Police Shooting in the United States

28 April 2020

GROUP 13

Yigao Fang	518370910032
Haorong Lu	518370910194
Hongdan Wang	518370910209

Abstract

Police shootings are particularly serious in the United States. Over four thousand cases of police shooting are recorded in the US during the past five years. Whether there exist some regular patterns inside the huge data samples are what we concern. With the help of Pearson statistic and Cochran's Rule, we can test that the occurrence of police shootings follows a Poisson distribution over the past five years, with a parameter k we could estimate using the sample mean. We also calculate the 95% confidence interval for the estimated parameter. We then use a similar method to test whether the police shooting samples in 2020 follows a Poisson distribution. Our hypothesis is rejected at a high significance level and hence we conclude that the data does not follow a Poisson distribution in the year 2020. Factors leading to this phenomenon are also discussed. Based on the database of the past years, we can predict the number of police shootings in 2020 and find out its predicting interval. Comparing with the actual numbers, we discover the influence of the outbreak of the Coronavirus on the number of police shootings in 2020.

Keywords: Poisson distribution, Pearson statistic, confidence interval, predicting interval.

Contents

1 Introduction

1.1 Background

In 2014, Michael Brown, an unarmed black man, was killed by the police officer in Ferguson. This began the protest of Black Lives Matter and rose the public focus on police accountability in the U.S. After then, it was found that the official data of police fatal shootings was incomplete. From Jan. 1, 2015, The Washington Post [2]started to record every fatal police shootings in the U.S. The data was obtained from local news reports, law enforcement websites and social media, and by monitoring independent databases like Killed by Police and Fatal Encounters. Besides, The Post contacts the department to get more details.

The term "fatal police shooting" means that the data covers only those who killed immediately by shootings from a police officer in the line of duty. The deaths of people in police custody and fatal shootings by off-duty police officers or non-shooting deaths such as complications will not be recorded.

In this project, we are going to find out the pattern of these data and then make a prediction to the future.

1.2 Methods

The methods we used are based on the article *London murders: a predictable pattern?* by David Spiegelhalter and Arthur Barnett [4], which analyzes the murders in London. The article estimates that the murders in London from 2004 to 2007 followed a Poisson distribution with parameter k = 0.44, and give a 95% prediction interval for murders in London in 2008 based on the previous data. In this paper, we will test whether the occurrence of fatal police shootings follows a Poisson distribution, and the dependence between the average number of shootings and the weekday. Also, we will also give a prediction interval for the number of fatal police shootings in 2020 based on the data for 2015 to 2019. Then we will compare it to the actual data to show the influence of COVID-19 on the number of fatal police shootings.

2 Results

2.1 Possion Distribution for Data between 2015 and 2019

To have a better understanding of the term "fatal police shooting", we have studied the pattern of police shootings in the US between January 1^{st}, 2015, and December 31^{st}, 2019. Among these five years, a total of 4938 cases of police shootings are recorded in the database of the Washington Post[2]. Figure 1 illustrates the number of police shootings in the US each day over the five years. We notice that there were five days on which nine police shootings occurred, and there were some days on which not a single case was recorded.

