DATA 8, FALL 2016 **FINAL EXAM** **A. Adhikari**

NAME (FIRST LAST): _____ **SID:** _____

TIME AND CONDITIONS: 3 hours; closed book/notes/internet; no calculator/computer

QUESTIONS AND ANSWERS

• There are 16 questions. Not all questions will take the same amount of time.

• You may answer any part of any question. If the answer to one part depends on another that you couldn't do, you can still provide an answer such as "The answer to part (a), divided by 2."

• When answers involve calculations that can't easily be done by mental arithmetic, please leave the arithmetic unsimplified, unless you need to carry out a straightforward calculation in order to complete the problem. Leave arithmetic expressions in any form that can be typed (perhaps laboriously) into a calculator to get the decimal answer.

• Explanations are expected to be concise. One or two clear sentences should be enough. Calculations and code are sufficient as explanations.

GRADING

• The exam is worth 100 points.

• Questions 1-6 are worth 5 points each. Questions 7-11 are worth 6 points each. Questions 12-16 are worth 8 points each.

• We will give partial credit, but only for substantial progress towards a correct answer. We get to decide what "substantial progress" means.

• Commit yourself to a single answer for each part of each question. If you give multiple answers (such as both True and False), please don't expect credit, even if the right answer is among those that you gave.

FORMAT

• Please **write your name on each page** in the space provided. This will identify your work should there be any mechanical problems during scanning.

• There is space for your answer below each question. **Please do not write outside the black boundary**; the scanner and Gradescope won't read it.

• If you need scratch paper, please use the backs of the pages of the exam, but be aware that they will not be graded.

• A reference sheet of code and formulas will be provided. But it does not contain everything that was covered in class.

HONOR CODE

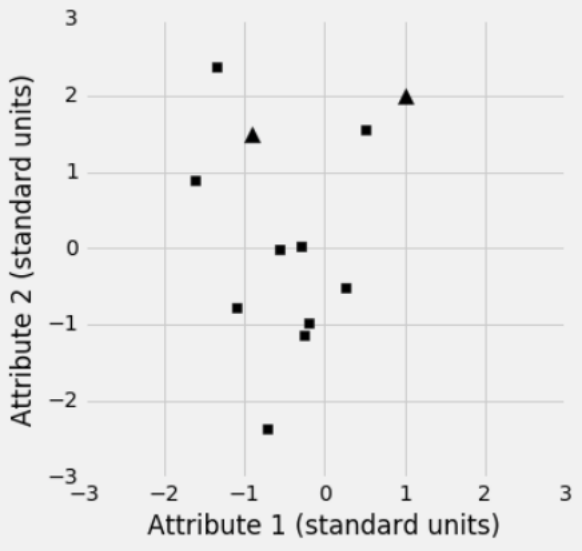
Data Science and the entire academic enterprise are based on one quality – integrity. We are all part of a community that doesn't fabricate evidence, doesn't fudge data, doesn't steal other people's work, doesn't lie and cheat. You trust that we will treat you fairly and with respect. We trust that you will treat us and your fellow students fairly and with respect. **Please abide by UC Berkeley's Honor Code:**

"As a member of the UC Berkeley community, I act with honesty, integrity, and respect for others."

Please sign here to commit to following the Honor Code: _____

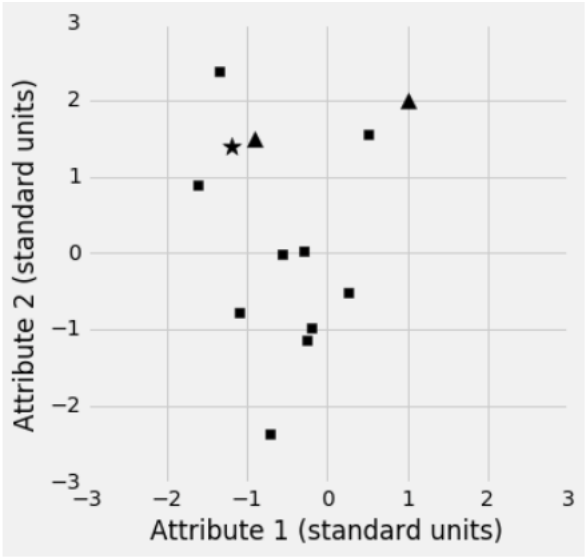
Name: _____

1. Each individual in a population belongs to one of two classes: a triangle or a square. Two attributes are going to be used to classify new individuals. The training set, consisting of 12 of points, is shown on the right. Both of the attributes have been measured in standard units so that distances are comparable on the two axes.



