Assignment 2 Report

Analysis 1

1a: My ID number is [12345678].

1b: I do have concerns about study error in this study. This is because the size of sample we use in this study is smaller than 500, which means that our final estimation may be of quite large error.

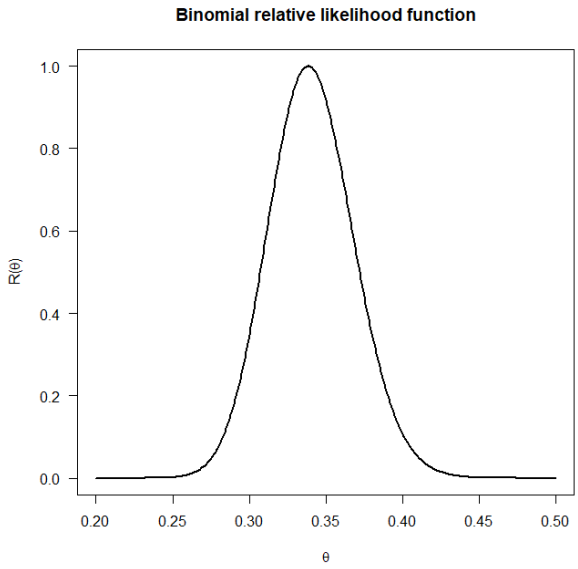
1c: The maximum likelihood estimate for Chicago is about 0.338565, while for San Francisco it is about 0.309013.

These were calculated by

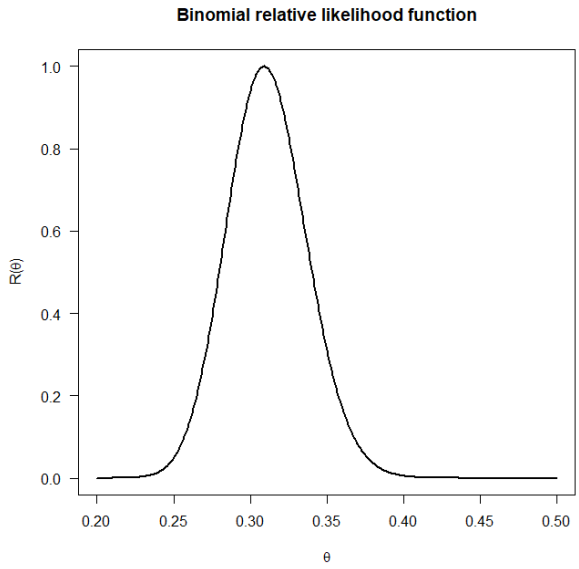
$$\theta_c = \frac{number_{Chicago,Female}}{number_{Chicago}} = \frac{151}{446} \approx 0.338565$$

$$\text{and } \theta_s = \frac{number_{San\ Francisco,Female}}{number_{San\ Francisco}} = \frac{144}{466} \approx 0.309013.$$

1d: Relative likelihood function plots:



(a)Chicago



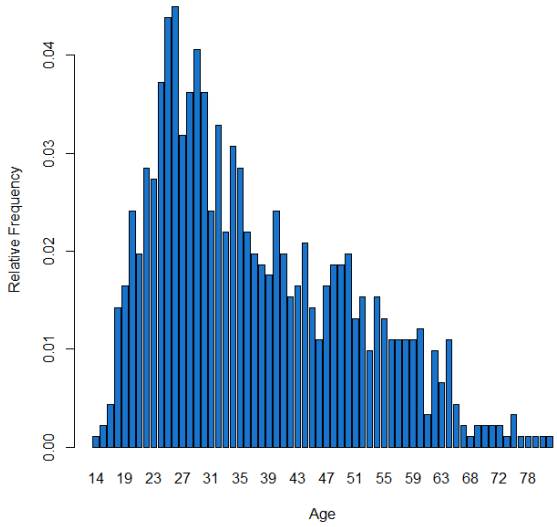
(b)San Fransisco

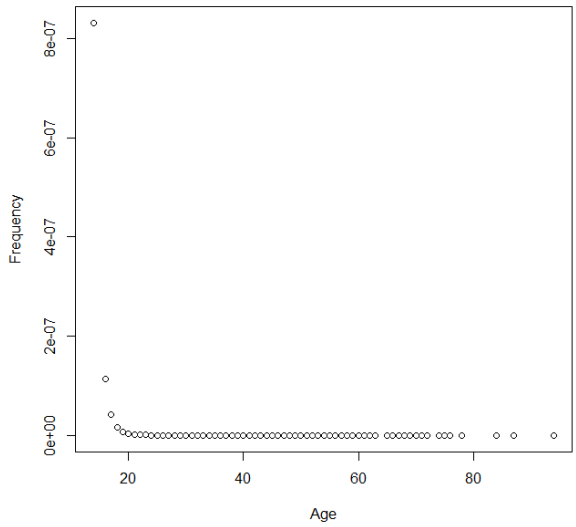
Analysis 2

2a: My ID number is [12345678].

