Data 8 Foundations of Data Science
Fall 2017

FINAL SOLUTIONS

INSTRUCTIONS

- You have 3 hours to complete the exam. Some questions are harder than others, so don't spend too long on any one question.
- 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 two official study guides 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. (16 points) Tables

In a bike share service, each bike has a unique number. The `trips` table describes the *bike* number (int), *month* number (int), and *duration* in seconds (int) of each trip in a year: each row is a trip. A bike might be involved in multiple trips. The `months` table contains the *number* (int) and *name* (string) of each month.

`trips`:

bike	month	duration
16	7	443
283	4	179
440	10	509
... (338351 rows omitted)		

`months`:

number	name
1	January
2	February
3	March
... (9 rows omitted)	

Complete the **Python expressions** below to compute each result. **You must fit your solution into the lines and spaces provided to receive full credit.** A blank can be filled with multiple expressions, such as two expressions separated by commas. The last line of each answer should evaluate to the result requested.

(a) (2 pt) The proportion of trips that lasted more than 600 seconds.

```
trips.where('duration', are.above(600)).num_rows / trips.num_rows
```

(b) (3 pt) The duration of the third shortest trip in month 3 (March).

```
trips.where('month', 3).sort('duration').column('duration').item(2)
```

(c) (3 pt) The number of bikes that were ridden for fewer than 100,000 seconds in the whole year.

```
trips.group('bike', sum).where(2, are.below(100000)).num_rows
```

(d) (4 pt) The name of the month with the fewest trips. Assume one month had fewer trips than any other (no ties).

```
t = trips.group('month').join('month', months, 'number')
t.sort('count').row(0).item('name')
```

(e) (4 pt) The number of bikes that were ridden more times in month 6 (June) than month 3 (March).

```
u = trips.pivot('month', 'bike')
sum(u.column('6') > u.column('3'))
```

2. (20 points) Sampling

Four times, you draw a simple random sample (without replacement) of 900 trip durations from the `trips` table from the previous question, which has 338,361 rows.

- The standard deviations of these 4 samples are 300, 292, 299, and 311 seconds, respectively.
- A 95% confidence interval for the mean, computed from the first sample using 10,000 resamples, is 620-660.

(a) (4 pt) Based on the SD of the first sample (300) and the sample size (900), what width should you expect for a 95% confidence interval of the mean? *Show your work & justify your answer in one or two sentences.*

$4 \times \frac{300}{\sqrt{900}} = 40$, because the SD of a large random sample is typically similar to the population SD, and the width of a 95% CI for a large random sample is typically 4 times the SD of the population divided by the square root of the sample size.

(b) (2 pt) Which of the following would be closest to the standard deviation of all 3,600 trip durations that appear in all four samples? Fill in the oval next to the correct answer.

- ○ 75
- ○ 150
- ● 300
- ○ 600
- ○ 1200

(c) (2 pt) Which of the following would be closest to a 95% confidence interval for the mean, computed from the first sample using 40,000 resamples?

- ○ 310-330
- ○ 610-670
- ● 620-660
- ○ 630-650
- ○ 635-645

(d) (2 pt) Which of the following would be closest to a 99.7% confidence interval for the mean, computed from the first sample using 10,000 resamples?

- ○ 310-330
- ● 610-670
- ○ 620-660
- ○ 630-650
- ○ 635-645

(e) (2 pt) Which of the following would be closest to a 95% confidence interval for the mean, computed from all 3,600 trip durations that appear in all four samples using 10,000 resamples.

- ○ 310-330
- ○ 610-670
- ○ 620-660
- ● 630-650
- ○ 635-645

(f) (2 pt) Which of the following would be closest to the most narrow range that must contain at least 75% of all durations in the first sample, according to Chebyshev's inequality?

- ● 40-1240

- ⬭ 340-940
- ⬭ 580-700
- ⬭ 600-680
- ⬭ 620-660

(g) **(2 pt)** Consider the original 95% confidence interval (620-660) computed from the first sample. Which of the following are correct interpretations of this interval? Fill in the oval next to **all** that are correct.

