ECE 20875: Python for Data Science
Fall 2019 Midterm Examination

Milind Kulkarni and Chris Brinton

Name: _____ Purdue username: _____

Section: ☐ Kulkarni ☐ Brinton

You have 60 minutes to complete the 5 (multi-part) questions in this exam. It is worth 100 points total, and the number of points for each question is listed.

You are free to consult printed out materials from the course website and handwritten notes for the exam. However, you are not permitted to use a computer, calculator, or any other resources.

For each question, you must show all the work you used to arrive at your answer. Use the space underneath each problem statement to write your solution to that question. If you need more than what is provided, feel free to use the back of the page.

Good luck!

Honor Pledge: (Please sign the following)
I affirm that the answers given on this test are mine and mine alone. I did not receive help from any person or material (other than those explicitly allowed).

X_____

Question	Points
1	/15
2	/20
3	/25
4	/20
5	/20
Total	**/100**

QUESTION I: BASH

1. [5 pts] Consider the following line of bash code:

```
./cmd1 > foo; ./cmd2 < foo > bar; cat bar
```

Of the following lines of bash code, circle the ones that are guaranteed to produce the same output as the above line, no matter what file(s) are in the directory to begin with and no matter what cmd1 and cmd2 do. We only care about the output of running the line of code, not the final state of any file(s) in the directory.

```
./cmd1 > foo; ./cmd2 < foo > bar && cat bar
```

```
./cmd1 &> foo; ./cmd2 < foo > bar; cat bar
```

```
./cmd1 | ./cmd2 > bar; cat bar
```

```
./cmd1 | ./cmd2 | cat
```

```
./cmd1 >> foo; ./cmd2 < foo > bar; cat bar
```

2. [10 pts] You run the following two bash commands, and get *different* results (different content in foo).

```
V=temp ; ./cmd > foo
```

```
export V=temp ; ./cmd > foo
```

Complete the file cmd in the box below by adding one or more bash commands, so that you would see the behavior above (different results from the two commands). To get full credit, your answer should use three or fewer lines.

```
#!/usr/bin/env bash

echo $V #anything that uses the value of $V in some way should work
```

QUESTION II: PYTHON BASICS & DATA STRUCTURES

1. [10 pts] Suppose you have a list of numbers called `data`. Write a *list comprehension* that produces a new list that has the squares of all the *non-negative* numbers in `data` (and filters out the negative numbers). For example, applying this comprehension to `[2, -3, 5, -1, 0]` should result in `[4, 25, 0]`

[x * x for d in data if d >= 0]

2. [10 pts] Fill in the missing part of the following function. Assume that `my_data` is a list of tuples, where each tuple has two elements. Each tuple in `my_data` should become a key/value pair in the output dictionary (in other words, the keys of `output` should be the first elements of the tuples, and the value for each key should be the corresponding second elements).

For example the result of running:

```
build_dictionary([('a', 1), ('b', 2), ('c', 3)]
```

should be a dictionary with key/value pairs that look like this:

```
{'a' : 1, 'b' : 2, 'c' : 3}
```

Assume that there will not be repeat keys in `my_data`. For full credit, your answer *should not use* `range` or `len`.

```
def build_dictionary (my_data) :
      output = {}

      for k, v in my_data :
         output[k] = v









      return output
```

QUESTION III: HIGHER ORDER FUNCTIONS

1. [15 pts] What do the three print statements print? (Write your answer next to each print statement)

```
def foo(x, f) :
   def bar(y) :
      return y * f(x)
   return bar

def a(z) :
   return 2 * z

b = foo(3, a) #line 1
c = foo(4, b) #line 2

print(a(5)) # 10

print(b(5)) # 30

print(c(5)) # 40
```

2. [10 pts] Write a function `multi_map` that takes in two arguments:

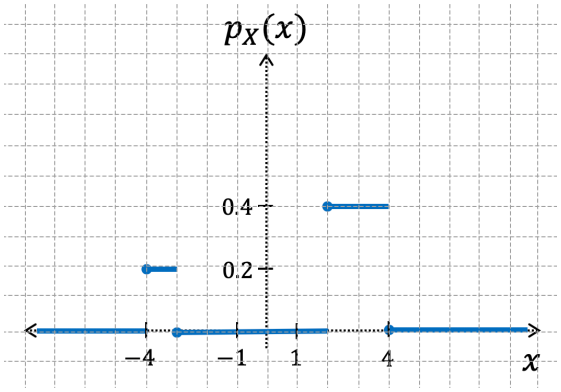
`data`: A list of integers
`funs`: A list of functions

