Midterm 1

This is a **60 minute** exam. You will have an additional 10 minutes to scan and upload your exams to Gradescope. Exams turned in after 10:10 am (UTC-7) will be marked as "Late" and will not be graded.

If, for whatever reason, you are unable to upload your exam to Gradescope, you may email them to taliasaravi@berkeley.edu but no later 10:10 am.

During this exam, you:

- may use your notes
- may not use the internet or any outside sources
- may not discuss the exam during the entire exam period and until all exams have been completed.

Please note there are two exam periods so you may not discuss the exam until both groups have finished testing.

Academic Integrity:

Berkeley Campus Code of Student Conduct (http://sa.berkeley.edu/student-code-of-conduct):

"The Chancellor may impose discipline for the commission or attempted commission (including aiding or abetting in the commission or attempted commission) of the following types of violations by students, as well as such other violations as may be specified in campus regulations: 102.01 Academic Dishonesty: All forms of academic misconduct including but not limited to cheating, fabrication, plagiarism, or facilitating academic dishonesty."

By signing, you agree that all exam work is your own and that you abided by the above rules. Not abiding by this honor code will lead to the penalties. The penalties for academic dishonesty may include failure in the course or potentially expulsion from the university. Please sign above question 1 on your exam.

Signed: ______________________________ Date: ________________

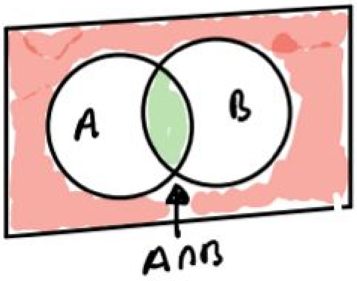
Problem 1 (20 pts):

We discussed several important topics including concepts, tools, theorems and definitions. In this question, we ask you to write a short discussion (3-4 paragraphs, all inclusive) on one of these topics (or any combination of topics). Your essay should describe the topic, in words and in math (including its theoretical foundations), explain its significance and applications, and provide one or more examples. The examples should not replicate examples from the textbook or lecture material (but smart variations are welcome). You will be graded based on thoroughness, accuracy and clarity of your answer, as well as the quality of original example(s). Example(s) based on lab data/material are very welcome.

BAYES' THEOREM AND CONDITIONAL PROBABILITY

Bayes' Theorem is an application of conditional probability.
Given a sample space:
what is the probability of A given B?

$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$



?

What is the probability of B given A.

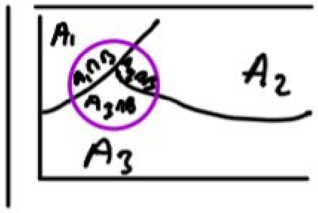
$$P(B|A) = \frac{P(A\cap B)}{P(A)}$$

Then, we manipulate these two equations in order to get Bayes' Theorem.

$$P(A\cap B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \Rightarrow \text{EQ \#1}$$

When we have a sample space which is a collection of 3 events: A_1, A_2, A_3. They are mutually exclusive and collectively exaustive.



$$P(B) = P(A_1\cap B) + P(A_2\cap B) + P(A_3\cap B)$$

$$P(B) = P(B|A_1) \times P(A_1) + P(B|A_2) \times P(A_2) + P(B|A) P(A_3)$$

$$P(B) = \sum_{i=1}^{n} P(B|A) \times P(A_i) \Rightarrow \text{EQ } \#2$$

$$P(A_i|B) = \frac{P(B|A_i) \times P(A_i)}{\sum_{i=1}^{n} P(B|A_i) \times P(A_i)} \Rightarrow$$

These are equation 1 and equation 2 together

Ex: What is the probability of two boys given at least one boy?

