S&DS 123 YData: An Introduction to Data Science
Spring 2019

FINAL Solutions

INSTRUCTIONS

- You have three hours to complete the exam.
- The exam is closed book, closed notes, closed computer, closed phone, and closed calculator, except one handwritten 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.

First name	
Last name (Surname)	
NetID	

This page was intentionally left blank.

Problem	Score	Out of
1		10
2		14
3		6
4		14
5		4
6		8
7		10
8		8
9		14
10		2
Total		90

1. **Tables (10 points)**

   Each row of the `trips` table describes a single bicycle rental in the San Francisco area. Durations are integers representing times in seconds. The first three rows out of 338,343 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 Python expressions that compute each of the following values. You may use up to two lines and introduce variables, and assume `numpy` is imported as `np`.

   (a) The average duration of a rental

```
np.mean(trips.column('duration'))
```

   (b) The average duration of a rental that lasted more than 300 seconds (5 minutes)

```
five_mins = trips.where('duration', are.above(300))
np.mean(five_mins.column('duration'))
```

   (c) The number of rentals that started at the `SF Caltrain` station

```
trips.where('start', 'SF Caltrain').num_rows
```

   (d) The average duration for rentals that started and ended at the same station

```
np.mean(trips.where(trip.column(0) == trip.column(1)).column(2))
```

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

```
trips.group('end').sort('count', descending=True).column(0).item(0)
```

2. **Hypothesis testing (14 points)**

   (a) Following the analysis of a well-designed completely randomized experiment it was reported that the observed effect had a p-value of 1% and the researchers rejected the null hypothesis. Which of the following statements best explains the conclusion of the researchers? *Circle all correct answers.*

      i. The observed result made sense to the experimenter since it was what they hoped would happen.

ii. The observed effect happened because the experiment was properly designed and carried out without bias.
iii. The experimenter carefully employed the basic principles of experimental design in conducting the study.
iv. The laws of probability say that this observed result would be expected to happen by chance.
v. The observed effect was sufficiently large so that it would rarely occur simply by chance.

Answer: v

(b) A hypothesis test is going to be carried out by researchers, and they have decided to use a p-value cutoff of 8%. If the null hypothesis is actually true, what is the probability that the test reaches the correct conclusion? *Circle all correct answers and explain your choice(s).*

i. 0%
ii. 4%
iii. 8%
iv. 92%
v. 96%
vi. 100%
vii. None of the above.

Explanation:

92% is the correct answer. Using a p-value cutoff of 8% means that when the null hypothesis is true, there is an 8% chance that we will incorrectly reject the null hypothesis. The remaining 92% would not reject the null, which would be a correct conclusion of the test.

(c) A study on the effect of caffeine involved asking subjects to take a memory test 20 minutes after drinking cola. The memory test had a total of 25 items on it. Some subjects were randomly assigned to drink caffeine-free cola, and some to drink regular cola (with caffeine). For each subjects, a test score (the number of items recalled correctly) was recorded. The subjects were not told which type of cola they had been given. Each row of the table **cola** contains the `Type` and `Test score` for a subject in the study. The first four rows out of 120 appear below.

Type	Test Score
Caffeine-free	20
Regular	12
Regular	21
Regular	15

- The researchers would like to carry-out a hypothesis test using these data. Suppose the researchers think the regular cola (with caffeine) group will perform better on the memory test. State an appropriate null and alternative hypothesis.
  Null: there is no difference in the average performance on the memory test between those who consume regular or caffeine-free cola. Alt: the average performance on the memory test is higher for those who consume regular cola.
- What is an appropriate test statistic to use in this setting? Explain. (You can write your solution as Python code or as an equation.)
  Test statistics that can detect if the average test score for the regular group is larger than for the caffeine-free group. For example, mean(score regular group) - mean(score caffeine-free group), and then large positive differences favor the alternative hypothesis.
- A permutation test is carried out to test the hypothesis. Fill in the blanks below with Python expressions for carrying-out this test.

