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.

<table>
<thead>
<tr>
<th>Last name</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>First name</td>
<td></td>
</tr>
<tr>
<td>Student ID number</td>
<td></td>
</tr>
<tr>
<td>CalCentral email</td>
<td>@berkeley.edu</td>
</tr>
<tr>
<td>GSI</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your left</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your right</td>
<td></td>
</tr>
<tr>
<td>All the work on this exam is my own. (please sign)</td>
<td></td>
</tr>
</tbody>
</table>

**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

The table `nba` contains data for the 2016-2017 NBA Season. All numerical values in the table are integers. All other values are strings.

The `nba` table contains 8 columns. The first few rows are shown below. Note that each prefix is an abbreviation for a team.

<table>
<thead>
<tr>
<th>player</th>
<th>prefix</th>
<th>position</th>
<th>age</th>
<th>salary</th>
<th>games</th>
<th>minutes</th>
<th>points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Al Horford</td>
<td>BOS</td>
<td>C</td>
<td>30</td>
<td>2.65401e+07</td>
<td>68</td>
<td>2193</td>
<td>952</td>
</tr>
<tr>
<td>Amir Johnson</td>
<td>BOS</td>
<td>PF</td>
<td>29</td>
<td>1.2e+07</td>
<td>80</td>
<td>1608</td>
<td>520</td>
</tr>
<tr>
<td>Avery Bradley</td>
<td>BOS</td>
<td>SG</td>
<td>26</td>
<td>8.26966e+06</td>
<td>55</td>
<td>1835</td>
<td>894</td>
</tr>
<tr>
<td>Demetrius Jackson</td>
<td>BOS</td>
<td>PG</td>
<td>22</td>
<td>1.45e+06</td>
<td>5</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>Gerald Green</td>
<td>BOS</td>
<td>SF</td>
<td>31</td>
<td>1.4106e+06</td>
<td>47</td>
<td>538</td>
<td>282</td>
</tr>
<tr>
<td>Isaiah Thomas</td>
<td>BOS</td>
<td>PG</td>
<td>27</td>
<td>6.58713e+06</td>
<td>76</td>
<td>2569</td>
<td>2199</td>
</tr>
<tr>
<td>Jae Crowder</td>
<td>BOS</td>
<td>SF</td>
<td>26</td>
<td>6.28641e+06</td>
<td>72</td>
<td>2335</td>
<td>999</td>
</tr>
<tr>
<td>James Young</td>
<td>BOS</td>
<td>SG</td>
<td>21</td>
<td>1.8252e+06</td>
<td>29</td>
<td>220</td>
<td>68</td>
</tr>
<tr>
<td>Jaylen Brown</td>
<td>BOS</td>
<td>SF</td>
<td>20</td>
<td>4.743e+06</td>
<td>78</td>
<td>1341</td>
<td>515</td>
</tr>
<tr>
<td>Jonas Jerebko</td>
<td>BOS</td>
<td>PF</td>
<td>29</td>
<td>5e+06</td>
<td>78</td>
<td>1232</td>
<td>299</td>
</tr>
</tbody>
</table>

Fill in the blanks in 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 enter anything you would like in 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'))
```

(b) (5 pt) The three-letter prefix of the team that had 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)
```
(c) (5 pt) The number of teams that have fewer than 5 players older than 30.

```
old = nba.__________________(___________________, are._________________________)
```

```
um_old = old.______________________(_______________________________)
```

```
num_old.__________________(____________________, ____________________)._______________________
```

(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(________________________________________________________________________________)
```

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
```
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.

```python
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`

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.

(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)]
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.

<table>
<thead>
<tr>
<th>Day</th>
<th>Sunday</th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
<th>Saturday</th>
</tr>
</thead>
<tbody>
<tr>
<td>Survey %</td>
<td>11.9</td>
<td>14.2</td>
<td>14.8</td>
<td>14.6</td>
<td>17.4</td>
<td>14.6</td>
<td>12.5</td>
</tr>
</tbody>
</table>

(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 chart, we expect 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, it is predicted by the law of averages, association is not the same as causation, random samples can be different from each other.

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

```
births.show(2)

<table>
<thead>
<tr>
<th>Day</th>
<th>US</th>
<th>Survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunday</td>
<td>9.62</td>
<td>11.9</td>
</tr>
<tr>
<td>Monday</td>
<td>14.94</td>
<td>14.2</td>
</tr>
</tbody>
</table>
```

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(7), births.column('US'), births.column('Survey')

(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.

```
observed_statistic = ____________________________
```

(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) / _____________
```
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.

![Histogram showing age distribution](histogram.png)

(a) (2 pt) Pick only one of the two options to complete the sentence, 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 ______________________________________________________________

(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.

(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!

\[ a = \] \hspace{0.5cm} \[ b = \]