



JOINT INSTITUTE
交大密西根学院

VE401, Probabilistic Methods in Engineering

Term Project 2:

Police Shootings in the United States

SP2019 Group 17

Group members

Hou Yichun	517370910128
Ma Ziqiao	517370910114
Shen Dinghao	517370910200
Wang Zibo	516021910050
Yang Zhe	517021911127

Instructed by

Dr. Horst Hohberger

April 29, 2019

UM-SJTU Joint Institute

1 Abstract

David Spiegelhalter and Arthur Barnett's *London Murder* analyses murder cases in London between 2004 and 2007. Also, *Washington Post* provides a database recording the occurrence of fatal police shootings. *Washington Post* claims that there have been 270 citizen killed by police shooting since 2019, as of March 15th.

In this project, we are going to use data above, applying mathematical methods we obtained in ve401, to do some data analysis. We will summarize the data sources and conclude how the term "fatal police shooting" is used here. We will use *Mathematica* to create a histogram of numbers of fatal police shooting from 2015 to 2018. We will test the hypothesis that the number of police shootings in the US between 2015 and 2018 follows a Poisson distribution. We will test whether there is evidence that the average number of police shootings depends on the weekday. We will calculate the $(1 - \alpha)$ -confidence interval of parameter k . We will check whether fatal police shooting data between January and March in 2019 follows a Poisson distribution and calculate the value of k . We will obtain 95% prediction intervals for the number of mass shootings in 2019 based on the data for 2015 to 2018, and plot the result.

Key Words Fatal police shootings Poisson distribution Confidence interval Prediction interval Mathematica

Contents

1	Abstract	1
2	Project Introduction	5
2.1	Problem Description	5
2.2	Project Objectives	5
3	Data Analysis	6
3.1	Objectives	6
3.2	Definitions and Notations	6
3.3	Summary of Source of Data	6
3.4	Key Term Characterization	6
3.5	Data Completeness	7
3.6	Date Histogram of Fatal Police Shooting	7
4	Goodness-of-Fit Test for Poisson Distribution Model	8
4.1	Objectives	8
4.2	Definitions and Notations	8
4.3	Goodness-of-Fit Test for 2015-2018	9
4.3.1	Ignoring Pearson Criteria	10
4.3.2	Considering Pearson Criteria	10
4.4	Bar Charts of Observed and Predicted Shooting Numbers	12
5	Dependence of Average Number of Police Shooting on Weekdays	14
5.1	Objectives	14
5.2	Raw Data Processing	14
5.3	Goodness-of-fit for the Average Number of Police Shootings Depends on The Weekday . .	15
5.4	Independence Test for Number of Shoot and Weekdays	17
6	Confidence Interval for Poisson Parameter	19
6.1	Objectives	19
6.2	Definitions and Notations	20
6.3	Verification of Confidence Interval for k	20
6.4	Calculation of Confidence Interval with certain α	21
7	Estimates and Predictions for Police Shooting in 2019	22
7.1	Objectives	22
7.2	Definitions and Notations	22
7.3	Verification of Poisson Distribution Model with Updated Data	22

7.3.1	Ignoring Pearson Criteria	22
7.3.2	Considering Pearson Criteria	23
7.4	Bar Charts of Observed and Predicted Shooting Numbers	25
7.5	Prediction Interval for Number of Fatal Police Shooting	26
7.5.1	Derivation of Nelson's Formula	26
7.6	95% Prediction Intervals for 2019	27
8	Conclusion	30
9	References	31
9.1	Works Cited	31
10	Appendix	31
10.1	Mathematica and Matlab Codes	31

List of Figures

1	The Histogram of Numbers of Fatal Police Shooting	7
2	Feb. 29 th , 2016	8
3	Observed Numbers of Fatal Police Shooting	13
4	Expected Numbers of Fatal Police Shooting	13
5	Number of Occurrence for Different Weekdays	15
6	Observed Numbers of Fatal Police Shooting	25
7	Expected Numbers of Fatal Police Shooting	25
8	Prediction Interval and Updated Observed Data in 2019	28
9	Prediction Interval and Updated Observed Data before Apr. 15th, 2019	29

List of Tables

1	Occurrence per day table	10
2	The Category Probabilities	11
3	The Expected Frequencies	11
4	The Merged Expected and Observed Frequencies	11
5	Number of Occurrence for Different Weekdays	14
6	Average Observed Occurrence	16
7	Average Observed and Expected Occurrence	16
8	Contingency Table	18
9	Expected Frequencies	18
10	Each Adders in Pearson Statistic	19
11	Number of Shoots v.s. Number of Such Days	22
12	The Category Probabilities	23
13	The Expected Frequencies	24
14	The Merged Expected and Observed Frequencies	24

2 Project Introduction

2.1 Problem Description

According to the article *London murders: a predictable pattern?*, published by David Spiegelhalter and Arthur Barnett [1], the pattern of London murders between April 2004 and September 2007 are analyzed based on data of the London Metropolitan Police and figures.

As to this project, we are required to analyze the pattern of Fatal Police Shootings in the U.S. with the similar statistical method from the article *London murders: a predictable pattern?*[1] as well as another article *How The Washington Post is examining police shootings in the United States.* [4].

The analysis is based on the data obtained from the Database of Fatal Police Shootings of the Washington Post [6] from January 2015 to April 2019. It includes processing the data, discussion on what kind of distribution the data belongs to, whether the number of police shootings depends on weekdays and confidence interval for the parameter of the distribution and plotting the corresponding figures with Mathematica.

By finding out the pattern of fatal police shooting, we can predict the number of mass shooting in the future, which may be helpful to public safety in the future as mass shooting has become a social issue recently in the United States.

2.2 Project Objectives

The project is based on *London murders: a predictable pattern?*, published by David Spiegelhalter and Arthur Barnett [1] and the data obtained from the Database of Fatal Police Shootings of the Washington Post [6].

The objectives of the project are listed below:

- Give a thorough and introductory summary of the data of “fatal police shooting”. Plot a Histogram with Mathematica to help the overview of the data.
- Find out whether the occurrence of police shootings in the United States from years 2015 to 2018 follows a Poisson distribution with Pearson’s Chi-squared Goodness-of-Fit Test.
- Find out whether the occurrence of police shootings depends on the weekday with Goodness-of-Fit Test classified by average number of shooting and exact number of shooting. Plot the corresponding figures with Mathematica.
- Prove and calculate the confidence interval for Poisson Parameter with the normal approximation to Poisson distribution.
- Find and give proper estimations and predictions for fatal police shooting numbers based on updated data in 2019. Plot the corresponding figures about prediction interval and observed Data to help discussion.

3 Data Analysis

3.1 Objectives

This section is oriented to question i) and ii).

The objectives of this section is give a thorough and introductory analysis of the data which the project bases on. Especially, we are interested in the term “fatal police shooting” used to describe the data, and we will represent the overview of data by date a histogram.

3.2 Definitions and Notations

Date Histogram DateHistogram is a histogram which features a horizontal axis of DateObjects. This figure gives a clear view of the number of fatal police shooting in a day within a long range of time [3].

3.3 Summary of Source of Data

The following summary cites some factual basis based on *How The Washington Post is examining police shootings in the United States*, officially written by Julie Tate et al [4].

Database Introduction The data we research on in the project comes from a database called the Washington Post, which compiles “a database of every fatal shooting in the United States by a police officer in the line of duty since Jan. 1, 2015”.

Source of Data The Washington Post collects the information by culling

- Local news reports;
- Law enforcement websites;
- Social media;
- Monitoring other independent databases, e.g. *Killed by Police* and *Fatal Encounters*.

We note that do not use the data logged by the *FBI* and the *Centers for Disease Control and Prevention*, because these databases acknowledge the incompleteness of their data.

3.4 Key Term Characterization

The database records only “Fatal Police Shooting”, because only those shootings in which **a police officer, in the line of duty, shoots and kills a civilian will be recorded**. The data are in favour of Black Lives Matter movement and police accountability discussion.

Hence, the Post is not tracking

- Deaths of people in police custody and non-shooting deaths (not shooting);

- Fatal shootings by off-duty officers (not police);
- Police shootings that did not lead to deaths (not fatal).

3.5 Data Completeness

The Post's database is updated regularly as fatal shootings are reported and as facts emerge about individual cases on [GitHub](#) [6]. Hence, we had access to the `csv` file recording all fatal police shootings.

However, we noticed that the database on GitHub is updated only to Feb. 14, 2019. For complete data, we refer to the official website of Washington Post, which reports the latest shooting on Apr. 15, 2019 [7].

3.6 Date Histogram of Fatal Police Shooting

With the application of `Mathematica`, we recreate a version of Figure 1 in [1] for the data between Jan. 1st, 2015 and Dec. 31st, 2018. The plot is a `DateHistogram` with a step length of 1 day, and vertical axis representing the number of fatal police shootings in that day (Figure 1).

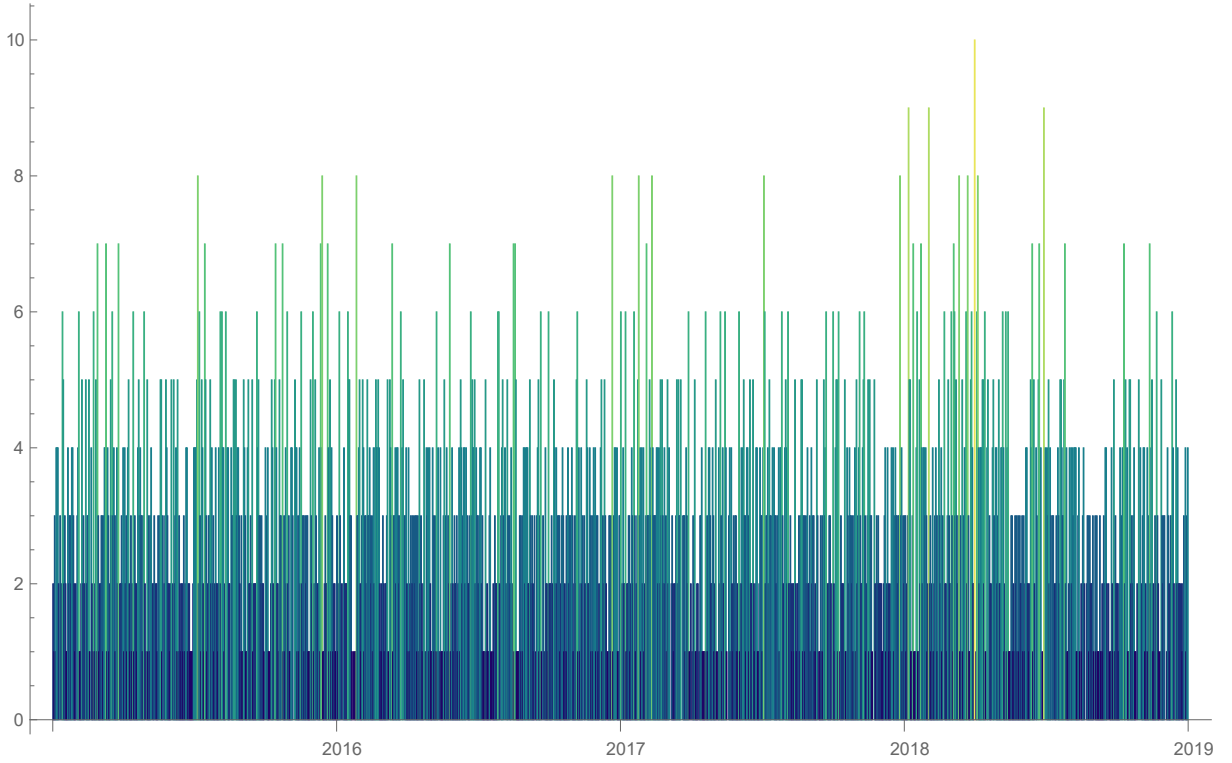


Figure 1: The Histogram of Numbers of Fatal Police Shooting

For convenience of observations, we plot the figure in a way that the lighter the color is, the more shoots observed in that day.

We noticed that 2016 is a leap year with Feb. 29th, check the data with command `Tally`, we observed that there is 1 shooting in Feb. 29th, 2016 (Figure 2).

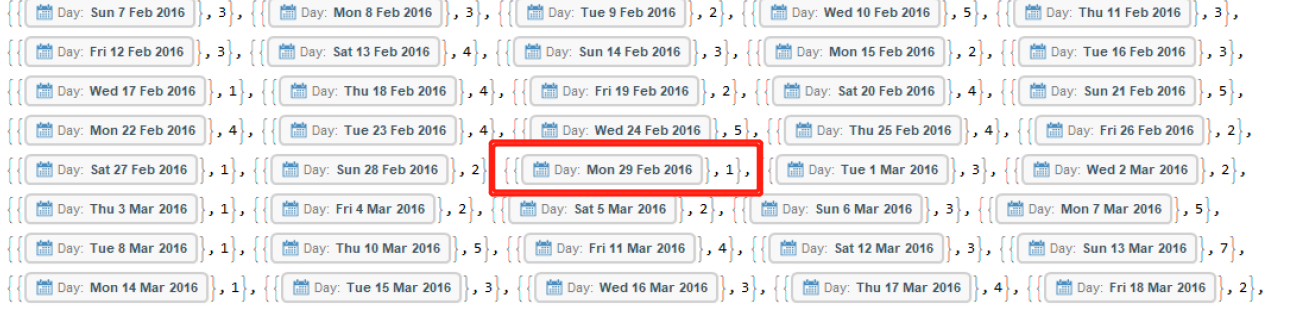


Figure 2: Feb. 29th, 2016

We specificate that this day should be kept.