G= girl

B= boy

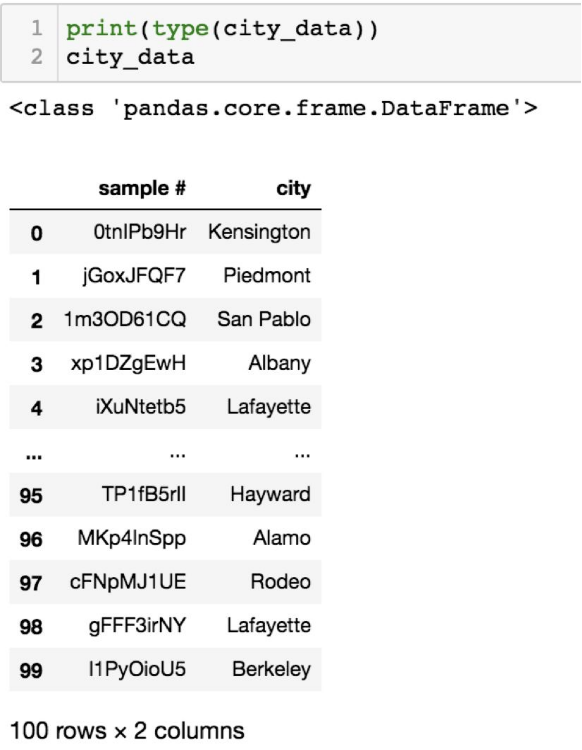
$$P(2B) = GB, \underset{\uparrow}{BB}, BG, GG = \frac{1}{4}$$

$$P(2B/\text{at least } 1B) = \frac{P(1B/2B) \cdot P(2B)}{P(1B)}$$

$$= \frac{1 \cdot 1/4}{3/4} = \frac{1}{3}$$

Problem 2 (20 pts):

Let's assume we work at EBMUD, and we have collected 100 samples of drinking water at different sites around the Bay Area to test the total dissolved solids level of the water (TDS). Each sample has a TDS measurement in ppb (parts per billion) and the city where the sample was taken associated with it. The data is stored in the data frames shown below:

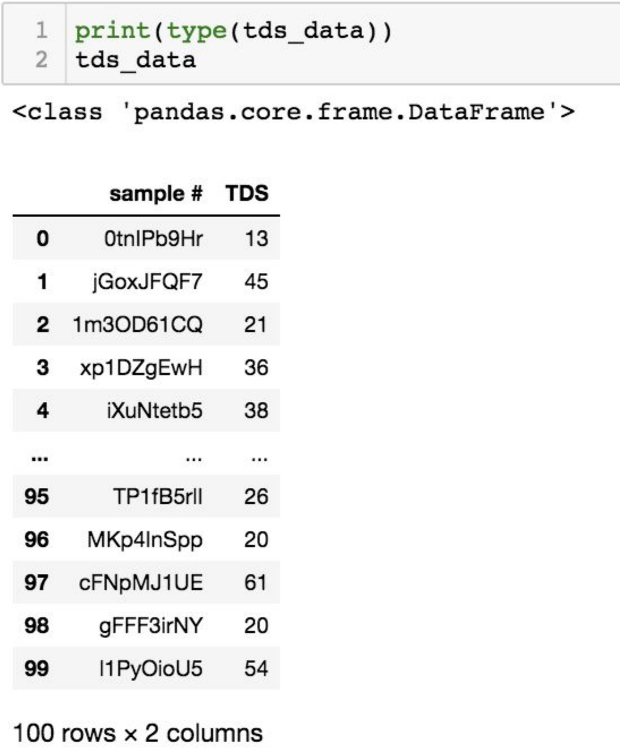


```
print(type(city_data))
city_data
```

```
<class 'pandas.core.frame.DataFrame'>
```

	sample #	city
0	0tnIPb9Hr	Kensington
1	jGoxJFQF7	Piedmont
2	1m3OD61CQ	San Pablo
3	xp1DZgEwH	Albany
4	iXuNtetb5	Lafayette
...	...	...
95	TP1fB5rll	Hayward
96	MKp4InSpp	Alamo
97	cFNpMJ1UE	Rodeo
98	gFFF3irNY	Lafayette
99	l1PyOioU5	Berkeley