`group_labels = ______________________________`

`scores = ______________________________`

```
test_stats = ____________________________

for i in np.arange(20000):

    shuffled_scores = scores.sample(with_replacement = ___________).column( ______________)

    shuffled_tbl = group_labels.with_column(‘Shuffled Scores’, shuffled_scores)

    means_tbl = shuffled_tbl.group(‘Type’, ____________________________)

    score_avgs = means_tbl.column( ____________________________)

    new_test_stat = ____________________________

    test_stats = ____________________________
```

Answer:

```
group_labels = cola.select(‘Type’)
scores = cola.select(‘Test Score’)
test_stats = make_array()
for i in np.arange(20000):
    shuffled_scores = scores.sample(with_replacement = False).column(‘Test Score’)
    shuffled_tbl = group_labels.with_column(‘Shuffled Scores’, shuffled_scores)
    means_tbl = shuffled_tbl.group(‘Type’, np.average)
    score_avgs = means_tbl.column(‘Shuffled Score average’)
    new_test_stat = score_avgs.item(1) - score_avgs.item(0) (consistent with test statistic
in previous question)
    test_stats = np.append(test_stats, new_test_stat)
```

- Using the array `test_stats` from your permutation test above, fill in the blanks below so the last line evaluates to the left and right endpoints of an approximate 99% confidence interval.

  ```
  left_end = ____________________________________

  right_end = ____________________________________

  make_array(left_end, right_end)
  ```

  Answer:

  ```
  left_end = percentile(.5, differences)
  right_end = percentile(99.5, differences)
  make_array(left_end, right_end)
  ```

- If you want to carry out a hypothesis test using your approximate 99% confidence interval above, what would have to happen for you to reject the null hypothesis using a p-value cutoff of 1%?
  Answer: If 0 is not in the approximate 99% confidence interval, you would reject the null hypothesis with a p-value cutoff of 1%.

3. **Confidence Intervals (6 points)**

   (a) What do we hope to capture within a confidence interval? **Circle all correct answers.**

   The unknown confidence level The parameter estimate The unknown statistic

   The unknown parameter The margin of error The sample size

   None of the above

   The unknown parameter

(b) If you reject the null hypothesis H_0 : mean $= x_0$ vs. the alternative H_1 : mean $\neq x_0$ with a p-value cutoff of 0.05, then x_0 would fall in the 90% confidence interval for the mean. (True or False or Not Enough Information)

Answer: ______________________________

False

(c) The confidence interval for a population mean is narrower for smaller confidence level C with everything else the same. (True or False or Not Enough Information)

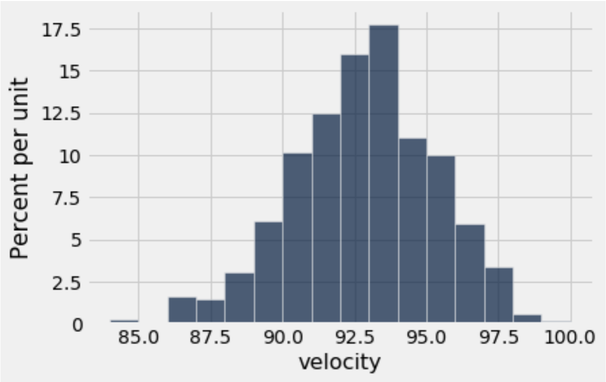
Answer: ______________________________

True

4. **Sampling (14 points)**

Data are collected on the average pitch velocities of all 700 Major League Baseball pitchers during the 2018 season. The first 10 rows of the data are shown below on the left; a histogram of all of the velocity data is shown on the right (the units are miles-per-hour).

index	player	velocity
1	Justin Verlander	94.6
2	Zack Greinke	89.2
3	J.A. Happ	91.9
4	Kevin Gausman	93.1
5	Mike Clevinger	93.5
6	Sean Newcomb	92.9
7	Jose Quintana	91.5
8	Luis Severino	97.3
9	Jon Lester	90.8
10	Max Scherzer	93.9
...	...	...



In the following questions, the 700 velocities are considered to be the true population.

