DS 100
Spring 2025

Intro to Data
Science

Practice Midterm

- You have approximately 80 minutes.
- This is a long exam, with many chances to demonstrate understanding. We encourage you to try the problems in the order that makes most sense to you, and to keep moving if you get stuck on one question.
- The exam is closed book, closed notes.
- Mark your answers ON THE EXAM ITSELF. If you are not sure of your answer you may wish to provide a *brief* explanation. All short answer sections can be successfully answered in a few sentences AT MOST.
- Make sure to write your answers clearly and legibly.
- Write your initials on the top right hand corner of each page.

First name	
Last name	
BMC/HC Email (What you use for gradescope)	

First and last name of student to your left	
First and last name of student to your right	

I agree to complete this exam without unauthorized assistance from any person, materials or device.

Signature:

For staff use only:

Total	/81

Q1. [18 pts] Multiple Choice

For the following T/F questions, if the answer is False, explain why.

(a) [2 pts] In an observational study, if the treatment and control groups differ in ways other than the treatment, we can make conclusions about causality.

○ True
○ False

(b) [2 pts] 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

○ True
○ False

(c) [2 pts] 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.

○ True
○ False

(d) [2 pts] 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.

○ True
○ False

(e) [2 pts] A/B testing is used to determine whether or not we believe two samples come from the same underlying distribution.

○ True
○ False

(f) [2 pts] 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.

○ True
○ False

(g) [2 pts] The law of averages states that if a chance experiment is repeated independently and under identical conditions, then, in the long run, the proportion of times that an event occurs gets closer and closer to the theoretical probability of the event.

○ True
○ False

(h) [1 pt] Which of the following is an example of categorical data?

○ Temperatures in degrees Fahrenheit of different cities.
○ The colors of cars in a parking lot.
○ Heights of students in a classroom.
○ The weights of packages shipped via a courier.

(i) [1 pt] Which of the following best illustrates the concept that association does not imply causation?

○ Ice cream sales are associated with the rate of drowning deaths.
○ Increased study time is associated with higher grades.
○ Height is associated with weight in adults.
○ Smoking is associated with lung cancer incidence.

(j) [1 pt] Which of the following figures are used to find associations between numerical variables:

○ Scatter Plot
○ Histogram
○ Line Plot
○ Bar Chart

(k) [1 pt] Which of the following figures are used to visualize categorical distributions:

○ Scatter Plot
○ Histogram
○ Line Plot
○ Bar Chart

Q2. [14 pts] Short Answers

(a) [4 pts] In the Monty Hall problem, should the participant change their guess after one of the doors is opened? Why or why not? How did we demonstrate this in class?

(b) [2 pts] How can we use `Table` in our code if we do not explicitly define the name `Table`.

(c) [4 pts] Briefly explain why we use `total variation distance`, and why does it included taking an absolute value?

(d) [4 pts] Imagine you are standining at a street corner and take as your sample the first ten people who pass by? Is this sample considered a random sample? Explain why or why not? If it isn't a random sample, then what type of sample is this?

Q3. [16 pts] Probability

Show your work for partial credit

(a) [4 pts] What is the probability of rolling a sum of 8 with two dice?

(b) [4 pts] If two cards are drawn from a standard deck of 52 cards without replacement, what is the probability that both cards are Aces?

(c) [4 pts] If you toss a fair coin three times, what is the probability that you will get exactly two heads?

(d) [4 pts] An urn contains 4 red balls and 6 blue balls. Two balls are drawn randomly without replacement. What is the probability that both balls are blue?

Q4. [33 pts] Programming

Complete the following questions.

(a) [15 pts] **Pokeman**

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

Table 1: Sample Pokémon Stats

(i) [5 pts] Find the name of the pokemon of type `Water` that has the highest HP
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.

```
water_pokemon = pokemon._____________(____________,___________________)
```

```
water_pokemon.________(__________,____________).column("Name").item(0)
```

(ii) [5 pts] Find the proportion of pokemon of type `Fire` in the dataset whose Speed is strictly less than 100

```
 fire_pokemon = pokemon._____________(_______________,_______________)
```

```
fire_pokemon._______(_________,____________)._________/________________
```

(iii) [5 pts] 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",__________________)._________(_________,____________)
```

‘

(b) [18 pts] **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.

(i) [4 pts] 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():
```

(ii) [4 pts] 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.

```
distances = ...

for i in np.arange(10000):


    new_distance = ...



    distances = ...
```

(iii) [1 pt] What type of plot should we use to visualize the values in `distances`?

(iv) [3 pts] 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 and seems far from the origin, 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:

Alternative Hypothesis:

Test Statistic:

(v) [6 pts] 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 of test statistics under the null hypothesis. Try to use the `distances` *array.*

```
p_value = ...
```