Data 8 Foundations of Data Science
Spring 2017

FINAL SOLUTIONS

INSTRUCTIONS

- You have 3 hours 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 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)**	

Omissions from the study guide:

Expression	**Description**
`np.std(array)`	Standard deviation of an array of values.
`np.median(array)`	Median of an array of values.

1. (13 points) Tables

The `cal` table describes the *name* (string), *position* (string), *class* (string), and *height* (int) of Cal basketball players in the 2016-17 season.

```
name           | position | class     | height
Ivan Rabb      | Forward  | Sophomore | 83
Charlie Moore  | Guard    | Freshman  | 71
... (15 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; you never need to call `print`.

(a) (2 pt) The proportion of all players whose position is `Forward`.

```
cal.where('position', 'Forward').num_rows / cal.num_rows
```

(b) (2 pt) The name of the shortest `Freshman`. Assume that one is shorter than the rest.

```
cal.where('class', 'Freshman').sort('height').row(0).item('name')
```

(c) (3 pt) Whether there are at least $\frac{3}{4}$ of players that are 80 inches tall or shorter. The result should be `True` or `False`.

```
percentile(75, cal.column('height')) <= 80
```

(d) (3 pt) The number of players that are (strictly) more than one standard deviation above the mean height.

```
a = cal.column('height')
cal.where('height', are.above(np.mean(a) + np.std(a))).num_rows
```

Another common correct answer:

```
a = np.mean(cal.column('height')) + np.std(cal.column('height'))
cal.where('height', are.above(a)).num_rows
```

(e) (3 pt) An array of all positions, sorted in increasing order of the average height for all players in that position.

```
t = cal.select('position', 'height')
```

```
t.group('position', np.average).sort(1).column(0)
```

2. (18 points) Experiments

The `cal` table described in the previous question has columns *name*, *position*, *class*, and *height*. Read the following code used to test a hypothesis about the `cal` table, then answer the questions below to interpret it.

```
def diff_of_means(t):
    forward_mean = t.where('position', 'Forward').column('height').mean()
    guard_mean   = t.where('position', 'Guard').column('height').mean()
    return forward_mean - guard_mean