First, the following discussions are generally based on a unit scale of days. Removing the leap day may cause inaccurate estimation results.

Secondly, `Mathematica` supports the leap day with the plotting system. It will be extra unnecessary work to remove that from the sample.

4 Goodness-of-Fit Test for Poisson Distribution Model

4.1 Objectives

This section is oriented to question iii).

In this section, we will test whether the occurrence of police shootings in US within 2015 to 2018 follows a Poisson distribution. We set the hypothesis:

H_0 : The occurrence of police shootings in US follows a Poisson distribution with parameter k

and wish not to reject it.

4.2 Definitions and Notations

Pearson's Chi-squared Goodness-of-Fit Test [2] Let (X_1, \dots, X_k) be a sample of size n from a categorical random variable with parameters (p_1, \dots, p_k) satisfying Pearson Criteria. Let $(p_{1_0}, \dots, p_{k_0})$ be a vector of null values. Then the test

$$H_0 : p_i = p_{i_0}, \quad i = 1, \dots, k$$

based on

$$\chi^2_{k-1} = \sum_{i=1}^k \frac{(X_i - np_{i_0})^2}{np_{i_0}}$$

is called an chi-squared goodness-of-fit test.

This method allows us to test if the data conforms to an arbitrary discrete or continuous distribution, particularly, Poisson distribution.

Pearson Criteria [2] Large n is required for the chi-squared distribution to be a good approximation to the true distribution of the Pearson statistic. The general criteria is:

- $E[X_i] = np_i \geq 1$ for all $i = 1, \dots, k$;
- $E[X_i] = np_i \geq 5$ for 80% of all $i = 1, \dots, k$.

4.3 Goodness-of-Fit Test for 2015-2018

To test the hypothesis, we first need to process the raw data from the *Database of Fatal Police Shootings* of the Washington Post. First, we screen out the data from 2015 to 2018. Then, we use *Mathematica* to count how many times of police shooting occurs a day, and set them into different categories.

The fulfilling codes with results are:

```
In[ ]:= csvPath =
    "M:\\JI Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Figures\\2\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]]];
For[i = 1, i <= Length[AllDate], i++,
    AllDate[[i]] = DateObject[StringCases[AllDate[[i, 1]],
        x : DatePattern[{"Year", "Month", "Day"}] >= DateList[x]][[1 ;; 1, 1 ;; 3]]];
Count1 = Tally[AllDate];
Count2 = Tally[Count1[[All, 2]]]
Out[ ]:= {{2, 324}, {1, 287}, {3, 310}, {4, 227},
    {6, 53}, {5, 116}, {7, 21}, {8, 11}, {9, 3}, {10, 1}}
```

To know the number of days of zero-shooting, we just need to use the total days minus obtained days. Note that there are 366 days in 2016!

```
In[ ]:= n = 365 + 366 + 365 + 365;
NumZero = n;
For[i = 1, i <= 10, i++, NumZero = NumZero - Count2[[i, 2]]];
Insert[Count2, {0, NumZero}, 1]
Out[ ]:= {{0, 108}, {2, 324}, {1, 287}, {3, 310}, {4, 227},
    {6, 53}, {5, 116}, {7, 21}, {8, 11}, {9, 3}, {10, 1}}
```

Reading from above coding results, we finally obtained the following data table recording occurrence of each numbers of fatal police shooting in a day.

Number of shooting	0	1	2	3	4	5	6	7	8	9	10
Occurrence of such day	108	287	324	310	227	116	53	21	11	3	1

Table 1: Occurrence per day table

4.3.1 Ignoring Pearson Criteria

There is an automatic goodness-of-fit test in `Mathematica`, which ignores the Pearson Criteria. For a rough result, we plug our data table in, and the results turned out to be:

$$\left\{ \{ \mathbf{k1} \rightarrow 2.69884 \}, 6, \frac{\text{Statistic}}{\text{Pearson } \chi^2} \left| \begin{array}{cc} \text{Statistic} & \text{P-Value} \\ 8.06532 & 0.233357 \end{array} \right. \right\}$$

Hence, we can read off that $\hat{k}_{2019} = 2.69884$. The null hypothesis is given by:

H_0 : the number of fatal police shooting follows a Poisson distribution with parameter $\hat{k}_{2019} = 2.70$

The P -value is $p = 0.233$, which is large. Hence, there is no reason to believe that the number of fatal police shooting is not Poisson distributed.

4.3.2 Considering Pearson Criteria

To get the Pearson Statistic, we have to first find an estimator of k , and then calculate out the observed frequency and the expected frequency of number of police shooting occurrence per day. The expected frequency is calculated with expected probability and the total number, while the expected probability can be calculated with the formula of Poisson distribution

$$f(x) = \frac{e^{-\hat{k}} \hat{k}^x}{x!}. \quad (1)$$

Step 1 Find an estimator for k .

The maximum-likelihood estimator for \hat{k} is given by the sample mean. Therefore, we can use the table above to calculate

$$\begin{aligned} \hat{k} &= \bar{X} \\ &= \frac{1}{365 \times 4 + 1} (0 \times 108 + 1 \times 287 + 2 \times 324 + 3 \times 310 + 4 \times 227 \\ &\quad + 5 \times 116 + 6 \times 53 + 7 \times 21 + 8 \times 11 + 9 \times 3 + 10 \times 1) \\ &= 2.6988 \end{aligned}$$

Step 2 To apply the multinomial distribution, we calculate each element:

$$P[X = 0] = \frac{e^{-\hat{k}} \hat{k}^0}{0!} = 0.0673$$

$$P[X = 1] = \frac{e^{-\hat{k}} \hat{k}^1}{1!} = 0.1816$$

$$P[X = 2] = \frac{e^{-\hat{k}} \hat{k}^2}{2!} = 0.2450$$

... ..

$$P[X = 10] = 1 - P[X = 1] - \dots - P[X = 9] = 0.0005$$

The results are shown as:

Num. of shoots	0	1	2	3	4	5	6	7	8	9	10
Category Prob.	0.0673	0.1816	0.2450	0.2204	0.1487	0.0803	0.0361	0.0139	0.0047	0.0014	0.0005

Table 2: The Category Probabilities

Consider the number of shoots as categories, the Category Random Variable is given by (0.0673, 0.1816, 0.2450, 0.2204, 0.1487, 0.0803, 0.0361, 0.0139, 0.0047, 0.0014, 0.0005).

Step 3 Upon this, we can calculate the expected frequencies by

$$E_i = np_i = (365 \times 4 + 1)p_i = 1461p_i.$$

For example, $E_1 = 1461 \times 0.0673 = 98.302$.

All results are given by

Categories	0	1	2	3	4	5	6	7	8	9	10
Expected Freq.	98.302	265.300	358.001	322.061	217.297	117.290	52.758	20.341	6.862	2.058	0.730

Table 3: The Expected Frequencies

Step 4 We noticed that the Pearson Criteria are not satisfied, because $E_{9-10} < 5$ and $E_{10} < 1$. We solve this problem by merging the last 2 categories to obtain:

Categories	0	1	2	3	4	5	6	7	8	9
Expected Freq.	98.302	265.300	358.001	322.061	217.297	117.290	52.758	20.341	6.862	2.788
Observed Freq.	108	287	324	310	227	116	53	21	11	4

Table 4: The Merged Expected and Observed Frequencies

Step 5 Then we can calculate the Pearson statistic:

$$\chi_{N-1-m}^2 = \sum_{i=1}^k \frac{(X_i - np_{i_0})^2}{np_{i_0}} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

The degrees of freedom is found by

$$N - 1 - m = 10 - 1 - 1 = 8,$$

and the statistic is found as

```
In[*]:= (*The Pearson Statistic is found by*)
Stat = 0;
For[i = 1, i ≤ 10, i++,
  Stat = Stat + (NewObservedE[i] - NewExpectedE[i])^2 / NewExpectedE[i];
Stat
Out[*]= 9.9049
```

$$\chi_8^2 = 9.9049$$

Step 6 Fix $\alpha = 0.05$, we test the null hypothesis with the $\chi_{0.05,8}^2 = 15.5073$. We found that

$$\chi_8^2 = 9.9049 < 15.5073 = \chi_{0.05,8}^2$$

Hence we are unable to reject H_0 .

Step 7 At last, we calculate the P -value of this test.

$$\begin{aligned} p &= P[\chi_8^2 \mid H_0] \\ &\leq P[\chi^2 \geq 9.9049] \\ &= 1 - P[\chi^2 \leq 9.9049] \\ &= 0.271764 \end{aligned}$$

The P -value is extremely large at 27% level of significance, thus we fail to reject H_0 .

Conclusions The above calculations and tests proved that

- There is no evidence to believe that Poisson distribution fails describe the number of fatal police shooting in a day. Basically, Poisson distribution gives a good approximation.
- The Poisson parameter $k = 2.6988$ in US fatal police shooting is larger than that of London Homicide $k = 0.44$.

4.4 Bar Charts of Observed and Predicted Shooting Numbers

The barcharts are plotted by *Mathematica* (See next page).

The observed numbers of fatal police shooting:

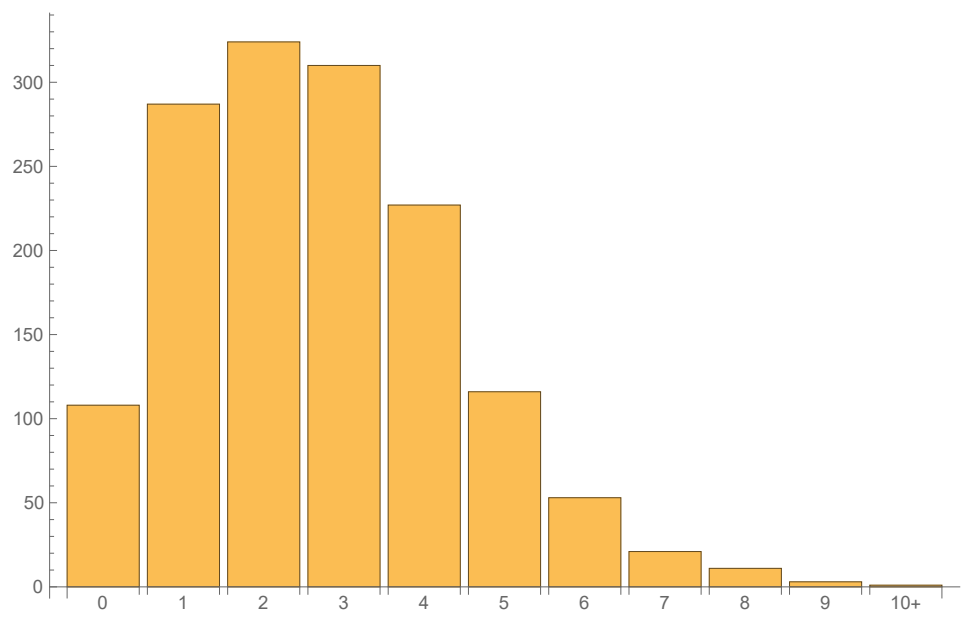


Figure 3: Observed Numbers of Fatal Police Shooting

The expected numbers of fatal police shooting:

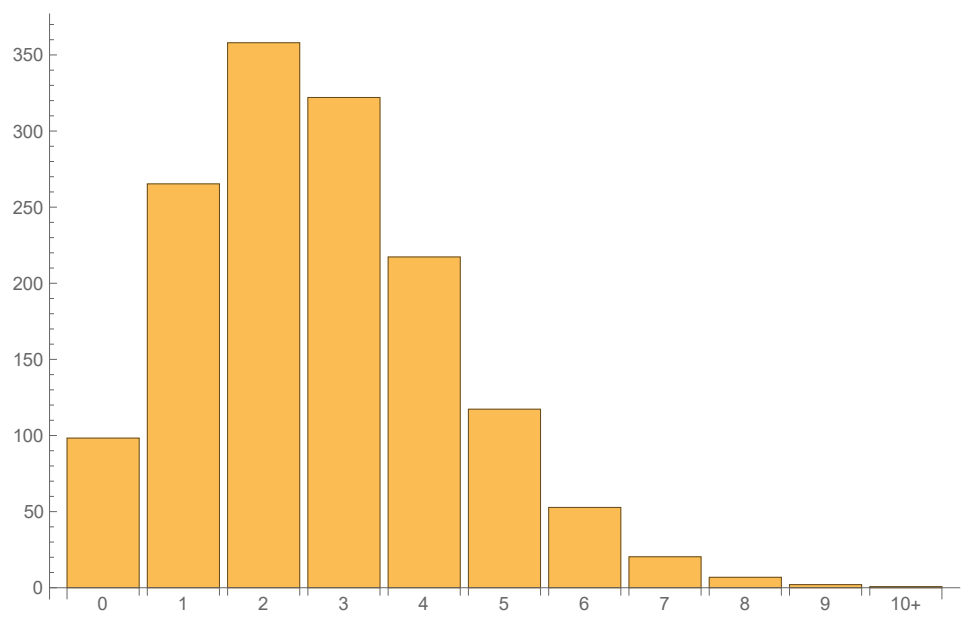


Figure 4: Expected Numbers of Fatal Police Shooting

5 Dependence of Average Number of Police Shooting on Week-days

This section is oriented to question iv).

