Name: Steven Ahemri **Student ID:** 12261552

DATA 118: Introduction to Data Science 1
Winter 2024
Homework 7: Review
Due: March 1st - 11:59 pm

Problem	Total Points	Points Received
Multiple Choice	10	
Problem 1	8	
Problem 2	3	
Problem 3	8	
Problem 4	6	
Problem 5	15	
Total	50	

Name: Steven Alnemri **Student ID:** 12261552

The following section contains 10 multiple choice questions, each worth 1 point. Please choose the best answer to each question and indicate your answer by filling in the corresponding bubble on the previous page.

1. A random sample of 5000 students were asked whether they prefer a 10 week quarter system or a 15 week semester system. Of the 5000 students asked, 500 students responded. The results of this survey...

   (a) should not be generalized to the entire student body because it suffers from non-response bias.

   (b) can be generalized to the entire student body because it is a random sample.

   (c) can be generalized to the entire student body because the sample size is large.

   (d) should not be generalized to the entire student body because it suffers from sampling bias.

2. The numerical estimate for a population parameter which we determine from a sample is called:

   (a) ~~Simulation~~

   (b) ~~Confidence Coefficient~~

   (c) Statistic

   (d) ~~Null~~

3. A researcher plans to obtain data by interviewing household members of victims who perished in a tornado. She will interview them, and people in the same town who didn't have relatives die over the next 10 years to see how closeness to a traumatic event might affect recovery time.

   (a) This is a prospective observational study.

   (b) This is a retrospective observational study.

   (c) This is an experiment.

   (d) This is a voluntary response sample.

4. What is the value of the following expression? `23 % 3`?

   (a) 7

   (b) 7.666666666666667

   (c) 3

   (d) 2

5. Which of these would you expect to be binomially distributed?

   (a) The number of clubs in 5 cards drawn without replacement from a deck

   (b) The number of boys in a family

   (c) The number of girls in a family of 5 children

   (d) The number of children in a family

Name: Steven Ahemri **Student ID:** 1226152

6. Valeria claims that she can distinguish between Coke and Pepsi 90% of the time. Eva claims that Valeria just guesses. To settle this, a bet is made. Valeria is to be given 20 small glasses, each having been filled with either Coke or Pepsi. She wins if she gets 18 or more correct. What is the probability that Valeria wins the bet if she has the ability she claims? That is, what is the probability of getting 18 or more correct (assuming Valeria has the ability she claims)?

   (a) $\binom{20}{18}(.9)^{18}(.1)^2$

   (b) $1-\binom{20}{19}(.9)^{19}(.1)^1$

   (c) $1-\binom{20}{19}(.9)^{19}(.1)^1-\binom{20}{20}(.9)^{20}(.1)^0$

   (d) $\binom{20}{18}(.9)^{18}(.1)^2+\binom{20}{19}(.9)^{19}(.1)^1+\binom{20}{20}(.9)^{20}(.1)^0$

7. Suppose we have a discrete distribution and events A, B, D contained in the sample space S such that $P(A) = 4/11, P(B) = 5/11, P(A \cap B) = 3/11$, and $D = S - (A \cup B)$. Find $P(A|B)$.

   (a) $\frac{20}{121}$

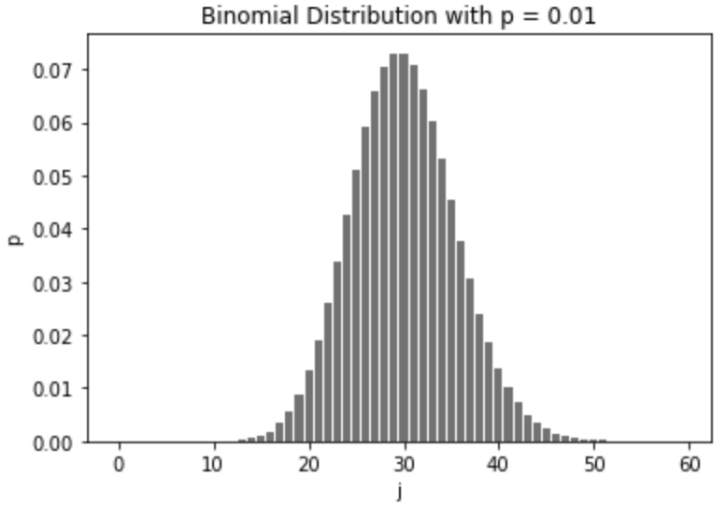
   $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3/11}{5/11} = 3/5$

