MATH 118: Statistics and Probability **(Due: 07/06/21)**

Homework #2

Instructor: Dr. Zafeirakis Zafeirakopoulos *Name:* Tuba Toprak *Student Id:* 161044116
Assistant: Gizem Süngü *161044116*

Course Policy: Read all the instructions below carefully before you start working on the assignment, and before you make a submission.

- It is not a group homework. Do not share your answers to anyone in any circumstance. Any cheating means at least -100 for both sides.
- Do not take any information from Internet.
- No late homework will be accepted.
- For any questions about the homework, send an email to gizemsungu@gtu.edu.tr.
- Submit your homework (both your latex and pdf files in a zip file) into the course page of Moodle.
- Save your latex, pdf and zip files as "Name_Surname_StudentId".{tex, pdf, zip}.
- The answer which has only calculations without any formula and any explanation will get zero.
- The deadline of the homework is 07/06/20 23:55.
- I strongly suggest you to write your homework on LaTeX. However, hand-written paper is still accepted **IFF** your hand writing is **clear and understandable to read**, and the paper is well-organized. Otherwise, I cannot grade your homework.
- You do not need to write your Student Id on the page above. I am checking your ID from the file name.

Problem 1: (10+10+10+10+10+10+40 = 100 points)

WARNING: Please show your OWN work. Any cheating can be easily detected and will not be graded.

For the question, please follow the file called manufacturing_defects.txt while reading the text below.

In each year from 2000 to 2019, the number of manufacturing defects in auto manufacturers were counted. The data was collected from 14 different auto manufactory companies. The numbers of defects for the companies are indicated in 14 columns following the year column. Assume that the number of manufacturing defects per auto company per year is a random variable having a Poisson(λ) and that the number of defects in different companies or in different years are independent.
(Note: You should implement a code for your calculations for each following subproblem. You are free to use any programming languages (Python, R, C, C++, Java) and their related library.)

(a) Give a table how many cases occur for all companies between 2000 and 2019 for each number of defects (# of Defects).
Hint: When you check the file you will see: # of Defects = {0, 1, 2, 3, 4}.

(b) Estimate λ from the given data.

(c) Update Table 1 in Table 2 with Poisson predicted cases with the estimated λ.

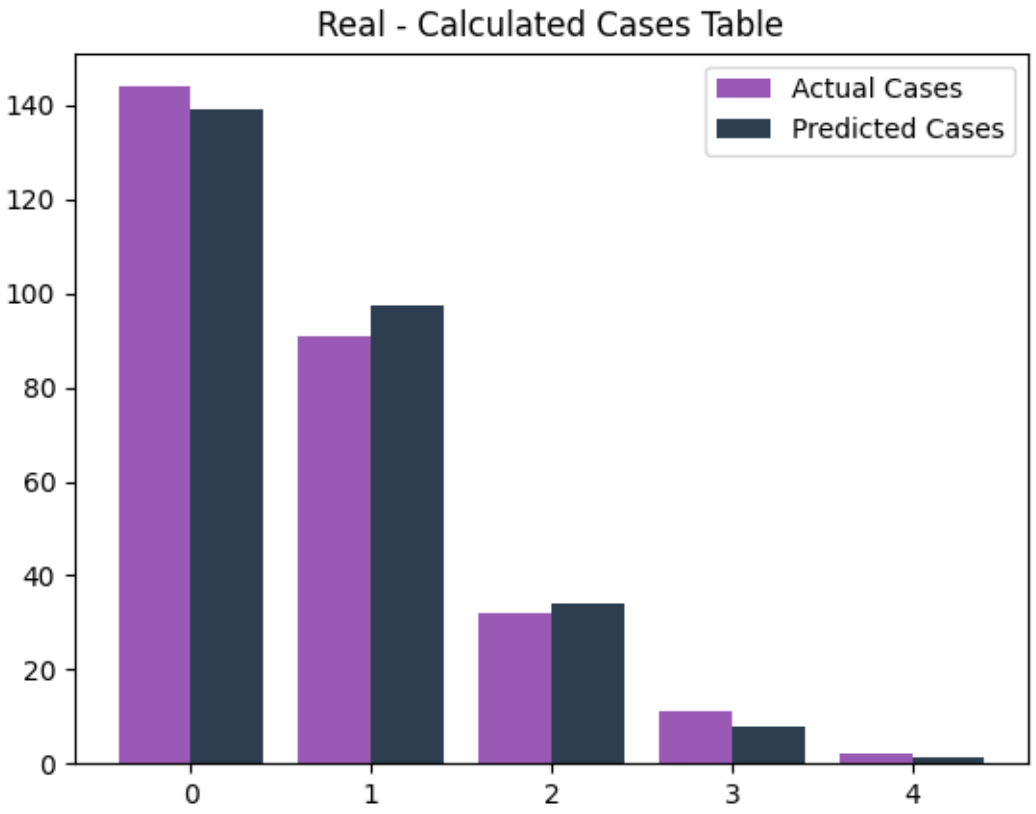
(d) Draw a barplot for the actual cases (Table 2 in column 2) and the predicted cases (Table 2 column 3) with respect to # of defecrs. You should put the figure.