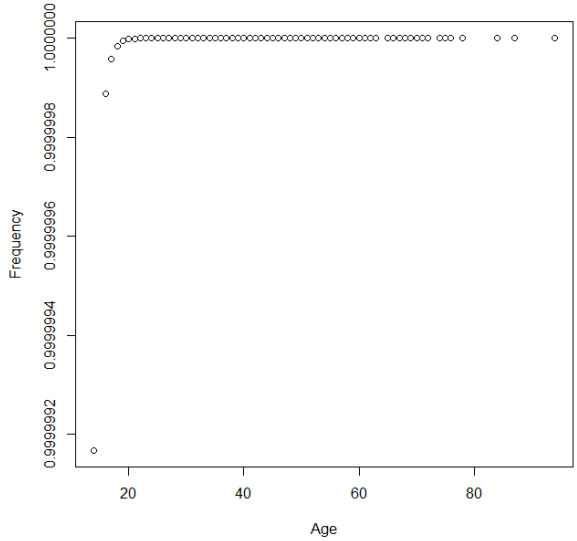
2b: The sample size, mean, median, and standard deviation are, respectively, 912, 37.02412, 34 and 13.46873.

2c: Relative frequency histogram:





2d: Empirical cumulative distribution function plot:

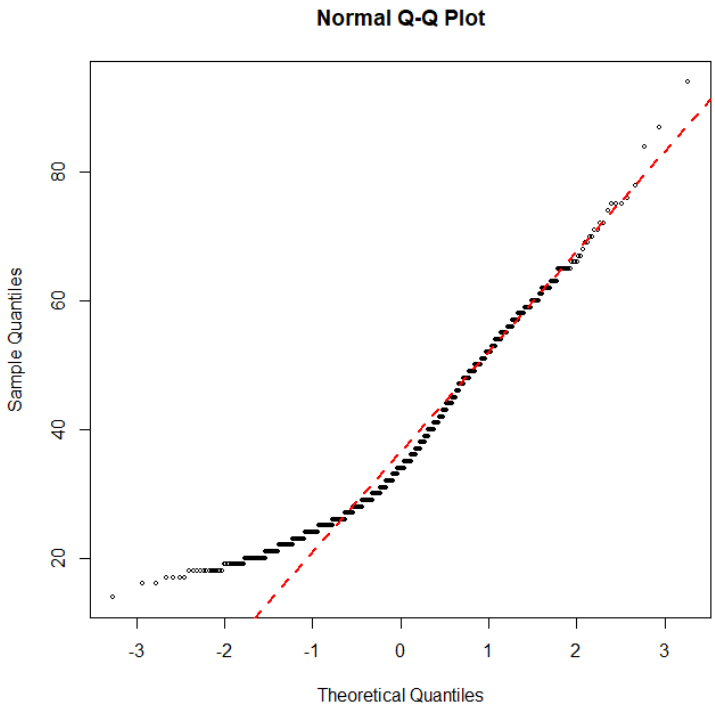


2e: Based on the plot in Analysis 2d, we can see that the relative frequency firstly increases with age then decreases with age while for data generated from an Exponential distribution

we would expect to see the relative frequency keeps decreasing when age increases. Besides, the data collected has its age reaching summit at around 27, which does not fit the model as well.

Ov**e**rall, the Exponential model does not fit our data well.

2f: Q-Q plot:



2g: Based on the Q-Q plot, balabala.

Analysis 3

3a: My ID number is [12345678].

3b: I do have concerns about sample error in this study. This is because the size of sample may limit the accuracy of our final estimation.

3c: The maximum likelihood estimate is 37.02412.

This was found by using the sample mean we calculated in Analysis part (b) with accordance to equation $E[Y]=\lambda$.

3d: The maximum likelihood estimate is about 0.1405.

From part (c), we obtain that $\hat{\lambda}=37.02412$, then we have approximation that

$$P(Y=k)=\frac{37.02412^k}{k!}e^{-37.02412},\quad k=0,1,\cdots$$

Then the estimation

$$\sum_{k=1}^{30}P(Y=k)=\sum_{k=1}^{30}\frac{37.02412^k}{k!}e^{-37.02412}\approx 0.1405$$

3e: R(39) = 6.789802×10^{-21}. Based on this, we can say that the value 39 is very implausible for λ based on our sample.