5.1 Objectives

In this section, we test whether the occurrence of police shootings depends on the weekday. We first test if the average numbers follow a uniform distribution by goodness-of-fit, and then test the dependence of number of shoot and weekdays.

5.2 Raw Data Processing

To test whether the occurrence of police shooting depends on weekday, we first need to process the raw data from the [Database of Fatal Police Shootings](#) of the Washington Post. We use Mathematica to count how many times of police shooting occurs for each weekday, and set them into different categories.

```
csvPath =
  "M:\\JI Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Figures\\2\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]];
For[i = 1, i <= Length[AllDate], i++,
  AllDate[[i]] = DayName[DateObject[StringCases[AllDate[[i, 1]],
    x : DatePattern[{"Year", "Month", "Day"}] -> DateList[x]]][[1 ;; 1, 1 ;; 3]]];
Count1 =
  Tally[
    AllDate]
Out[1223]= {{Friday}, 572}, {{Saturday}, 518}, {{Sunday}, 561},
  {{Monday}, 517}, {{Tuesday}, 586}, {{Wednesday}, 616}, {{Thursday}, 573}}
```

We get the total number of police-shooting occurrence for each weekday. They are 517, 586, 616, 573, 572, 518 and 561 from Monday to Sunday respectively. To make it clearer, we set a table.

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Number of Occurrence	517	586	616	573	572	518	561

Table 5: Number of Occurrence for Different Weekdays

Upon the data, we create a histogram similar to Figure 3 of *London murders: a predictable pattern?* based on data above with Mathematica.

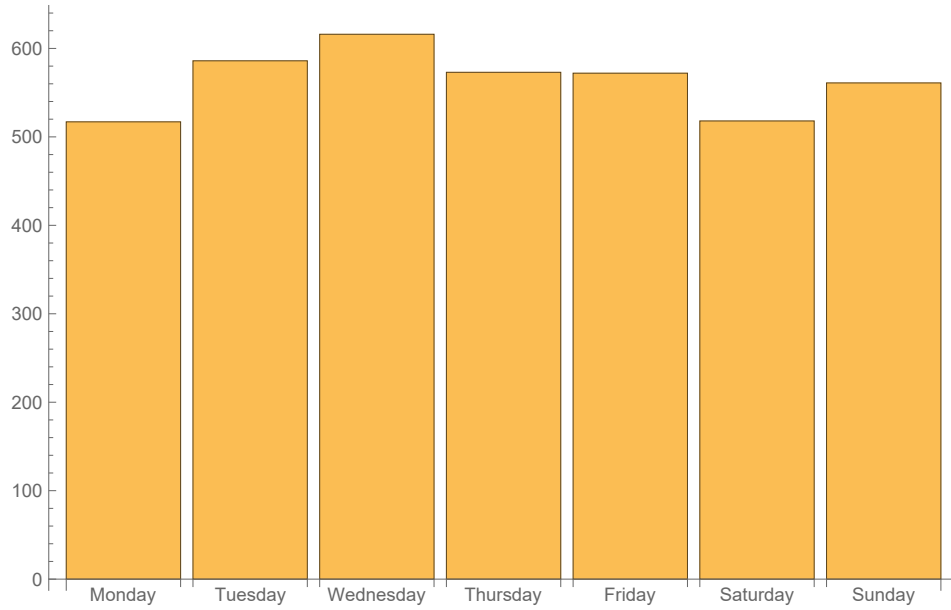


Figure 5: Number of Occurrence for Different Weekdays

The distribution seems almost even, but is it really the case? The following tries to verify it.

5.3 Goodness-of-fit for the Average Number of Police Shootings Depends on The Weekday

Actually, we can interpret our testing object into another form, which is whether the occurrence of police shooting is randomly distributed among weekdays. Therefore, we want to test whether the data are uniformly distributed.

We first set our null hypothesis

$$H_0 : \text{The data follow a multinomial distribution with parameters } (p_1, \dots, p_7) = \left(\frac{1}{7}, \dots, \frac{1}{7}\right)$$

Step 1 We calculate the average occurrence of each weekday, by processing the data above:

$$\text{Average Shoot} = \frac{\text{Total Shoot}}{\text{Number of Days}}$$

In the time period, there is a total of 208 Tuesdays and Wednesdays, and 209 for the rest. By code, we find out the average:


```

In[66]:= AllShoot = Length[AllDate]
AvgShoot = N[AllShoot / (365 * 4 + 1)]
Count4 = Count3;
Count4[[1]] = N[Count3[[1]] / 209];
Count4[[2]] = N[Count3[[2]] / 208];
Count4[[3]] = N[Count3[[3]] / 208];
For[i = 4, i ≤ 7, i++, Count4[[i]] = N[Count3[[i]] / 209]];
Count4

Out[66]= 3943

Out[67]= 2.69884

Out[73]= {2.47368, 2.81731, 2.96154, 2.74163, 2.73684, 2.47847, 2.68421}

```

Hence we received the observed data as the following table.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Total
Observed Occurrence	2.474	2.817	2.962	2.742	2.737	2.478	2.684	18.894

Table 6: Average Observed Occurrence

Step 2 We then find the expected occurrence. This can be found by

$$E_i = np_i$$

For uniform distribution, all E_i s are given by

$$E = np = \frac{3943}{1461} = 2.699$$

Thus we gain a table of expected average occurrence of police shooting for different weekdays.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Total
Observed Occurrence	2.474	2.817	2.962	2.742	2.737	2.478	2.684	18.894
Expected Occurrence	2.699	2.699	2.699	2.699	2.699	2.699	2.699	18.892

Table 7: Average Observed and Expected Occurrence

Step 3 We then find the Pearson statistic.

It follows a chi-squared distribution with

$$7 - 1 = 6$$

degrees of freedom.

The observed test statistic is

$$\sum_{i=1}^7 \frac{(O_i - E_i)^2}{E_i} = \frac{(2.474 - 2.699)^2}{2.699} + \dots + \frac{(2.684 - 2.699)^2}{2.699} = 0.069 \quad (2)$$

Step 4 Fix $\alpha = 0.05$, we test the null hypothesis with the $\chi^2_{0.05,6} = 12.592$. We found that

$$\chi^2_6 = 0.069 < 12.592 = \chi^2_{0.05,6}$$

0.069 is much less than 12.592. Hence we are unable to reject H_0 .

Step 5 At last, we calculate the P -value of this test.

$$\begin{aligned} p &= P[\chi^2_6 \mid H_0] \\ &\leq P[\chi^2 \geq 0.069] \\ &= 1 - P[\chi^2 \geq 0.069] \\ &= 0.999993 \end{aligned}$$

The P -value is extremely large at more than 99% level of significance, thus we fail to reject H_0 .

Therefore, there is no evidence that the average occurrence of police shooting is not randomly distributed.

We draw our final conclusion that there is no evidence that the average number of police shootings depends on the weekday.

5.4 Independence Test for Number of Shoot and Weekdays

But is the number of shoots really independent of weekdays? To find out, we perform independence test. We set H_0 that there is no dependence.

Step 1 We first process the rough data by Mathematica. We want the numbers of the day with certain number of shoots. The results are:

```
In[1316]:= (*Finding contingency tables*)
csvPath = "M:\\JII Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Figures\\2\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]];
For[i = 1, i <= Length[AllDate], i++,
  AllDate[[i]] = DateObject[StringCases[AllDate[[i, 1]], x : DatePattern[{"Year", "Month", "Day"}] => Datelist[x]] [[
    1 ;; 1, 1 ;; 3]]];
Count1 = Tally[AllDate];
For[i = 1, i <= Length[Count1], i++, Count1[[i, 1]] = DayName[Count1[[i, 1]]];
Count2 = Tally[Count1]

Out[ ]:= {{{{Friday}, 2}, 50}, {{{Saturday}, 1}, 48}, {{{Sunday}, 3}, 53}, {{{Monday}, 1}, 51}, {{{Tuesday}, 4}, 43},
{{{Wednesday}, 4}, 27}, {{{Thursday}, 4}, 27}, {{{Tuesday}, 2}, 40}, {{{Wednesday}, 6}, 18}, {{{Thursday}, 5}, 26},
{{{Saturday}, 3}, 41}, {{{Sunday}, 2}, 43}, {{{Tuesday}, 1}, 41}, {{{Thursday}, 2}, 48}, {{{Friday}, 3}, 54},
{{{Monday}, 4}, 30}, {{{Wednesday}, 3}, 51}, {{{Friday}, 1}, 36}, {{{Saturday}, 2}, 50}, {{{Monday}, 3}, 35},
{{{Thursday}, 1}, 39}, {{{Sunday}, 5}, 16}, {{{Monday}, 2}, 48}, {{{Tuesday}, 3}, 41}, {{{Wednesday}, 2}, 45},
{{{Friday}, 5}, 14}, {{{Tuesday}, 5}, 19}, {{{Wednesday}, 1}, 37}, {{{Friday}, 4}, 26}, {{{Sunday}, 1}, 35},
{{{Monday}, 6}, 6}, {{{Saturday}, 7}, 4}, {{{Thursday}, 3}, 35}, {{{Wednesday}, 7}, 2}, {{{Thursday}, 6}, 6},
{{{Saturday}, 4}, 34}, {{{Friday}, 7}, 3}, {{{Wednesday}, 5}, 18}, {{{Saturday}, 5}, 10}, {{{Sunday}, 4}, 40},
{{{Monday}, 5}, 13}, {{{Tuesday}, 8}, 4}, {{{Thursday}, 7}, 6}, {{{Friday}, 6}, 9}, {{{Tuesday}, 6}, 6}, {{{Monday}, 8}, 2},
{{{Monday}, 7}, 4}, {{{Saturday}, 6}, 4}, {{{Wednesday}, 8}, 2}, {{{Sunday}, 7}, 1}, {{{Tuesday}, 7}, 1}, {{{Sunday}, 6}, 4},
{{{Friday}, 8}, 2}, {{{Saturday}, 9}, 1}, {{{Thursday}, 9}, 1}, {{{Sunday}, 10}, 1}, {{{Thursday}, 8}, 1}, {{{Friday}, 9}, 1}}
```

The results are summarized in the contingency table.

Numbers of shoots	0	1	2	3	4	5	6	7	8	9	10	Total
Monday	20	51	48	35	30	13	6	4	2	0	0	209
Tuesday	13	41	40	41	43	19	6	1	4	0	0	208
Wednesday	8	37	45	51	27	18	18	2	2	0	0	208
Thursday	20	39	48	35	27	26	6	6	1	1	0	209
Friday	14	36	50	54	26	14	9	3	2	1	0	209
Saturday	17	48	50	41	34	10	4	4	0	1	0	209
Sunday	16	35	43	53	40	16	4	1	0	0	1	209
Total	108	287	324	310	227	116	53	21	11	3	1	1461

Table 8: Contingency Table

Step 2 Find the expected frequencies.

The expected frequencies are found by

$$E_{ij} = n \cdot \widehat{p_{ij}}$$

Hence the table is given by:

Numbers of shoots	0	1	2	3	4	5	6	7	8	9	10
Monday	15.45	41.06	46.35	44.35	32.47	16.59	7.58	3.00	1.57	0.43	0.14
Tuesday	15.38	40.86	46.13	44.13	32.32	16.51	7.55	2.99	1.57	0.43	0.14
Wednesday	15.38	40.86	46.13	44.13	32.32	16.51	7.55	2.99	1.57	0.43	0.14
Thursday	15.45	41.06	46.35	44.35	32.47	16.59	7.58	3.00	1.57	0.43	0.14
Friday	15.45	41.06	46.35	44.35	32.47	16.59	7.58	3.00	1.57	0.43	0.14
Saturday	15.45	41.06	46.35	44.35	32.47	16.59	7.58	3.00	1.57	0.43	0.14
Sunday	15.45	41.06	46.35	44.35	32.47	16.59	7.58	3.00	1.57	0.43	0.14

Table 9: Expected Frequencies

Step 3 Then we find the Pearson statistics.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

The individual adders are presented by:

Numbers of shoots	0	1	2	3	4	5	6	7	8	9	10
Monday	1.34	2.41	0.06	1.97	0.19	0.78	0.33	0.33	0.12	0.43	0.14
Tuesday	0.37	0.00	0.81	0.22	3.53	0.37	0.32	1.32	3.78	0.43	0.14
Wednesday	3.54	0.36	0.03	1.07	0.87	0.13	14.48	0.33	0.12	0.43	0.14
Thursday	1.34	0.10	0.06	1.97	0.92	5.33	0.33	2.99	0.21	0.76	0.14
Friday	0.14	0.62	0.29	2.10	1.29	0.41	0.27	0.00	0.12	0.76	0.14
Saturday	0.16	1.17	0.29	0.25	0.07	2.62	1.69	0.33	1.57	0.76	0.14
Sunday	0.02	0.89	0.24	1.69	1.74	0.02	1.69	1.34	1.57	0.43	5.13

Table 10: Each Adders in Pearson Statistic

The degrees of freedom is

$$(11 - 1) \times (7 - 1) = 60$$

Hence the statistic is

$$\chi_{60}^2 = \sum_{i=1}^7 \sum_{j=1}^{11} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 85.02$$

Step 4 Fix $\alpha = 0.05$, we test the null hypothesis with the $\chi_{0.05,60}^2 = 79.08$. We found that

$$\chi_6^2 = 85.02 > 79.08 = \chi_{0.05,60}^2$$

Hence can reject H_0 .