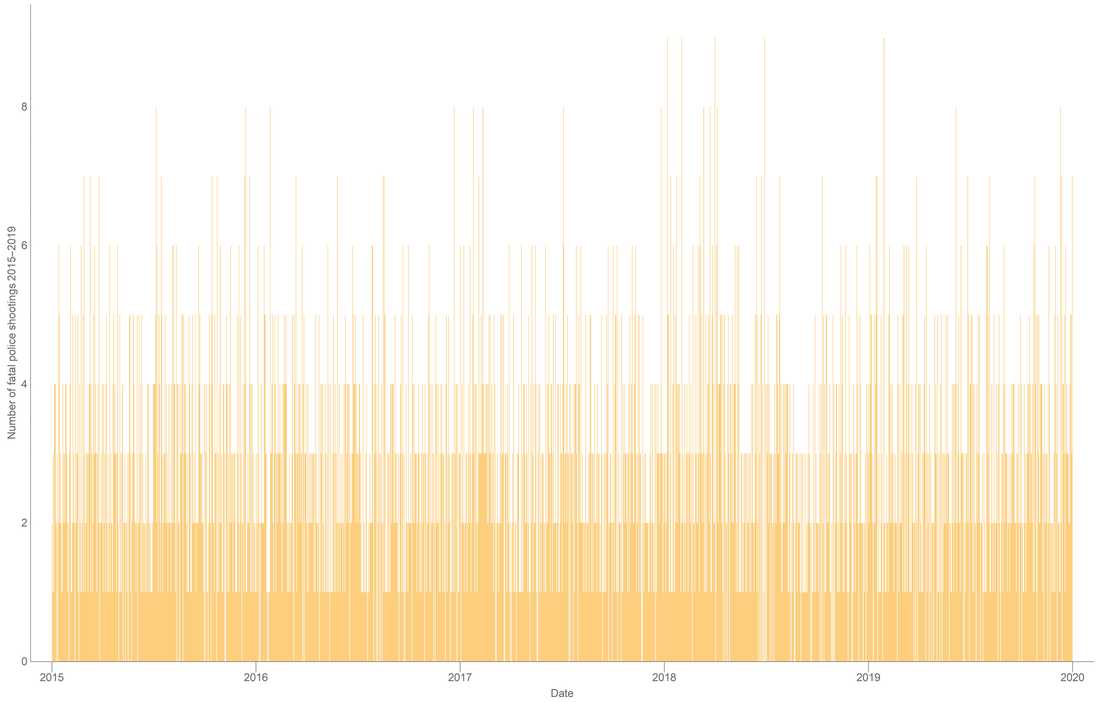

Figure 1: Number of Fatal Police Shootings in US between 2015 and March 2019

This figure seems quite mass, but if we consider the police shootings as random events and look into the distribution of the 4938 cases in four years(1826 days), we discover that the number of shootings each day would follow a Poisson distribution with parameter k. To estimate the value of k, we first list Table 1 which illustrates the observed frequencies of the different number of police shootings per day

Number of police shootings	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9 or more**
Observed Frequency	139	348	414	382	280	151	66	28	13	5

Table 1: The observed frequency of numbers of police shootings per day

According to the concepts of Poisson distribution, a maximum likelihood estimator for the Poisson parameter k is the sample mean $\overline{X}$, which is

$$\begin{aligned}\hat{k} &= \overline{X}\\ &= \frac{1}{1826}(348 + 414 \times 2 + 382 \times 3 + 280 \times 4 + 151 \times 5 + 66 \times 6 + 28 \times 7 + 13 \times 8 + 5 \times 9)\\ &= \frac{2469}{913} \approx 2.704\end{aligned}$$

We have the density function of the Poisson distribution

$$f_X(x) = \frac{k^x e^{-k}}{x!} \tag{1}$$

Based on Equation (1) and the estimate value $\hat{k}$, we insert the value of x and can then calculate the expected number of days with the following numbers of police shootings per day, as shown below in Table 2

x	0	1	2	3	4	5	6	7	8	9 or more
Expected(E_i)	122.19	330.44	446.80	402.76	272.29	147.27	66.38	25.64	8.67	3.56
Observed(O_i)	139	348	414	382	280	151	66	28	13	5