(a) On the graph, mark **one** new point (not in the training set) that a 3-nearest neighbor classifier using this training set would classify as a triangle. You don't have to provide reasoning.

(b) The training set is provided again for your reference. This time, the graph also contains a new point not in the training set, shown as a star. **Circle the three nearest neighbors (in the training set) of the star**, and classify it using two different classifiers below. Just underline the right shape. You don't have to provide reasoning.



1-nearest neighbor: Triangle Square

3-nearest neighbor: Triangle Square

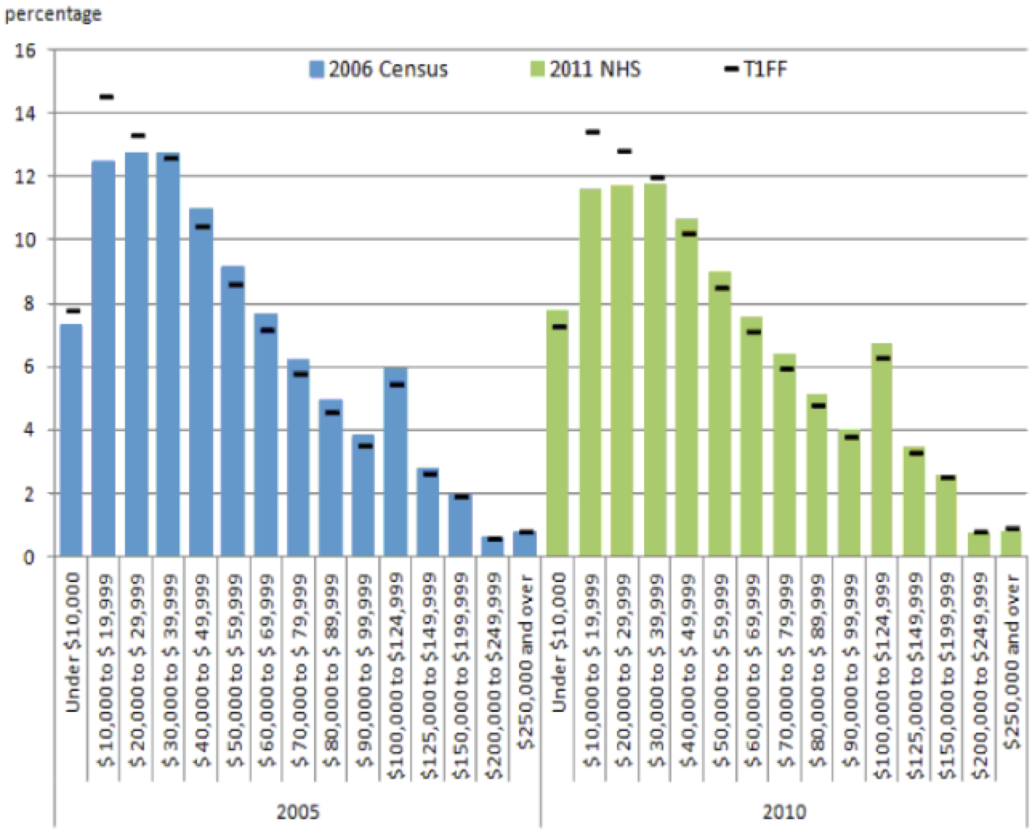
(c) Suppose a new point is below average in both attributes. In which class would it be placed by the 3-nearest neighbor classifier? Explain briefly.

