Data 8 Foundations of Data Science
Spring 2017

PRACTICE MIDTERM SOLUTIONS

INSTRUCTIONS

- You have 45 minutes to complete the exam.
- The exam is closed book, closed notes, closed computer, closed calculator, except one hand-written 8.5" × 11" crib sheet of your own creation and the official study 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	
BearFacts 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)**	

1. (12 points) Expressions

(a) (10 pt) An array of integers named `ca` contains the (estimated) population of California every 10 years. It has 11 items. The first item is the population of California in 1900. The last is the population in 2000.

```
array([ 1485053,  2377549,  3426861, ..., 23667902, 29760021, 33871648])
```

Write a Python expression below each of the following descriptions that computes its value. The first one is provided as an example. Do not include numbers above (e.g., 1485053) in your solutions.

- The population in 1900.

  ```
  ca.item(0)
  ```

- The population change from 1940 to 1960, expressed as a number of persons (not a proportion).

  ```
  ca.item(6) - ca.item(4)
  ```

- Whether the population ever grew by less than 500000 in a decade represented by `ca`. (`True` or `False`)

  ```
  min(np.diff(ca)) < 500000
  ```

- The *annual* (yearly) growth rate from 1920 to 1930.

  ```
  (ca.item(3)/ca.item(2)) ** (1/10) - 1
  ```

- The population in 1924, assuming a fixed exponential annual growth rate from 1920 to 1930. *You may use the name `g` for the annual growth rate from 1920 to 1930 (computed above).*

  ```
  ca.item(2) * (1+g)**4
  ```

- The number of items in `ca` that are at least twice as large as the population in 1960.

  ```
  np.count_nonzero(ca >= 2 * ca.item(6))
  ```

(b) (2 pt) You have 1000 different shirts in your huge closet, but only 2 of them are red. Each morning, you pick one uniformly at random, wear it, then put it back at night. Write a Python expression to compute the chance that, during a 30-day month, you *never* wear a red shirt.

```
(998/1000)**30
```

2. (10 points) Tables

(a) (2 pt) The table named `twins` has a row for each pair of twins that contains the height of each twin in inches. The following code computes the average absolute difference in heights among the twins.

```
def diff(height1, height2):
    return abs(height1 - height2)
diffs = Table(['Absolute Differences'])
for i in np.arange(twins.num_rows):
    diffs.append([diff(twins.column(0).item(i), twins.column(1).item(i))])
np.mean(diffs.column(0))
```

Complete the expression below to compute the same result **without calling `diff` or using a `for` statement**.

```
np.mean(np.abs(twins.column(0) - twins.column(1)))
```

(b) (8 pt) Each row of the `trip` table from lecture describes a single bicycle rental in the San Francisco area. Durations are integers representing times in seconds. The first three rows out of 338343 appear below.

Start	End	Duration
Ferry Building	SF Caltrain	765
San Antonio Shopping Center	Mountain View City Hall	1036
Post at Kearny	2nd at South Park	307

Write a Python expression below each of the following descriptions that computes its value. You *may* use up to two lines and introduce variables.

- The average duration of a rental that lasted more than 2 minutes.

```
two_mins = trip.where('Duration', are.above(120))
two_mins.column('Duration').sum() / two_mins.num_rows
```

- The number of rentals that started at the `SF Caltrain` station.

```
trip.where(0, 'SF Caltrain').num_rows
```

- The name of the station where the most rentals ended (assume no ties).

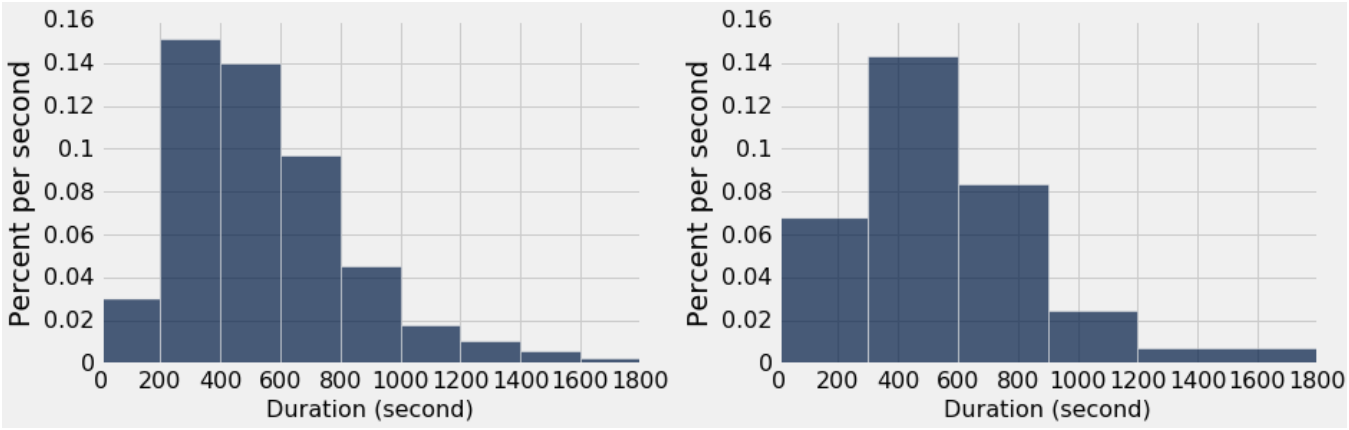
```
trip.group('End').sort('count', descending=True).column(0).item(0)
```

