DATA 8 Fall 2020

Sample Exam.

MIDTERM

INSTRUCTIONS

This is your exam. Complete it either at exam.cs61a.org or, if that doesn't work, by emailing course staff with your solutions before the exam deadline.

This exam is intended for the student with email address `<EMAILADDRESS>`. If this is not your email address, notify course staff immediately, as each exam is different. Do not distribute this exam PDF even after the exam ends, as some students may be taking the exam in a different time zone.

For questions with **circular bubbles**, you should select exactly *one* choice.

- ◯ You must choose either this option
- ◯ Or this one, but not both!

For questions with **square checkboxes**, you may select *multiple* choices.

- ☐ You could select this choice.
- ☐ You could select this one too!

You may start your exam now. Your exam is due at <DEADLINE> Pacific Time. Go to the next page to begin.

For fill-in-the-blank coding questions, you can put anything inside the blanks, including commas, parentheses, and periods.

The exam is worth 100 points.

If you encounter any logistical problems during the exam, please contact us at data8berkeley@gmail.com. We will not be answering any questions related to the contents of the exam.

(a) Your name:

(b) Your @berkeley.edu email address:

(c) The Berkeley Honor Code states: "As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others." Do you agree to follow the honor code on this exam?

◯ Yes

◯ No

1. (7 points) Experiments

(a) (1 pt) We have learned that an association between two factors does not necessarily mean that one causes the other ("correlation does not imply causation"). Which of the following would be the best example to demonstrate that? Choose one.

- ● A retrospective study found an association between poor reading comprehension and having experienced a heart attack in the past 10 years, but this is not because strong reading skills prevents heart attacks; it is because poverty tends to cause weaker reading comprehension (because schools in poor areas tend to be weaker) and poverty tends to increase the risk of heart attacks (because poor people have worse access to medical care).
- ○ A retrospective study found an association between poor reading comprehension and having experienced a stroke in the past 10 years, but this is not because strong reading skills prevent strokes; it is because strokes often cause a cognitive decline and worse reading comprehension.
- ○ Data shows an association between smoking and lung cancer. This is because smoking causes lung cancer.
- ○ Jon Snow discovered that S&V customers died from cholera at a higher rate than Lambeth customers, but this was due to Lambeth's water being cleaner than S&V's; it had nothing to do with dirty water causing cholera.

(b) (1 pt) Suppose we have discovered an association between two variables in a dataset. Which of the following would be the best way to test whether it is causal? Choose one.

- ○ Use hypothesis testing to check whether the association is statistically significant.
- ● Run a randomized controlled experiment.
- ○ Brainstorm some potential confounding factors and test whether any of them has an association with both variables.
- ○ If both variables are numerical, use a scatter plot to check for a trend.

(c) (1.5 pt) Which of the following must be true, for an experiment to count as a randomized controlled experiment? Select all that apply.

- ■ There is a control group.
- ☐ The experimenters control who is selected to participate in the experiment.
- ■ Randomness is used to determine whether each participant will be part of the control group or treatment group.
- ☐ Each participant is informed whether they are in the treatment group or not.
- ☐ The distribution of ages of the participants in the experiment are representative of the distribution of ages in the population at large.

(d) **(1.5 pt)** Many people believe that taking vitamin C supplements regularly causes you to be less likely to get sick in day-to-day life. As the leader of a research team, you would like to find out if this is true or not. Which of the following studies would be a persuasive way to answer this question? Select all that apply.

- ☐ Recruit many participants. Ask them all to take vitamin C supplements regularly. For each participant, track whether they get sick over the next year.
- ■ Recruit many participants. For each participant, randomly flip a coin. If it comes up heads, ask them to take vitamin C supplements regularly; if tails, ask them to not take vitamin C. Track whether they get sick over the next year.
- ☐ Recruit many participants. For each participant, ask them about whether they take vitamin C supplements and about whether they got sick over the past year. Eliminate anyone who has taken vitamin C supplements irregularly over the past year; keep only people who either haven't taken vitamin C supplements at all or have taken vitamin C regularly.
- ☐ Partner with a local hospital. Analyze the electronic records of patients who are sick to determine what fraction regularly take vitamin C supplements regularly. Then, perform a survey of a random sample of people in the population at large and ask them how whether they regularly take vitamin C supplements. Compare these two proportions.
- ☐ Recruit many sick patients from a local hospital. For each participant, randomly flip a coin. If it comes up heads, ask them to take vitamin C supplements regularly; if tails, ask them to not take vitamin C. Track how long it takes for them to recover from their sickness.

