Data 8 Spring 2022

Lab: Midterm Review

Tables

You are given the following table called `pokemon`. For the following questions, fill in the blanks.

Name	Type	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
Bulbasaur	Grass	318	45	49	49	65	65	45	1	False
Ivysaur	Grass	405	60	62	63	80	80	60	1	False
Venusaur	Grass	525	80	82	83	100	100	80	1	False
VenusaurMega Venusaur	Grass	625	80	100	123	122	120	80	1	False
Charmander	Fire	309	39	52	43	60	50	65	1	False
Charmeleon	Fire	405	58	64	58	80	65	80	1	False
Charizard	Fire	534	78	84	78	109	85	100	1	False
CharizardMega Charizard X	Fire	634	78	130	111	130	85	100	1	False
CharizardMega Charizard Y	Fire	634	78	104	78	159	115	100	1	False
Squirtle	Water	314	44	48	65	50	64	43	1	False

... (790 rows omitted)

1. Find the name of the pokemon of type `Water` that has the highest HP.

```
water_pokemon = pokemon._____________(___________,____________________)

water_pokemon._________(___________,_____________).column("Name").item(0)
```

```
water_pokemon = pokemon.where("Type", are.equal_to("Water"))
water_pokemon.sort("HP",descending=True) .column("Name").item(0)
```

2. Find the proportion of pokemon of type `Fire` in the dataset whose Speed is strictly less than 100.

```
fire_pokemon = pokemon._____________(_________________,_________________)
fire_pokemon._______(________,_____________)._________/_________________
```

```
fire_pokemon = pokemon.where("Type", "Fire")
fire_pokemon.where("Speed",
are.below(100)).num_rows/fire_pokemon.num_rows
```

3. Create a table containing `Type` and `Generation` that is sorted in decreasing order by the average `HP` for each pair of `Type` and `Generation`.

```
d = pokemon._____________(_____________________________,_________________)

d.sort("HP mean",_____________________)._________(_________,_____________)
```

```
d = pokemon.group(make_array("Type", "Generation"), np.mean)

d.sort("HP mean", descending=True).select("Type", "Generation")
```

or

```
D = pokemon.group(make_array("Type", "Generation"), np.mean)

d.sort("HP mean", descending=True).select("Type", "Generation")
```

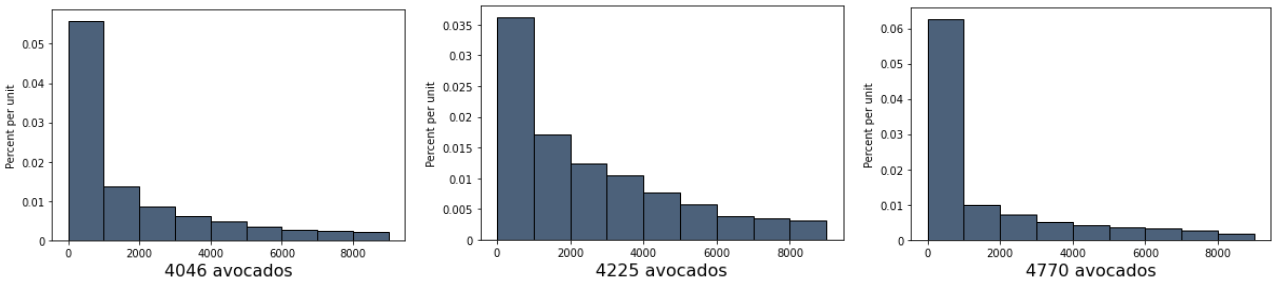
4. Return an array that contains ratios of legendary to non-legendary pokemons for each generation.

```
t = pokemon._____________(___________________,_____________________)

ratio = t._____________(_________________)/t._____________(_____________)
```

```
t = pokemon.pivot("Legendary", "Generation")

ratio = t.column("True")/t.column("False")
```

Histograms

Everyone knows Zoomers love their avocado toast, so it comes as no surprise that this Hass Avocado dataset was a popular selection among your peers! Each type of avocado (PLU 4046, PLU 4225, PLU 4770) has a corresponding histogram below; each data point in the histogram represents the number of avocados sold in one order. All bars are **1000 units wide**, but take note: the **density scale is different in each histogram**. There are **16812 orders shown in the PLU 4046 histogram** and **19572 orders shown in the bay PLU 4225** histogram.