Name: _____

2. The figure below appears on the website of the Canadian National Household Survey. The graphs attempt to display the distribution of family income: the graph on the left shows the incomes in 2005 and the one on the right shows incomes in 2010.

Distribution of after-tax income of census family units for Canada, 2005 and 2010

Description for figure 2



In each of the two graphs, the eleventh bar from the left is unusually tall compared to the tenth bar. Explain why.

Name: _____

3. In a population of tiny birds, the diameter of the egg and the weight of the hatchling (the baby bird that hatches from the egg) follows the regression model. The summary statistics in the sample are:

correlation = 0.75

	mean	SD
egg diameter (mm)	23	0.5
bird weight (gm)	6	0.4

(a) Find the regression estimate of the weight of a bird that hatches from an egg of diameter 24 mm.

(b) If you use the sample to make a bootstrap prediction interval at $x = 24$ mm, the interval is for predicting the height of the

(i) regression line (ii) true line in the regression model

at $x = 24$. Pick one option and explain your choice.

Name: _____

4. A data science class has 500 students. As part of an assignment, each student tests the fairness of a coin using data from his/her own set of tosses of the coin. All 500 students test the same coin, and they all test the same pair of hypotheses:

Null: The coin is fair.
Alternative: The coin is not fair.

All of the students use the 5% cutoff for the P-value. You can assume that all the students perform the same test based on the same large number of tosses.

Suppose that, unknown to the students, the coin is fair. About how many students will conclude that the coin is **not** fair? Pick one option and justify your choice.

(i) No students (ii) 5 students (iii) 10 students (iv) 25 students (v) 250 students

Name: _____

5. In a population, 85% of the people are in Class A and the remaining 15% are in Class B. For people in Class A, a classifier has an accuracy of 90% (that is, among Class A people, 90% are classified as Class A and 10% as Class B). For people in Class B, the accuracy of the classifier is 98%.

One person is picked at random from the population.

(a) What is the chance that the person is classified correctly?

(b) Given that the person is classified correctly, what is the chance that the person is in Class B?

Name: _____

6. A new function that takes a numerical argument is defined as follows:

```
def my_function(c):
    if c < -2:
        return 4
    elif c > 2:
        return 4
    else:
        return abs(c) + 2
```

(a) Draw the plot generated by the following code. You don't have to worry about exactly what labels Python will put on the axes. Just make sure the horizontal and vertical coordinates of your points are clear.

```
t = Table().with_column('x', np.arange(-3, 3.1, 1))
t.with_column('y', t.apply(my_function, 'x')).scatter(0, 1)
```

(b) Pick the option that best completes the sentence, and explain your choice.

The expression **minimize(my_function)** evaluates to

(i) -3 (ii) 0 (iii) 1 (iv) 2 (v) 3 (vi) 3.1 (vii) 4

Name: _____

7. A hospital system has data on the systolic and diastolic blood pressures (both measured in millimeters of mercury) of hundreds of thousands of patients. Assume that the scatter plot of the two variables is roughly football shaped with an unknown correlation coefficient r.

The table **bp** consists of one row for each of 300 patients sampled at random from the population of patients. The table has two columns. Column **Systolic** contains the systolic blood pressures and column **Diastolic** contains the diastolic blood pressures.

(a) Complete the code below so that the last line evaluates to an array consisting of the end points of an approximate **90%** bootstrap confidence interval for r, based on 10,000 repetitions of the bootstrap process. You may use a function **corr** that takes as its arguments two numerical arrays of the same length and returns the correlation between them. You do not need to define **corr**.

```
r_values = make_array()

for i in np.arange(_____):
    resample = bp._____
    new_r = corr(resample._____, resample._____)
    r_values = np.append(_____)

left_end = percentile(_____)
right_end = percentile(_____)

make_array(_____)
```

(b) How would you use the interval constructed in part **(a)** to test whether or not $r = 0.6$? Your answer should include the cutoff for the P-value. [No code is required for this answer. Just explain in words.]

