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:

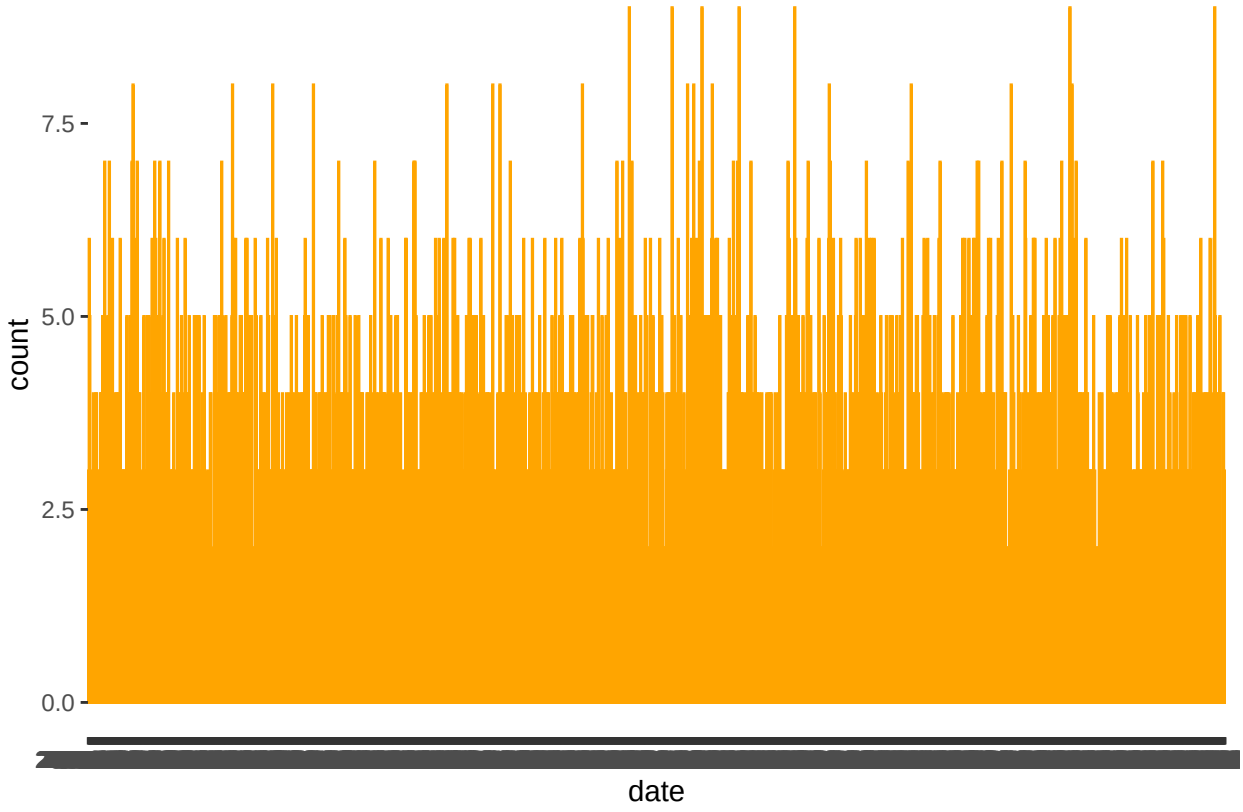
- id: a unique identifier for each victim
- name: the name of the victim
- date: the date of the fatal shooting in YYYY-MM-DD format
- manner of death:
  - shot
  - shot and Tasered
- armed: indicates that the victim was armed with some sort of implement that a police officer believed could
  - undetermined: 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.