1. Calculate each quantity described below or write *Unknown* if there is not enough information above to express the quantity as a single number (not a range). Show your work!
   a. The percentage of PLU 4225 orders sold of less than 1000 avocados is equal to the percentage of PLU 4046 orders sold of less than 1000 avocados.
   False. As noted, the y-axis for the two histograms is different.

   b. The number of PLU 4046 orders containing at least 2500 avocados but less than 3000 avocados.
   Unknown: We can't tell how heights are distributed within a bin.

   c. The number of PLU 4770 orders containing at least 1000 avocados but less than 2000 avocados.
   Unknown: The total number of 4770 orders is unknown, so the size of any subset is unknown.

   d. The number of PLU 4225 orders that contained less than 8000 avocados
   (100 percent - (1000 orders * 0.004 percent/order)) * 19572 = 18789 orders

2. If the PLU 4225 histogram were redrawn, replacing the three bins from 0-1000, 1000-2000, 2000-3000 with one bin from 0-to-3000, what would be the height of its bar?
   The bin contains 0.036 * 1000 + 0.0175 * 1000 + 0.0125 * 1000 = 66 percent, and the width is 3000, so the height is 0.02 percent/unit.

Probability

1. A fair coin is tossed five times. Two possible sequences of results are HTHTH and HTHHH. Which sequence of results is more likely? Explain your answer and calculate the probability that each sequence appears.

They are equally likely since the coin is fair. By the multiplication rule, the probability that either of the two sequences appears is (1/2)**5.

2. For questions 2 - 4, assume we have a biased coin such that the probability of getting heads is ⅕ and the probability of getting tails is 4/5. The coin is tossed 3 times. What is the probability that you get exactly 2 heads?

Here we need to enumerate all the outcomes that fall into this event and count their probabilities. The outcomes are HHT, HTH, THH. There is a total of 3 events. The probability for each of them is (⅕)^2 * (⅘)^1. So the final result then will be 3 * (⅕)^2 * (⅘)^1.

3. Once again, we toss the same biased coin 3 times. What is the probability I get no heads?
(⅘)^3

4. Again, we toss the same biased coin 3 times. What is the probability I get at least 1 heads?
Hint: There are two ways of calculating this probability. One is significantly easier to calculate than the other.
1 - (⅘)^3

Simulation and Hypothesis Testing

Achilles the turtle sits on the number line. Achilles loves long random walks that last a total of 100 times steps. At each time step, Achilles moves based on the following scheme: He flips a coin and moves one step to the right if the coin comes up heads or one step to the left if the coin comes up tails.