Analysis 4

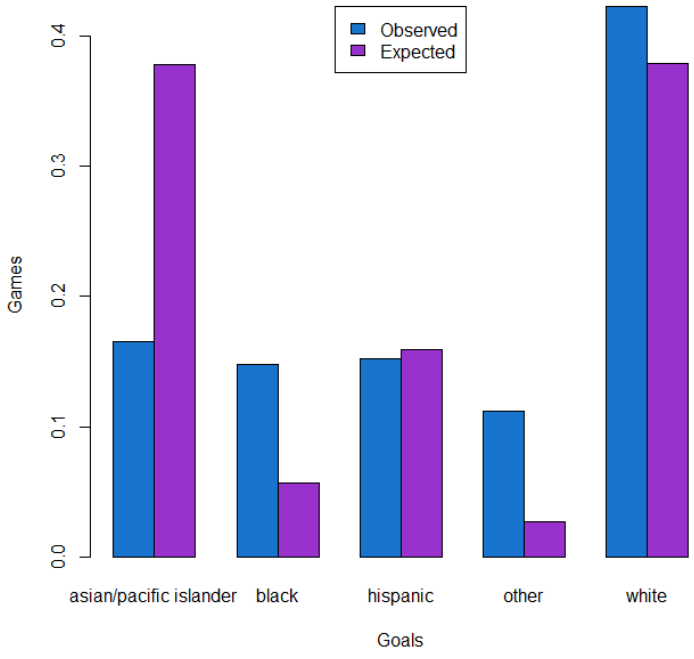
4a: My ID number is [12345678]. I will be analyzing the subject.race variate for San Francisco.

4b: I do have concerns about measurement error in the subject.race variate. This is because the size of sample is relatively small, besides, the population races make-up of San Francisco cannot represent the population races make-up of whole USA.

4c: Table of observed and expected frequencies (reminder: your Report should only contain one such table for your chosen variate):

Race	Observed Frequency	Expected Frequency
Asian/Pacific Islander	0.1652362	0.378
Black	0.1480687	0.057
Hispanic	0.1523605	0.159
White	0.4227468	0.379
Other	0.1115880	0.027

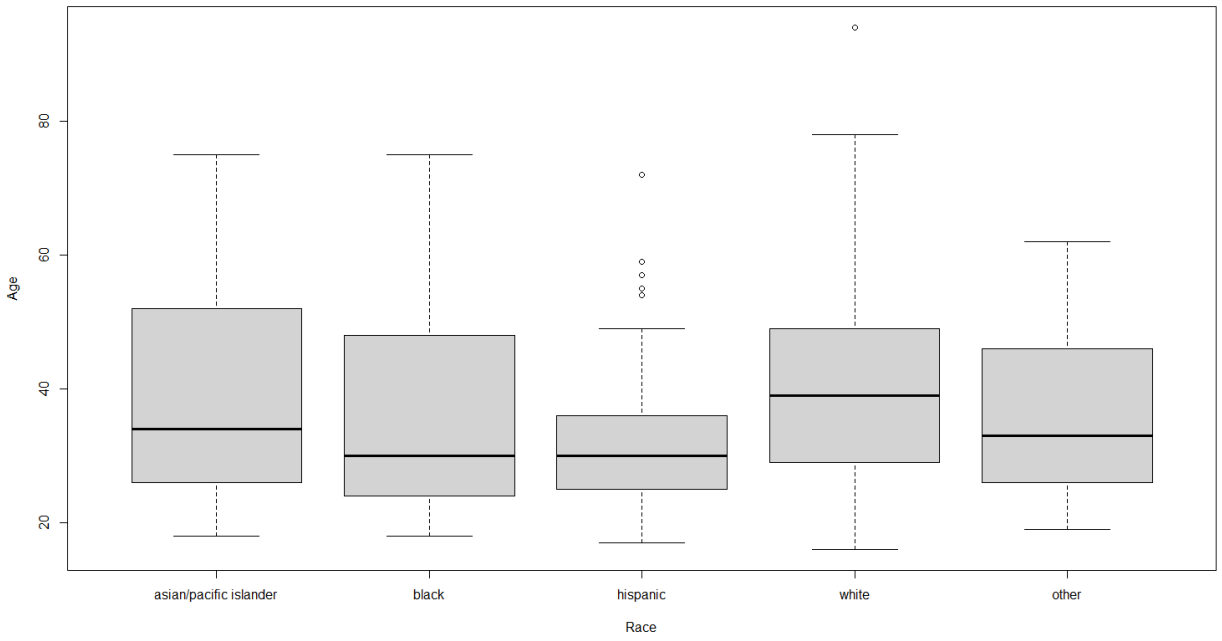
4d: Grouped barplot of observed and expected frequencies:



4e: Based on the results of Analyses 4c and 4d, we notice that the observed frequency and expected frequency of Hispanic and White races are similar, while the frequencies of

Asian/Pacific Islander, Black and Other races have quite significance difference. Overall, the observed data does not appear consistent with the expected frequencies.

4f: Boxplot of subject.age:



4g: Based on the results of Analysis 4f, we observe that subject.age does not appear to be similar across the categories of subject.race. In particular, we notice the age of Hispanic seems to be more concentrate, while the other categories have more disperse age distributions, but the average ages of different categories are quite close.