UM-SJTU Joint Institute
Probabilistic Methods in Engineering
(VE401)

Term Project

Group 1

Name: Hao Liang ID: 520370910094
Name: Chen Yunuo ID: 519021910839
Name: Yu Weixin ID: 520370910104
Name: Jin Yihan ID: 519370910073
Name: Zhenjie Sun ID: 520021911223
Date: April 18, 2022

Contents

1 Synopsis

A goodness-of-fit test for Poisson distribution is performed based on data span of 7 years as well as on data from the year 2020, respectively. Based on the tests, we conclude that there is no strong evidence that the shootings do not follow a Poisson distribution or the COVID-19 has caused a significant effect.

Apart from the overall distribution, analysis of whether the rate of occurrence of shootings depend on weekday or month is made using Fisher tests.The conclusion is that the average number of shootings depends on weekday but not on month.

In the fifth part, we calculate the confidence intervals from the data of the years 2015 to 2018 based on our result in the former part that X follows a Poisson distribution and the assumption that X mean follows a normal distribution.

A goodness-of-fit test for Poisson distribution is performed based on data span of January - March 2022. Based on the tests, we conclude that there is no sufficient evidence to reject the hypothesis that the shootings follow a Poisson distribution.

In the seventh part, we use the Nelson's method for Poisson Distribution to give a prediction interval on the number of fatal police shootings in 2022 and make a comparison with the number of cases so far, finding that the 95% prediction interval gives a quite fair prediction.

In the last part, we found that, feature "state" has no missing information, which means that we can analyse the dataset precisely and adequately with this feature. Then, with State Bar Graph of Fatal Police Shootings and State-wise Distribution graph of Fatal Police Shootings in USA, we intuitively find that Fatal Police Shootings show distinct regional character. So, we made a linear regression about Fatal Police Shootings and the ratio of arrested people in each state. To test whether it is significant, we use mathematica to analyze the data. After observing parameter table and scatter diagram, we find that there is no enough evidence to reject the $H_0 : \beta_1 = 0$, so we cannot prove the regression is significant.

2 Introduction

2.1 Background

We have learnt several kinds of distributions and hypothesis test methods in VE401, which are widely used in research and production. Practice is the best way to acquire knowledge so that we apply what we learnt to study the fatal shootings fired by the police in the United States, though not many researches have done in this field.

Fatal police shooting happens almost every day in the United States, which is in the headlines all the time. According to the data from the *The Washington Post*, there are nearly 1000 fatal shootings, the number remaining relatively constant, from the police nationwide every year since the Post started to collect statistics[1]. This leads to severe social turmoil. According to a research based on the online Newsbank archives, there are "1228 protests against police brutality in 170 cities from 1990 through the end of 2018."[2] We want to figure out whether the number of fatal shootings follows a specific distribution and whether it depends on weekday and month.

Figure 1: Protest erupts in Charlotte, N.C. after a fatal police shooting.

Our data comes from *The Washington Post*'s database, which contains records of every fatal shooting in the line of duty by a police officer in the United States since January 1, 2015 [1]. The term "fatal police shooting" here means people were shot dead in the line of duty by police in the United States. However, the data do not include people who died in police custody, fatal shootings by off-duty police or were not shot dead [1].

2.2 Introduction to data

The Washington Post also records whether the victims were armed and had mental illness or not as well as victims' race, gender and some other detailed information. Let's first focus on the number of fatal police shootings happened every day. What should be mentioned here is that 2016 and 2020 are leap years. There is 1 fatal shooing on February 29, 2016 and 3 fatal shootings on February 29,2020.

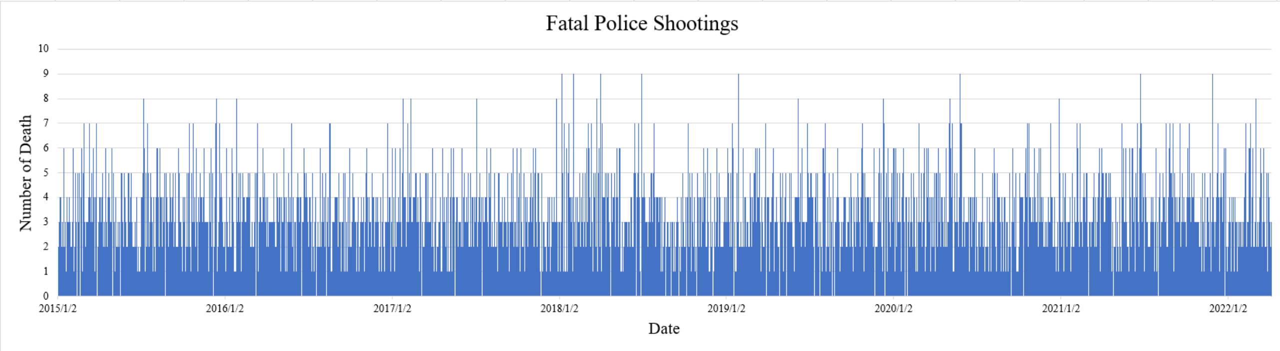


Figure 2: Number of fatal police shootings from February 2015 to April 2022 (Histogram).

To highlight the trend in number of fatal shots fired each day, I also use Mathematica to plot a line chart. From Fig.2 (shown in the next page), it's quite clear that there are commonly 1 to 5 fatal shots fired from the police in the line of duty every day.

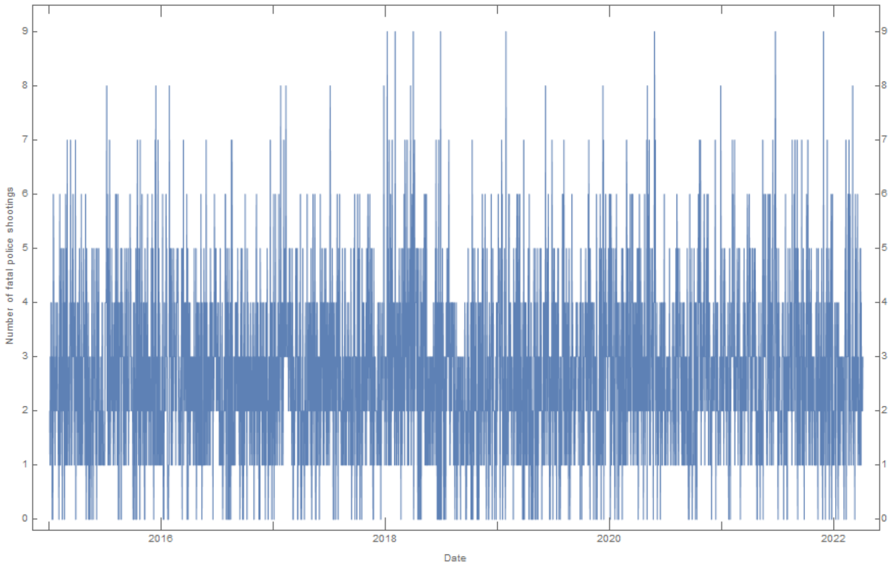
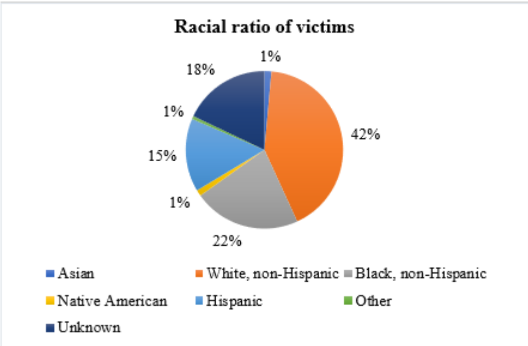
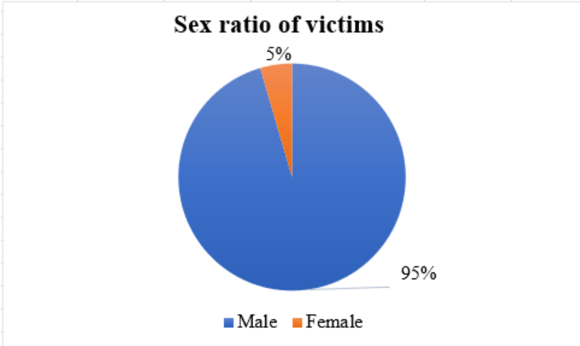


Figure 3: Number of fatal police shootings from February 2015 to April 2022 (Line chart).

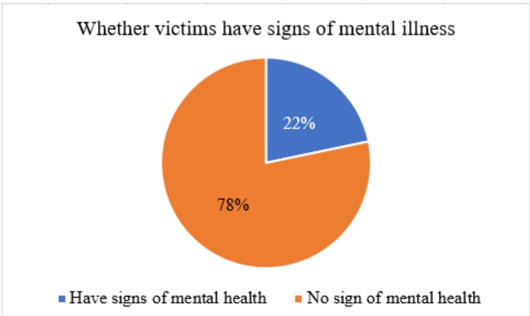
Let's now zoom in other detailed data. Fig. 3 (shown in the next page) shows that most victims are white, followed by black people and asian. Second, most of the victims are male, accounting for 95% of all victims. Third, only less than a quarter of victims have signs of mental illness and 78% of victims are mentally healthy.



(a) Racial ratio of victims



(b) Sex ratio of victims



(c) Whether victims have mental illness

Figure 4: Data of race, gender and mental health of victims.

Figure 5 is a histogram, plotted by Mathematica and the Freedman-Diaconis Rule applied, of victims' age. It's apparent that most victims are in their twenties to forties.

Figure 5: Distribution of victims' age.

3 Goodness-of-Fit Test for Poisson Distribution

If fatal police shooting happened randomly, the number of shootings each day should follow a Poisson distribution. In this section, whether the shootings follow a Poisson Distribution is tested. Since the COVID-19 pandemic has influenced the US society in many ways, such as economy and unemployment rate, we doubt whether the statistical results of fatal police shootings would be affected. Thus, the particular data set starting from the year 2020 is tested on with the null hypothesis that the fitted Poisson distribution model for the whole data span (2015.1 to 2022.4) still works.

3.1 Test on Data from 2015.1 to 2022.4

We want to find out whether the fatal police shootings follow a Poisson distribution by a Goodness-of-fit test. Formally, we set the null hypothesis to be:

H_0: The fatal police shootings from 2015.1 to 2022.4 follow Poisson distribution.

