Data 8 Foundations of Data Science
Spring 2018

MIDTERM SOLUTIONS

INSTRUCTIONS

- You have 45 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except the official midterm exam reference guide provided with the exam.
- Mark your answers **on the exam itself**. We will *not* grade answers written on scratch paper.

Last name	
First name	
Student ID number	
CalCentral email (`_@berkeley.edu`)	
GSI	
Name of the person to your left	
Name of the person to your right	
All the work on this exam is my own. **(please sign)**	

Question 0 (2 points) Write your name and SID in the space provided on one side of every page of the exam.

1. (18 points) Basketball Bonanza

Assume we have a tables for the 2016-2017 NBA Season. Assume that if a column contains numbers, then they are integers, and otherwise, it is a column of strings.

The `nba` table contains 8 columns. The first few rows are shown below.

player	prefix	position	age	salary	games	minutes	points
Al Horford	BOS	C	30	2.65401e+07	68	2193	952
Amir Johnson	BOS	PF	29	1.2e+07	80	1608	520
Avery Bradley	BOS	SG	26	8.26966e+06	55	1835	894
Demetrius Jackson	BOS	PG	22	1.45e+06	5	17	10
Gerald Green	BOS	SF	31	1.4106e+06	47	538	262
Isaiah Thomas	BOS	PG	27	6.58713e+06	76	2569	2199
Jae Crowder	BOS	SF	26	6.28641e+06	72	2335	999
James Young	BOS	SG	21	1.8252e+06	29	220	68
Jaylen Brown	BOS	SF	20	4.743e+06	78	1341	515
Jonas Jerebko	BOS	PF	29	5e+06	78	1232	299

Fill in the blanks of the Python expressions to compute the described values. You must use only the lines provided. The last line of each answer should evaluate to the value described. Assume that the statements `from datascience import *` and `import numpy as np` have been executed. You may add anything you would like to the blanks below, but you may not add code outside of the blanks.

(a) (3 pt) The age of the oldest NBA player.

```
______________________(nba._____________________________________('age'))
```

```
max(nba.column('age'))
```

(b) (5 pt) The three-letter prefix of the team which has the highest paid player with the position center (C) in the NBA. You may assume there is only one such player.

```
centers = nba._____________('position', ____________________)
```

```
centers._____________(_____________________________________).column(_________________).item(0)
```

```
centers = nba.where(position, are.equal_to('C'))
centers.sort('salary', descending=True).column('player').item(0)
```

(c) (5 pt) The number of teams that have fewer than 5 players older than 30.

```
old = nba._________________(_________________, are._____________________)


num_old = old._____________________(_______________________________)


old._____________________(_________________________, ____________)._________________________

old = nba.where('age', are.above(30)).group('prefix')
num_old = old.group('prefix')
num_old.where('count', are.below(5)).num_rows
```

(d) (5 pt) The number of positions for which the total points scored by CLE players in that position was higher than the total points scored by BOS players in that position.

```
positions = nba.pivot('prefix',_____________________________________________________)


sum(positions_____________________________________________________________________________)
positions = nba.pivot('prefix', 'position', 'points', sum)
sum(positions.column('CLE')> positions.column('BOS'))
```

2. (5 points) The Range of a Sample

The function `data_range` takes as its argument an array of numbers. The function returns the *range* of the numbers in the array, that is, the maximum value minus the minimum value.

The table `survey` consists of one row for each of the respondents to a survey. The column `Age` contains the ages of the respondents, measured in years.

Use the function `data_range` to simulate the range of the ages of a sample of size 55 drawn at random with replacement from the survey respondents. The last line should evaluate to an array consisting of 5000 simulated values of the range.

```
ranges = _________________________________

for k in _________________________________:

   simulated_range = data_range(_____________________________________________________________)

   ____________________________________________________________________________
ranges
ranges = make_array()
for k in np.arange(5000)
   simulated_range = data_range(survey.sample(55).column('Age'))
   ranges = np.append(ranges, simulated_range)
ranges
```

3. (2 points) All Children

In a population, 40% of the people are children, 50% are women, and 10% are men. The array `proportions` contains the corresponding proportions.

```
proportions = make_array(0.4, 0.5, 0.1)
```

A sample of 10 people is drawn at random with replacement from the population. The chance that all 10 people in the sample are children is one of the four options below. Fill in the bubble of the correct option.

