MATH 118: Statistics and Probability **(Due: 22/06/20)**

Homework #4

Instructor: Dr. Zafeirakis Zafeirakopoulos *Name: Ahmed Semih Özmekik* *Student Id: 171044039*
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 22/06/20 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 airplane_crashes.txt while reading the text below.

In each year from 1993 to 2012, the number of airplane crashes in airline companies were counted. The data was collected from 14 different airline companies. The numbers of crashes for the airline companies are indicated in 14 columns following the year column. Assume that the number of crashes per airline company per year is a random variable having a Poisson(λ) and that the number of crashes in different airline company 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 1993 and 2012 for each number of crashes (# of Crashes).
Hint: When you check the file you will see: # of Crashes = {0, 1, 2, 3, 4}.

\# of Crashes	\# 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.

This calculation was done with code, you can check below for calculation. λ is 0.7.

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

\# of Crashes	\# 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	34.065751840090684
3	11	7.948675429354493
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 crashes. You should put the figure.

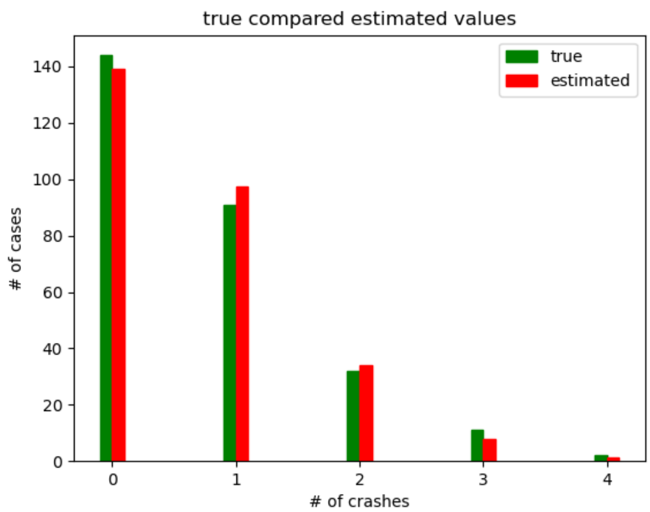

Figure 1: Actual vs. Predicted Cases

(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.

Difference list: [4.95611494, -6.33071954, -2.06575184, 3.05132457, 0.6089818] They are very close, and we can say the distribution fit the data well.

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

In my opinion, airplane transportation seems not dangerous, according to the the poisson distribution = 0.503.

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

CRASH = 5

def input():
    file = open("airplane_crashes.txt", "r")
    X = []
    for line in file:
        stripped_line = line.strip()
        line_list = stripped_line.split()
        X.append(line_list)
    file.close()
    return X
def num_of_case(X, crash_num):
    count = 0
    for line in X:
        count += line[1:].count(str(crash_num))
    return count

X = input()

company_count = len(X[0]) - 2
year_count = len(X)

real_table = print_table_a(X)

# find lambda
mean = find_lambda(X, company_count, year_count)
print(mean)

# find estimations
estimated_table = estimate([i for i in range(len(real_table))], mean, company_count, year_count)
print(estimated_table)

# plot
barplot(real_table, estimated_table)

# difference
print(np.array(real_table) - np.array(estimated_table))
```

- The code block for (a)

```
def print_table_a(X):
    table_a = []
        for i in range(0, CRASH):
            table_a.append(num_of_case(X, i))
    print("{}\t{}".format(i, table_a[i]))
    return table_a
```

- The code block for (b)

```
def find_lambda(X, cc, yc):
    total_case = 0
        for i in range(0, CRASH):
            total_case += i * num_of_case(X, i)
    return total_case / (cc * yc)
```

- The code block for (c)

```
def pdf(X, mean):
r   eturn (math.exp(-mean) * (mean**X) / math.factorial(X))

def estimate(Xe, mean, cc, yc):
    E = []
    for x in Xe:
        E.append(pdf(x, mean) * cc * yc)
    return E
```

- The code block for (d)

```
def barplot(R, E):
    w = 0.1
    np_X = np.arange(CRASH)
    real_plt = plt.bar(np_X, R, w, label="true")
    est_plt = plt.bar(np_X + w, E, w, label="estimated")
    for i in range(0, CRASH):
        real_plt[i].set_color('g')
        est_plt[i].set_color('r')
    plt.title("true compared estimated values")
    plt.xlabel("# of crashes")

    plt.ylabel("# of cases")
    plt.xticks(np_X + w/2, [i for i in range(0, CRASH)])
    plt.legend(loc="best")
    plt.show()
```