We have looked at the patterns of shootings between 2015.1.1 and 2022.4.5. There are a total 7246 shootings over the 2652 days. The average value of shootings per day can serve as a good estimator for the parameter k for Poisson distribution, and thus we estimate k as

$$\hat{k} = \overline{X} = 7246/2652 = 2.73. \tag{1}$$

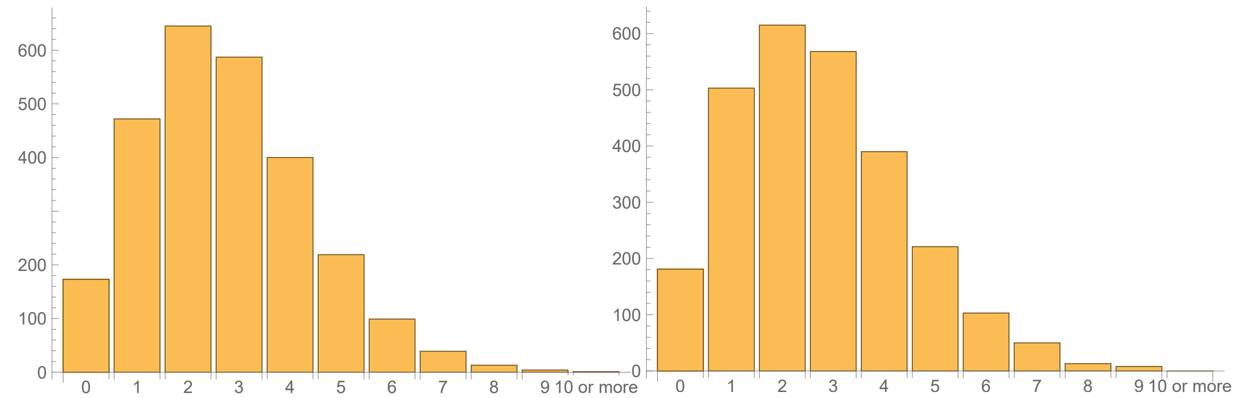
By plugging $\hat{k}$ and x into the probability density function for Poisson distribution

$$f_x(x) = \frac{k^x e^{-k}}{x!}, \tag{2}$$

the probability that exactly x shootings occur on each day is calculated. Multiplying with the total days, the expected pattern in terms of the number of days x shootings occur is computed. The probability and expected pattern are shown in Table 1 and Figure 6. The bar chart in Figure 6 is plotted by Mathematica.

Table 1: Expected and observed number of days with $x = 0, 1, 2, ...$ numbers of fatal police shootings per day, as well as the expected probability, for the shootings in the United States between January 1, 2015 and April 5,2022, recorded by *The Washington Post*'s database

	Numbers of days with the following numbers of fatal police shootings per day:										
	0	1	2	3	4	5	6	7	8	9	10 or more
Probability (%)	6.52	17.80	24.30	22.11	15.09	8.24	3.74	1.46	0.49	0.15	0.05
Expected	172	472	645	587	400	219	99	39	13	4	1
Observed	181	503	615	568	390	221	103	50	13	8	0



(a) Predicted number of fatal police shootings in the 7 years (b) Actual number of fatal police shootings in the 7 years

Figure 6: The frequency of fatal police shootings occurrence between January 1,2015 and April 5, 2022: (a) predicted; (b) observed.

From Table 1 and Figure 6, we can see that the expected number of days is close to that of the observed ones despite some fluctuations. For example, 6.52% days are expected to be have no police shootings at all, while the observed value is 6.83% (181/2652). Statistically, a χ^2 test based on Pearson Statistic is performed. Since all of the expected value is greater than 1, and only 2 out of 11 of them is less than 5, the Cochran's Rule is satisfied.

The Pearson statistic is calculated as

$$\chi_9^2 = \sum_{x=0}^{x>=10} \frac{(O_i - E_i)^2}{E_i} = 12.95, \tag{3}$$

where 9 is the degrees of freedom of the χ^2 statistic, since there are 11 categories (0 to ≤ 10) in total, and another 1 degree of freedom is subtracted because $\hat{k}$ is estimated from the same group of data.

Using Mathematica, the P-value for the test is thus $0.16 > 0.05$. As a result, the null hypothesis is not rejected. If it follows exactly Poisson distribution with $\hat{k} = 2.73$, the probability of getting this χ^2 value is 16%, which is fairly large.

Hence, there is no evidence that the fatal police shootings in the United States do not follow a Poisson Distribution.

3.2 Analysis of COVID-19 Effect

Although there is no evidence against that the fatal police shootings from 2015 to 2022 follow a Poisson distribution, we still want to figure out whether the pattern applies specifically to the years after 2020 since the COVID-2019 has affected the US society significantly.

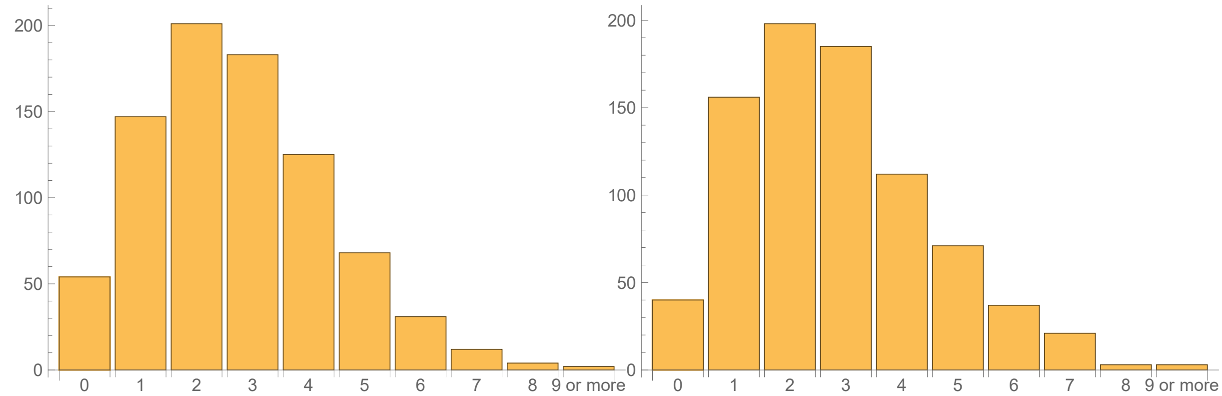
Formally, we set the Null hypothesis to be

H_0: The fatal police shootings from the year 2020 follows a Poisson distribution with parameter 2.73.

Since the total number of days is reduced, we merge the category $x = 9$ and $x \geq 10$ to satisfy Cochran's rule. Following the same procedure as described in the previous section, the expected pattern and the observed data are presented in Table 2 and Figure 7. The bar chart in Figure 7 is plotted by Mathematica.

Table 2: Expected and observed number of days with $x = 0, 1, 2, ...$ numbers of fatal police shootings per day for the shootings in the United States between January 1, 2020 and April 5,2022, recorded by *The Washington Post*'s database

Numbers of days with the following numbers of fatal police shootings per day:									
	0	1	2	3	4	5	6	7	8	9 or more
Expected	54	147	201	183	125	68	31	12	4	2
Observed	40	156	198	185	112	71	37	21	3	3



(a) Predicted number of fatal police shootings in the 7 years (b) Actual number of fatal police shootings in the 7 years

Figure 7: The frequency of fatal police shootings occurrence between January 1,2015 and April 5, 2022: (a) predicted; (b) observed.

Following the same calculation as before, the Pearson statistic is calculated to be 14.39, and the P-value is $0.072 > 0.05$. The P-value is still large, and thus the null hypothesis can not be rejected. But the data is less likely ($0.072 < 0.16$) to be observed if the fatal police shootings in 2020 still follows Poisson distribution with $k = 2.73$.

Moreover, if we only focus on the year after 2020, the estimated parameter $\hat{k}$ is 2.82, which is larger than the estimator for the 7 years. Thus, one may guess the occurrence rate for fatal police shootings increases a little bit after the year 2020. However, according to the previous test which gives the P-value $0.072 > 0.05$, we still do not have enough evidence to conclude any significant effect of COVID-19 on the fatal police shootings pattern.

4 Analysis of Shooting's Dependence on Weekday and Month

The days in a week is related to people's working situation and month is related to weather and season. One may suspect the fatal police shootings may be influenced by these factors. In this section, fatal police shootings' dependence on weekday and month is investigated.

4.1 Dependence on Weekday

A bar chart showing the frequency of occurrence of fatal police shootings on each weekday is plotted using Matlab, as shown in Figure 8.

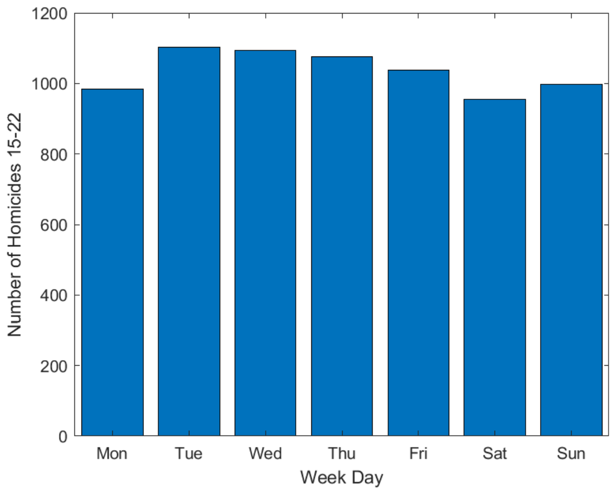


Figure 8: Frequency of occurrence of fatal police shootings on each weekday.

From Figure 8, we can see that number of occurrence of shootings on each weekday fluctuates, in which the highest frequency is on Tuesday, and the lowest is on Saturday.

Formally, a Fisher test with the null hypothesis

H_0: The probability of shootings' occurrence on each weekday is equal

is performed. Note that the null hypothesis is given prior to the plot of Figure 8.

The expected number of shootings each weekday in the total 7246 murders is $7246/7 = 1035$ days. The Pearson statistic is given by

$$\chi_6^2 = \sum_{i=1}^{i=7} \frac{(O_i - E_i)^2}{E_i} = 19.19, \tag{4}$$

and thus the P-value is calculated by Mathematica as $0.0039 < 0.05$. Thus, we reject the null hypothesis at significance level 0.0039.

We conclude that there is strong evidence that the average number of shootings depends on weekday. However, since we should not base our hypothesis on our pre-test of the same data set, and we do not have prior guess about on which specific weekday is the shootings most or least often, we do not conclude formally about the effect of specific weekday.

4.2 Dependence on Month

A date histogram showing the frequency of occurrence of fatal police shootings in each month is plotted using Mathematica, as shown in Figure 9. The data from the whole years of 2015 to 2021 is included.