(e) **(2 pt)** Suppose you perform a randomized controlled experiment where people in the treatment group take vitamin C regularly and people in the control group do not take vitamin C. In each group you measure the proportion that get sick over the next year. Which of the following statements are correct? Select all that apply.

- ☐ If the control group has a higher proportion of sickness, then it is reasonable to conclude that vitamin C causes a reduction in the chance of getting sick.
- ■ If the control group has a higher proportion of sickness and a hypothesis test finds that this difference is statistically significant, then it is reasonable to conclude that there is an association between taking vitamin C and not getting sick.
- ■ If the control group has a higher proportion of sickness and a hypothesis test finds that this difference is statistically significant, then it is reasonable to conclude that vitamin C causes a reduction in the chance of getting sick.

2. (8 points) Visualization

We have a dataset of all buildings in Berkeley, with three attributes for each building: its size (in square feet), its type (residential or commercial), and its estimated value (sale price) if sold (in dollars).

(a) (1 pt) Select all that are correct:

- ☐ The size attribute is a categorical variable.
- ■ The type attribute is a categorical variable.
- ☐ The value attribute is a categorical variable.

(b) (1 pt) Select all that are correct:

- ■ The size attribute is a numerical variable.
- ☐ The type attribute is a numerical variable.
- ■ The value attribute is a numerical variable.

(c) (1.5 pt) The best visualization to understand the distribution of building sizes is: (choose one)

- ○ A bar chart
- ○ A line plot
- ○ A scatter plot
- ● A histogram

(d) (1.5 pt) The best visualization to understand the distribution of building types is: (choose one)

- ● A bar chart
- ○ A line plot
- ○ A scatter plot
- ○ A histogram

(e) (1.5 pt) The best visualization to check for an association between building size and building type is: (choose one)

- ○ A bar chart
- ○ A line plot
- ○ A scatter plot
- ○ A histogram
- ● Two histograms, overlaid

(f) (1.5 pt) The best visualization to check for an association between building size and building value is: (choose one)

- ○ A bar chart
- ○ A line plot
- ● A scatter plot
- ○ A histogram
- ○ Two histograms, overlaid

3. (11 points) Histograms

Sokka really likes to drink cactus juice. Katara and Aang decided to record how many liters he drank per day in a month (30 days). They make a histogram of Sokka's cactus juice consumption for a month using 5 bins, as shown below. The heights of 4 of the bins are given below.

The heights of the bins, in order, are 5, 4.444, ?, 5.333, and 5. Note the height of bin 3 is not given; you'll have to calculate this in one of the questions below.

Write a mathematical expression that evaluates to the quantity described. Do not use any Python array or table operations. You do not need to simplify your answer (e.g., `(2+5)/4` is ok). If there is not enough information, write "Not enough information". You do not need to justify your answer.

(a) (2 pt) The percent of the month that Sokka drank less than 2 liters of cactus juice.

```
5*2
```

(b) (2 pt) The percent of the month that Sokka drank less than 4 liters of cactus juice.

Not enough information

(c) (2.5 pt) The percent of the month that Sokka drank less than 5 liters of cactus juice.

```
2*5 + 3*4.444
```

(d) (2.5 pt) The height of Bin 3 in the histogram, assuming there were 12 days where Sokka drank 5 - 8 liters of cactus juice (not including 8 liters).

```
(100 * 12/30) / 3
```

(e) (2 pt) The number of days that Sokka drank less than 2 liters of cactus juice in the month. (not the percent)

```
30 * 2*5/100
```

4. (12 points) Programming in Python

For each question below, write Python code to answer the question using what we have taught you in this class. If we ran your Python code, it should evaluate to the answer to the question.

