Stat 231 Assignment 2
MingMing Z — Student No. 21059539

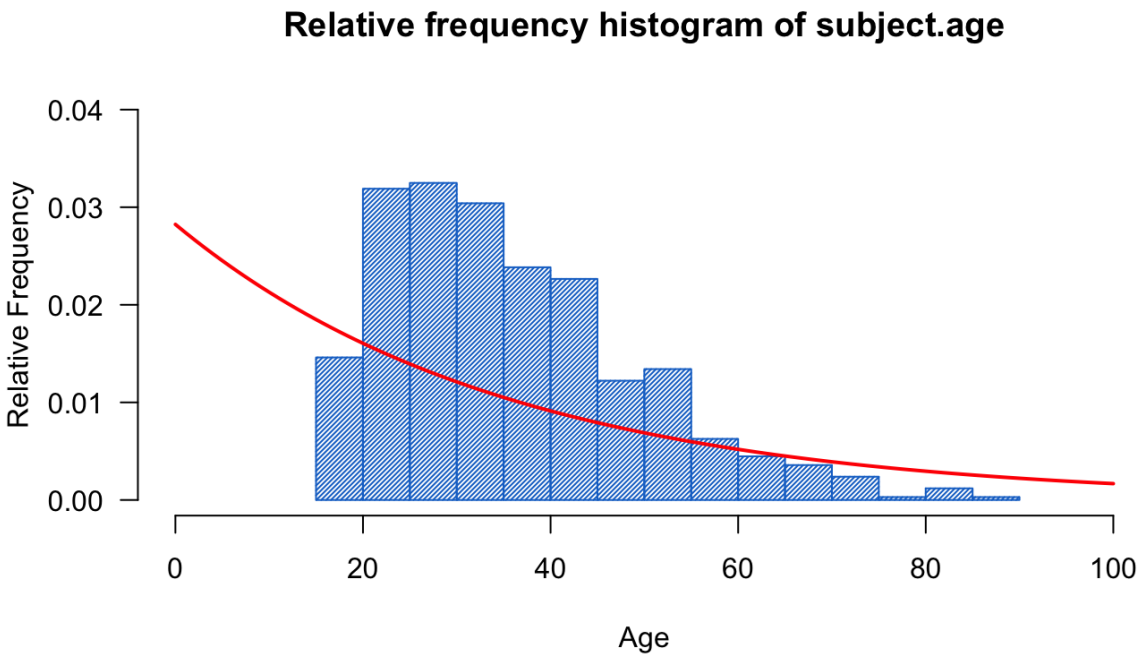
Analysis 1

1a: My ID number is 21058539. I will be analyzing the data for San Antonio.