(a) Which of the following is the best estimate of the mean of the distribution? **Circle your response.**

i. 2.5
ii. 5
iii. 90
iv. 93 Answer
v. 95

(b) Which of the following is the best estimate of the standard deviation of the distribution? **Circle your response.**

i. 1
ii. 2.5 Answer
iii. 5

iv. 10

v. 93

(c) Does the distribution approximately have a normal distribution? Answer yes or no, and explain.
Yes, the shape is roughly bell-shaped and with a mean of 93 and standard deviation of 2.3

(d) Suppose that you have a sample of 100 pitchers from this population and compute the average of their velocities to be 93 with a standard deviation of 2.5. Construct a 95% confidence interval for the average velocity of the population. Show your work.

$$\left(93 - 2 \times \frac{2.5}{\sqrt{100}}, 93 + 2 \times \frac{2.5}{\sqrt{100}}\right) = (92.5, 93.5)$$

(e) Suppose that we have 100 random samples from this population, each containing 50 pitchers. A 95% confidence interval for the average population velocity is computed for each of these samples. We expect that the true average of the population will fall into how many of these intervals? **Circle your response.**

i. 0

ii. all of them

iii. 95 Answer

iv. 97.5

v. It is impossible to predict given the available information

(f) Suppose that we have 100 random samples from this population, each containing 50 pitchers. A 95% confidence interval for the average population velocity is computed for each of these samples. We expect that the sample average will fall into how many of these intervals? **Circle your response.**

i. 0

ii. all of them Answer

iii. 95

iv. 97.5

v. It is impossible to predict given the available information

(g) Suppose that we have 100 random samples from this population, each containing 50 pitchers. A 95% confidence interval for the average population velocity is computed for each of these samples. We expect that the average velocity of the population of pitchers from the 2019 season (the season after the season where the data were collected) will fall into how many of these intervals? **Circle your response.**

i. 0

ii. all of them

iii. 95

iv. 97.5

v. It is impossible to predict given the available information Answer

5. **Sample size (4 points)**

(a) To assess the accuracy of a laboratory scale, a standard weight that is known to weigh exactly 1 gram is repeatedly weighed a total of n times and the sample mean $\bar{x}$ is computed. Suppose the scale readings are normally distributed with unknown population mean μ and population standard deviation $\sigma = 0.01$ g. How large should n be so that a 95% confidence interval for the mean is no wider than 0.0002?
$0.0002 = \frac{4\times 0.01}{\sqrt{n}} \implies n = \left(\frac{4\times 0.01}{.00002}\right)^2 = 40,000$ (numerical answer is not required)

(b) Suppose that we wish to estimate the fraction p of the vote that a political candidate has obtained, by counting a random sample of the ballots that have been cast. Which of the following constraints on the sample size ensure that a 95% confidence interval constructed from the random sample will have a width no larger than one-half of one percent (0.005)? **Circle each statement that is true.**

i. Sample size smaller than 2000
ii. Sample size at least 1600
iii. Sample size at least 2000
iv. Sample size at least 400
v. No sample size will guarantee this

$n \geq \left(\frac{4\cdot\frac{1}{2}}{.005}\right)^2 = 160,000$, none of these answers are correct [this was an error on the exam, (v.) should have been 'None of the above']

6. **Causality (8 points)**

Measles was declared eliminated in the United States in 2000 but the highly contagious disease has returned in recent years in communities with low vaccination rates. Outbreaks have been largely confined to regions with low immunization rates. A non-profit organization decides to perform a public service experiment. They compile a list of all adults in the greater New York region and send a mailing offering for free a new measles vaccine to the first 1,000 respondents. Those that respond and receive the vaccine form the treatment group. The second 1,000 respondents form a control group, and do not receive the vaccine. A year later the non-profit contacts both the treatment and control groups to ask whether measles was contracted. The results are used to estimate the effectiveness of the vaccine.

Circle True or False for each of the following statements. Justify each answer.

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

False. It is not randomized as it is biased towards adults that respond quickly to free vaccine offers on a volunteer basis. It is also not a controlled experiment as the subjects are not randomly assigned to the treatment and control groups.

(b) True or False: Due to the design of the study, confounding factors will make it impossible to determine whether the new vaccine is effective.

True. An example of such a factor would be a culture in favor of vaccines. This could lead one to quickly accept a new vaccine, while already being immune to measles and therefore not contracting it.

Subsequently, the non-profit decides to perform another experiment. This time they compile a list of all adults in the U.S., and they now choose a *random subset* of 1,000 adults from this list and mail them an offer for a free measles vaccine. Those that respond and receive the vaccine form the treatment group. Another random set of 1,000 adults is chosen as a control group, and do not receive the vaccine. A year later the non-profit contacts both the treatment and control groups to ask whether measles was contracted. The results are used to estimate the effectiveness of the vaccine.