differences = make_array()
for i in np.arange(10000):
    shuffled_heights = cal.sample(with_replacement=False).column('height')
    shuffled = cal.select('position').with_column('height', shuffled_heights)
    differences = np.append(differences, abs(diff_of_means(shuffled)))
```

(a) (4 pt) Circle all of the following hypotheses that could potentially be tested using the `differences` array.

(A) (Correct) Among guards and forwards on this team, there is an association between height and position.
(B) Whether a player is a guard or a forward is like flipping a fair coin.
(C) The forwards on Cal's team have historically been taller than the guards, on average.
(D) (Correct) The heights of guards and forwards are like random samples from the same distribution.

(a) This permutation test is used to test association. This is the alternative. (b) The differences array does not include counts of positions. (c) The height of current players doesn't tell you about historical trends. (d) This is the main assumption that motivates the null hypothesis.

(b) (4 pt) Circle one option among (A), (B), and (C) for each blank in this description of the hypothesis test:
In this ____(i)____, the null hypothesis states that the heights and positions of players are ____(ii)____.

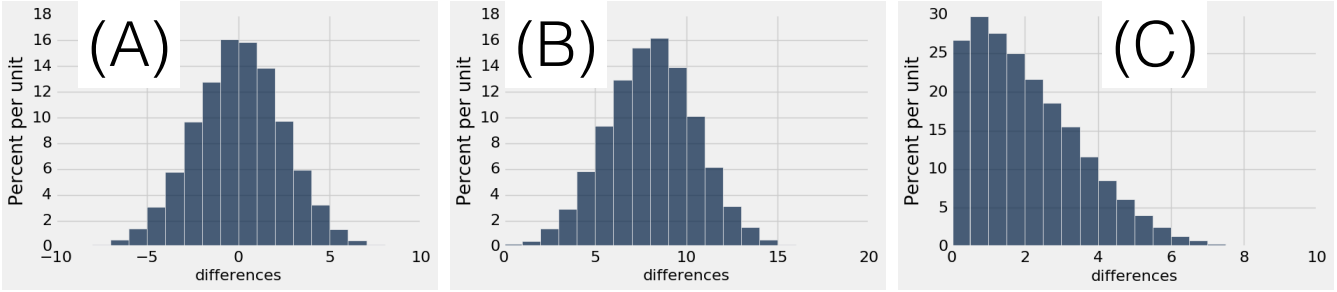
(i): (A) (Correct) permutation test
(B) confidence interval test
(C) bootstrap resampling test

(ii): (A) drawn at random from the same population
(B) (Correct) paired up at random
(C) normally distributed

Note: heights and positions are not drawn from the same distribution. Instead, heights are drawn from the same distribution for each position.

(c) (2 pt) What test statistic is being used to test the null hypothesis in (b)? Describe it in English, not code.
The (absolute) difference in means between guards and forwards.

(d) (2 pt) Circle the letter for the chart below that could plausibly be a histogram of the `differences` array.



(C); There's an absolute value in the test statistic, and random differences in means for a permutation test will mostly be near zero. We get the above-mean half of a normal distribution.

(e) (4 pt) Write a Python expression to compute a P-value for this test, using the null hypothesis you defined in part (b) and the test statistic you described in part (c).

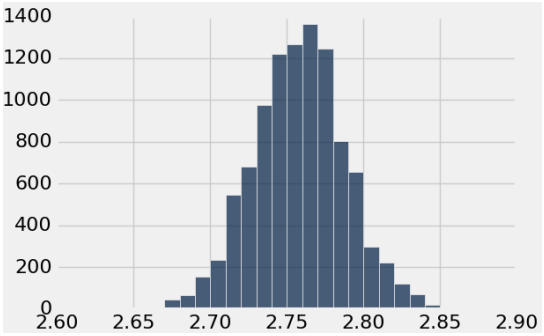
```
np.count_nonzero(differences >= abs(diff_of_means(cal)))/10000
```

(f) (2 pt) Using a P-value cutoff of 5% to determine significance, the null hypothesis from (b), the test statistic from (c), and the null distribution from (d), what should you conclude if the observed value of the test statistic is 6?
We would reject the null hypothesis because only about 1.5% of the differences are 6 or greater. Therefore, we conclude that there is an association between height and position among guards and forwards on this team.

3. (14 points) Sampling

This histogram shows sample means for 2,500 random samples. Each sample contains 10,000 trip distances, measured in miles, drawn at random from the distances of 1.4 million trips for New York taxis in January 2016.