   (b) $\frac{4}{11}$

   (c) $\frac{3}{5}$

   (d) $\frac{5}{6}$

8. This is a Binomial Distribution with probability of success $p = 0.01$. What is n?



   (a) 300

   (b) 3000

   Bin → N as n ↑

   (c) 3

   (d) 30

9. Suppose a researcher is computing 12 hypothesis tests. Assuming the outcome of each test is independent from every other test, and the null hypothesis is true is all cases. What is the probability of making at least one Type I error in 12 tests, assuming a significance level of 0.05?

   (a) $1 - 0.05$

   (b) $(1 - 0.05)^{12}$

   (c) $1 - (1 - 0.05)^{12}$

   (d) $(0.05)^{12}$

   $P(\text{At least one type I}) = 1 - P(\text{No type I})$
   $= 1 - (1-0.05)^{12}$

10. The following variables have been defined as follows:

```
tut = 4 > 8
bup = 7 != 7.0 or 5 > 3
sus = 4 * bup + tut
glob = np.arange(sus, sus*3, 2)
```

   start stop step

   What is the value of `glob[2]`? Assume `NumPy` has been imported as `np` prior to defining these variables. (Hint: Writing intermediate steps may help you.)

   (a) 8

   (b) 7

   (c) 6

   (d) 9

   tut = False

   bup = True or True

   sus = 4 x True + False = 4

   glob = np.array([4, 6, 8, ...
   0 1 2

Name: Steven Alnemri Student ID: 1226552

The following section contains 5 short answer questions, worth a total of 40 points. Please read and respond to each question carefully. Show all of your work!

Question 1. (8 pts)

Pandas: The `donuts` and `cashiers` dataframes below are loaded into memory. Each row of `donuts` consists of information regarding a transaction at the donut shop, with the columns "cashierID" (ID number referencing the cashier working), "Day" (day of the week the transation occurred), "Bill" (amount paid), "Tip" (amount patron tipped the cashier), and"Flavor" (flavor of donut(s) purchased). Each row of the `cashiers` dataframe contains information about a cashier working at the donut shop: their "cashierID" , "Name" , and "Age".

donuts

	cashierID	Day	Bill	Tip	Flavor
0	1	Fri	6.07	0.70	Glazed
1	3	Sun	3.98	0.01	Chocolate
2	1	Sat	7.23	0.71	Vanilla
3	1	Sun	4.24	0.85	Chocolate
4	3	Tues	7.03	2.35	Glazed
...	...	...	...	...	...
95	1	Sun	5.77	0.62	Chocolate
96	2	Wed	3.80	1.83	Vanilla
97	3	Mon	6.76	0.24	Vanilla
98	2	Fri	6.98	1.04	Glazed
99	2	Tues	5.73	0.26	Chocolate

100 rows × 5 columns

cashiers

	cashierID	Name	Age
0	1	Bart	28
1	2	Homer	16
2	3	Lisa	34

Write Pandas expressions to compute each result. For example, if the result prompt is, "Return the sum of all bills charged by the donut shop," then you would fill in: `sum(donuts['Bill'])`. The last line of each answer should evaluate to the result requested; you never need to call `print`. Each part should be answerable with one or two lines of code. (Points will be taken off for overly long answers).

(a) (2pts) Return a dataframe containing only the rows from the donuts table that correspond to transactions where Glazed donuts were purchased.

```
donuts_df = donuts.loc[cashiers['Flavor'] == 'Glazed']

donuts_df
```

(b) (2pts) Add a new column called `total` to the `donuts` dataframe, which contains the sum of the Bill and Tip in each row.