100 rows × 2 columns



```
print(type(tds_data))
tds_data
```

```
<class 'pandas.core.frame.DataFrame'>
```

	sample #	TDS
0	0tnIPb9Hr	13
1	jGoxJFQF7	45
2	1m3OD61CQ	21
3	xp1DZgEwH	36
4	iXuNtetb5	38
...	...	...
95	TP1fB5rll	26
96	MKp4InSpp	20
97	cFNpMJ1UE	61
98	gFFF3irNY	20
99	l1PyOioU5	54

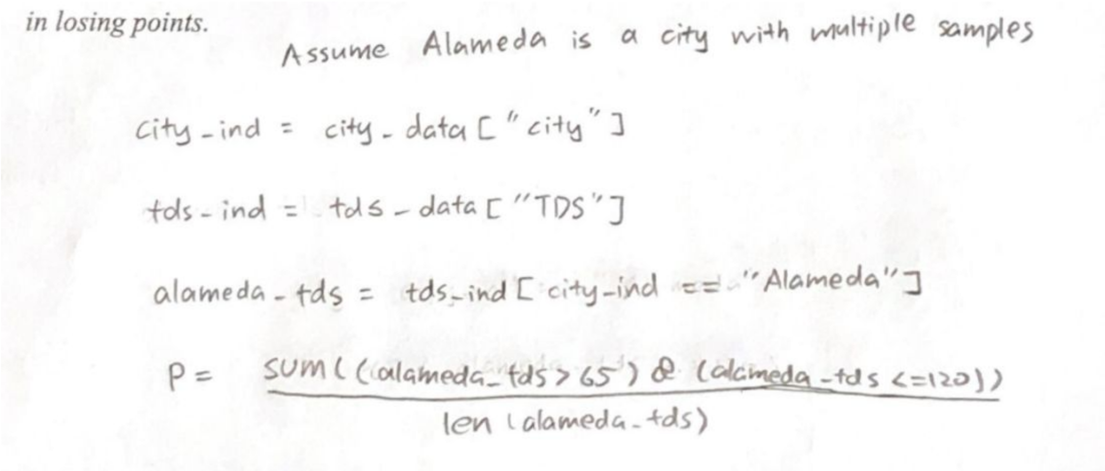
100 rows × 2 columns

sample #: a string, the unique number assigned to a water sample.
city: a string, the city where the sample was taken.
TDS: an int, the total dissolved solids detected in the sample in parts per billion.

(see next page for the first part of Problem 2)

a) While looking through the dataset, we see that the data for Alameda looks like it has much higher values for the total dissolved solids compared to other samples. The system average of TDS is 65 ppb. Write code to find the probability that a randomly selected sample from Alameda has a TDS that is larger than the system average (not inclusive of the system average), but not larger than 120 ppb, using the data frames given above.