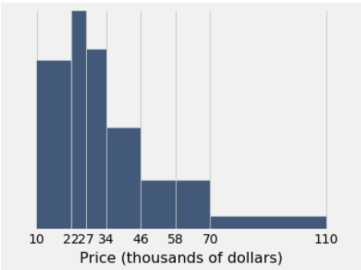
Name: _____

8. The prices of 152 cars are summarized in the table below. Prices are in thousands of dollars. Each interval includes the left end point but not the right.

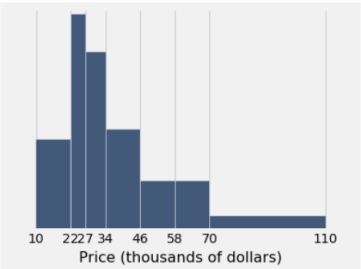
interval	[10, 22)	[22, 27)	[27, 34)	[34, 46)	[46, 58)	[58, 70)	[70, 110)
number of cars	26	26	30	29	14	14	13

(a) One of the graphs below is a histogram of these data. Which is it, and why? [No, you don't need vertical scales or a calculator.]

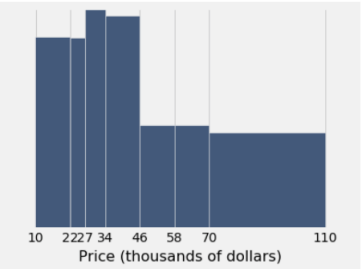
(i)



(ii)



(iii)



(b) The prices are sorted in increasing order and placed in the array **prices**. Thus the expression **len(prices)** evaluates to 152. Here are the first 20 entries of **prices**.

```
11.85,  14.07,  14.59,  16.34,  16.39,  16.91,  17.05,  18.24,  18.25,  18.56,
18.60,  18.68,  18.94,  19.01,  19.04,  19.08,  19.14,  19.14,  19.24,  19.32
```

What does the following expression evaluate to, and why?

percentile(10, prices)

Name: _____

9. Researchers studying health insurance in the United States have gathered data on whether or not people are insured.

There are several thousand people in the study. The table **insured** contains one row for each person. The table has three columns in the following order: the column **Name** contains the person's name; **Zip Code** contains the zip code of the person's home address; and **Insured** is a 0/1 variable where 1 means "insured" and 0 means "not insured".

The table **states** consists of one row for each zip code in the United States. The first column is labeled **Zip Code** and contains the zip code; the second column is labeled **State** and contains the name of the state (such as **California**, or **New York**) in which that zip code is located.

Write Python code in each of the following parts. You can use multiple lines of code. The last line of your code should evaluate to the element described in the question.

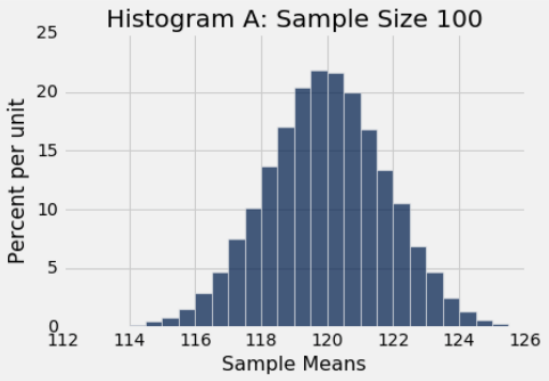
(a) the proportion of insured people in the study

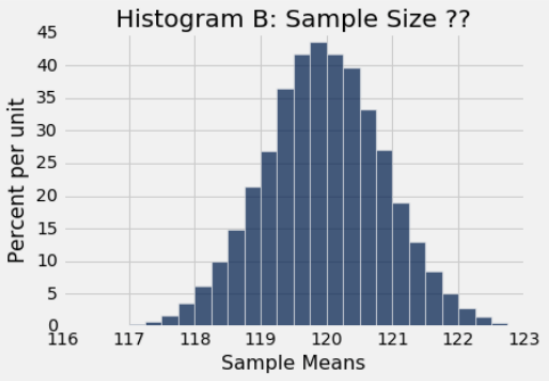
(b) a state that has the largest number of insured people among the all states represented in the study

(c) a state that has the largest proportion of insured people among the all states represented in the study

Name: _____

10. A population consists of more than half a million people. Histogram A below is an empirical histogram of the mean weight (in pounds) of a random sample of 100 people drawn with replacement from the population, based on 25,000 repetitions of the sampling process. Histogram B is an empirical histogram of the mean weight of a random sample drawn with replacement from the population, also based on 25,000 repetitions, but the sample size is unknown.