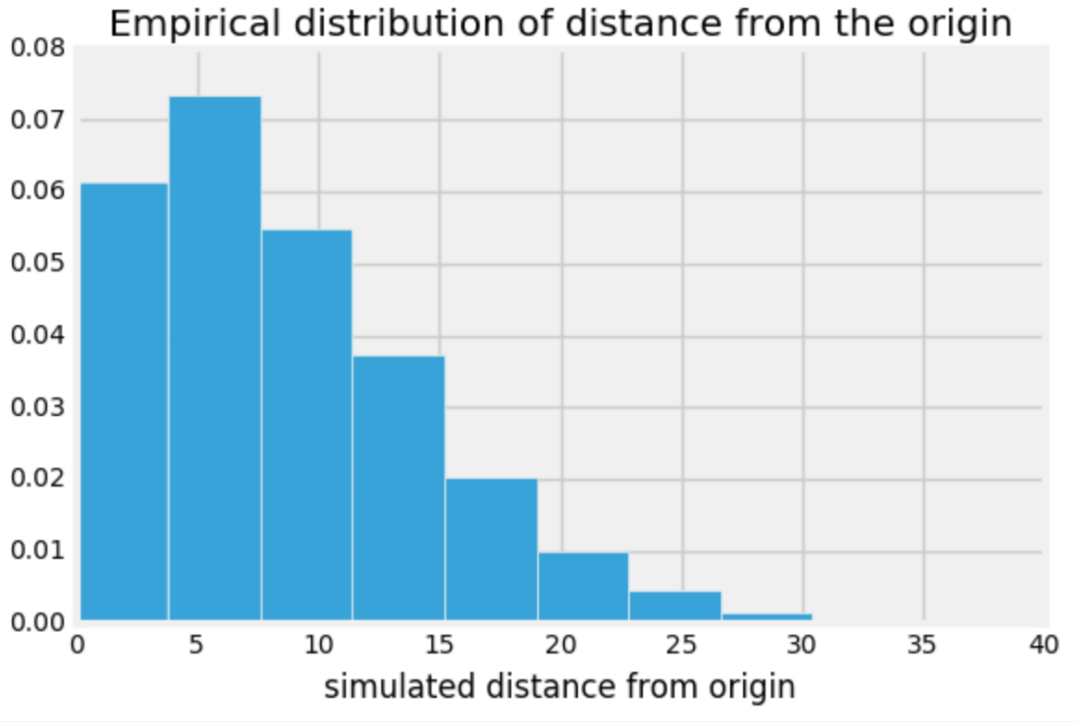
1. Assuming that Achilles' coin is fair, write a function called `one_walk` that simulates one random walk of 100 time steps and returns
how far from the origin Achilles ends up at the end of his walk. You may assume that Achilles always starts from the origin.

```
def one_walk():
      dist = 0
      for i in np.arange(100):
            move = np.random.choice(make_array(1, -1))
            dist = dist + move
      return abs(dist)

OR

      return abs(sum(np.random.choice(make_array(1,-1), 100)))
```

2. Assuming that Achilles' coin is fair, we would like to simulate what would happen if Achilles took 10000 different random walks. Complete the simulation below and keep track of how far Achilles ends up from the origin in each of his walks in an array called `distances`. The histogram shown below is an example of a histogram plotted from `distances`.

```
distances = make_array()

for i in np.arange(10000):
    new_distance = ____________________________________________________
    distances = ____________________________________________________
```

```
for i in np.arange(10000):
    new_distance = one_walk()
    distances = np.append(distances, new_distance)
```



3. Achilles goes for a walk and claims that at the end of his walk, he ended up 30 steps away from the origin. You notice this is strange, so you want to run a hypothesis test to test whether or not Achilles used a fair coin. Fill in the blanks below for the null and alternative hypotheses and test statistic.

Hint: When considering your alternative hypothesis, note that we do not really care about whether the coin is biased towards heads or towards tails.

Null Hypothesis:
The coin Achilles uses his walk is fair. The fact that he ended up so far away from the origin is merely due to chance.

Alternative Hypothesis:
Achilles' coin was unfair.

Test Statistic:
Absolute difference from the origin at the end of a walk. Can be thought of as how far he walked in either direction.

4. Write the code to calculate the p-value given the test statistic listed above and using a 5% p-value cut-off. Then, describe the different conclusions that you would arrive at depending on the p-value.

Hint: We simulated an array in part(b) of test statistics under the null hypothesis. Try to use the `distances` *array.*

```
p_value = _____________________________________________________________________
```

```
p_value =  np.count_nonzero(distances >= 30) / len(distances)
```

If the p value is less than the 5% cutoff, we would consider this to be evidence against the null hypothesis while if it is above 5% we would say that there is not strong evidence against the null.

True/False

Respond with true or false to the following questions. If your answer is false, explain why.

1. In the U.S. in 2000, there were 2.4 million deaths from all causes, compared to 1.9 million in 1970, which represents a 25% increase. The data shows that the public's health got worse over the period 1970-2000.

False, because the population also got bigger between 1970 and 2000. It would be more appropriate to look at the total number of deaths compared to the total population at each year. In fact, the U.S. population in 1970 was 203 million, while in 2000 it was 281 million.

2. A company is interested in knowing whether women are paid less than men in their organization. They share *all* their salary data with you. An A/B test is the best way to examine the hypothesis that all employees in the company are paid equally.

False, there is no room for statistical inference here. We have access to the whole population so the answer can simply be retrieved by directly looking at the data. No need for an A/B test here.

3. Consider a randomized control trial where participants are randomly split into treatment and control groups. There will be no systematic differences between the treatment and control groups if the process is followed correctly.