(a) (1 pt) This year my salary is $BASE_SALARY. I just learned I will receive a 12% raise. What will my salary be next year, in dollars?

```
1.12 * BASE_SALARY
```

(b) (1 pt) The absolute value of the difference between the two variables `x` and `y`. `x` and `y` are numbers.

```
abs(x-y)
```

(c) (1 pt) The smaller of the two variables `x` and `y`. `x` and `y` are numbers.

```
min(x,y)
```

(d) (1 pt) An array containing the numbers 6, 9, A_NUM, and 12, in that order.

```
make_array(6, 9, A_NUM, 12)
```

(e) (1 pt) The average of the numbers in the array `arr`.

```
np.mean(arr) or np.average(arr) or sum(arr)/len(arr)
```

(f) (1 pt) An array containing the first 100 squares, i.e., 1, 4, 9, 16, . . . , 10000. The square of a number is that number to the power of 2.

```
np.arange(1, 101)**2 or np.arange(1, 101) * np.arange(1, 101)
```

(g) (2 pt) Assume you are given a function `simulate_once()` that, when called, returns the result of a single simulation. Write code that creates an array containing the result of NUM_SIMS simulations and assigns it to `RESULTS`.

```
RESULTS = make_array()
for i in np.arange(NUM_SIMS):
  r = simulate_once()
  RESULTS = np.append(RESULTS, r)
```

(h) **(2 pt)** In one sentence, please describe what the following function returns.

```
def mystery(x):
    for i in np.arange(1, x+1):
        if i*i >= x:
            return i
    return 0
```

Returns the square root of a number x, rounded up.

(i) **(2 pt)** The following function is intended to return an array containing the 100 even numbers 2, 4, 6, 8, . . . , 200. However it has two bugs. Describe both bugs.

```
def buggy():
    EVENS = make_array()
    for i in np.arange(100):
        np.append(EVENS, 2*i)
    return EVENS
```

Needs to re-assign to EVENS, i.e., `EVENS = np.append(EVENS, 2*i)`**. Should use** `np.arange(1, 101)`**.**

5. (28 points) Table Manipulations

The table `TRUCKS` displays recent applications for food truck permits in San Francisco, one row per application:

Applicant	Day of Week	Start Time	End Time	Hot Dishes	Latitude	Longitude
Kettle Corn Star	Tuesday	10	18	Y	37.7862	-122.405
Kettle Corn Star	Monday	10	13	Y	37.7862	-122.405
Kettle Corn Star	Monday	15	18	Y	37.7862	-122.405
Off the Grid Services, LLC	Monday	08	15	Y	37.7778	-122.397

... (3106 rows omitted)

The 3rd and 4th columns record the hour when service starts and ends, as an integer. For instance, 15 refers to the time 15:00, which is 3 pm. The 'Hot Dishes' columns indicates whether or not the food truck prepares hot dishes on-site.

For each question below, write Python code to answer the question using the table operations we have taught you in this class.

We show the shape of our staff solution, but you are not required to follow it.

You should not use for-loops, define new functions, or need more than a few lines of code.

(a) (2 pt) The number of applications that list DAY_WEEK in the Day of Week column.

Our staff solution has the form `____.____(____, ____).____`

```
TRUCKS.where("Day of Week", "DAY_WEEK").num_rows
```

(b) (2 pt) How many applications has each applicant submitted? The result should be a table with just two columns, Applicant and count, and one row per applicant. Make sure there are no other columns in the output.

Our staff solution has the form `____.____(____)`

```
TRUCKS.group("Applicant")
```

(c) (3 pt) List all days of the week that the applicant EXAMPLE_APP has submitted an application for. The result should be a table with just two columns, Applicant and Day of Week, and one row per day of the week with an application from EXAMPLE_APP. No day of the week should appear in the result more than once. Make sure there are no other columns in the output.

Our staff solution has the form `____.____(____).____(____, ____).____(____, ____)`

```
TRUCKS.group(['Applicant', 'Day Of Week']).where('Applicant',
'EXAMPLE_APP').select('Applicant', 'Day of Week')
```

(d) (3 pt) Which applicant has the most permit applications? The result should be a string with the name of the applicant. (you can resolve a tie in any way you want)

Our staff solution has the form `____.____(____).____(____, ____).____(____).____(____)`

```
TRUCKS.group('Applicant').sort('count', descending=True).column('Applicant').item(0)
```

(e) **(2 pt)** A bar chart that shows, for each day of the week, the number of applications for that day of week.