Table 2: Numbers of days with the following numbers of police shootings per day

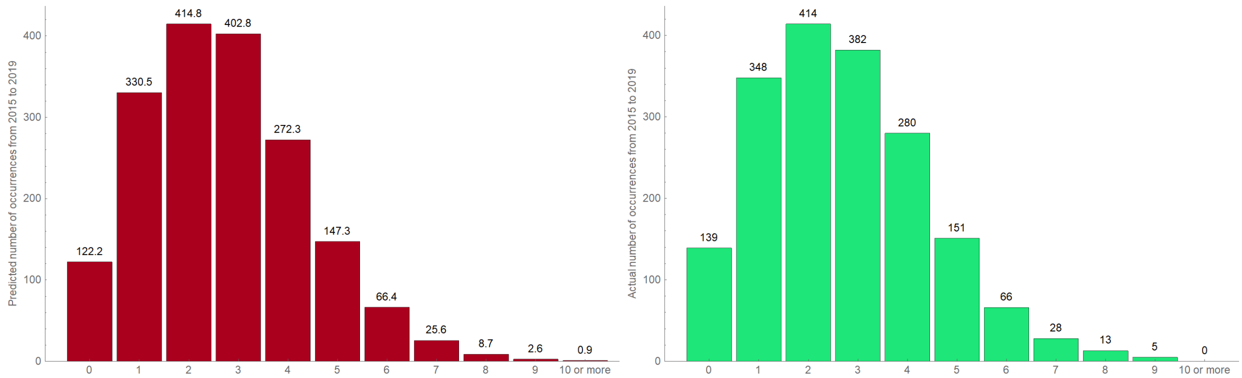

Figure 2: Number of Police Shooting Per Day

As illustrated in Figure 2, the left side shows the predicted number of police shootings over the four years, while the right side shows the actual number of shootings. We find that these two graphs are quite similar, indicating our theoretical model fits the actual case well. Now we will test the hypothesis H_0, which is

H_0 : The occurrence of police shootings follows a Poisson distribution with parameter $k = 2.704$

We choose $N = 10$. The expected numbers obey Cochran's Rule. Hence the Pearson statistic

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} = 9.99 \tag{2}$$

follows a chi-squared distribution with $N-1-m = 8$ degrees of freedom. We want to realize $\alpha = 0.05$. The calculation gives

$$X^2_{0.05,8} = 15.51 > 9.99 \tag{3}$$

Hence we are unable to reject H_0 at the 5% level of significance.
The P-value for the test is given by

$$P - value = 2(1 - P[X_8^2 < 9.99]) = 0.531 \tag{4}$$

2.2 Relationship between Number of Police Shootings and Weekday

	Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**	**Saturday**	**Sunday**
Observed(O_i)	668	742	757	732	692	662	685
Expected(E_i)	705.43						

Table 3: Oberved and Expected Numbers of Shootings on Each Weekday

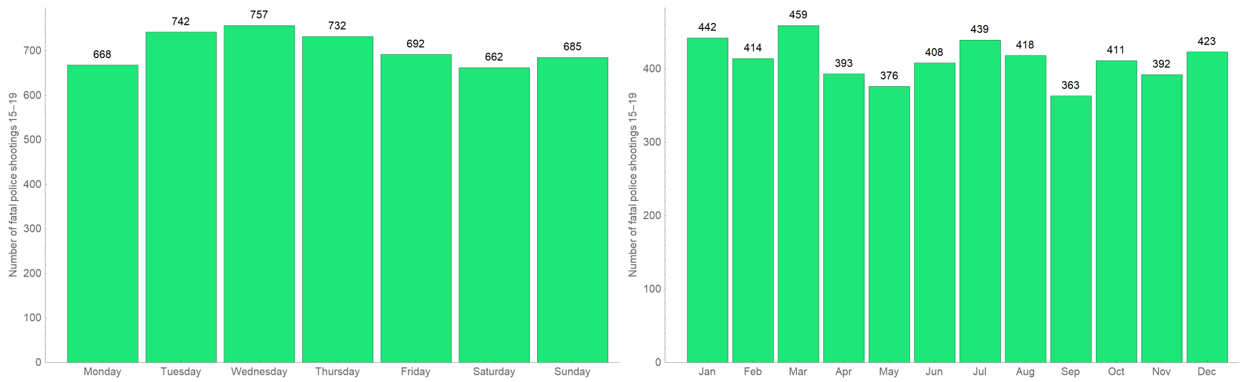

Figure 3: Number of Police Shooting Per Day

We will test whether there is evidence to show that the average number of police shootings depends on the weekday. Hence we test the hypothesis H_0, which is

$$H_0 : \text{Each weekday is equally likely for police shootings}$$

We choose $N = 7$. The expected numbers obey Cochran's Rule. Hence the Pearson statistic

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} = 12.17 \tag{5}$$

follows a chi-squared distribution with $N - 1 = 6$ degrees of freedom. We want to realize $\alpha = 0.05$. The calculation gives

$$X^2_{0.05,6} = 12.6 > 12.17 \tag{6}$$

Hence we are unable to reject H_0 at the 95% level of significance.
The P-value for the test is given by

$$P - value = 2(1 - P[X^2_6 < 12.17]) = 0.117 \tag{7}$$

Hence we conclude that the average number of police shootings does not depend on the weekdays.