- ⬭ 95% of the trips in the sample are between 620 and 660.
- ⬭ 95% of the trips in the population are between 620 and 660.
- ⬭ There is a 95% chance that the sample mean is between 620 and 660.
- ⬭ There is a 95% chance that the population mean is between 620 and 660.
- ● None of the above

(h) **(2 pt)** Suppose that 1,000 times, we repeat the procedure to create a 95% confidence interval for the mean using 10,000 resamples (using the data from the first sample each time). About how many of these intervals do we expect will contain the mean of the first sample?

All 1,000 of them.

(i) **(2 pt)** Suppose we draw 1,000 different simple random samples of size 900 from the `trips` table. For each sample, we generate a 95% confidence interval for the mean. About how many of these intervals do we expect will contain the mean of the population?

About 950 of them.

3. (20 points) Hypothesis Testing

Looking at the `trips` table from Question 1, you wonder if bike trips have a different duration during the winter than the summer. You decide to compare the duration of trips taken in December to trips taken in July and ignore the other 10 months. Let's use a hypothesis test to perform the comparison.

(a) (4 pt) Which would be a good null hypothesis for this purpose? Fill in the oval next to **all** that are good.

- ● There is no association between the duration of the trip and whether it was taken in July or December. Any difference is due to chance.
- ○ There is an association between the duration of the trip and whether it was taken in July or December. The difference is not due to chance.
- ● The duration of trips taken in July come from the same underlying distribution as the duration of trips taken in December.
- ○ Trips taken in July have the same duration as trips taken in December.
- ○ Trips taken in July have a different duration than trips taken in December.
- ○ There is a positive correlation between the duration of the trip and the month.
- ○ The temperature outside does not affect the duration of the trip.
- ○ The temperature outside does affect the duration of the trip.

(b) (2 pt) You decide use the absolute difference between the two sample means as the test statistic. Fill in the blanks below to obtain Python code to compute this test statistic for a table `t` of trips like the one from Question 1:

```
def test_statistic(t):
    july_avg = t.where(month,  7).column('duration').mean()

    dec_avg  = t.where(month, 12).column('duration').mean()

    return abs(july_avg - dec_avg)
```

(c) (6 pt) Fill in the code below to do the hypothesis test (via the bootstrap method):

```
stats = make_array()
for i in np.arange(10000):
    months = trips.select('month')
    simulated_durations = trips.select('duration').sample(with_replacement = False
    simulated_outcomes = Table().with_columns(

        'month', months.column('month'),

        'duration', simulated_durations.column('duration'))

    simulated_stat = test_statistic(simulated_outcomes)

    stats = np.append(stats, simulated_stat)
```

(d) (4 pt) Write a Python expression to compute the P-value for this test, assuming the code above has already been run:

```
sum(stats >= test_statistic(trips)) / 10000
```

(e) **(3 pt)** In a sentence or two, write a statement defining what the P-value represents in this context.

The P-value is the chance, calculated assuming the null hypothesis, that the test statistic is equal to the value that was observed in the data or is even larger.

(f) **(2 pt)** Suppose you decide to use a P-value cutoff of 5% in the hypothesis test, and the computed P-value is `0.0078` (assuming parts (b)-(d) are correctly implemented). What should you conclude about the duration of trips in December and July?

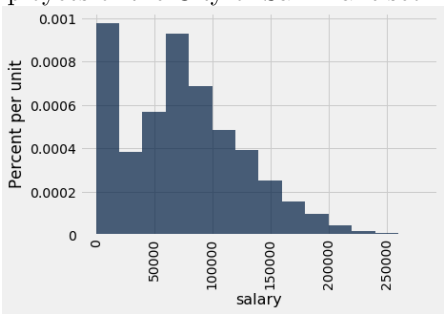
We reject the null hypothesis. There is evidence that the average duration of trips in July differs from the average duration of trips in December.

(g) **(2 pt)** Suppose you run all of the code above a second time (but with the same `trips` table). Which of the following is true? Fill in the oval next to **all** that are true.

- ○ The second time you are guaranteed to get exactly `0.0078`, the same as the first time.
- ● The second time you might not get exactly `0.0078`, but it will probably be close.
- ○ There's a significant chance that the second time you will get something very different from `0.0078`.
- ○ Which of these are true depends on the P-value cutoff you use.

