UM-SJTU Joint Institute

Probabilistic Methods in Engineering

(ECE4010J)

Probabilities, Statistics, Prejudice: A Study of Fatal Police Shootings

Spring2025

Group:	Term Project Group 7		
Name:	Yuxuan Cao	ID:	523370910207
	Yukai Wang		523370910239
	Run Qian		523370910258
	Tianyi Wang		522370910235
	Yicheng Liao		523370910056

Date: April 21, 2025

Abstract

Fatal police shootings in the United States have long been a critical issue, raising concerns about social security and trust. This project analyzes the patterns of fatal police shootings in the US from 2015 to 2025. We use software like *Mathematica* to visualize the data according to different classifications, apply the methods of hypothesis testing to test the Poisson distribution and weekdays' or months' independence, calculate the confidence intervals for parameters, and predict intervals for the number of observations. Furthermore, we extend our analysis to the difference in black victimization rates between democratic and republican states, and other external factors' influence represented by body camera's carrying.

Our analysis reveals that the daily frequency of fatal police shootings follows a Poisson distribution with a maximum-likelihood estimated parameter of 2.85464. A goodness-of-fit test supports this assumption, but the COVID-19 pandemic influences the distribution. The multinomial statistics indicate that variations in shooting frequencies on weekdays and months are not a random chance, but they are dependent. Further analysis displays the statistical significance of a higher proportion of black victims in democratic states than in republican states, and the proportion of black victims in incidents involving body cameras is statistically significantly higher than in unrecorded incidents.

This study provides valuable insights into the patterns and influencing factors of fatal police shootings, and also serves as a good practice for various statistical analysis methods.

Keywords: Fatal Police Shootings in the US, Hypothesis Testing, Confidence Interval.

Contents

1 Introduction

In recent years, fatal police shootings in the United States have attracted increasing public attention and academic research. According to the Washington Post's "Fatal force" database, [1], which systematically records all known fatal police shootings in the United States from 2015 to 2024, it is a huge shock that 94% of days will have shooting accidents in the United States. And the frequency seems random. But is it a random fact? Inspired by Spiegelhalter and Barnett's analysis of the London homicide case [2], **our goal is to investigate whether these shooting incidents exhibit a probabilistic pattern**, especially whether their occurrence can be modeled using a Poisson distribution.

In addition to testing the distribution hypothesis, we also studied the time trends, including weekday and monthly changes, and evaluated the impact of external shocks, especially during the COVID-19 pandemic. By constructing confidence intervals and prediction intervals, we quantified the uncertainties related to future events. Furthermore, we explore deeper social factors impacting the incidents, particularly in the racial composition of victims in states with different political tendencies and the relationship with the use of body cameras. These surveys aim to provide a statistical basis for understanding the frequency and social background of fatal police shootings, and give us insights that may inform public cognition.

2 The Definition and Data Overview

Firstly, it is essential to give an exact definition of the fatal police shooting. We regulate the term used in this project, which reflects the reporting methods employed by the popular press, instead of official terms. It refers to **cases in which a person is shot and killed by a law enforcement officer in the course of duty**. The definition does not depend on whether the shooting was lawful and only focuses on the fact that the fatal shooting occurred.

The data in this project is accessed from [1], which documents all known fatal police shootings in the United States since 2015. This data set covers all reported fatal shootings in the United States from 2015 to 2024. The data comes from news reports, public records, social media, and law enforcement databases, which are regularly updated for accuracy and completeness.

The data set is detailed, which records information from multiple dimensions about each fatal police shooting. Each record included **the time and place** of the incident, **the type of threat** the police believed existed (such as point, move, and so on), **the race**, **age**, **gender**, and **whether the person was armed during the incident**. The data set also included additional perspectives, such as whether there was a body camera and whether the incident was mental-illness-related. These detailed dimensions provide the basis for our statistical analysis.

3 Visualization of Incident Frequency Over Time

To better analyze data trends, we visualized the distribution of fatal police shootings over time. We processed dates in the data using the `DateHistogram` function in *Mathematica*. First, the CSV file containing the event information is imported through the `Import` function. Since the data in this column is automatically identified in `DateObject` format, it can be used directly for statistical analysis. Next, the histogram is plotted using the `DateHistogram` function (Figure 1). The histogram shows the daily frequency of fatal police shootings from 2015 to 2024.

To further understand the distribution of fatal police shootings, two additional analyses are conducted from Figure 1. First, the gap between two events is relatively small; the longest gap

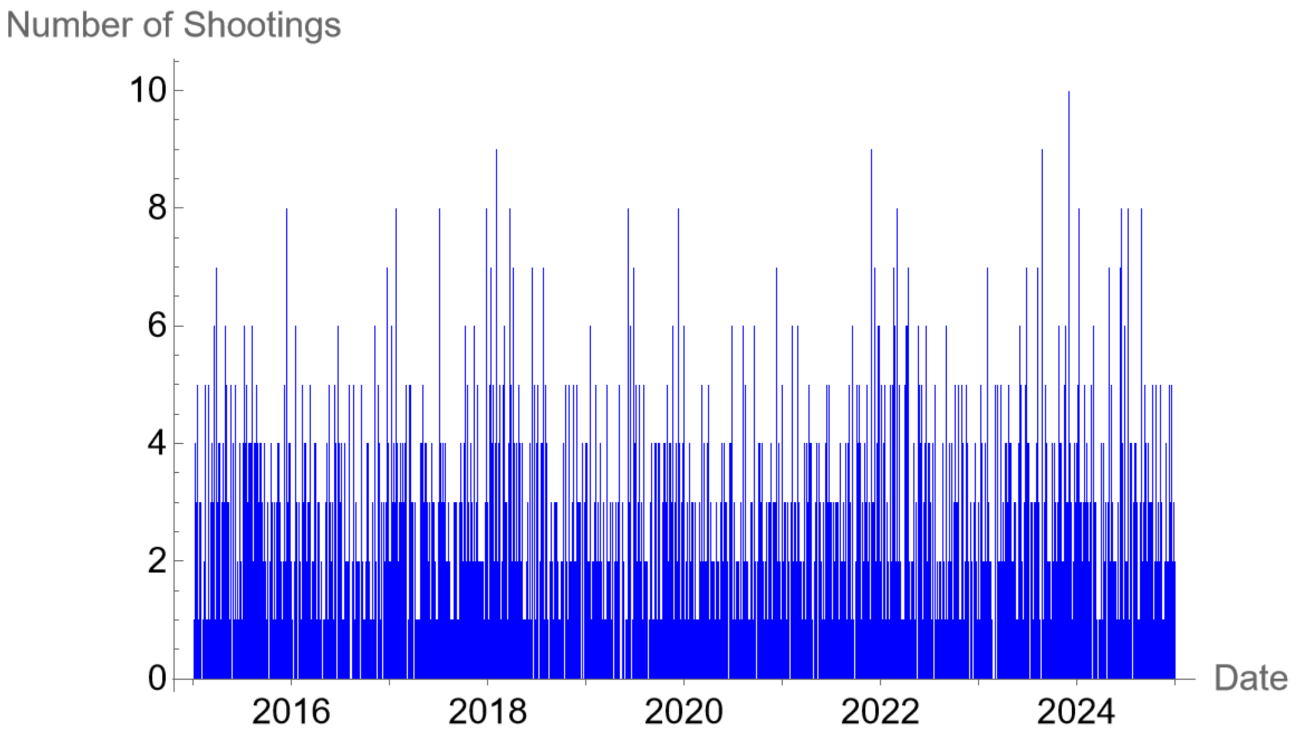

Figure 1: Daily frequency of fatal police shootings from 2015 to 2024.

between two consecutive events is found to be **3 days**. This indicates a relatively continuous pattern of fatal police shootings.

Also, we can find that on **Dec 3rd, 2023**, a maximum of **10 events** occurred. This unusually high day count may reflect the occurrence of some special events, and it may be necessary to investigate the background factors surrounding this date further.

4 Verification of Poisson Distribution (2015-2024)

4.1 Data Evaluation from 2015 to 2024

Based on the previous research from [2], which estimated that London homicides follow a Poisson distribution, we intend to verify whether the occurrence of police shootings in the US would also follow a **Poisson distribution**.

After collecting and processing the data from [1], we have obtained the days with different numbers of shooting incidents from 2015 to 2024. The table is shown below.

Number of incidents	Number of days	Number of incidents	Number of days
0	233	6	163
1	641	7	73
2	801	8	28
3	786	9	11
4	575	10	1
5	341		

Table 1: Number of days of different numbers of shooting.