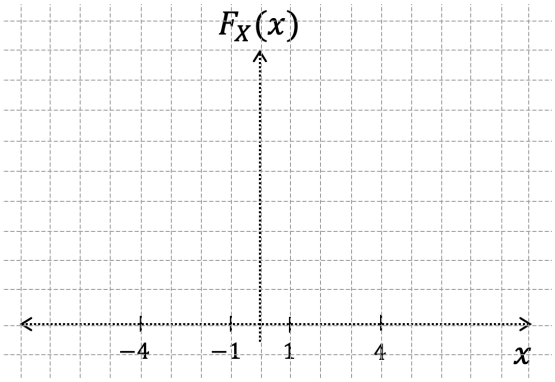
And returns a list where the first element is the result of calling the first function in `funs` on the first element of `data`, the second element is the result of calling the second function in `funs` on the second element of data, and so on. (Assume the two lists have the same length.)

```
def multi_map(data, funs) :
      output = []


     for i in range(len(data)) :
        output.append(funs[i](data(i)))




      return output
```

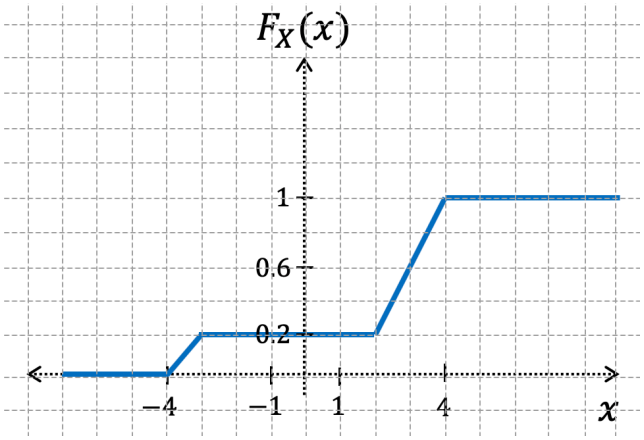
QUESTION IV: PROBABILITY DISTRIBUTIONS & HISTOGRAMS

1. [10 pts] Consider a continuous random variable X distributed according to the PDF on the left below. Sketch of the CDF on the right. Be sure to clearly indicate where it hits 0 and 1.





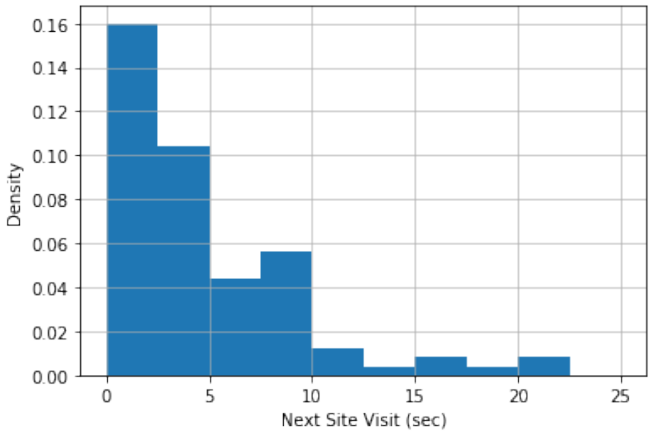
Solution



Rubric

- **Showing the CDF is between F=0 (+1) and F=1 (+1)**
- **Having ranges where F is constant (+1) and ranges where it is increasing linearly (+2)**
- **Correct values of each piece (+1 each, 5 total, for +5)**

2. [10 pts] A company is interested in how frequently viewers visit their website. They measure the time between the next 100 people visiting and obtain the following histogram:



A. Using this histogram as an estimator, approximate the probability that the next site visit will be between 5 and 10 sec. (You only need to be correct to one significant digit.)

The density between 5 and 7.5 is slightly above 0.04, and between 7.5 and 10 is slightly below 0.06. Approximating both as 0.05, we have

$$\hat{p}[between\ 5\ and\ 10] = d_{5-7.5} \cdot w_{5-7.5} + d_{7.5-10} \cdot w_{7.5-10} = 2 \cdot 2.5 \cdot 0.05 = 0.25$$

where d stands for the density of a bin and w for the width of a bin. Note that the y-axis is measured in density, not frequency (or count).

B. Assume the true population follows an exponential distribution. How do we explain the fact that our estimate is not always decreasing as we move right, e.g., the density between 20 and 22.5 sec appears higher than between 17.5 and 20 sec?

While the population density would be monotonically decreasing for higher values of the random variable (i.e., a longer next site visit is always less probable), what we have is a sample of the population, which is always an imperfect representation of it. The general trend is clear, but with this bin width we would need many more than 100 samples for the values between consecutive bins to always be accurate.

C. Would it be advantageous to use 100 bins instead of 10 bins? Explain.

No. As discussed in (B), our histogram is experiencing some issues with the accuracy of the density even for 10 bins. Increasing the number of bins (decreasing the bin width) increases precision at the expense of lower accuracy density estimation in each bin, so this would make the problem worse. If a histogram is being used as an estimator, it is also never a good idea to have the number of samples equal to the number of bins.