We conclude that **although the average number of shooting is independent from weekdays, the exact number of shooting is dependent on weekdays.**

6 Confidence Interval for Poisson Parameter

6.1 Objectives

The section is oriented to question 3).

The objectives of this section are to verify the provided expression of the confidence interval for Poisson Parameter and to calculated it numerically based on the data of "fatal police shooting" from year 2015 to 2018.

For the verification part, we will prove $(1-\alpha)100\%$ confidence interval for k is $\hat{k} \pm z_{\alpha/2} \sqrt{\hat{k}/n}$ with normal approximation to Poisson distribution. The method and available condition of normal approximation to Poisson distribution will be discussed.

For the calculation part, we will apply the \hat{k} and n obtained in question 3) and calculate the confidence interval by the formula verified before under the condition that $\alpha = 0.05$.

6.2 Definitions and Notations

Central Limit Theorem Let X_1, \dots, X_n be a random sample of size n from an arbitrary distribution with mean μ and variance σ^2 . Then under some general conditions, for large n , \bar{X} is approximately normal with mean μ and variance σ^2/n . [2]

Required sample size n for Central Limit Theorem The value of n required to let the Central Limit Theorem provide a good approximation [2]:

1. **Well-behaved** (nearly symmetric densities that look close to that of a normal distribution)

$$n > 4$$

2. **Reasonably behaved** (no prominent mode, densities look like uniform densities)

$$n > 12$$

3. **Ill-behaved** (much of the weight of the densities is in the tails, irregular appearance)

$$n > 4$$

Normal approximation to Poisson distribution For sufficiently large values of n , (say $n > 1000$), the normal distribution with mean k and variance k is an excellent approximation to the Poisson distribution [9].

$$F_{\text{Poisson}}(x; k) \approx F_{\text{normal}}(x; \mu = k, \sigma^2 = k)$$

Confidence Interval Let X_1, \dots, X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2 . [2] A $100(1-\alpha)\%$ confidence interval on μ is given by

$$\bar{X} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$$

.

6.3 Verification of Confidence Interval for k

From the question 3), n is calculated as

$$n = 365 \times 3 + 366 = 1461$$

Due to Central Limit Theorem mentioned in Section 6.2, it is concluded that for $n = 1461 (> 100)$, we can apply a good normal approximation to Poisson distribution.

$n = 1461$ also satisfies the requirement of the normal distribution being an excellent approximation to Poisson distribution— $n > 1000$.

By approximation, we get

$$\mu = k, \quad \sigma^2 = k$$

Hence, the standard deviation is

$$\sigma = \sqrt{k}$$

From Section 6.2, a $100(1-\alpha)\%$ confidence interval for normal distribution on μ is given by

$$\bar{X} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}$$

Since a maximum-likelihood estimator for k is the sample mean, $\hat{k} = \bar{X}$. Also, $\mu = k$ and $\sigma = \sqrt{k}$ hold.

Therefore, a $(1-\alpha)100\%$ confidence interval for k is

$$\begin{aligned} \hat{k} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}} \\ = \hat{k} \pm \frac{z_{\alpha/2} \cdot \sqrt{k}}{\sqrt{n}} \\ = \hat{k} \pm z_{\alpha/2} \cdot \sqrt{\hat{k}/n} \end{aligned}$$

The given expression of a $(1-\alpha)100\%$ confidence interval for k is proved.

6.4 Calculation of Confidence Interval with certain α

Determine $\alpha = 0.05$.

Then $z_{\alpha/2} = z_{0.025} = 1.96$ by reading the value of z from the z -value table.

From question 3), we get \hat{k} and n from the data of "fatal police shooting" from year 2015 to 2018.

$$\hat{k} = \frac{3943}{1461} = 2.6988$$

$$n = 1461$$

Apply the $z_{0.025}$, \hat{k} and n to the formula proved in the Section 6.3, a 95% confidence interval for k is

$$\begin{aligned} \hat{k} \pm z_{\alpha/2} \cdot \sqrt{\hat{k}/n} \\ = 2.6988 \pm 1.96 \cdot \sqrt{2.6988/1461} \\ = 2.6988 \pm 0.0842 \\ = [2.6146, 2.7830] \end{aligned}$$

Therefore, $[2.6146, 2.7830]$ is the 95% confidence interval for k based on the data of the years 2015 to 2018.

7 Estimates and Predictions for Police Shooting in 2019

7.1 Objectives

This section is oriented to question vi) and vii).

The objectives of this section are to find and give proper estimations and predictions for fatal police shooting numbers based on updated data in 2019.

We will first verify the Poisson Distribution Model with which we applied Goodness-of-fit tests in the past sections and failed to reject. Comments and relevant figures will be given.

Then, we refer to *Improved closed-form prediction intervals for binomial and Poisson distributions*[5] published by K.Krishnamoorthy and J.Peng, we introduce the Nelson's Formula for prediction interval. It will be further applied with the given data set and derive 95% prediction intervals with figures.

7.2 Definitions and Notations

Estimate An estimate is a statistical statement on the value of an unknown but fixed population parameter, which in this section, refers to the Poisson parameter \hat{k}_{2019} . [2]

Prediction A prediction is a statistical statement on the value of a random quantity X , which in this section, refers to the number of mass shooting in 2019. A $100(1 - \alpha)\%$ prediction interval is defined by $P[L_1 \leq X \leq L_2] = 1 - \alpha$. [2]

7.3 Verification of Poisson Distribution Model with Updated Data

In this subsection we would like to apply similar goodness-of-fit test for a Poisson distribution, and verify the validity of our previous conclusions.

We sort out the data from the official website using **Mathematica**, and obtained the following table for 2019, between Jan. and Mar.

Number of shoots in a day	0	1	2	3	4	5	6	7	8	9
Number of such days	8	16	22	18	11	7	5	2	0	1

Table 11: Number of Shoots v.s. Number of Such Days

7.3.1 Ignoring Pearson Criteria

Similarly, we first apply the automatic goodness-of-fit test in **Mathematica**, which ignores the Pearson Criteria. The output results are:

$$\left\{ \{k1 \rightarrow 2.73333\}, 4, \frac{\text{Statistic}}{\text{Pearson } \chi^2} \mid \frac{\text{P-Value}}{2.23499 \quad 0.692629} \right\}$$

Hence, we can read off that $\hat{k}_{2019} = 2.73$. The null hypothesis is given by:

H_0 : the number of fatal police shooting follows a Poisson distribution with parameter $\hat{k}_{2019} = 2.73$

The P -value is $p = 0.693$, which is extremely large. Hence, there is no reason to believe that the number of fatal police shooting is not Poisson distributed.

7.3.2 Considering Pearson Criteria

Considering the Pearson Criteria, we perform the following tests.

Step 1 Find an estimator for k .

The maximum-likelihood estimator for \hat{k}_{2019} is given by the sample mean.

$$\begin{aligned}\hat{k}_{2019} &= \bar{X} \\ &= \frac{1}{31 + 28 + 31}(8 \times 0 + 16 \times 1 + 22 \times 2 + \dots + 1 \times 9) \\ &= 2.73\end{aligned}$$

Step 2 To apply the multinomial distribution, we calculate each element:

$$\begin{aligned}P[X = 0] &= \frac{e^{-\hat{k}}\hat{k}^0}{0!} = 0.0650023 \\ P[X = 1] &= \frac{e^{-\hat{k}}\hat{k}^1}{1!} = 0.177673 \\ P[X = 2] &= \frac{e^{-\hat{k}}\hat{k}^2}{2!} = 0.24282 \\ &\dots \quad \dots \\ P[X = 9] &= 1 - P[X = 1] - \dots - P[X = 8] = 0.0021\end{aligned}$$

The results are shown as:

Number of shoots	0	1	2	3	4	5	6	7	8	9
Category probabilities	0.0650	0.1777	0.2428	0.2212	0.1512	0.0826	0.0376	0.0147	0.0050	0.0021

Table 12: The Category Probabilities

Consider the number of shoots as categories, the Category Random Variable is given by (0.0650023, 0.177673, 0.24282, 0.221236, 0.151178, 0.0826438, 0.0376488, 0.014701, 0.00502283, 0.00152545).

Step 3 Upon this, we can calculate the expected frequencies by

$$E_i = np_i = 90p_i.$$

The results are given by

Categories	0	1	2	3	4	5	6	7	8	9
Expected Frequencies	5.850	15.991	21.854	19.911	13.606	7.438	3.388	1.323	0.452	0.187

Table 13: The Expected Frequencies

Step 4 We noticed that the Pearson Criteria are not satisfied, because $E_{6-9} < 5$ and $E_{8-9} < 1$. We solve this problem by merging the last 4 categories to obtain:

Categories	0	1	2	3	4	5	6
Expected Frequencies	5.850	15.991	21.854	19.911	13.606	7.438	5.350
Observed Frequencies	8.000	16.000	22.000	18.000	11.000	7.000	8.000

Table 14: The Merged Expected and Observed Frequencies

Step 5 Then we can calculate the Pearson statistic:

$$\chi_{N-1-m}^2 = \sum_{i=1}^k \frac{(X_i - np_{i_0})^2}{np_{i_0}} = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

The degrees of freedom is found by

$$N - 1 - m = 7 - 1 - 1 = 5,$$

and the statistic is found as

$$\chi_5^2 = 2.81152$$

Step 6 Fix $\alpha = 0.05$, we test the null hypothesis with the $\chi_{0.05,5}^2 = 11.0705$. We found that

$$\chi_5^2 = 2.81152 < 11.0705 = \chi_{0.05,5}^2$$

Hence we are unable to reject H_0 .

Step 7 At last, we calculate the P -value of this test.

$$\begin{aligned} p &= P[\chi_5^2 \mid H_0] \\ &\leq P[\chi^2 \geq 2.81152] \\ &= 1 - P[\chi^2 \geq 2.81152] \\ &= 1 - 0.271 \\ &= 0.729 \end{aligned}$$

The P -value is extremely large at 73% level of significance, thus we fail to reject H_0 .

The above calculations and tests proved that there is no evidence to believe that Poisson distribution fails describe the number of fatal police shooting in a day. Basically, Poisson distribution gives a good approximation.

7.4 Bar Charts of Observed and Predicted Shooting Numbers

The barcharts are plotted by `Mathematica`.