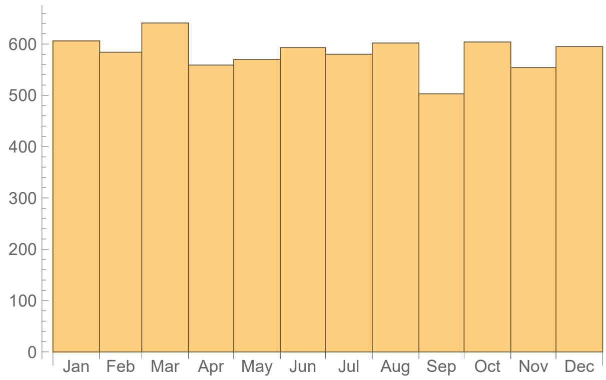


Figure 9: Frequency of occurrence of fatal police shootings in each month.

From Figure 9, we can see that the fluctuation of shootings each month fluctuates only a little bit, and thus it seems from the date histogram that the fatal police shootings is independent of month.

Again, formally, a Fisher test with the null hypothesis

H_0: The probability of shootings' occurrence on each day in each month is equal

is performed. Note that the null hypothesis is given prior to the plot of Figure 9.

The leap year and difference of number of days in every month is taken into account. The Pearson statistic is given by

$$\chi^2_{11} = \sum_{i=1}^{i=12} \frac{(O_i - E_i)^2}{E_i} = 19.55, \tag{5}$$

and thus the P-value is calculated by Mathematica as $0.052 > 0.05$. Thus, the null hypothesis is not rejected. Namely, there is no strong evidence that the average number of fatal police shootings depend on month.

5 Confidence intervals from the data of the years 2015 to 2018

5.1 Derivation of the Confidence Interval

Since X follows a Poisson distribution, which we have studied extensively,

$$f(x) = \frac{(k)^x}{x!} e^k.$$

its variance of X should be Var X = k, i.e,$\sigma^2 = k$. And the mean of the distribution is also k, according to the property of Poisson distribution.

Considering the very large sample size of our sample, it is safe to assume that $\overline{X}$, (the estimator for k), follows a normal distribution. According to the slides of VE401, a $(1-\alpha)100\%$ confidence interval on the sample mean of a normal distribution is given by

$$\overline{X} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}. \tag{6}$$

By plugging $\sigma^2 = \hat{k}$ and $\overline{X} = \hat{k}$ into the equation, we then get, the formula of $(1-\alpha)100\%$ confidence interval for k,

$$\hat{k} \pm z_{\alpha/2} \sqrt{\hat{k}/n}. \tag{7}$$

5.2 Confidence Interval from the data of the years 2015 to 2018

By using mathmetica, it is easy to calculate out that

$$z_{\alpha/2} = z_{0.025} \approx 1.96.$$

We take the value of $\overline{X}$ as $\hat{k}$, by plugging the data into the equation, we then have the table,

Year	Days	n	$\hat{k}$	95% confidence interval of $\hat{k}$
2015	365	995	2.72603	$(2.6207, 2.82588)$
2016	366	963	2.63115	$(2.51504, 2.71994)$
2017	365	987	2.70411	$(2.58508, 2.79026)$
2018	365	998	2.73425	$(2.59604, 2.80122)$
All	1461	3943	2.69884	$(2.63044‘, 2.73300)$

Table 3: 95% confidence interval from the data of the years 2015 to 2018

6 Poisson Distribution in the data of January - March 2022

We will test whether the data of January - March 2022 follows a Poisson Distribution in this part. From the data we know that there are in total 243 cases, and the total days is 31+28+31=90, then we get $\overline{X}$ =243/90=2.7. From the property of the Poisson Distribution we know that $\hat{k} = \overline{X} = 2.7$. In the observation result, the

Then we calculate the estimated value for each category,

$$E_i = nP[X = i] = \frac{ne^{-\hat{k}}\hat{k}^i}{i!}, \quad \text{for } i \in [0, 7], i \in N,$$

and

$$E_8 = 1 - \sum_{i=0}^{7} E_i$$

. The calculated result is shown as follows.

Number of fatal police shootings X (Category i)	Observed Value O_i	Expected Value E_i]
0	5	6.0485
1	17	16.3309
2	24	22.0468
3	23	19.8421
4	8	13.3934
5	5	7.23244
6	5	3.2546
7	2	1.25535
$\geqslant$8	1	0.595906

Table 4: Raw data of expected and observed data for different categories a of January - March 2022

According to the Pearson Statistic, E_i should satisfy two criteria,

$$E_i \geqslant 1 \qquad \text{for all } i = 1, \ldots, N, \tag{8}$$

$$E_i \geqslant 5 \qquad \text{for 80\% of all } i = 1, \ldots, N, \tag{9}$$

We adjust the data accordingly.

Number of fatal police shootings X (Category i)	Observed Value O_i	Expected Value E_i]
0	5	6.0485
1	17	16.3309
2	24	22.0468
3	23	19.8421
4	8	13.3934
5	5	7.23244
$\geqslant$6	8	5.105856

Table 5: Adjusted data of expected and observed data for different categories a of January - March 2022

In the latter table,

$$O_6' = O_6 + O_7 + O_8 = 8, \tag{10}$$

$$E_6' = E_6 + E_7 + E_8 = 5.105856. \tag{11}$$

H_0 : the number of fatal police shootings follows a categorical distribution with parameters E_i.

$$X^2 = \sum_{i=0}^{N-1} \frac{(O_i - E_i)^2}{E_i} \tag{12}$$

follows a chi-squared distribution.Then we get

$$X^2 = 5.38625$$

. Since there are N = 7 categories, it has $7 - 1 - m = 7 - 1 - 1 = 5$ degree of freedom, where m is the number of estimated parameters. We can reject H_0 at α level of significance if $X^2 > \chi^2_{\alpha,5}$. When $\alpha = 0.37058$,

$$\chi^2_{\alpha,8} \approx 5.38625 = X^2,$$

so the P-value of the test is 0.37058, which is much larger than 0.05. Therefore, we can conclude that the data follow a Poisson distribution.

7 Prediction Interval for Fatal Police Shootings in 2022

In this section, we will make predictions of the police shootings in 2022 based on the prediction interval for the number of observations in 2020 and 2021.

First, we derive the Nelson's formula for Poisson Distribution. Given that $\hat{\lambda} = X/n$ and $v\hat{a}r(m\hat{\lambda} - Y) = m^2\hat{\lambda}(1/n + 1/m)$, we derive that

$$\frac{(m\hat{\lambda} - Y)}{\sqrt{v\hat{a}r\left(m\hat{\lambda} - Y\right)}} = \frac{\frac{m}{n}X - Y}{\sqrt{m^2\hat{\lambda}(\frac{1}{n} + \frac{1}{m})}} = \frac{\frac{1}{n}X - \frac{1}{m}Y}{\sqrt{\frac{X}{n}(\frac{1}{n} + \frac{1}{m})}} = \frac{\hat{Y} - Y}{\sqrt{m\hat{Y}(\frac{1}{n} + \frac{1}{m})}} \sim N(0,1)$$

where $\hat{Y} = mX/n$, for $X = 1, 2...$, and is $0.5m/n$ when $X = 0$.

Therefore, the α prediction interval is

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

Based on the data for 2020 and 2021 where $X = 2074, n = 731, m = 365$, we can obtain the 95% prediction interval

$$[L, U] = 1035.58 \pm 77.23 = [959, 1112]$$

and the plot for the prediction interval is as follows:

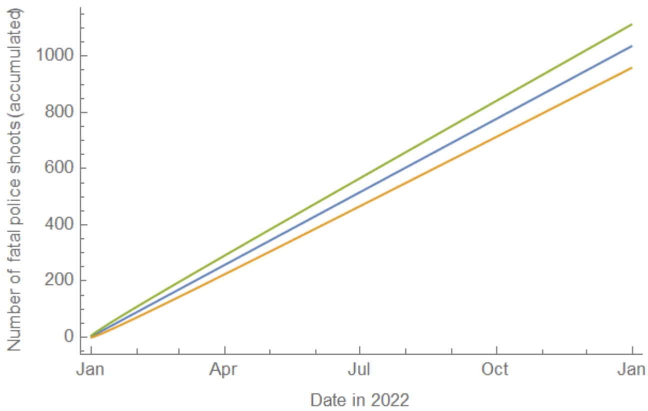


Figure 10: 95% Prediction Interval for Police Shootings in 2022

Comparing the prediction interval with the data for 2022 so far, we can clearly see that the accumulated number of shots is always within the 95% prediction interval. Hence we can say that Nelson's method for Poisson Distribution gives a quite fair prediction for this event.

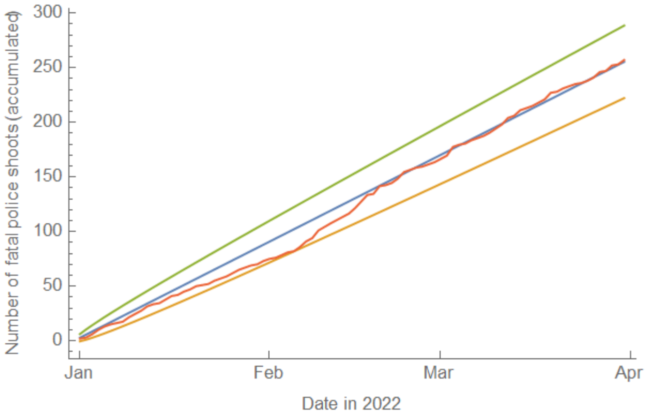
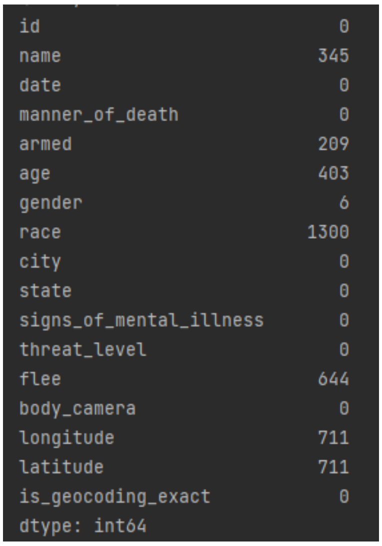


Figure 11: Predicted and Observed Police Shootings in 2022 so far.

8 Further Analyze: Regional Character

Caution: Graphs in this part are plotted by Python. All codes are attached in the Appendix part.
We first counted the number of missing information in each feature. Here is the result.



```
id                        0
name                    345
date                      0
manner_of_death           0
armed                   209
age                     403
gender                    6
race                   1300
city                      0
state                     0
signs_of_mental_illness   0
threat_level              0
flee                    644
body_camera               0
longitude               711
latitude                711
is_geocoding_exact        0
dtype: int64
```

Figure 12: the number of missing information in each feature

We can find that the features "city" and "state" have no missing information, and Fatal Police shootings show apparent regional differences.So, in order to analyse the data precisely and adequately we choose to further analyse and visualize the geographic factors.