`donuts['total'] =` `donuts['Bill'] + donuts['Tip']`

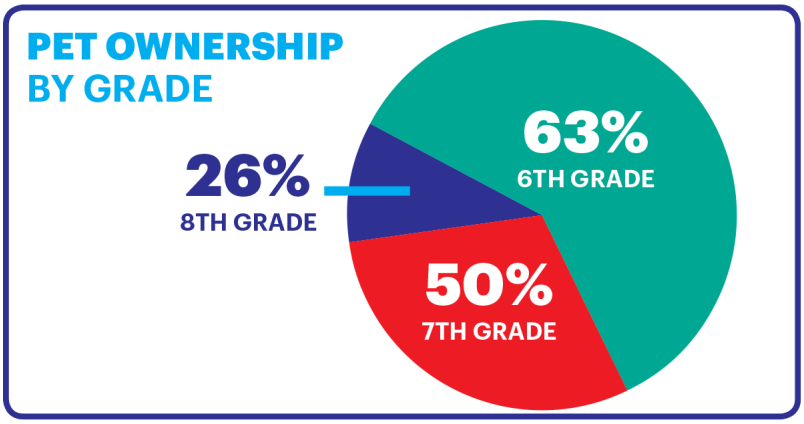
(c) (4pts) Return the total of the tips earned by the cashier named Homer. [Note: for full points you must combine the dataframes and search on the string "Homer"]

```
together = pd.merge( donuts , cashiers , how = inner )

Sum (together.loc[ together['Name']=='Homer'] [ 'Tip' ])
```

Question 2. (3 pts)

Explain what is wrong, missing, or misleading about the following graph.
This graph was taken from a 2017 *Scholastic Math* article titled "Fake News, Fake Data: How bad data and misleading graphs are fueling fake news":



The proportion of Pet ownership by grade is not properly represented by the graphic. 7th Graders take 50% but less than 50% in the pie chart

Name: Steven Alnemri **Student ID:** 12261552

Question 3. (8 pts)

Probability

(a) (2 pts) A statistics professor recorded the number of siblings for each student in a statistics class. The teacher claims that the probability distribution for the number of siblings is represented in the table below.

Number of Siblings	0	1	2	3	4	5
$P(X = x)$	0.12	0.55	0.17	0.10	0.05	0.05
		67	84	94	99	1.04

Determine whether or not the probability distribution is valid, and explain your reasoning.

No, the distribution does not add up to 1, it is > 1.0.

(b) Your data science professor recorded the number of hours, rounded to the nearest hour, students spent studying for the final exam. She constructs the following probability distribution.

Number of Hours Spent Studying	0	1	2	3
$P(X = x)$	$\frac{1}{6}$	$\frac{5}{12}$	$\frac{1}{4}$	$\frac{1}{6}$

(i) (2 pts) What is the probability a student selected at random studied exactly 2 hours?

$\frac{1}{4}$

(ii) (2 pts) What is the probability a student selected at random studied at least 2 hours?

$\frac{1}{4} + \frac{1}{6} = \frac{3}{12} + \frac{2}{12} = \frac{5}{12}$

(iii) (2 pts) Select 2 students at random with replacement. What is the probability that both students studied at least 2 hours?

$\frac{5}{12} \cdot \frac{5}{12} = \frac{25}{144}$

Name: Steven Alnemri Student ID: 12261552

Question 4. (6 pts)

Hemoglobin (Hb) is a protein produced by red blood cells that is important for transporting oxygen in the blood. It can be measured in the clinic to help diagnose anemia. The table below shows ranges of Hb levels in the blood (in females) and the corresponding symptom description:

Hb Values (g/dl)	Symptom Severity
12 - 16	Normal
10 - 12	Mild
8 - 10	Moderate
6.5 - 8	Severe
< 6.5	Life-threatening

The upper bounds of these ranges are exclusive. Any observation other than the aforementioned ranges is considered abnormal.