It is claimed that the occurrence of shooting X follows a Poisson distribution. We want to know if there is evidence that justifies that it is false.

From Example 14.5 in the slides [3], the sample mean is the maximum-likelihood estimator for k, which is the parameter of the assumed Poisson distribution:

$$\hat{k} = \overline{X} = \frac{1}{3653}(0 \cdot 233 + 1 \cdot 641 + 2 \cdot 801 + 3 \cdot 786 + \cdots + 9 \cdot 11 + 10 \cdot 1) = 2.85464.$$

To apply the **multinomial distribution**, we first calculate the expected probability:

	Probability		**Probability**
$P[X=0] = \frac{e^{-\hat{k}}\hat{k}^0}{0!}$	0.05758	$P[X=6] = \frac{e^{-\hat{k}}\hat{k}^6}{6!}$	0.04327
$P[X=1] = \frac{e^{-\hat{k}}\hat{k}^1}{1!}$	0.16436	$P[X=7] = \frac{e^{-\hat{k}}\hat{k}^7}{7!}$	0.01765
$P[X=2] = \frac{e^{-\hat{k}}\hat{k}^2}{2!}$	0.23459	$P[X=8] = \frac{e^{-\hat{k}}\hat{k}^8}{8!}$	0.00630
$P[X=3] = \frac{e^{-\hat{k}}\hat{k}^3}{3!}$	0.22323	$P[X=9] = \frac{e^{-\hat{k}}\hat{k}^9}{9!}$	0.00200
$P[X=4] = \frac{e^{-\hat{k}}\hat{k}^4}{4!}$	0.15931	$P[X=10] = \frac{e^{-\hat{k}}\hat{k}^{10}}{10!}$	0.00057
$P[X=5] = \frac{e^{-\hat{k}}\hat{k}^5}{5!}$	0.09095	$P[X \geq 11]$	0.00019

Table 2: Probability of Poisson distribution with assumed k.

Then we calculate the expected frequencies $E_i = np_i$. We observe that $E_9 > 5$ and $1 < E_{10} < 5$. Therefore, **Cochran's Rule** is satisfied. We could apply **Pearson's Test**. The table with expected and observed data is shown below. And we also create two graphs to compare them.

Number of incidents X (Category i)	**Expected Frequency** E_i	**Observed Frequency** O_i
0	210.34	233
1	600.41	641
2	856.96	801
3	815.46	786
4	581.96	575
5	332.24	341
6	158.07	163
7	64.47	73
8	23.01	28
9	7.306	11
10	2.08	1

Table 3: Expected and observed frequencies.

We set the **null hypothesis**

- H_0: The number of shootings follows a Poisson distribution with parameter $k = 2.85464$

For $N = 11$ categories, the statistic

$$X^2 = \sum_{i=0}^{N-1} \frac{(O_i - E_i)^2}{E_i} = 15.01$$

follows a **chi-squared distribution** with $N - 1 - m = 9$ degrees of freedom.

The critical value for $\alpha = 0.05$ is $\chi^2_{0.05,9} = 16.9$. Since $X^2 < \chi^2_{0.05,9}$, **we can not reject H_0 at the 5% level of significance**.

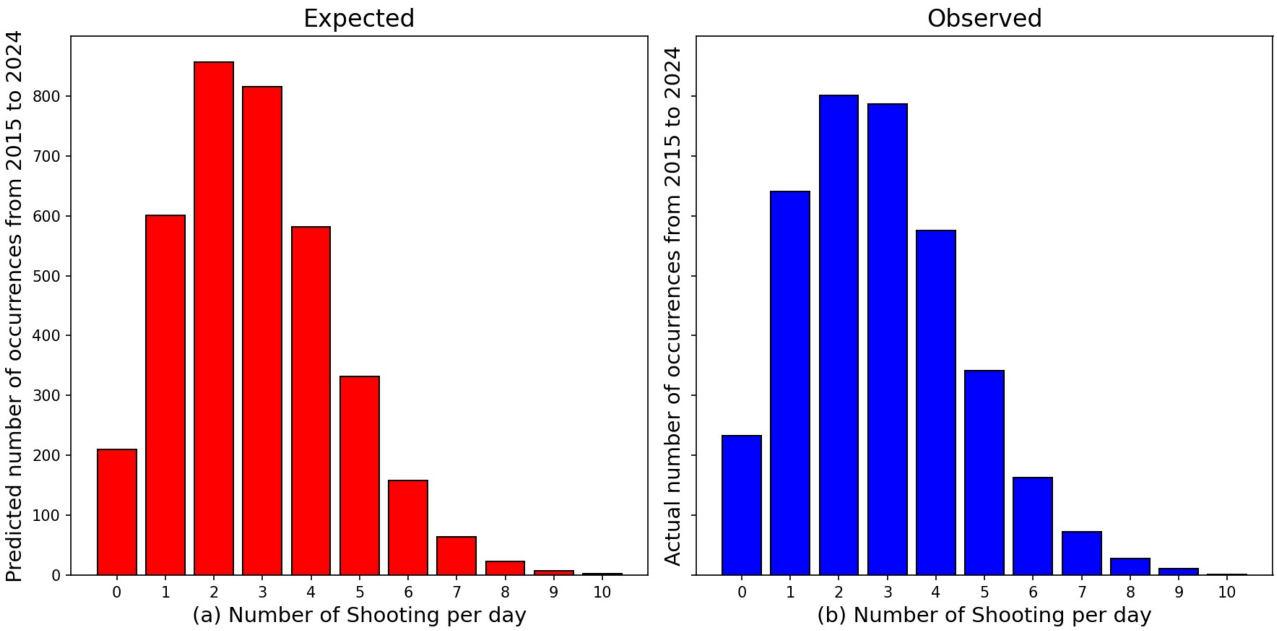

Figure 2: Comparison of the expected and observed frequencies.

4.2 Influence may brought by COVID-19

COVID-19 had a huge impact on people all over the world. From [4] we know that gun violence incidence during the COVID-19 pandemic is higher than before the pandemic in the US. To verify this opinion, we wonder whether there is an actual influence on the frequency of shooting brought by COVID-19.

We first should ensure the exact time interval of COVID-19. In March 2020, the WHO officially declared the COVID-19 outbreak a pandemic. In May 2023, the WHO and former U.S. President Joe Biden declared an end to the pandemic. Thus, based on this fact, we select two sets of data, one ranging from March 2016 to May 2019, the other ranging from March 2020 to May 2023. Thus, we ensure consistency of the comparison to make the result more persuasive.

The data from the test group could be summarized in the following graph.

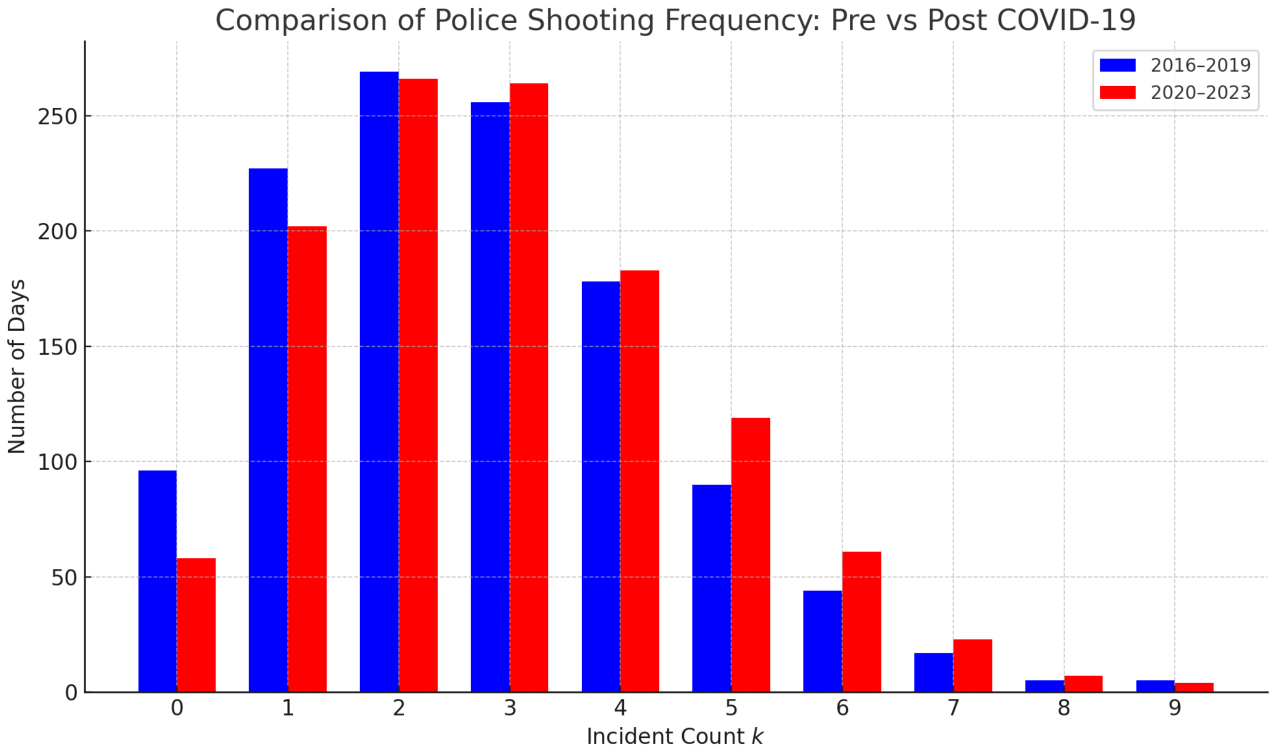

Figure 3: Comparison of incidents before and after the pandemic.