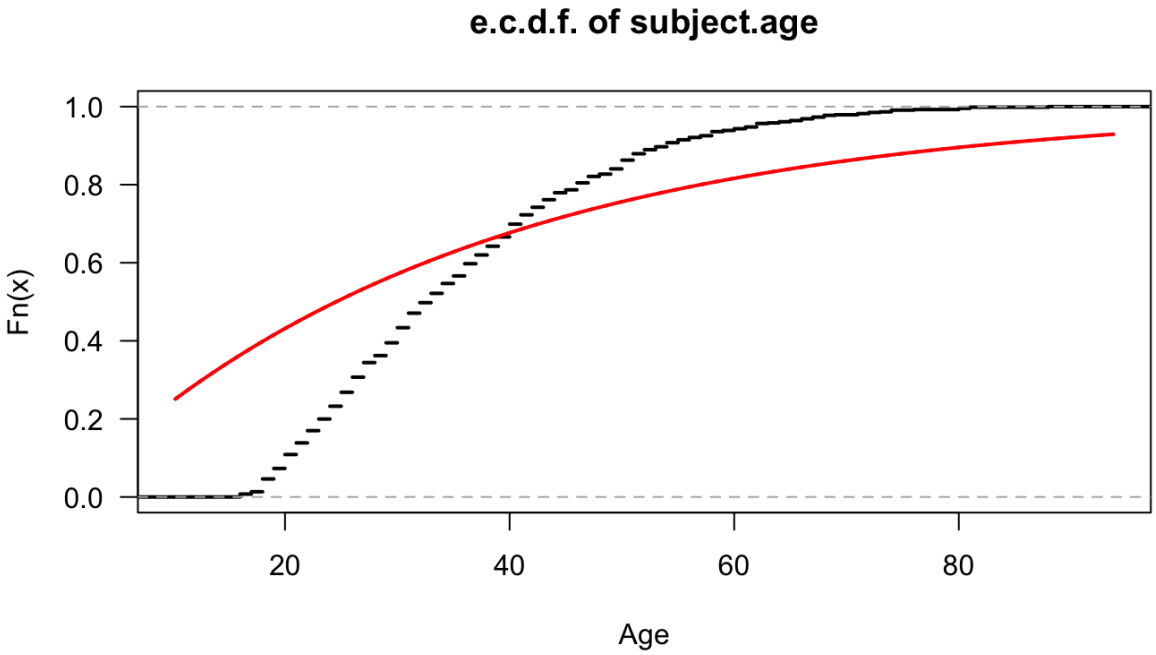
1b: I do not have much of a concern about measurement error in the `subject.age` variate, although I do recognise some causes of concern, since measurement errors should always be considered. This is because depending on how the data is collected, respondents may choose to lie about their age (e.g. underage driving) resulting in a difference between the true value of the age variate and the observed value. However, considering the large sample population (over one million traffic stops) and relatively large sample size (between 900 and 1100 traffic stops) and how the data comes from traffic stops from police departments (ID checks would ensure the validity of age), the reliability of the overall dataset makes concerns like this quite negligible in the broader scheme.

1c: The sample size is 671 traffic stops in San Antonio. The sample mean, sample median, sample standard deviation, and sample skewness of `subject.age` are, respctively, 35.4, 33.0, 13.4, 0.943. (rounded to 3 s.f.)

1d: Relative frequency histogram of `subject.age` for San Antonio with a superimposed PDF curve corresponding to an Exponential distribution with sample mean of `subject.age`.



1e: Empirical cumulative distribution function plot of `subject.age` for San Antonio with a superimposed CDF curve corresponding to an Exponential distribution with sample mean of `subject.age`.