(a) Pick one option and justify your choice:
The SD of the 25,000 sample means used to construct Histogram A is closest to

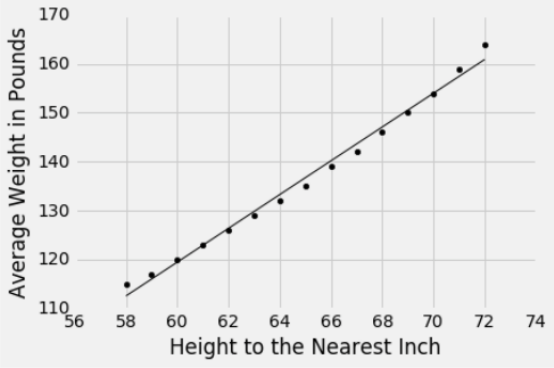
(i) 1 pound (ii) 2 pounds (iii) 3 pounds (iv) 4 pounds (v) 10 pounds (vi) 20 pounds

(b) Pick one option and justify your choice:
The size of each of the 25,000 samples whose means were used to construct Histogram B is closest to

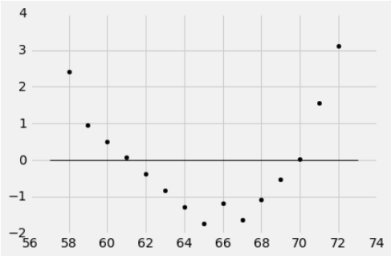
(i) 100 (ii) 200 (iii) 400 (iv) 800 (v) 1600

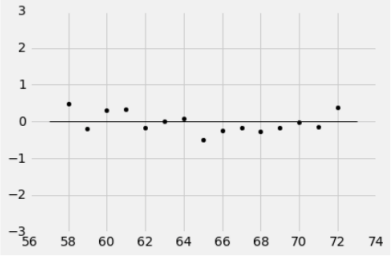
Name: _____

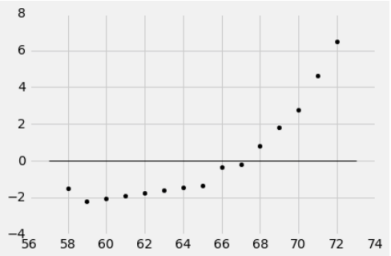
11. The plot on the right shows 15 points along with the regression line. The data represent thousands of women in the United States, grouped by height to the nearest inch. For example, all the women whose heights are 62 inches to the nearest inch form one group. The value on the horizontal axis is the height to the nearest inch, and the value on the vertical axis is the average weight of women in the corresponding group. The correlation is about 0.995.



(a) One of the graphs below is the residual plot of this regression. Which is it, and why?







(b) If you draw a scatter plot consisting of one point for each of the thousands of women, with her height on the horizontal axis and her weight on the vertical, will your scatter show a correlation of about 0.995, more than 0.995, or less than 0.995? Pick one option and explain your choice with reference to the scatter plot of heights to the nearest inch and average weights given in this problem.

Name: _____

12. In a large random sample of U.S. households, the median annual income is $54,000. This original sample is bootstrapped 5,000 times and the sample median is recorded for each of the bootstrap samples. The middle 95% interval of these values is ($53,000, $55,000).

(a) True or false (explain your answer):

The interval ($53,000, $55,000) is an approximate bootstrap 95% confidence interval for the median income of all the households in the sample.

(b) Pick the option that you think best completes the sentence, and explain your choice.

The percent of all U.S. households with annual incomes in the range ($53,000, $55,000)

(i) is about 95%. (ii) is about 50%. (iii) cannot be approximated based on the information given.