By observation, it seems that the two datasets roughly follow a Poisson distribution and the parameters are different. So, the pandemic might have an impact on it. To further explore, we could first assume these two categorizations (data before the pandemic and after the pandemic) are independent. And the contingency table is listed below.

Incident Count k	**2016–2019** Before pandemic	**2020–2023** After pandemic	$n_{i.}$
0	$n_{1,1} = 96$	$n_{1,2} = 58$	$n_{1.} = 154$
1	$n_{2,1} = 227$	$n_{2,2} = 202$	$n_{2.} = 429$
2	$n_{3,1} = 269$	$n_{3,2} = 266$	$n_{3.} = 535$
3	$n_{4,1} = 256$	$n_{4,2} = 264$	$n_{4.} = 520$
4	$n_{5,1} = 178$	$n_{5,2} = 183$	$n_{5.} = 361$
5	$n_{6,1} = 90$	$n_{6,2} = 119$	$n_{6.} = 209$
6	$n_{7,1} = 44$	$n_{7,2} = 61$	$n_{7.} = 105$
7	$n_{8,1} = 17$	$n_{8,2} = 23$	$n_{8.} = 40$
8	$n_{9,1} = 5$	$n_{9,2} = 7$	$n_{9.} = 12$
9	$n_{10,1} = 5$	$n_{10,2} = 4$	$n_{10.} = 9$
	$n_{.1}$=1187	$n_{.2} = 1187$	$n = 2374$

Table 4: Contingency table of incident counts before and after the pandemic.

Then, we try to use cell probabilities to check the independence. Suppose that

- p_{ij} is the probability of falling into the ith row and jth column.
- $p_{i.}$ is the probability of falling in the ith row.
- $p_{.j}$ is the probability of falling in the jth column.

If the categorizations are independent, we could conclude that the incidents are independent of the pandemic period. Therefore, we assume the **null hypothesis** :

- $H_0 : p_{ij} = p_{i.}p_{.j}, \quad i = 1, \dots, 10; \quad j = 1, 2; \quad (r = 10, c = 2).$

  which is equivalent to

- H_0 : The shooting incidents are independent of the pandemic period.

Then, natural estimates for the row and column probabilities are

$$\hat{p_{ij}} = \hat{p_{i.}}\hat{p_{.j}} = \frac{n_{i.}}{n} \cdot \frac{n_{.j}}{n} = \frac{n_{i.}n_{.j}}{n^2},$$

and the expected number of elements in (i, j) is

$$E_{ij} = n \cdot \hat{p_{ij}} = \frac{n_{i.}n_{.j}}{n}.$$

Now we compare the observed frequencies O_{ij} to the expected frequencies E_{ij}. The table is shown below:

Incident Count k	**Observed** O_{ij} (2016-2019),(2020-2023)	**Expected** E_{ij} (2016-2019),(2020-2023)
0	96, 58	77.0, 77.0
1	227, 202	214.5, 214.5
2	269, 266	267.5, 267.5
3	256, 264	260.0, 260.0
4	178, 183	180.5, 180.5
5	90, 119	104.5, 104.5
6	44, 61	52.5, 52.5
7	17, 23	20.0, 20.0
8	5, 7	6.0, 6.0
9	5, 4	4.5, 4.5

Table 5: Comparison of observed frequencies and expected counts.

We use the **Pearson statistic** again:

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 19.2,$$

which follows a **chi-squared distribution** with $(r - 1)(c - 1) = 9$ degrees of freedom.

The critical value for $\alpha = 0.05$ is $\chi^2_{0.05,9} = 16.9$. Since $X^2_{(r-1)(c-1)} > \chi^2_{0.05,9}$, **we decide to reject H_0 at the 5% level of significance.**

We could say that there is evidence existing an effect on the shooting frequency in the US due to the COVID-19 pandemic.

5 Analysis of Number of Shootings' Dependence on Weekdays and Months

You may wonder whether the number of shootings depends on weekdays or months, and indeed, we investigate their relations. We first use *Mathematica* to plot the corresponding figures counted by weekdays and months, which are shown in Figure 4 and Figure 5, respectively.

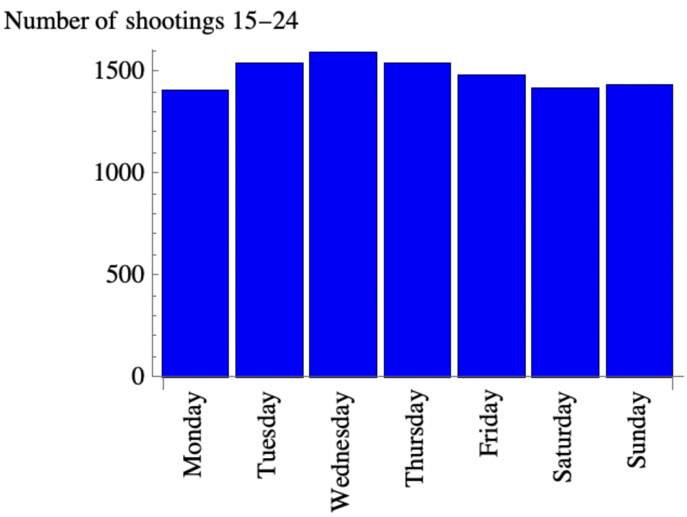

Figure 4: Number of shootings on different weekdays.

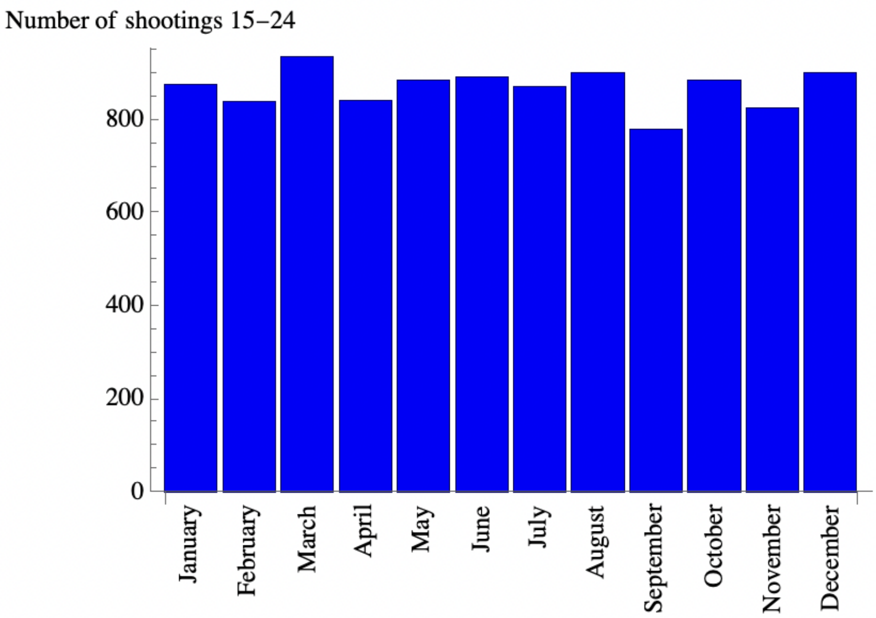

Figure 5: Number of shootings in different months.

Due to the limitation of the precision of the axis, we need to rigorously test whether the occurrence of shooting is independent of weekdays and months. We can treat this classification as categorical data, therefore, we use the test for multinomial distribution, *i.e.*, **Pearson's chi-squared goodness-of-fit test**.