- The number of stations for which the average duration ending at that station was at least 300 seconds.

```
np.count_nonzero(trip.select(1, 2).group(0, np.mean).column(1) >= 300)
```

3. (11 points) Distributions

The two histograms of bike trip durations below were both generated by `trip.hist(...)` using different bins.



(a) (8 pt) Write the proportion of trips that fall into each range of durations below. *Show your work.* If it is not possible to tell from the histograms, instead write **Not enough information**.

- Between 200 (inclusive) and 400 (exclusive) seconds

  `0.0015 * 200 == 0.30` or 30%

- Between 300 (inclusive) and 900 (exclusive) seconds

  `0.0014 * 300 + 0.0008 * 300 == 0.66` or 66%

- Between 400 (inclusive) and 900 (exclusive) seconds

  `0.0014 * 200 + 0.0008 * 300 == 0.52` or 52%

- Between 200 (inclusive) and 300 (exclusive) seconds

  `0.0015 * 200 + 0.0014 * 200 - 0.0014 * 300 == 0.16` or 16%
  Alternatively: `0.0007 * 300 - 0.0003 * 200 == 0.15` or 15%

(b) (3 pt) A study followed 369 people with cardiovascular disease, randomly selected from hospital patients. A year later, those who owned a dog were four times more likely to be alive than those who didn't.

- Circle *True* or *False*: This study is a randomized controlled experiment. Answer: False; the experimenters did not control who owned a dog.

- Circle *True* or *False*: This study shows that dog owners live longer than cat owners on average. Answer: False; the experiment compares those who owned a dog to those who didn't (no mention of cats)

- Circle *True* or *False*: This study shows that for someone with cardiovascular disease, adopting a dog will probably cause them to live longer. Answer: False; an observational study does not show causation.

4. (10 points) Sampling and Hypothesis Testing

Chancellor Dirks claims to have quantified the drink consumption habits of Cory Hall Tea 1 patrons. He claims that the choice of each customer is a random sample from a distribution over five outcomes: Fruit Tea, Coffee, Milk Tea, Classic Tea, and Sparkling Juice, with probability of 0.2, 0.15, 0.2, 0.4, and 0.05, respectively.

(a) (2 pt) Guy Fieri rejects one part of the Chancellor's claim: he can't believe customers choose Classic Tea with a 40% chance. (He doesn't care at all about the other chances.) State null and alternative hypotheses that he should use to investigate the issue.

Null: Customers choose a drink at random from some distribution that gives Classic Tea a 40% chance.

Alternative: Customers choose a drink in some other way.

(b) (2 pt) Now Guy needs a sample of customer choices. Each beverage cup contains a mark describing its original contents. Should he look in the garbage can outside of Tea 1 at the end of the day and count the proportion of cups that contained Classic Tea? Why or why not?

No. That's a convenience sample. Cups are not thrown into a particular garbage according to a known random distribution.

(c) (2 pt) Alternatively, Tea 1 offers to give Guy a uniform random sample of 10 orders from its database of all past orders. He replies, "That's not enough, I need a *large* random sample." They ask why. How should he respond to justify his request of a large random sample?

Only for a large random sample will the proportions of different choices be similar to the population distribution that Guy wants to investigate. Using only a small sample, even if the alternative hypothesis is true, it will be less likely that he can reject the null hypothesis based on the sample evidence.

(d) (2 pt) Should Guy use the total variation distances (TVD) between Dirks' claimed distribution and the observed distribution as a test statistic? If not, why not?

No. TVD measures differences in any proportion, but Guy only cares about Classic Tea. A better choice of test statistic would be the absolute difference between 40% and the observed proportion of Classic Tea orders.

(e) (2 pt) Guy chooses as his test statistic the absolute difference between 40% and the observed proportion of Classic Tea orders. Tea 1 provides Guy with a random sample of 100 orders. 40 of the orders are Classic Tea. He then simulates the test statistic 100,000 times and computes a p-value. Based on this information, circle which of the following is true. Briefly justify your answer.

- (a) The null hypothesis will certainly be rejected using a 5% p-value threshold.
- (b) The null hypothesis will certainly *not* be rejected using a 5% p-value threshold.
- (c) Can't tell without actually running the simulation and looking at the empirical histogram.

Justification: (b): The observed test statistic is 0, so 100% of simulated statistics will be greater than or equal to 0. The P-value will be 1.