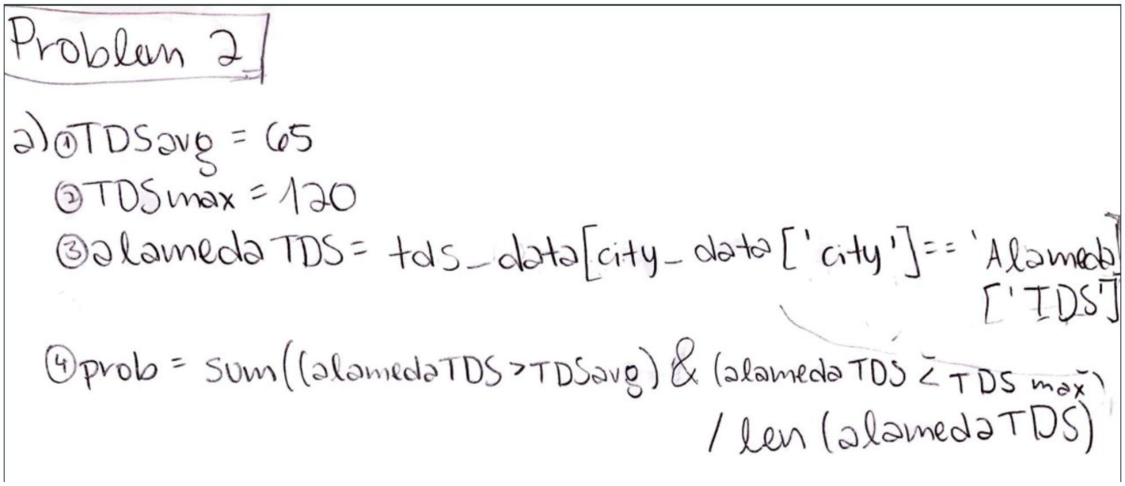
Notes for writing code: *Please make sure your code is legible, and your syntax is clear. Additionally, please number each line of code. Please make sure brackets, parentheses, and curly brackets are distinguishable. Failure to make these distinguishable will result in losing points.*



in losing points.

Assume Alameda is a city with multiple samples

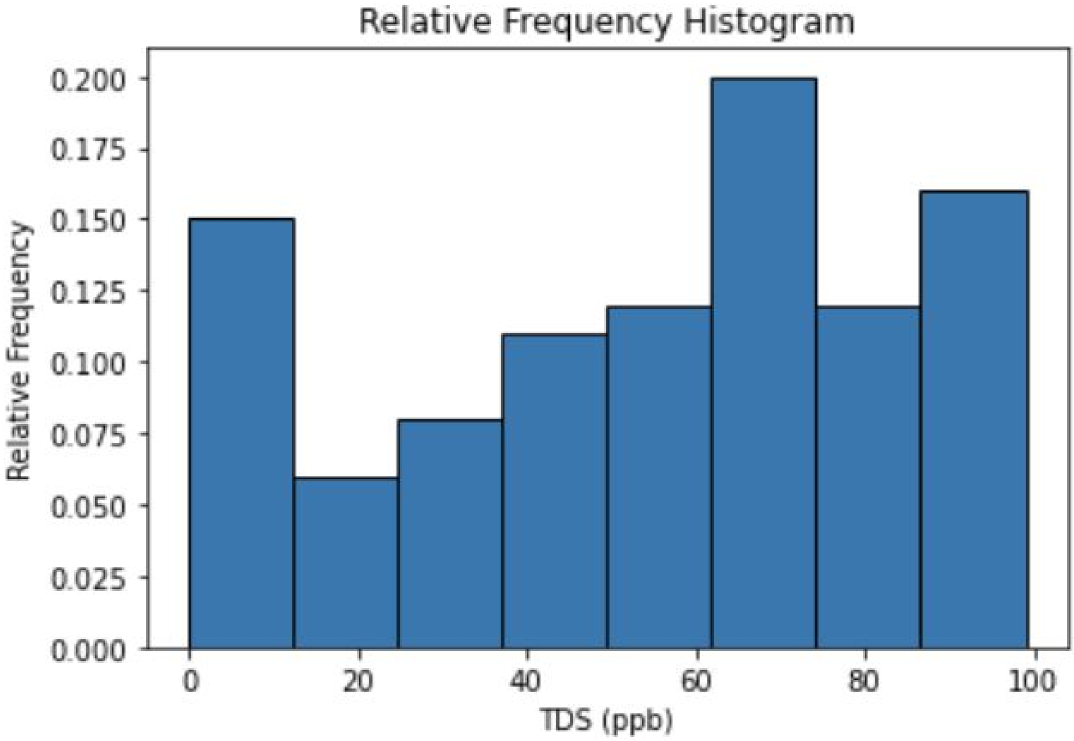
```
city_ind = city_data["city"]

tds_ind = tds_data["TDS"]

alameda_tds = tds_ind[city_ind == "Alameda"]

P = sum((alameda_tds > 65) & (alcmeda_tds <=120))
    ---------------------------------------------
              len(alameda_tds)
```



