SJTU Joint Institute Date: April 30, 2020

VE401 Probabilistic Methods in Engineering

Police Shootings in the United States

Group 46: Junqi Chen, Ruiyu Li, Shuang Chen

Abstract

Analogous to the analysis of London murders between 2004 and 2007 by David Spiegelhalter and Arthur Barnett in *London murders: a predictable pattern?* [4], we investigate the pattern of the occurrence of fatal police shootings in the US between 2015 and 2020 using data from the *Database of Fatal Police Shootings* [6]. We first introduce the source of data and characterize how the term ”fatal police shooting” is used here. We then use Mathematica to visualize the daily occurrence of police shootings from 2015 to 2019. Moreover, utilizing a Pearson's Chi-squared Goodness-of-fit test, we conclude that there is no enough evidence that the average number of police shootings depends on the weekday. Next, we test the hypothesis that the occurrence of police shootings in the US has followed a Poisson distribution in the years 2015-2019 and conclude that there is no enough evidence that the occurrence of police shootings in the US has not followed a Poisson distribution with parameter $k \approx 2.704$. We also prove a formula of a $(1 - \alpha)100\%$confidence interval for parameter k of a Passion distribution and calculate such an in-terval using the data of the years 2015 to 2019. Base on past data from 2015 to 2020, we predict the shooting occurrences in 2020 and find out it follows Poisson distribution with $k \approx 2.86$. With Nelson's formula, we calculate the prediction interval for shooting occurrence in 2020 and plot the curves with given cases to check the correctness. Fi-nally, some possible analysis for relation between shooting occurrences and the outbreak of coronavirus in 2020 is given in the last section.

Keywords: Fatal police shootings, Pearson's Chi-squared Goodness-of-fit test, Poisson distribution, ,Confidence interval, Prediction interval

Contents

1 Introduction

In this project, we will analyze the occurrence of fatal police shootings in the US using data from the Database of Fatal Police Shootings of the Washington *Post*[6].

We first clarify the source of data we use and visualize the number of fatal police shootings in the US each day between January 1^{st} 2015 and December 31^{st} 2019. Then we perform a Pearson's Chi-squared Goodness-of-fit test to determine whether there is an evidence that the average number of police shootings depends on the weekday. Next, we test the hypothesis that the occurrence of police shootings in the US has followed a Poisson distribution in the years 2015-2019. Using a goodness-of-fit test, we conclude that there is no enough evidence that the occurrence of police shootings in the US has not followed a Poisson distribution with parameter $k \approx 2.704$. We also prove a formula of a (1 - α)100% confidence interval for parameter k of a Passion distribution and calculate such an interval using the data of the years 2015 to 2019. Later we predict the shooting occurrence in 2020 base on past data from 2015 to 2020 and find out it follows Poisson distribution with $k \approx 2.86$. With Nelson's formula, we are going to calculate the prediction interval for shooting occurrence in 2020 and plot the curves with given cases to check the correctness. Finally, some possible analyze for relation between shooting occurrence and the outbreak of coronavirus in 2020 is given in the last section.

2 Fatal Police Shooting Data

2.1 Source of Data

The data of fatal police shootings comes from the database established and updated by *Washington Post.* The database contains data of every fatal shooting in the United States by a police officer in the line of duty since Jan. 1, 2015. The database also includes detailed information about each police shootings, such as the age, gender, and race of the victim and the places where the shooting happened. The data is collected from local news reports, law enforcement websites, social media, and independent databases [1]. The database is acknowledged by the officials that the database is more complete than the data logged by The FBI and the Centers for Disease Control and Prevention.

The data used in this project is accessed on April 28^{th}, 2020.

2.2 Characterization of the Term "Fatal Police Shooting"

According to *Washington Post*, the definiton of "fatal police shooting" here refers to "those shootings in which a police officer, in the line of duty, shoots and kills a civilian", while "deaths of people in police custody, fatal shootings by off-duty officers or non-shooting deaths" are excluded[1].

2.3 Daily Occurrence of Police Shootings

After obtaining the data from the database, we use Matlab to process data and use Mathematica to plot Figure 1, which visualizes the number of fatal police shooting in the US each day between January 1^{st} 2015 and December 31^{st} 2019.

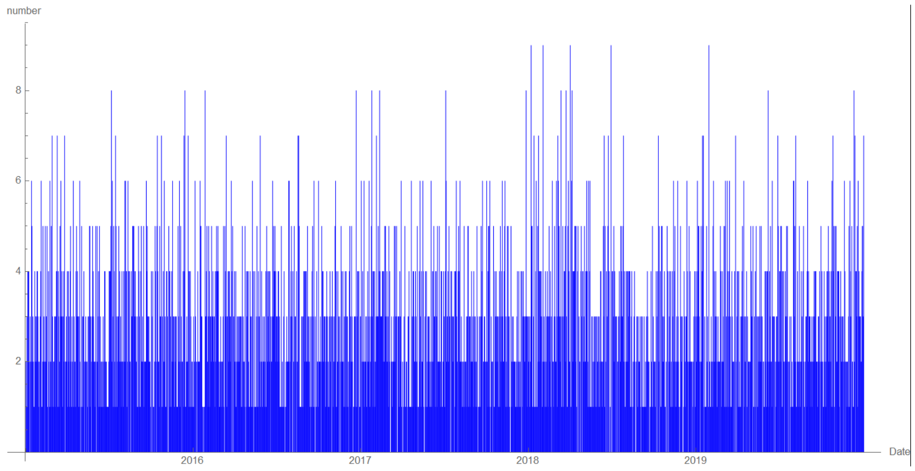


Figure 1: Number of fatal police shooting each day between January 1st 2015 and December 31st 2019

A total of 4938 cases happened in 1826 days from 2015-2019. The highest frequency is 9 cases per day and occurred five times while there are 139 days when no fatal police shooting was reported.

3 Dependence of Average Number of Police Shootings on Weekdays

In this section, we are going to test whether there is evidence that the average number of police shootings depends on the weekday using a Pearson's Chi-squared Goodness-of-fit test.

Firstly, by Mathematica, we plot the frequency of occurrence of police shootings on different weekdays and different months in the US from January 1st 2015 to December 31st 2019, shown in Figure 2 and 3. We can see that the number of occurrence of police shootings has some differences between different weekdays. Hence, we'd like to test the dependence of average number of police shootings on weekdays using the Police Shooting data from January 1st 2015 to December 31st 2019.

Let (X_1, X_2, X_3, X_4, X_5, X_6, X_7) denote the number of police shootings from Monday to Sunday and the sample size n = $\sum_{i=1}^{7} x_i$ = 4938. If the average number of police shootings is independent of weekdays, the probability of a police shooting occurring on each weekday p_i, i = 1, ..., 7 is equal to the frequency of each weekday in the data used in the test and $\sum_{i=1}^{7} p_i = 1$. Thus, we'd like to perform **a Pearson's Chi-squared Goodness-of-fit test** to determine whether the data follow an approximately uniform distribution on Ω = {Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday}. If we used the data that contains the same number of each weekday, $p_i = \frac{1}{7}$ for $i \in [1, 7]$. However, since the number of each weekday may not be the same from 2015 - 2019 and the test statistic is sensitive to each p_i, we count the number of each days and use the frequency of each weekday as the corresponding p_i for the sake of preciseness. The number of each day from 2015 to 2019 is shown below:

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday	Total
Total number of days	261	261	260	261	261	261	261	1826

Table 1: Number of each weekdays from January 1st 2015 to December 31st 2019

According to Table 1, we set the null hypothesis to be

$$H_0\text{: the data follow a multinomial distribution}$$
$$\text{with parameters } (p_1,...,p_7) = (\tfrac{261}{1826}, \tfrac{261}{1826}, \tfrac{260}{1826}, \tfrac{261}{1826}, \tfrac{261}{1826}, \tfrac{261}{1826}, \tfrac{261}{1826})$$

Then, we calculate the expected frequencies of occurrence of police shootings using $E_i = np_i = 4938p_i$, shown in Table 2. It turns out that the expected E_is satisfy the Cochran's rule, implying that the sample size n is sufficiently large to apply a Pearson's test.

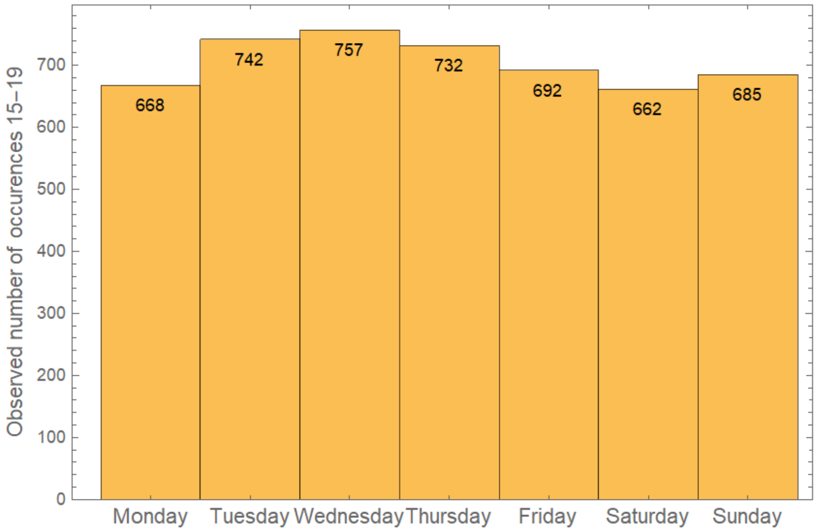


Figure 2: Number of police shootings on different Weekdays between January 1^{st} 2015 and December 31^{st} 2019

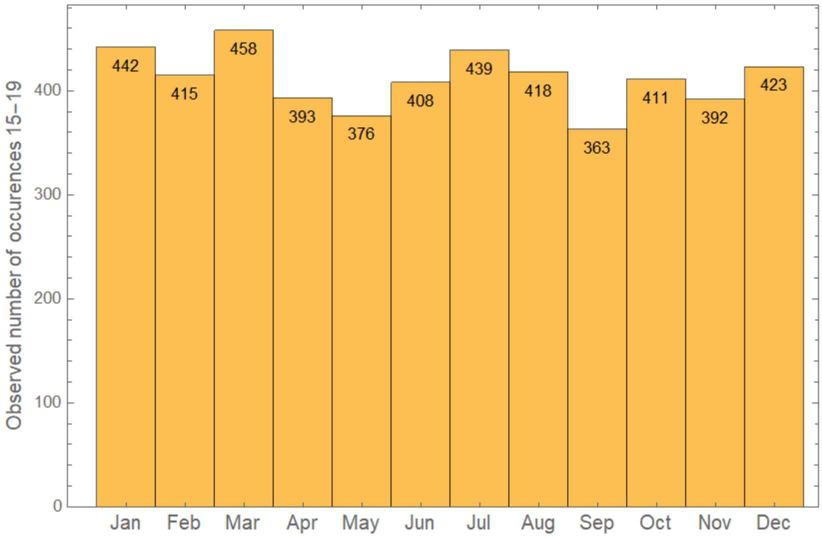


Figure 3: Number of police shootings on different months between January 1^{st} 2015 and December 31^{st} 2019

Based on the expected occurrence (E_i) we calculated and the observed occurrence (O_i) obtained from the data listed in Table 2, we calculate the test statistic as follows:

$$X^2_{k-1} = X^2_{7-1} = \sum_{i=1}^{K} \frac{(X_i - np_{i_0})^2}{np_{i_0}} = \sum_{i=1}^{7} \frac{(O_i - E_i)^2}{E_i} \approx 12.586. \tag{1}$$

This statistic follow a chi-squared distribution with 7-1 = 6 degrees of freedom.

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Observed occurrence (O_i)	668	742	757	732	692	662	685
Expected occurrence (E_i)	705.81	705.81	703.11	705.81	705.81	705.81	705.81

Table 2: Observed and expected numbers of occurrence for different weekdays

According to the table for cumulative chi-squared distribution [2], $\chi_{0.05,6} = 12.592$. Since $12.586 > 12.592$, the p-value of the test is larger than 5%. Therefore, we may not decide to reject H_0 at the 5% level of significance and conclude that there is no enough evidence that the average number of police shootings depends on the weekdays. However, since the critical value 12.592 is really close to our test statistic 12.586, further verification of the conclusion needs to be performed.

4 Goodness-of-Fit Test for the Poisson Distribution

In this section, utilizing a goodness-of-fit test for a Poisson distribution, we are going to investigate whether the occurrence of police shootings in the US has followed a Poisson distribution in the years 2015-2019.

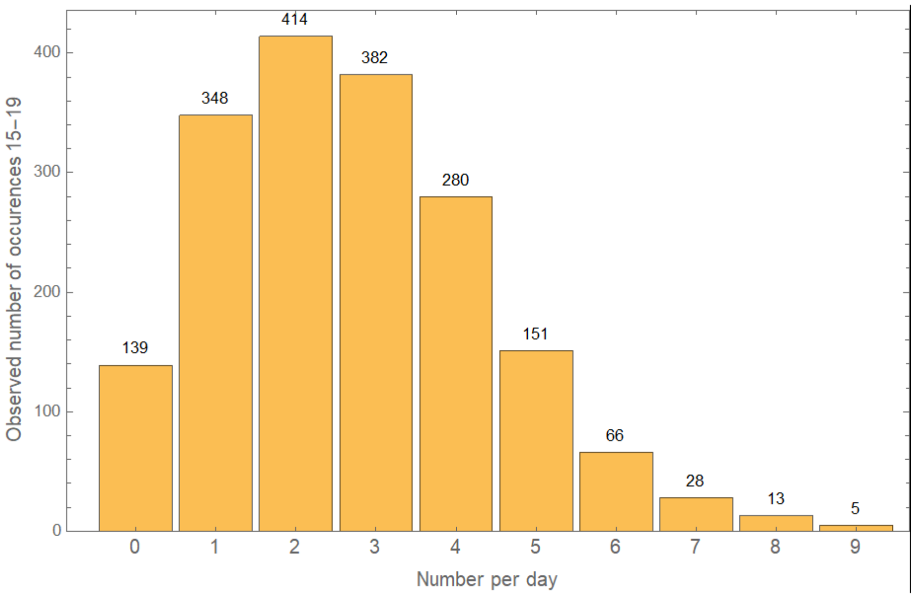


Figure 4: Observed number of days with different numbers of police shootings between January 1st 2015 and December 31st 2019

From the data, we count how many times of police shootings occur a day, and obtain the observed number of days that has a certain number of occurrence of police shootings from 2015 to 2019, shown in Figure 4 and Table 4. Let X denote the number of police shootings a day. The total number of days from 2015 to 2019 is calculated as n = 5 × 365 + 1 = 1826.

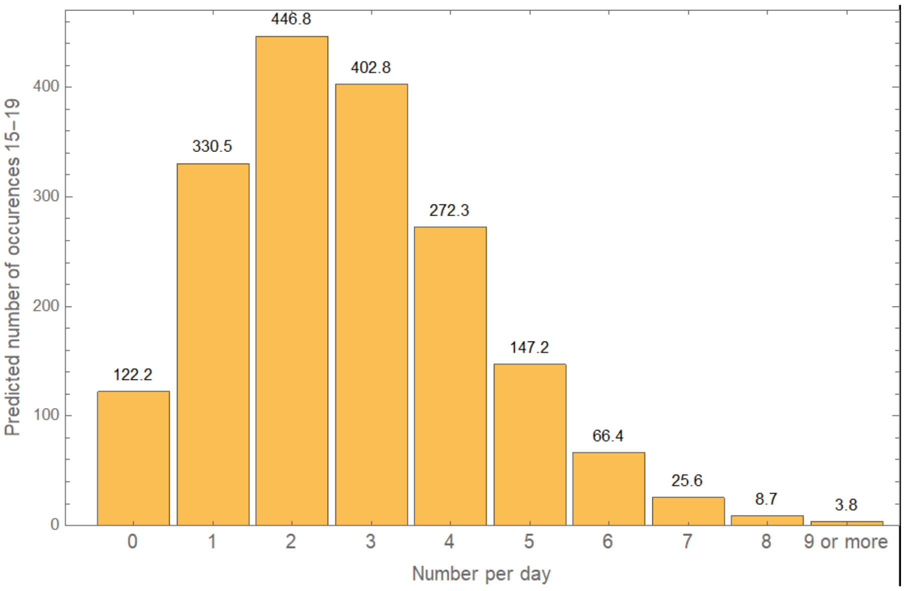


Figure 5: Expected number of days with different numbers of police shootings between January 1^{st} 2015 and December 31^{st} 2019

A maximum-likelihood estimator for the parameter k of the assumed Poisson distribution is given by the sample mean. So we get

$$\widehat{k} = \bar{X} = \frac{\sum_{i=0}^{9} X_i O_i}{n} = \frac{4938}{1826} \approx 2.704. \tag{2}$$

We set the null hypothesis as

H_0: the occurrence of police shootings in the US has followed a Poisson distribution with parameter $k = 2.704$.

To apply the multinomial distribution, we calculate each expected probability using

$$P[X = i] = \frac{e^{-\widehat{k}}\widehat{k}^i}{i!} \text{for } i = 1, 2, 3, ..., 8 \text{ and}$$
$$P[X \geq 9] = 1 - P[X = 1] - P[X = 2] - \cdots - P[X = 8]$$

The expected frequencies are calculated as

$$E_i = np[X = i].$$

The results are shown in Table 3 and Figure 5. The Cochran's rule is obeyed here, indicating that the sample size n is sufficiently large so that we can apply the Pearson's test.

Category i	0	1	2	3	4	5	6	7	8	9
X	0	1	2	3	4	5	6	7	8	≥9
O_i	139	348	414	382	280	151	66	28	13	5
P[X= i]	0.0669	0.1810	0.2447	0.2206	0.1491	0.0806	0.0363	0.0140	0.0047	0.0021
E_i	122.2	330.5	446.8	402.8	272.3	147.2	66.4	25.6	8.7	3.8

Table 3: Observed and expected numbers of days with different number of police shootings

For $N = 10$ categories, the statistic

$$X^2 = \sum_{i=1}^{10} \frac{(O_i - E_i)^2}{E_i} \approx 9.77 \tag{3}$$

follows a chi-squared distribution with N - 1 - m = 10 - 1 - 1 =8 degrees of freedom.

From the table[2], when $\alpha = 0.05$, we have $\chi_{0.05,8} = 15.507$. Since $9.77 < 15.507$, we are unable to reject H_0 at 5% level of significance.

According to table[3], $\chi_{0.5,8} = 7.34$ and $\chi_{0.25,8} = 10.2$. Since $9.77 \in (7.34, 10.2)$, the P-value of the test is approximately 0.29. Then it's reasonable for us to conclude that there is no enough evidence that the occurrence of police shootings in the US has not followed a Poisson distribution with parameter $k \approx 2.704$ in the years 2015-2019.

5 Confidence Interval for Parameter k of a Poisson Distribution

5.1 Formula of a Confidence Interval for k

In this section, we are going to show that a (1 - α)100% confidence interval for parameter k of a Passion distribution can be given by

$$\hat{k} \pm z_{\alpha/2}\sqrt{\hat{k}/n}. \tag{4}$$

Let X_1, X_2,..., X_n be a random sample of X with size n from a Poisson distribution with parameter k. As the random variable X follows a Poisson distribution, E$[X] = k$, Var$[X] = k$ and $\hat{k} = \overline{X}$. Since the sample size n in our case is large enough, by the central limit theorem, $\overline{X}$ is approximately normally distributed with mean k and variance $\frac{k}{n}$. Hence,

$$\frac{\hat{k} - k}{k/n}$$

is approximately standard-normally distributed. It follows that a (1 - α)100% confidence interval for k is given by:

$$\hat{k} \pm z_{\alpha/2}\sqrt{k/n}. \tag{5}$$

The interval of Equation 5 depends on the unknown parameter k, which we are actually trying to estimate. Thus, we replace k by $\hat{k}$. Though, the number $z_{\alpha/2}$ is no longer accurate, the interval is still valid and reasonable because our sample size n is large enough to allow the central limit theorem to hole and then the difference between $z_{\alpha/2}$ and the a correct value is negligible. Hence, we obtain that when the sample size is sufficiently large, a (1 - α)100% confidence interval for parameter k of a Passion distribution can be given by

$$\hat{k} \pm z_{\alpha/2}\sqrt{\hat{k}/n}. \tag{6}$$

5.2 Confidence Interval for k of the Poisson Distribution of Police Shootings from 2015 to 2019

We have obtained that $\hat{k} = 2.704$ for the police shootings from 2015 to 2019. Besides, the sample size $n = 365 \times 5 + 1 = 1826$. Thus, we can calculate a 95% confidence interval for k based on the data of the years 2015 to 2019 by Equation 6, i.e.

$$2.704 \pm 1.96\sqrt{2.704/1826} \approx 2.704 \pm 0.075. \tag{7}$$

Therefore, the 95% confidence interval for k based on the data of the years 2015 to 2019 is (2.629, 2.779).

6 Estimates and Predictions for Police Shootings in 2020

In this section, our aim is giving a prediction for occurrence of fatal police shootings in 2020 based on data from 2015 to 2019, where the occurrence follows the Poisson distribution model as previous sections. Also, we introduce the Nelson's Formula for prediction interval[5] and apply the given data to derive a 95% prediction intervals and plots.

6.1 Fatal Police Shootings in 2020

We have checked that the occurrence of shootings from 2015 to 2019 follows a Poisson distribution. Here we are going to test whether the shooting data recorded so far also follow Poisson distribution:

H_0: The occurrence of police shootings in 2020 follows a Poisson distribution.

Given the data, we count the number of daily occurrence of shootings from January 1^{st} to April 15^{th} 2020, where the overall number of days is:

$$n = 106$$

We get the following table:

Occurrence of shootings X	0	1	2	3	4	5	6	7	8 or more
Observed Number of days O_i	7	18	27	18	14	13	8	0	1

Table 4: Observed occurrence per day from January $1^{st}$to April 15^{th} 2020

And the estimator $\hat{k}$ can be calculated as:

$$\hat{k} = \overline{X} = \frac{\sum_{i=0}^{8} X_i O_i}{n} \approx 2.86$$

Hence we can calculate the expected occurrence with the formula of Poisson distribution:

$$f(x) = P[X = x] = \frac{\hat{k}^x}{x!} e^{-\hat{k}} = \frac{2.86^x}{x!} e^{-2.86}$$

So expected number of days with zero occurrence:

$$P[X = 0] = \frac{2.86^0}{0!} e^{-2.86} \approx 6.08$$

Similarly, we can have following table for expected number of days E_i:

Occurrence of shootings X	0	1	2	3	4	5	6	7	8 or more
Expected Number of days E_i	6.08	17.38	24.84	23.67	16.91	9.67	4.61	1.88	0.97

Table 5: Expected occurrence per day from January $1^{st}$to April 15^{th} 2020

Applying Pearson's statistics,

$$\chi^2_{N-1-m} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where $N-1-m = 9-1-1 = 7$, hence we can calculate that $\chi^2_7 \approx 7.74$. Checking the table of Cumulative Chi-squared Distribution[2], we have $\chi^2_{0.05,7} = 14.1$. Since

$$\chi^2_{N-1-m} < \chi^2_{0.05,7},$$

we are unable to reject H_0 at 95% level of significance.
Hence we draw a conclusion: **there is no evidence that the model of Poisson distribution doesn't work for occurrence of shootings in 2020.**

6.2 Nelson's Formula and Prediction Intervals

6.2.1 Nelson's Formula and Nelson's Prediction Interval

Here we are going to derive the Nelson's prediction interval formula.

Nelson's Formula: Let X be the total counts in a sample of size n from a Poisson distribution with mean λ. Note that $X \sim Poisson(n\lambda)$. Y denote the future total counts that can be observed in a sample of size m from the same Poisson distribution so that $Y \sim Poisson(m\lambda)$.
So the 100(1-α)% Nelson prediction interval is:

$$[\lceil L \rceil, \lfloor U \rfloor]$$

where

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

The proof is as follow.

Estimators: We define estimators for λ and Y:

$$\hat{\lambda} = \frac{X}{n}$$

$$\hat{Y} = \begin{cases} \frac{mX}{n} = m\widehat{\lambda} & X = 1, 2, \cdots \\ \frac{m}{2n} & X = 0 \end{cases}$$

Since

$$\mathrm{E}[\widehat{\lambda}] = \frac{n\lambda}{n} = \lambda$$

$$Y - \mathrm{E}[\widehat{Y}] = m\lambda - \frac{m}{n}\mathrm{E}[X] = m\lambda - \frac{m}{n} \cdot n\lambda = 0$$

for $X > 0$, hence both estimators are unbiased.

Proof: Hence we are going to proof Nelson's formula. According to properties of Poisson distribution, we have

$$\mathrm{E}[X] = n\lambda \quad \mathrm{E}(Y) = m\lambda$$

$$\mathrm{Var}(X) = n\lambda \quad \mathrm{Var}(Y) = m\lambda$$

So we deduce further for mean and variance of $m\widehat{\lambda} - Y$:

$$\mathrm{E}(m\widehat{\lambda} - Y) = m\lambda - m\lambda = 0$$

And since X, Y are independent, i.e. $\mathrm{Cov}(X, Y) = 0$, we have

$$\begin{aligned}\mathrm{Var}(m\widehat{\lambda} - Y) &= \mathrm{Var}\left(\frac{m}{n}X - Y\right) \\ &= \frac{m^2}{n^2}\mathrm{Var}(X) + \mathrm{Var}(Y) \\ &= \frac{m^2}{n^2} \cdot n\lambda + m\lambda \\ &= m^2\lambda\left(\frac{1}{n} + \frac{1}{m}\right)\end{aligned}$$

So that we can get:

$$\widehat{\mathrm{Var}}(m\widehat{\lambda} - Y) = m^2\widehat{\lambda}(1/n + 1/m)$$

According to Central Limit Theorm, we have

$$\frac{m\widehat{\lambda} - Y}{\sqrt{\widehat{\mathrm{Var}}(m\widehat{\lambda} - Y)}}$$

follows a standard normal distribution. So a $100(1-\alpha)\%$ prediction interval is given by

$$\begin{aligned}
1-\alpha &= P\left[-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right] \\
&= P\left[-z_{\alpha/2} \leq \frac{m\widehat{\lambda}-Y}{\sqrt{\operatorname{var}(m\widehat{\lambda}-Y)}} \leq z_{\alpha/2}\right] \\
&= P\left[-z_{\alpha/2}\sqrt{\operatorname{var}(m\widehat{\lambda}-Y)}+m\widehat{\lambda} \leq Y \leq z_{\alpha/2}\sqrt{\operatorname{var}(m\widehat{\lambda}-Y)}+m\widehat{\lambda}\right] \\
&= P\left[-z_{\alpha/2}\sqrt{m\widehat{Y}\left(\frac{1}{m}+\frac{1}{n}\right)}+\widehat{Y} \leq Y \leq z_{\alpha/2}\sqrt{m\widehat{Y}\left(\frac{1}{m}+\frac{1}{n}\right)}+\widehat{Y}\right]
\end{aligned}$$

i.e.

$$\widehat{Y} \pm z_{\alpha/2}\sqrt{m\widehat{Y}\left(\frac{1}{m}+\frac{1}{n}\right)}$$

which is valid with the assumption of large sample size. For the police shooting cases, the occurrence is supposed to be integer, so we get the upper and lower bound for the prediction interval, i.e.

$$[\lceil L\rceil, \lfloor U\rfloor] \; with \; [L,U] = \widehat{Y} \pm z_{\alpha/2}\sqrt{m\widehat{Y}\left(\frac{1}{m}+\frac{1}{n}\right)}$$

Hence the Nelson's formula is proved.

6.2.2 Application to Fatal Police Shooting Data

Given the formula, we are allowed to derive the prediction interval of the number of shooting occurrence in 2020. In the police shooting cases, some variables need to be clarified:

X: The number of shooting occurrence from 2015 to 2019.

$$X = 4938$$

Y: The number of shooting occurrence from 2015 considering 2020.

n: The number of days between 2015 and 2019.

$$n = 365 \times 5 + 1 = 1826$$

m: The number of days considering 2020.

According to Nelson's formula, we have

$$\widehat{Y} = \frac{mX}{n} \approx 2.704m$$

Set $\alpha = 0.05$ so that $z_{\alpha/2} = 1.96$, applying the Nelson's formula:

$$L = \widehat{Y} - z_{\alpha/2}\sqrt{m\widehat{Y}\left(\frac{1}{m}+\frac{1}{n}\right)} = 2.704m - \sqrt{10.3876864m + 0.005688766m^2}$$

$$U = \widehat{Y} + z_{\alpha/2}\sqrt{m\widehat{Y}\left(\frac{1}{m}+\frac{1}{n}\right)} = 2.704m + \sqrt{10.3876864m + 0.005688766m^2}$$

so the 95% prediction interval for the number of police shootings in 2020 is $[\lceil L\rceil, \lfloor U\rfloor]$.
We plot the prediction interval of cumulative shooting occurrence number in 2020 as shown below:

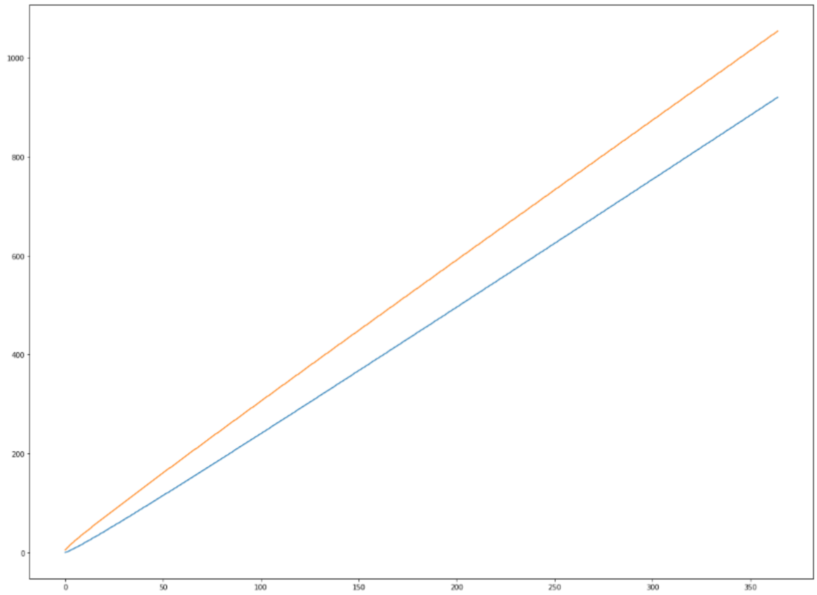


Figure 6: Prediction Interval for 2020

Hence we have 95% confidence to say that the true result will fall within this interval.
To check the correctness of the Nelson's formula as well as prediction interval, the observed number of shooting occurrence (form Jau 1^{st} to Apr 15^{th}) is included:

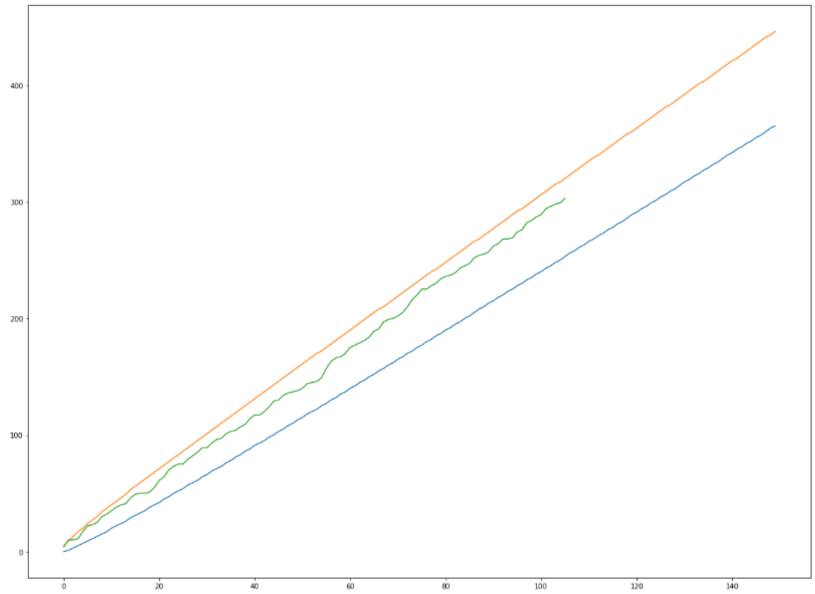


Figure 7: Prediction Interval with observed data in 2020

where the x-axis represents the number of days in 2020 and the y-axis represents the cumulative number of shooting occurrence in this year. From the graph shown above, we can see that the green line, which represents the observed shooting occurrence so far, falls mostly within the prediction interval. So we can confidently draw a conclusion that: **The Nelson's formula provides an acceptable prediction interval for future estimation.**

6.3 Comment on the impact of Coronavirus

The outbreak of novel coronavirus in 2020 has a huge impact on the world. We are going to look into whether the fatal police shootings in 2020 is influenced by the coronavirus.
Following are some possible point according to the data of 2020 so far:

i) Comparing the estimator $\hat{k}$ in 2020 with past years,we have

$$\hat{k}_{2020} = 2.86 > \hat{k}_{2015-2019} = 2.7$$

There are two possible explanation towards this phenomenon: Firstly, we calculate $\hat{k}$ by only data from Jau 1^{st} to Apr 15^{th} 2020. $\hat{k}$ may be closer to the real k (approximately as $\hat{k}_{2015-2019}$) with more occurrence observed and time passing; Secondly, since $\hat{k}$ represents the expectation of shooting occurrence, we can conclude that the outbreak of coronavirus increase the shooting occurrence to some extent. The inner reason for this phenomenon requires further study in sociology , psychology and etc. For example, the police may be nervous about the spread of coronavirus, which increases the tendency for fatal police shootings.

ii) Even though the coronavirus causes a higher $\hat{k}$, it still lies within the prediction interval. So we can confidently conclude that the outbreak of coronavirus does not make serious affect on the Poisson distribution model. We can still use the Poisson distribution to give a proper prediction on future occurrence of shootings.

7 Conclusion

In this project, we analyze the Police Shooting data to derive a proper model for shooting occurrence, using which we are allowed to perform future prediction.

We first give an overview for shooting occurrence by plotting the daily occurrence from 2015 to 2019 and further investigate that shooting occurrence is independent on weekdays. Then we find that there is no evidence to reject the model of Poisson distribution with $k_{2015-2019} = 2.704$ for daily occurrence. Also, with the confidence interval formula for k, we have 95% confidence to say that the parameter k in Poisson distribution for data in 2015-2019 falls into (2.629,2.779). Given the model and past data, we predict the shooting occurrence in 2020 and successfully find out it follows the Poisson distribution with $k_{2020} = 2.86$.

Applying the Nelson's formula, we calculate the prediction interval for shooting occurrence in 2020, which is approximately $[\lceil 2.7m - \sqrt{10.4m + 0.057m^2}\rceil, \lfloor 2.7m + \sqrt{10.4m + 0.057m^2}\rfloor]$ with 95% confidence to state that the real occurrence will fall within. We further plot the prediction interval with given data in 2020 to show that the Nelson's formula provides an acceptable prediction interval for future estimation. Finally, we give possible explanation on the relation between shooting occurrence and the outbreak of coronavirus in 2020.

8 Reference

1 The Washington Post. Fatal Force. https://www.washingtonpost.com/graphics/2019/national/police-shootings-2019/. Web. Accessed April 28th, 2020

2 (n.d.). Retrieved from https://people.richland.edu/james/lecture/m170/tbl-chi.html

3 Dinov, I. (n.d.). Retrieved from http://socr.ucla.edu/Applets.dir/ChiSquareTable.html

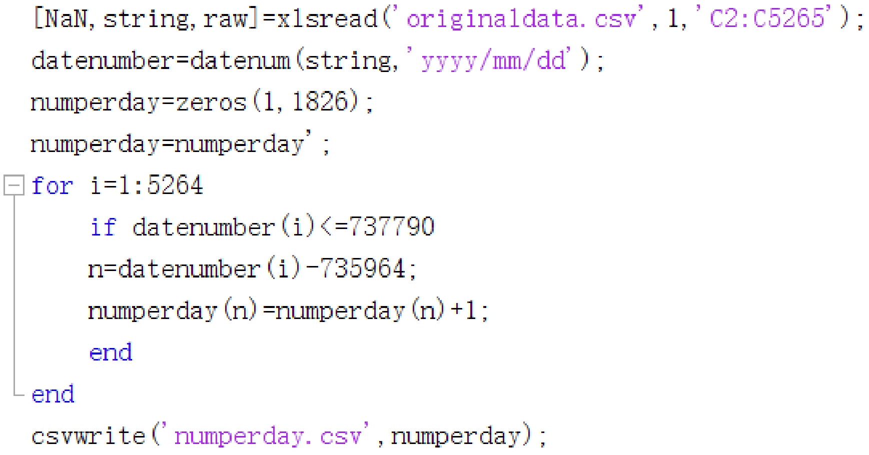
4 D. Spiegelhalter and A. Barnett. London murders: a predictable pattern? *Significance*, 6(1):5–8, 2009. http://onlinelibrary.wiley.com/doi/10.1111/j.1740-9713.2009.00334.x/abstract.

5 K.Krishnamoorthy and J.Peng. Improved closed-form prediction intervals for binomial and Poisson distributions. *Journal of Statistical Planning and Inference*,141(5):1709–1718,2011. http://www.sciencedirect.com/science/article/pii/S0378375810005215.

6 Washingtonpost. (2020, April 28). washingtonpost/data-police-shootings. Retrieved from https://github.com/washingtonpost/data-police-shootings

9 Appendix

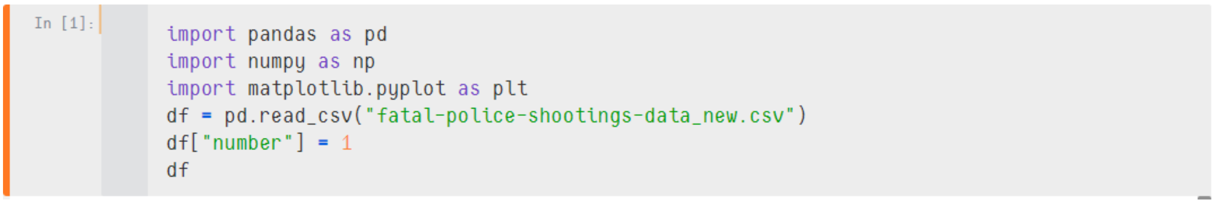
9.1 Matlab Code for Data Processing



```
[NaN,string,raw]=xlsread('originaldata.csv',1,'C2:C5265');
datenumber=datenum(string,'yyyy/mm/dd');
numperday=zeros(1,1826);
numperday=numperday';
for i=1:5264
    if datenumber(i)<=737790
    n=datenumber(i)-735964;
    numperday(n)=numperday(n)+1;
    end
end
csvwrite('numperday.csv',numperday);
```

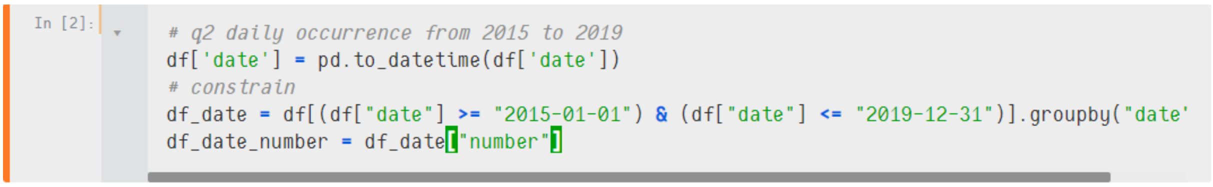
9.2 Python Code for Data Processing

9.2.1 Import data



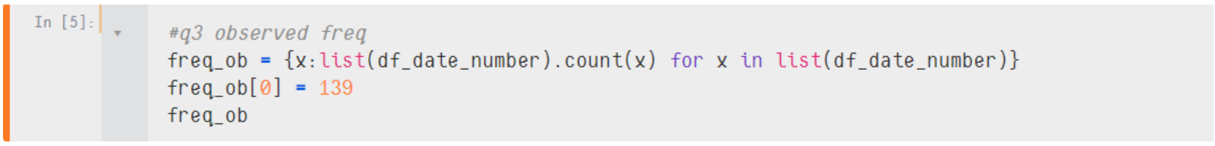
```
In [1]:
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
df = pd.read_csv("fatal-police-shootings-data_new.csv")
df["number"] = 1
df
```

9.2.2 Daily occurrence from 2015 to 2019

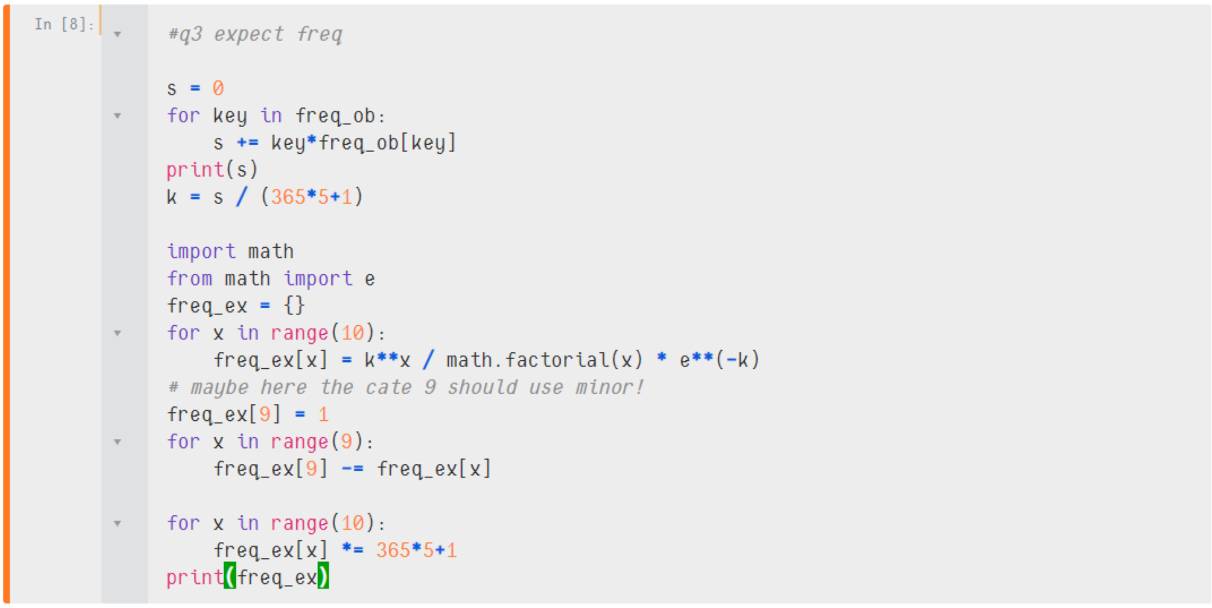


```
In [2]:
# q2 daily occurrence from 2015 to 2019
df['date'] = pd.to_datetime(df['date'])
# constrain
df_date = df[(df["date"] >= "2015-01-01") & (df["date"] <= "2019-12-31")].groupby("date'
df_date_number = df_date["number"]
```

9.2.3 Observed and expected occurrence from 2015 to 2019



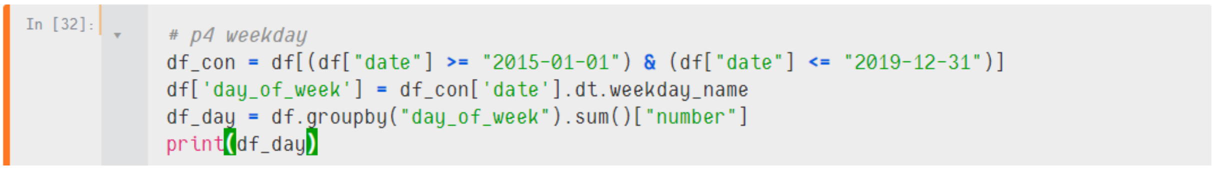
```
In [5]:
#q3 observed freq
freq_ob = {x:list(df_date_number).count(x) for x in list(df_date_number)}
freq_ob[0] = 139
freq_ob
```



In [8]:

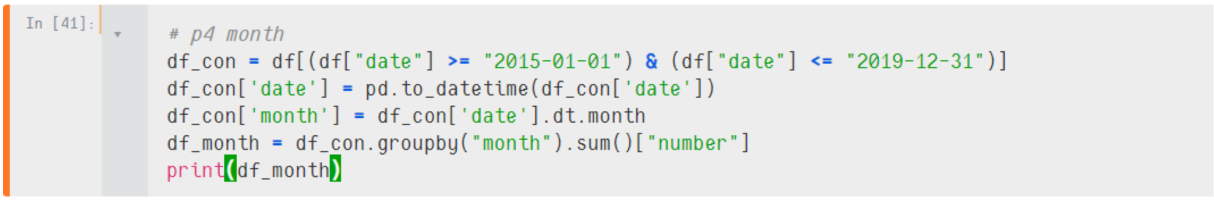
```
#q3 expect freq

s = 0
for key in freq_ob:
    s += key*freq_ob[key]
print(s)
k = s / (365*5+1)

import math
from math import e
freq_ex = {}
for x in range(10):
    freq_ex[x] = k**x / math.factorial(x) * e**(-k)
# maybe here the cate 9 should use minor!
freq_ex[9] = 1
for x in range(9):
    freq_ex[9] -= freq_ex[x]

for x in range(10):
    freq_ex[x] *= 365*5+1
print(freq_ex)
```

9.2.4 Weekly and monthly occurrence from 2015 to 2019



In [32]:

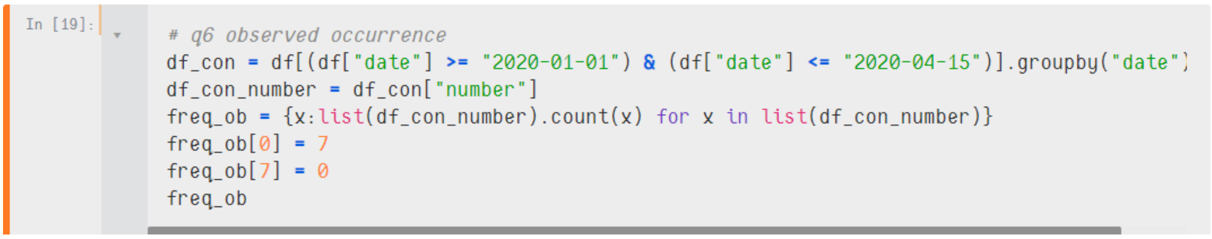
```
# p4 weekday
df_con = df[(df["date"] >= "2015-01-01") & (df["date"] <= "2019-12-31")]
df['day_of_week'] = df_con['date'].dt.weekday_name
df_day = df.groupby("day_of_week").sum()["number"]
print(df_day)
```



In [41]:

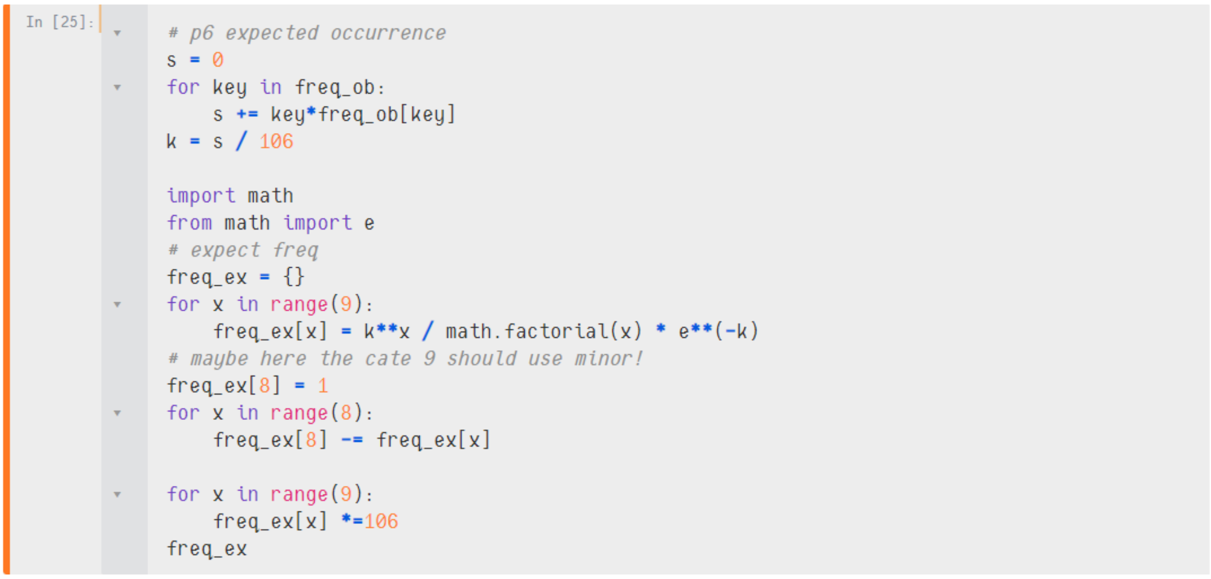
```
# p4 month
df_con = df[(df["date"] >= "2015-01-01") & (df["date"] <= "2019-12-31")]
df_con['date'] = pd.to_datetime(df_con['date'])
df_con['month'] = df_con['date'].dt.month
df_month = df_con.groupby("month").sum()["number"]
print(df_month)
```

9.2.5 Observed and expected occurrence in 2020



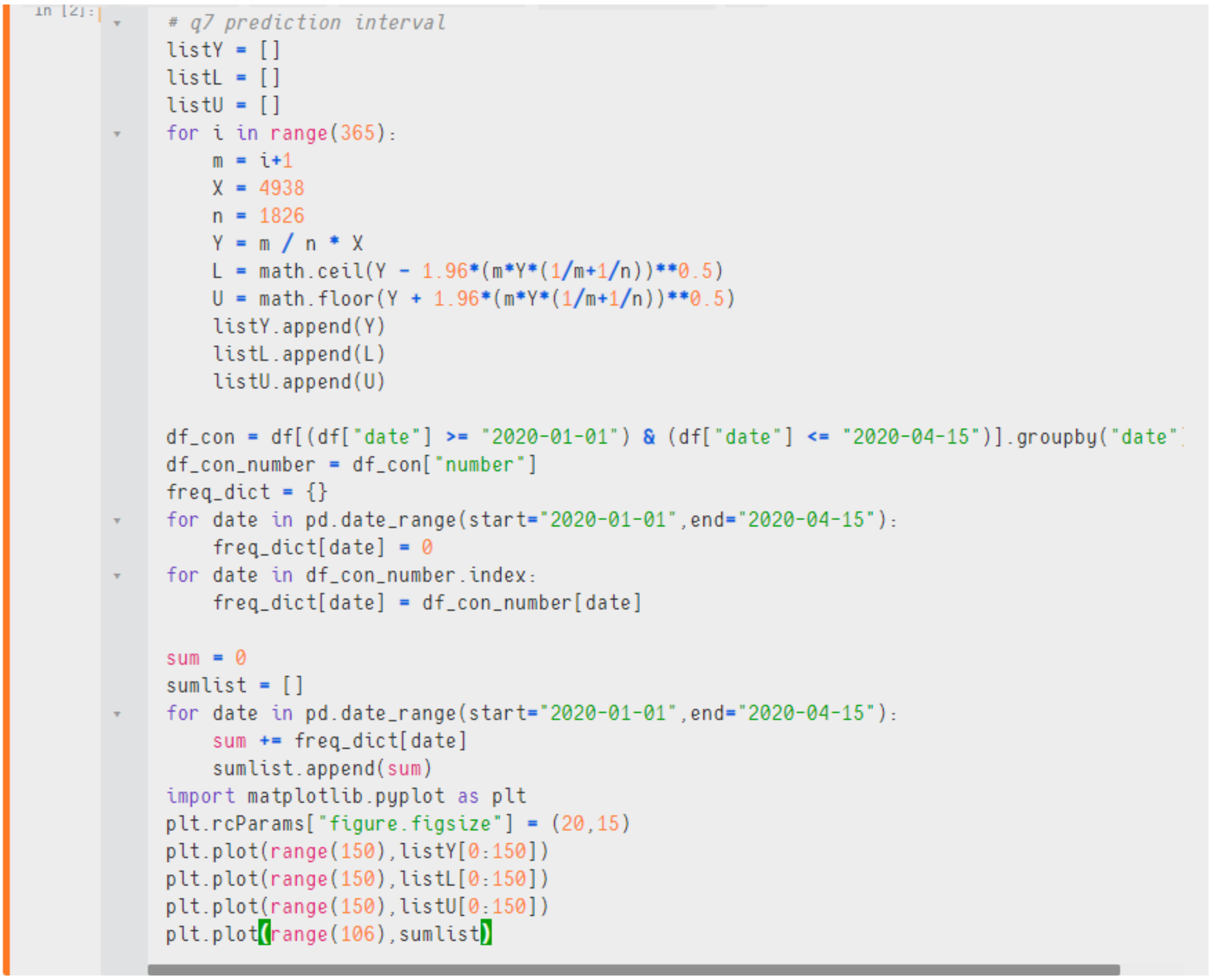
In [19]:

```
# q6 observed occurrence
df_con = df[(df["date"] >= "2020-01-01") & (df["date"] <= "2020-04-15")].groupby("date"
df_con_number = df_con["number"]
freq_ob = {x:list(df_con_number).count(x) for x in list(df_con_number)}
freq_ob[0] = 7
freq_ob[7] = 0
freq_ob
```



```
In [25]:
# p6 expected occurrence
s = 0
for key in freq_ob:
    s += key*freq_ob[key]
k = s / 106

import math
from math import e
# expect freq
freq_ex = {}
for x in range(9):
    freq_ex[x] = k**x / math.factorial(x) * e**(-k)
# maybe here the cate 9 should use minor!
freq_ex[8] = 1
for x in range(8):
    freq_ex[8] -= freq_ex[x]

for x in range(9):
    freq_ex[x] *=106
freq_ex
```

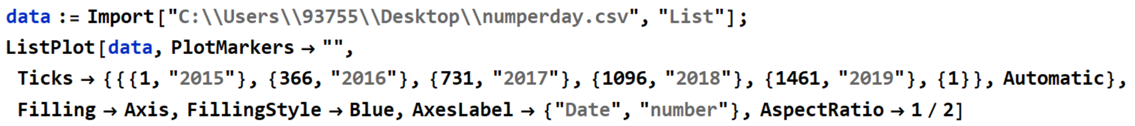
9.2.6 Prediction interval for 2020



```
In [2]:
# q7 prediction interval
listY = []
listL = []
listU = []
for i in range(365):
    m = i+1
    X = 4938
    n = 1826
    Y = m / n * X
    L = math.ceil(Y - 1.96*(m*Y*(1/m+1/n))**0.5)
    U = math.floor(Y + 1.96*(m*Y*(1/m+1/n))**0.5)
    listY.append(Y)
    listL.append(L)
    listU.append(U)

df_con = df[(df["date"] >= "2020-01-01") & (df["date"] <= "2020-04-15")].groupby("date"
df_con_number = df_con["number"]
freq_dict = {}
for date in pd.date_range(start="2020-01-01",end="2020-04-15"):
    freq_dict[date] = 0
for date in df_con_number.index:
    freq_dict[date] = df_con_number[date]

sum = 0
sumlist = []
for date in pd.date_range(start="2020-01-01",end="2020-04-15"):
    sum += freq_dict[date]
    sumlist.append(sum)
import matplotlib.pyplot as plt
plt.rcParams["figure.figsize"] = (20,15)
plt.plot(range(150),listY[0:150])
plt.plot(range(150),listL[0:150])
plt.plot(range(150),listU[0:150])
plt.plot(range(106),sumlist)
```

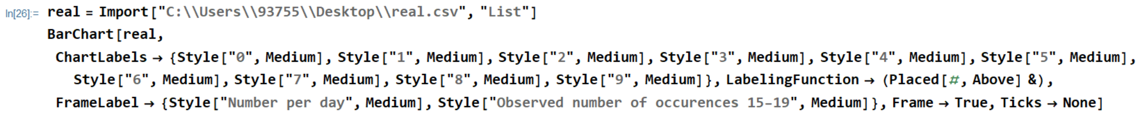
9.3 Mathematica Code for Plotting

9.3.1 Figure 1



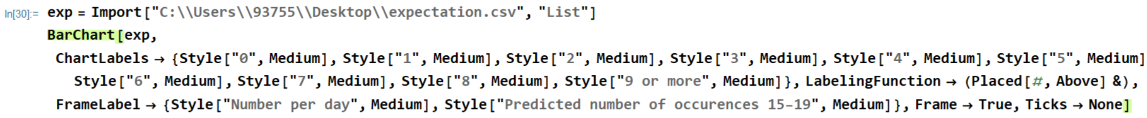
```
data := Import["C:\\Users\\93755\\Desktop\\numperday.csv", "List"];
ListPlot[data, PlotMarkers → "",
 Ticks → {{{1, "2015"}, {366, "2016"}, {731, "2017"}, {1096, "2018"}, {1461, "2019"}, {1}}, Automatic},
 Filling → Axis, FillingStyle → Blue, AxesLabel → {"Date", "number"}, AspectRatio → 1 / 2]
```

9.3.2 Figure 2



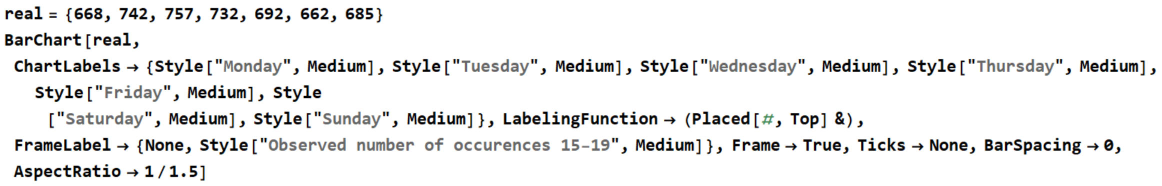
```
In[26]:= real = Import["C:\\Users\\93755\\Desktop\\real.csv", "List"]
         BarChart[real,
          ChartLabels → {Style["0", Medium], Style["1", Medium], Style["2", Medium], Style["3", Medium], Style["4", Medium], Style["5", Medium],
            Style["6", Medium], Style["7", Medium], Style["8", Medium], Style["9", Medium]}, LabelingFunction → (Placed[#, Above] &),
          FrameLabel → {Style["Number per day", Medium], Style["Observed number of occurences 15-19", Medium]}, Frame → True, Ticks → None]
```

9.3.3 Figure 3



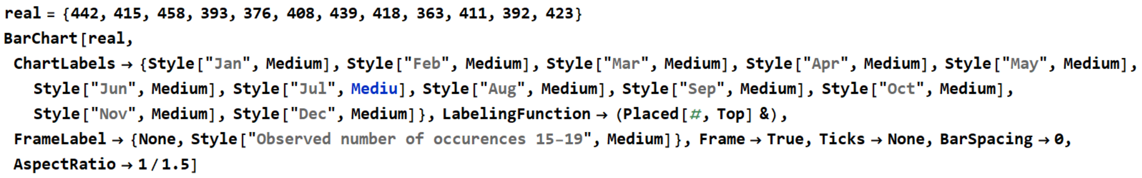
```
In[30]:= exp = Import["C:\\Users\\93755\\Desktop\\expectation.csv", "List"]
         BarChart[exp,
          ChartLabels → {Style["0", Medium], Style["1", Medium], Style["2", Medium], Style["3", Medium], Style["4", Medium], Style["5", Medium],
            Style["6", Medium], Style["7", Medium], Style["8", Medium], Style["9 or more", Medium]}, LabelingFunction → (Placed[#, Above] &),
          FrameLabel → {Style["Number per day", Medium], Style["Predicted number of occurences 15-19", Medium]}, Frame → True, Ticks → None]
```

9.3.4 Figure 4



```
real = {668, 742, 757, 732, 692, 662, 685}
BarChart[real,
 ChartLabels → {Style["Monday", Medium], Style["Tuesday", Medium], Style["Wednesday", Medium], Style["Thursday", Medium],
   Style["Friday", Medium], Style
    ["Saturday", Medium], Style["Sunday", Medium]}, LabelingFunction → (Placed[#, Top] &),
 FrameLabel → {None, Style["Observed number of occurences 15-19", Medium]}, Frame → True, Ticks → None, BarSpacing → 0,
 AspectRatio → 1 / 1.5]
```

9.3.5 Figure 5



```
real = {442, 415, 458, 393, 376, 408, 439, 418, 363, 411, 392, 423}
BarChart[real,
 ChartLabels → {Style["Jan", Medium], Style["Feb", Medium], Style["Mar", Medium], Style["Apr", Medium], Style["May", Medium],
   Style["Jun", Medium], Style["Jul", Mediu], Style["Aug", Medium], Style["Sep", Medium], Style["Oct", Medium],
   Style["Nov", Medium], Style["Dec", Medium]}, LabelingFunction → (Placed[#, Top] &),
 FrameLabel → {None, Style["Observed number of occurences 15-19", Medium]}, Frame → True, Ticks → None, BarSpacing → 0,
 AspectRatio → 1 / 1.5]
```