2.3 Confidence Intervals for the Parameter k

We are now interested in the confidence interval for the parameter k of this Poisson distribution. The $(1-\alpha)100\%$ confidence interval for the parameter k is defined as

$$P[\overline{X} - L \leq \hat{k} \leq \overline{X} + L] = 1 - \alpha \tag{8}$$

Since in this case, the sample size is huge enough for us to assume that $\overline{X}$ follows a normal distribution, we have

$$\frac{\alpha}{2} = P[z > z_{\alpha/2}] = \frac{1}{\sqrt{2\pi}} \int_{z_{\alpha/2}}^{\infty} e^{-\frac{x^2}{2}} dx \tag{9}$$

Since k is the parameter of Poisson distribution, $\overline{X}$ is normally distributed with mean k and variance k/n. Hence $Z = \dfrac{\overline{X} - k}{\sqrt{k/n}}$ follows a standard normal distribution. Hence we have

$$\frac{\alpha}{2} = P[\frac{L}{\sqrt{k/n}} \leq z] \tag{10}$$

$$L = z_{\alpha/2}\sqrt{\hat{k}/n} \tag{11}$$

and the $(1-\alpha)100\%$ confidence interval for k is given by

$$\hat{k} \pm z_{\alpha/2}\sqrt{\hat{k}/n} \tag{12}$$

According to former calculations, we have $\hat{k} = 2.704, n = 1826, z_{0.025} = 1.96$. Insert these values into Equation (12), we have the 95% confidence interval

$$\hat{k} \pm z_{\alpha/2}\sqrt{\hat{k}/n} = 2.704 \pm 0.075 = [2.629, 2.779] \tag{13}$$

2.4 Possion Distribution for Data in 2020

We have seen that the sample data between 2015 and 2019 follows a Poisson distribution, but whether data of 2020 also follows a Poisson distribution is also what we care about. We assume that the data between January to April 15^{th}, 2020 follows a Poisson distribution. The maximum likelihood estimator for k is the sample mean, which is

$$\hat{k} = \overline{X} = \frac{265}{106} \approx 2.49 \tag{14}$$

Based on Equation (1) and the estimate value $\hat{k}$, we can calculate the expected number of days with the following numbers of police shootings per day, as shown below in Table 4

	0	**1**	**2**	**3**	**4**	**5**	**6 or more**
Expected(E_i)	8.70	21.75	27.19	22.66	14.16	7.08	4.45
Observed(O_i)	17	20	20	18	14	9	8

Table 4: Numbers of days with the following numbers of police shootings per day

Now we will test the hypothesis H_0

H_0 : The occurrence of police shootings follows a Poisson distribution with parameter $k = 2.49$

We choose $N = 7$. The expected numbers obey the Cochran's Rule. Hence the Pearson statistic

$$X^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} = 14.27 \tag{15}$$

follows a chi-squared distribution with $N-1-m = 5$ degrees of freedom. We want to realize $\alpha = 0.05$. The calculation gives

$$X^2_{0.05,5} = 11.07 < 14.27 \tag{16}$$

Hence we are able to reject H_0 at the 95% level of significance.

The result is astonishing. We discover an entirely different result from what we have obtained before. The test result indicates that the number of police shootings in 2020 does not follow a Poisson distribution. The different result may results from the small sample size. Comparing with 4938 cases in 1826 days, we only have 265 cases in 106 days now. The sample is quite small and may lead to the inaccuracy of the result.

Another important factor is the influence of COVID-19. The impact of the environment will strongly affect the sample data. This influence will be discussed later in the Discussion section.

Consequently, we wonder whether this phenomenon is normal or not. Can the number of shootings in 2020 be predicted using the database of 2015 to 2019? Hence we will calculate the predicting interval in the next part.

2.5 Predicting Intervals for Data in 2020

Based on the calculation above, we have already examined the pattern of police shootings in the US. One of the most important purposes of analyzing the data is to utilize the current sample to predict the future, and we are now able to predict the number of police shootings that will happen in 2020 using the current database.