(c) Pick the option that you think best completes the sentence, and explain your choice.

If you calculate the mean of each of the 5,000 bootstrap samples and take the middle 95% interval of the 5,000 means, the center of the new interval will be

(i) less than $54,000. (ii) about $54,000. (iii) more than $54,000.

Name: _____

13. The "handedness" of a person refers to whether the person mainly uses their left hand or right hand; some people are equally at ease with both hands and are called "ambidextrous". In a study of whether handedness is is related to gender, a random sample of 1,000 people was taken in a county. There were 488 men and 512 women in the sample, and the distributions of handedness of males and females came out as follows:

	male	female
right handed	0.875	0.915
left handed	0.106	0.079
ambidextrous	0.019	0.006

(a) To test whether or not handedness and gender are related, we need null and alternative hypotheses. Does the null hypothesis say that the two distributions displayed above are the same? If not, which two distributions does it compare, and what does it say about them?

(b) State the alternative hypothesis.

(c) Justify a choice of test statistic and find its observed value in the sample.

(d) To carry out the test, the process starts with (pick one option and justify your choice):

(i) drawing 512 times at random with replacement from the distribution of males in the table above.
(ii) drawing 488 times at random with replacement from the distribution of females in the table above.
(iii) permuting all 1000 people and labeling the first 488 "male" and the remaining 512 "female".

Name: _____

14. The code below generates a plot.

```
data = Table().with_columns(
    'x', make_array(-1, 2, 0),
    'y', make_array( 2, -4, 0))
def mse(slope):
    intercept = 0
    predictions = slope*data.column('x') + intercept
    return np.mean((predictions - data.column('y'))**2)
slopes = Table().with_column('potential slope', np.arange(-3, 1, 1))
mses = slopes.apply(mse, 'potential slope')
slopes.with_column('MSE', mses).scatter('potential slope', 'MSE')
```

(a) Draw the plot. Don't worry about the labels that Python will put on the axis. Just make sure that you provide coordinates of some points so that it is clear what you are plotting.

(b) Consider the following four equations for lines. Among these, which has the lowest mean-squared error in predicting the **'y'** column of **data** based on the **'x'** column, **according to the plot you made**?

(i) y = -3*x + 0 (ii) y = -2*x + 0 (iii) y = -1*x + (iv) y = 0*x + 0

Name: _____

15. A random sample of 1,000 12-year-olds in a state took a multiple choice test. One of the questions had five possible answers, one of which was correct. Test results showed that 180 of the 1000 students got that question right.

This alarmed some educators, who said, "The kids did worse than they would have by random guessing!" But other educators said the results were like random guessing, allowing for chance variation.

Show how to perform a statistical test to see which educators' viewpoint is better supported by the data, in the following steps.

(a) State the null hypothesis as a clearly specified chance model.

(b) State the alternative hypothesis. Keep in mind that the goal of the statistical test is to decide between the two viewpoints of the educators.

(c) Suppose the test is performed using as its test statistic the number of students who get the answer right. Draw a sketch of the empiricial distribution of this statistic under the null hypothesis. Mark the observed value of the test statistic in a reasonable place on the horizontal axis (it doesn't have to be exact but it should make sense).

(d) On the sketch above, shade the area corresponding to the P-value. In the space below, explain why you chose to shade that region.

Name: _____

16. Bootstrapping is a way of replicating a sample so that you get a sample that is similar but most likely not exactly the same as the original sample. However, there is a chance that a bootstrap sample is exactly the same as the original. In this problem you will find that chance.

(a) The original sample consists of four people: John, Paul, George, and Ringo. This sample will be bootstrapped. Find the chance that all four people appear in the bootstrap sample. Your answer should just be an arithmetic expression; no code is needed.

(b) The original sample consists of N people. The sample will be bootstrapped. Write a Python function called **same** that takes N as its argument and returns the chance that all N people appear in the bootstrap sample. [There are many different ways of writing this code. Any correct way is fine.]