4. (11 points) Variability of a Sample

A histogram of the salaries of all employees of the City of San Francisco in 2014 appears below.



The mean of the entire population is about \$75,500 and the standard deviation is about \$51,700.

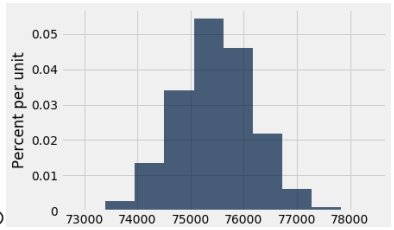
(a) (2 pt) Suppose we took a large random sample, of size 5000, from this population. What would you expect the mean of this sample to be, approximately? Fill in the oval next to the correct answer.

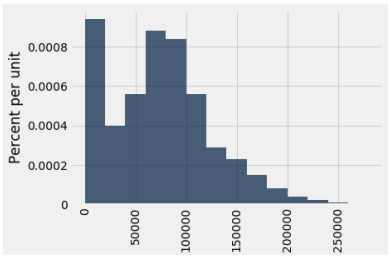
- ○ 75500×51700
- ○ $75500 \times \sqrt{51700}$
- ● 75500
- ○ $75500/\sqrt{51700}$
- ○ $75500/51700$
- ○ $75500/5000$
- ○ $75500/\sqrt{5000}$
- ○ None of the above

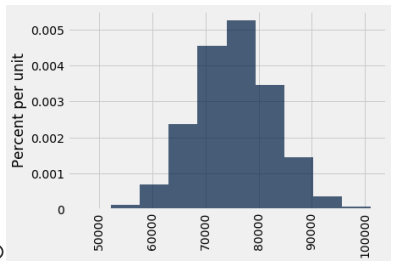
(b) (2 pt) What would you expect the standard deviation of the sample from part (a) to be, approximately?

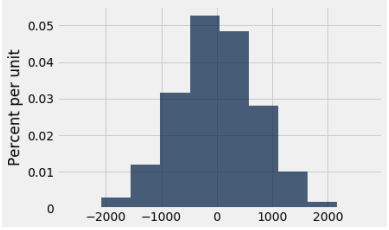
- ○ 51700×5000
- ○ $51700 \times \sqrt{5000}$
- ● 51700
- ○ $51700/5000$
- ○ $51700/\sqrt{5000}$
- ○ None of the above

(c) (2 pt) We plot a histogram of the sample from part (a). Fill in the oval next to the histogram below that is the most likely to have been generated in this way.

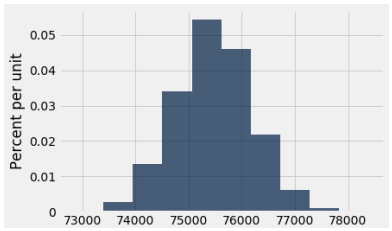
○


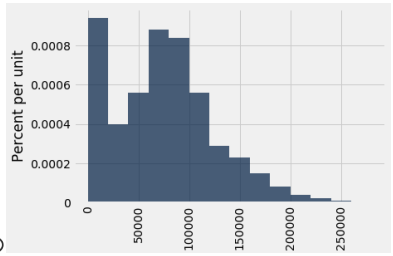
●


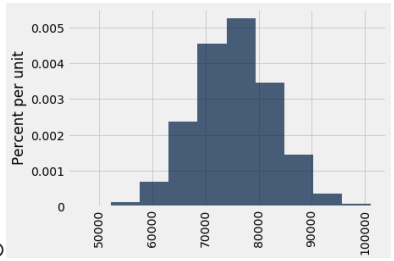
○


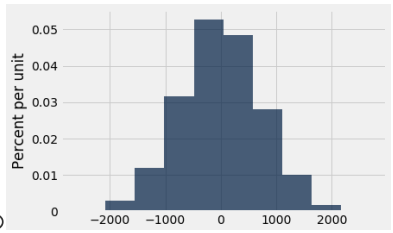
○


(d) (3 pt) Suppose we take the sample from part (a), and we resample from it with replacement. We do this 10,000 times and obtain 10,000 resamples. We compute the mean of each resample and then plot a histogram of these 10,000 means. Fill in the oval next to the histogram below that is the most likely to have been generated in this way.