Nelson's formula[3] is a useful tool to find the prediction interval for the numbers. Assume X be the total occurrence in an n size sample we already have, which follows a Poisson distribution with parameter k. Y denotes the future occurrence in an m size sample, which we are going to predict.

According to the definition of Nelson's formula,

$$\hat{Y} = \frac{mX}{n}, \quad v\hat{a}r(\hat{Y} - Y) = m\hat{Y}(\frac{1}{n} + \frac{1}{m})$$

Nelson's approximate predicting interval is based on the asymptotic result that

$$\frac{\hat{Y} - Y}{\sqrt{v\hat{a}r(\hat{Y} - Y)}} \sim N(0, 1) \tag{17}$$

Let z be the value of CDF for a normal distribution. A (1-2α)100% predicting interval for Y satisfies,

$$\left|\frac{\hat{Y} - Y}{\sqrt{v\hat{a}r(\hat{Y} - Y)}}\right| \leq z_{1-\alpha}$$

So a (1-2α)100% predicting interval for Y is given by $Y \pm z_{1-\alpha}\sqrt{m\hat{Y}(\frac{1}{n}+\frac{1}{m})}$.
Hence we have derived the Nelson's formula

$$[L,U] = \hat{Y} \pm z_{\alpha/2}\sqrt{m\hat{Y}(\frac{1}{m}+\frac{1}{n})} \tag{18}$$

As an example, we apply the data from the years 2015 to 2019 and predict the number of police shootings in 2020. Based on the data over the past five years, we have $n = 1826$, $X = 4938$, hence $\hat{Y}$ is given by

$$\hat{Y} = \frac{mX}{n} = 2.704m \tag{19}$$

Insert $\hat{Y}$ into Equation (18), taking $\alpha = 0.05$, we have

$$L = 2.704m - 1.96\sqrt{2.704m^2(\frac{1}{1826}+\frac{1}{m})} \tag{20}$$

$$U = 2.704m + 1.96\sqrt{2.704m^2(\frac{1}{1826}+\frac{1}{m})} \tag{21}$$

Take the ceil of L and the floor of U, we can obtain the 95% prediction intervals $[\lceil L\rceil, \lfloor U\rfloor]$ for the number of police shootings in 2020, as illustrated in the left graph of Figure 4 below:

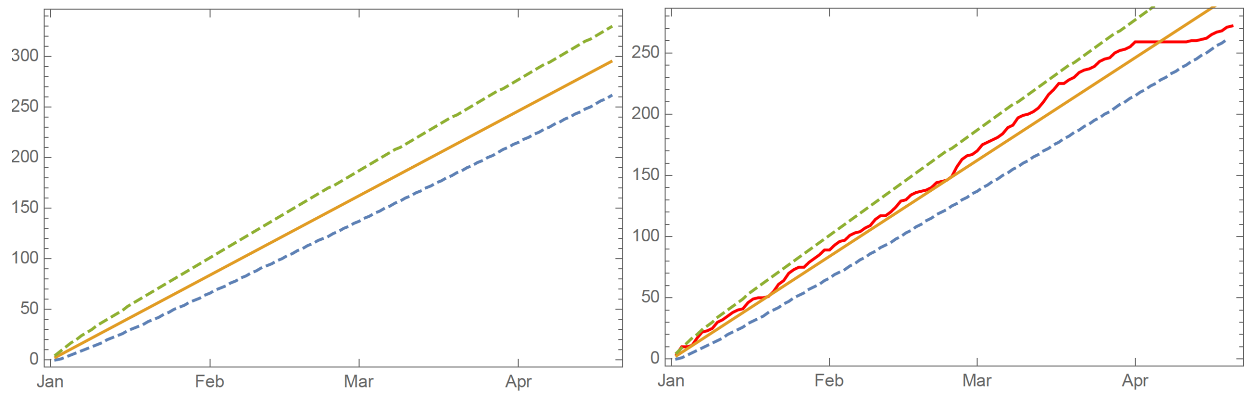

Figure 4: Predicted Number of Police Shootings during 2020 (with 95% prediction limits)

We have seen that the sample of shootings in 2020 is influenced by some outer factors, especially the outbreak of COVID-19. Will this circumstance influence the accuracy of our prediction? We will discover it now.