Circle True or False for each of the following statements. Justify each answer.

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

False. The study is still not fully randomized as there is non-response bias present. Many of those who respond to the mail offer are likely in favor of vaccines already and therefore are not likely to contract measles with or without taking the new vaccine.

(d) True or False: Due to the design of the study, confounding factors will make it impossible to determine whether the new vaccine is effective.

True. Same example as before

7. **Regression (10 points)**

(a) The least-squares regression line is

i. the line that makes the sum of the squares of the vertical distances of the data points to the line as small as possible

ii. the line that best splits the data in half, with half of the points above the line and half below the line.

iii. the line that makes the square of the correlation in the data as large as possible.

iv. all of the above

v. i and ii

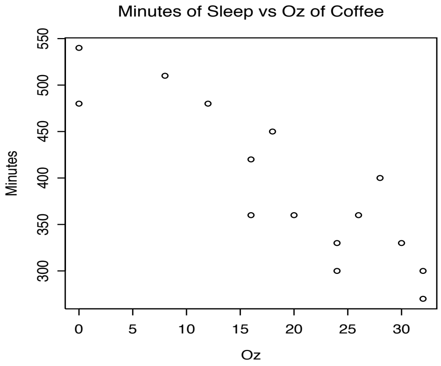
vi. i and iii

vii. ii and iii

viii. None of the above

i. the line that makes the sum of the squares of the vertical distances of the data points to the line as small as possible

(b) The following scatterplot represents the relationship between the number of ounces of coffee consumed and the number of minutes of sleep per day.

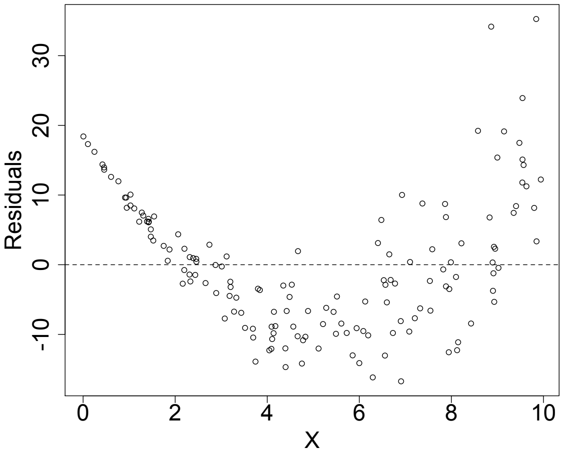


Variable	Mean	Stand. Deviation
Ounces of coffee	$\bar{x} = 19.07$	$s_x = 10.53$
Minutes of sleep	$\bar{y} = 392.67$	$s_y = 83.7$

- Given $r^2 = 0.7756$ (the estimated correlation squared) and the summary information in the table above, find the Least Squares regression Line for predicting the amount of sleep in minutes using ounces of coffee consumed.
  $r = -\sqrt{0.7756} = -0.88$, slope $= r \times \frac{83.7}{10.53} \approx -7$, intercept $= 392.67 -$ slope $\times\ 19.07 \approx 526$

  minutes of sleep = slope × ounces of coffee + intercept
- One of the points on the scatterplot has the coordinates (12 oz, 480 min.), calculate the residual value at 12 oz for this observation.
  $480 -$ (slope $\times\ 12 +$ intercept) ≈ 38

(c) Suppose a Least-squares line was fit on explanatory variable X and response variable Y, with the residuals plotted in the figure below against X. Which linear model assumptions, if any, appear to be violated given the residual plot below?



The assumption of linearity and the assumption of constant variance are violated

8. **Probability (8 points)**

   (a) Suppose you have three tickets: Red, Green, and Blue. What is the chance that you would draw a Blue ticket and then a Green ticket if you sample **without** replacement?
   Answer: $P(BG) = (\frac{1}{3}) \times (\frac{1}{2}) = 1/6$. You could also note the six different options when drawing two tickets without replacement: RG, RB, GR, GB, BR, BG. BG is one of the six options.

   (b) Suppose you have three tickets: Red, Green, and Blue. What is the chance that you would draw a Blue ticket and then a Green ticket if you sample **with** replacement?
   Answer: $P(BG) = (\frac{1}{3}) \times (\frac{1}{3}) = 1/9$. You could also note the six different options when drawing two tickets without replacement: RR, RG, RB, GG, GR, GB, BB, BR, BG. BG is one of the nine options.

