CS 133 Foundations of Data Science Final Exam
Boise State University
Department of Computer Science
Dr. Kennington

Name:

Instructions: You have 120 minutes to complete this exam. There are 15 questions, each worth a varied number of points for a total of 50 points. This exam is closed book, but you may use one page (front and back) of notes. Please write legibly. Remember, the point and goal of this exam is for you to prove to me that you know your stuff. Good luck!

Part I: Functions & Simulations

1. (3 points) What is the general purpose of functions in Python?

Consider the following Python function then answer questions 2-6 related to it below (the text between the quotes in the function constitutes a comment):

```
1  def bootstrap_median(original_sample, label, replications):
2      """
3      Returns an array of bootstrapped sample medians:
4      original_sample: table containing the original sample
5      label: label of column containing the variable
6      replications: number of bootstrap samples
7      """
8      just_one_column = original_sample.select(label)
9      medians = make_array()
10     for i in np.arange(replications):
11         bootstrap_sample = just_one_column.sample()
12         resampled_median = percentile(50, bootstrap_sample.column(0))
13         medians = np.append(medians, resampled_median)
14
15     return medians
```

2. (4 points) What is the **general** purpose of the `bootstrap_median` function? How do you know? What does that tell you about how functions should be written?

3. (3 points) What parameters/variables does the `bootstrap_median` function require? Are any of them optional?

4. (3 points) What is the purpose of the `for` loop (lines 10-13) in the `bootstrap_median` function?

5. (4 points) Line 11 of the `bootstrap_median` function could be rewritten as:

```
bootstrap_sample = just_one_column.sample(with_replacement=False)
```

Assuming the default behavior is `with_replacement=True`, Why doesn't the original function use the above line of code instead?

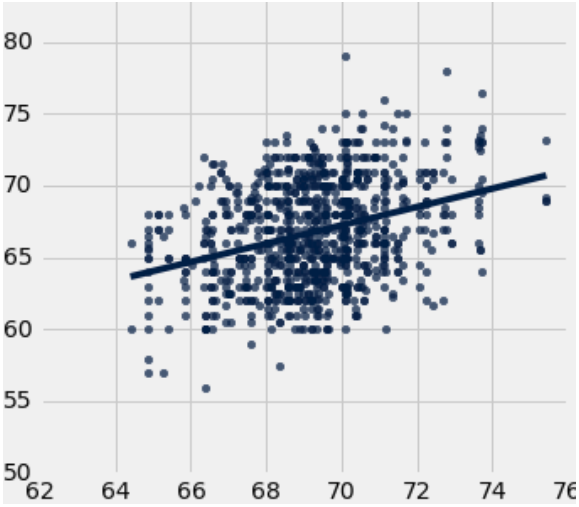
6. (3 points) How can someone use the `bootstrap_median` function to determine a p-value for a particular data set?

7. (3 points) Suppose you have several thousand points of data that you have collected. How can you determine if the mean of your data is somewhere close to the true mean?

8. (3 points) What does it mean when we say that we can reject the null hypothesis beacuse our p-value is less than a 0.05 threshold?

Part II: Regression & Distributions

Consider the following figure then answer questions 9-10 related to it below:



The correlation of the data in the figure is 0.32 and the regression is modeled with the linear function `y = 0.64 x + 22.64`.

9. (3 points) What would you predict the value of the dependent variable to be, given that the independent variable is 100?

10. (3 points) What does the correlation tell us about this data? What does the regression line tell us about this data?

11. (3 points) What is the general difference between correlation and regression?

12. (3 points) When we use linear regression to model and predict data, what kinds of assumptions are we making about the data?

13. (3 points) What do the *mean* and *standard deviation* represent for a normal distribution? For any dataset, can we calculate those values? Does it always make sense to use a normal distribution when modeling a dataset?

14. (4 points) Suppose you collect two sets of data, each with 500 samples. After plotting each dataset (i.e., each dataset has an x independent variable and a y dependent variable), you find that one of the datasets is normally distributed, but the other dataset is not–it looks more like a heavily-skewed exponential distribution. You then decide to resample 200 samples from each data set several thousand times, calculate the mean of each resampling, then plot each sample mean on the x axes and the percent of times the means were calculated on the y axes. What do you expect your two new plots to look like? Why do you expect that?

Part III: Privacy, Bias, & Ethics

15. (5 points) Identify a problem in data science where data scientists need to be vigilant when it comes to privacy, bias, or ethics. For the problem you identified, why is it a problem? What do you think data scientists should do to combat the problem?