● 

○ 

○ 

○ 

(e) (2 pt) Fahad, Maddy, and Vinitra are arguing about whether we can use the Central Limit Theorem (CLT) to help think about what the histogram should look like in part (d).

- Fahad believes that we can not use the CLT in part (d), as there is a large spike of people whose salary is very low in the city of San Francisco (as evidenced by the histogram of the population), so the distribution is not normal.
- Vinitra believes we can not use the CLT in part (d), since we are looking at the empirical histogram of sample means and we have no idea what that probability distribution looks like.
- Maddy believes that both of these concerns are invalid, and the CLT is helpful for part (d).

Who is right? Fill in the oval next to the best answer.

○ Fahad is right.
○ Vinitra is right.
● Maddy is right.
○ They are all wrong.

5. (3 points) True or False

Circle True or False for each of the following statements. Don't justify your answer.

(a) (1 pt) Circle True or False: With k-nearest neighbors classification, increasing the value of k will always improve accuracy on the test set.

(b) (1 pt) Circle True or False: With k-nearest neighbors classification, increasing the value of k will never improve accuracy on the test set.

(c) (1 pt) Circle True or False: With nearest neighbors classification, the test set should be selected to include some data points from the training set and some data points that aren't in the training set.

6. (2 points) Multiple choice

Fill in the bubble next to **all** samples that are a random sample of Data 8 students.

- ○ All Data 8 students who attended lab the first week
- ○ Every 10th person starting with the first person in Wheeler Hall on a random day of Data 8 lecture
- ● 200 students picked randomly from the Data 8 course roster
- ○ All sophomores in Data 8
- ○ None of the above

7. (6 points) Causality

In Fall 2018, the instructors for Data 8 decide to perform an experiment. They compile a list of all students who did not receive a passing grade on the midterm and send each of them an email with a list of the resources available to Data 8 students to learn the material and offer to meet briefly with them to discuss approaches to improve their performance in the course. At the end of the semester, the instructors measure how much these students' final exam score increased compared to their midterm score and use this to evaluate whether sending this email is helpful or not.

Circle True or False for each of the following statements. Don't justify your answer.

(a) (1 pt) True or False: This is a randomized controlled experiment.

(b) (2 pt) True or False: Due to the design of the study, confounding factors will make it impossible to determine whether contacting the students caused any improvement in final exam scores.

In Spring 2019, the instructors decide to perform another experiment. They do the same thing, but instead of contacting students who failed the midterm, they randomly select 100 students from the course roster and contact those 100 students.

Answer the following questions about the Spring 2019 experiment.

(c) (1 pt) True or False: This is a randomized controlled experiment.

(d) (2 pt) True or False: Due to the design of the study, confounding factors will make it impossible to determine whether contacting the students caused any improvement in final exam scores.

8. (12 points) Probability

Professor Wagner is putting up decorations for the holidays, and he has a bag of ornaments. In each question below, he starts with 2 red ornaments, 2 green ornaments, and 1 yellow ornament in the bag, and each ornament is chosen with equal probability. Do not show your work or justify your answer; just give the final answer.

(a) (3 pt) If he draws two ornaments with replacement, what is the probability that the second ornament drawn is green?

2/5

(b) (3 pt) If he draws two ornaments without replacement, what is the probability that both ornaments drawn are green?

2/5 * 1/4 = 2/20

(c) (3 pt) He draws two ornaments without replacement. Given that the first ornament drawn is green, what is the probability that the second ornament is red?