```
sample_means.hist(bins=np.arange(2.65, 2.9, 0.01))
```



(a) (2 pt) What quantity is measured by the horizontal axis of this histogram?

(A) Total miles for a single randomly chosen trip
(B) Total miles for a single randomly chosen sample
(C) (Correct) Average miles for 10,000 randomly chosen trips
(D) Average miles for 2,500 randomly chosen samples
(E) None of the above

(b) (2 pt) What quantity is measured by the vertical axis of this histogram?

(A) Percent of trips per sample
(B) Percent of trips per mile
(C) Percent of trips per sample mean
(D) Percent of sample means per trip
(E) (Correct) Percent of sample means per mile

(c) (2 pt) The percent of sample means represented by the tallest bar (with height about 1400) is closest to:

(A) 1.4 percent
(B) 7 percent
(C) (Correct) 14 percent
(D) 28 percent
(E) 70 percent

(d) (2 pt) How would the height of the tallest bar change if we drew 10,000 random samples instead of 2,500?

(A) Grow by about 2 times
(B) Grow by about 4 times
(C) Shrink by about 2 times
(D) Shrink by about 4 times
(E) (Correct) Not much change

Increasing samples makes the empirical distirbution evem more similar to the probability distribution, but the shape stays the same since there are lots of samples under both scenarios.

(e) **(2 pt)** How would the height of the tallest bar change if the bars all had width 0.02 instead of 0.01? The new histogram would be generated by `sample_means.hist(bins=np.arange(2.65, 2.9, 0.02))`.

(A) Grow by about 2 times
(B) Grow by about 4 times
(C) Shrink by about 2 times
(D) Shrink by about 4 times
(E) (Correct) Not much change

Since the histogram is drawn to the density scale, bar heights do not increase with the width because the density remains the same.

(f) **(2 pt)** The standard deviation of taxi trip distances in the population of 1.4 million trips is closest to:

(A) 0.01 miles
(B) 0.03 miles
(C) 0.1 miles
(D) 0.3 miles
(E) (Correct) 3 miles

The SD of the distribution of sample means is about 0.03. You can see this from the inflection point of the normal curve, which is about 0.03 away from the mean. The sample mean SD is (Population SD) / $\sqrt{10000}$, so the population SD is around $0.03 \times \sqrt{10000} = 3$

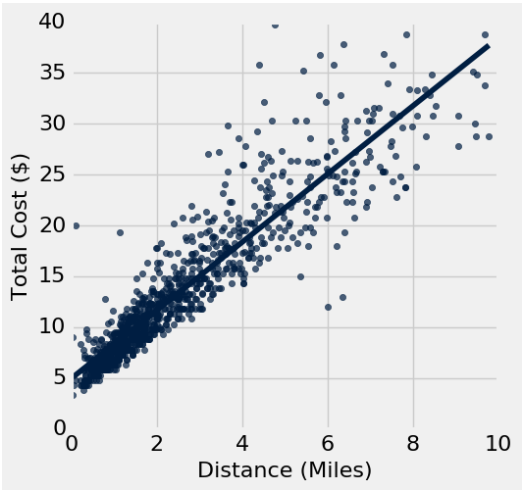
(g) **(2 pt)** In order to construct a 95% confidence interval for the mean trip distance in the population, such that the width of the interval is 0.4 miles or less, the minimum sample size required is closest to:

(A) 9
(B) 90
(C) (Correct) 900
(D) 9,000
(E) 90,000

A 95% confidence interval of a normally-distributed statistic (such as a sample mean) will extend from 2 SDs below to 2 SDs above the mean: 4 SDs total. So, a width of 0.4 will come from a distribution of the statistic with an SD of 0.1. Since the population SD is 3, a sample of size 900 will have an SD of $3/\sqrt{900} = 0.1$.

4. (22 points) Linear Regression

This scatter plot of a sample of 1,000 trips for New York taxis in January 2016 compares distance and cost. The regression line is shown. Two trips of the same length can vary in cost because of waiting times, special fees, taxes, tolls, tips, discounts, etc.




