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

Homework #2

Instructor: Dr. Zafeirakis Zafeirakopoulos *Name:* Ali Bahar *Student Id:* 171044066
Assistant: Gizem Süngü *171044066*

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.
λ = total number of event / total number of cases
λ = total number of defects / total number of cases
total number of defects = (0 * 144) + (1 * 91) + (2 * 32) + (3 * 11) + (4 * 2) = 196
total number of cases = 144 + 91 + 32 + 11 + 2 = 280
λ = 196 / 280 = 0.7

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

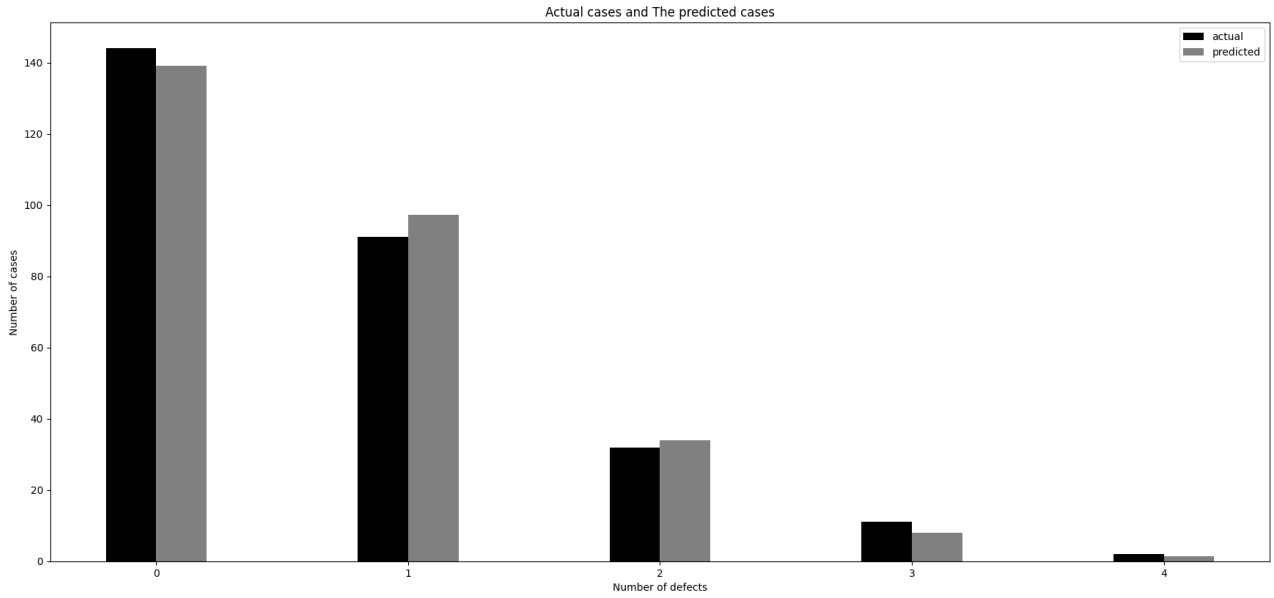
Table 1: Actual cases

\# of Defects	\# of cases in all companies between the years	Predicted \# of cases in all companies between the years
0	144	139.04
1	91	97.33
2	32	34.06
3	11	7.94
4	2	1.39

Table 2: Actual vs. Predicted Cases

(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.



(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.

differences are:

- 144 - 139.04 = 4.96
- 91- 97.33 = -6.33
- 32 - 34.06 = -2.6

- 11 - 7.94 = 3.06
- 2- 1.39 = 0.61

I can say that distribution fit the data well.

(f) According to your estimations above, write your opinions considering your barplot and Table 2. Which company do you prefer to buy a car? Why?

(f Updated) 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.
We do not know that what kind of defects the vehicles have, so i assume that this defect might cause to fatal accident.
We do not know how many vehicle is produced at total.
If these defects could be detected in company, then there is no problem.
If these defects could not be detected in company, we should know how many vehicle is produced at total to be able to say whether it is dangerous or not.

(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
  Paste here. Your code should read the file and compute other things which the following subproblems need.

```
import math
import matplotlib . pyplot as plt
import numpy as np

filep = open("manufacturing_defects.txt", "r")

file_txt = filep.read()

filep.close()

file_lines = file_txt.split('\n')

```

- The code block for (a)
  Paste here. Your code should compute the values in Table 1 column 2.

```
cases = {}
for curr_line in file_lines:
    line_items = curr_line.split('\t')

    for element in line_items[2:]:
        x = int(element)

        if x in cases:
            cases[x] += 1
        else:
            cases[x] = 1

print("\# of Defects\t|\# of cases in all company between the years")
print(64 * "-")
for case in cases:
    print(case, end='')
    print("\t\t|", end='')
    print(cases[case])

```

- The code block for (b)
  Paste here. Your code should compute λ.

```
total_number_of_defects = 0
total_number_of_cases = 0
for case in cases:
    total_number_of_defects += case * cases[case]
    total_number_of_cases += cases[case]

the_lambda = total_number_of_defects / total_number_of_cases

print(f"Total number of defects is {total_number_of_defects}")
print(f"Total number of cases {total_number_of_cases}")
print(f"Lambda is {the_lambda}")

```

- The code block for (c)
  Paste here. Your code should compute the values in Table 2 column 3.

```
predicted_cases = {}
for case in cases:
    predicted_cases[case] = (((math.e ** (-1 * the_lambda)) * (the_lambda ** case)) /
    math.factorial(case)) * total_number_of_cases

print("\# of Defects\t|\# of cases in all company between the years\t|Predicted \# of
    cases in all companies between the years")
print(122 * "-")
for case in cases:
    print(case, end='')
    print("\t\t|", end='')
    print(cases[case], end='')
    print("\t\t\t\t\t\t|", end='')
    print(predicted_cases[case])

```

- The code block for (d)
  Paste here. Your code should draw the barplot.

```
a_range = np.arange(len(cases))

actual_plt = plt.bar(a_range - 0.1, list(cases.values()), 0.2, label='actual', color = '
    #000000')
predicted_plt = plt.bar(a_range + 0.1, list(predicted_cases.values()), 0.2, label='
    predicted', color = '#808080')

plt.xlabel("Number of defects")
plt.ylabel("Number of cases")
plt.title("Actual cases and The predicted cases")
plt.xticks(a_range)
plt.legend()
plt.tight_layout()
plt.show()

```