The observed numbers of fatal police shooting:

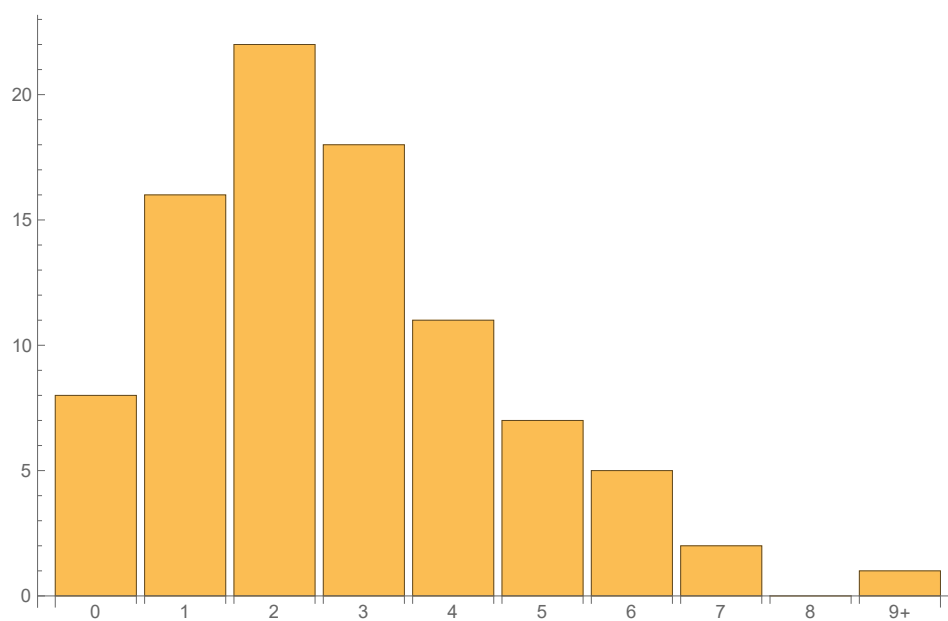


Figure 6: Observed Numbers of Fatal Police Shooting

The expected numbers of fatal police shooting:

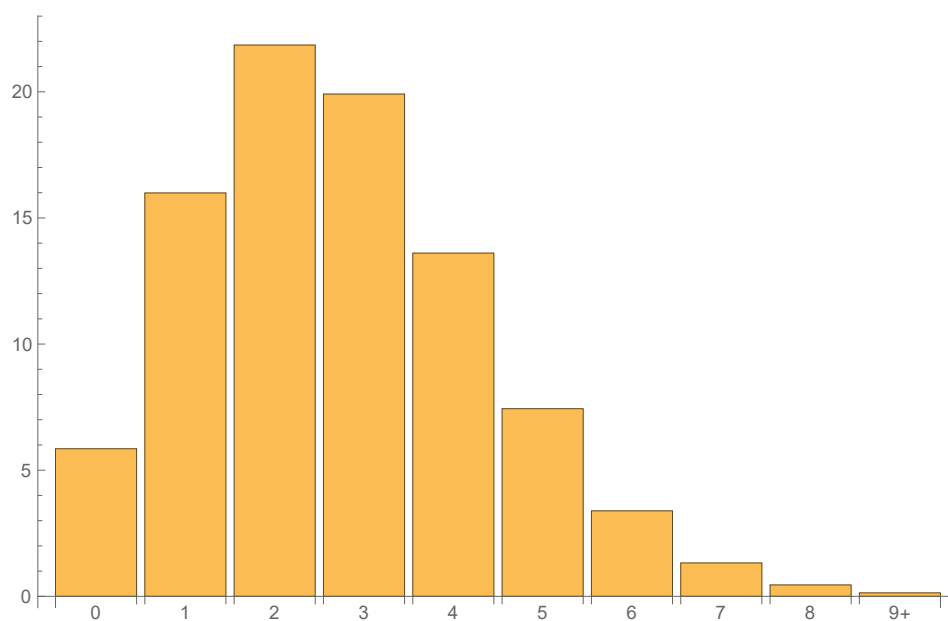


Figure 7: Expected Numbers of Fatal Police Shooting

7.5 Prediction Interval for Number of Fatal Police Shooting

7.5.1 Derivation of Nelson's Formula

The $100(1 - \alpha)\%$ Nelson prediction interval we want to derive is retrieved from [5,(18)]:

$$[[L], [U]] \text{ with } [L, U] = \hat{Y} \pm z_{\alpha/2} \sqrt{m\hat{Y}\left(\frac{1}{m} + \frac{1}{n}\right)}$$

The proof is as follows.

Model Specification [5] Let X be the total counts in a sample of size n from a Poisson distribution with mean λ . Note that $X \sim \text{Poisson}(n\lambda)$. Let Y denote the future total counts that can be observed in a sample of size m from the same Poisson distribution so that $Y \sim \text{Poisson}(m\lambda)$.

The conditional distribution of X , conditionally given $\frac{n}{n+m}$, is binomial with number of trials s and the success probability $\frac{n}{n+m}$, binomial $(s, \frac{n}{n+m})$, and the conditional distribution of Y given the sum $X + Y$ is binomial $(s, \frac{n}{n+m})$ [5].

Estimators [5,8] The estimators used in this formula needs specifications.

- The maximum-likelihood estimator $\hat{\lambda}$ by the sample mean:

$$\hat{\lambda} = \frac{X}{n}$$

- The variance estimate $\widehat{\text{var}}(\hat{Y} - Y)$:

$$\widehat{\text{var}}(\hat{Y} - Y) = m^2 \hat{\lambda} \left(\frac{1}{m} + \frac{1}{n} \right)$$

- Estimate of essentially random Y denoted as \hat{Y} :

$$\hat{Y} = \begin{cases} \frac{mX}{n} & = m\hat{\lambda} & X = 1, 2, \dots \\ \frac{m}{2n} & & X = 0 \end{cases}$$

As X assumes these values with positive probabilities, the coverage probabilities of the above PI are expected to be much smaller than the nominal level when p of the binomial model is at the boundary $p = 0$ and $p = n$ [8]. In Poisson model, when $p = 0$ the approximation parameter $\lambda = np = 0$ is at boundary as well. To overcome the poor coverage probabilities at the boundary, we can define $\hat{\lambda} = \frac{0.5}{n}$ when $X = 0$.

Statistic The statistic we use here the asymptotic result which is standard normal:

$$Z = \frac{m\hat{\lambda} - Y}{\sqrt{\text{var}(m\hat{\lambda} - Y)}}$$

Now, a $100(1 - \alpha)\%$ prediction interval is given by:

$$\begin{aligned}
1 - \alpha &= P[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}] \\
1 - \alpha &= P[-z_{\alpha/2} \leq \frac{m\hat{\lambda} - Y}{\sqrt{\text{var}(m\hat{\lambda} - Y)}} \leq z_{\alpha/2}] \\
1 - \alpha &= P[-z_{\alpha/2} \leq \frac{Y - m\hat{\lambda}}{\sqrt{\text{var}(m\hat{\lambda} - Y)}} \leq z_{\alpha/2}] \\
1 - \alpha &= P[-z_{\alpha/2}\sqrt{\text{var}(m\hat{\lambda} - Y)} \leq Y - m\hat{\lambda} \leq z_{\alpha/2}\sqrt{\text{var}(m\hat{\lambda} - Y)}] \\
1 - \alpha &= P[-z_{\alpha/2}\sqrt{\text{var}(m\hat{\lambda} - Y)} + m\hat{\lambda} \leq Y \leq z_{\alpha/2}\sqrt{\text{var}(m\hat{\lambda} - Y)} + m\hat{\lambda}] \\
1 - \alpha &= P[-z_{\alpha/2}\sqrt{m\hat{Y}(\frac{1}{m} + \frac{1}{n})} + \hat{Y} \leq Y \leq z_{\alpha/2}\sqrt{m\hat{Y}(\frac{1}{m} + \frac{1}{n})} + \hat{Y}]
\end{aligned}$$

Hence, taking the ceiling of the lower prediction bound and the floor of the upper prediction bound for integer,

$$[[L], [U]] \text{ with } [L, U] = \hat{Y} \pm z_{\alpha/2}\sqrt{m\hat{Y}(\frac{1}{m} + \frac{1}{n})}$$

7.6 95% Prediction Intervals for 2019

This formula allows us to find a prediction interval of the number of mass shootings in 2019. We first explain for each notations.

- X : the number of mass shootings between 2015 and 2018;

$$X = 3943$$

- Y : the number of mass shootings in 2019;
- n : number of days between 2015 and 2018;

$$n = 1461$$

- m : the number of days considered in 2019;

Hence we can calculate \hat{Y} by

$$\hat{Y} = \frac{mX}{n} = \frac{3943}{1461}m \approx 2.699m$$

Plug in the above data, we can find the prediction interval formula:

$$\begin{aligned}
[L] &= \hat{Y} - z_{\alpha/2}\sqrt{m\hat{Y}(\frac{1}{m} + \frac{1}{n})} \\
[U] &= \hat{Y} + z_{\alpha/2}\sqrt{m\hat{Y}(\frac{1}{m} + \frac{1}{n})}
\end{aligned}$$

The calculations are performed with $\alpha = 0.05$, i.e. $z_{\alpha/2} = 1.96$:

$$\begin{aligned}
L &= \hat{Y} - z_{\alpha/2} \sqrt{m\hat{Y}\left(\frac{1}{m} + \frac{1}{n}\right)} \\
&= \frac{3943}{1461}m - 1.96 \sqrt{\frac{3943}{1461}m^2\left(\frac{1}{m} + \frac{1}{1461}\right)} \\
&= 2.69884m - 3.21991 \sqrt{0.000684463m^2 + m} \\
\lceil L \rceil &= \lceil 2.69884m - 3.21991 \sqrt{0.000684463m^2 + m} \rceil
\end{aligned}$$

$$\begin{aligned}
U &= \hat{Y} + z_{\alpha/2} \sqrt{m\hat{Y}\left(\frac{1}{m} + \frac{1}{n}\right)} \\
&= \frac{3943}{1461}m + 1.96 \sqrt{\frac{3943}{1461}m^2\left(\frac{1}{m} + \frac{1}{1461}\right)} \\
&= 2.69884m + 3.21991 \sqrt{0.000684463m^2 + m} \\
\lfloor U \rfloor &= \lfloor 2.69884m + 3.21991 \sqrt{0.000684463m^2 + m} \rfloor
\end{aligned}$$

The plot is performed by **Mathematica**, and the figure is as follows.

For the whole year of 2019, the plot gives:

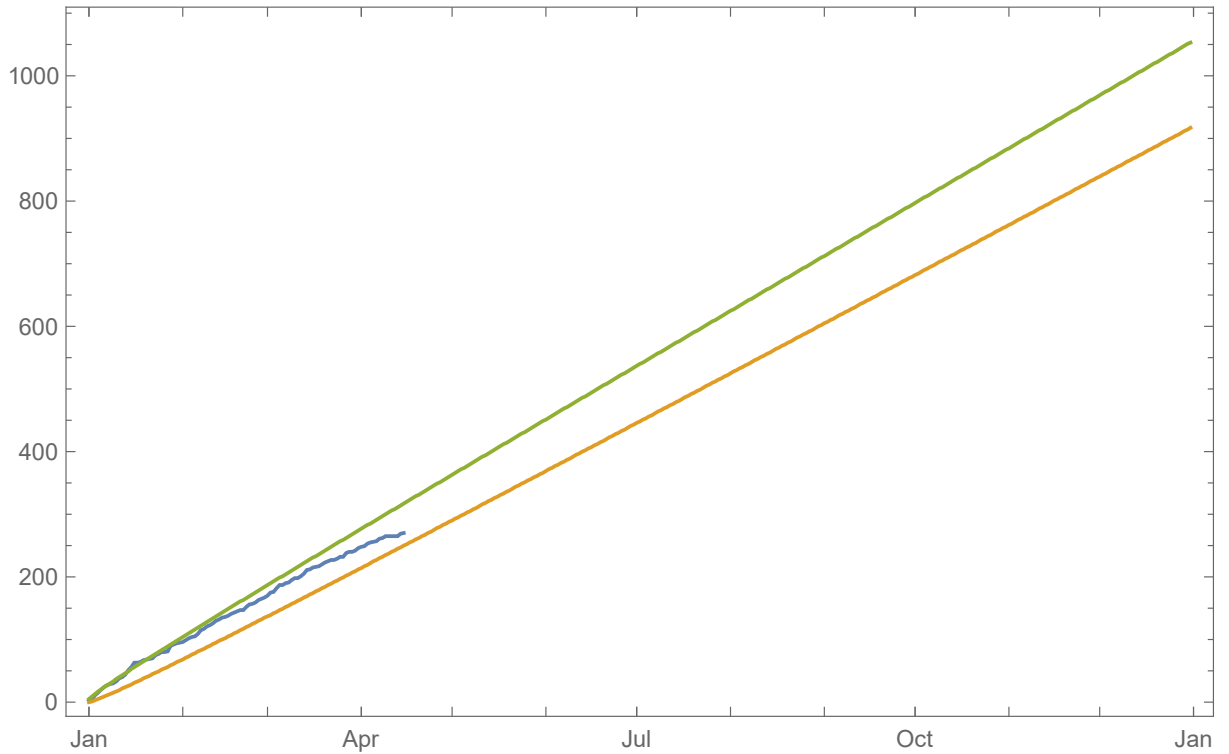


Figure 8: Prediction Interval and Updated Observed Data in 2019

For the clearer view, we plot within the updated dates in 2019:

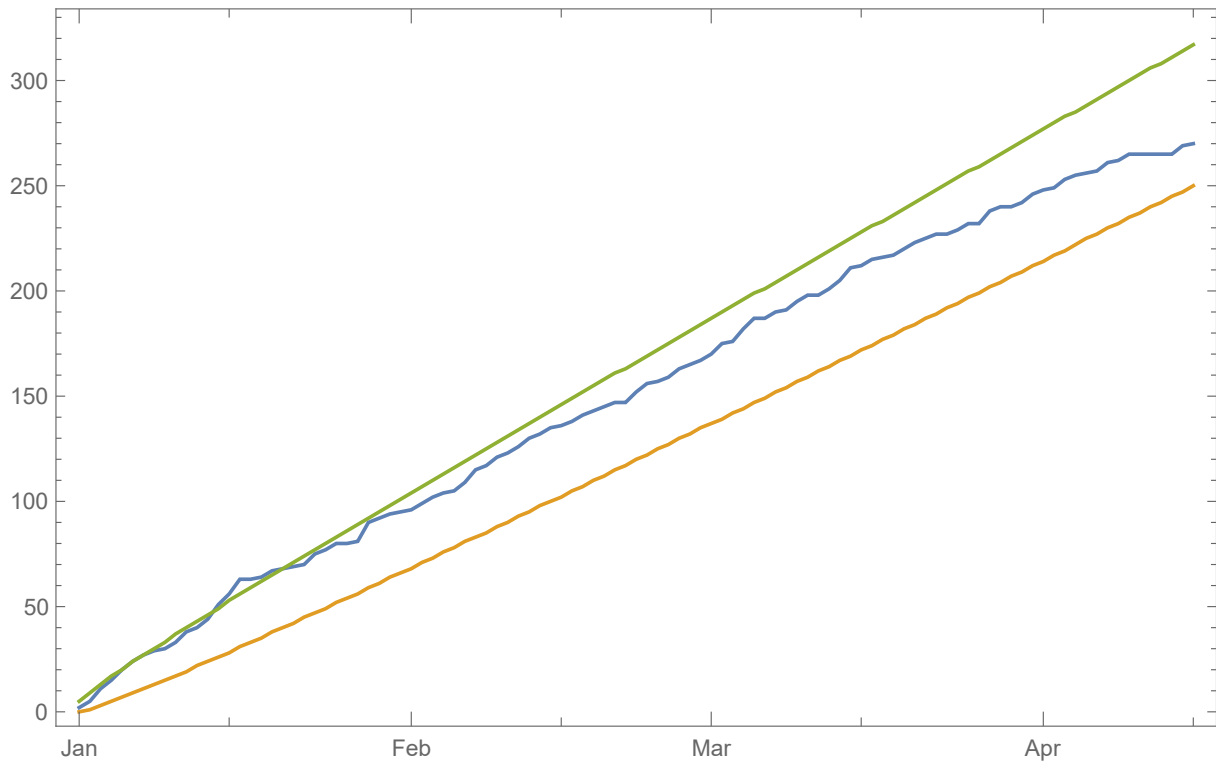


Figure 9: Prediction Interval and Updated Observed Data before Apr. 15th, 2019

Comments With the graphs above, we basically conclude that the prediction interval gives a good prediction for the mass shooting in 2019, because the updated observed data generally lie within the band.

However, in the middle of January, the observed total shootings rise above the upper bound. This is statistically acceptable, because the confidence level is 95%.

We observed that at beginning, the observed data is closer to the upper bound. However, as the numbers go up, the observed data track back to the middle. This is because only with large sample size, an approximate normal distribution is acceptable for the estimator of k .

8 Conclusion

Analogous to *London murders: a predictable pattern?* by David Spiegelhalter and Arthur Barnett, we develop our project based on the *Database of Fatal Police Shootings* of the Washington Post.

We first plot the everyday police shooting data between January 1st, 2015 and December 31st, 2015 to gain a straightforward impression.

Then we test whether the occurrence of fatal police shooting from 2015 to 2018 follows a Poisson distribution, and our conclusion is that there's no evidence that it does not follow a Poisson distribution.

Someone believes that weekday may influence the fatal police shooting rate, so we do a test based on Pearson statistic and draw a conclusion that the average number of shooting is independent of weekdays.

Since we state that the police shooting from 2015 to 2018 follows a Poisson distribution, we are interested in the value of k . Our result is that there're at least 95% possibility that k is in the interval [2.6146, 2.7830]. With the training data from 2015 to 2018, we state the fatal police shooting follows a Poisson distribution.

Then we use testing data from 2019 to test our statement. These data pass the test, which indicates that the data follows a Poisson distribution, and we calculate out the value of \hat{k}_{2019} as 2.73. Finally, we get the 95% prediction intervals for the number of mass shootings in 2019 based on the data of 2015 to 2018, which is shown in last section.

In conclusion, we make use of fatal police shooting data from 2015 to 2018 to set a model and do prediction on 2019. That's a basic method to get preparation for what will happen in the future based on statistic. We hope this project can contribute a little for the analysis of fatal police shooting in US.

9 References

9.1 Works Cited

- [1] D. Spiegelhalter and A. Barnett. London murders: a predictable pattern? *Significance*, 6(1):5-8, 2009. <http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2009.00334.x/abstract>. [Online; accessed 15-April-2019].
- [2] H. Hohberger. "ve401_main.pdf"(2018). UMJI-SJTU, Shanghai. [Online; accessed 25-Feb-2019].
- [3] DateHistogram. Wolfram. 2019. <https://reference.wolfram.com/language/ref/DateHistogram.html>.
- [4] J. Tate, J. Jenkins, S. Rich, J. Muyskens, K. Elliott, T. Mellnik and A. Williams. How The Washington Post is examining police shootings in the United States. *The Washington Post*. July 7, 2016. https://www.washingtonpost.com/national/how-the-washington-post-is-examining-police-shootings-in-the-2016/07/07/d9c52238-43ad-11e6-8856-f26de2537a9d_story.html?utm_term=.d357cd563cc4. [Online; accessed 15-April-2019].
- [5] K.Krishnamoorthy and J.Peng. Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference*, 141(5):1709-1718, 2011. <http://www.sciencedirect.com/science/article/pii/S0378375810005215>. [Online; accessed 15-April-2019].
- [6] Washington Post. Data-police-shootings. GitHub. <https://github.com/washingtonpost/data-police-shootings>. [Online; accessed 15-April-2019].
- [7] Washington Post. Police Shootings 2019. https://www.washingtonpost.com/graphics/2019/national/police-shootings-2019/?utm_term=.b08ecc24520c. [Online; accessed 21-April-2019].
- [8] H. Wang. 2008. Coverage probability of prediction intervals for discrete random variables. *Computational Statistics and Data Analysis* 53, 17-26.
- [9] Wikipedia. Poisson Distribution, *wikipedia*, the free encyclopedia, 2019. https://en.wikipedia.org/wiki/Poisson_distribution [Online; accessed 20-April-2019].

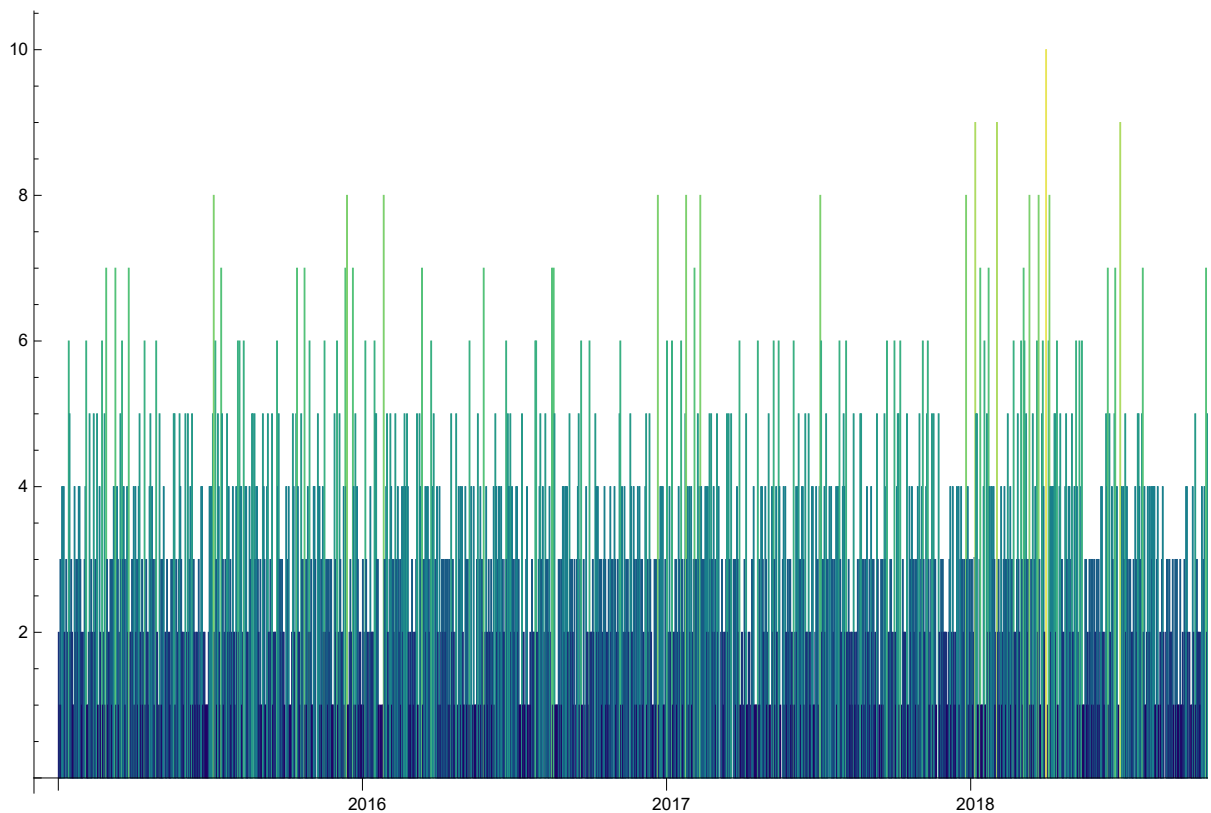
10 Appendix

10.1 Mathematica and Matlab Codes


```

In[569]:= csvPath =
  "M:\\\\JI Courses\\\\Sophomore_Year\\\\2019_Spring\\\\VE401\\\\Proj2\\\\Figures\\\\2\\\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]]];
For[i = 1, i <= Length[AllDate], i++, AllDate[[i]] =
  DateObject[StringCases[AllDate[[i, 1]], x : DatePattern[{"Year", "Month", "Day"}] >=
    DateList[x]][[1 ;; 1, 1 ;; 3]][[1]]];
DateHistogram[AllDate, "Day",
  ColorFunction -> "BlueGreenYellow",
  DateTicksFormat -> {" ", " ", "Year"}]

```



```

In[*]:= csvPath =
  "M:\\JI Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Figures\\2\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]];
For[i = 1, i <= Length[AllDate], i++,
  AllDate[[i]] = DateObject[StringCases[AllDate[[i, 1]],
    x : DatePattern[{"Year", "Month", "Day"}] -> DateList[x]][[1 ;; 1, 1 ;; 3]]];
Count1 = Tally[AllDate];
Count2 = Tally[Count1[[All, 2]]]

Out[*]= {{2, 324}, {1, 287}, {3, 310}, {4, 227},
  {6, 53}, {5, 116}, {7, 21}, {8, 11}, {9, 3}, {10, 1}}

(*Occurrence of number of shoot in a day*)
n = 365 + 366 + 365 + 365;
NumZero = n;
For[i = 1, i <= 10, i++, NumZero = NumZero - Count2[[i, 2]];
Insert[Count2, {0, NumZero}, 1]

Out[*]= {{0, 108}, {2, 324}, {1, 287}, {3, 310}, {4, 227},
  {6, 53}, {5, 116}, {7, 21}, {8, 11}, {9, 3}, {10, 1}}

In[306]:= (*The fit done by Mathematica*)
Needs["HypothesisTesting`"];
TestData = Join[Table[0, {i, 1, 108}], Table[1, {i, 1, 287}],
  Table[2, {i, 1, 324}], Table[3, {i, 1, 310}], Table[4, {i, 1, 227}],
  Table[5, {i, 1, 116}], Table[6, {i, 1, 53}], Table[7, {i, 1, 21}],
  Table[8, {i, 1, 11}], Table[9, {i, 1, 3}], Table[10, {i, 1, 1}]];
PearsonChiSquareTest[TestData, PoissonDistribution[k1],
  {"FittedDistributionParameters", "DegreesOfFreedom", "TestDataTable"}]

Out[308]= {{k1 -> 2.69884}, 6, 

|                  |                  |
|------------------|------------------|
| Statistic        | P-Value          |
| Pearson $\chi^2$ | 8.06532 0.233357 |

}

(*The fit done considering Pearson Criteria 3.1.2*)

In[325]:= Count2 = {{0, 108}, {1, 287}, {2, 324}, {3, 310},
  {4, 227}, {5, 116}, {6, 53}, {7, 21}, {8, 11}, {9, 3}, {10, 1}};
k2 = 0;
For[i = 1, i <= 11, i++, k2 = k2 + Count2[[i, 1]] * Count2[[i, 2]];
k2 = N[k2/n]

Out[328]= 2.69884

In[329]:= e = N[Exp[1]];
CRV = {};
For[i = 0, i <= 10, i++, CRV = Insert[CRV, e^(-k2) * k2^i / Factorial[i], -1]];
CRV[[11]] = 1;
For[i = 1, i <= 10, i++, CRV[[11]] = CRV[[11]] - CRV[[i]]];
CRV

Out[334]= {0.0672838, 0.181588, 0.245038, 0.220439, 0.148732,
  0.0802808, 0.0361108, 0.0139224, 0.0046968, 0.00140843, 0.000499723}

