Data 8 **Name:** _____
Summer 2018
Worksheet 3
7/21/2018
Conceptual Office Hours

This worksheet can serve as a general overview of materials you have learned this week and give you an opportunity to practice solving them in an exam-like format. NOTE: This worksheet is a collection of problems that come from resources all students have access to, resources only staff members of previous iterations of the course have access to, and some that I have made up myself. This is not necessarily representative of what will be tested on the actual exams. For any issues or questions, contact Robert Sweeney Blanco at robertsweeney-blanco@berkeley.edu.

For all problems, you may assume you have imported datascience and numpy as np.

You're at The Wizarding World of Harry Potter at Universal Studios Hollywood where a wizard sells you an enchanted coin that has a 40% chance of landing heads. When you get home, you flip the coin 100 times and get 50 heads. You begin to think the wizard is a fraud. Design a hypothesis test to determine whether it is believable that the coin is enchanted. (Some of these questions have multiple possible answers)

1. What is your null hypothesis?

   The coin has a 40% chance of landing heads and any difference from this is due to random chance.

2. What is your alternative hypothesis?

   The difference is not due to chance, there is some bias in favor of heads.

3. What is your test statistic?

   The number of heads.

4. Fill in the code to run the simulation. Note: The answers to these depend on your answers above!

```
proportions = <1>
collection = <2>
for i in np.arange(1000):
    sample = 100 * sample_proportions(<3>, proportions).item(<4>)
    collection = <5>
```

4.1) make_array(.4,.6)

4.2) make_array()

4.3) 100

4.4) 0

4.5) np.append(collection, sample)

5. Assuming you have access to *collection* from your code above. Write code to compute the p-value.

   np.count_nonzero(collection >= 50) / len(collection)

6. Pretend the line above evaluated to .024. Using the standard p-value cutoff of 0.05, is this result statistically significant or not?

   Yes, it is below our p-value cutoff

7. What is your conclusion with regards to whether the coin is enchanted or not?

   The null hypothesis is rejected in favor of the alternative because the original sample statistic is too unreasonable to occur under the null.

Suppose you work in marketing for movie theaters and decide to test whether advertising popcorn before a movie will increase the number of people that will get up and buy popcorn before the movie begins. There are two theaters (Theater 1 and Theater 2), each with 100 people inside. Theater 1 plays an advertisement for popcorn before the movie, Theater 2 does not. Suppose 29 people from Theater 1 get up and buy popcorn whereas only 22 people from Theater 2 get up. The results are recorded in the table *popcorn* shown below. Each person in the theater is assigned an ID, shown in the column *ID*.

ID	Theater	Got up
0	1	False
1	1	False
2	1	True
3	1	False
4	1	True
5	1	False
6	1	False
7	1	True
8	1	False
9	1	False

... (190 rows omitted)

8. What kind of testing methodology is required here?

   Permutation test

9. What is your null hypothesis?

   More people bought popcorn in Theater 1 due to random chance.

10. What is your alternative hypothesis?

    Theater 1 having more people buy popcorn was not due to random chance.

11. What is your test statistic?

The number of people who got popcorn from Theater 1 minus the number of people who got popcorn from Theater 2

12. Fill in the code below

```
differences = <1>
repetitions = 5000
for i in np.arange(repetitions):
    shuffled = popcorn.sample(<2>).column(2)
    with_shuffled = popcorn.with_column('Shuffled', <3>)
    shuffled_results = with_shuffled.group('Theater', <4>).column(3)
    simulated_stat = <5>
    differences = np.append(<6>, <7>)
```

12.1) make_array()

12.2) with_replacement = False

12.3) shuffled

12.4) np.count_nonzero

12.5) shuffled_results.item(0) - shuffled_results.item(1)

12.6) differences

12.7) simulated_stat

13. Assuming you have access to *collection* from your code above. Write code to compute the p-value.

    np.count_nonzero(differences >= 7) / 5000

14. Pretend the line above evaluated to .125. Using the standard p-value cutoff of 0.05, is this result statistically significant or not?

    It is not statistically significant.

15. What is your conclusion with regards to whether the marketing strategy is effective or not?

    We fail to reject the null hypothesis. It is reasonable that 7 more people got popcorn after the ad under the null hypothesis.

France has been trying to improve its consumption of clean energy. As a data scientist, you want to know if France's efforts have paid off, or if their energy consumption is like the rest of Europe's and variation is due to random chance. You have access to the table below, called `energy`. Additionally, you know that France consumed approximately 2,826,000 kilowatt-hours of energy.

Energy Source	Europe Average	France
Oil	0.34	0.28
Gas	0.20	0.16
Coal	0.16	0.04
Renewables	0.13	0.12
Nuclear	0.13	0.40

16. What would be your Null hypothesis?

France's energy use is no different from the rest of Europe's, any variation from the European energy consumption levels is due to random chance.

17. What would be your alternative hypothesis?

France's energy consumption levels are different from the rest of Europe's, and there are other factors contributing to this besides "noise in the data."

18. What statistic would you use to test the hypothesis? Why?

Total Variation Distance. We want to find out if the distribution of energy use in France across the set of sources is any different from a random draw from the European distribution.

19. Fill in the function to compute the test statistic. Call it on `energy`. Use as many arguments as you need, and fill in each line corresponding to the numbers you see in the code.

```
def test_stat(dist1, dist2 ):
    return <1>(np.abs(<2>)) / <3>

observed_value = test_stat(<4>, <5>)
```

19.1) sum

19.2) dist1-dist2

19.3) 2

19.4) energy.column(1)

19.5) energy.column(2)

20. Write code to help you simulate one value of the test statistic.

```
total_energy = 2826000
random_distribution = <1>(<2>, <3>)
simulated_value = <4>(<5>, <6>)
```

20.1) sample_proportions

20.2) total_energy

20.3) energy.column(1)

20.4) test_stat

20.5). energy.column(2)

20.6) random_distribution (5 and 6 can be any order)