Problem 2

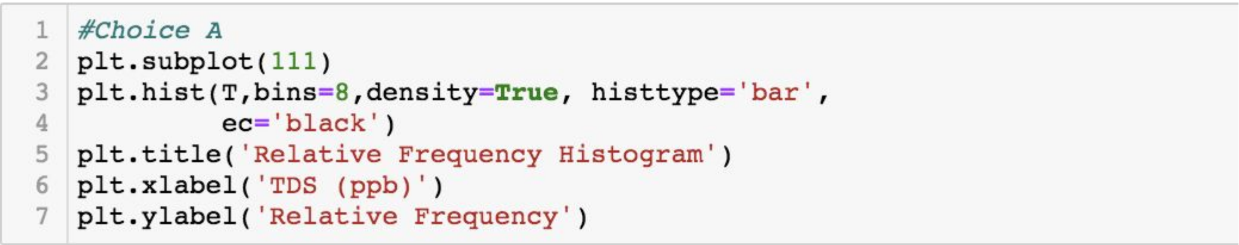
```
a) ① TDSavg = 65
   ② TDSmax = 120
   ③ alamedaTDS = tds_data[city_data['city'] == 'Alameda]
                                                   ['TDS']
   ④ prob = sum((alamedaTDS > TDSavg) & (alamedaTDS < TDSmax)
                                      / len(alamedaTDS)
```

b) We develop a relative frequency histogram for the TDS data. We create the figure below:

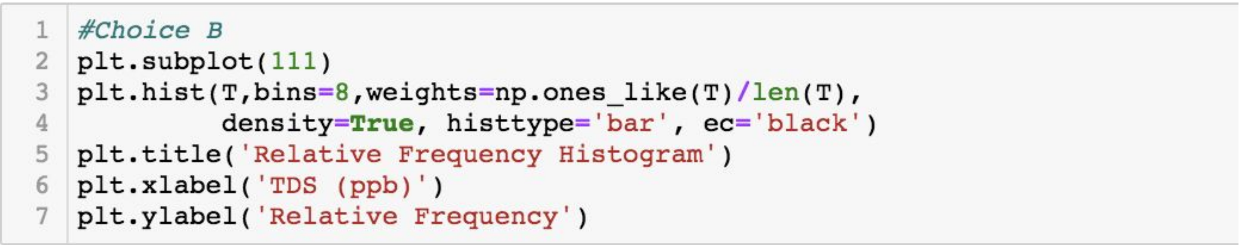


We first create the variable T, which represents the values in the TDS column. What would be your input into python to create this figure? *(see next page for options)*

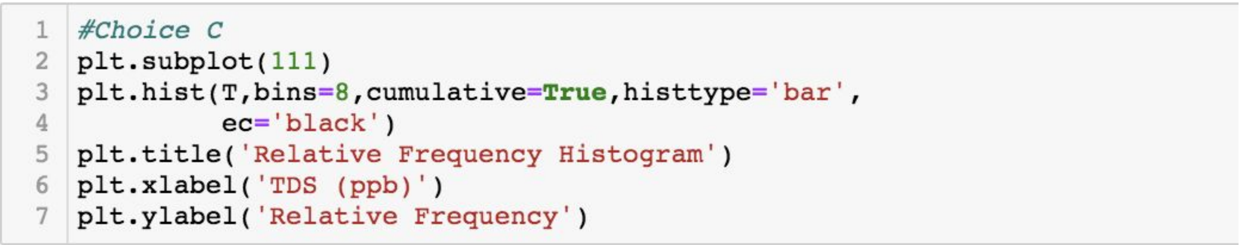
Choice E is correct.