(a) (3 pts) Complete the function below called `hb_symptom()` which takes as input a patient's blood Hb level as a float point number and returns the descriptor of symptom severity for anemia. Write the appropriate conditional statements in the blanks.

```
def hb_symptom(Hb) :
    '''A function that determines the anemia symptom based on an input Hb level'''
    if Hb < 6.5 :
        return 'Life Threatening'
    elif Hb < 8 :
        return 'Severe'
    elif Hb < 10 :
        return 'Moderate'
    elif Hb < 12 :
        return 'Mild'
    elif Hb < 16 :
        return 'Normal'
    else :
        return 'High'
```

(b) (3 pts) The below dataframe, which is defined as `df`, lists the Hb levels for a set of patients. Complete the below code such that a new column in `df` is created called `Symptom` based on the values in the `Hb` column. Utilize your previously defined function to do this.

	Patient ID	Hemoglobin
0	1	6
1	2	15
2	3	6
3	4	13
4	5	5
5	6	15
6	7	14
7	8	11
8	9	12
9	10	6
10	11	7
11	12	12

```
import pandas as pd

df['Symptom'] = hb_symptom(df['Hemoglobin'])
```

Name: Steven Alnemri **Student ID:** 12261552

Question 5. (15 pts)

A local school wants to assess the efficacy of a new standardized test prep course. The average math score in the US on this standardized test is 66. We know that these scores are normally distributed with standard deviation of 15. You are hired to decide if there is evidence of higher math scores for students who have taken the new prep course. After collecting the data you see the prep course students had an average math score of 90.

normal
Sd = 15

(a) (1pts) State the null and alternative hypotheses.

H_0: The average math score between students with prep > (greater than) without prep are the same

$\mu_1 > \mu_2$

(b) (1pts) Is this a 1 or 2 tailed test? Why?

Single-tailed. We only want to know if students with preparation have a higher mean score

(c) (1pt) What is the test statistic?

90

(d) (4pts) To test our hypothesis, we need to create an empirical distribution of test statistics under the null hypothesis. Fill in the blanks in the following code to create and plot this distribution. Note 'loc' corresponds to mean and 'scale' corresponds to standard deviation.

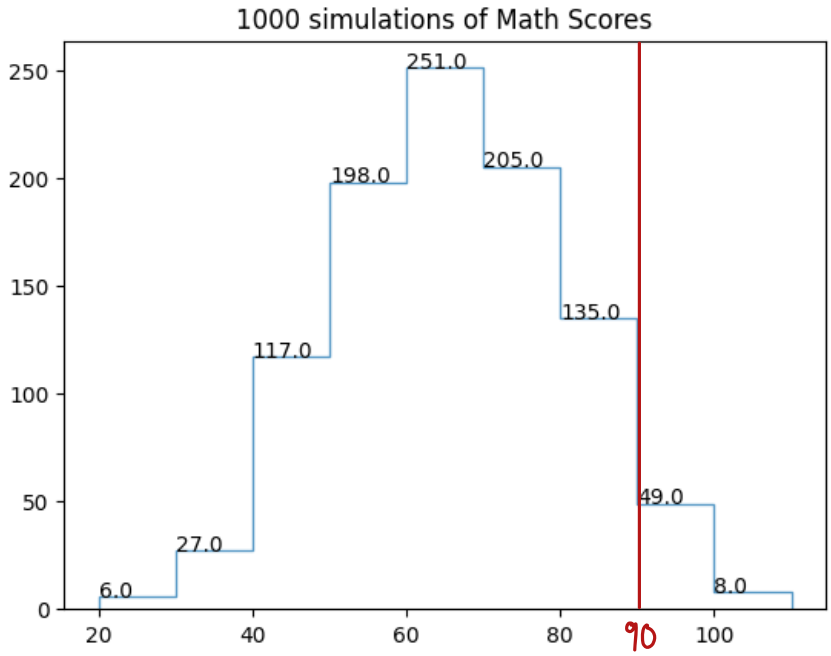
we assume null is true

```
import numpy as np
import matplotlib.pyplot as plt

sim = np.random.normal( loc = 66 , scale = 15 , 1000)

plt.hist( sim )
```

(e) (2pt) The empirical distribution from part d is shown below. Indicate where our test statistic lies on this distribution using a vertical line through the image and label it with its value.