False, randomization can give rise to significantly different treatment and control groups merely by chance, meaning there is still the possibility for systematic differences between the treatment and control groups.

4. A researcher considers the following scheme for splitting a people into control and treatment groups. People are arranged in a line and for each person, a fair, six-sided die is rolled. If the die comes up to be a 1 or a 2, the person is allocated to the treatment group. If the die comes up to be a 3, 4, 5 or 6 then the person is allocated to the control group. This is a randomized control experiment.

True, because the participants were randomly assigned to each group through the roll of a die. This makes it a randomized controlled experiment!

5. You are conducting a hypothesis test to check whether a coin is fair. After you calculate your observed test statistic, you see that its p-value is below the 5% cutoff. At this point, you can claim with certainty that the null hypothesis can not be true.

False, remember the definition of a p-value: A p-value expresses the probability, under the null Hypothesis, that you observe a value for your test statistic that is at least as extreme as your observed test statistic in the direction of the alternative. Assuming that this probability is non 0 then we can not claim that the null can never be true. It could be the case that we simply got an unusual sample from our null.

6. You roll a fair die a large number of times. While you are doing that, you observe the frequencies with which each face appears and you make the following statement: As I increase the number of times I roll the die, the probability histogram of the observed frequencies converges to the empirical histogram.

False, the statement should be: As I increase the number of times I roll the die, the **empirical histogram** of the observed frequencies converges to the **probability histogram** of a fair die.

Multiple Choice

1. Gary is playing with a coin and he wants to test whether his coin is fair. His experiment is to toss the coin 100 times. He chooses the following Null Hypothesis:

> **Null Hypothesis:** The coin is fair and any deviation observed is due to chance.

For each of the alternative hypotheses listed below, determine whether or not the test statistic is valid.

a. **Alternative Hypothesis:** The coin is biased towards heads.
Test Statistic: # of heads
Correct

b. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: # of heads
Incorrect, we want large values of our test statistic to favor the alternative hypothesis. In this case, the coin could be unfair in either direction, so we are not accounting for the case when it is biased towards tails

c. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: |# of heads - expected # of heads|
Correct

d. **Alternative Hypothesis:** The coin is biased towards heads.
Test Statistic: |# of heads - expected # of heads|
Incorrect, this is the opposite case of part (b). We see that this test statistic will also account for a bias towards tails (because of the absolute value)

e. **Alternative Hypothesis:** The coin is not fair.
Test Statistic: ½ - proportion of heads
Incorrect, without the absolute value, we will not achieve large values of our test statistic leaning towards the alternative hypothesis

Fun with Functions

1. Write a function called `compute_pvalue` that, given an empirical distribution in the form of an array and the observed value of your test statistic, calculates the p-value for that test statistic. You may assume that large values of your test statistic provide evidence against the null hypothesis.

```
def compute_pvalue(empirical_dist, observed_ts):
      return np.count_nonzero(empirical_dist >=
observed_ts)/len(empirical_dist)
```

2. Now write a function called `is_significant` that takes in an empirical distribution, the observed test statistic and a p-value cutoff, returns `True` if the p-value of the observed test statistic is statistically significant based on the cutoff provided and `False` otherwise.

Hint: Use the function you defined in Question 1!

```
def is_significant(empirical_dist, observed_ts, cutoff):
      observed_pval = compute_pvalue(empirical_dist, observed_ts)
      return observed_pval <= cutoff
```

More Hypothesis Testing

Chloe is a big fan of Trader Joes' frozen mac n cheese, but she noticed that the cheese used in it varies from box to box. A Trader Joe's employee provides her with some data about the 4 different cheeses used and the probability of them being used in each box:

Cheese	**Probability**
Velveeta	0.05
Gruyère	0.55
Sharp Cheddar	0.25
Monterey Jack	0.15

Chloe is suspicious about this distribution. After all, Velveeta is much cheaper to use than Gruyère, and she has also never bought a box that uses Gruyère. Chloe decides to buy many boxes throughout the next month and tracks the type of cheese used in each box. She uses this to conduct a hypothesis test.