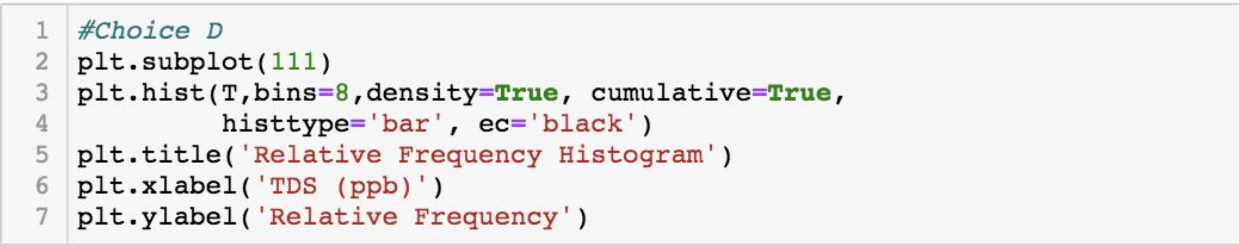
```
#Choice A
plt.subplot(111)
plt.hist(T,bins=8,density=True, histtype='bar',
         ec='black')
plt.title('Relative Frequency Histogram')
plt.xlabel('TDS (ppb)')
plt.ylabel('Relative Frequency')
```



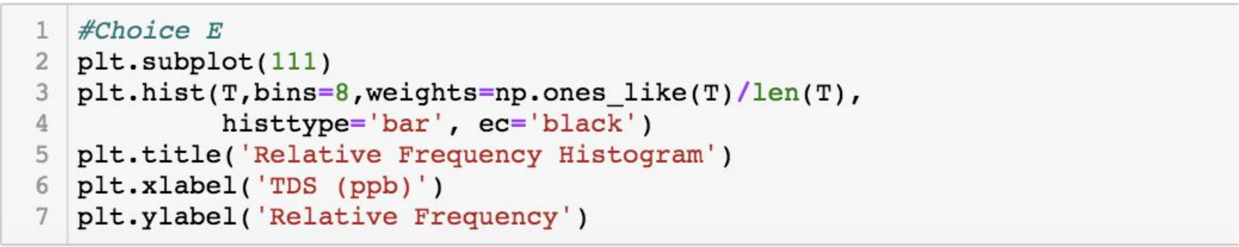
```
#Choice B
plt.subplot(111)
plt.hist(T,bins=8,weights=np.ones_like(T)/len(T),
         density=True, histtype='bar', ec='black')
plt.title('Relative Frequency Histogram')
plt.xlabel('TDS (ppb)')
plt.ylabel('Relative Frequency')
```



```
#Choice C
plt.subplot(111)
plt.hist(T,bins=8,cumulative=True,histtype='bar',
         ec='black')
plt.title('Relative Frequency Histogram')
plt.xlabel('TDS (ppb)')
plt.ylabel('Relative Frequency')
```



```
#Choice D
plt.subplot(111)
plt.hist(T,bins=8,density=True, cumulative=True,
         histtype='bar', ec='black')
plt.title('Relative Frequency Histogram')
plt.xlabel('TDS (ppb)')
plt.ylabel('Relative Frequency')
```



```
#Choice E
plt.subplot(111)
plt.hist(T,bins=8,weights=np.ones_like(T)/len(T),
         histtype='bar', ec='black')
plt.title('Relative Frequency Histogram')
plt.xlabel('TDS (ppb)')
plt.ylabel('Relative Frequency')
```

Problem 3 (25 pts):

A factory runs three shifts per day. Of all the items produced by the factory, 35% of them are produced on the first shift, 50% on the second shift, and 15% on the third shift. Of all the items produced on the first shift, 10% are defective, while 20% of the items produced on the second shift and 15% of the items produced on the third shift are defective.

a) Determine the probability of an item being defective.