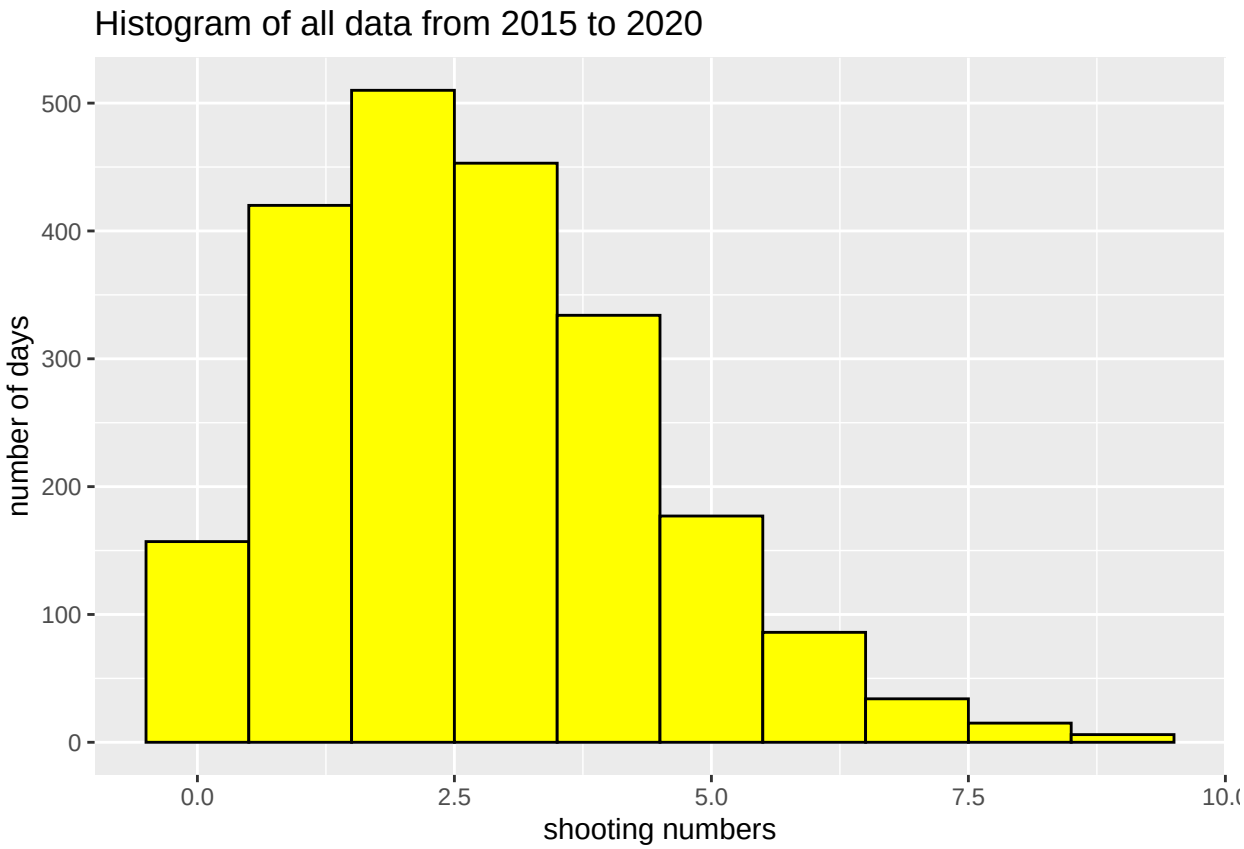
  - M: Male
  - F: Female
  - None: unknown
- race:
  - W: White, non-Hispanic
  - B: Black, non-Hispanic
  - A: Asian
  - N: Native American
  - H: Hispanic
  - O: Other
  - None: unknown
- city: the municipality where the fatal shooting took place. Note that in some cases this field may contain a county 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

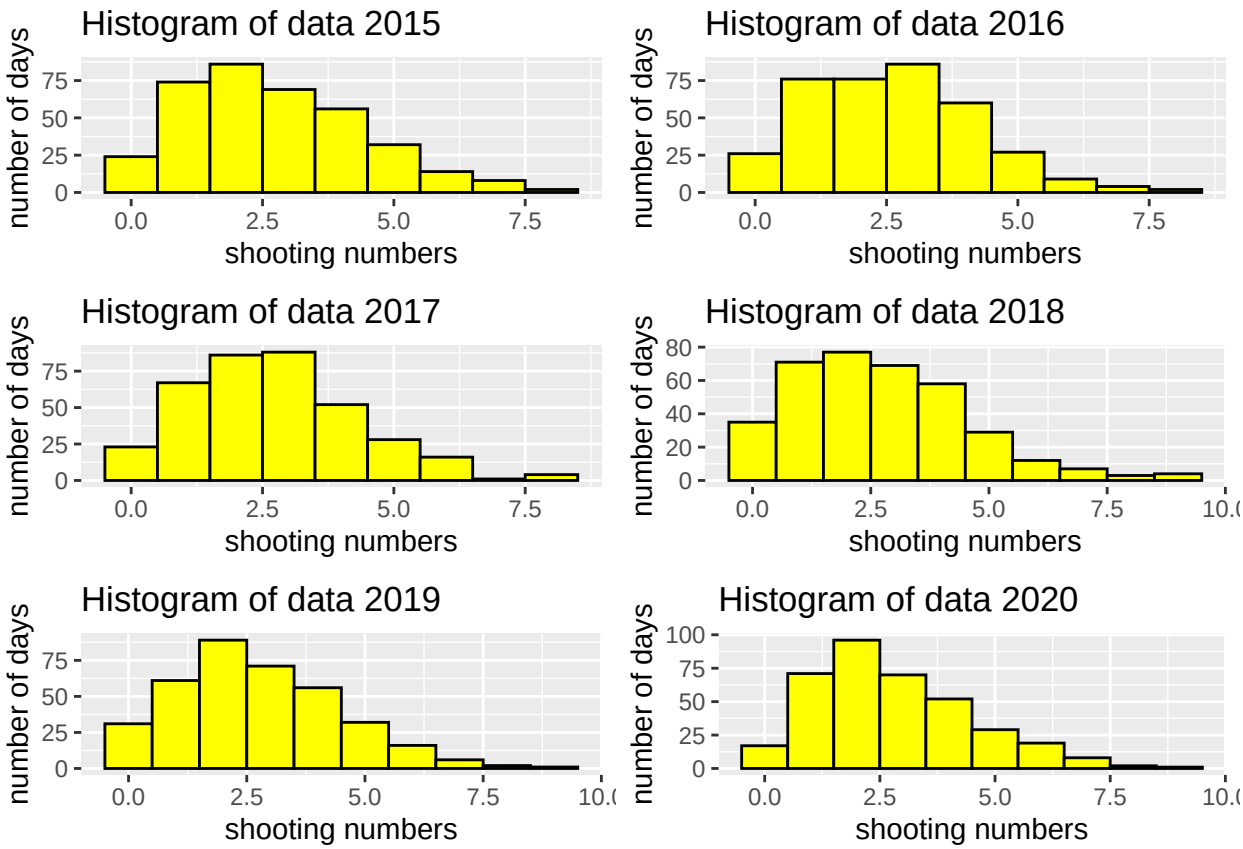
- body camera: News reports have indicated an officer was wearing a body camera and it may have recorded some portion of the incident.
- latitude and longitude: the location of the shooting expressed as WGS84 coordinates, geocoded from addresses. The coordinates are rounded to 3 decimal places, meaning 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,

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```




Histogram of all data from 2015 to 2020
number of days
500
400
300
200
100
0
0.0
2.5
5.0
7.5
shooting numbers



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

Denote X as the number of fatal shootings happened each day. First we want to estimate the value of average fatal shootings per day $\bar{X}$. Since sample mean is an unbiased estimator, we use following estimator:

$$\hat{k} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n}(X_i),$$

where n is the number of days in a year. If the value of $\hat{k}$ didn't grow significantly in 2020, there is no evidence that fatal shooting is more frequent. We will use bootstrap instead of Monte Carlo method to give a 95% confidence interval for $\hat{k}$. 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.

What we want to study after that is the distribution of X. If the distribution of X in 2020 didn't differ significantly from distributions in past few years, there is no evidence that the police had changed their gun-using habits. From Fig.2, 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 fatal shootings follows a Poisson distribution with parameter $\hat{k}$.The alternative hypothesis is that the number of everyday fatal shootings follows a Poisson distribution with parameter $\hat{k} + 0.5$.

$$H_0 : X \sim \text{Poisson}(k), k = \hat{k} H_1 : X \sim \text{Poisson}(k), k \geq \hat{k} + 0.5$$

After that, referring to Goodness-of-Fit Test for a discrete distribution, we can use the Person statistic X^2_{k-1} such that:

$$T(X)_1 = \sum_{i=0}^{k} \frac{(E_i - O_i)^2}{E_i},$$

to test whether certain data follows a Poisson distribution. Here, E_i is the expected number of days at which i fatal shootings happened, while O_i is the observed number of days at which i fatal shootings happened. The value of E_i can be expressed as $E_i = n \cdot \mathrm{P}[X = i]$. Monte Carlo methods will be applied to obtained a rejection region for this test statistic and to calculate the power of the test. We can decide whether to reject the null hypothesis by comparing the computed $T(X)_1$ and the critical value we get by using Monte Carlo method.

We are also interested in the ratio of mean to variance of the data,

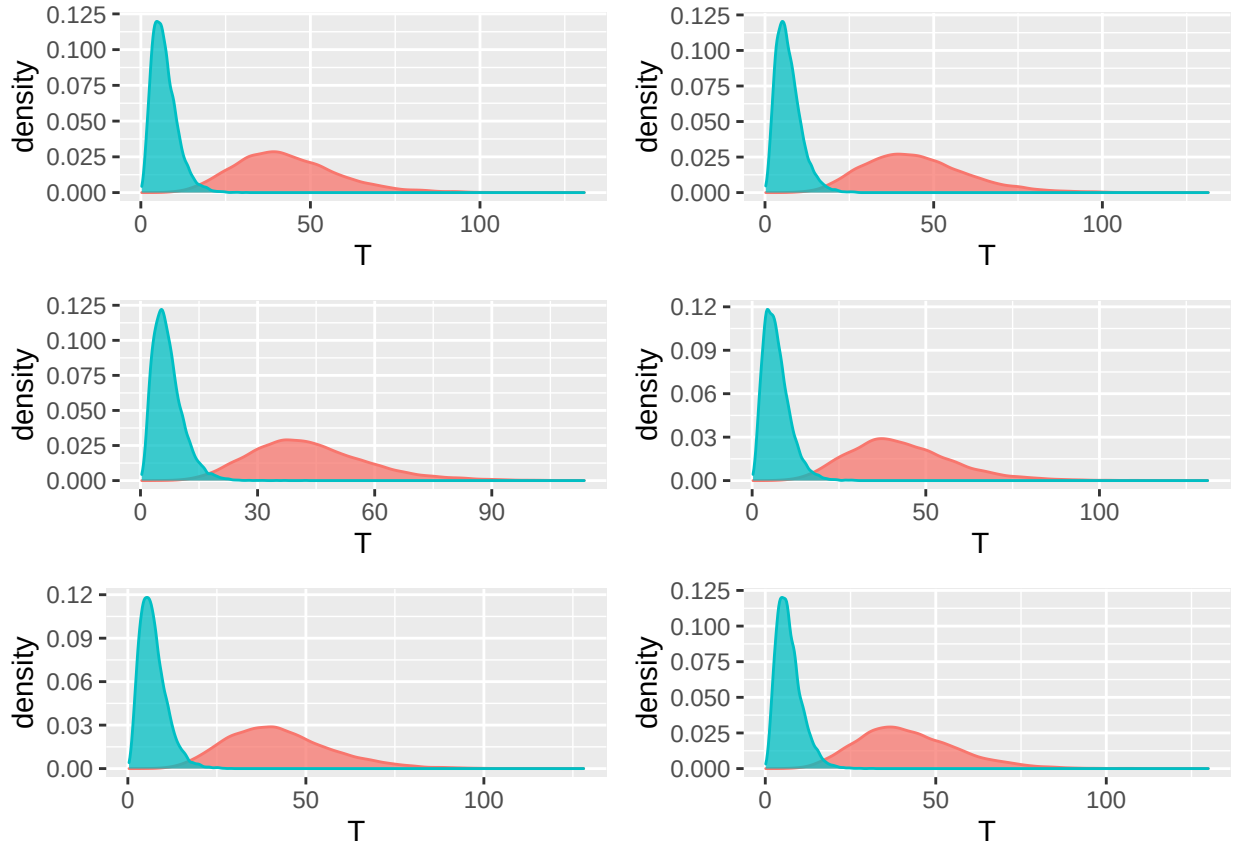
$$T(X)_2 = \frac{\bar{X}}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2},$$

since the mean equals the variance for data following Poisson distribution, which will be proved by Monte Carlo method. We will estimate the ratio of mean to variance of our data and provide a 95% confidence interval for this value using bootstrap. We expect such confidence interval to include 1, which indicates evidences that our fatal shooting data may follows a Poisson distribution.

Simulation

We will randomly generate 10000 samples. Each sample is drawn from Poisson distribution with parameter $\hat{k}$. Each sample size equals to number of days one year. For each sample, the X^2_{k-1} statistic will be calculate and the resulting 10000 value will form a sample distribution of X^2_{k-1}. The 2.5% and 97.5% quantile will be the critical value for our test.

After that, we draw the six density plots corresponding to the years from 2015 to 2020, which is shown below.



After that, we calculate the critical regions as the interval between 2.5% and 97.5% quantile of each data set.

We also generate another 10000 samples drawn from Poisson distribution with parameter $k^* = \hat{k} + 0.5$. Following the same procedure, we can similarly calculate 10000 values of statistic X^{*2}_{k-1}.

Finally, we can get the power of our test as the probability of the 10000 X^{*2}_{k-1} falling in the rejection region $(-\infty, 2.5\%\text{ quantile}) \cup (97.5\%\text{ quantile}, \infty)$.

The rejection region and corresponding power is shown in the table below:

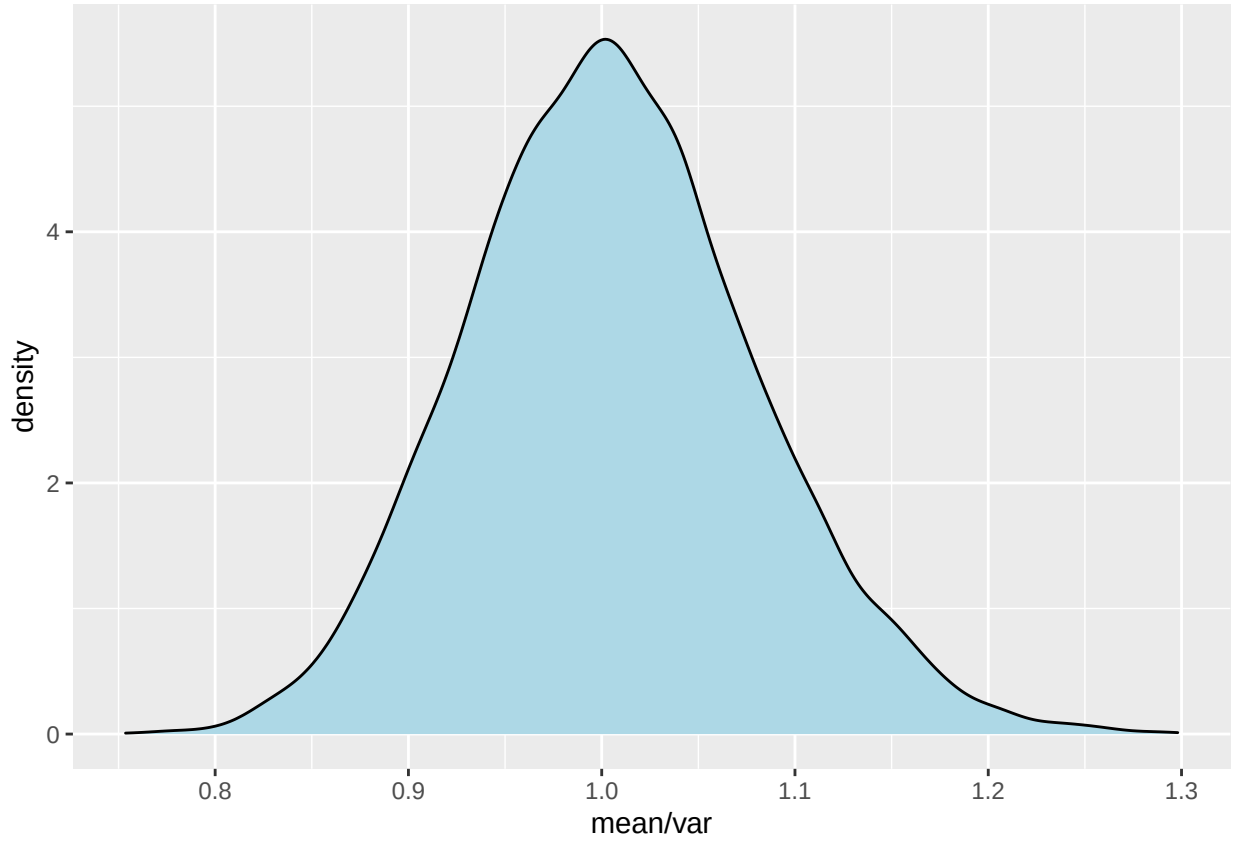
Year	Rejection Region	Power
2015	$(-\infty, 1.715) \cup (15.976, +\infty)$	0.9874
2016	$(-\infty, 1.639) \cup (16.042, +\infty)$	0.9922
2017	$(-\infty, 1.756) \cup (16.270, +\infty)$	0.9893
2018	$(-\infty, 1.660) \cup (16.129, +\infty)$	0.9879
2019	$(-\infty, 1.748) \cup (16.215, +\infty)$	0.9887
2020	$(-\infty, 1.678) \cup (15.865, +\infty)$	0.9877

Table 1: Results for Monte Carlo Hypothesis Test

From table 1, we find that the power of our test is extremely close to 1, which indicates that the simulation of our test is powerful and successful.

We then continue to prove that the mean and variance of a Poisson distribution are equal, and taking one sample into account is enough to show the relation between the two.

To ensure the consistency, we reuse the sample generated to simulate data in 2015. For the 10000 samples, we calculate the value of their mean divided by variance. The density plot of the 10000 values is shown below:



From the plot, we can obviously see that it follows a normal distribution and reaches the peak at 1. Still, to get the 95% confidence interval, we calculated the two-tailed quantile from 2.5% to 97.5%, which is $(0.8699, 1.1630)$. We can simply find that 1 lies in the confidence interval, thus we have confidence that the value of mean divided by variance is 1.

Therefore, in the analysis part, we can calculate the real mean and variance. If their division is close to 1, then we will have great confidence to believe that the pattern of shootings follows a Poisson distribution.

Analysis

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:

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

2015: bootstrap for mean/var

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_mean_var, type = c("norm", "basic", "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 95%   ( 0.8031,  1.0544 )   ( 0.7983,  1.0455 )   ( 0.8274,  1.0746 )
## Calculations and Intervals on Original Scale
```

bootstrap for mean:

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_mean, type = c("norm", "basic", "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 95%   ( 2.552,  2.896 )   ( 2.543,  2.893 )   ( 2.548,  2.899 )
## Calculations and Intervals on Original Scale
```

2016: bootstrap for mean/var

bootstrap for mean:

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
```

```
## CALL :
## boot.ci(boot.out = boot_mean, type = c("norm", "basic", "perc"))
##
## Intervals :
## Level      Normal              Basic              Percentile
## 95%   ( 2.448,  2.787 )   ( 2.445,  2.795 )   ( 2.445,  2.795 )
## Calculations and Intervals on Original Scale
```

2017: bootstrap for mean/var

bootstrap for mean:

2018: bootstrap for mean/var

bootstrap for mean:

2019: bootstrap for mean/var

bootstrap for mean:

2020: bootstrap for mean/var

bootstrap for mean:

Years	Normal	Basic	Percentile
2015	(0.8031, 1.0544)	(0.7983, 1.0455)	(0.8274, 1.0746)
2016	(0.887, 1.173)	(0.875, 1.161)	(0.908, 1.195)
2017	(0.873, 1.178)	(0.860, 1.168)	(0.897, 1.205)
2018	(0.6557, 0.8948)	(0.6514, 0.8871)	(0.6815, 0.9172)
2019	(0.7816, 1.0212)	(0.7752, 1.0127)	(0.8040, 1.0415)
2020	(0.8108, 1.0813)	(0.7964, 1.0708)	(0.8306, 1.1050)

Table 2: Bootstrap for mean/var of each year

Years	Normal	Basic	Percentile
2015	(2.552, 2.896)	(2.543, 2.893)	(2.548, 2.899)
2016	(2.448, 2.787)	(2.445, 2.795)	(2.445, 2.795)
2017	(2.537, 2.864)	(2.543, 2.863)	(2.540, 2.860)
2018	(2.523, 2.905)	(2.532, 2.912)	(2.512, 2.893)
2019	(2.562, 2.914)	(2.556, 2.915)	(2.559, 2.918)
2020	(2.630, 2.965	(2.625, 2.962)	(2.633, 2.970)

Table 3: Bootstrap for mean of each year

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.