4/8

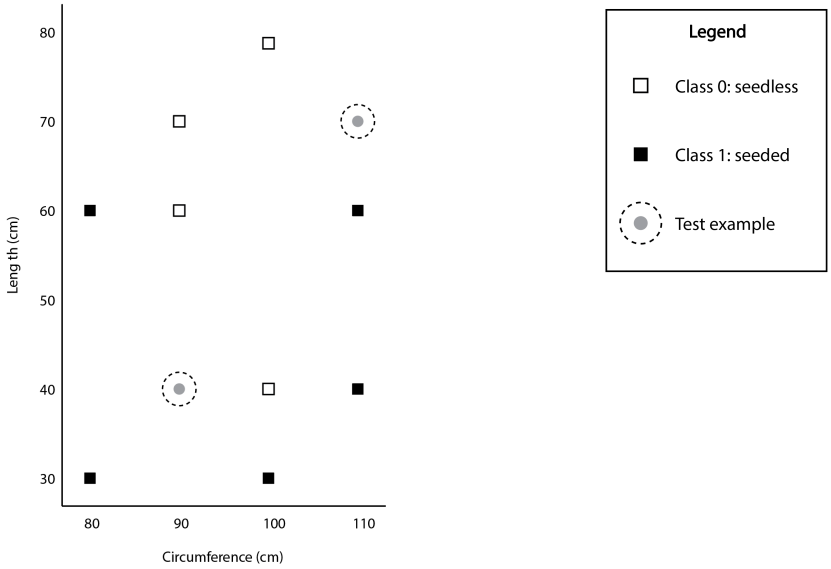
(d) (3 pt) If he draws three ornaments without replacement, what is the probability that he draws the yellow ornament within those three?

$1 - \frac{4\times3\times2}{5\times4\times3} = 3/5$

9. (8 points) Nearest-neighbor classifiers

We want to predict whether a watermelon is a seedless (class 0) or seeded (class 1) variety, based on two attributes.

We have a training set of 9 examples and a test set with 2 examples, as shown in the plot below. No two melons have the same length and circumference.



(a) (2 pt) The first watermelon in the test set has circumference 90cm and length 40cm. What class will a 1-nearest neighbor classifier predict for this watermelon?

- ● Class 0: Seedless
- ○ Class 1: Seeded

(b) (1 pt) What class will a 3-nearest neighbor classifier predict for this watermelon?

- ○ Class 0: Seedless
- ● Class 1: Seeded

(c) (1 pt) What class will a 5-nearest neighbor classifier predict for this watermelon?

- ○ Class 0: Seedless
- ● Class 1: Seeded

(d) (2 pt) Suppose we have two more watermelons: one with length 41cm and circumference 100cm, and the other with length 52cm and circumference 100cm. What is the distance between these two watermelons, as computed by a nearest-neighbor classifier that uses the length and circumference as its two attributes?

$\sqrt{(41-52)^2+(100-100)^2} = 11$

(e) (1 pt) What class will a 3-nearest neighbor classifier predict for a watermelon with length 41cm and circumference 100cm?

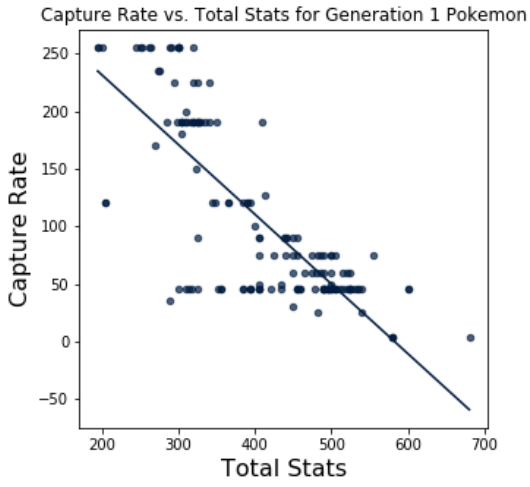
- ○ Class 0: Seedless
- ● Class 1: Seeded

(f) (1 pt) What class will a 3-nearest neighbor classifier predict for a watermelon with length 52cm and circumference 100cm?

- ● Class 0: Seedless
- ○ Class 1: Seeded

10. (24 points) Regression

This scatter plot shows a population of pokemon. For each pokemon, we plot its total stats (a measure of its power) and capture rate (which affects how likely you are to catch the pokemon when you throw a pokeball at it). The plot also shows the regression line through this data.



The mean of total stats is 407.1, and the standard deviation of total stats is 99.4. The mean of capture rate is 106.2, and the standard deviation is 76.9. The correlation coefficient is -0.78.

