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

Homework #2

Instructor: Dr. Zafeirakis Zafeirakopoulos *Name:Akif Kartal* *Student Id:171044098*
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}.

(Solution)

Check Table 1 in next page.

\# 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.

(Solution)

We need to calculate average of the manufacturing defects. Thus;

$$\lambda = \frac{0*144 + 91*1 + 2*32 + 3*11 + 4*2}{144 + 91 + 32 + 11 + 2} = \frac{196}{280} = 0.7$$

(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	139.04
1	91	97.33
2	32	34.07
3	11	7.95
4	2	1.39

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 defecrs. You should put the figure.

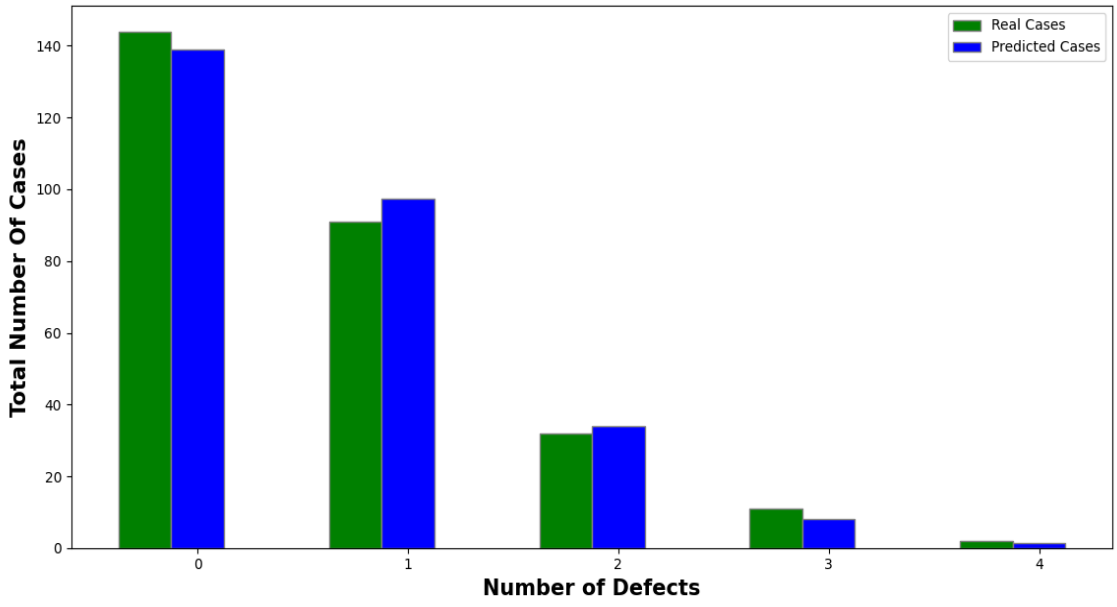

Figure 1: BarPlot

(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.

(Solution)

As can be seen in barplot the actual cases and the values of the poisson predicted cases are very close to each other. This means that poisson distribution fit the data very well. As a result **this is a very good example of poisson distribution**. According to the barplot actual cases is higher than predicted cases in 0, 3 and 4 but since 3 and 4 case number is very low, still poisson 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?

(Solution)

If we observe both barplot and Table 2, we will see that total number of 3 and 4 defects in 20 year is very low. And also from Table 2 we observe that actual cases are very close to predicted cases this means we will not have a surprise or bad actual case numbers for defects. This shows us we can trust road transportation and road transportation is not dangerous for us. Also when we look at the average number of defects for a company in 20 years is 0.7 and also this is very good result in order to trust road transportation. Lastly in road transportation we shouldn't be fear about car company defects but we should be careful about traffic accidents that happens because of people.

(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.

```
//these code blocks are common for all subproblems
class HW2:
  def __init__(self):
        //private class members
        self.__lines = [[]]
        self.__defects = []
        self.__lamda = 0.0
        self.__predictedCases = []

    def get_results(self):

        self.__q1()
        self.__q2()
        self.__q3()
        self.__q4()

  def __readFile(self):
        file = open("manufacturing_defects.txt", "r")
        for line in file:
            if len(line) > 0:
                self.__lines.append(line.split()[2:])
        self.__lines = self.__lines[1:len(self.__lines) - 1]

hw2 = HW2()
hw2.get_results()
```

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

```
  //read file
  def __readFile(self):
  file = open("manufacturing_defects.txt", "r")
  for line in file:
    if len(line) > 0:
      self.__lines.append(line.split()[2:])
  self.__lines = self.__lines[1:len(self.__lines) - 1]

  //calculate number of defects
  def __q1(self):
  self.__readFile()
  for defect in range(5):
    total = 0
    for line in self.__lines:
      for item in line:
        if int(item) == defect:
          total += 1
    self.__defects.append(total)
  print("Number of defects:")
  print(self.__defects)
```

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

```
  def __q2(self):
        numerator = 0
        for i in range(5):
            numerator += i*self.__defects[i]
        denominator = sum(self.__defects)
        self.__lamda = numerator / denominator
        print("lambda:",self.__lamda)
```

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

```
def __q3(self):
  total = sum(self.__defects)
  for i in range(5):
    self.__predictedCases.append(round(total*scipy.stats.poisson.pmf(i ,self.__lamda),2))
  print(self.__predictedCases)
```

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

```
  def __q4(self):
        barWidth = 0.25
        fig = plt.subplots(figsize=(12, 8))
        real = self.__defects
        predict = self.__predictedCases
        br1 = np.arange(len(real))
        br2 = [x + barWidth for x in br1]


        plt.bar(br1, real, color='g', width=barWidth,
                edgecolor='grey', label='Real Cases')
        plt.bar(br2, predict, color='b', width=barWidth,
                edgecolor='grey', label='Predicted Cases')


        plt.xlabel('Number of Defects', fontweight='bold', fontsize=15)
        plt.ylabel('Total Number Of Cases', fontweight='bold', fontsize=15)
        plt.xticks([r + barWidth for r in range(len(real))],
                   ['0', '1', '2', '3', '4'])

        plt.legend()
        plt.show()
```

Result from code pieces is following;

```
Number of defects: [144, 91, 32, 11, 2]
Lambda: 0.7
Predicted Cases: [139.04, 97.33, 34.07, 7.95, 1.39]
```

Figure 2: Question a,b and c

References that was used

- https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.poisson.html
- https://www.geeksforgeeks.org/bar-plot-in-matplotlib/