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

Homework #2

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

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}.

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

Table 1: Actual cases

(b) Estimate λ from the given data.

- total number of defect = (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

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

\# of Defects	\# of cases in all companies between the years	Predicted \# of cases in all companies between the years
0	144	139.04388506159466
1	91	97.33071954311626
2	32	34.06575184009068
3	11	7.948675429354496
4	2	1.3910182001370366

Table 2: Actual vs. Predicted Cases

(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 defects. You should put the figure.

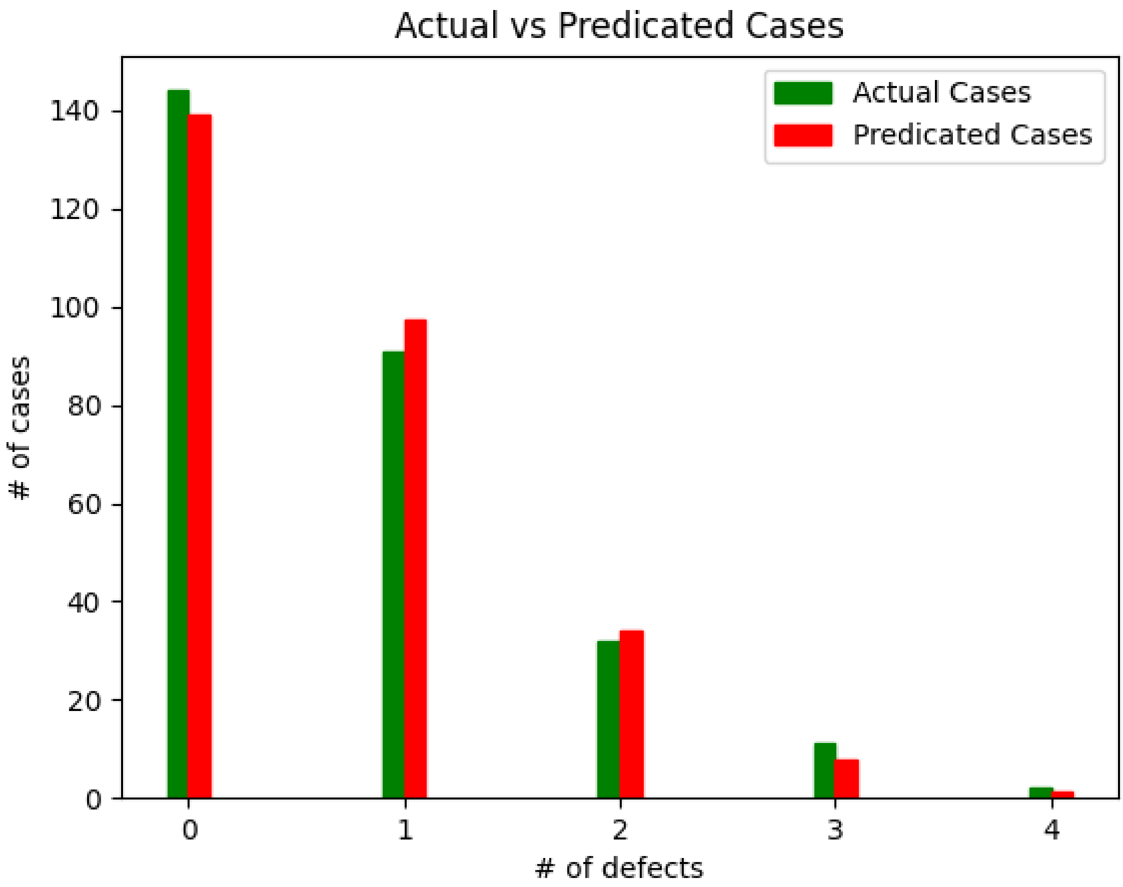

Figure 1: A barplot

(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.

- We should compare the actual cases and predicted cases to determine whether poisson distribution fits the data well or not.
- Number of actual cases: 280
- Number of predicted cases: 279
- As we can see, difference is not big here and we can see that poisson predicts this fairly well. When we evaluate in terms of performance, it can make very close prediction using Poisson distribution. Using the Poisson distribution can be efficient in terms of creating a predicted distribution data.

(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.

- We need to compare the numbers of real cases and predicted cases to decide whether road transportation is dangerous or not.
- According to actual number of cases

  - Total number of crash: 196
  - Total number of crash-free: 144
  - Crash rate = 196 / (144 + 196) = 0.576

- According to predicted number of cases

  - Total number of crash: 191
  - Total number of crash-free: 139
  - Crash rate = 191 / (139 + 191) = 0.578

- As we can see, we can say that the risk of crashes increased and we can say that road transportation is not safe.

(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

Listing 1: The common code - Import modules and File operations

```
import scipy
import matplotlib.pyplot as plot
from scipy import stats
import numpy as np

DEFECT = 5

def read():
    file = open("manufacturing_defects.txt", "r")
    data = []

    for line in file:
        row = []
        for i in line.split():
            if i.isdigit():
                row.append(int(i))
        data.append(row)

    return data

# counts defect for given defect_id
def count_defects(data, defect_id):
    total = 0

    for i in data:
        total += i[1:].count(defect_id)

    return total

data = read()
num_of_company = len(data[0]) - 2 # -2 for first(row number) and second(↩
    year) column
num_of_year = len(data)

# print table
table = print_table1(data)

# find lambda
mean = find_lambda(data, num_of_company, num_of_year)
```

```
    print(mean)

    # find estimations
    estimated_table = estimate(table, mean, num_of_company, num_of_year)
    print(estimated_table)

    # draw barplot
    barplot(table, estimated_table)
```

- The code block for (a)

Listing 2: The code block a - Compute the values in Table 1 and Print table

```
def print_table1(data):
    table = []

    for i in range(0, DEFECT):
        table.append(count_defects(data, i))

    print(table)
    return table
```

- The code block for (b)

Listing 3: The code block b - Compute Lambda

```
def find_lambda(data, num_of_company, num_of_year):
    total = 0

    for r in data:
        for i in range(len(r) - 2):
            total += r[i+2]

    return total / (num_of_company * num_of_year)
```

- The code block for (c)

Listing 4: The code block c - Compute the values in Table 2 and Print table

```
    def estimate(table, mean, num_of_company, num_of_year):
    predicated_cases_poisson = []
    total = 0

    for i in table:
        total += i

    for i in range(len(table)):
        predicated_cases_poisson.append(total * scipy.stats.poisson.pmf(i, mean))

    return predicated_cases_poisson
```

- The code block for (d)

Listing 5: The code block d - Draw the barplot

```
def barplot(table, estimated_table):
    w = 0.1

    real_plt = plot.bar(np.arange(DEFECT), table, w, label = "Actual Cases")
    est_plt = plot.bar(np.arange(DEFECT) + w, estimated_table, w, label = "Predicated Cases")

    for i in range(0, DEFECT):
        real_plt[i].set_color(’g’)
        est_plt[i].set_color(’r’)

    plot.title("Actual vs Predicated Cases")
    plot.xlabel("# of defects")
    plot.ylabel("# of cases")
    plot.xticks(np.arange(DEFECT) + w/2, [i for i in range(0, DEFECT)])
    plot.legend(loc = "best")
    plot.show()
```