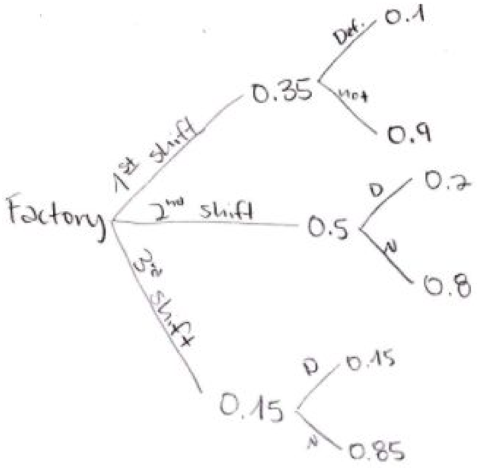
b) An item is sampled at random from the day's production, and it turns out to be defective. What is the probability that it was manufactured during the first shift?

c) On a particular Wednesday, 4 items were produced by the second shift. Let X be the number of items that are defective.

i) What is the PMF for X? Plot the results.

ii) What is the value for the expectation, $E(X)$?

iii) What is the value for the variance, *Var(X)*?



a) $P(D) = P(1 \wedge D) + P(2 \wedge D) + P(3 \wedge D)$

$= (0.35)(0.1) + (0.5)(0.2) + (0.15)(0.15)$

$= \boxed{0.1575 = 15.75\%}$

b) $P(1 \mid D) = \frac{P(1 \wedge D)}{P(D)} = \frac{(0.35)(0.1)}{0.1575} = \boxed{0.222 = 22.2\%}$

c) 4 items from second shift, x = # defective

$P(D) = 0.2$

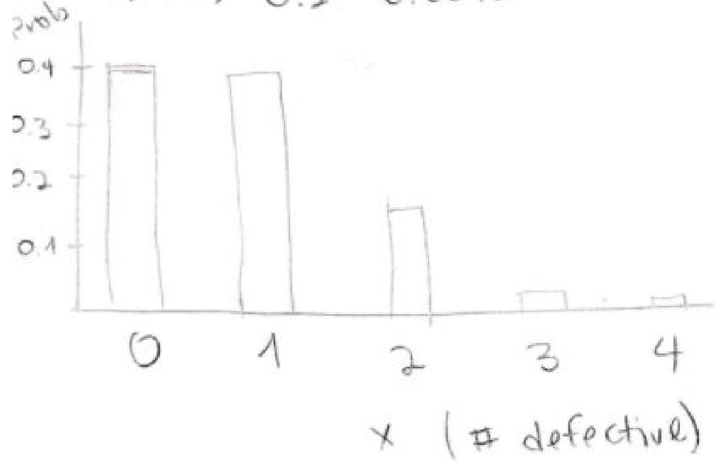
$P_D(0) = 0.8^4 = 0.41$

$P_D(1) = 0.8^3 \cdot 0.2 = (0.10)(4 \text{ combos}) = 0.41$

$P_D(2) = 0.8^2 \cdot 0.2^2 = (0.026)(6 \text{ combos}) = 0.16$

$P_D(3) = 0.8 \cdot 0.2^3 = (0.0064)(4 \text{ combos}) = 0.026$

i) $P_D(4) = 0.2^4 = 0.0016$



ii) $E(x) = \Sigma P(x) \cdot x$

$= 0(0.41) + 1(0.41) + 0.16(2) + 3(0.026) + 4(0.0016)$

$\boxed{0.8144}$

iii) $Var(x) = \Sigma\, x^2 P(X = x) - \mu_x^2$

$1(0.41) + 4(0.16) + 9(0.026) + 16(0.0016) - (.8144)^2$

$\boxed{0.646}$

Problem 4 (35 pts):

A stock solution of acid supplied by a certain vendor contains small amounts of several impurities, including copper and nickel. Let X denote the amount of copper and let Y denote the amount of nickel, in parts per ten million, in a randomly selected bottle of solution. Assume that the joint probability density function of X and Y is given by:

$$f(x,y) = \begin{cases} c(x+y)^2 & 0 < x < 1 \; and \; 0 < y < 1 \\ 0 & otherwise \end{cases}$$

a) Find the value of the constant c so that $f(x,y)$ is a joint density function.