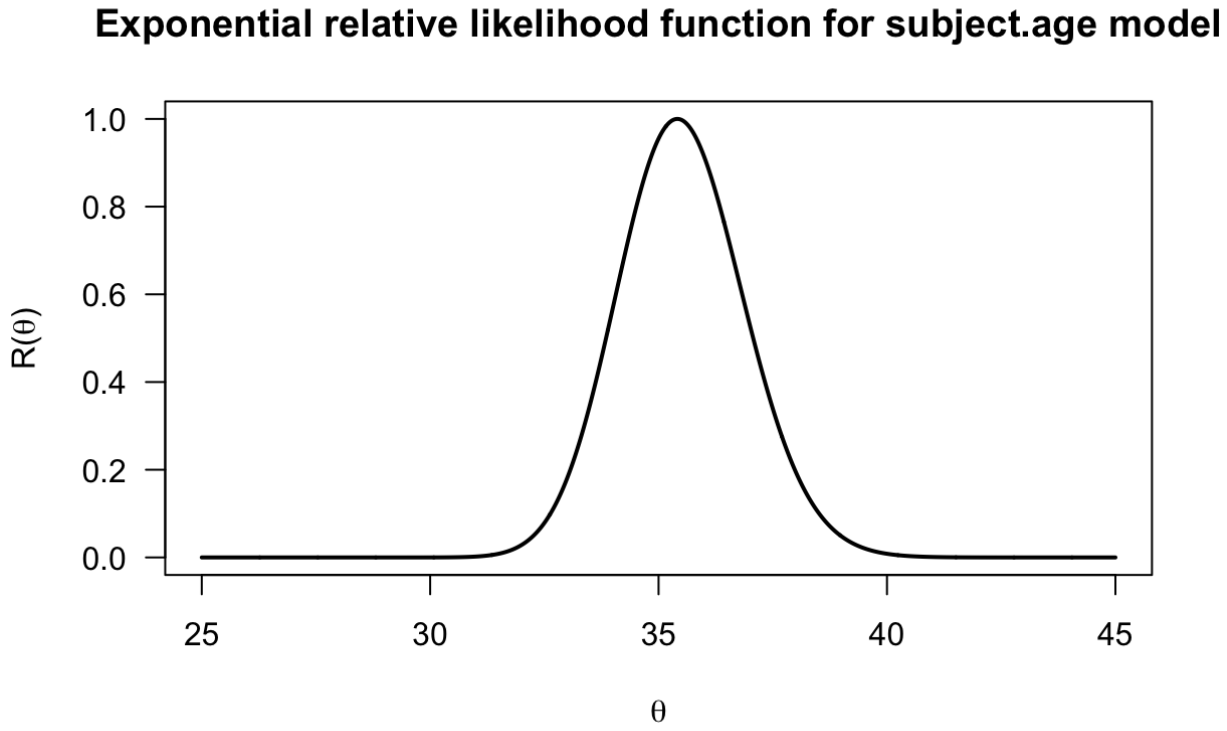


1f: Based on the plot in Analysis 1d, we can see that there exists a right tail based on the right skewed histogram (skewness is close to 1 from 1c), while for data generated from an Exponential distribution we would also expect to see a right skew, but with a much more exaggerated long right tail. From the plot in 1e, we can see the imposed exponential distribution CDF curve does not have a reasonable fit to the actual distribution. Additionally, based on descriptive statistics in 1c, we know the sample mean is 35.4 and sample standard deviation is 13.4, which are not similar, meaning our sample does not follow the property of an exponential distribution where the mean and standard deviation are equal. Overall, the Exponential model would not be that appropriate for this particular sample distribution, as the imposed curves have relatively large discrepancies to the actual plots.

1g: We know for an exponential distribution, $\hat{\theta} = \bar{y}$, the sample mean.
So the maximum likelihood estimate of θ is 35.4.

1h: The maximum likelihood estimate is 0.494.
This was found by invariance property. We assumed $A \sim Exponential(\theta)$,
so $P(A > 25) = 1 - P(A \leq 25) = 1 - (1 - e^{-\frac{25}{\theta}}) = e^{-\frac{25}{\theta}}$
Using the invariance property, the MLE is $e^{-\frac{25}{\hat{\theta}}} = e^{-\frac{25}{35.4}} = 0.494$
(`pexp(25, rate = 1 / agem, lower.tail = FALSE)` in R code)

1i: Relative likelihood function plot.