First, for **the number of shootings by weekdays**, we set the **null hypothesis** to be:

H_0: The number of shootings by weekdays follows a multinomial distribution with parameters $(p_1, \cdots, p_7) = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$.

and it satisfies **Cochran's Rule** because

$$E[X_i] = \frac{10428}{7} \approx 1489.71 > 5 \text{ for all } i = 1, \cdots, 7.$$

Now, the table recording the observed and expected frequencies is listed below:

Weekday	Observed Frequency O_i	Expected Frequency E_i
Monday($i = 1$)	1409	1489.71
Tuesday($i = 2$)	1541	1489.71
Wednesday($i = 3$)	1594	1489.71
Thursday($i = 4$)	1541	1489.71
Friday($i = 5$)	1484	1489.71
Saturday($i = 6$)	1421	1489.71
Sunday($i = 7$)	1438	1489.71

Table 6: The observed and expected frequencies of shootings by weekdays.

The statistic

$$X^2 = \sum_{i=1}^{7} \frac{(X_i - np_i)^2}{np_i} = \frac{52639}{2607} \approx 20.191$$

follows **a chi-squared distribution with** 6 **degrees of freedom**, and the critical value is

$$\chi^2_{0.01,6} = 16.8.$$

Therefore, we **reject** H_0 **at** 1% **significance level** because $X^2 > \chi^2_{0.01,6}$. This implies there is evidence that the average number of police shootings depends on the weekdays actually.

Next, for **the number of shootings by months**, we set the **null hypothesis** to be:

H_0: The number of shootings by months follows a multinomial distribution with parameters $(p_1, \cdots, p_{12}) = (\frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12}, \frac{1}{12})$.

and it also satisfies **Cochran's Rule** because

$$E[X_i] = 869 > 5 \text{ for all } i = 1, \cdots, 12.$$

Similarly, we list the table recording the observed and expected frequencies as follows:

Month	Observed Frequency O_i	Expected Frequency E_i
January($i = 1$)	875	869
February($i = 2$)	838	869
March($i = 3$)	936	869
April($i = 4$)	841	869
May($i = 5$)	885	869
June($i = 6$)	891	869
July($i = 7$)	871	869
August($i = 8$)	900	869
September($i = 9$)	780	869
October($i = 10$)	885	869
November($i = 11$)	826	869
December($i = 12$)	900	869

Table 7: The observed and expected frequencies of shootings by months.

The statistic

$$X^2 = \sum_{i=1}^{12} \frac{(X_i - np_i)^2}{np_i} = \frac{18962}{869} \approx 21.821$$

follows **a chi-squared distribution with** 11 **degrees of freedom**, and the critical value is

$$\chi^2_{0.05,11} = 19.7.$$

Therefore, we **reject** H_0 **at** 5% **significance level** because $X^2 > \chi^2_{0.05,11}$. This implies there is evidence that the average number of police shootings depends on the months actually.

In summary, we can believe that **the average number of police shootings depends on weekdays and months**.

6 Calculation of the Confidence Interval for the Parameter k of Poisson Distribution

6.1 Verification of the Given Confidence Interval

Next, we explore the confidence interval of the parameter k in the Poisson distribution. Recall that if $X_1, \ldots, X_n$ is a random sample of size n taken from a normal distribution with mean μ and variance σ^2, then the sample mean $\overline{X}$ follows a normal distribution with mean μ and variance σ^2/n.

Considering "the sample size is very large", it's reasonable to apply the **Central Limit Theorem** and **approximate** $\overline{X}$ **follows a normal distribution**.

Recall the statistical properties of Poisson Distribution, *i.e.*, $E[X] = Var[X] = k$.

Therefore, $\overline{X}$ **follows a normal distribution with mean** k **and variance** k/n. Also, note that in this circumstance we choose $\overline{X}$ to be the estimator for k, which equivalently means $\hat{k}$ follows a normal distribution with mean $\hat{k}$ and variance $\hat{k}/n$.

Recall that when the variance is known, a two-sided $100(1-\alpha)\%$ confidence interval on mean μ is given by

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

In our analysis, the variance for $\hat{k}$ is $\hat{k}/n$, which implies its standard deviation to be $\sqrt{\hat{k}/n}$. Then **a** $(1-\alpha)100\%$ **confidence interval for** k is indeed given by

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

6.2 Calculation Result

Import the data from [1] into *Mathematica* and through simple calculation, we get the results of the sample size $n = 3651$ and the sample mean $\hat{k} = 2.8562$.

We then choose to calculate the 95% confidence interval ($\alpha = 0.05$), for it's the most commonly used statistic. Then, plugging in the formula in (6.1), we get **a** 95% **confidence interval for the average daily fatal police shootings from the years 2015 to 2024 is [2.80138, 2.91102]**.

7 Verification of Poisson Distribution (2025)

This part will continue research in Verification of Poisson Distribution (2015-2024). We want to check whether the data in 2025 also follows a Poisson distribution and calculate $\hat{k}$.

First, we need to calculate the parameter of the assumed Poisson distribution:

$$\hat{k} = \overline{X} = \frac{1}{90}(119 + 101 + 102) = \frac{322}{90} = 3.5778.$$

Consider it as the multinomial distribution, we first calculate the expected probability:

	Probability		**Probability**
$P[X=0] = \frac{e^{-\hat{k}}\hat{k}^0}{0!}$	0.02793	$P[X=6] = \frac{e^{-\hat{k}}\hat{k}^6}{6!}$	0.08136
$P[X=1] = \frac{e^{-\hat{k}}\hat{k}^1}{1!}$	0.09995	$P[X=7] = \frac{e^{-\hat{k}}\hat{k}^7}{7!}$	0.04159
$P[X=2] = \frac{e^{-\hat{k}}\hat{k}^2}{2!}$	0.17876	$P[X=8] = \frac{e^{-\hat{k}}\hat{k}^8}{8!}$	0.01860
$P[X=3] = \frac{e^{-\hat{k}}\hat{k}^3}{3!}$	0.21319	$P[X=9] = \frac{e^{-\hat{k}}\hat{k}^9}{9!}$	0.00739
$P[X=4] = \frac{e^{-\hat{k}}\hat{k}^4}{4!}$	0.19069	$P[X=10] = \frac{e^{-\hat{k}}\hat{k}^{10}}{10!}$	0.00264
$P[X=5] = \frac{e^{-\hat{k}}\hat{k}^5}{5!}$	0.13645	$P[X \geq 11]$	0.00145

Table 8: Probability of Poisson distribution with assumed k.

Then we calculate the expected frequencies $E_i = np_i$. We observe that $E_9 > 5$ and $1 < E_{10} < 5$. Therefore, **Cochran's Rule** is satisfied. We could apply **Pearson's Test**. Table 9 shows the expected and observed frequencies. We also create two graphs to compare them.

Number of incidents X (Category i)	**Expected Frequency** E_i	**Observed Frequency** O_i
0	2.51	0
1	9.00	10
2	16.09	21
3	19.19	20
4	17.17	16
5	12.28	7
6	7.32	9
≥ 7	6.46	7

Table 9: Expected and observed frequencies.

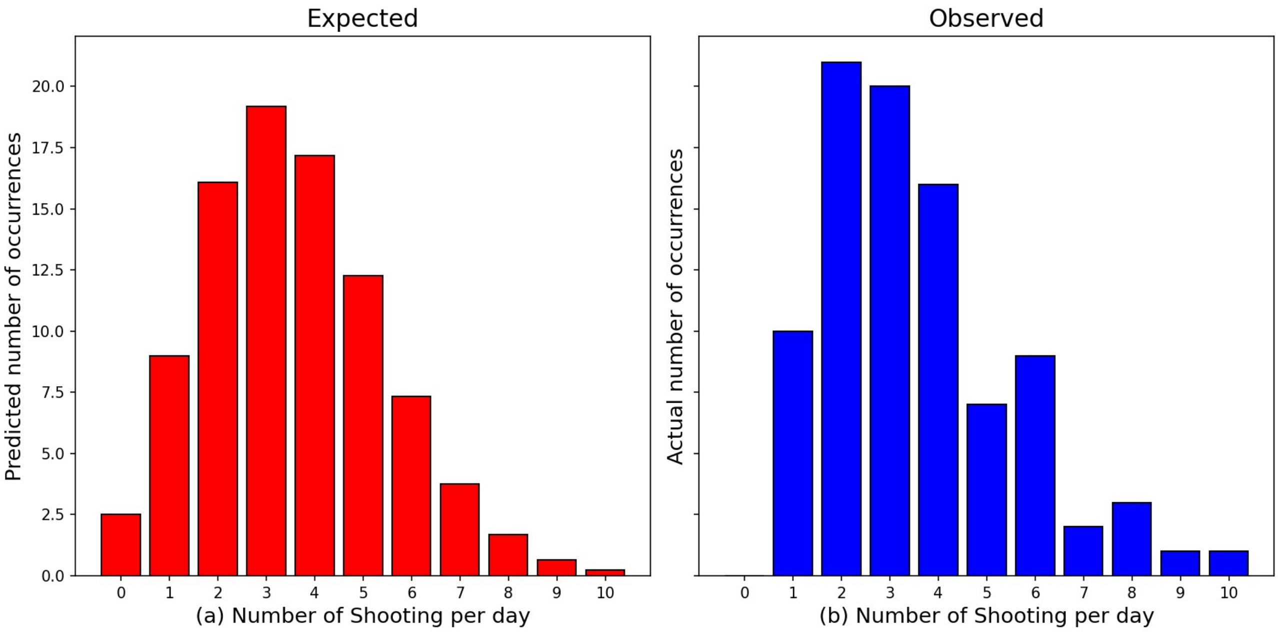

Figure 6: Comparison of the expected and observed frequencies.