b) Compute the marginal density function $f_X(x)$.

c) Compute the conditional density function $f_{Y|X}(y|x)$.

d) Compute the conditional expectation $E(Y|X = 0.5)$.

e) Are X and Y independent? Explain.

$$f(x,y) = \begin{cases} c(x+y)^2 & 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{otherwise} \end{cases}$$

a) $\int^1\int^1 c(x^2+y^2+2xy)\,dy\,dx = 1$

$$c\int^1 \left[x^2 y + y^3/3 + \frac{2xy^2}{2}\right]_0^1 dx = 1$$

$$c\int^1 x^2 + 1/3 + x \, dx = 1$$

$$c\left[x^3/3 + x/3 + x^2/2\right]_0^1 = 1$$

$$c\left(2/3 + 1/2\right) = 1$$

$$\underline{c = 6/7}$$

b) marginal density of x

$$f_x(x) = \int_0^1 6/7\,(x^2+y^2+2xy)\,dy$$

$$= 6/7\left[x^2y + y^3/3 + 2xy^2/2\right]_0^1$$

$$= 6/7\,(x^2+1/3+x)$$

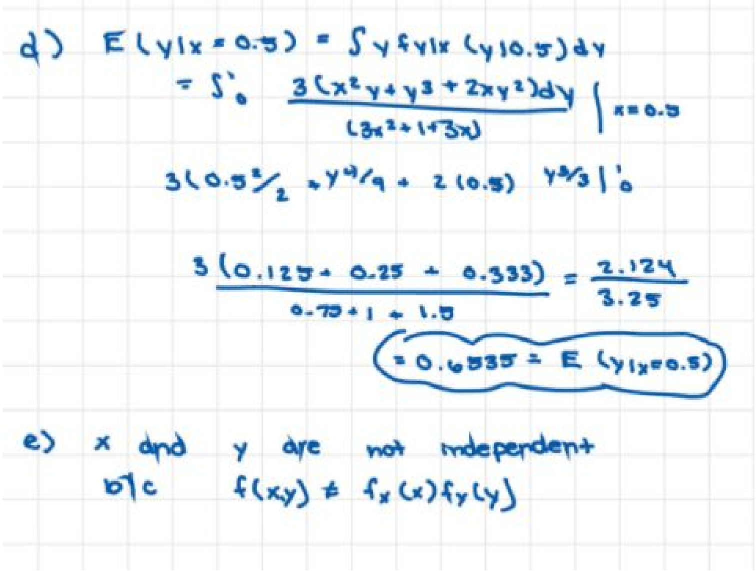
$$\underbrace{f_x(x) = 2/7\,(3x^2+1+3x)}$$

c) conditional pdf of y|x)

$$f_{y|x}(y|x) = \frac{f_{xy}(x,y)}{f(x)} = \frac{6/7\,(x+y)^2}{2/7(3x^2+1+3x)}$$

$$f_{y|x}(y|x) \boxed{\frac{3(x+y)^2}{(3x^2+1+3x)}} \quad 0<y<1$$

8



d) $E(y|x=0.5) = \int y f_{y|x}(y|0.5)\,dy$

$$= \int_0^1 \frac{3(x^2y+y^3+2xy^2)dy}{(3x^2+1+3x)}\Bigg|_{x=0.5}$$

$$3(0.5^2/_2 + y^4/4 + 2(0.5)\ y^3/3 \Big|_0^1$$

$$\frac{3(0.125+0.25+0.333)}{0.75+1+1.5} = \frac{2.124}{3.25}$$

$$\boxed{= 0.6535 = E(y|x=0.5)}$$

e) x and y are not independent

b/c $f(xy) \neq f_x(x)f_y(y)$

Near-perfect solution. Just, on:

4b need to show the interval on which f_x is defined as well

4e need to determine f_y(y) in order to receive full credit