Our staff solution has the form `____.____(____).____(____, ____)`

```
TRUCKS.group("Day of Week").barh('Day of Week', 'count')
```

(f) **(3 pt)** Create a table that shows the average AVG_BY of all trucks, broken down by day of the week. (For instance, the table should show the average AVG_BY of all applicants for Monday, the average AVG_BY of all applicants for Tuesday, and so on.)

Our staff solution has the form `____.____(____).____(____, ____)`

```
TRUCKS.group('Day of Week', np.mean).select('Day of Week', 'AVG_BY mean')
```

(g) **(3 pt)** Create a table that shows the average AVG_BY of all trucks that close strictly before CLOSE_BEFORE:00, broken down by day of the week and by whether the truck serves hot dishes or not.

Our staff solution has the form `____.____(____, ____).____(____, ____)`

```
TRUCKS.where("End Time", are.below(CLOSE_BEFORE)).pivot("Hot Dishes", "Day
of Week", values='AVG_BY', collect=np.mean)    OR    TRUCKS.where("End Time",
are.below(CLOSE_BEFORE)).group(["Hot Dishes", "Day of Week"], np.mean)
```

(h) **(4 pt)** Create a table that shows, for each applicant, how many different days of the week they have applied for. For instance, Kettle Corn Star has applied for 2 days of the week (Monday and Tuesday). Your table should have two columns, 'Applicant' and 'Number of Days'.

Our staff solution has the form `____.____(____).____(____).____(____, ____)`

```
TRUCKS.group(["Applicant", "Day of Week"]).group("Applicant").relabeled('count',
'Number of Days')
```

(i) **(2 pt)** Parham, who collected the data, forgot to ask each applicant to include their phone numbers on their application. He asks each applicant to give him their phone number and stores it in a two column table called `PHONE_NUMBERS_TBL`. The table has a column called "Applicant Name" that contains the unique names of the applicants and a column called "Phone Number" that contains the phone numbers of each applicant. Complete the line of code so that it evaluates to a table that has all the same columns as the `TRUCKS` table as well as a column called "Phone Number" that contains the phone number for that applicant.

Our staff solution has the form `____.____(____, ____, ____)`

```
TRUCKS.join("Applicant", PHONE_NUMBERS, "Applicant Name")
```

(j) (4 pt) San Francisco Food Truck Secretary Connor has devised his own formula for determining whether or not to accept the applications from a food truck. He has implemented this formula in a function called `ACCEPT` that takes three arguments, the minimum end time for all of an applicant's applications on Monday, on Thursday and on Saturday. `ACCEPT(mon_end, thu_end, sat_end)` returns a boolean. Fill out the following code so that result evaluates to a two column table, where the first column contains the name of each applicant and the second contains the value of Connor's formula for that applicant.

Our staff solution has the form:

```
earliest_close = TRUCKS.____(____, ____, ____, ____)
DECISIONS = earliest_close.____(____, ____, ____, ____)
result = earliest_close.select("Applicant").____("Decision", ____)
```

earliest_close = TRUCKS.pivot("Day of the Week", "Applicant", "End Time", min) DECISIONS = earliest_close.apply(ACCEPT, "Monday", "Thursday", "Saturday") result = earliest_close.select("Applicant").with_column("Decision", DECISIONS)

6. (8 points) Probability

(a) Amy is competing on a Japanese game show. First Amy randomly chooses one of three kitchen implements (a spatula, a single chopstick, or a soup ladle, all equally likely), then randomly chooses one of three food items (a raw egg, a pancake, or a glass of orange juice, all equally likely). She has to run through a maze while carrying the chosen food on her chosen implement.

Answer each of the following questions. Express your answer as a Python expression (e.g., `(4/5) * (3/7)`).

i. (1 pt) Carrying something on a chopstick sounds awful. What is the probability that Amy picks the chopstick?