```

```

In[335]:= ExpectedE = {};
For[i = 0, i ≤ 10, i++, ExpectedE = Insert[ExpectedE, n * CRV[[i + 1]], -1]];
ExpectedE

Out[337]= {98.3016, 265.3, 358., 322.062, 217.298,
117.29, 52.7579, 20.3407, 6.86203, 2.05772, 0.730095}

In[ ]:= (*We see that the Pearson Criteria 3.1.2 is not satisfied,
therefore we merge the last 2 categories and obtain the following*)
Count3 = {{0, 108}, {1, 287}, {2, 324},
{3, 310}, {4, 227}, {5, 116}, {6, 53}, {7, 21}, {8, 11}, {9, 4}};
NewCRV = {0.06728375768633715`, 0.18158785527530966`, 0.24503795802551198`,
0.220439121262741`, 0.1487322818512984`, 0.08028081962213136`,
0.03611079988250787`, 0.01392244880578161`, 0.004696801475119516`,
0.0014084332052928933` + 0.000499722907968528`};
NewExpectedE = {98.30156997973859`, 265.29985655722743`, 358.000456675273`,
322.0615561648646`, 217.29786378474694`, 117.29027746793392`, 52.757878628344`,
20.34069770524693`, 6.862026955149613`, 2.057720912932917` + 0.7300951685420194`};
NewObservedE = {108, 287, 324, 310, 227, 116, 53, 21, 11, 3 + 1};
(*The Degree of Freedom is found by*)
DegreesOfFreedom = Length[NewExpectedE] - 1 - 1

Out[ ]:= 8

In[ ]:= (*The Pearson Statistic is found by*)
Stat = 0;
For[i = 1, i ≤ 10, i++,
Stat = Stat + (NewObservedE[[i]] - NewExpectedE[[i]])^2 / NewExpectedE[[i]]];
Stat

Out[ ]:= 9.9049

(*Fix alpha=0.05, we test the null hypothesis, and fail to reject it*)

In[ ]:= Solve[CDF[ChiSquareDistribution[DegreesOfFreedom], x] == 0.95, x]


$$\text{Solve: Solve was unable to solve the system with inexact coefficients. The answer was obtained by solving a corresponding exact system and numericizing the result.}$$


Out[ ]:= {{x → 15.5073}}

In[ ]:= (*The P value is found by:*)

In[ ]:= PValue = 1 - CDF[ChiSquareDistribution[DegreesOfFreedom], Stat]

Out[ ]:= 0.271764

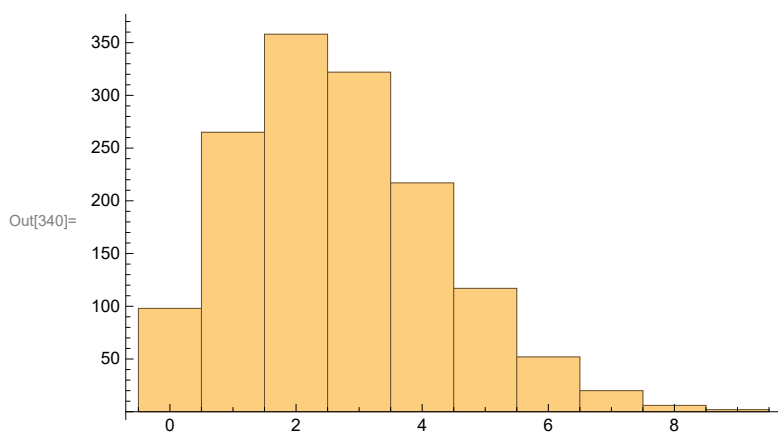
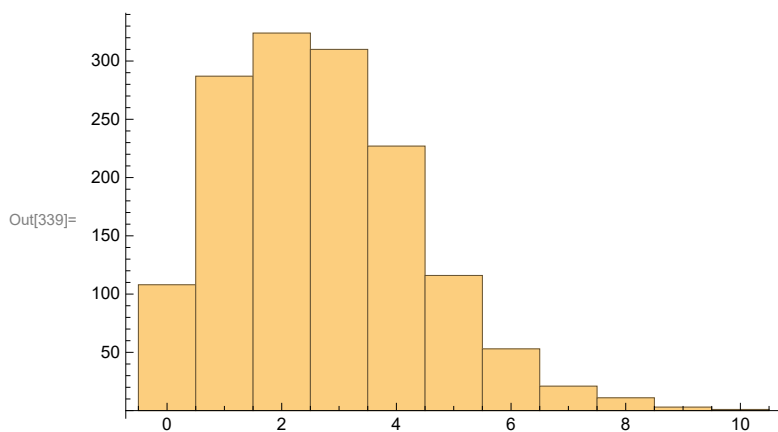
(*The Histograms are created as*)

```

```

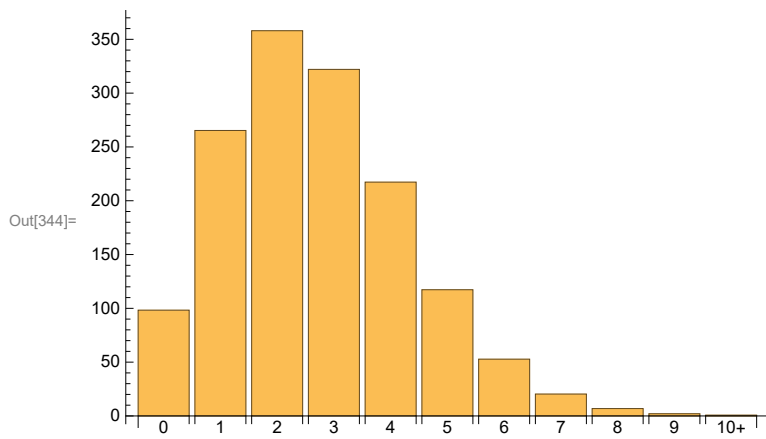
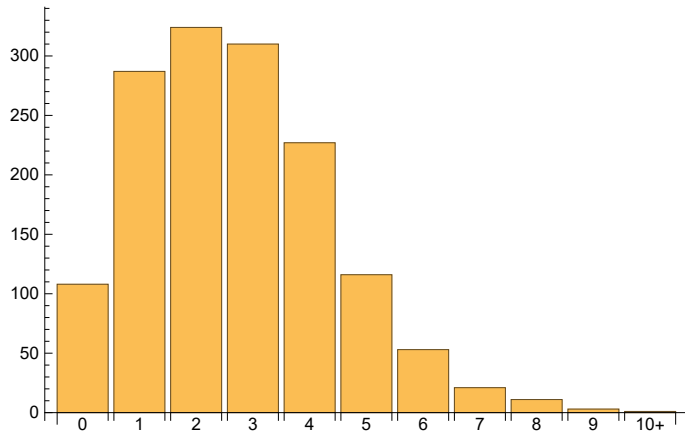
In[338]:= ExpectedData = Join[
  Table[0, {i, 1, 98.30156997973859`}],
  Table[1, {i, 1, 265.29985655722743`}], Table[2, {i, 1, 358.000456675273`}],
  Table[3, {i, 1, 322.0615561648646`}], Table[4, {i, 1, 217.29786378474694`}],
  Table[5, {i, 1, 117.29027746793392`}], Table[6, {i, 1, 52.757878628344`}],
  Table[7, {i, 1, 20.34069770524693`}], Table[8, {i, 1, 6.862026955149613`}],
  Table[9, {i, 1, 2.057720912932917`}],
  Table[10, {i, 1, 0.7300951685420194`}]];
Histogram[TestData]
Histogram[ExpectedData]

```



(*The BarCharts are created as*)

```
In[341]:= Bar0 = {108, 287, 324, 310, 227, 116, 53, 21, 11, 3, 1};  
BarE = ExpectedE;  
BarChart[Bar0, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10+"}]  
BarChart[BarE, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10+"}]
```



```

In[37]:= csvPath =
  "M:\\JI Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Codes\\2\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]]];
For[i = 1, i <= Length[AllDate], i++,
  AllDate[[i]] = DayName[DateObject[StringCases[AllDate[[i, 1]],
    x : DatePattern[{"Year", "Month", "Day"}] => DateList[x]][[1 ;; 1, 1 ;; 3]]]];
Count1 =
  Tally[
    AllDate]

```

```

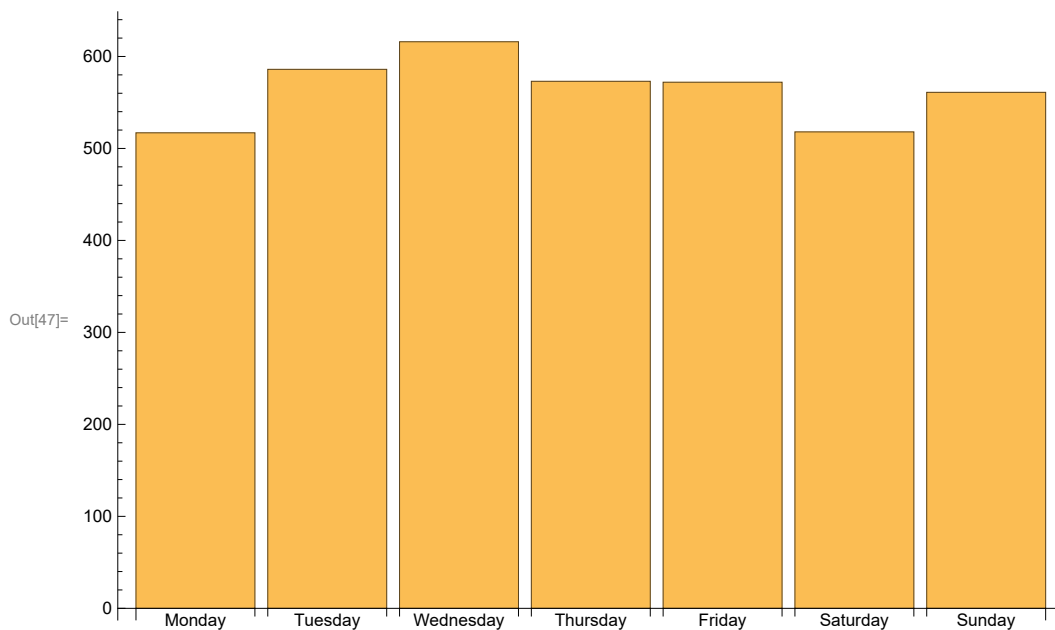
Out[43]= {{{Friday}, 572}, {{Saturday}, 518}, {{Sunday}, 561},
  {{Monday}, 517}, {{Tuesday}, 586}, {{Wednesday}, 616}, {{Thursday}, 573}}

```

```

In[44]:= (*Adjust the Orders and make the barchart*)
Count1 = {{Monday, 517}, {Tuesday, 586}, {Wednesday, 616},
  {Thursday, 573}, {Friday, 572}, {Saturday, 518}, {Sunday, 561}};
Count2 = Count1[[All, 1]];
Count3 = Count1[[All, 2]];
BarChart[Count3, ChartLabels -> Count2]

```



```

In[ ]:= (*The following applies Pearson's Goodness-of-
  fit to test dependence of shooting and days, the latter part is in excel file*)

```

```

In[66]:= AllShoot = Length[AllDate]
AvgShoot = N[AllShoot / (365 * 4 + 1)]
Count4 = Count3;
Count4[[1]] = N[Count3[[1]] / 209];
Count4[[2]] = N[Count3[[2]] / 208];
Count4[[3]] = N[Count3[[3]] / 208];
For[i = 4, i ≤ 7, i++, Count4[[i]] = N[Count3[[i]] / 209]];
Count4

Out[66]= 3943

Out[67]= 2.69884

Out[73]= {2.47368, 2.81731, 2.96154, 2.74163, 2.73684, 2.47847, 2.68421}

```

```

In[1316]:= (*Finding contingency tables*)
csvPath =
  "M:\\JI Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Figures\\2\\data.csv";
SystemOpen@csvPath
Data = Import[csvPath];
PosDate = Position[Data, "2019-01-01"];
AllDate = Data[[2 ;; PosDate[[1, 1]] - 1, PosDate[[1, 2]] ;; PosDate[[1, 2]]]];
For[i = 1, i <= Length[AllDate], i++,
  AllDate[[i]] = DateObject[StringCases[AllDate[[i, 1]],
    x : DatePattern[{"Year", "Month", "Day"}] :> DateList[x]][[1 ;; 1, 1 ;; 3]]];
Count1 = Tally[AllDate];
For[i = 1, i <= Length[Count1], i++, Count1[[i, 1]] = DayName[Count1[[i, 1]]]];
Count2 = Tally[Count1]

