DOES THE OCCURRENCE OF FATAL POLICE SHOOTING IN THE UNITED STATES FOLLOW A PREDICTABLE PATTERN?

Tanrui Wu 518370910221
Taoyue Xia 518370910087
Xingyu Zhu 518370910023

2021-07-18

Introduction

Each year, There are thousands of fatal police shootings happening in the United States. So do they follow a predictable pattern? In our project, we analyse the data of those shootings from 2015 to 2021 and try to find some rules behind them. Since the shootings are independent; happening one does not change the probability of when the next one will happen, and the shootings occur with an almost constant rate within a fixed interval of time. We make our assumption that the number of shootings per day follows a Possion distribution and try to test it during this project.

What means "fatal police shooting"? From the detailed information provided on the website, The Washington Post. Fatal force[1]. We can have a general idea of the meaning of the term "fatal police shooting". Among the term, the "police" stands for not only on-duty police officers, but it also can be off-duty officers or deputies of the County sheriff. In the cases that fatal police shooting occur, most of the suspects were shot and killed immediately, but there are also people shocked by stun gun first and shot later. Also, they all show threat to the "police" to some extent, before being shot. And however the process was, they died in the end, which corresponds to the word, "fatal".

In this project, we first explain the meaning of "fatal", summarize the data, and visualize the data from 2015 to 2020. Then we test whether the number of shootings per day in the last 6 years follows a Poisson distribution, and calculate the confidence interval of the parameter k, which is also the expected shooting number per day. Finally, we test the data in 2021 to see whether there is any factor, for example, Coronavirus, has influenced the occurrence of fatal police shootings, and make some comparisons between the observed data and our expectations.

Data

The data we download from the website[2] records the information of every fatal police shooting in the U.S each day from January of 2015 to July of 2021. It includes the name, gender, age and race of the suspects, and record how he/she is killed, and the condition when the fatal shooting happened, for example,the threat level, whether he/she was armed, whether he/she showed sign of mental illness and so on. And most importantly, it contains the location and date of the cases.

The file `fatal-police-shootings-data.csv` contains data about each fatal shooting in CSV format. Each row has the following variables:

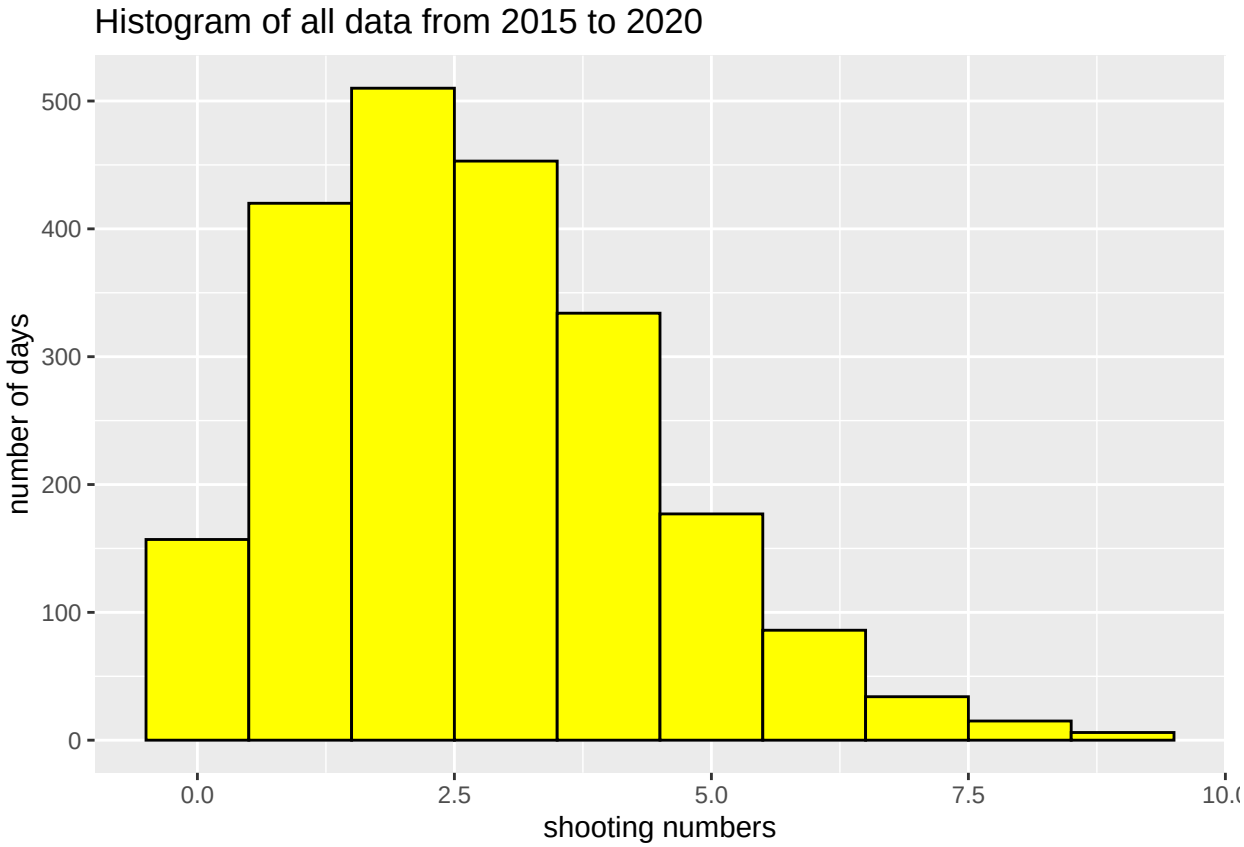
- id: a unique identifier for each victim

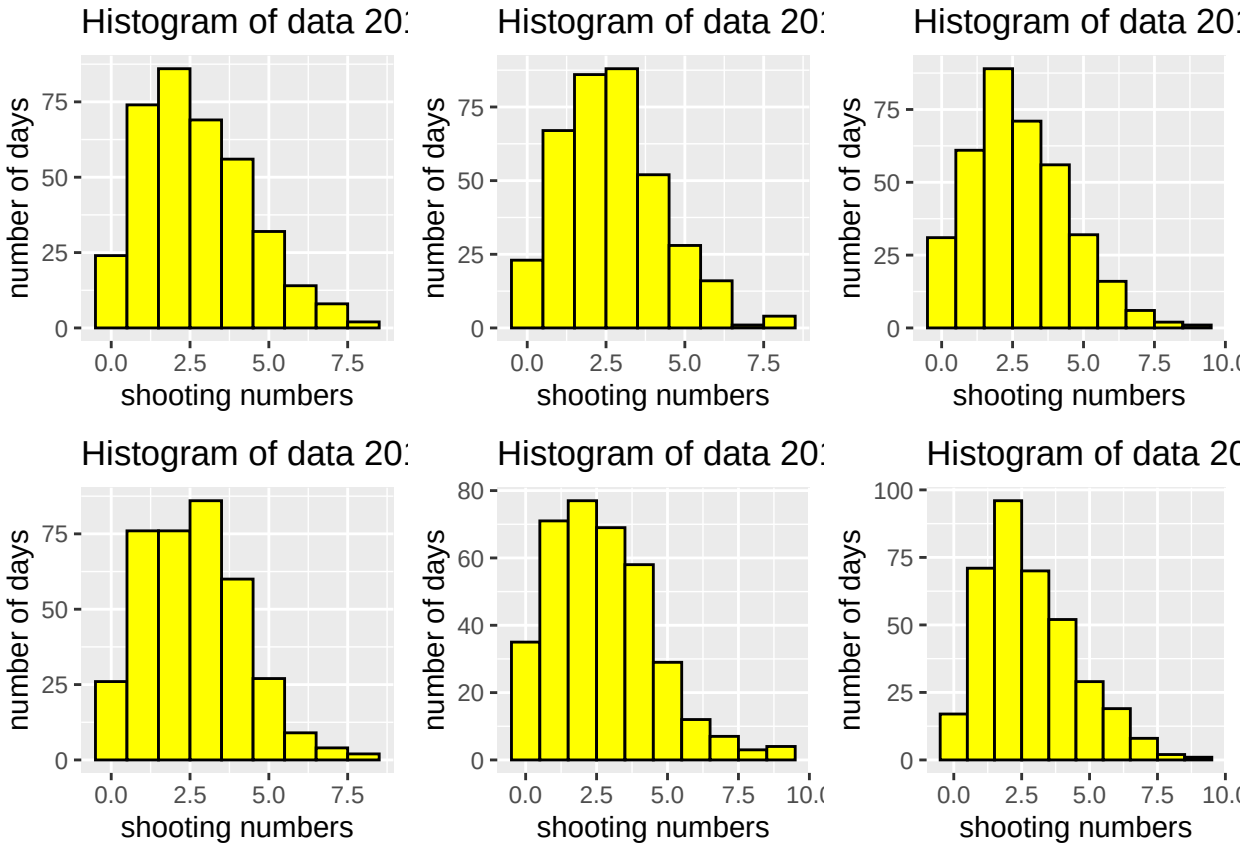
- name: the name of the victim
- date: the date of the fatal shooting in YYYY-MM-DD format
- manner of death:
  - shot
  - shot and Tasered
- armed: indicates that the victim was armed with some sort of implement that a police officer believed could
  - undetermined: it is not known whether or not the victim had a weapon
  - unknown: the victim was armed, but it is not known what the object was
  - unarmed: the victim was not armed
- age: the age of the victim
- gender: the gender of the victim. The Post identifies victims by the gender they identify with if reports indicate that it differs from their biological sex.
  - M: Male
  - F: Female
  - None: unknown
- race:
  - W: White, non-Hispanic
  - B: Black, non-Hispanic
  - A: Asian
  - N: Native American
  - H: Hispanic
  - O: Other
  - None: unknown
- city: the municipality where the fatal shooting took place. Note that in some cases this field may contain a county name if a more specific municipality is unavailable or unknown.
- state: two-letter postal code abbreviation
- signs of mental illness: News reports have indicated the victim had a history of mental health issues, expressed suicidal intentions or was experiencing mental distress at the time of the shooting.
- threat level: The threat level column was used to flag incidents for the story by Amy Brittain in October 2015. http://www.washingtonpost.com/sf/investigative/2015/10/24/on-duty-under-fire/ As described in the story, the general criteria for the attack label was that there was the most direct and immediate threat to life. That would include incidents where officers or others were shot at, threatened with a gun, attacked with other weapons or physical force, etc. The attack category is meant to flag the highest level of threat. The other and undetermined categories represent all remaining cases. Other includes many incidents where officers or others faced significant threats.
- flee: News reports have indicated the victim was moving away from officers
  - Foot
  - Car

  - Not fleeing

- body camera: News reports have indicated an officer was wearing a body camera and it may have recorded some portion of the incident.
- latitude and longitude: the location of the shooting expressed as WGS84 coordinates, geocoded from addresses. The coordinates are rounded to 3 decimal places, meaning they have a precision of about 80-100 meters within the contiguous U.S.
- is geocoding exact: reflects the accuracy of the coordinates. true means that the coordinates are for the location of the shooting (within approximately 100 meters), while false means that coordinates are for the centroid of a larger region, such as the city or county where the shooting happened.

To have a general impression of the data from 2015 to 2020, we first plot the number of fatal police shooting per day from 2015 to 2020,





From the diagram, we can see that the number of fatal police shooting per day varies from 0 to 9. There are only few days that no shooting happened. The shape of "Histogram of data" is very similar to Poisson distribution. However, it is hard to find any obvious issues directly from the diagram, so we need to do further analysis in detail.

Methods

Hypothesis test

Since in 2020 the coronavirus broke out, we will look into the police shooting events happened from 2015 to 2019.

Now that we want to study the distribution of the number of everyday shootings, we first construct a histogram to give a rough knowledge. After gathering the five-year data together, we find that the data shown by the histogram fit a Poisson distribution to some extent, which motivate us to use hypothesis test to prove our view.

Now we give the null hypothesis that the number of everyday shootings from 2015 to 2019 follows a Poisson distribution. Since the mean of a random variable that follows a Poisson distribution equals to k, we first approximate the parameter k using sample mean, which can be expressed as:

$$\hat{k} = \frac{N}{n} = \frac{1}{n}\sum_{i=0}^{max}(i \times O_i),$$

where N stands for total shooting numbers, n denotes for total days in these five years, and O_i is the number of days counting for each number of shootings from 0 to max. After that, referring to Cochran's Rule, we can

construct a statistic X such that:

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

follows a Chi-squared distribution, where E_i is the expected days for each shooting number.
Finally, we can decide whether to reject the null hypothesis by comparing the computed X^2 with $\chi^2_{0.05,8}$ at $\alpha = 0.05$ level.

Bootstrap Confidence Intervals

After estimating the value of the parameter k, we also want to give a confidence interval for $\hat{k}$. In previous part, we introduce one estimator for parameter k:

$$\hat{k}_1 = \frac{N}{n} = \frac{1}{n}\sum_{i=0}^{max}(i \times O_i).$$

Since the variance of a random variable that follows a Poisson distribution also equals to k, we can also approximate the parameter k using sample variance, which can be expressed as:

$$\hat{k}_2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \bar{X})^2,$$

where X_k is sample value of shooting numbers in one day.

We will use bootstrap instead of Monte Carlo method to give a 95% confidence interval for $\hat{k}_1$ and $\hat{k}_2$. This is because it is not realistic to let time go back and draw multiple samples for the number of shooting per day, making it hard for us to use Monte Carlo methods. Using bootstrap, we can estimate the sampling distribution of $\hat{k}$ by drawing multiple samples from the original sample.

Suppose that X_1^*, X_2^*,..., X_n^* is a sample randomly picked with replacement from the original sample which has n data. We calculate the statistic

$$\hat{k}_1^* = \frac{N*}{n}\hat{k}_2^* = \frac{1}{n}\sum_{k=1}^{n}(X_k^* - \bar{X}^*)^2$$

with the newly-picked sample X_1^*, X_2^*,..., X_n^*.

We do such re-sampling 10000 times. Then the resulted 10000 $\hat{k}^*$ values form the sampling distribution of $\hat{k}$. We will than take the basic bootstrap confidence intervals, which is calculated as follows:

$$[2\hat{k} - \hat{k}^*_{0.975}, 2\hat{k} - \hat{k}^*_{0.025}].$$

Simulation

2015:

The critical value is:

```
##      2.5%     97.5%
##  1.715012 15.975985
```

The power of the test:

```
## [1] 0.9874
```

2016:

The critical value is:

```
##      2.5%     97.5%
##  1.638511 16.042217
```

The power of the test:

```
## [1] 0.9922
```

2017:

The critical value is:

```
##      2.5%     97.5%
##  1.755625 16.269545
```

The power of the test:

```
## [1] 0.9893
```

2018:

The critical value is:

```
##      2.5%     97.5%
##  1.660038 16.128671
```

The power of the test:

```
## [1] 0.9879
```

2019:

The critical value is:

```
##      2.5%     97.5%
##  1.748418 16.215108
```

The power of the test:

```
## [1] 0.9887
```

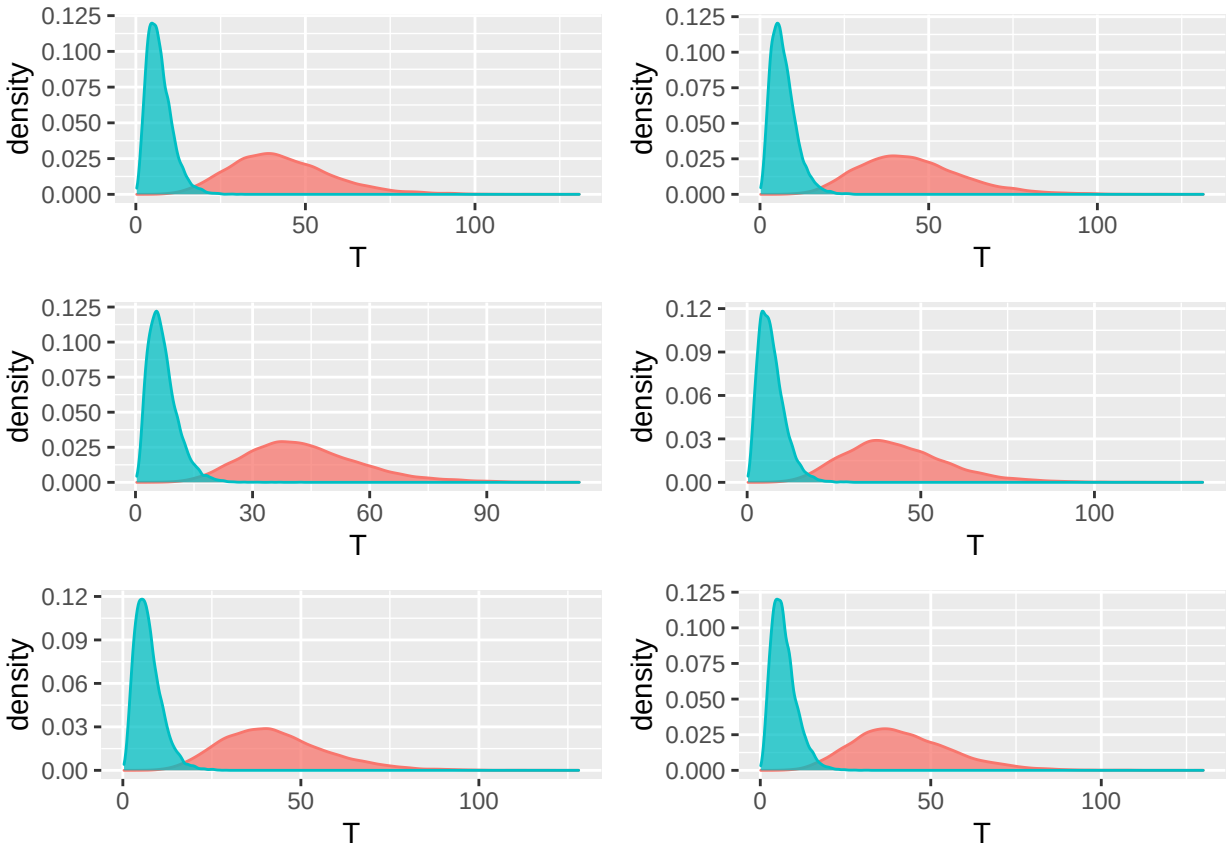
2020:

The critical value is:

```
##      2.5%     97.5%
##  1.677705 15.865349
```

The power of the test:

```
## [1] 0.9877
```



Reference

1 The Washington Post. Fatal force. https://www.washingtonpost.com/graphics/investigations/police-shootings-database/. Web.

2 "washingtonpost/data-police-shootings", GitHub, 2021. [Online]. Available: https://github.com/washingtonpost/data-police-shootings.