- ◯ `proportions.item(0) ** 10`
- ◯ `proportions.item(0) * 10`
- ◯ `sample_proportions(10, proportions).item(0) ** 10`
- ◯ `sample_proportions(10, proportions).item(0) * 10`

Option 1

4. (5 points) Vaccine Effectiveness

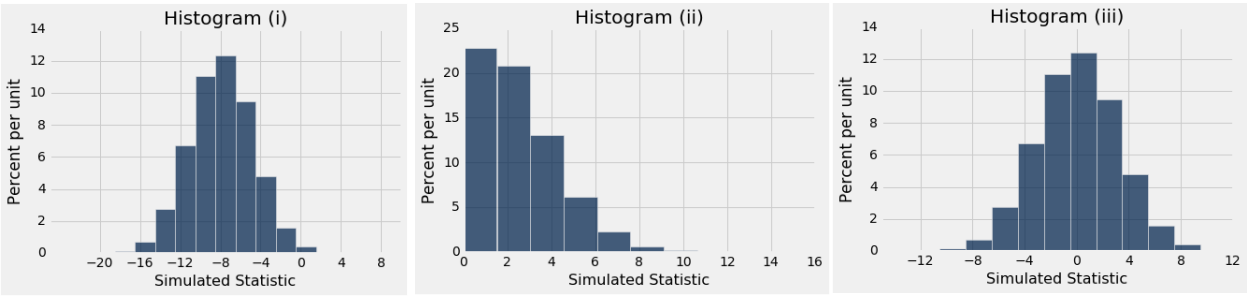
Researchers are studying the effectiveness of a particular flu vaccine. A large random sample was taken from the population of people who took the vaccine in 2016. Among the sampled people, 48% did not get the flu. Another large random sample was taken in 2017, from among the people who took the vaccine that year. Among these sampled people, 40% did not get the flu.

(a) (3 pt) A researcher thinks the vaccine was less effective in 2017 than in 2016. To test this, a null hypothesis is needed. Exactly one of the following choices is the correct null hypothesis. Fill in the bubble of the correct choice.

- ◯ The vaccine was less effective in the 2017 population than in the 2016 population, due to chance.
- ◯ The vaccine was equally effective in the two samples but its effectiveness was different in the two populations due to chance.
- ◯ The vaccine was equally effective in the two populations but its effectiveness was different in the two samples due to chance.

Option 3

(b) (2 pt) The researcher says, "The observed value of my test statistic is $40\% - 48\% = -8\%$." To perform the test, the statistic is simulated under the null hypothesis. One of the figures below is the empirical histogram of the simulated values. Which is it? Fill in the bubble of the correct histogram.



- ◯ Histogram (i)
- ◯ Histogram (ii)
- ◯ Histogram (iii)

Histogram (iii) The statistic is the difference between the two sample percents. Under the null hypothesis, this could be positive or negative depending on the sample. This rules out (ii). Under the null hypothesis, the two sample percents are expected to be equal and hence the difference is expected to be 0. This rules out (i). Only (iii) has all the right properties.

5. (10 points) Births and Days of the Week

Births in the United States are more common on weekdays than on weekends. The chart below shows the percent of births on each day of the week in the U.S. in 2016. The total number of births was 3,945,875. The chart also shows the percent of babies born on each day to the respondents of an internet survey. The number of respondents was 14,600.

Day	Sunday	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
U.S. %	9.62	14.94	16.30	16.06	16.05	15.91	11.12
Survey %	11.9	14.2	14.8	14.6	17.4	14.6	12.5

(a) (2 pt) Complete the statement below using the options given. Fill in the bubble of the day of the week that corresponds to the first blank. Next, fill in the bubble of the reasoning that should be in the second blank.

If we draw a large number of times at random from the distribution of U.S. births in the display, we expect that the largest number of sampled births will be on _____ because _____.

First Blank:

- ○ Monday
- ○ Tuesday
- ○ Wednesday
- ○ Thursday
- ○ Friday
- ○ Saturday
- ○ Sunday

Second Blank:

- ○ there are confounding factors.
- ○ random samples can be different from each other.
- ○ it is predicted by the law of averages.
- ○ association is not the same as causation.

If we draw 14,600 times at random from the distribution of U.S. births in the chart, we expect that the largest number of sampled births will be on Tuesday because it is predicted by the law of averages

(b) (3 pt) The table `births` contains the data in the chart. The first two rows are shown below.

```
births.show(2)
```

Day	US	Survey
Sunday	9.62	11.9
Monday	14.94	14.2

... (5 rows omitted)

Complete the statement below to make it a null hypothesis that can be used to test whether or not the survey results are like random draws from the distribution of births in the U.S. Fill in the bubble of the correct option to complete each of the blanks.

Null Hypothesis: The births in the survey are like _____ random draws from the distribution in _____.

First Blank:

- ○ 14,600
- ○ 3,945,875

Second Blank:

- ○ `(1/7) * np.ones`
- ○ `births.column('US')`
- ○ `births.column('Survey')`

1st blank: 14,600, 2nd blank: births.column('US')

(c) (3 pt) Fill in the line of code so that `observed_statistic` is the observed value of the test statistic that is appropriate for testing the hypotheses in Part (b) above. Assume you have access to a variable `null_proportions`, which is equal to your answer to the second blank of Part (b).

```
observed_statistic = ____________________________________________________________________
```

```
observed_statistic = sum(abs((births.column('Survey')/100)- null_proportions)))/ 2
```

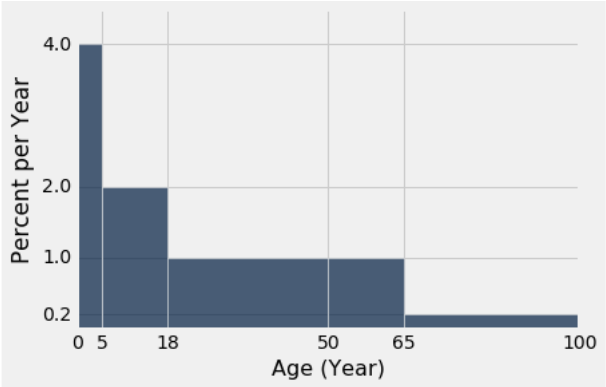
(d) (2 pt) The array `simulated_statistics` contains 10,000 values of the test statistic simulated under the null hypothesis. Complete the expression below so that it evaluates to the P-value of the test.

```
__________________________(simulated_statistics _________ observed_statistic)/ ______________
np.count_nonzero(simulated_statistics >= observed_statistic)/ 10000
```

6. (8 points) Ages

Last Sunday's San Francisco Chronicle reported data from the Centers for Disease Control about the effectiveness of the flu vaccine. The histogram below shows the distribution of the ages of the people who took the flu vaccine this year and did not get the flu.

As usual, bins include the left endpoint but not the right. The numbers on the vertical axis are the heights of the bars. For example, the height of the bar over the 65-100 bin is 0.2. Units are provided in the axis labels.



(a) (2 pt) For the following question, pick **one** of the two options to complete the statement and fill in the blank for that option.

"The percent of people in the 0-5 bin is two times the percent of people in the 5-18 bin." This quoted statement is

(i) True because _____

(ii) False because _____

(ii) False because the area of the 0-5 bar is less than the area of the 5-18 bar.

(b) (2 pt) Define school-age children to be those who are at least 5 years old but less than 18 years old. Fill in the blank with a number or arithmetic expression: _____% of the people are school-age children.

(18 -5) years × 2 percent per year = 26 percent

(c) (4 pt) Define adults to be people who are at least 18 years old. **Among the adults,** the proportion who are less than 50 years old is equal to a/b where a and b are whole numbers. Fill in the blanks with any two numbers or arithmetic expressions that result in the correct proportion. Show your work!

a = _____ b = _____

$a = 32$, $b = 54$ The area of the 18-50 bar is 32%. The area of the 0-5 bar is 20% and the area of the 5-18 bar is is 26%. So the area of the bars to the right of 50 is (100 - 46)% = 54%.