\# of Defects	\# of cases in all company between the years
0	144
1	91
2	32
3	11
4	2

Table 1: Actual cases

\# of Defects	\# of cases in all companies between the years	Predicted \# of cases in all companies between the years
0	144	139.043
1	91	97.330
2	32	34.065
3	11	7.948
4	2	1.391

Table 2: Actual vs. Predicted Cases



(e) According to the barplot in (c), does the poisson distribution fit the data well? Compare the values of the actual cases and the values of the poisson predicted cases, and write your opinions about performance of the distribution.

The fit generally is very good, as might be expected with values of the actual cases and values of the pois-

son predicted cases so close together.According to the barplot, some values increase while some values decrease.
for 0 defects different = % 3.56 fits very good
for 1 defects different = % 6.96 fit very good
for 2 defects different = % 6.44 fit very good
for 3 defects different = % 38.37 fit not good
for 4 defects different = % 43.77 fit not good

(f) According to your estimations above, write your opinions considering your barplot and Table 2. Do you think that road transportation is dangerous for us? Whether yes or no, explain your reason.

yes, I think it's dangerous because when I look at my barplot, I think there should be only 0 cases as there may be loss of life.

(g) Paste your code that you implemented for the subproblems above. Do not forget to write comments on your code.
Example:

- The common code block for all subproblems

```
import pandas as pd
import numpy as numpy
import math
import matplotlib.pyplot as plt
table = pd.read_csv('manufacturing_defects.txt', sep="\t", header=None)  #read to file
manufacturing = len(table.loc[0])-2   # find manifacturing size
time = len( numpy.array(table[1]))           #  find time
numberOfCases = 5    # cases = 0,1,2,3,4
actual_case = []   # to record actual cases
occurrences = 0
case = 0
all = 0
pre_case = []  ## to record pre cases

for x in range(numberOfCases):
    for y in range(2,manufacturing+2):    #to calculate the cases given in the range
        occurrences += numpy.count_nonzero(table[y] == x)   #to calculate how many cases there are
    actual_case.append(occurrences) # for barplot record
    all += occurrences   # to calculate lambda
    case += x*occurrences  # to calculate predicate cases
    occurrences = 0  #reset
```

- The code block for (a)

```
print("# of Defects ------ # of cases in all company between the years")
for x in range(numberOfCases):
    print("      ",x,"            ",end =" ")
    print("            ",actual_case[x],"                                ")  #To print the number of cases
```

- The code block for (b)

```
lamb = (case/all)            #to calculate lambda
print("Lambda: ",lamb)
```

- The code block for (c)

```
print("# of Defects ------ # of cases in all company between the years ------ Predicted \#of cases in all comp
for x in range(numberOfCases):
    print("       ",x,"            ",end =" ")
    print("            ",actual_case[x],"                                ",end =" ")
    pre_case.append(((pow(lamb,x) * math.exp(-lamb)) / (math.factorial(int(x))))*all)  # to calculate  predica
    print(pre_case[x])
```

- The code block for (d)

```
X_axis = numpy.arange(numberOfCases)   # creating the bar plot
plt.bar(X_axis - 0.2, actual_case, 0.4, label = 'Actual Cases',color = "#9B59B6") # to set actual case bar pl
plt.bar(X_axis + 0.2, pre_case, 0.4, label = 'Predicted Cases',color = "#2C3E50")  # to set predicate case ba
plt.title("Real - Calculated Cases Table")  # bar plot title
plt.legend()
plt.show()
```