We set the **null hypothesis** to be:

- H_0: The number of shootings follows a Poisson distribution with parameter $k = 3.5778$.

For $\boldsymbol{N = 8}$ categories, the statistic

$$X^2 = \sum_{i=0}^{N-1} \frac{(O_i - E_i)^2}{E_i} \approx 6.9343$$

follows **a chi-squared distribution with** $N - 1 - m = 6$ **degrees of freedom**.

The critical value for $\alpha = 0.05$ is $\chi^2_{0.05,6} = 12.6$. Since $X^2 < \chi^2_{0.05,6}$, **we can not reject** H_0 **at the** 5% **level of significance**.

Therefore, we can conclude that the data of January-March 2025 **follows a Poisson distribution with a parameter** $\boldsymbol{\hat{k}_{2025} = 3.5778}$. Besides, we can notice that $\hat{k}_{2025}$ is much larger than the value we get for 2015-2024, which might suggest that **American society is becoming more and more turbulent**.

8 Prediction Interval for 2025

8.1 Derive Nelson's formula

Prediction intervals based on the observations give us another insight into the prediction, which deserves further exploration. Thus, we try to use it to predict the interval for 2025.

First, we want to derive Nelson's formula in [5].

Assumption

Assume X is the total count in a past sample, which follows a Poisson distribution with parameter $n\lambda$, i.e., $X \sim \text{Poisson}(n\lambda)$. The future sample total count Y follows a Poisson distribution with parameter $m\lambda$, i.e., $Y \sim \text{Poisson}(m\lambda)$.

Nelson's method is based on the following assumptions:

1. Use $\hat{\lambda} = \frac{X}{n}$ as the estimate of λ.

2. Assume the distribution of $m\hat{\lambda} - Y$ can be approximated by a normal distribution.

Then we start to prove

1. **Calculate the Expectation and Variance of Y:** We know that $Y \sim \text{Poisson}(m\lambda)$. Replacing λ with $\hat{\lambda}$, we get:
$$E(Y) = m\hat{\lambda} = m \cdot \frac{X}{n},$$
and
$$\text{Var}(Y) = m\hat{\lambda} = m \cdot \frac{X}{n}.$$

2. **Construct Normal Approximation:**
$$\frac{(m\hat{\lambda} - Y) - E(m\hat{\lambda} - Y)}{\sqrt{\text{Var}(m\hat{\lambda} - Y)}} \sim N(0,1),$$
where the expectation $(m\hat{\lambda} - Y) = 0$ and the variance $\text{Var}(m\hat{\lambda} - Y)$ can be expressed as:
$$\text{Var}(m\hat{\lambda} - Y) = \text{Var}\left(m \cdot \frac{X}{n} - Y\right) = \left(\frac{m}{n}\right)^2 \text{Var}(X) + \text{Var}(Y) = \left(\frac{m}{n}\right)^2 (n\lambda) + m\lambda = m\lambda\left(\frac{m}{n} + 1\right).$$
Replacing λ with $\hat{\lambda}$, we get:
$$\text{Var}(m\hat{\lambda} - Y) = m \cdot \frac{X}{n}\left(\frac{m}{n} + 1\right).$$

3. **Construct Prediction Interval:** Based on the normal approximation, we have:
$$\frac{m\hat{\lambda} - Y}{\sqrt{m \cdot \frac{X}{n}\left(\frac{m}{n} + 1\right)}} \sim N(0,1).$$
Therefore, the prediction interval can be expressed as:
$$m\hat{\lambda} - Y \approx \pm z_{1-\alpha}\sqrt{m \cdot \frac{X}{n}\left(\frac{m}{n} + 1\right)},$$
which simplifies to:
$$Y \approx m\hat{\lambda} \pm z_{1-\alpha}\sqrt{m \cdot \frac{X}{n}\left(\frac{m}{n} + 1\right)}.$$
Substituting $\hat{\lambda} = \frac{X}{n}$, we get:
$$Y \approx m \cdot \frac{X}{n} \pm z_{1-\alpha}\sqrt{m \cdot \frac{X}{n}\left(\frac{m}{n} + 1\right)}.$$
Thus, the final prediction interval is:
$$\boldsymbol{\left[\left\lceil \frac{mX}{n} - z_{1-\alpha}\sqrt{\frac{mX}{n}\left(\frac{m}{n} + 1\right)} \right\rceil, \left\lfloor \frac{mX}{n} + z_{1-\alpha}\sqrt{\frac{mX}{n}\left(\frac{m}{n} + 1\right)} \right\rfloor\right]}.$$

8.2 Prediction Interval

Here, we take the data for 2023 and 2024 as X and use the corresponding result Y as the prediction interval for 2025. Then, we get a graph shown below:

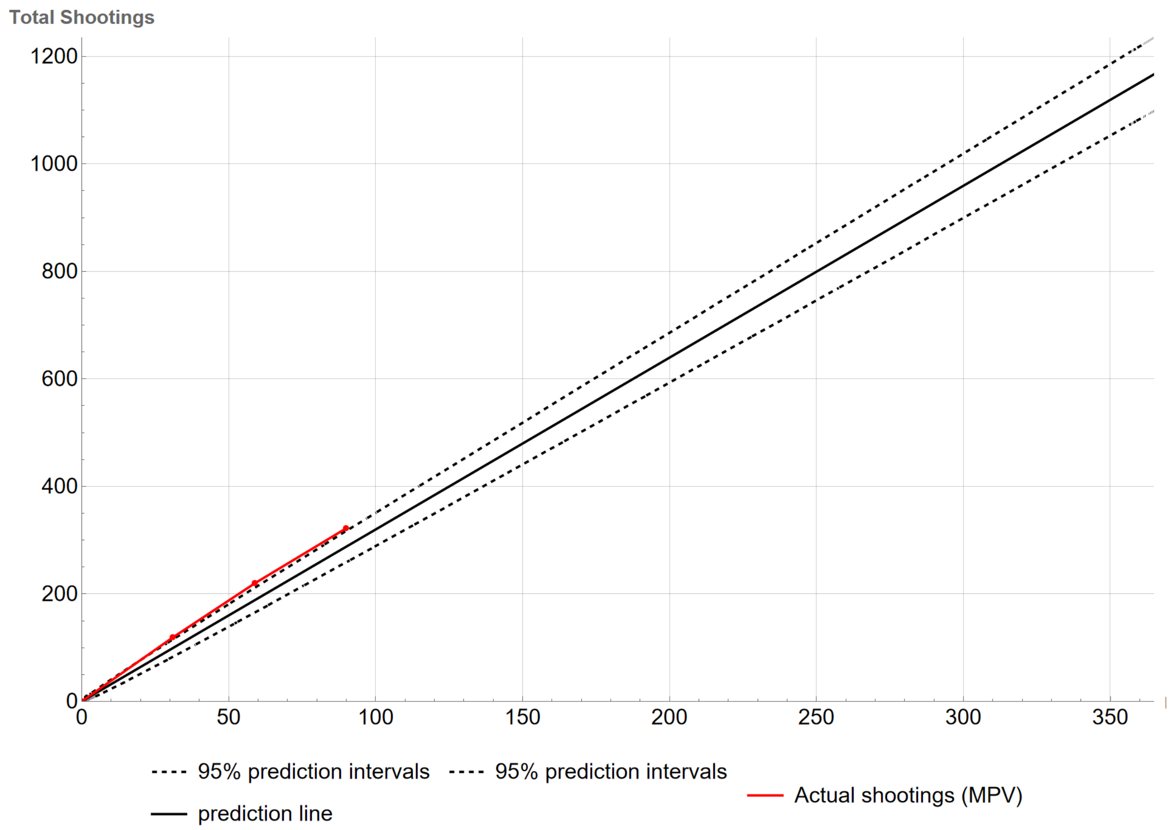

Figure 7: Actual number of shootings along with the prediction intervals.