```

```

Out[1324]= {{{{Friday}}, 2}, 50}, {{{Saturday}}, 1}, 48}, {{{Sunday}}, 3}, 53}, {{{Monday}}, 1}, 51},
{{{Tuesday}}, 4}, 43}, {{{Wednesday}}, 4}, 27}, {{{Thursday}}, 4}, 27},
{{{Tuesday}}, 2}, 40}, {{{Wednesday}}, 6}, 18}, {{{Thursday}}, 5}, 26},
{{{Saturday}}, 3}, 41}, {{{Sunday}}, 2}, 43}, {{{Tuesday}}, 1}, 41},
{{{Thursday}}, 2}, 48}, {{{Friday}}, 3}, 54}, {{{Monday}}, 4}, 30},
{{{Wednesday}}, 3}, 51}, {{{Friday}}, 1}, 36}, {{{Saturday}}, 2}, 50},
{{{Monday}}, 3}, 35}, {{{Thursday}}, 1}, 39}, {{{Sunday}}, 5}, 16},
{{{Monday}}, 2}, 48}, {{{Tuesday}}, 3}, 41}, {{{Wednesday}}, 2}, 45},
{{{Friday}}, 5}, 14}, {{{Tuesday}}, 5}, 19}, {{{Wednesday}}, 1}, 37},
{{{Friday}}, 4}, 26}, {{{Sunday}}, 1}, 35}, {{{Monday}}, 6}, 6}, {{{Saturday}}, 7}, 4},
{{{Thursday}}, 3}, 35}, {{{Wednesday}}, 7}, 2}, {{{Thursday}}, 6}, 6},
{{{Saturday}}, 4}, 34}, {{{Friday}}, 7}, 3}, {{{Wednesday}}, 5}, 18},
{{{Saturday}}, 5}, 10}, {{{Sunday}}, 4}, 40}, {{{Monday}}, 5}, 13}, {{{Tuesday}}, 8}, 4},
{{{Thursday}}, 7}, 6}, {{{Friday}}, 6}, 9}, {{{Tuesday}}, 6}, 6}, {{{Monday}}, 8}, 2},
{{{Monday}}, 7}, 4}, {{{Saturday}}, 6}, 4}, {{{Wednesday}}, 8}, 2}, {{{Sunday}}, 7}, 1},
{{{Tuesday}}, 7}, 1}, {{{Sunday}}, 6}, 4}, {{{Friday}}, 8}, 2}, {{{Saturday}}, 9}, 1},
{{{Thursday}}, 9}, 1}, {{{Sunday}}, 10}, 1}, {{{Thursday}}, 8}, 1}, {{{Friday}}, 9}, 1}}

```

(*Independence Test is performed in the Excel File*)


```

In[252]:= csvPath = "M:\\\\JI
           Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Codes\\6\\data2019.xlsx";
           SystemOpen@csvPath
           Data = Import[csvPath];
           Count1 = Data[[1]][[2 ;; -16]];
           Count2 = Tally[Count1[[All, 2]]]

Out[256]= {{2., 22}, {3., 18}, {6., 5}, {4., 11}, {5., 7}, {1., 16}, {7., 2}, {0., 8}, {9., 1}}

In[345]:= (*The fit done by Mathematica*)
           Needs["HypothesisTesting`"];
           TestData = Join[Table[0, {i, 1, 8}], Table[1, {i, 1, 16}], Table[2, {i, 1, 22}],
                           Table[3, {i, 1, 18}], Table[4, {i, 1, 11}], Table[5, {i, 1, 7}],
                           Table[6, {i, 1, 5}], Table[7, {i, 1, 2}], Table[9, {i, 1, 1}]];
           PearsonChiSquareTest[TestData, PoissonDistribution[k1],
                                {"FittedDistributionParameters", "DegreesOfFreedom", "TestDataTable"}]

Out[347]= {{k1 -> 2.73333}, 4, 

|                  | Statistic | P-Value  |
|------------------|-----------|----------|
| Pearson $\chi^2$ | 2.23499   | 0.692629 |

}

In[348]:= (*The fit done considering Pearson Criteria 3.1.2*)
           Count2 =
             {{0, 8}, {1, 16}, {2, 22}, {3, 18}, {4, 11}, {5, 7}, {6, 5}, {7, 2}, {8, 0}, {9, 1}};
           k2 = 0;
           n = 31 + 28 + 31;
           For[i = 1, i <= 10, i++, k2 = k2 + Count2[[i, 1]] * Count2[[i, 2]]];
           k2 = N[k2/n]

Out[352]= 2.73333

In[374]:= e = N[Exp[1]];
           CRV = {};
           For[i = 0, i <= 9, i++, CRV = Insert[CRV, e^(-k2) * k2^i / Factorial[i], -1]];
           CRV

Out[377]= {0.0650023, 0.177673, 0.24282, 0.221236, 0.151178,
           0.0826438, 0.0376488, 0.014701, 0.00502283, 0.00152545}

In[378]:= ExpectedE = {};
           For[i = 1, i <= 10, i++, ExpectedE = Insert[ExpectedE, n * CRV[[i]], -1]];
           ExpectedE

Out[380]= {5.8502, 15.9906, 21.8538, 19.9112, 13.606, 7.43794, 3.38839, 1.32309, 0.452055, 0.137291}

In[381]:= (*We see that the Pearson Criteria 3.1.2 is not satisfied,
           therefore we merge the last 4 categories and obtain the following*)
           Count3 = {{0, 8}, {1, 16}, {2, 22}, {3, 18}, {4, 11}, {5, 7}, {6, 8}};
           NewCRV = {0.06500225396303454`, 0.17767282749896107`, 0.2428195309152468`,
                     0.22123557261166932`, 0.15117764128464067`, 0.08264377723560358`,
                     0.037648831851774964` + 0.01470097243735975` + 0.007098592201709374`};
           NewExpectedE = {5.850202856673108`, 15.990554474906496`, 21.853757782372213`,
                           19.91120153505024`, 13.60598771561766`, 7.4379399512043225`,
                           3.3883948666597465` + 1.3230875193623775` + 0.6388732981538436`};
           NewObservedE = {8, 16, 22, 18, 11, 7, 8};
           (*The Degree of Freedom is found by*)
           DegreesOfFreedom = Length[NewExpectedE] - 1 - 1

Out[385]= 5

```

```

In[386]:= (*The Pearson Statistic is found by*)
Stat = 0;
For[i = 1, i ≤ 7, i++,
  Stat = Stat + (NewObservedE[[i]] - NewExpectedE[[i]])^2 / NewExpectedE[[i]];
Stat

```

Out[388]= 2.81152

```

In[293]:= (*Fix alpha=0.05, we test the null hypothesis, and fail to reject it*)
Solve[CDF[ChiSquareDistribution[DegreesOfFreedom], x] == 0.95, x]

```

Solve: Solve was unable to solve the system with inexact coefficients. The answer was obtained by solving a corresponding exact system and numericizing the result.

Out[293]= { {x → 11.0705} }

```

In[294]:= (*The P value is found by:*)
PValue = 1 - CDF[ChiSquareDistribution[DegreesOfFreedom], Stat]

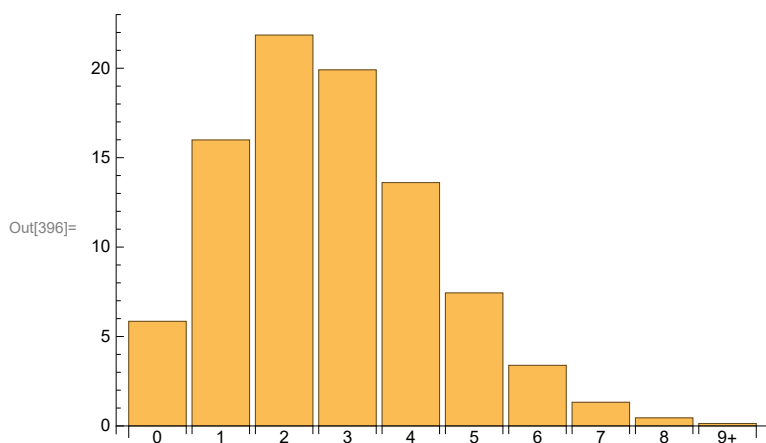
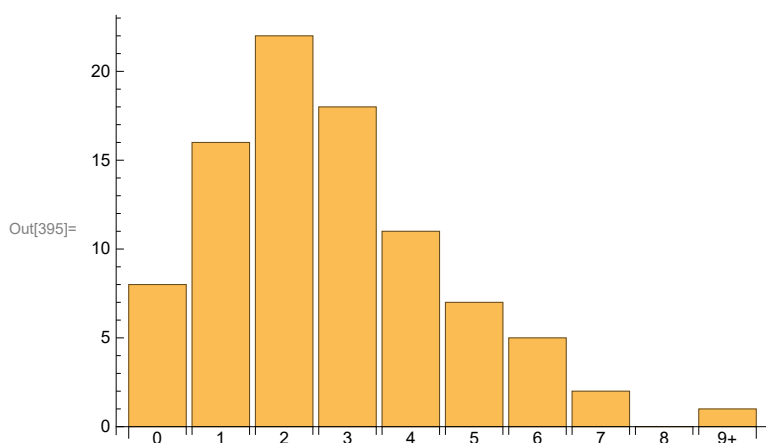
```

Out[294]= 0.729016

```


In[393]:= Bar0 = {8, 16, 22, 18, 11, 7, 5, 2, 0, 1};
BarE = ExpectedE;
BarChart[Bar0, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9+"}]
BarChart[BarE, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9+"}]

```



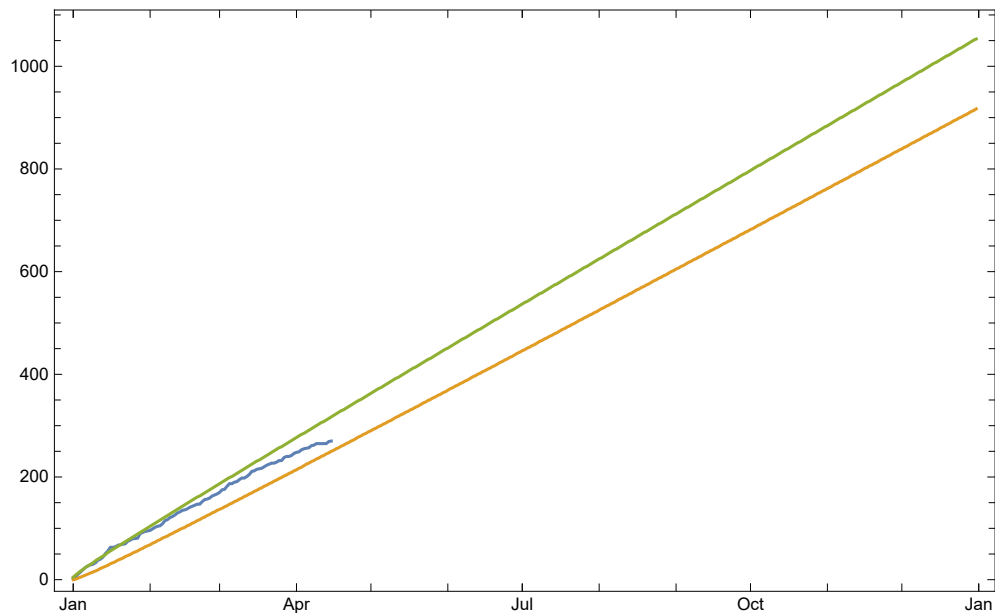
```

In[579]:= (*Observed*)
csvPath = "M:\\\\JI
Courses\\Sophomore_Year\\2019_Spring\\VE401\\Proj2\\Codes\\6\\data2019.xlsx";
SystemOpen@csvPath
Data = Import[csvPath];
AllDate = Data[[1]][[2 ;; -1]];
For[i = 1, i <= Length[AllDate], i++,
  AllDate[[i, 1]] = DateObject[AllDate[[i, 1]][[1 ;; 1, 1 ;; 3]]][[1]];
Count2 = Count1;
For[i = 2, i <= Length[AllDate], i++,
  Count2[[i, 2]] = Count2[[i - 1, 2]] + Count2[[i, 2]]
];

(*Predicted full year*)
Count1 = {};
Countu = {};
DayTemp =  Tue 1 Jan 2019 00:00:00 GMT+8. ;
For[m = 1, m <= 365, m++,
  l = Ceiling[3943 / 1461 * m - 1.96 * Sqrt[a * m^2 * (1 / m + 1 / 1461)]];
  u = Floor[3943 / 1461 * m + 1.96 * Sqrt[a * m^2 * (1 / m + 1 / 1461)]];
  LowTemp = {DayPlus[DayTemp, m - 1], l};
  UpTemp = {DayPlus[DayTemp, m - 1], u};
  Count1 = Insert[Count1, LowTemp, -1];
  Countu = Insert[Countu, UpTemp, -1];
];
(*Plot*)
DateListPlot[{Count2, Count1, Countu}]


```

Out[590]=



```

In[591]:= (*Predicted till Apr. 15th*)
Count1 = {};
Countu = {};

DayTemp =  Tue 1 Jan 2019 00:00:00 GMT+8. ;

For[m = 1, m ≤ 31 + 28 + 31 + 15, m++,
  l = Ceiling[3943 / 1461 * m - 1.96 * Sqrt[a * m^2 * (1 / m + 1 / 1461)]];
  u = Floor[3943 / 1461 * m + 1.96 * Sqrt[a * m^2 * (1 / m + 1 / 1461)]];
  LowTemp = {DayPlus[DayTemp, m - 1], l};
  UpTemp = {DayPlus[DayTemp, m - 1], u};
  Count1 = Insert[Count1, LowTemp, -1];
  Countu = Insert[Countu, UpTemp, -1];
];
(*Plot*)
DateListPlot[{Count2, Count1, Countu}]

```