```
np.average(t.column('Distance'))    = 3
np.std(t.column('Distance'))        = 2
np.average(t.column('Cost'))        = 13
np.std(t.column('Cost'))            = 6
correlation(t, 'Distance', 'Cost') = 0.9
```

(a) (3 pt) Convert a trip total cost of $9 to standard units.
(9-13)/6 = -2/3

(b) (3 pt) What is the slope of the regression line for this sample in dollars per mile?
0.9 * 6 / 2 = 2.7

(c) (3 pt) What is the intercept of the regression line for this sample in dollars?
13 - 2.7 * 3 = 4.9

(d) (3 pt) If instead we fit a regression line to estimate distance in miles from total cost in dollars, what would be the slope of that line in miles per dollar? Write *not enough info* if it's impossible to say.
0.9 * 2 / 6 = 0.3

(e) (2 pt) Circle one of (A) *True*, (B) *False*, or (C) *Not Enough Info* to describe the following statement:

The total cost values in this sample are normally distributed.
False: Most values are between 0 and 2 with a tail to the right.

(f) (2 pt) Circle one of (A) *True*, (B) *False*, or (C) *Not Enough Info* to describe the following statement:

All of the total cost values in this sample are within 3 standard deviations of the mean.
False: One value is 40, which is larger than 3 * 6 + 13 = 31.

(g) (2 pt) Circle one of (A) *True*, (B) *False*, or (C) *Not Enough Info* to describe the following statement:

At least 88% of the total cost values in this sample are within 3 standard deviations of the mean.
True: By Chebyshev's inequality, it must be true. At least 8/9 or 88.89% are within 3 SDs of the mean for any distribution.

(h) (2 pt) Circle one of (A) *True*, (B) *False*, or (C) *Not Enough Info* to describe the following statement:

The residual costs have a similar average magnitude for short trips (1 mile) and long trips (5+ miles).
False: There is less variability for shorter rides, and therefore smaller error magnitudes.

(i) (2 pt) You compute a 95% confidence interval from this sample to estimate the height (fitted value) of the population regression line at 6 miles. Which one of the following could plausibly be the result?

(A) 5 to 7
(B) 7 to 19
(C) 12 to 14
(D) 15 to 35
(E) (Correct) 24 to 26

The regression estimate for this sample is 25, so that will be the center of the confidence interval. Since it's a large sample, most resampled estimates will be near to this value. 15 to 35 is not plausible because a resampled regression line would almost never vary so much that it would go through those extreme values.

5. (13 points) Classification

You want to predict whether a final survey response comes from a first-year student (class 1) or not (class 0) based on responses to two questions. The average responses from a random sample of 200 surveys are below.

Survey questions:
- What fraction of **lectures** did you attend?
- What fraction of the **text** did you read?

Class	Count	Lecture average	Text average
1: First-year	80	75%	64%
0: Other	120	67%	68%

(a) (2 pt) A *constant* classifier is one that always guesses the same class label, regardless of the example attributes. What's the accuracy on this sample of the best constant classifier for predicting the class?

(A) 50%
(B) (Correct) 60%
(C) 70%
(D) 80%
(E) 90%

Always predicting 0 would classify 120/200 correctly.

(b) (2 pt) Among the following, what is the best reason to expect that a nearest-neighbor classifier that uses this sample as a training set will have higher accuracy on a held-out test set than any constant classifier?

(A) The test set may have a different distribution of classes than this sample.
(B) A nearest-neighbor classifier is designed to generalize to unseen examples.
(C) A nearest-neighbor classifier can predict different classes for different examples.
(D) The attributes (lecture and text) are associated with each other.
(E) =¿ The attributes (lecture and text) are both associated with the class.

An association between attributes and the class can typically be used to make better predictions about the class.

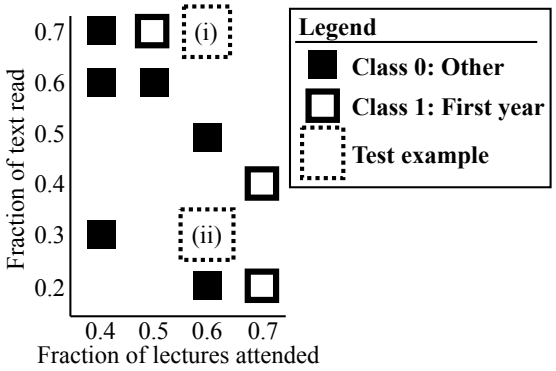
(c) (2 pt) Two roommates always attended exactly the same lectures. One read $\frac{1}{2}$ the textbook, and the other read $\frac{9}{10}$. What is the distance between these two roommates used by a nearest-neighbor classifier that includes as attributes both the fraction of lectures attended and the fraction of text read?

$\sqrt{\left(\frac{4}{10}\right)^2 + 0^2} = \frac{4}{10} == \frac{2}{5}$

(d) (3 pt) For the small training set of 9 examples shown below, how will a k-nearest-neighbor classifier label each of the two test examples (i) and (ii)? For each example and each value of k, circle either **0** or **1**. If it is impossible to determine the predicted label because of tied distances, **circle both 0 and 1**.

Circle your answers in this table:

	Prediction for (i)	**Prediction for** (ii)
Using 1 nearest	**0** (Correct)**1**	(Correct)**0** **1**
Using 3 nearest	(Correct)**0** **1**	**0** (Correct)**1**
Using 5 nearest	(Correct)**0** **1**	(Correct)**0** **1**



(e) (4 pt) Your nearest-neighbor classifier is correct $\frac{4}{5}$ of the time on the test set, but it's so slow that you can only use it for $\frac{3}{4}$ of test examples. The rest of the time you use a constant classifer that always guesses "1: First-year", which is correct only $\frac{2}{5}$ of the time on the test set. For a randomly chosen test example,

- What is the chance that it will be classified correctly?
  `(3/4 * 4/5 + 1/4 * 2/5) = 7/10`
- What is the chance that you used your nearest-neighbor classifier, given that it was classified correctly?
  `(3/4 * 4/5) / (3/4 * 4/5 + 1/4 * 2/5) = 6/7`