Since we can easily observe that the red line exceeds the upper bound of the prediction interval (more shootings than we've expected), this image further supports the viewpoint we put forward in Question 6: American society might become more and more turbulent. However, in the research [6], they find that for the average number of shooting incidents per month, the data from Mapping Police Violence (the database we used for 2025) is originally larger than that from The Washington Post (the database we used for the previous years) and over time, the percentage differences in the number of reported shooting incidents among the two data sources have shown an increasing trend. Thus, the sudden increase in the number of fatal shootings may be simply caused by the **difference between the two databases**. Therefore, whether this deviation reflects troubles in the American society or not still **needs further investigation**.

9 Further Analysis

In addition to the above conclusions, more interesting analyses can be done by delving into the data.

There is "Systemic Racism" in police killings according to [7, p. 420]. They use **Multilevel Logistic Regression analysis** on MPVD (Mapping Police Violence Database) and show that the white victims are more dangerous to the police. They are more likely to show mental disease and bear arms. Still, the percentage of black victims is larger compared to their proportion of the population.

This conclusion is fun, and here we go further.

9.1 Black Victims in Democratic vs. Republican States

By comparison, Democrats were thought to pay more attention to minority rights. Then, a question arises:

Do black victims have significantly lower rates in Democratic-controlled states than in Republican states? We set our research hypothesis below:

Hypothesis: In the Democratic-controlled states, the black victimization proportions are lower than those of Republican-controlled states.

9.1.1 Data Preparation

Data are collected on the total black population in the 2020 United States census data. States are categorized as either Republican-controlled or Democratic-controlled based on their governors' party. Here, Texas and Alabama are taken as the representatives of Republican states, and California and New York are taken as the representatives of Democratic states.

And here we follow the definition of Black race in [1]. The people with multiple ethnic groups, including blacks, are treated as black. Those records without race recognition cannot be taken into consideration.

State	Political Affiliation
Texas	Republican-controlled
Alabama	Republican-controlled
California	Democratic-controlled
New York	Democratic-controlled

Table 10: Political affiliation of states.

Note. Data source: [8]

The following is their population of blacks:

State	Percentage of Black Population	Total Population	Black Population
Texas (TX)	13.60%	29,145,505	3,963,789
Alabama (AL)	26.60%	5,024,279	1,336,458
California (CA)	6.50%	39,538,223	2,569,984
New York (NY)	17.70%	20,201,249	3,575,621

Table 11: Black population statistics by state (2020 Data).

Note. Data source: [9] The table uses 2020 U.S. Census data to represent the demographic composition of states for the period 2014-2025. Assuming the population proportion remains relatively stable over this time frame.

9.1.2 Hypotheses

Definition of The black victimization Proportions:

$$p_{\text{Black, Party}} = \frac{\text{Number of Black victims in Party's states}}{\text{Number of Black population in Party's states}}.$$

Using the definition of Black Victimization Proportions, we test the following **null hypothesis**:

- H_0: No significant difference exists in the proportion of black victims between Democratic-controlled and Republican-controlled states. *i.e.*,

$$H_0 : p_{\text{Black, Democratic}} - p_{\text{Black, Republican}} = 0.$$

9.1.3 Large-sample Test for Differences in Proportions

The statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}.$$

Here, we take:

- $\hat{p}_1$: $p_{\text{Black, Democratic}}$.
- $\hat{p}_2$: $p_{\text{Black, Republican}}$.
- n_1: Total black population in California and New York.
- n_2: Total black population in Texas and Alabama.

Then $Z = 1.41939$ is calculated using the *Mathematica* code provided in Appendix A. The corresponding significance level p-value = 0.0778921.

9.1.4 Conclusion

The p-value of **0.0779** is smaller than **the significance level of 0.10** but greater than **the significance level of 0.05**. It shows that the difference in the black victimization rate between Democratic-controlled and Republican-controlled states is statistically significant at the 10% level but not significant enough at the 5% level.

9.1.5 Extension

Give a definition of β_{state}:

$$\beta_{state} = \frac{p_{\text{Black, state}}}{p_{\text{Other, state}}} = \frac{\frac{\text{Number of Black victims in state}}{\text{Number of Black population in state}}}{\frac{\text{Number of Other victims in state}}{\text{Number of Other victims in state}}}.$$

The β_{State} for the four states is calculated and plotted using the Mathematic code provided in Appendix A.

State	Black Victim Proportion	Other Victim Proportion	Beta
Texas (TX)	0.0000585299	0.0000304983	1.91912
Alabama (AL)	0.0000433983	0.0000317261	1.36791
California (CA)	0.0000805452	0.0000263469	3.0571
New York (NY)	0.0000257298	0.00000577422	4.45598

Table 12: Black and other victim proportions with beta values by state.

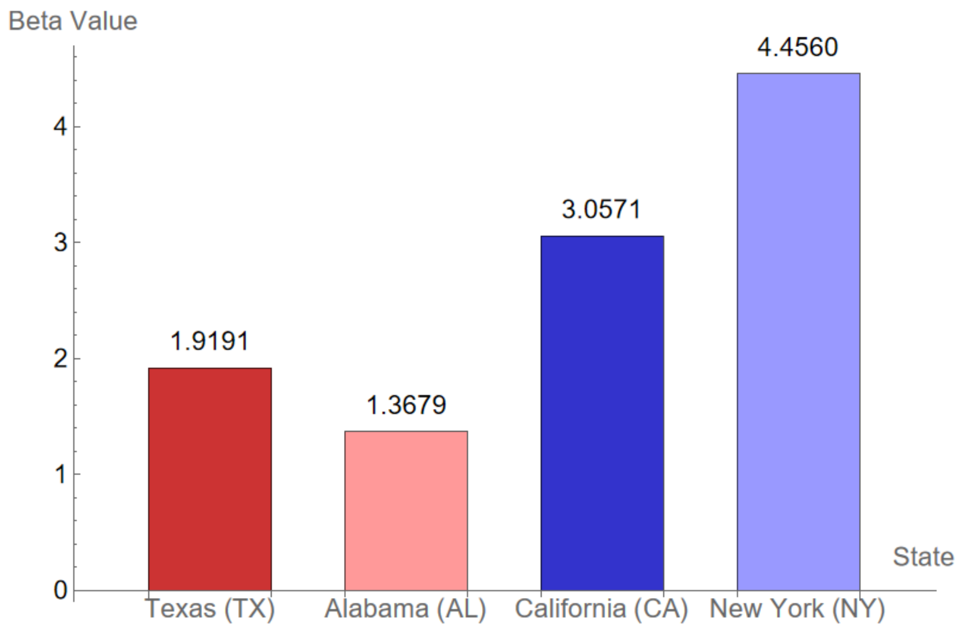

Figure 8: Beta values for four states.

According to Figure 8, the hypothesis is wrong, and it is more likely that:

In the Democratic-controlled states, the black victimization proportions are higher than those of Republican-controlled states. It may be due to many conflict factors, like the different degrees of urbanization or the different law enforcement intensity between Republican-controlled states and Democratic-controlled states.

To be more specific, states controlled by the Democratic Party usually have a higher degree of urbanization. Over there, the density of population is much larger than the county controlled by the Republican Party. Due to the dense population and high social complexity, the crime rate may be higher and influence our research. What's more, the social contradictions may be more serious and lead to an increasing tendency of risk of black people becoming victims.

9.2 Influence of Body-Worn Cameras on Racial Disparities in Fatal Shootings

According to [10], protests against police violence promote the probability of building up a Citizen Review Board. For example, they promote the usage of body-worn cameras.

The relationship between the year and the proportion of body camera usage is plotted with Python code provided in Appendix, based on data in [1].