1. Write the correct null hypothesis for this experiment
   - Null Hypothesis: The types of cheese in the frozen Mac n Cheese boxes are distributed according to the probability distribution provided by the employee.
   - Alternative Hypothesis: The types of cheese in the frozen mac n cheese boxes are not distributed according to the probability distribution provided by the employee.

```
observed_proportions = make_array(0.2, 0.3, 0.45, 0.05)
employee_proportions = make_array(0.05, 0.55, 0.25, 0.15)
```

The array `observed_proportions` contains the proportions of cheese that Chloe observed in 20 boxes of Mac n Cheese.

2. Chloe wants to use the mean as a test statistic, but Katherine suggests that she uses the TVD (total variation distance) instead. Which test statistic should Chloe use in this case? Briefly justify your answer. Then write a line of code to assign the observed value of the test statistic to `observed_stat`.
Katherine is correct, we should use the total variation distance, because she's comparing two categorical distributions (the observed distribution and the one provided by Trader Joes)

```
observed_stat = sum(abs(observed_proportions - employee_proportions)) / 2
```

3. Define the function `one_simulated_test_stat` to simulate a random sample according to the null hypothesis and return the test statistic for that sample.

```
def one_simulated_test_stat():
      sample_prop = sample_proportions(20, employee_proportions)
      return sum(abs(employee_proportions - sample_prop)) / 2
```

4. Chloe simulates the test statistic 10,000 times and stores the results in an array called simulated_stats. The observed value of the test statistic is stored in `observed_stat`. Complete the code below so that it evaluates to the p-value of the test:

```
np.count_nonzero(simulated_stats >= observed_statistic) / 10000
```

5. Given that the computed p-value is 0.0825, which of the following are true? Select all that may apply.

a. Using an 8% p-value cutoff, the null hypothesis should be rejected
b. Using a 10% p-value cutoff, the null hypothesis should be rejected.
c. There is an 8.25% chance that the null hypothesis is true
d. There is an 8.25% chance that the alternative hypothesis is true

A/B Testing

1. Choose True/False for each of the statements below, and explain your answer.

a) A/B testing is used to determine whether or not we believe two samples come from the same underlying distribution.
True, this is the definition of A/B testing.

b) To conduct a permutation test, you should sample your data with replacement with a sample size equal to the number of rows in the original table.
False, you should sample your data *without* replacement--otherwise, you would not get a permutation of your data.

c) A/B testing is the same as using total variation distance as a test statistic for a hypothesis test.
False, total variation distance is just a test statistic which computes the distance between two distributions. It does not involve taking a random permutation of your data.

2. You and your friend are both huge Warriors fans, and are watching a game together. You're enjoying the absolute dominance the Warriors are displaying as a team. You attribute the Warriors' recent success to Stephen Curry's shooting skills, but your friend argues that Klay Thompson has

actually been the better shooter as of late. Because you are a data scientist who loves Stephen Curry, you decide to test your friend's claim.

You decide to compare the two players by taking a sample of games from this year and finding the field goal percentages (the percentage of shots a player makes) for both Curry and Thompson. You record these in a table, and at a glance, it seems that Curry has better field goal percentages. You decide to run an A/B test to determine if this is just due to chance.

a) Choose a null hypothesis for your A/B test.
   The distribution of field goal percentages for Stephen Curry and Klay Thompson are the same throughout the year. The difference we see in our sample is just due to random chance.

b) Choose an alternative hypothesis.
   Stephen Curry is making more shots than Klay Thompson throughout the year; the difference we view in our sample is not just due to chance.

c) Choose a test statistic.
   A natural test statistic would be the difference between the average field goal percentage for Curry and the average field goal percentage for Thompson.

d) Describe how you would use a permutation test to simulate your test statistic and calculate a p-value.

   First, calculate the observed value of the test statistic. Then, begin a simulation: create an empty array for simulated values of the test statistic, and initialize a for loop. In each iteration, shuffle the field goal percentages. Then, calculate the test statistic for the new shuffled groups. Store this value in your empty array, and repeat to get an empirical distribution. Larger values of the test statistic support the alternative hypothesis in this case, so calculate a p-value by finding the proportion of values in your empirical distribution which are larger than or equal to the observed value of the test statistic.