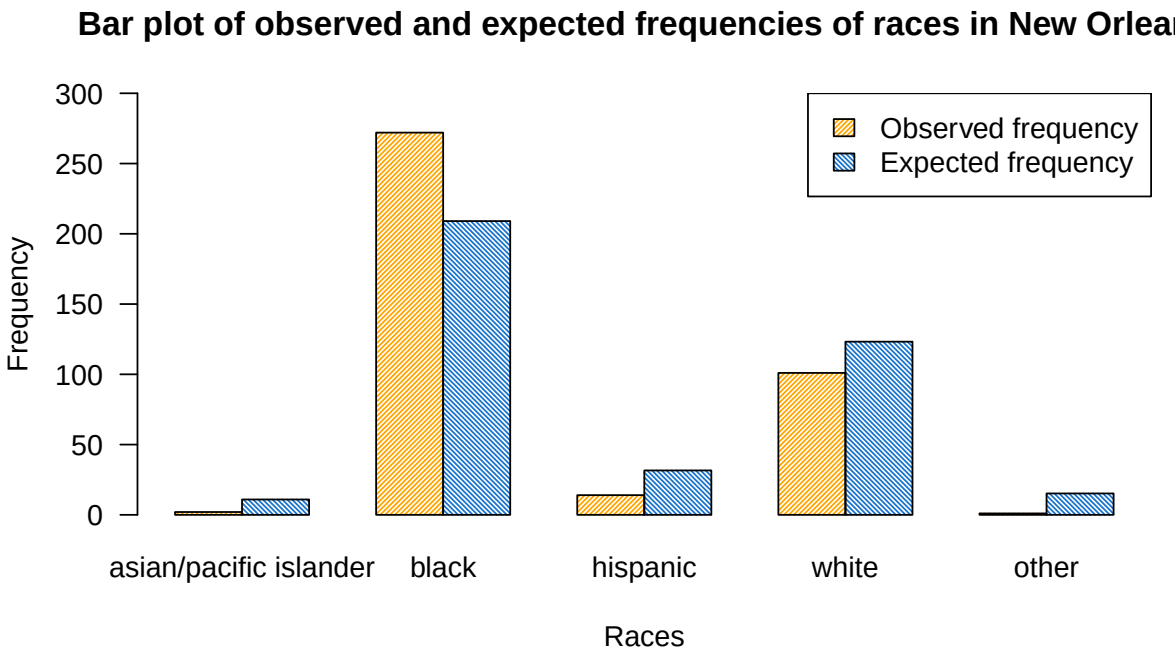
Analysis 2:

2a: My ID number is 21058539. I will be analyzing the `subject.race` variate for New Orleans.

2b. Table of observed and expected frequencies:

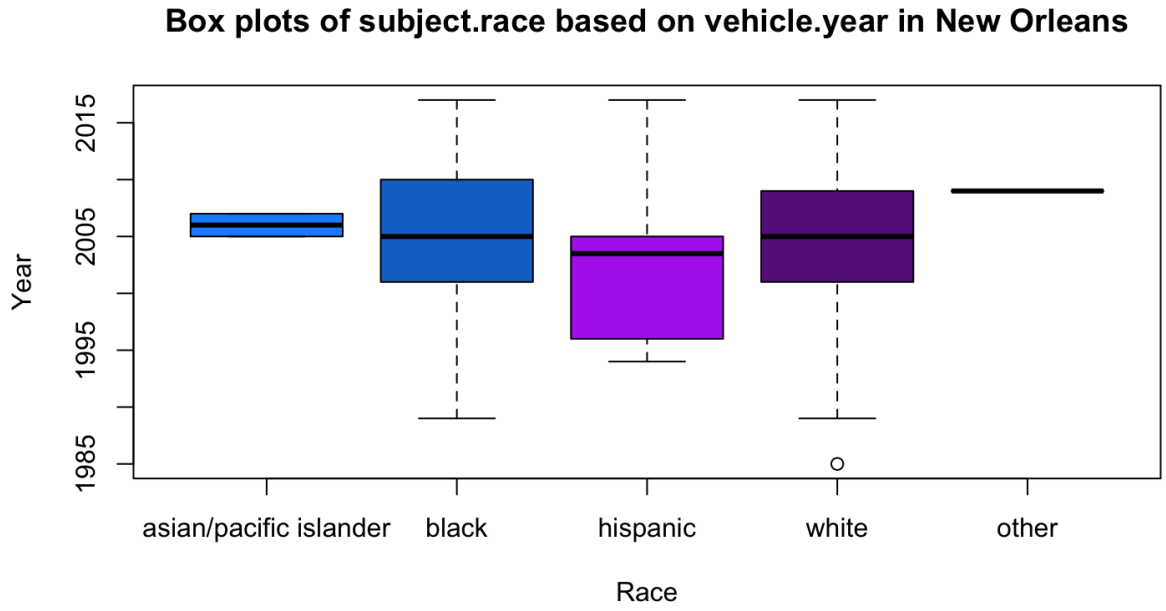
Subject Race	Observed Frequency	Expected Frequency
Asian/Pacific Islander	2	10.92
Black	272	209.04
Hispanic	14	31.59
White	101	123.24
Other	1	15.21

2c: Grouped barplot of observed and expected values.