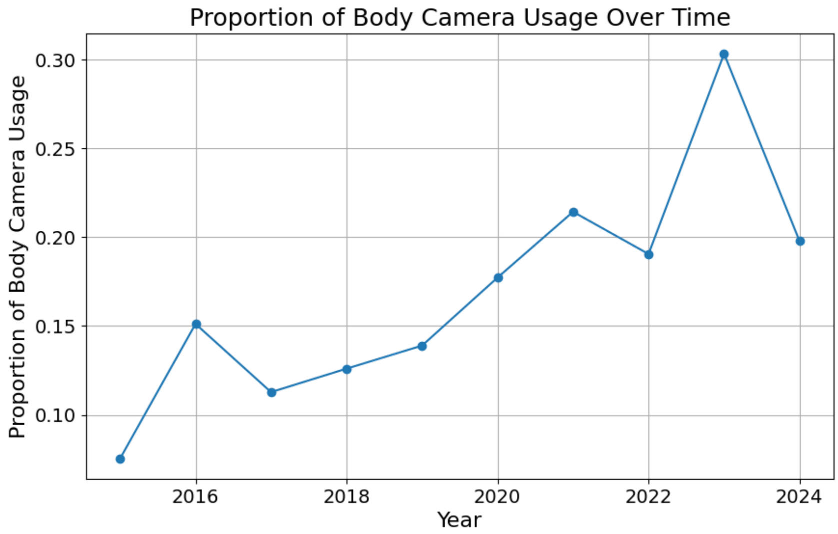

Figure 9: Body camera usage with year shift.

We can view that as the year goes by, the usage of body cameras is increasing.

Here, we explore the difference in the proportion of black victims in body camera usage police shooting incidents and unrecorded police shooting incidents.

9.2.1 Hypotheses

- H_0: No significant difference in the proportion of black victims in body camera usage police shooting incidents and unrecorded police shooting incidents, thus $p_1 = p_2$.

Here, p_1 refers to the proportion of black victims in camera usage police shooting incidents. p_2 refers to the proportion of black victims in unrecorded police shooting incidents.

9.2.2 Confidence Interval for $p_1 - p_2$

Here, we use the **95% confidence interval** for the difference:

$$(p_1 - p_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

Here, $z_{\alpha/2} = 1.96$. The confidence interval is calculated using the *Mathematica* code provided in Appendix A:

$$\text{CI for } (p_1 \text{ - } p_2) : [0.0833386, 0.130105].$$

9.2.3 Conclusion

0 isn't included in the confidence interval, which shows a statistically significant difference in the proportion of black victims for body camera usage incidents and unrecorded incidents.

Furthermore, the proportion of black victims in body camera usage incidents p_1 **is statistically significantly higher than the proportion in unrecorded incidents** p_2**.**

10 Conclusion

This project conducts a comprehensive statistical analysis of fatal police shootings in the United States from 2015 to 2024 by using probability and statistical tools to explore any possible influencing factors.

First, we analyze the data and find that the daily frequency of fatal police shootings follows a Poisson distribution with a maximum-likelihood estimated parameter $\hat{k} = 2.85464$. By applying a chi-squared test ($X^2 \approx 15.01 < \chi^2_{0.05,9} = 16.9$), we fail to reject the null hypothesis, supporting the assumption of Poisson distribution. Taking COVID-19 into consideration, we use another chi-squared test ($X^2 \approx 19.2 > \chi^2_{0.05,9} = 16.9$) to show that the pandemic significantly changes the distribution of fatal shootings.

Then, we observe significant variations in shooting frequencies across weekdays ($X^2 \approx 20.19 > \chi^2_{0.01,6} = 16.8$) and months ($X^2 \approx 21.82 > \chi^2_{0.05,11} = 19.7$). These results suggest that temporal factors influence incident rates.

We also use the same methodology to analyze the data for 2025 to verify its accuracy. This verification helps ensure that the statistical patterns observed in the previous years are consistently true for the current year and can provide an indication of any trends or shifts in the distribution of fatal police shootings. The final result did not fall within the interval, which perhaps implies the problems emerging in society and also makes us pay attention to the issue of uniformity among different databases.

For further analysis, we also discuss the comparison of black victimization rates between Democratic states and Republican ones, as well as the relationship between carrying a body camera and black victimization. Data analysis shows that Democratic states have a higher proportion of black victims than Republican ones. Moreover, the proportion of black victims in body camera usage incidents is statistically significantly higher than the proportion in unrecorded incidents, which indicates that systemic biases exist even under monitoring.

Acknowledgements

Thanks to Prof. Horst Hohberger's dedication and unique insight on Probability and Statistics, we gained a lot from this course. Also, we appreciate the collaboration of our team members, as it is our mutual support that makes this project successful.

References

[1] The Washington Post, "Fatal force," https://github.com/washingtonpost/data-police-shootings, 2024, accessed April 1st, 2025.

[2] D. Spiegelhalter and A. Barnett, "London murders: a predictable pattern?" *Significance*, vol. 6, no. 1, pp. 5–8, 2009. [Online]. Available: https://doi.org/10.1111/j.1740-9713.2009.00334.x

[3] H. Hohberger, *ECE4010J Probabilistic Methods in Engineering.* Shanghai, China: University of Michigan - Shanghai Jiao Tong University Joint Institute, 2025, spring term.

[4] P. Ssentongo et al., "Gun violence incidence during the covid-19 pandemic is higher than before the pandemic in the united states," *Scientific Reports*, vol. 11, no. 20654, 2021. [Online]. Available: https://doi.org/10.1038/s41598-021-98813-z

[5] K.Krishnamoorthy and J. Peng, "Improved closed-form prediction intervals for binomial and poisson distributions," *Journal of Statistical Planning and Inference*, vol. 141, pp. 1709 – 1718, 2011. [Online]. Available: https://doi.org/10.1016/j.jspi.2010.11.021

[6] B. P. Comer and J. R. Ingram, "Comparing fatal encounters, mapping police violence, and washington post fatal police shooting data from 2015–2019: A research note," *Criminal Justice Review*, vol. 48, no. 2, pp. 249–261, 2023.

[7] R. T. DeAngelis, "Systemic racism in police killings: New evidence from the mapping police violence database, 2013–2021," *Race and Justice*, vol. 14, no. 3, pp. 413–422, 2024. [Online]. Available: https://doi.org/10.1177/21533687211047943

[8] National Governors Association, "Governors' bios and terms," 2015–2024, retrieved from National Governors Association website. [Online]. Available: https://www.nga.org/governors/bios/

[9] U.S. Census Bureau, "Quickfacts," 2023, retrieved from U.S. Census Bureau QuickFacts website. [Online]. Available: https://www.census.gov/quickfacts/fact/table

[10] S. Olzak, "Does protest against police violence matter? evidence from us cities, 1990 through 2019," *American sociological review*, vol. 86, no. 6, pp. 1066–1099, 2021. [Online]. Available: https://doi.org/10.1177/00031224211056966

A Appendix: Code

```
X = 2337;
n = 731;
alpha = 0.05;
z1alpha = InverseCDF[NormalDistribution[0, 1], 1 - alpha];

f1[m_] := Ceiling[(m*X)/n - z1alpha*Sqrt[(m*X)/n*(m/n + 1)]];
f2[m_] := Floor[(m*X)/n + z1alpha*Sqrt[(m*X)/n*(m/n + 1)]];
f3[m_] := m*X/n;
points = {{0, 0}, {31, 119}, {59, 220}, {90, 322}};

plotFunction =
  Plot[{f1[m], f2[m], f3[m]}, {m, 1, 365},
   PlotStyle -> {Directive[Black, Dashed, Thick],
     Directive[Black, Dashed, Thick], Directive[Black, Thick]},
   AxesLabel -> {"m", Style["Total Shootings", FontSize -> 16, Bold]},
    PlotLegends ->
    Placed[LineLegend[{"95% prediction intervals",
       "95% prediction intervals", "prediction line"},
      LegendLayout -> "Row", LegendMarkerSize -> {30, 10},
      LabelStyle -> {FontSize -> 18} ], Below],
   GridLines -> Automatic, ImagePadding -> {{60, 10}, {40, 30}},
   BaseStyle -> {FontSize -> 18} ];

plotPoints =
  ListLinePlot[points, PlotStyle -> Directive[Red, Thick],
   Mesh -> All, MeshStyle -> Directive[PointSize[Medium], Red],
   PlotLegends ->
    Placed[LineLegend[{"Actual shootings (MPV)"},
      LegendMarkerSize -> {30, 10}, LabelStyle -> {FontSize -> 18} ],
     Below]];

Show[plotFunction, plotPoints, PlotRange -> All,
 PlotRangePadding -> {{0, 0}, {0, 0.2}}]
```