1/3

ii. (1 pt) What is the probability that Amy has to carry the pancake on a spatula?

1/9

iii. (1 pt) What is the probability that Amy does not have to carry the pancake on a spatula?

1 - (1/9)

iv. (2 pt) After Amy makes her choices, Bill randomly chooses one of the remaining two implements (both equally likely) and one of the remaining two food items (both equally likely). Then, Candice is stuck with the last implement and the last food item. What is the probability that Candice has to carry the glass of orange juice on the spatula?

1/9

v. (3 pt) Gregory is responsible for cooking the pancakes used in the game show. He cooks batches of 12 pancakes at a time and each pancake has a 1/5 chance of getting burnt. Gregory says the probability that at least three pancakes in a batch get burnt is 44.17%. What is the probability that exactly one or two pancakes in a batch get burnt?

1 - 0.4417 - (0.8)12**

(b) Amy is competing on a Japanese game show. First Amy randomly chooses one of four kitchen implements (a fork, a knife, a single chopstick, or a soup ladle, all equally likely), then randomly chooses one of four food items (a raw egg, a pancake, a muffin, or a cup of coffee, all equally likely). She has to run through a maze while carrying the chosen food on her chosen implement.

Answer each of the following questions. Express your answer as a Python expression (e.g., `(4/5) * (3/7)`).

i. (1 pt) Carrying something on a knife sounds awful. What is the probability that Amy picks the knife?

```
1/4
```

ii. (1 pt) What is the probability that Amy has to carry the egg on the ladle?

```
1/16
```

iii. (1 pt) What is the probability that Amy does not have to carry the egg on the ladle?

```
1 - (1/16)
```

iv. (2 pt) After Amy makes her choices, Bill randomly chooses one of the remaining three implements (both equally likely) and one of the remaining three food items (both equally likely). Then, Tom randomly chooses one of the remaining two implements (both equally likely) and one of the remaining two food items (both equally likely). Then, Candice is stuck with the last implement and the last food item. What is the probability that Candice has to carry the cup of coffee on the fork?

```
1/16
```

v. (3 pt) Gregory is responsible for cooking the pancakes used in the game show. He cooks batches of 10 pancakes at a time and each pancake has a 1/4 chance of getting burnt. Gregory says the probability that at least three pancakes in the batch get burnt is 47.44%. What is the probability that one or two pancakes get burnt?

```
1 - 0.4744 - (0.75)**10
```

7. (20 points) Hypothesis Testing

In the Avatar world, the world population is 30% Fire Nation, 40% Earth Kingdom, 25% Water Tribe, and 5% Air Nomads. Avatar Korra and her friends are trying to determine if the proportions of these four tribes in AVATAR_CITY City matches that in the world population.

Korra has selected a random sample of 300 people living in AVATAR_CITY City and calculated the proportion of people in the sample from each tribe:

Nationality	Proportion
Fire Nation	0.36
Earth Kingdom	0.45
Water Tribe	0.09
Air Nomads	0.10

Provide a null and alternative hypothesis that Korra can use to test if AVATAR_CITY City has the same proportions as the world population.

(a) (2 pt) Null hypothesis:

The distribution of tribes in AVATAR_CITY City is the same as the distribution in the world population.

(b) (1.5 pt) Alternative hypothesis:

AVATAR_CITY City has a different distribution of tribes than the world population.

(c) (2 pt) Which of the following are valid test statistics? Select all that apply. Assume that `world_pop` holds an array of proportions for the world population (i.e., `make_array(0.3, 0.4, 0.25, 0.05)`) and `city_sample` holds an array of proportions for the sample from AVATAR_CITY City.

- ■ `sum(np.abs(world_pop - city_sample)) / 2`
- ■ `sum(np.abs(world_pop - city_sample)) / 3`
- ■ TVD
- ☐ The difference in means
- ☐ `sum(world_pop - city_sample) / 2`
- ☐ `sum(world_pop - city_sample) / 3`

(d) (7.5 points)

Fill in the blanks below so that the code correctly performs a hypothesis test using the above defined test statistic. Assume that we have defined `test_statistic()` correctly to compute a valid test statistic.

```
world_pop = make_array(________(1)__________)
observed_sample = make_array(0.36, 0.45, 0.09, 0.10)
OBS_STAT = test_statistic(world_pop, observed_sample)

simulated_stats = _______(2)________
for i in np.arange(NUM_SIMS):
    one_sample = ________(3)________( ______(4)_______)
    test_stat = _______(5)___________(_____(6)______, one_sample)
    simulated_stats = np.append(________(7)_______, test_stat)

p_value = _____(8)_______(simulated_stats ____(9)_____) / ____(10)____
p_value
```

i. Blank 1

```
make_array(0.30, 0.40, 0.25, 0.05)
```

ii. Blank 2

```
make_array()
```

iii. Blank 3

```
sample_proportions
```

iv. Blank 4

```
300, world_pop
```

v. Blank 5

```
test_statistic
```

vi. Blank 6

```
world_pop
```

vii. Blank 7

```
simulated_stats
```

viii. Blank 8

```
np.count_nonzero or sum
```

ix. Blank 9

```
>= OBS_STAT
```

x. Blank 10

```
NUM_SIMS
```

(e) **(3 pt)** Suppose the result of running the above cell leads to `p_value = 0.0348`. Which of the following conclusions could be justified? Select all that apply.

- [x] If we use a p-value cutoff of 10%, we should reject the null hypothesis.
- [ ] If we use a p-value cutoff of 1%, we should reject the null hypothesis.
- [ ] Someone could reasonably conclude that AVATAR_CITY City's population has the same distribution as the world population.
- [x] Someone could reasonably conclude that AVATAR_CITY City's population has a different distribution from the world population.
- [ ] Avatar Korra did not sample randomly. This bias led to a high concentration of Air Nomads selected.

(f) **(2 pt)** Suppose instead now that Korra only wants to focus on the proportion of AVATAR_CITY City residents that are Water Tribe members. Her new null hypothesis is that 15% of city residents are Water Tribe, and her alternative hypothesis is that less than 15% of city residents are Water Tribe. She takes a random sample of 300 residents from AVATAR_CITY City and stores the number of them who are Water Tribe in `num_water_tribe`.

Which of the following is a good choice for an observed test statistic? Select all that apply.

- [x] `0.15*300 - num_water_tribe`
- [x] `0.15 - num_water_tribe/300`
- [ ] `abs(num_water_tribe/300 - 0.15)`
- [ ] `abs(0.15 - num_water_tribe)/300`

(g) **(2 pt)** Korra performed the hypothesis test from the previous part, using a different, but still valid, test statistic. She got the following histogram.

Note: the bins were generated using `np.arange(0.08, 0.22, 1/100)`

Which of the following could be the p-value, using the observed statistic of 0.09? Choose the best answer.

- (●) 0.001
- ( ) 0.05
- ( ) 0.1
- ( ) 0.5

8. (6 points) Testing A Hypothesis

Shmuel has collected some data about cells that have been infected with the ILLNESS virus and cells that have not been infected. For each cell, he has collected data about the number of nuclei in the cell. In some cases, ILLNESS may cause some cells to form multiple nuclei. Shmuel would like to determine if the number of nuclei in a cell is higher in cells that have been infected with ILLNESS compared to those that have not been infected.

(a) (2 pt) What should Shmuel's null hypothesis be?

The distribution of the number of nuclei in cells infected with ILLNESS is the same as the distribution of the number of nuclei in uninfected cells.

(b) (2 pt) What should Shmuel's alternative hypothesis be?

The number of nuclei in cells infected with ILLNESS is higher, on average, than the number of nuclei in uninfected cells.

(c) (2 pt) What is a reasonable test statistic that Shmuel can use?

The average number of nuclei for cells that have been infected - the average number of nuclei for cells that have not been infected

9. (0 points) Last Words

(a) If there was any question on the exam that you thought was ambiguous and required clarification to be answerable, please identify the question (including the title of the section, e.g., Experiments) and state your assumptions. Be warned: We only plan to consider this information if we agree that the question was erroneous or ambiguous and we consider your assumption reasonable.

No more questions.