UM-SJTU Joint Institute

VE401

Probabilistic Methods in Eng

Investigation on Police Shootings in the United States

Author
Edric Guo 714370290031
Gerik Guo 714370290032
Ningfeng Yang 5133709104
Congfei Zhang 5133709241

Supervisor
Dr. Horst Hohberger

October 8, 2017

Abstract

Inspired by the article "London murders: a predictable pattern?" by David Spiegelhalter and Arthur Barnett, this report analyzes the occurrence of fatal police shootings between 2015 and 2016 in the US in a analogous manner. The data is obtained from the Database of Fatal Police Shootings of the Washington Post. Next, the term "fatal police shooting" is defined. The number of shootings each day is consequently displayed with bar graph. Whether the occurrence of police shootings follows a Poisson distribution and whether the number of fatal police shootings depends on the days of the week are investigated. The predicted and actual number of days with n shootings, n = 0, 1, 2, ..., are given side by side for comparison. The confidence interval for the parameter k of the Poisson distribution is then calculated. Lastly, data from the year 2017 is compiled. Tests to determine whether 2017's data follows a Poisson distribution, whether the calculated estimator $\hat{k}$ falls into the confidence interval, and whether each of the three years 2015's, 2016's, 2017's data falls into the 95% confidence interval obtained from Nelson's formula are performed.

Contents

1 Introduction

Eight gun shots in a single day. From on-duty police officers. 1954 shootings in two years, averaging about three lives per day. Shocking figures — or are they? By creating graphs and performing statistical tests on the fatal police shooting data, this report aims to find a pattern of the shootings.

2 Source of Data and the Meaning of Fatal Police Shooting

The data of the source comes from Washington Post. The Post began tracking every fatal shooting in the United States by a police officer in the line of duty since Jan.1, 2015. It's database is updated regularly as new news reports, public records, social media and other credible sources emerge. In 2016, the Post started to file open-records requests with departments for additional information. Note that the data solely represents the number of fatal police shooting, in which a police officer, in the line of duty, shoots and kills a civilian. It does not include deaths of people in police custody, fatal shootings by off-duty officers or non-shooting deaths. [1]

3 Police Shooting Data between January 1st, 2015 and December 31st, 2016 [2]

The pattern of fatal police shooting in America between 2015 and 2016 is examined. Figure 1 shows the number of fatal police shooting in each day over this two year period. Note that 2016 is a leap year and the data for February 29, 2016 is included by creating a bar for the data of each day in the years 2015 to 2016 manually and processing/counting the number of police shootings on that specific day using a c++ program thereafter.

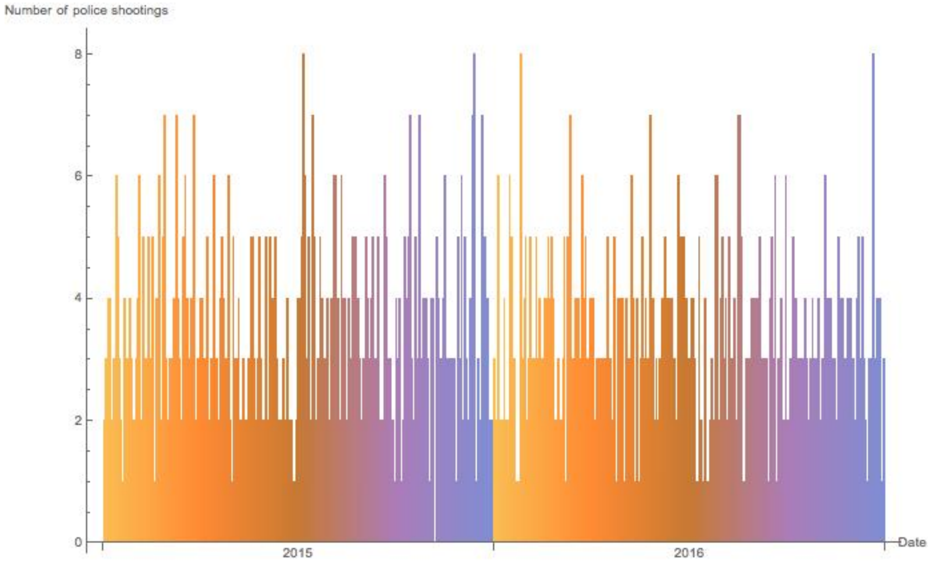

Figure 1: Prediction of Shooting in year 2016 and 2016

4 Test on the Occurrence of Police Shootings in the US in the Years 2015 to 2016 Follows a Poisson Distribution

Spiegelhalter and Barnett estimated that the London homicides follow a Poisson distribution with parameter $\hat{k} = 0.44$ from their data. For the police shooting data in the years 2015 to 2016, the estimator for k is $\hat{k} = \bar{X} = \frac{1954}{365+366} = 2.67$. To apply the goodness-of-fit test, we first calculate

$$\text{P[X = 0]} = \frac{e^{-\hat{k}} * \hat{k}^0}{0!} = 0.069$$
$$\text{P[X = 1]} = \frac{e^{-\hat{k}} * \hat{k}^1}{1!} = 0.185$$

$$\text{P[X = 2]} = \frac{e^{-\hat{k}} * \hat{k}^2}{2!} = 0.247$$
$$\text{P[X = 3]} = \frac{e^{-\hat{k}} * \hat{k}^3}{3!} = 0.220$$
$$\text{P[X = 4]} = \frac{e^{-\hat{k}} * \hat{k}^4}{4!} = 0.147$$
$$\text{P[X = 5]} = \frac{e^{-\hat{k}} * \hat{k}^5}{5!} = 0.078$$
$$\text{P[X = 6]} = \frac{e^{-\hat{k}} * \hat{k}^6}{6!} = 0.034$$
$$\text{P[X = 7]} = \frac{e^{-\hat{k}} * \hat{k}^7}{7!} = 0.013$$
$$\text{P[X >= 8] = 1 - P[X = 0] - P[X = 1] - ... - P[X = 7]} = 0.007$$

Table 1: Expected and Observed Number of Fatal Police Shooting in 2015 and 2016 (731 days)

	Number of days								
	0	1	2	3	4	5	6	7	>=8
Expected	50.44	135.24	180.56	160.82	107.46	57.02	24.85	9.50	5.12
Observed	50	149	163	155	115	60	23	12	4

The following test is then used

H_0: The occurrence of police shootings follows a Poisson distribution with k = 2.67

For N = 9 categories, the statistic

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

then follows a chi-squared distribution with N - 1 - m = 9 - 1 - 1 = 7 degrees of freedom.
Next, we want to realize $\alpha = 0.05$ and therefore reject H_0 if $X^2 > \chi^2_{0.05,7} = 14.07$. The calculated

$$X^2 = \frac{(50-50.44)^2}{50.44} + \frac{(149-135.24)^2}{135.24} + \frac{(163-180.56)^2}{180.56} + \frac{(155-160.82)^2}{160.82} + \frac{(115-107.46)^2}{107.46} + \frac{(60-57.02)^2}{57.02} + \frac{(23-24.85)^2}{24.85} + \frac{(12-9.50)^2}{9.50} + \frac{(4-5.12)^2}{5.12} = 5.048 < 14.07 = \chi^2_{0.05,7}.$$

Therefore, we are unable to reject H_0 at the 5% level of significance. And we take that the data follows a Poisson distribution with k = 2.67 in the years 2015 to 2016.
The following graph shows the frequency of occurrence of days with different numbers of fatal police shootings in the US between 2015 and 2016.

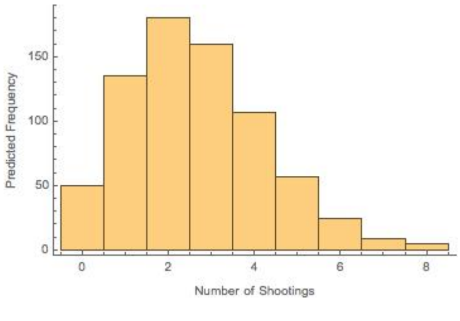

Figure 2: Predicted number of days for different numbers of shootings

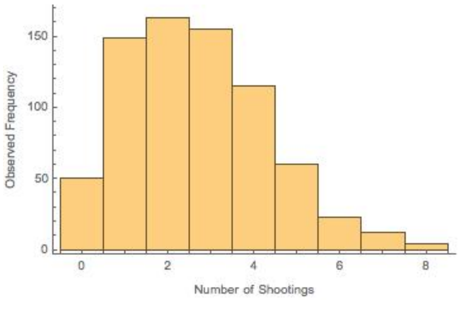

Figure 3: Observed number of days for different numbers of shootings

5 Dependency of the Number of Fatal Police Shootings on Days of the Week

Given that the data records the number of fatal police shootings on each day between 2015 and 2016, we want to test if the average numbers of shootings depends on the weekday. The following numbers of shootings for each weekday is compiled and shown in the table below.

Table 2: The Number of Shootings For Each Weekday

	Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.
Number of Shootings	249	294	307	285	273	269	277

The following histogram also shows this weekly distribution.

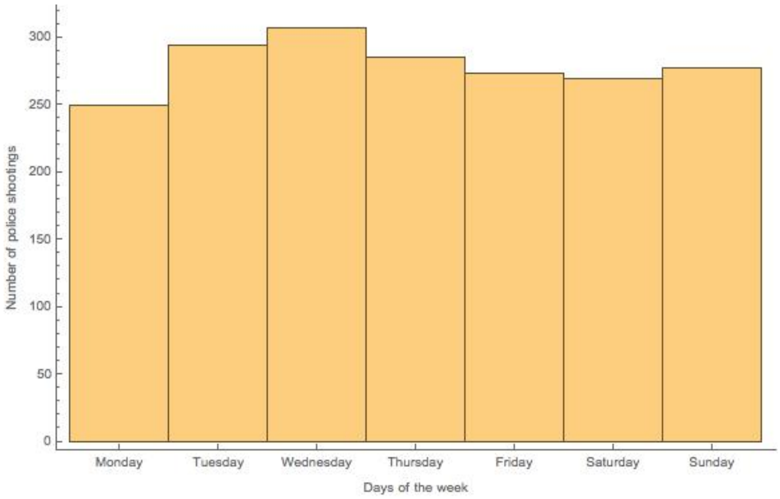

Figure 4: Distribution of the number of police shootings for days of the week.

We want to test if the data follows a discrete uniform distribution on Ω = 1,2,...,7 at a 5% significance level. We test

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

We have $E_i = 1954 * \frac{1}{7} = 279.1$ for i = 1,...7.
The observed test statistic is

$$\sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} = \frac{(249-279.1)^2}{279.1} + ... + \frac{(277-279.1)^2}{279.1} = 7.47.$$

This statistic follows a chi-squared distribution with 7-1 = 6 degrees of freedom. Since $\chi^2_{0.05,6}$ = 12.59, the P-value of the test is greater than 5%. Thus, there is not enough evidence to reject H_0. We conclude that there is no evidence that the number of police shootings is dependent on days of the week.

6 Graphs on Expected and Observed Number of Days with n (n∈ 0,1,2,3...) Shootings

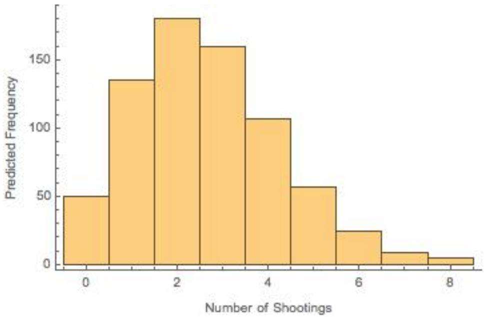

Figure 5: Predicted number of days for n = 0,1,2,3... occurrences of police shootings.

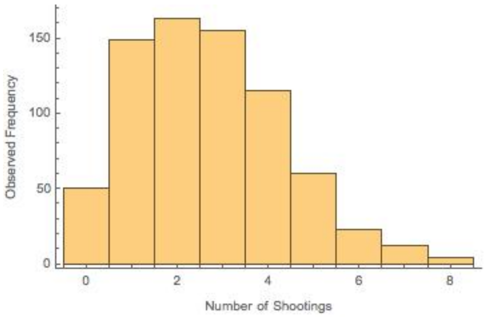

Figure 6: Observed number of days for n = 0,1,2,3... occurrences of police shootings.

From the two graphs we can see that the predicted occurrences of n shootings fairly resemble the actual occurrences of n shootings in the years 2015 to 2016. In both cases, the shape of the graph follows a Poisson distribution. The most commonly observed number of shootings in a given day is 2 in the years 2015 - 2016, and the observed frequencies decrease as the number of shootings increase or decrease in a given day.

7 Confidence Intervals for the Parameter k of a Poisson Distribution

We know that for a Poisson distributed random variable with parameter k, its expected value and variance are both k. Furthermore, the Central Limit Theorem tells us that for large sample size n, $(Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}})$ follows, at least approximately, a standard normal distribution N(0,1). Now, the equation and theorem above imply that, for large n:

$$(Z = \frac{\hat{k}-k}{\sqrt{\frac{k}{n}}})$$

also follows, at least approximately, a standard normal distribution N(0,1). Then we have

$$(P\left[-z_{\alpha/2} \leq \frac{\hat{k}-k}{\sqrt{\frac{k}{n}}} \leq z_{\alpha/2}\right] \approx 1-\alpha)$$

and we manipulate the quantity inside the parentheses:

$$(-z_{\alpha/2} \leq \frac{\hat{k}-k}{\sqrt{\frac{k}{n}}} \leq z_{\alpha/2})$$

to get the formula for a (1-α)100% confidence interval for k. Multiplying through the inequality by the quantity in the denominator, we get:

$$(-z_{\alpha/2}\sqrt{\frac{k}{n}} \leq \hat{k} - k \leq z_{\alpha/2}\sqrt{\frac{k}{n}})$$

Subtracting through the inequality by $(\hat{k})$, we get:

$$(-\hat{k} - z_{\alpha/2}\sqrt{\frac{k}{n}} \leq -k \leq -\hat{k} + z_{\alpha/2}\sqrt{\frac{k}{n}})$$

And, upon dividing through by -1, and thereby reversing the inequality, we get the claimed (1-α)100% confidence interval for k

$$(\hat{k} - z_{\alpha/2}\sqrt{\frac{k}{n}} \leq k \leq \hat{k} + z_{\alpha/2}\sqrt{\frac{k}{n}})$$

Although it appears that we need to know the parameter k in order to estimate the confidence interval of k for the Poisson distribution, we can replace the k's that appear in the endpoints of the interval with the sample mean $\bar{X}$, or namely the estimator for k, $\hat{k}$, to get an (approximate) (1-α)100% confidence interval for k:

$$(\hat{k} - z_{\alpha/2}\sqrt{\frac{\hat{k}}{n}} \leq k \leq \hat{k} + z_{\alpha/2}\sqrt{\frac{\hat{k}}{n}})$$

Thus, a (1-α)100% confidence interval for k is given by $\hat{k} \pm z_{\alpha/2}\sqrt{\frac{\hat{k}}{n}}$.
Now, to calculate a 95% confidence interval for k, we use the previous sample size n = 731, the previous estimator for k, $\hat{k} = \bar{X} = 2.67$, and $z_{\alpha/2} = z_{0.05/2} = z_{0.025} = 1.96$.

$$[2.67 - 1.96\sqrt{2.67/731}, 2.67 + 1.96\sqrt{2.67/731}] = [2.55, 2.79]$$

8 Fit Data of Year 2017 into a Poisson Distribution

We first assume data in year 2017 (until 2017.7.28) follows a Poisson distribution with parameter k_{2017}, we can estimate the parameter k_{2017} by moment generation method. i.e.

$$k_{2017} = \bar{X}$$

where $\bar{X}$ denotes the mean of police shoots happened per day in the observed data.
From the data we can get

$$\hat{X} = 2.75$$

So we can assume that data from year 2017 follows a Poisson distribution with parameter

$$k_{2017} = \sum_{i=1}^{8} \frac{(O_i - E_i)^2}{E_i} = 2.75$$

. In order to test the goodness of fit, we take this fit as the H_0 hypothesis.
Based on the hypothesis, we do the prediction of police shooting in year 2017 We find the possibility of having n shooting each day by the equation:

$$P[X = x] = \frac{e^{-k_{2017}} * k_{2017}^x}{x!}$$

Where x ranges from 1 to 8. Then we can find the predicted data by multiplying the probability and total days, as shown in the following table.

Table 3: Expected and Observed Number of Fatal Police Shooting in 2017 (207 days)

	Number of days								
	0	1	2	3	4	5	6	7	>=8
Expected	13.19	36.31	49.99	45.88	31.59	17.40	7.98	3.14	1.08
Observed	13	37	46	48	36	14	10	1	2

By Chi-squared test, we get

$$\chi^2_{2017} = 4.46$$

As mentioned before, the DOF of this Chi-squared distribution is 7. Thus we can find the p value of observing a distribution that is extremer than this χ^2_{2017} value is about 0.30, thus we can not reject H_0, which means this is a fair enough curve fit.

9 Prediction of Police Shooting from Data of Year 2015 and 2016

By Nelson's formula, we can calculate the 95% confidence interval based on the predicted value. denote L as the lower bound and U as the upper bound.

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

For 95% CI, we have $Z_{1-\alpha}$=1.96.
For year 2015, we have $m = n = 365$, and for year 2016, we have $m = n = 366$.
As described before, data in year 2015 follows a Poisson distribution with parameter $k_{2015} = 2.72$, and year 2016 follows a Poisson distribution with parameter $k_{2016} = 2.63$. Plot the prediction along with the confidence interval.

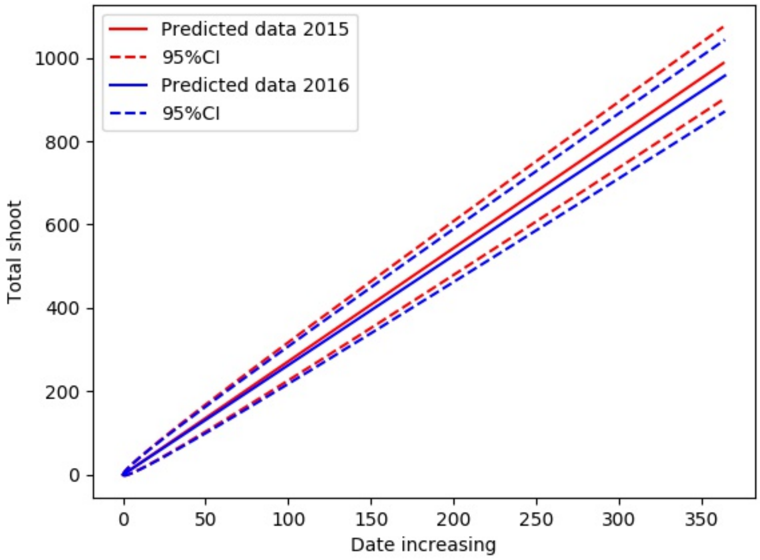

Figure 7: Prediction of Shooting in year 2016 and 2016

10 Comparison between Prediction and Data Observed in Year 2017

Plot the observed data in year 2017 along with the predicted data interval, we get this figure. There exists slight difference in the prediction of year 2015 and 2016.

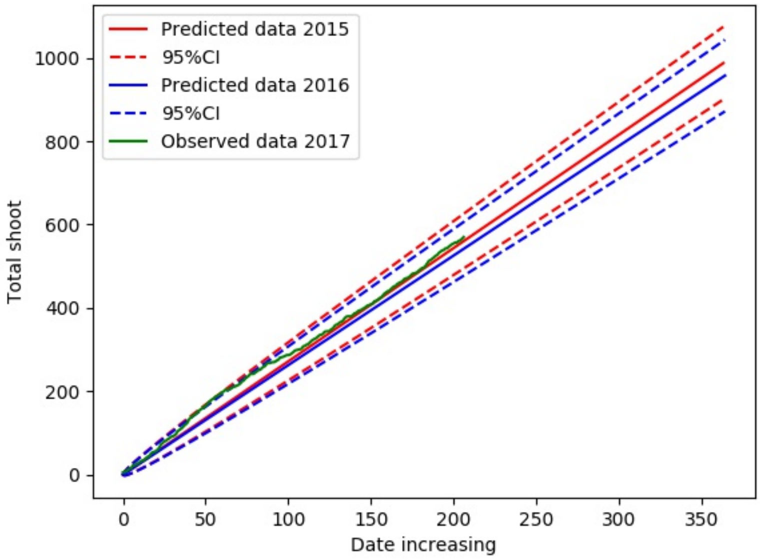

Figure 8: Data of 2017, Compared to Prediction

From the plot we can find that most of the observed data in 2017 falls in the confidence interval. Counting the points, we find that:
98.07% of data from year 2017 is in the 95% CI of year 2015,
87.93% of data from year 2017 is in the 95% CI of year 2016

References

[1] S. Rich J. Muyskens K. Elliott T. Mellnik J. Tate, J. Jenkins and A. Williams. How the washington post is examining police shootings in the united states. 2016. `https://www.washingtonpost.com/national/how-the-washington-post-is-examining-police-shootings-in-the-united-states/2016/07/07/d9c52238-43ad-11e6-8856-f26de2537a9d_story.html?utm_term=.a1f6b9c92f02`. Accessed on Aug 3, 2017.

[2] The Washington Post. `https://github.com/washingtonpost/data-police-shootings`. Accessed on Aug 3, 2017.