Listing 1: Mathematica Code for Q7.

```
SetDirectory["ROADTOTHEDATA"];
dataPopulation = Import["population_states.xlsx", {"Data", 1}];
blackPopulation = dataPopulation[[1 ;;, 4]];
stateNames = dataPopulation[[1 ;;, 1]];
stateBlackPopulation =
    AssociationThread[stateNames -> blackPopulation];
TXBlackPopulation = stateBlackPopulation["Texas"];
ALBlackPopulation = stateBlackPopulation["Alabama"];
CABlackPopulation = stateBlackPopulation["California"];
NYBlackPopulation  = stateBlackPopulation["New York"];

RepublicanBlackPopulation = TXBlackPopulation + ALBlackPopulation;
DemocraticBlackPopulation = CABlackPopulation + NYBlackPopulation;

Print["Republican Black Population: ", RepublicanBlackPopulation];
Print["Democratic Black Population: ", DemocraticBlackPopulation];
```

```
(* Output: Republican Black Population: 5.30025*10^6 *)
(* Output: Democratic Black Population: 6.14561*10^6 *)

dataShootings = Import["fatal-police-shootings-data.csv"];
headers = First[dataShootings];
Content = Rest[dataShootings];

targetStates = {"TX", "AL", "CA", "NY"};
BlackNation = "B";

countB =
  Association[
   Table[state ->
     Count[Content, {_, _, s_, BlackNation} /; s == state], {state,
     targetStates}]];

countOther =
  Association[
   Table[state ->
     Count[Content, {_, _, s_, race_} /;
       s == state && race != BlackNation], {state, targetStates}]];

Print["Number of events with race B by state: ", countB];
Print["Number of events with other races by state: ", countOther];

TXVictimBlack = Lookup[countB, "TX", 0];
TXVictimOthers = Lookup[countOther, "TX", 0];
ALVictimBlack = Lookup[countB, "AL", 0];
ALVictimOthers = Lookup[countOther, "AL", 0];
CAVictimBlack = Lookup[countB, "CA", 0];
CAVictimOthers = Lookup[countOther, "CA", 0];
NYVictimBlack = Lookup[countB, "NY", 0];
NYVictimOthers = Lookup[countOther, "NY", 0];

RepublicanBlackVictim = TXVictimBlack + ALVictimBlack;
DemocraticBlackVictim = CAVictimBlack + NYVictimBlack;

Print["Republican Black Victim Numbers: ", RepublicanBlackVictim];
Print["Democratic Black Victim Numbers: ", DemocraticBlackVictim];

(* Output: Number of events with race B by state: <|TX->232,AL->58,CA
   ->207,NY->92|> *)
(* Output: Number of events with other races by state: <|TX->768,AL
   ->117,CA->974,NY->96|> *)
(* Output: Republican Black Victim Numbers: 290 *)
(* Output: Democratic Black Victim Numbers: 299 *)

P1 = RepublicanBlackVictim / RepublicanBlackPopulation;
P2 = DemocraticBlackVictim / DemocraticBlackPopulation;

Z = ((P1 - P2) - 0)/ (Sqrt[
     P1 * (1 - P1)/RepublicanBlackPopulation +
```

```
      P2 * (1 - P2) /DemocraticBlackPopulation]);
Print["Z:", Z];
pValue = 1 - CDF[NormalDistribution[], Z];
Print["P-Value: ", pValue];

(* Output: Z:1.41939 *)
(* Output: P-Value: 0.0778921 *)

otherPopulation = dataPopulation[[1 ;;, 3]] - dataPopulation[[1 ;;,
   4]];

stateOtherPopulation =
  AssociationThread[stateNames -> otherPopulation];
TXOtherPopulation = stateOtherPopulation["Texas"];
ALOtherPopulation = stateOtherPopulation["Alabama"];
CAOtherPopulation = stateOtherPopulation["California"];
NYOtherPopulation  = stateOtherPopulation["New York"];

TXPBlack = TXVictimBlack/TXBlackPopulation;
ALPBlack = ALVictimBlack/ALBlackPopulation;
CAPBlack = CAVictimBlack/CABlackPopulation;
NYPBlack = NYVictimBlack/NYBlackPopulation;

TXPOther = TXVictimOthers/TXOtherPopulation;
ALPOther = ALVictimOthers/ALOtherPopulation;
CAPOther = CAVictimOthers/CAOtherPopulation;
NYPOther = NYVictimOthers/NYOtherPopulation;

TXBeta = TXPBlack/TXPOther;
ALBeta = ALPBlack/ALPOther;
CABeta = CAPBlack/CAPOther;
NYBeta = NYPBlack/NYPOther;

Print["TXBeta: ", TXBeta]; Print["TXPBlack: ", TXPBlack];
Print["TXPOther: ", TXPOther];
Print["ALBeta: ", ALBeta]; Print["ALPBlack: ", ALPBlack];
Print["ALPOther: ", ALPOther];
Print["CABeta: ", CABeta]; Print["CAPBlack: ", CAPBlack];
Print["CAPOther: ", CAPOther];
Print["NYBeta: ", NYBeta]; Print["NYPBlack: ", NYPBlack];
Print["NYPOther: ", NYPOther];

(* Output: TXBeta:1.91912 *)
(* Output: TXPBlack:0.0000585299 *)
(* Output: TXPother:0.0000304983 *)

(* Output: ALBeta:1.36791 *)
(* Output: ALPBlack:0.0000433983 *)
(* Output: ALPother:0.0000317261 *)

(* Output: CABeta:3.0571 *)
(* Output: CAPBlack:0.0000805452 *)
(* Output: CAPother:0.0000263469 *)
```

```

(* Output: NYBeta:4.45598 *)
(* Output: NYPBlack:0.0000257298 *)
(* Output: NYPother:5.77422x10^-6 *)

betaValues = <|"Texas (TX)" -> 1.91912, "Alabama (AL)" -> 1.36791,
   "California (CA)" -> 3.0571, "New York (NY)" -> 4.45598|>;

BarChart[betaValues, ChartLabels -> Automatic,
 ChartStyle -> {RGBColor[178/255, 34/255, 52/255],
   RGBColor[1, 0.5, 0.5],
   RGBColor[60/255, 59/255, 110/255],
   RGBColor[135/255, 206/255, 250/255]
 AxesLabel -> {"State", "Beta Value"}, BarSpacing -> 0.6,
 ImageSize -> 600, LabelStyle -> {FontSize -> 16},
 ImagePadding -> {{50, 10}, {50, 20}},
 Epilog ->
  Table[Text[
    Style[NumberForm[betaValues[[i]], {Infinity, 4}], FontSize -> 16,
     FontColor -> Black], {i, betaValues[[i]] + 0.1}, {0, -1}], {i,
    Length[betaValues]}]
```

Listing 2: Mathematica Code for Q8 Black Victims in Democratic vs. Republican States Mathematica Code.

```python
import pandas as pd
import matplotlib.pyplot as plt

data = pd.read_csv(file_path)

dates = pd.to_datetime(data['date'])
body_camera = data['body_camera']
years = dates.dt.year
yearly_stats = pd.DataFrame({'Year': years, 'BodyCamera': body_camera})
yearly_stats = yearly_stats.groupby('Year')['BodyCamera'].apply(lambda x: (x == True).mean()).reset_index()

plt.figure(figsize=(10, 6))
plt.plot(yearly_stats['Year'], yearly_stats['BodyCamera'], marker='o', linestyle='-')
plt.xlabel("Year", fontsize=16)
plt.ylabel("Proportion of Body Camera Usage", fontsize=16)
plt.title("Proportion of Body Camera Usage Over Time", fontsize=18)
plt.grid(True)
plt.xticks(fontsize=14)
plt.yticks(fontsize=14)
plt.show()
```

Listing 3: Proportion of Body Camera Usage vs. Year Python Code.

```
SetDirectory["ROADTOTHEDATA"];

dataShootings = Import["body_camera_usage_Race.csv"];
```

```
headers = First[dataShootings];
content = Rest[dataShootings];

Race = "B";
Recorded = "TRUE";
Unrecorded = "FALSE";

countRecordedB = Count[content, {Race, Recorded}];
totalRecorded = Count[content, {_, Recorded}];

countUnrecordedB = Count[content, {Race, Unrecorded}];
totalUnrecorded = Count[content, {_, Unrecorded}];

p1 = N[countRecordedB/totalRecorded];
p2 = N[countUnrecordedB/totalUnrecorded];
Print["p1 = ", p1];
Print["p2 = ", p2];
z = 1.96;
se = Sqrt[(p1 (1 - p1)/totalRecorded) + (p2 (1 - p2)/
       totalUnrecorded)];
ciLow = p1 - p2 - z*se;
ciHigh = p1 - p2 + z*se;

Print["CI: [", ciLow, ", ", ciHigh, "]"];

(* Output: p1 = 0.326087 *)
(* Output: p2 = 0.219365 *)
(* Output: CI: [0.0833386, 0.130105] *)
```

Listing 4: CI calculation for body camera usage and Race.