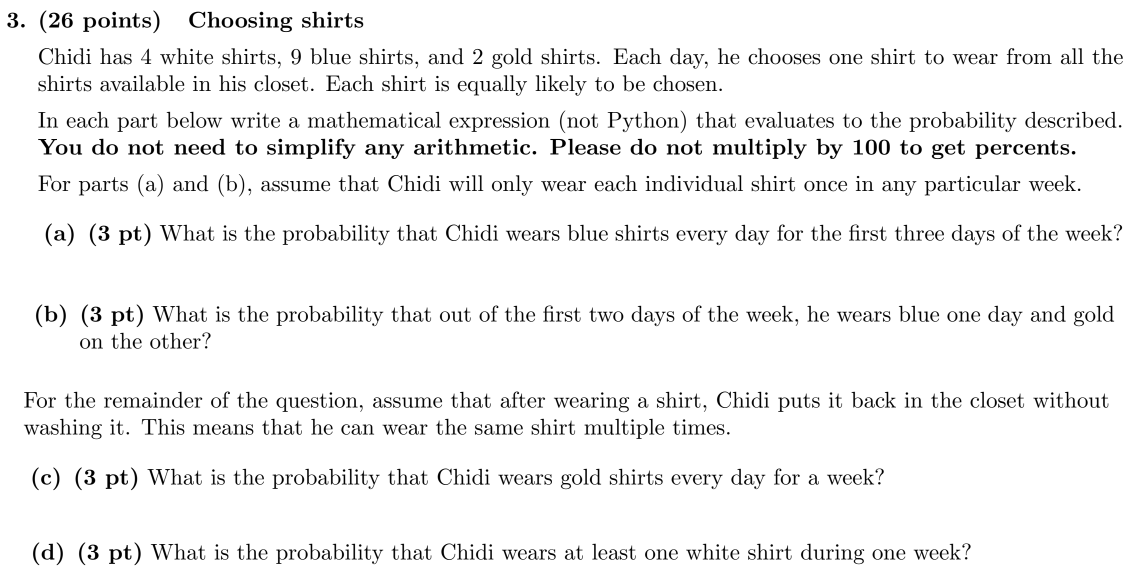
a) What is the population of the study? What is the sample used in the study?

population: all men, 18 y/olds in the Netherlands

sample: 18 year old men

b) Is there a need to apply inference techniques to predict the mean score, max score, etc? Why or why not? Are those values parameters or statistics?

Old exam question (FA19 MT1)



3. (26 points) Choosing shirts

Chidi has 4 white shirts, 9 blue shirts, and 2 gold shirts. Each day, he chooses one shirt to wear from all the shirts available in his closet. Each shirt is equally likely to be chosen.

In each part below write a mathematical expression (not Python) that evaluates to the probability described. **You do not need to simplify any arithmetic. Please do not multiply by 100 to get percents.**

For parts (a) and (b), assume that Chidi will only wear each individual shirt once in any particular week.

(a) (3 pt) What is the probability that Chidi wears blue shirts every day for the first three days of the week?