We made linear regression about Fatal Police Shootings and the ratio of arrested people in each state. Here we will first introduce why we made this regression, and then do quantitative analysis to test this hypothesis $H_0 : \beta_1 = 0$.

8.1 Intuitive Qualitative Analysis

8.1.1 State-wise Distribution

For this part, we first plot the bar graph for number of Fatal Police Shootings vs. States. From the diagram, we found that California and Texas have relatively numerous number of Fatal Police Shootings. And Florida is not far behind.

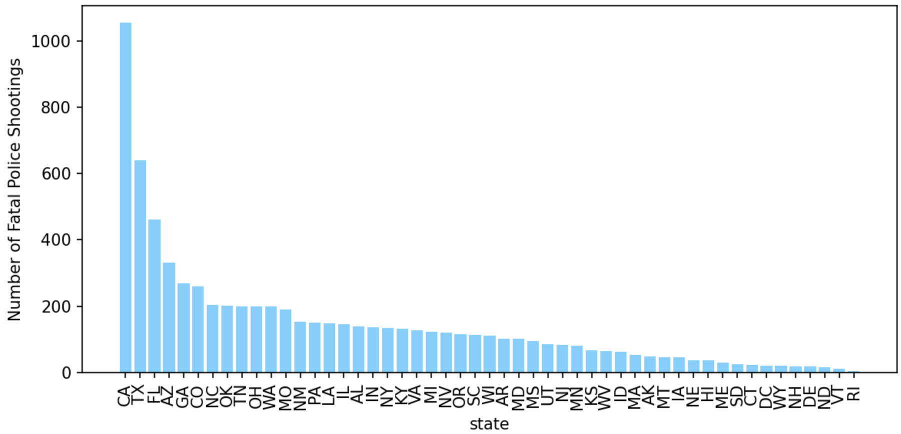


Figure 13: State Bar Graph of Fatal Police Shootings

So, we conclude that, Fatal Police Shootings have distinct regional character.

We call it State-wise, as it seems that Fatal Police Shootings mostly appear in several states (eg. California, Texas and Florida). Accordingly, we plotted one State-wise Distribution of Fatal Police Shootings, in order to visualize this feature.

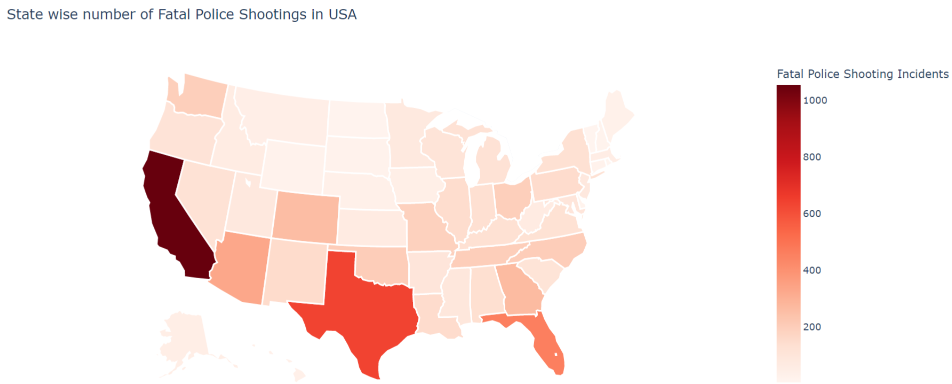


Figure 14: State-Wise Distribution of Fatal Police Shootings

8.2 Quantitative Analysis

Why does Fatal Police Shootings show such distinct regional character? We found another dataset on Kaggle, named Violent Crime Rates by US State. From this dataset, we can visualize statistics of criminals in arrests per 100,000 residents for assault, murder, and rape in each of the 50 US states. So, similarly, we visualize total quantity of these crimes for each state below.

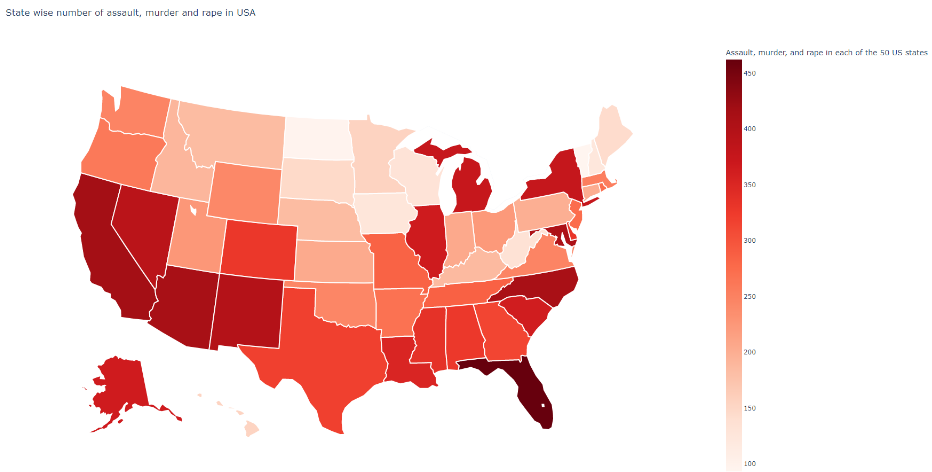


Figure 15: State-wise number of assault, murder and rape in USA

We can see some similar trends in these two graphs. For example, the color is much deeper in California and Florida. So, we made regression about Fatal Police Shootings and the ratio of arrested people in each state. Here we use the number of arrested number per 10,000 people to represent the ratios.

To test the hypothesis $H_0 : \beta_1 = 0$, we use Mathematica to analyse this two data. Corresponding Mathematica code is in the Appendix.

After making linear regression on Fatal Police Shootings and Arrested Ratios, we reach the normal expression that

$$\text{Fatal Police Shootings} = -89.2127 + 0.880871 * \text{Arrested Ratios}$$

where a 95% confidence interval for the intercept is (-216.097, 37.6712). And a 95% confidence interval for the slope is (0.431915, 1.32983).

We found that the covariance of Fatal Police Shootings and Arrested Ratio is 8520.57. The correlation coefficient of Fatal Police Shootings and Arrested Ratios is 0.494813. The parameter table is as follows.

	Estimate	Standard Error	t-Statistic	P-Value
1	-89.2127	63.1064	-1.41369	0.163907
x	0.880871	0.223291	3.94496	0.000259358

The p-value of slope is smaller than 0.05, but the p-value of intercept is larger than 0.05. As a result, this regression may not be perfectly significant. And we draw the point set as below

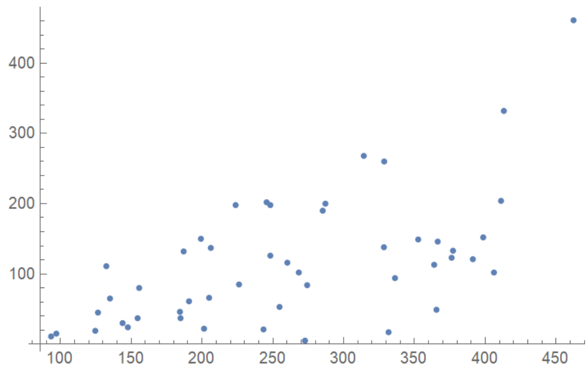


Figure 16: point set of Fatal Police shootings vs. Arrested Ratios

8.3 Discussion

8.3.1 Improvement

From the quantitative analysis and parameter table of the Linear Regression about Fatal Police Shootings and Arrested Ratio, we can reach a conclusion that, the dependence is not strong enough.

To some content, we can prove the regression that Fatal Police Shootings have strong linear dependence on the ratio of arrested people in each state is significant.

However, we realize that there are indeed some flaws. For example, we use the exact number of Fatal Police Shootings, which means that we ignore the population difference between each state.

As a result, we made a trial to minimize this effect of population difference between each state by dividing the population in each state in 2020.

Similarly, we calculated the number of Fatal Police Shootings per 10,000 people in each state, which we will call Fatal Police Shootings ratio hereinafter. Here is the modified distribution map, namely the map for state-wise Fatal Police Shootings ratio in USA.

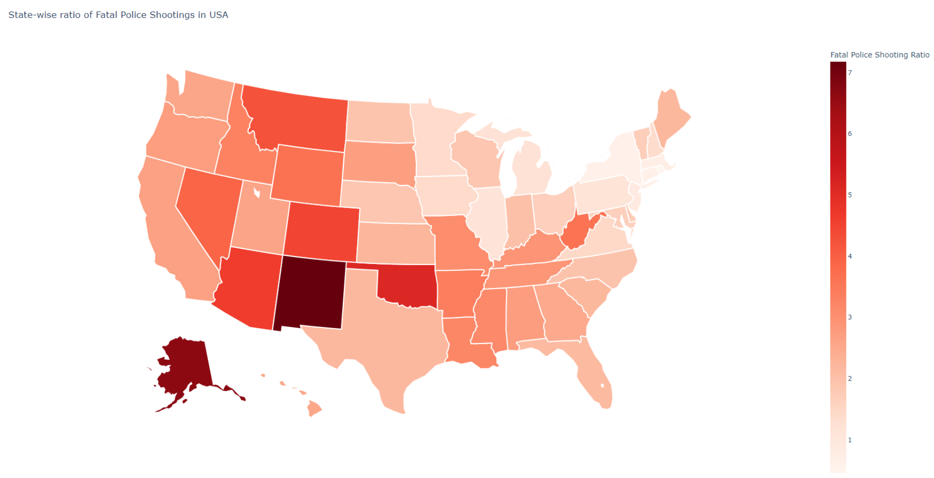


Figure 17: State-wise ratio of assault, murder and rape in USA

And we also made quantitative analysis on the Fatal Police Shooting ratio and Arrested People ratio. But we still cannot reject H_0, which means that we still cannot prove the regression is significant. The code is also in the appendix.

	Estimate	Standard Error	t-Statistic	P-Value
1	1.62825	0.566868	2.87236	0.00604924
x	0.00340141	0.00200576	1.69582	0.0963971

For this regression, P-value for slope is larger than 0.05. So, we cannot reject our hypothesis that $H_0 : \beta_1 = 0$. so we cannot prove the regression is significant.

8.3.2 Another Discovery: Precise Precise Coordinates Analysis

Then we noticed that, besides feature "state", we have a more precise feature for individual Fatal Police Shooting, the feature "longitude" and "latitude". So, we plot one heat-map for all individual Fatal Police Shooting in USA. Here is the result.

Caution: We first abandoned the inaccurate coordinate information according to feature "is_geocoding_exact".

