INSTRUCTIONS

- You have 90 minutes to complete the exam. There are 60 points available.
- 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.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Last name</td>
<td></td>
</tr>
<tr>
<td>First name</td>
<td></td>
</tr>
<tr>
<td>Student ID number</td>
<td></td>
</tr>
<tr>
<td>BearFacts email <a href="mailto:@berkeley.edu">@berkeley.edu</a></td>
<td></td>
</tr>
<tr>
<td>GSI</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your left</td>
<td></td>
</tr>
<tr>
<td>Name of the person to your right</td>
<td></td>
</tr>
</tbody>
</table>

*All the work on this exam is my own.*
*(please sign)*
1. (13 points) True/False

(a) (6 pt)
Circle either True or False for each statement below.

- The best choice of a test set for a classifier is a subset (or a small part of) the training set.
  True or False
  False

- We are unable to determine causation given a scatter plot of a dataset with two columns, and their correlation.
  True or False
  True

- A right-skewed histogram has a mean less than the median.
  True or False
  False

- A class has 1,000 people in it. The average on their final exam is a 75 percent, with a standard deviation of 5 percent. Exactly 3 people must have a score above a 90 percent.
  True or False
  False

- The slope of the regression line is equal to the correlation coefficient if and only if SD(Y) ≥ SD(X).
  True or False
  False

- A permutation test is implemented in exactly the same way as an A/B test; the only difference is the statement and interpretation of the null hypothesis.
  True or False
  False

(b) (7 pt)
Which of the following statistics would you expect to have an approximately normal sampling distribution, according to the central limit theorem? You may assume that each sample is drawn with replacement from a very large population of numbers. (Circle the letter for all that apply.)

- (a) The maximum of a sample of size 3
- (b) The mean of a sample of size 3
- (c) The minimum of a sample of size 1000
- (d) The mean of a sample of size 1000
- (e) The difference between the maximum and the minimum of a sample of size 1000
- (f) Five times the mean of a sample of size 1000
- (g) The sum of a sample of size 1000

D, F, G
2. (10 points) Histograms

(a) (6 pt)
Consider the following histograms created from ages of 714 passengers onboard the Titanic.

![Histograms](image)

Note that these histograms are drawn from the same data, with different bins.

For each of the problems below, show your work for an answer or write “not enough information”.

a. What proportion of passengers were between 0 and 20 years old or at least 60 years old? 0.8% * 10 + 1.4% * 10 + 0.25% * 10 + 0.1% * 10 = 25.5% (approximately).

b. What proportion of passengers were between 30 and 35 years old? Not enough information.

c. About how many passengers were between 10 and 15 years old? 714 * (0.7% * 15 - 0.8% * 10) = 17.85

(b) (4 pt) Recall the Warplanes problem from lecture. We are interested in estimating the number of enemy airplanes based on their serial numbers. To make this problem more realistic, suppose the serial numbers don’t necessarily start at 1. Now, to estimate the number of planes, we’ll use the spread of our sample (the largest number - the smallest) instead of just the max. Match the following descriptions with the histograms below by writing the letter of the corresponding histogram in the space provided.

a. Histogram of the observed serial numbers in a sample (with replacement) of size 5. d
b. Empirical histogram of the spread of many repeated samples of size 5. c
c. Histogram of the observed serial numbers in a sample (with replacement) of size 30. e
d. Empirical histogram of the spread of many repeated samples of size 30. a
3. (15 points) Regression

(a) (8 pt)
Suppose you are an early natural scientist trying to understand the relationship between the length of time (t, in seconds) an initially-stationary object above Earth's surface spends in free fall and the distance (d, in meters) it travels in that time. You run experiments in which you drop an iron ball 10 times from a very tall cliff; each time, you choose a time randomly between 0 and 50 seconds and measure the distance it has fallen at that time. You have three hypotheses: (i) distance is a linear function of falling time; (ii) distance is a quadratic function of falling time; or (iii) distance is a 9th-degree polynomial function of falling time. (If you don’t remember what quadratic and polynomial functions are, just think of them as kinds of curves; the graph of a quadratic function looks like a bowl, and the graph of a 9th-degree polynomial looks really wiggly. The details of how these functions work aren’t that important for understanding this question.) To test these, you decide to find the function that fits the data most closely under each hypothesis, in the sense of minimizing the mean squared error. You plot the curves and the data, getting the following three pictures:

(i) [Image of the best linear fit curve]
(ii) [Image of the best quadratic fit curve]
(iii) [Image of the best 9th-degree polynomial fit curve]

a. Rank the curves by average squared residual, least to greatest.

iii, ii, i

b. Suppose you ran another copy of the experiment, drew the curves from the first copy of the experiment (the ones displayed in the pictures above) over the 10 points from the second copy of the experiment (not pictured), and computed the difference between the predicted value and the actual y-value of each of these 10 new points. (This is like testing your predictions on a fresh test set.) Rank the curves by the average squared residual you would expect to see, least to greatest.

ii, i, iii

c. Informally, which hypothesis do you think is most supported by these data? Why?

Hypothesis 2 - it fits the data well, and the errors it makes in the training data appear to be random. Unlike hypothesis 3, the fitted curve doesn’t violate our intuition about how falling balls work.
(b) (7 pt) Assume there is a table called $t$ that contains three columns: “Month”, “Ice Cream”, and “# Murders”. Each row in this table represents a month, the amount of ice cream consumed in that month (in pounds), and the number of murders in that month. Without using the correlation function defined in *Inferential Thinking*, how would you compute the correlation of ice cream and murders?

a. **Come up with a Python expression to compute this correlation.**
   Note: You may assume that a function called `standard_units` has been defined for you. It takes an array of numbers as its argument and returns an array of those numbers converted to standard units.

   $$\text{np.mean}(\text{standard_units}(t.\text{column}(\text{murders})) \ast \text{standard_units}(t.\text{column}(\text{ice cream})))$$

b. **Circle all of the statements below that are true.**

   (A) A possible value for the correlation is 1.5.

   (B) If the correlation is between -0.05 and 0.05, there is little association between ice cream and murder rates.

   (C) If we get a correlation of -0.9, we know that a linear regression of murders per month on ice cream consumption would be able to predict murders per month fairly well.

   (D) If we get a correlation of 1, we know that ice cream consumption causes an increase in murder rates.

Only (C) is true. (A): Correlations can’t be bigger than 1.5. (B): Two things can be associated despite having a low correlation; for example, data on a symmetric bowl-shaped (parabolic) curve will have 0 correlation but a clear association. (C): A correlation of -0.9 tells us precisely that the best-fit line will predict murders per month well. (D): A correlation of 1 tells us only that there is an exact linear relationship; it doesn’t tell us about the causal nature of the relationship. For example, it could be that a common factor (like heat) causes both murder and ice cream consumption.
4. (8 points) Confidence Intervals

A group of basketball aficionados want to recruit the perfect team of players from a large random sample of athletes, based on the correlation between their height and shoe size. The measurements of each player’s height and shoe size is stored in a table called `player_measurements`, with the heights in the column called ‘height’ and the shoe sizes in the column ‘shoe size’. You can assume that the sampling scheme is essentially equivalent to random sampling with replacement. The scatter plot of the two variables is football-shaped.

The aficionados would like to construct a bootstrap confidence interval for the correlation between the heights and shoe sizes of the players in the entire population of athletes available for recruiting. Fill in the missing pieces for the function below that will compute this bootstrap confidence interval, as follows.

Assume that the function `corr(table, column_name_x, column_name_y)` returns the correlation between the arrays `table[column_name_x]` and `table[column_name_y]`, just as in class.