Rubric

- In (A), +2 for adding probabilities in the range together
- In (A), +1 for using density instead of frequency
- In (A), +1 for the answer being correct (between 0.20 and 0.30)
- In (B), +3 for explaining that our estimate is based on a sample from the population
- In (C), +2 for explaining the impact of increasing the number of bins
- In (C), +1 for saying it is not advantageous

QUESTION V: CONFIDENCE INTERVALS & HYPOTHESIS TESTING

1. [10 pts] Suppose we have reason to believe that a system has bias in it, which we test by measuring whether it produces an output significantly higher or lower than 0. We collect 25 samples, which produce an average output of $\bar{x} = 2$. We also know that $\sigma = 5$.

A. Formulate the null and alternative hypotheses for this statistical test.

$H_0: \mu = 0$ (the system does not have bias in it)
$H_1: \mu \neq 0$ (the system does have bias in it)

B. What is the standard error, and the z-score of the sample average?

With $n = 25$ samples and a population $\sigma = 5$,

$$SE = \frac{\sigma}{\sqrt{n}} = \frac{5}{5} = 1$$

and

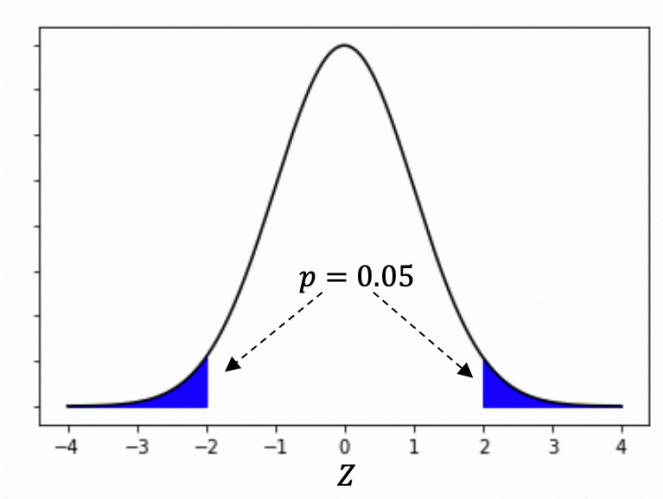
$$z = \frac{\bar{\bar{x}} - \mu}{SE} = \frac{2}{1} = 2$$

C. Find the p-value (Hint: remember the 68-95-99.7% rule), and sketch the result of the test on a normal curve below. Is the result significant at $\alpha = 0.1$? How about 0.01?

For the standard normal distribution, we know that

$$p = P(Z < -2 \text{ or } Z > 2) = 1 - P(-2 \leq Z \leq 2) = 0.05$$

which is the p-value.



Since $0.01 < p < 0.1$, the result is significant at $\alpha = 0.1$ but not at $\alpha = 0.01$.

Rubric

- In (A), +1 for the null hypothesis, and +1 for the alternative hypothesis
- In (B), +2 for the standard error, and +2 for the z-score
- In (C), +1 for using the rule, +1 for the p-value, +1 for the chart, and +1 for the conclusions

2. [10 pts] Consider the distribution of the number of Facebook friends people have. We collect several hundreds of samples and find a 95% confidence interval for the mean of $\mu \in (200, 300)$.

A. What is the sample mean we are working with?

By the confidence interval formula $\mu \in \bar{x} \pm z_\alpha \cdot (\sigma/\sqrt{n})$, we see that the sample mean is always the midpoint. Hence $\bar{x} = 250$.

B. Circle which of the following could be the 90% confidence interval:

$\mu \in (230, 320)$ $\mu \in (220, 280)$ $\mu \in (175, 325)$ $\mu \in (201, 299)$

With a lower confidence (decreasing accuracy), we can have a narrower interval (increasing precision). The interval will always be symmetric about $\bar{x}$.

C. If we collect 25 times more samples (i.e., from n to $25n$), what will be the new 95% confidence interval? (Assume the sample variance stays the same.)

By the confidence interval definition, $50 = z_\alpha \cdot (\sigma/\sqrt{n})$. If n increases by a factor of 25, we have

$$z_\alpha \cdot (\sigma/\sqrt{25n}) = (1/5) \cdot z_\alpha \cdot (\sigma/\sqrt{n}) = 50/5 = 10.$$

Therefore, the new confidence interval is

$$\bar{x} \pm z_\alpha \cdot (\sigma/\sqrt{25n}) = 250 \pm 10 = (240, 260)$$

Rubric

- **In (A), +2 for realizing it is the midpoint, +1 for the correct answer.**
- **In (B), +1 for each one correct, up to a maximum of +3.**
- **In (C), +2 for making it narrower, and +2 for the correct answer.**