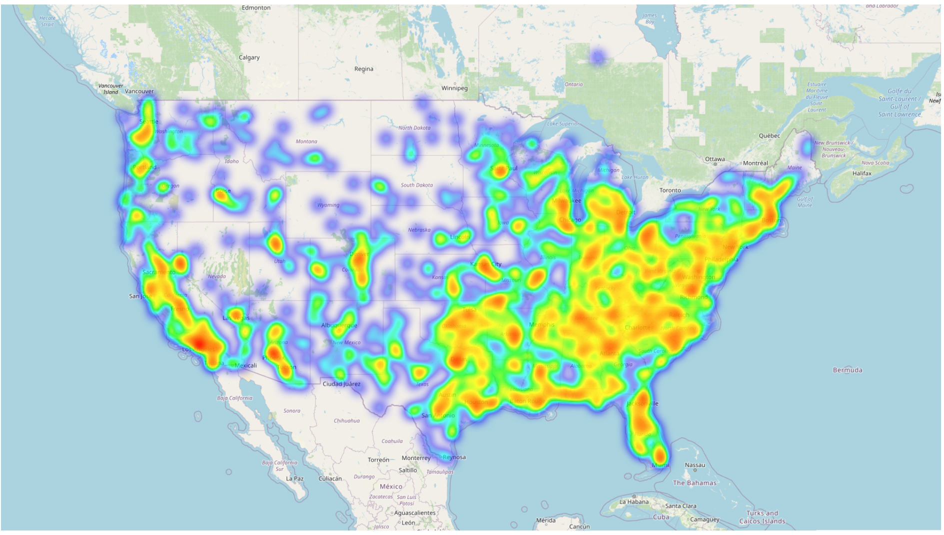


Figure 18: Heat-map of Fatal Police Shootings in USA

From the heat-map above, we quickly realized that out hypothesis and analyze are limited by the boundaries of states.

Regardless of limits of states, we can find that, there seems to be one north-south boundary in the middle part of USA. More precisely, eastern United States have much more Fatal Police Shootings than western United States, although there are some peak values in the southwest part of western United States.

The extremely huge values in certain states (eg. California, Texas and Florida) is too large, which made the gradient color of other states not apparent and distinct. So, the boundary of states disturbed our estimation on the average value of eastern United and the general distribution of Fatal Police Shootings.

These are still Intuitive Qualitative Analysis. And if we want to analyze the geographic factors without the concept of states, we will need much more information and features. If we have more chance to discuss the geographic factors with more resources, we will go deeper in this direction of study in the future.

9 Conclusion

We have tested whether the fatal police shootings in the United States follow a Poisson distribution. Separate tests on the whole data span (between January, 1, 2015 and April, 5, 2022) and from the year 2020 is performed separately. Based on our tests, there is no evidence that the shootings do not happen at constant rate or the COVID-19 pandemic has affected the rate significantly.

Apart from the test on the overall distribution, we have also looked into the effect of different weekday or month. We have tested on whether the average rate of occurrence depends on weekday on month, and there is evidence that it depends on weekday but not on month.

A goodness-of-fit test for Poisson distribution is performed based on data span of January - March 2022. Based on the tests, we conclude that there is no sufficient evidence to reject the hypothesis that the shootings follow a Poisson distribution.

By assuming that the police shootings follow the Poisson Distribution, we apply the Nelson's formula and achieve a 95% prediction interval for data in 2022. Comparing with the data so far, the observation result conforms with the prediction, suggesting this method may be quite useful in predicting such cases.

For further analysis, we assume that Fatal Police Shootings have strong linear dependence on the ratio of arrested people in each state. We made linear regression about Fatal Police Shootings and the ratio of arrested people in each state.And we test the hypothesis based on quantitative analysis by Mathematica. But we cannot reject $H_0 : \beta_1 = 0$, which means we cannot prove the regression is significant. Then we tried to improve the origin dataset, but we still cannot reject H_0, which proves that our regression is still not significant.

References

[1] “Fatal Force: Police Shootings Database.” The Washington Post, WP Company, 22 Jan. 2020, https://www.washingtonpost.com/graphics/investigations/police-shootings-database/.

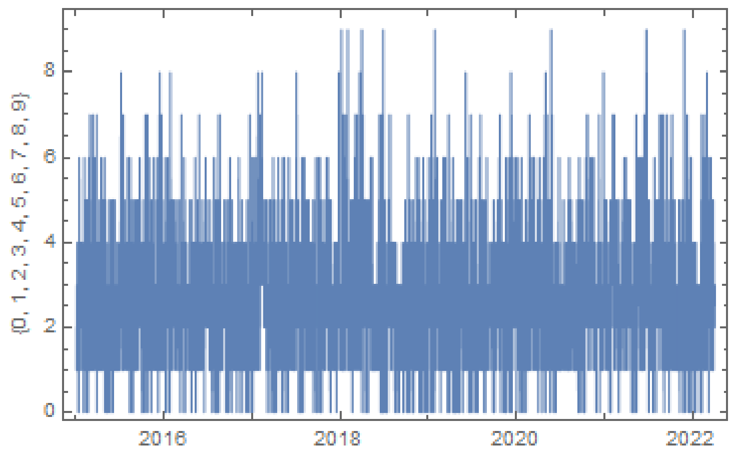
[2] Olzak, Susan. ”Does Protest Against Police Violence Matter? Evidence from US Cities, 1990 through 2019.” American Sociological Review 86.6 (2021): 1066-1099.

A Appendix

A.1 List plot of all data

```
In[5]:= data := Import["E:\\22SP\\VE401\\project\\Data.xlsx", {"Data", 1, Range[1, 2651]}]
                |导入                                                          |范围

In[6]:= DateListPlot[data, FrameLabel → {Automatic, Range[0, 9]},
        |日期列表图           |边框标签       |自动          |范围
         FrameLabel → {"Date", "Number of fatal police shootings"}]
         |边框标签
```

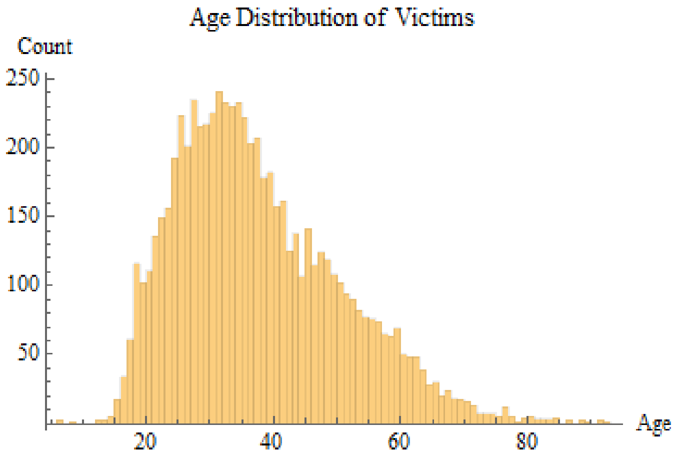
Out[6]=



A.2 Histogram of victims' age

```
In[10]:= Age := Import["E:\\22SP\\VE401\\project\\age.xlsx", {"Data", 1, Range[1, 6843], 1}]
                导入                                                     范围

In[14]:= Histogram[Age, {"Raw", "FreedmanDiaconis"}, AxesLabel → {HoldForm[Age], HoldForm[Count]},
         直方图                                          坐标轴标签      保持表达式       保持…       计数
          PlotLabel → HoldForm[Age Distribution of Victims],
          绘图标签       保持表达式
          LabelStyle → {FontFamily → "Times New Roman", 12, GrayLevel[0]}]
          标签样式         字体系列                                灰度级
```

Out[14]=



A.3 Code for Goodness-of-fit for Poisson Distribution

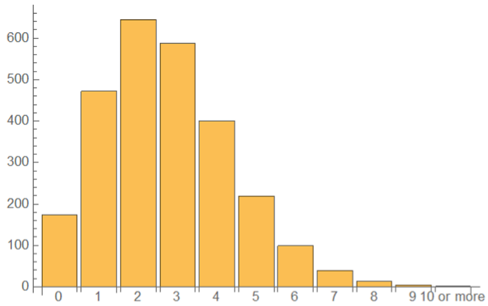
```
k = 2.73;
x = Table[i, {i, 0, 9}]
    表格
fx = (k^x * Exp[-k]) / x!;
         指数形式
AppendTo[fx, 1 - Total[fx]]
附加              总计
expect = fx * 2652
```

```
Out[•]= {0.0652193, 0.178049, 0.243036, 0.221163, 0.150944, 0.0824153, 0.037499, 0.0146246, 0.00499065, 0.00151383, 0.000545242}

Out[•]= {172.962, 472.185, 644.533, 586.525, 400.303, 218.565, 99.4473, 38.7844, 13.2352, 4.01467, 1.44598}

In[•]:= expect = {173, 472, 645, 587, 400, 219, 99, 39, 13, 4, 1};

In[•]:= observedData := {181, 503, 615, 568, 390, 221, 103, 50, 13, 8, 0};
```

$$X = N\left[\text{Total}\left[\frac{(\text{observedData} - \text{expect})^{\wedge}2}{\text{expect}}\right]\right]$$

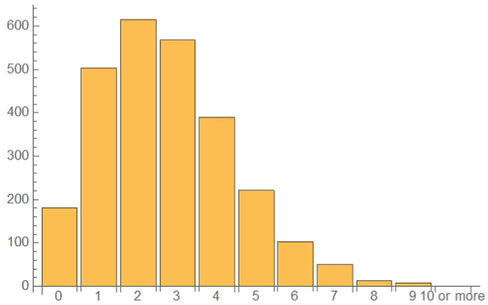
```
Out[•]= 12.9487

In[•]:= 1 - CDF[ChiSquareDistribution[9], X]
             卡方分布
Out[•]= 0.164939

In[•]:= BarChart[expect, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10 or more"}]
        条形图                 图表标签
```

Out[•]=




```
In[•]:= BarChart[observedData, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10 or more"}]
        条形图                      图表标签
```

Out[•]=



In[9]:=
```
k = 2.73;
x = Table[i, {i, 0, 8}];
    表格
fx = (k^x * Exp[-k]) / x!;
        指数形式
AppendTo[fx, 1 - Total[fx]]
附加               总计
expect = fx * 826
```

Out[12]= {0.0652193, 0.178049, 0.243036, 0.221163, 0.150944, 0.0824153, 0.037499, 0.0146246, 0.00499065, 0.00205907}