Complete the function `r_ci` that takes the following 5 arguments:

1. `table`: the table containing the data
2. `column_name_x`: the label (string) of the column containing variable x
3. `column_name_y`: the label (string) of the column containing variable y
4. `L`: a floating-point number, strictly between 0 and 100, specifying the level of confidence
5. `rep`: the number of repetitions of the bootstrap resampling procedure

The function should return an array consisting of the two endpoints of an approximate `L%` confidence interval, constructed using the bootstrap percentile method, for the correlation between the two variables in the population.

```python
def r_ci(table, column_name_x, column_name_y, L, rep):
    slopes = []
    ...
    bootstrap_sample = ...
    slopes.append(corr(bootstrap_sample, column_x, column_name_y))
    return np.array(..., ...)
```

1. 
2. 
3. 
4. 

```
1. for i in range(ref)
2. table.sample(with_replacement=True)
3. np.percentile(slopes, 0 + (100-L)/2)
4. np.percentile(slopes, 100-(100-L)/2)
```
5. (14 points) Hypothesis Testing

Inspired by Mark Zuckerberg, Henry decides to leave Berkeley and found a startup. He opens the Data8 ISP, providing internet access to Berkeley students. Data8 ISP promises connection speeds of 400 megabits per second (mbps). The fine print says that Data8 ISP only promises that the each customer’s connection speed is a random sample from a Normal distribution with mean 400 mbps and standard deviation 20 mbps.

(a) (3 pt) Joseph buys this new service, but he finds that the connection seems to be much slower than advertised. He decides to conduct a hypothesis test to determine whether the connection really is slower than advertised. To conduct the test, he buys 30 separate connections whose speeds he can test. What should his null hypothesis be? What should his alternative hypothesis be?

Null hypothesis: each connection speed comes from a Normal distribution with mean 400 and standard deviation 20. Alternative hypothesis: the connection speeds come from a distribution with a smaller mean.

(b) (4 pt) Suppose Joseph tests the connection speed on his 30 computers, and he obtains an average connection speed of 402 megabits. That’s his test statistic. Now he wants to compute a P-value. What distribution should he compute in order to find the P-value corresponding to this test statistic?

He should repeatedly (say 10,000 times) simulate drawing 30 numbers from a Normal distribution with mean 400 and standard deviation 20 and taking the average of those 30 numbers. This will give him 10,000 averages that he would have seen if the null hypothesis were true. He could then compare his test statistic to that distribution to compute a P-value.

(c) (3 pt) Would you expect the p-value of Josephs test to be large or small? In a hypothesis test, what does the p-value represent?

Large. It’s the probability of seeing an average as extreme or more extreme than the observed average of 402, given the null hypothesis is true. 402 megabits is a pretty reasonable average to see if our null is true. (In fact, since the observed average is bigger than 400, if we decide to use a 1-sided test, then the P-value will be more than 1/2!)

(d) (4 pt) Lets say that Joseph wants to minimize the risk of blaming Henry for a problem that doesn’t really exist. He decides that he’s willing to accept a 1% chance of concluding that the connection is slower than advertised if it’s really as fast as advertised. What should his significance level be? In a hypothesis test, what does the significance level represent?

0.01. It’s the probability of rejecting the null hypothesis, if the null hypothesis is actually true.

(e) (3 pt) Regardless of your answer to the above questions, assume that after performing his test, Joseph obtains a p-value of 0.03. If he chooses a significance level of 0.05, which hypothesis should he reject, if any? What about with a significance level of 0.005?

With a significance level of 0.05, he should reject the null hypothesis that Henry’s promise is genuine. With a significance level of 0.005, he fails to reject that hypothesis.
6. (1 points)  Extra Credit

Here’s an extra credit problem! (Yes, we know this is a practice exam.)

Guess the lowest positive integer that no one else will guess.

In other words, if five people guess 1, ten guess 2, four guess 3, three guess 4, none guess 5, one guesses 6, two guess 7, one guesses 8, and one guesses 9, the person who guessed 6 wins.

Write your guess here: