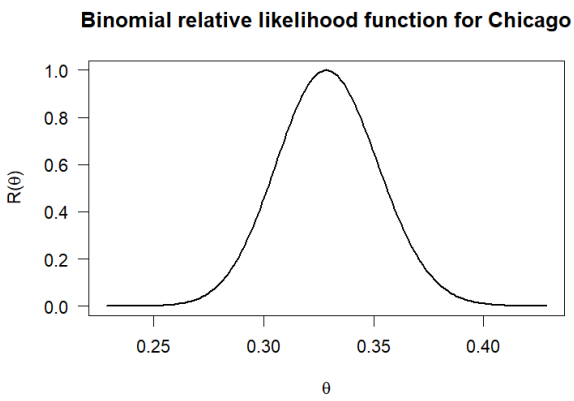
I explored subject sex and age, as well as subject race and vehicle make for the dataset. In particular, we will extend our analysis of the subject.sex variate by introducing a probability model and carrying out maximum likelihood estimation. I then explored subject.age, assessing the fit of various probability models and again employing maximum likelihood estimation. Finally, I combine the subject age variate with the subject race or vehicle make variate explored in file *Subject age variate*.

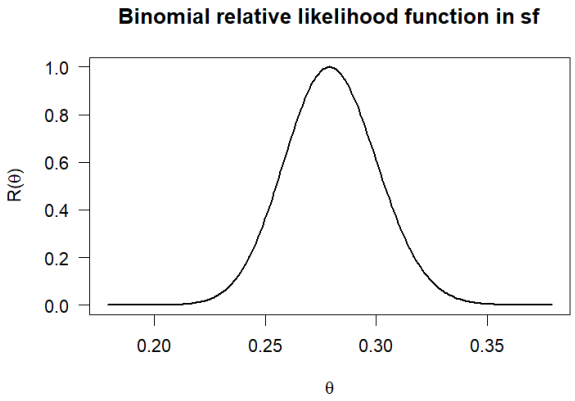
I have concerns about study error in this study. This is because the primary dataset from which the dataset was sampled is before 2024(Study population) which is different from the individual who could be the subject of a traffic stop in Chicago or San Francisco in the year 2024 (Target population). Since traffic conditions vary with traffic volume, number of people, season and time period.

The maximum likelihood estimate for Chicago is 0.329, while for San Francisco it is 0.279. These were calculated by m.l.e =y/n

137/417 130/466

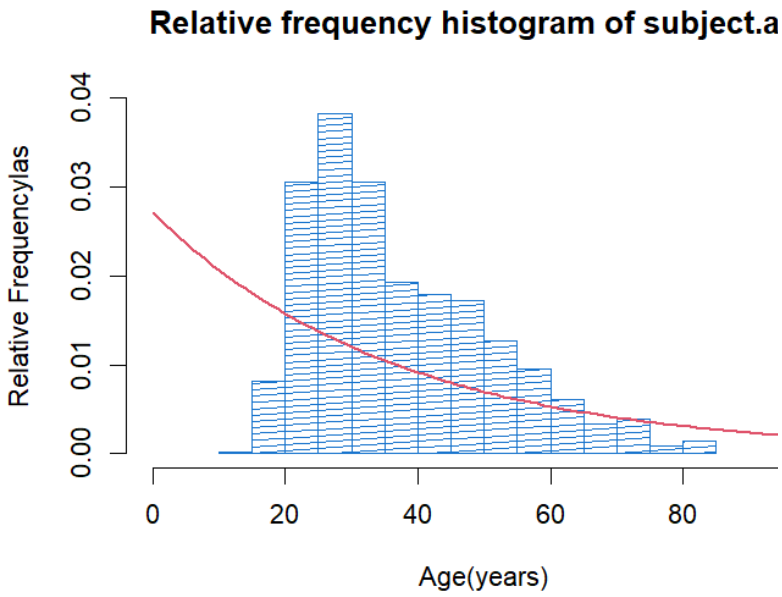
Relative likelihood function plots:



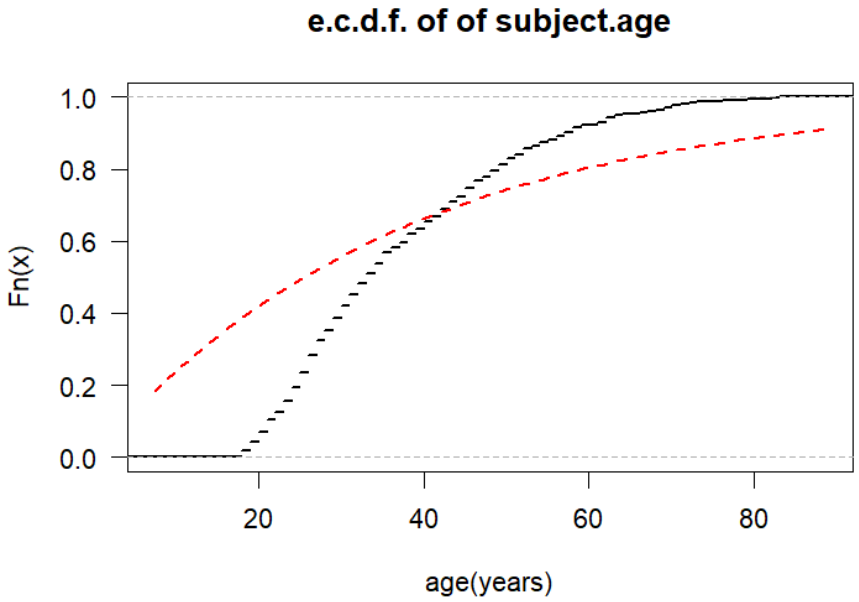



The sample size, mean, median, and standard deviation are, respectively, 883.000, 36.814, 33.000, 14.014.

Relative frequency histogram: (2b)




Empirical cumulative distribution function plot: (2c)



In 2b, we can get the mean=36.814, however, the standard deviation=14.014, which are not close to mean. Since the mean and standard deviation of  the random variable are not the same, we cannot generate Exponential model.

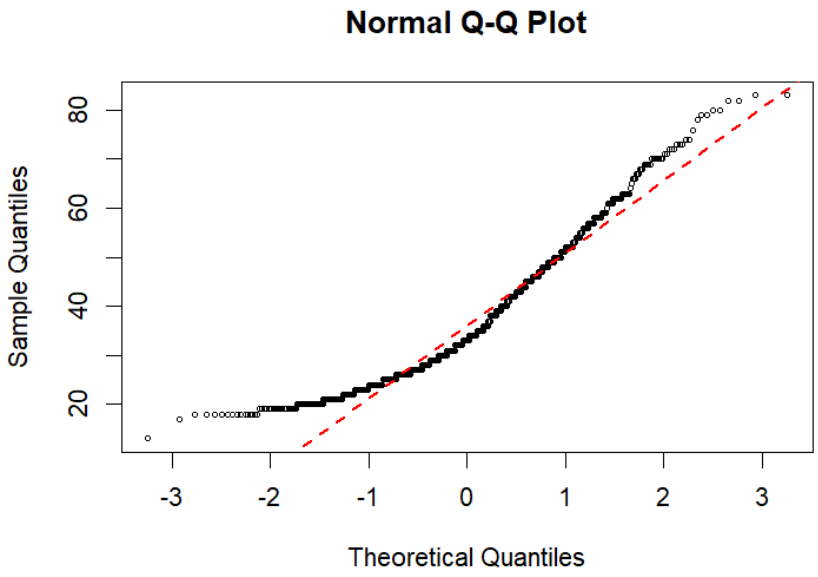
Based on Analysis 2b, median is 33.000. According to the image in 2d, when Fn(x)=0.5, age=23≠33. which does not match the real case.

Based on Analysis 2c, when age is around 28, the relative frequency is largest, mode ≈ 28, which means the data is most concentrated at age=28. Based on the image in 2d, the red line has the highest growth rate at age<10, which means for data generated from an Exponential distribution, we would expect to see the highest relative frequency occur at age<10.

Based on the plot in Analysis 2d, we can see that the image starts to rise gradually from around age=18, with the rate of increase slowing down as age increases. At around age=80 Fn(x) reaches 1 and remains flat. While for data generated from an Exponential distribution, we would expect to see that the image rises slowly from age 0 to 100, with the rate of increase slowing down much less than in the actual image, and at age=85, Fn(x) does not reach 1. does not reach 1.

Overall, the Exponential model doesn't fit well.

Q-Q plot: (2f)




Based on the Q-Q plot, the points do not appear to lie reasonably along a straight line. This indicates non-normal. The QQ plot appears U-shaped, suggesting asymmetry. There are more points in the right tails, suggesting positive skewness.

I do not have concerns about sample error in this study. This is because the target population is the age of a subject in a traffic stop chosen at random from the study population. So, there are no difference in attributes between the study population and sample.

The maximum likelihood estimate is 36.814, which was found by Pomle <- mean.

By invariance property of m.l.e, the maximum likelihood estimate of P(X⩽30) can be calculated by substituting the $\theta$ *hat* into $\theta$.

So, this was found by R code: probabilitymle <- ppois(30, Pomle).

```
#3d
probabilitymle <- ppois(30, Pomle)
probabilitymle
```

R(39) = 1.149788e-24. Based on this, we can say that R(39) is very close to 0. So, the relative likelihood function for 39 based on my sample is very implausible.

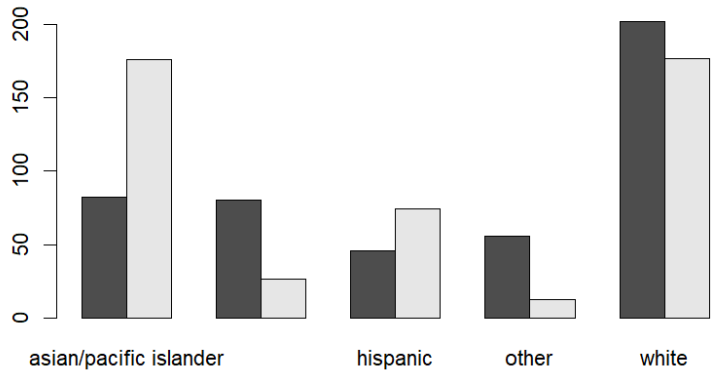
I will be analyzing the subject.race variate for San Francisco below.

I do have concerns about measurement error in the subject.race variate. This is because the police may identify the subject.race wrong, which cause the number of the race different from actual value.

Table of observed and expected frequencies:

| Race | Observed Frequency | Expected Frequency |
|---|---|---|
| Asian/Pacific Islander | 82 | 176.148 |
| Black | 80 | 26.562 |
| Hispanic | 46 | 74.094 |
| White | 202 | 176.614 |
| Other | 56 | 12.582 |

Grouped bar-plot of observed and expected frequencies:



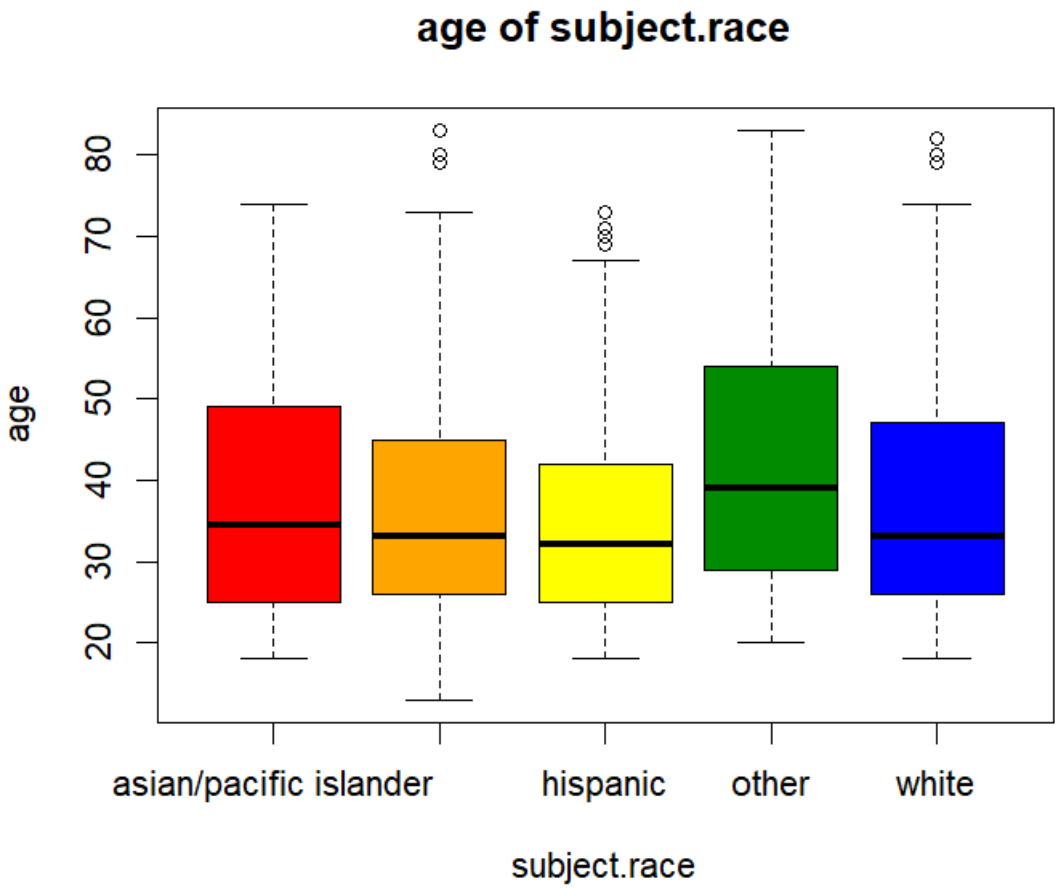

Based on the results of Analyses 4c and 4d, we noticed that in the observed data, whites make up the largest portion of the total at 200, far outnumbering the other three races. In the expected data, however, asian/pacific islander and white both have larger shares, both around 175, and far outnumber the other three races.

In asian/pacific islander, there is also a big difference between observed data and expected data. Observed data shows that the frequency of asian/pacific islander is about 80, while expected frequency is more than twice as high as observed. The observed data shows a frequency of about 80 for asian/pacific islander, while the expected frequency is more than twice as high, about 175.

The difference between observed data and expected data in the three groups "black", "hispanic", and "other". The observed data shows that black>other>Hispanic, while the expected data is hispanic>black>other.

Overall, the observed data do not appear consistent with the expected frequencies.

Boxplot of subject.age: (4f)




Based on the results of Analysis 4f, we observe that subject.age does mostly appear to be similar across the categories of subject.race.

In particular, we notice that among the five subject.races, the maximum value of age (excluding outliers) is around 70-80, and the minimum value is around 20. black's minimum value of age is lower than that of the other four races, which is around 5. Hispanic's maximum value of age is less than 70, which is the smallest among all the races, and other's maximum value is greater than 80, which is the largest among all the races. races, and other's highest value is greater than 80, the largest of all the races.

The median age of the five subject.races is very close to each other, at about 35. other has a slightly higher median, at about 40. The 5 lower quartiles of subject.races are also very close, all lying around 25, with other's being on the higher side at 30. The four races other than other share similar upper quartiles at around 45, with others reaching a value of 55.

Also, all the 5 box-plots appear to have positive skewness. The data are concentrated more on the top side of the distribution, and there are some outliers on the top side.