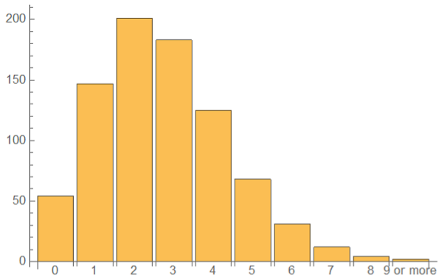
Out[13]= {53.8711, 147.068, 200.748, 182.681, 124.68, 68.0751, 30.9742, 12.0799, 4.12227, 1.70079}

In[3]:=
```
expect = {54, 147, 201, 183, 125, 68, 31, 12, 4, 2};
observedData = {40, 156, 198, 185, 112, 71, 37, 21, 3, 3};
X = N[Total[(observedData - expect)^2 / expect]]
    ⋯ 总计
1 - CDF[ChiSquareDistribution[8], X]
    ⋯   卡方分布
BarChart[expect, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9 or more"}]
条形图                 图表标签
BarChart[observedData, ChartLabels → {"0", "1", "2", "3", "4", "5", "6", "7", "8", "9 or more"}]
条形图                       图表标签
```

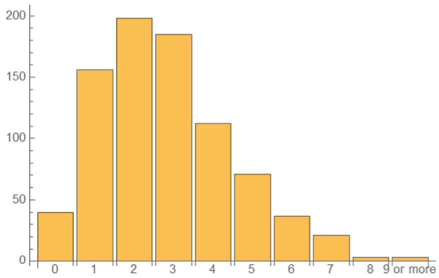
Out[5]= 14.3929

Out[6]= 0.0720815

Out[7]=



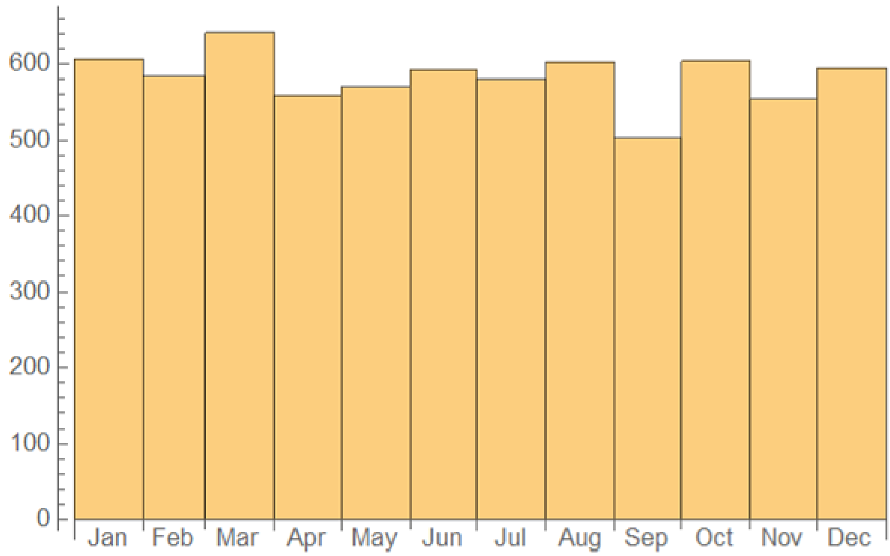
Out[8]=



A.4 Code for Bar Chart for Frequency of occurrence of fatal police shootings in each month and on each weekday

```
In[∘]:= date =
          Import[
          导入
           "D:\\Documents\\VE401\\Project\\fatal-police-shootings-data-
             dateform.csv", {"Data", 1 ;; 6991, 3}];

In[∘]:= DateHistogram[date, "Month", DateReduction → "Year",
        日期直方图                    日期约化
          DateFunction → (DateList[{#, {"Year", "Month", "Day"}}] &)]
          日期函数          日期列表
```

Out[∘]=




```
data=table2array(fatalpoliceshootingsdatadateform);
a=weekday(data);
Y=groupcounts(a);
Y=[Y(2:7);Y(1)];
X = categorical({'Mon','Tue','Wed','Thu','Fri','Sat','Sun'});
X = reordercats(X,{'Mon','Tue','Wed','Thu','Fri','Sat','Sun'});
bar(X,Y);
xlabel('Week Day');
ylabel('Number of Homicides 15-22');
```

A.5 Code for State Wise Distribution Graph

```
import pandas as pd
import plotly
from plotly.graph_objs import *
from plotly.offline import init_notebook_mode, iplot,plot
df = pd.read_csv('fatal-police-shootings-data.csv', parse_dates=['date'])
state_dict = {'AK': 'Alaska', 'AL': 'Alabama', 'AR': 'Arkansas', 'AZ': 'Arizona', 'CA': '
                                                California', 'CO': 'Colorado',
              'CT':'Connecticut', 'DC': 'District of Columbia', 'DE': 'Delaware', 'FL': '
                                                              Florida', 'GA': 'Georgia',
              'HI': 'Hawaii', 'IA': 'Iowa', 'ID': 'Idaho', 'IL': 'Illinois', 'IN': 'Indiana',
                                                              'KS': 'Kansas',
              'KY': 'Kentucky', 'LA': 'Louisiana', 'MA': 'Massachusetts', 'MD': 'Maryland', '
                                                              ME': 'Maine',
              'MI': 'Michigan', 'MN': 'Minnesota', 'MO': 'Missouri', 'MS': 'Mississippi', 'MT'
                                                              : 'Montana',
              'NC': 'North Carolina', 'ND': 'North Dakota', 'NE': 'Nebraska', 'NH': 'New
                                                              Hampshire', 'NJ': 'New Jersey',
              'NM': 'New Mexico', 'NV': 'Nevada', 'NY': 'New York', 'OH': 'Ohio', 'OK': '
                                                              Oklahoma', 'OR': 'Oregon',
              'PA': 'Pennsylvania', 'RI': 'Rhode Island', 'SC': 'South Carolina', 'SD': 'South
                                                               Dakota',
              'TN': 'Tennessee', 'TX': 'Texas', 'UT': 'Utah', 'VA': 'Virginia', 'VT': 'Vermont
                                                              ', 'WA': 'Washington',
              'WI': 'Wisconsin', 'WV': 'West Virginia', 'WY': 'Wyoming'}
states_counter = df['state'].value_counts()
states_data = pd.DataFrame()
states_data['state'] = states_counter.index
states_data['counts'] = states_counter.values
states_data['state_code'] = states_data['state'].apply(lambda x: state_dict[x])
data = dict(
        type='choropleth',
        locations=states_data['state'],
        z=states_data['counts'],
        autocolorscale=False,
        locationmode='USA-states',
        text=states_data['state_code'],
        colorscale='Reds',
        marker=dict(line=dict(color='rgb(255, 255, 255)', width=2)),
        colorbar=dict(
            title="Fatal Police Shooting Incidents")
        )
layout = dict(
        title='State wise number of Fatal Police Shootings in USA',
        geo=dict(
            scope='usa',
            countrycolor='rgb(255, 255, 255)',
            projection=dict(type='albers usa'),
            showlakes=True,
            lakecolor='rgb(255, 255, 255)'),
        )
figure = dict(data=data, layout=layout)
iplot(figure, filename='Fatal Police Shootings in USA')
```

A.6 Code for Confidence intervals from the data of the years 2015 to 2018

In[◦]:=
```
data = Import["D:\\fatal-police-shootings-data.csv"];
        导入      偏导
```

In[◦]:=
```
allDates = data[[2 ;;, 3]];
days = {365, 366, 365, 365};
dates2015 = Length[Select[allDates, StringStartsQ[#, "2015"] &]];
            长度    选择          字符串始端匹配判定
dates2016 = Length[Select[allDates, StringStartsQ[#, "2016"] &]];
            长度    选择          字符串始端匹配判定
dates2017 = Length[Select[allDates, StringStartsQ[#, "2017"] &]];
            长度    选择          字符串始端匹配判定
dates2018 = Length[Select[allDates, StringStartsQ[#, "2018"] &]];
            长度    选择          字符串始端匹配判定
dates = dates2015 + dates2016 + dates2017 + dates2018;
date = {dates2015, dates2016, dates2017, dates2018}
z = InverseCDF[NormalDistribution[], 0.975]
    逆累积分…   正态分布
```

Out[◦]= {994, 958, 981, 985}

Out[◦]= 1.95996

In[◦]:=
```
k = dates / 4 / Mean[days];
                平均值
{dates, 4 * Mean[days], k}
            平均值
interval = {k - z √(k / dates), k + z √(k / dates)}
```

Out[◦]= $\left\{3918, 1461, \frac{1306}{487}\right\}$

```
{2.6304478017753725`, 2.733001890216414`}
i = Range[1, 4, 1]
    范围
```

In[◦]:=
```
k = date[[i]] / days[[i]]

interval = Table[{k[[i]] - z √(k[[i]] / Length[splitDates[[i]]]),
           表格
   k[[i]] + z √(k[[i]] / Length[splitDates[[i]]])}, {i, 1, 4}]
```

Out[◦]= $\left\{\frac{994}{365}, \frac{479}{183}, \frac{981}{365}, \frac{197}{73}\right\}$

Out[◦]= {{2.6207, 2.82588}, {2.51504, 2.71994}, {2.58508, 2.79026}, {2.59604, 2.80122}}

A.7 Code for Poisson Distribution in the data of January - March 2022

```
In[•]:= data = Import["D:\\fatal-police-shootings-data.csv"];
              导入      偏导
        allDates = data[[2 ;;, 3]];
        dates = Select[allDates, StringStartsQ[#, {"2022-01", "2022-02", "2022-03"}] &];
                选择             字符串始端匹配判定
        n = Length[dates];
            长度
        days = 31 + 28 + 31;
        parameter = n / days;
        i = Range[0, 7, 1];
            范围
        ExpectationResult = days * PDF[PoissonDistribution[2.7], i];
                                   …   泊松分布
        E8 = 90 - 8 * Mean[ExpectationResult];
                      平均值
        ExpectationResult = Join[ExpectationResult, {E8}]
                            连接
        Observation = KeySort[Counts[Values[Counts[dates]]]];
                      键排序   关联计数 获取值   关联计数
        ObservationResult = Join[{90 - 8 * Mean[Values[Observation]]}, Values[Observation]]
                            连接          …    获取值                  获取值

Out[•]= {6.0485, 16.3309, 22.0468, 19.8421, 13.3934, 7.23244, 3.2546, 1.25535, 0.595906}

Out[•]= {5, 17, 24, 23, 8, 5, 5, 2, 1}
```

$$X = \sum_{j=1}^{9} ((\text{ExpectationResult}[[j]] - \text{ObservationResult}[[j]])^2 / \text{ExpectationResult}[[j]])$$

```
Out[•]= 5.39755

In[•]:=
        P = 1 - CDF[ChiSquareDistribution[6], X]
                …   卡方分布
Out[•]= 0.493925

In[•]:= ExpectationResult = {6.048496146577478`, 16.330939595759194`, 22.046768454274908`,
          19.842091608847422`, 13.393411835972014`, 7.232442391424884`, 5.105856}
        ObservationResult = {5, 17, 24, 23, 8, 5, 8}

Out[•]= {6.0485, 16.3309, 22.0468, 19.8421, 13.3934, 7.23244, 5.10586}
```

$$X = \sum_{j=1}^{7} ((\text{ExpectationResult}[[j]] - \text{ObservationResult}[[j]])^2 / \text{ExpectationResult}[[j]])$$

```
In[•]:= 5.386252257174421`
        P = 1 - CDF[ChiSquareDistribution[5], X]
                …   卡方分布
Out[•]= 5.38625
```

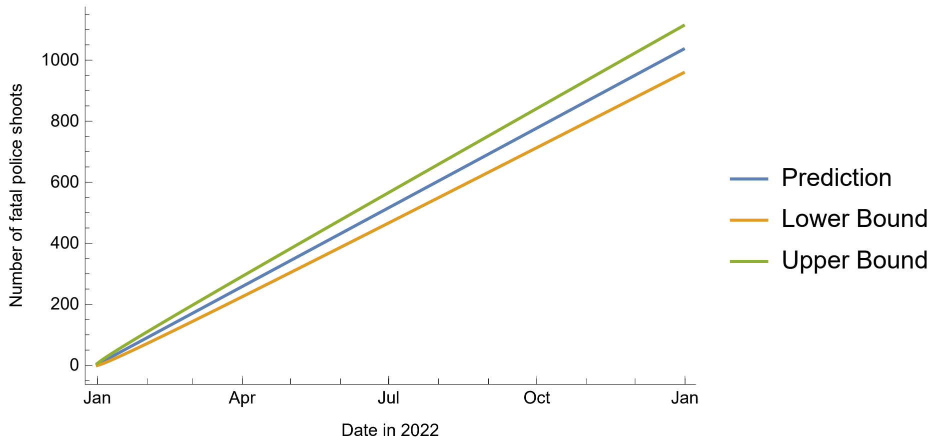
Out[ ]= 0.37058

A.8 Code for the prediction interval of fatal police shootings in 2022

In[41]:=

```
X = 2074;
n = 731;
zα/2 = InverseCDF[NormalDistribution[], 0.975];
        逆累积分…      正态分布
Y = Function[m, m * N[X / n]];
    纯函数              数值运算
lower = Function[m, Y[m] - zα/2 √(m * Y[m] (1/m + 1/n))];
        纯函数
upper = Function[m, Y[m] + zα/2 √(m * Y[m] (1/m + 1/n))];
        纯函数
Prediction = TimeSeries[Table[Y[m], {m, 1, 365}], {"Jan 1,2022"}];
             时间序列     表格
LowerBound = TimeSeries[Table[lower[m], {m, 1, 365}], {"Jan 1,2022"}];
             时间序列     表格
Upperbound = TimeSeries[Table[upper[m], {m, 1, 365}], {"Jan 1,2022"}];
             时间序列     表格
fig = DateListPlot[
      日期列表图
   {Prediction, LowerBound, Upperbound},
   PlotLegends → {"Prediction", "Lower Bound", "Upper Bound"},
   绘图的图例
   Frame → {True, True, False, False},
   边框     真     真     假     假
   FrameLabel → {"Date in 2022", "Number of fatal police shoots"}
   边框标签                        数
  ]
```

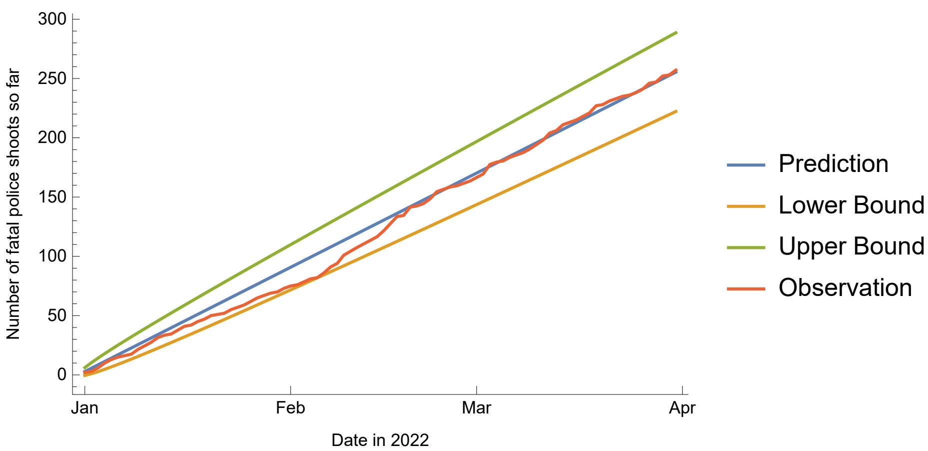
Out[50]=



In[51]:=

```
data = Import["D:\\Desktop_Files\\401\\project\\fatal-police-shootings-data.csv"];
        导入       偏导
allDates = data[[2 ;;, 3]];
dates = Select[allDates, StringStartsQ[#, {"2022-01", "2022-02", "2022-03"}] &];
         选择               字符串始端匹配判定
Observation = Accumulate[TimeSeries[Tally[dates]]];
               累加        时间序列     重复次数
X = 2074;
n = 731;
zα/2 = InverseCDF[NormalDistribution[], 0.975];
        逆累积分…    正态分布
Y = Function[m, m * N[X / n]];
     纯函数          数值运算
lower = Function[m, Y[m] - zα/2 Sqrt[m * Y[m] (1/m + 1/n)]];
         纯函数
upper = Function[m, Y[m] + zα/2 Sqrt[m * Y[m] (1/m + 1/n)]];
         纯函数
Prediction = TimeSeries[Table[Y[m], {m, 1, 90}], {"Jan 1,2022"}];
              时间序列    表格
LowerBound = TimeSeries[Table[lower[m], {m, 1, 90}], {"Jan 1,2022"}];
              时间序列    表格
Upperbound = TimeSeries[Table[upper[m], {m, 1, 90}], {"Jan 1,2022"}];
              时间序列    表格
fig = DateListPlot[
       日期列表图
  {Prediction, LowerBound, Upperbound, Observation},
  PlotLegends → {"Prediction", "Lower Bound", "Upper Bound", "Observation"},
  绘图的图例
  Frame → {True, True, False, False},
  边框      真     真     假     假
  FrameLabel → {"Date in 2022", "Number of fatal police shoots so far"}
  边框标签                              数
 ]
```

Out[64]=



A.9 Code for the linear regression of Fatal Police Shootings and Arrested ratio

```
dataset = Import["D:\\22SP\\401\\Project\\crime and FPS2.csv"]
              导入       编导
head = dataset[[1]]
```

Out[26]= {{total_crime, FPS_counts}, {416.6, 1055}, {319.2, 640}, {462.3, 461}, {413.1, 332}, {314.2, 268}, {328.6, 260}, {411.1, 204}, {245.6, 202}, {287.1, 200}, {223.7, 198}, {248.2, 198}, {285.2, 190}, {398.5, 152}, {199.2, 150}, {352.6, 149}, {366.4, 146}, {328.4, 138}, {206.2, 137}, {377.2, 133}, {187., 132}, {248.2, 126}, {376.2, 123}, {391.2, 121}, {260.2, 116}, {363.9, 113}, {132.4, 111}, {268.3, 102}, {406.1, 102}, {336.2, 94}, {226.1, 85}, {274.2, 84}, {155.6, 80}, {205., 66}, {135., 65}, {190.8, 61}, {254.7, 53}, {365.5, 49}, {184.4, 46}, {126.5, 45}, {184.8, 37}, {154.5, 37}, {143.9, 30}, {147.6, 24}, {201.4, 22}, {243.4, 21}, {124.6, 19}, {331.7, 17}, {97.1, 15}, {93.4, 11}, {272.7, 5}}

Out[27]= {total_crime, FPS_counts}

```
In[28]:= crime = Table[{dataset[[i]][[1]]}, {i, 2, Length[dataset]}]
                 表格                               长度
```

Out[28]= {{416.6}, {319.2}, {462.3}, {413.1}, {314.2}, {328.6}, {411.1}, {245.6}, {287.1}, {223.7}, {248.2}, {285.2}, {398.5}, {199.2}, {352.6}, {366.4}, {328.4}, {206.2}, {377.2}, {187.}, {248.2}, {376.2}, {391.2}, {260.2}, {363.9}, {132.4}, {268.3}, {406.1}, {336.2}, {226.1}, {274.2}, {155.6}, {205.}, {135.}, {190.8}, {254.7}, {365.5}, {184.4}, {126.5}, {184.8}, {154.5}, {143.9}, {147.6}, {201.4}, {243.4}, {124.6}, {331.7}, {97.1}, {93.4}, {272.7}}

```
In[29]:= FPS = Table[{dataset[[i]][[2]]}, {i, 2, Length[dataset]}]
               表格                               长度
```

Out[29]= {{1055}, {640}, {461}, {332}, {268}, {260}, {204}, {202}, {200}, {198}, {198}, {190}, {152}, {150}, {149}, {146}, {138}, {137}, {133}, {132}, {126}, {123}, {121}, {116}, {113}, {111}, {102}, {102}, {94}, {85}, {84}, {80}, {66}, {65}, {61}, {53}, {49}, {46}, {45}, {37}, {37}, {30}, {24}, {22}, {21}, {19}, {17}, {15}, {11}, {5}}

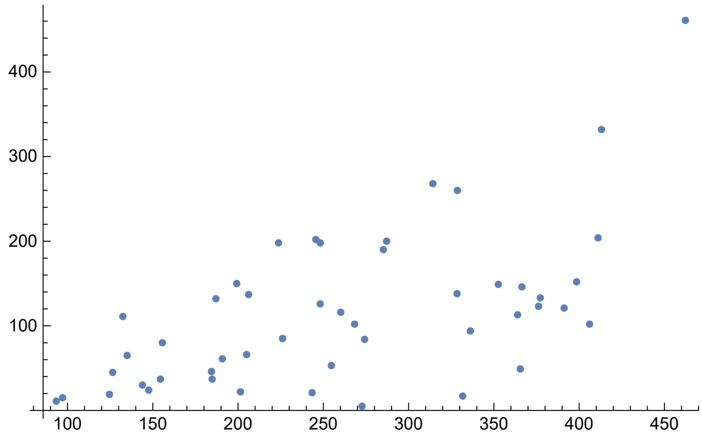
```
In[30]:= N[Covariance[crime, FPS]]
         [··  协方差
```

Out[30]= {{8520.57}}

```
In[31]:= N[Correlation[crime, FPS]]
         [··  相关
```

Out[31]= {{0.494813}}

In[33]:=
```
ListPlot[dataset]
```
绘制点集

Out[33]=


In[36]:=
```
modifyset = dataset[[2 ;;]]
LModel = LinearModelFit[modifyset, x, x]
```
线性拟合模型

Out[36]=
```
{{416.6, 1055}, {319.2, 640}, {462.3, 461}, {413.1, 332}, {314.2, 268}, {328.6, 260},
 {411.1, 204}, {245.6, 202}, {287.1, 200}, {223.7, 198}, {248.2, 198}, {285.2, 190},
 {398.5, 152}, {199.2, 150}, {352.6, 149}, {366.4, 146}, {328.4, 138}, {206.2, 137},
 {377.2, 133}, {187., 132}, {248.2, 126}, {376.2, 123}, {391.2, 121}, {260.2, 116},
 {363.9, 113}, {132.4, 111}, {268.3, 102}, {406.1, 102}, {336.2, 94}, {226.1, 85},
 {274.2, 84}, {155.6, 80}, {205., 66}, {135., 65}, {190.8, 61}, {254.7, 53}, {365.5, 49},
 {184.4, 46}, {126.5, 45}, {184.8, 37}, {154.5, 37}, {143.9, 30}, {147.6, 24},
 {201.4, 22}, {243.4, 21}, {124.6, 19}, {331.7, 17}, {97.1, 15}, {93.4, 11}, {272.7, 5}}
```

Out[37]= FittedModel[$-89.2127 + 0.880871\,x$]

In[38]:=
```
Normal[LModel]
```
转换为普通表达式

Out[38]= $-89.2127 + 0.880871\,x$

In[39]:=
```
LModel["ParameterConfidenceIntervals"]
```

Out[39]= {{-216.097, 37.6712}, {0.431915, 1.32983}}

In[40]:=
```
LModel["ParameterTable"]
```

Out[40]=

	Estimate	Standard Error	t-Statistic	P-Value
1	-89.2127	63.1064	-1.41369	0.163907
x	0.880871	0.223291	3.94496	0.000259358

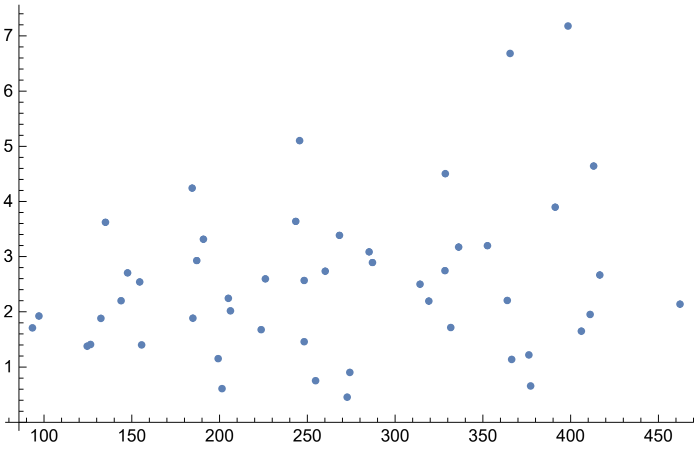
A.10 Code for the linear regression of Fatal Police Shooting ratio and Arrested ratio

```
In[44]:= dataset = Import["D:\\21SP\\401\\Project\\modify.csv"]
                   导入      偏导
         head = dataset[[1]]

Out[44]= {{total_crime, ratio}, {416.6, 2.6683}, {319.2, 2.19588}, {462.3, 2.14038},
  {413.1, 4.64238}, {314.2, 2.50189}, {328.6, 4.50317}, {411.1, 1.95414},
  {245.6, 5.10184}, {287.1, 2.894}, {223.7, 1.67804}, {248.2, 2.56967}, {285.2, 3.08696},
  {398.5, 7.1782}, {199.2, 1.15361}, {352.6, 3.19896}, {366.4, 1.13951},
  {328.4, 2.74666}, {206.2, 2.019}, {377.2, 0.658375}, {187, 2.92953}, {248.2, 1.45979},
  {376.2, 1.22056}, {391.2, 3.89742}, {260.2, 2.73762}, {363.9, 2.20771},
  {132.4, 1.88336}, {268.3, 3.38699}, {406.1, 1.65123}, {336.2, 3.1743},
  {226.1, 2.5981}, {274.2, 0.904296}, {155.6, 1.40191}, {205, 2.24652}, {135, 3.62376},
  {190.8, 3.31683}, {254.7, 0.753921}, {365.5, 6.68129}, {184.4, 4.24266},
  {126.5, 1.4105}, {184.8, 1.88631}, {154.5, 2.54248}, {143.9, 2.20206},
  {147.6, 2.70677}, {201.4, 0.610104}, {243.4, 3.64045}, {124.6, 1.37928},
  {331.7, 1.71726}, {97.1, 1.92531}, {93.4, 1.71053}, {272.7, 0.455631}}

Out[45]= {total_crime, ratio}

In[46]:= crime = Table[{dataset[[i]][[1]]}, {i, 2, Length[dataset]}]
                 表格                                长度

Out[46]= {{416.6}, {319.2}, {462.3}, {413.1}, {314.2}, {328.6}, {411.1}, {245.6},
  {287.1}, {223.7}, {248.2}, {285.2}, {398.5}, {199.2}, {352.6}, {366.4},
  {328.4}, {206.2}, {377.2}, {187}, {248.2}, {376.2}, {391.2}, {260.2}, {363.9},
  {132.4}, {268.3}, {406.1}, {336.2}, {226.1}, {274.2}, {155.6}, {205}, {135},
  {190.8}, {254.7}, {365.5}, {184.4}, {126.5}, {184.8}, {154.5}, {143.9},
  {147.6}, {201.4}, {243.4}, {124.6}, {331.7}, {97.1}, {93.4}, {272.7}}

In[47]:= FPS = Table[{dataset[[i]][[2]]}, {i, 2, Length[dataset]}]
               表格                                长度

Out[47]= {{2.6683}, {2.19588}, {2.14038}, {4.64238}, {2.50189}, {4.50317}, {1.95414}, {5.10184},
  {2.894}, {1.67804}, {2.56967}, {3.08696}, {7.1782}, {1.15361}, {3.19896},
  {1.13951}, {2.74666}, {2.019}, {0.658375}, {2.92953}, {1.45979}, {1.22056},
  {3.89742}, {2.73762}, {2.20771}, {1.88336}, {3.38699}, {1.65123}, {3.1743},
  {2.5981}, {0.904296}, {1.40191}, {2.24652}, {3.62376}, {3.31683}, {0.753921},
  {6.68129}, {4.24266}, {1.4105}, {1.88631}, {2.54248}, {2.20206}, {2.70677},
  {0.610104}, {3.64045}, {1.37928}, {1.71726}, {1.92531}, {1.71053}, {0.455631}}

In[48]:= N[Covariance[crime, FPS]]
         ·· 协方差

Out[48]= {{32.9015}}

In[49]:= N[Correlation[crime, FPS]]
         ·· 相关

Out[49]= {{0.237752}}
```

In[50]:= `ListPlot[dataset]`

绘制点集

Out[50]=



In[51]:=
```
modifyset = dataset[[2 ;;]]
LModel = LinearModelFit[modifyset, x, x]
```

线性拟合模型

Out[51]=
```
{{416.6, 2.6683}, {319.2, 2.19588}, {462.3, 2.14038}, {413.1, 4.64238}, {314.2, 2.50189},
 {328.6, 4.50317}, {411.1, 1.95414}, {245.6, 5.10184}, {287.1, 2.894},
 {223.7, 1.67804}, {248.2, 2.56967}, {285.2, 3.08696}, {398.5, 7.1782},
 {199.2, 1.15361}, {352.6, 3.19896}, {366.4, 1.13951}, {328.4, 2.74666},
 {206.2, 2.019}, {377.2, 0.658375}, {187, 2.92953}, {248.2, 1.45979},
 {376.2, 1.22056}, {391.2, 3.89742}, {260.2, 2.73762}, {363.9, 2.20771},
 {132.4, 1.88336}, {268.3, 3.38699}, {406.1, 1.65123}, {336.2, 3.1743},
 {226.1, 2.5981}, {274.2, 0.904296}, {155.6, 1.40191}, {205, 2.24652}, {135, 3.62376},
 {190.8, 3.31683}, {254.7, 0.753921}, {365.5, 6.68129}, {184.4, 4.24266},
 {126.5, 1.4105}, {184.8, 1.88631}, {154.5, 2.54248}, {143.9, 2.20206},
 {147.6, 2.70677}, {201.4, 0.610104}, {243.4, 3.64045}, {124.6, 1.37928},
 {331.7, 1.71726}, {97.1, 1.92531}, {93.4, 1.71053}, {272.7, 0.455631}}
```

Out[52]= `FittedModel[ 1.62825 + 0.00340141 x ]`

In[53]:= `Normal[LModel]`

转换为普通表达式

Out[53]= `1.62825 + 0.00340141 x`

In[54]:= `LModel["ParameterConfidenceIntervals"]`

Out[54]= `{{0.488483, 2.76801}, {-0.000631435, 0.00743426}}`

In[55]:= `LModel["ParameterTable"]`

Out[55]=

	Estimate	Standard Error	t-Statistic	P-Value
1	1.62825	0.566868	2.87236	0.00604924
x	0.00340141	0.00200576	1.69582	0.0963971

A.11 Code for Heat-map

```
import folium
import webbrowser
import pandas as pd
import numpy as np
from folium.plugins import HeatMap
df = pd.read_csv('fatal-police-shootings-data.csv', parse_dates=['date'])
df=df.dropna(subset=['latitude', 'longitude'])
LAT_new = df['latitude'].values
LNG_new = df['longitude'].values

is_geocoding_exact = df['is_geocoding_exact'].values

accuracy = np.zeros(len(df['is_geocoding_exact']))

for i in range(len(df['is_geocoding_exact'])):
    if is_geocoding_exact[i]:
        accuracy[i] = 1
print(accuracy.shape, np.sum(accuracy))
data1 = [[LAT_new[i], LNG_new[i], accuracy[i]] for i in range(len(df))]
Center = [np.mean(np.array(LAT_new, dtype='float32')), np.mean(np.array(LNG_new, dtype='
                                                float32'))]
m = folium.Map(location=Center, zoom_start=6)
HeatMap(data1, radius=15).add_to(m)
m.save('states_heatmap.html')
webbrowser.open('states_heatmap.html', new=2)
```