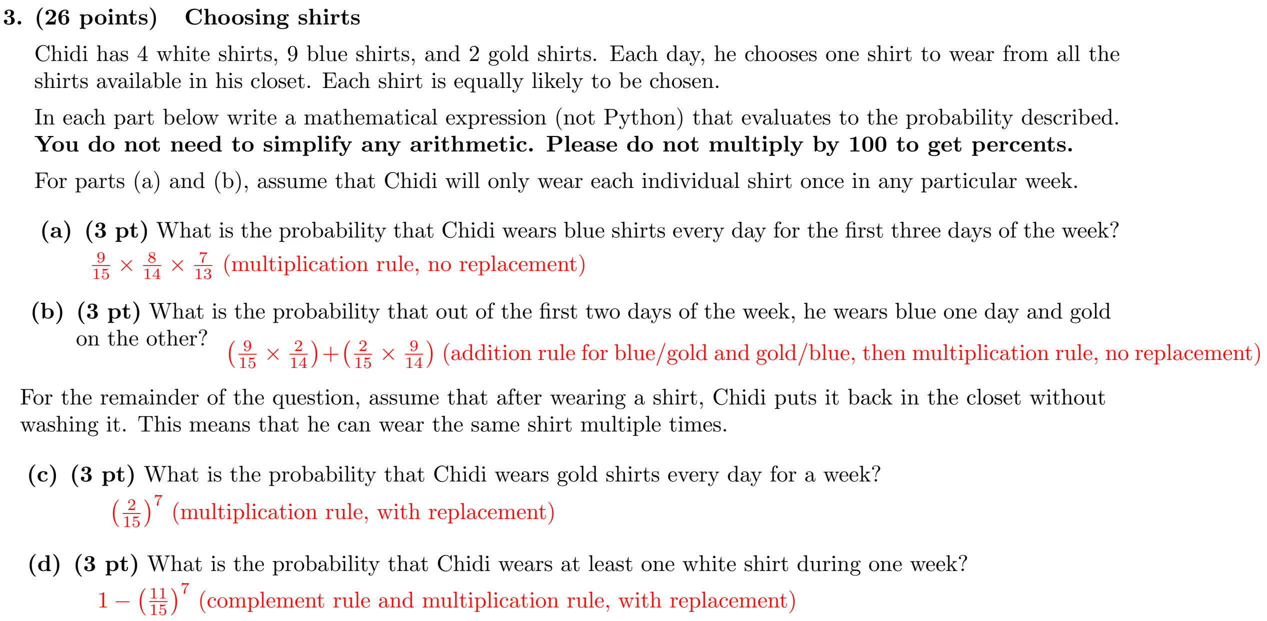
(b) (3 pt) What is the probability that out of the first two days of the week, he wears blue one day and gold on the other?

For the remainder of the question, assume that after wearing a shirt, Chidi puts it back in the closet without washing it. This means that he can wear the same shirt multiple times.

(c) (3 pt) What is the probability that Chidi wears gold shirts every day for a week?

(d) (3 pt) What is the probability that Chidi wears at least one white shirt during one week?

Old exam question (FA19 MT1)



3. (26 points) Choosing shirts

Chidi has 4 white shirts, 9 blue shirts, and 2 gold shirts. Each day, he chooses one shirt to wear from all the shirts available in his closet. Each shirt is equally likely to be chosen.

In each part below write a mathematical expression (not Python) that evaluates to the probability described. **You do not need to simplify any arithmetic. Please do not multiply by 100 to get percents.**

For parts (a) and (b), assume that Chidi will only wear each individual shirt once in any particular week.

(a) (3 pt) What is the probability that Chidi wears blue shirts every day for the first three days of the week?

$\frac{9}{15} \times \frac{8}{14} \times \frac{7}{13}$ (multiplication rule, no replacement)

(b) (3 pt) What is the probability that out of the first two days of the week, he wears blue one day and gold on the other?

$(\frac{9}{15} \times \frac{2}{14}) + (\frac{2}{15} \times \frac{9}{14})$ (addition rule for blue/gold and gold/blue, then multiplication rule, no replacement)

For the remainder of the question, assume that after wearing a shirt, Chidi puts it back in the closet without washing it. This means that he can wear the same shirt multiple times.

(c) (3 pt) What is the probability that Chidi wears gold shirts every day for a week?

$(\frac{2}{15})^7$ (multiplication rule, with replacement)

(d) (3 pt) What is the probability that Chidi wears at least one white shirt during one week?

$1 - (\frac{11}{15})^7$ (complement rule and multiplication rule, with replacement)

End of Section

- Please complete the anonymous Feedback form so I can improve my teaching:
  - **https://tinyurl.com/feedbackD8Kevin**

- Solutions and notes will be posted as soon as possible.

- Email me if you have any questions: kevinmiao@berkeley.edu

- **No tutor OH next week ☹**