As illustrated in the right graph of Figure 4, the right curve denotes the actual accumulation number of police shootings in 2020, while the two dashed curves denote the 95% prediction interval we calculate. According to the figure, we find that the data for 2020 is generally within the 95 % confidence interval before April, but after April 1^{st} there is a sharp decrease in the growth rate of the number of shootings (even 10 days of zero increase). This phenomenon may result from the outbreak of COVID-19. However, the actual curve is still within the 95% predicting interval, as we can see from the picture. Hence if we set up hypothesis H_0 that COVID-19 has no impact on the shooting cases, we do not have enough evidence to reject H_0 at 95% of confidence.

Nevertheless, from the graph, we can still observe a huge influence of the break of Coronavirus on the number of the police shooting in the US, which will be discussed in detail in the Discussion section.

3 Discussion and Conclusion

3.1 Discussion

In section 2.1 and 2.4 we discussed whether the occurence of police shooting follows a Poisson distribution in 2015-2019 and 2020. We set up hypothesis H_0 that the distribution follows a Poisson distribution, and we fail to reject H_0 for 2015-2019 at p-value 0.531, while we successfully reject H_0 for 2020. This indicates that the number of police shootings in 2020 is less likely to follow a Poisson distribution than the past few years.

We discussed the two reasons that may cause this change. Firstly, the number of police shootings may be affected by the COVID-19 pandemic occurs in 2020. From Figure 4 we can see a distinct drop in the increase of police shootings, which begins on April 1st. But according to CDC data[4], the first case appears as early as in January, and the outbreak begins in the middle of March. A distinct drop in the number of shooting cases appears at the beginning of April, which is more concerned with the ”stay at home” order. Washington government announced the order on March 31st[5], and on April 1st the drop appears. So the COVID-19 may not directly affect the number of shooting cases, but the stay home order does.

Secondly, the sample size we used for prediction is different. The sample size for 2020 data is 110, but the sample size for the 2015-2019 data is 1826. Abnormal samples, such as samples since April in 2020, will have a larger influence on small sample sizes than large sample sizes. A small sample size makes the ratio of observed and expected cases to be larger, so the Pearson static becomes greater, and then the hypothesis is rejected.

Although the number of observed cases does not fall out of Nelson’s PI, the difference of rejection in 2.1 and 2.4 shows that COVID-19 have some influence on the police shootings. Besides, a smaller sample size may also cause the data of 2020 to not follow a Poisson distribution.

3.2 Conclusion

During this project, we analyze the pattern of police shootings in the US over the past five years. We are 95% confident to claim that the number of shootings per day follows a Poisson distribution with parameter $k = 2.704 \pm 0.075$. We also reject the hypothesis that the number of shootings depends on weekdays, and are 95% confident to claim that the number of police shootings is equally distributed in the seven weekdays. Based on the current data from 2015 to 2019, we are able to estimate the predicting interval for the number of police shootings in 2020, and compare with the actual results. We find that all the data are within the 95% predicting interval. Therefore our prediction is quite accurate. The influence of the outbreak of Coronavirus indeed exists, but its impact does not exceed the 95% predicting interval. Hence we claim with 95% confidence that the influence of COVID-19 is ignorable.

References

[1] D. Spiegelhalter and A. Barnett. London murders: a predictable pattern? Significance, 6(1):5–8, 2009. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2009.00334.x/abstract.

[2] The Washington Post. (2020). Database of Fatal Police Shootings. Retrieved from https://github.com/washingtonpost/data-police-shootings

[3] 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.

[4] Centers for Disease Control and Prevention. "Cases of Coronavirus Disease (COVID-19) in the U.S.". https : // www.cdc.gov/coronavirus/2019 - ncov/cases - updates/cases - in - us.html. Accessed April 28 th, 2020.

[5] Washington State Coronavirus Response. "Inslee issues additional guidance on 'Stay Home, Stay Healthy' order and proclamation for retired workers to return to essential jobs". https : // www.governor.wa.gov/news - media/inslee - issues - additional - guidance - stay - home - stay - healthy - order - and - proclamation - retired. Accessed April 28 th, 2020.