   (c) Three students work independently on a homework problem. The probability that the first student solves the problem is 0.75. The probability that the second student solves the problem is 0.5. The probability that the third student solves the problem is 0.9. What is the probability that exactly one of the three students solves the problem correctly?
   (0.75)(1-0.5)(1-0.9) + (1-0.75)(0.5)(1-0.9) + (1-0.75)(1-0.5)(0.9)

   (d) You roll a fair 6-sided die 10 times. What is the chance that every roll is less than or equal to 5?
   Answer: $(\frac{5}{6})^{10}$

9. **Classification (14 points)**

   The k-nearest neighbor classifier can be described as a 2-step procedure. The first step is as follows.

   Step 1: To classify test point p, first find the k training points that are closest to p, and determine their classes.

   (a) What is the second step, in words?
   Find the most common classification among the k closest training points.

   Here is the template of code to implement the k-nearest neighbor classifier, where `step2` implements the second step:

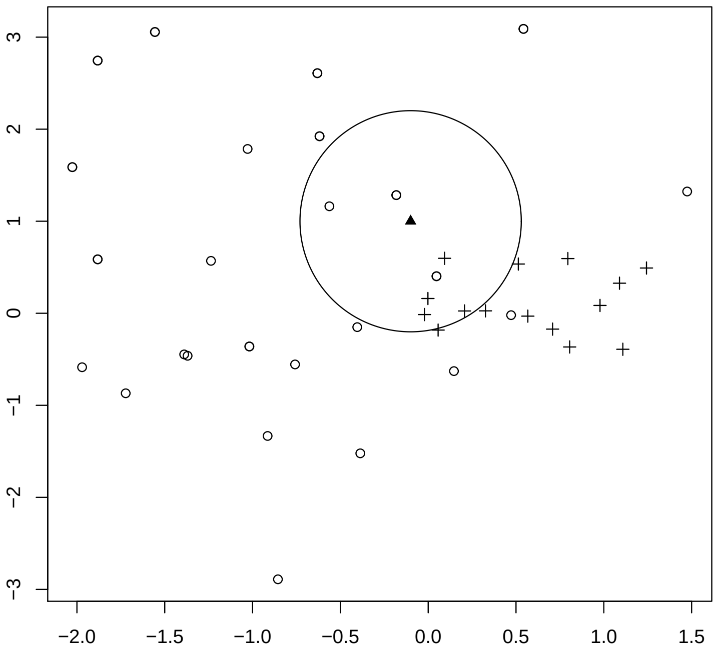
```
def classify(training, p, k):
    kclosest = closest(training, p, k)
    topkclasses = kclosest.select('Class')
    return step2(topkclasses)
```

   (b) Write Python code to implement the second step.

```
def step2(topkclasses):
```

```
topkclasses.group('Class').sort('count', descending=True).column(0).item(0)
```

(c) The following figure shows a training data set.



The solid triangle in the center of the circle is the test point p, which is not contained in the training data. The circles are the training points labeled '0' and the plus signs are the training points labeled '1'.

i. What is the 1-nearest neighbor prediction for the class of p ('0' or '1')? 0

ii. What is the 3-nearest neighbor prediction for the class of p ('0' or '1')? 0

iii. What is the 7-nearest neighbor prediction for the class of p ('0' or '1')? 1

(d) When a k-nearest neighbor classifier is applied to an independent set of test points, the error rate generally satisfies which of the following:

A. The test error is higher than the error rate of the classifier on the training data Answer

B. The test error is about the same as the error rate of the classifier on the training data

C. The test error is lower than the error rate of the classifier on the training data

Circle the correct answer above, and ***explain your answer below*****.**

Since the test data are not used to build the model, but the training data are, the test error rate is generally higher than the training error rate.

10. **Distributions (2 points)**

A city has 500,000 households. The annual incomes of these households have an average of $65,000 and an SD of $45,000. The distribution of the incomes: **Circle your response.**

(a) is roughly normal because the number of households is large.

(b) is not close to normal.

(c) may be close to normal, or not; we can't tell from the information given.

(ii) The distribution does not follow the 68-95-99.7 rule for normal distributions.

Thank you for being a part of the inaugural YData class - we hope you have a wonderful summer!