In the parts below, it is OK to write your answer as a Python expression that evaluates to the correct answer.

(a) (1 pt) Fill in the oval next to the correct statement:

- ○ There appears to be a positive association.
- ● There appears to be a negative association.

(b) (2 pt) Pikachu has a total stats of 320. What is Pikachu's total stats, in standard units?

$(320 - 407.1)/99.4$

(c) (2 pt) Charmander has a capture rate of 45. What is Charmander's capture rate, in standard units?

$(45 - 106.2)/76.9$

(d) (1 pt) Fill in the oval next to the correct statement:

- ○ The slope of the regression line is greater than zero.
- ● The slope of the regression line is smaller than zero.

(e) (2 pt) Calculate the slope of the regression line, in standard units.

-0.78

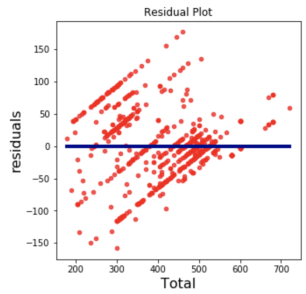
(f) (3 pt) Calculate the slope of the regression line, in original units.

$-0.78 \times 76.9/99.4$

(g) (2 pt) Suppose we encounter a new pokemon with total stats 200. If we use the regression line to predict the capture rate of this pokemon, which of the following could plausibly be the prediction? Fill in the oval next to the correct answer.

○ 253 ○ 122 ● 231 ○ 280 ○ 120

(h) (2 pt) Below is the residual plot for the linear regression fit on all the data.



Does the association between total stats and capture rate appear linear? Justify your answer in one or two sentences.

No. There is a definite pattern in the residuals: they are negative on the left side and positive on the right side.

In parts (i)-(k), assume that there is a true linear relationship between capture rate and total stats, and that each data point was generated by finding a point on the true line and then adding random error to the capture rate according to the model described in lecture.

(i) (2 pt) Suppose we want to test whether or not the slope of this true line is actually 0, based on the data we have. Specify a null hypothesis and alternative hypothesis we can use to perform a hypothesis test of this question.

Null hypothesis:

The slope of the true line is 0.

Alternative hypothesis:

The slope of the true line is not 0.

(j) (4 pt) Assume the data are in a table called `pokemon`:

```
name       | total stats | capture rate
pikachu    | 320         | 190
charmander | 309         | 45
... (703 rows omitted)
```

Assume a `slope(t, col1, col2)` function is defined for you. Its arguments are a table `t`, the label `col1` of the column for the x values, and the label `col2` of the column for the y values. The return value is the slope of the regression line for the data in the table in original units.

Fill in the blank below to write Python code to perform the hypothesis test mentioned in part (i).

```
slopes = []
for i in np.arange(10000):

    s = slope(pokemon.sample(), 'total stats', 'capture rate')
    slopes.append(s)
m = slope(pokemon, 'total stats', 'capture rate')
print ('95% confidence interval for the slope:')
print(percentile(2.5, slopes), percentile(97.5, slopes))
```

(k) (2 pt) Suppose we encounter a new pokemon with total stats 450. We generate a 95% prediction interval for the predicted capture rate of this pokemon using the bootstrap method (i.e., we resample the data 10,000 times, find a regression line for each resample, and use them to make 10,000 predictions). Assume this interval is 80.42-87.03. How should we interpret this interval? Fill in the oval next to **all** correct answers.

- ● We are 95% confident that the height of the true line at x=450 is between 80.42 and 87.03.
- ○ We are 95% confident that the height of the regression line for all the data (the one shown at the start of the question) at x=450 is between 80.42 and 87.03.
- ○ We are 95% confident that the slope of the true line is between 80.42 and 87.03.
- ○ We are 95% confident that the slope of the regression line for all the data (the one shown at the start of the question) is between 80.42 and 87.03.
- ○ None of the above

(l) (1 pt) Circle True or False: Based on the plot at the start of the question, the capture rates appear to be normally distributed.

11. (0 points) (Optional) Go Bears!

Draw a data visualization about UC Berkeley.