A Mathematica Code

```
data = Import["H:\\VE401_project_2\\fatal-police-shootings-data.csv",
"Data"][[2 ;; 4939, 3]]; (*Change to your own file path*)
Show[DateHistogram[data, "Day", Frame -> {True, True, False, False},
FrameLabel -> {"Date",
"Number of fatal police shootings 2015-2019"},
DateTicksFormat -> "Year"], ImageSize -> 1000]

```

Mathematica Code for Figure 1

```
datelist = DateRange[{2015, 1, 1}, {2019, 12, 31}, {1, "Day"}];
result = Range[Length[datelist]];
For[i = 1, i < Length[datelist] + 1, i++,
result[[i]] = Count[data, DateString[datelist[[i]], "ISODate"]]];
For[i = 0, i < 10, i++, Print[Count[result, i]]]
Histogram[result, LabelingFunction -> Above, ImageSize -> Large,
Frame -> {True, True, False, False},
FrameLabel -> {"", "Actual number of occurrences from 2015 to 2019"}]

BarChart[{139, 348, 414, 382, 280, 151, 66, 28, 13, 5, 0},
LabelingFunction -> Above, ImageSize -> Large,
Frame -> {True, True, False, False},
FrameLabel -> {"", "Actual number of occurrences from 2015 to 2019"},
ChartLabels -> {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9",
"10 or more"},
ChartStyle -> {Interpreter["Color"]["RGB 30 230 120"]}]

BarChart[{122.2, 330.5, 414.8, 402.8, 272.3, 147.3, 66.4, 25.6, 8.7,
2.6, 0.9}, LabelingFunction -> Above, ImageSize -> Large,
Frame -> {True, True, False, False},
FrameLabel -> {"",
"Predicted number of occurrences from 2015 to 2019"},
ChartLabels -> {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9",
"10 or more"},
ChartStyle -> {Interpreter["Color"]["RGB 170 0 30"]}]

```

Mathematica Code for Figure 2

```

BarChart[{668, 742, 757, 732, 692, 662, 685},
LabelingFunction -> Above, ImageSize -> Large,
Frame -> {True, True, False, False},
FrameLabel -> {"", "Number of fatal police shootings 15-19"},
ChartLabels -> {"Monday", "Tuesday", "Wednesday", "Thursday",
"Friday", "Saturday", "Sunday"},
ChartStyle -> {Interpreter["Color"]["RGB 30 230 120"]}]

BarChart[{442, 414, 459, 393, 376, 408, 439, 418, 363, 411, 392, 423},
LabelingFunction -> Above, ImageSize -> Large,
Frame -> {True, True, False, False},
FrameLabel -> {"", "Number of fatal police shootings 15-19"},
ChartLabels -> {"Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul",
```

```
"Aug", "Sep", "Oct", "Nov", "Dec"},
ChartStyle -> {Interpreter["Color"]["RGB 30 230 120"]}]

```

Mathematica Code for Figure 3

```
z = InverseCDF[NormalDistribution[0, 1], 0.975]
halfPI = z*Sqrt[m*Yba*(1/1826 + 1/m)]
L = Ceiling[Yba - halfPI]
U = Floor[Yba + halfPI]

data0 = Import[
 "C:\\Users\\lenovo\\Desktop\\fatal-police-shootings-data.csv"];
data = Table[data0[[i, 3]], {i, 4940, 5211}];
cases = Table[
Count[data,
DateString[
DatePlus["2020-01-01", i], {"Year", "-", "Month", "-",
"Day"}]], {i, 0, 109}];

data2020 = Table[{DatePlus["2020-01-01", i], \!\(
\*UnderoverscriptBox[\(\[Sum]\), \(n =
1\), \(i\)]\(cases[\([\)\(n\)\(]\)]\)\)}, {i, 110}];
plot1 = DateListPlot[data2020, Joined -> True, PlotStyle -> Red]
ltable = Table[{DatePlus["2020-01-01", m], L}, {m, 1, 109}];
utable = Table[{DatePlus["2020-01-01", m], U}, {m, 1, 109}];
ytable = Table[{DatePlus["2020-01-01", m], Yba}, {m, 1, 109}];
plot2 = DateListPlot[{ltable, ytable, utable},
PlotStyle -> {Dashed, Automatic, Dashed}]

Show[plot1, plot2]

```

Mathematica Code for Figure 4