(f) (2pt) If our significance level is 0.09, what can we say about our null and/or alternative hypotheses?

$(49+8)/(6+27+117+198+251+205+135+49+8) = 0.057$

we reject the null hypothesis. There is significant evidence supporting that students with prep have higher average score.

(g) (4pt) The school district decides to move forward with the prep course. After 2 years, they realize that their participants' math scores tend to be lower than their reading scores. They rehire you and provide you with the past 2 years of test score data to assess the difference in these scores. Using the code on the following page, calculate a 95% Basic Bootstrap Confidence Interval for this difference. (Note: You may or may not need to use all of the provided information, and you can give your answer rounded to two decimal places).

Basic Bootstrap: $[\hat{\theta} - U_1, \hat{\theta} - L_1]$

where $\hat{\theta}$ = original Sample diff = mydiff = -3.08

U_1 = np.quantile (bootstrap_diff − mydiff, 0.975) = 0.58715

L_1 = " " 0.025) = -.58715

Basic Bootstrap: $[-3.08 - 0.58715, -3.08 - (-0.58715)]$

$= [-3.67, -2.49]$

Name: Steven Alnemri **Student ID:** 12261552

```
In [1]: import pandas as pd
        import numpy as np

        testdat = pd.read_csv("StudentsPerformance.csv")
```

```
In [2]: def test_resampling(my_data,nboots):
            outcomes = np.array([])
            for i in np.arange(nboots):
                bootstrap_sample=my_data.sample(n=len(my_data),replace=True)
                bootstrap_diff=np.mean(bootstrap_sample['math score'] - bootstrap_sample['reading score'])
                outcomes = np.append(outcomes, bootstrap_diff)
            return outcomes
        bootstrap_results = test_resampling(testdat,1000)
```

```
In [3]: mydiff = np.mean(testdat['math score'] - testdat['reading score'])
        mydiff
Out[3]: -3.08
```

X

```
In [4]: np.quantile(bootstrap_results-mydiff,0.005)
Out[4]: -0.8080149999999998

In [5]: np.quantile(bootstrap_results-mydiff,0.995)
Out[5]: 0.8210150000000002

In [6]: np.quantile(bootstrap_results,0.005)
Out[6]: -3.8880149999999998

In [7]: np.quantile(bootstrap_results,0.995)
Out[7]: -2.258985

In [8]: np.quantile(mydiff-bootstrap_results,0.005)
Out[8]: -0.8210150000000002

In [9]: np.quantile(mydiff-bootstrap_results,0.995)
Out[9]: 0.8080149999999998

In [10]: np.quantile(bootstrap_results-mydiff,0.05)
Out[10]: -0.4991500000000001

In [11]: np.quantile(bootstrap_results-mydiff,0.95)
Out[11]: 0.48224999999999996

In [12]: np.quantile(bootstrap_results,0.05)
Out[12]: -3.5791500000000003

In [13]: np.quantile(bootstrap_results,0.95)
Out[13]: -2.59775

In [14]: np.quantile(mydiff-bootstrap_results,0.05)
Out[14]: -0.4822500000000002

In [15]: np.quantile(mydiff-bootstrap_results,0.95)
Out[15]: 0.49915
```

```
In [16]: np.quantile(bootstrap_results-mydiff,0.025)
Out[16]: -0.5490249999999999
```

O

```
In [17]: np.quantile(bootstrap_results-mydiff,0.975)
Out[17]: 0.5871500000000001

In [18]: np.quantile(bootstrap_results,0.025)
Out[18]: -3.629025

In [19]: np.quantile(bootstrap_results,0.975)
Out[19]: -2.4928500000000002
```

O

```
In [20]: np.quantile(mydiff-bootstrap_results,0.025)
Out[20]: -0.5871500000000002

In [21]: np.quantile(mydiff-bootstrap_results,0.975)
Out[21]: 0.5490249999999999
```