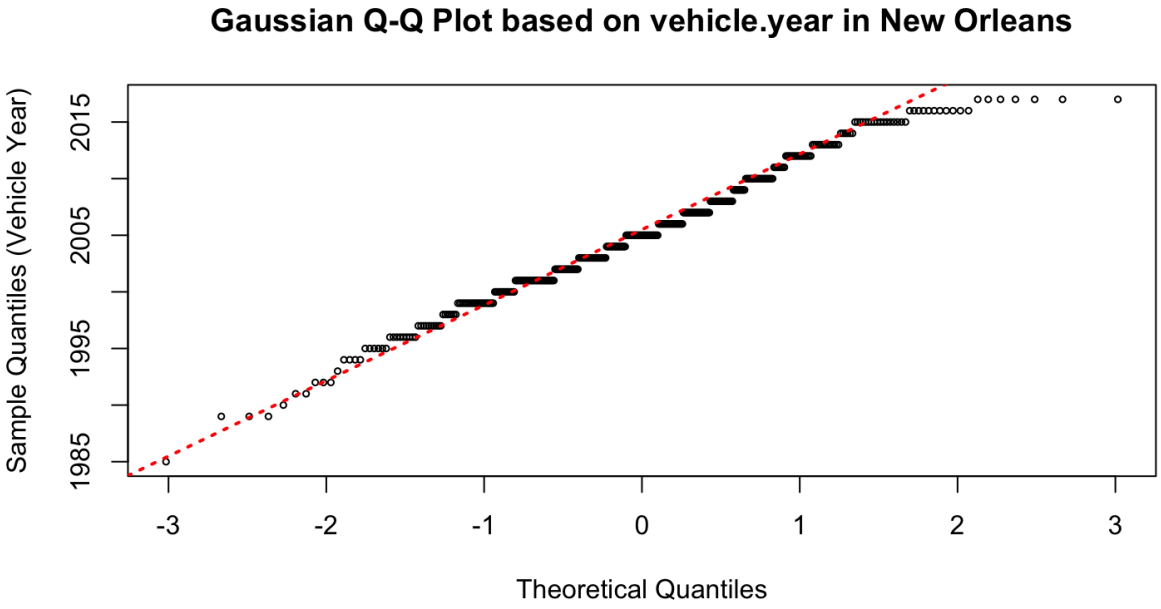
2d: Based on the results of Analyses 2b and 2c, we notice there are differences with observed and expected frequencies, especially for the minority races in New Orleans. The expected frequencies provided a notable overestimate for Asian/Pacific Islanders and Others when compared to the sample frequency, while the expected frequency is lower for Blacks compared to the observed frequency. However, the overall distribution of frequencies do align to an extent, where the Black and White races are still the majority. Overall, the observed data do appear consistent with the expected frequencies, but are less consistent for the minority race groups.

2e: Boxplot of `vehicle.year` and `subject.race` for New Orleans



2f: Based on the results of Analysis 2e, we observe that the distribution of `vehicle.year` does appear to be relatively similar across the categories of `subject.race`. In particular, we notice that the median falls around the same range (around 2005) for all the races, with Hispanics differing the most with having older vehicles, albeit still having a similar median with rest of the groups. But since the sample sizes for Asian/Pacific Islanders and Others are extremely small, it may not be right to conclude that the distribution is similar across all groups without more data (NEI).

2g: Q-Q plot of `vehicle.year` for New Orleans



2h: Based on the Q-Q plot, we can see that the points appear to lie reasonably along a straight line, with a suggestion of a left tail based on the slight U-shape bend at the right end of the data points. This implies that the data is relatively symmetric, except for a slight negative skew. Additionally, we can infer that the kurtosis may be lower or close to 3 based on the distribution of the end points, which makes sense with the Gaussian interpretation.

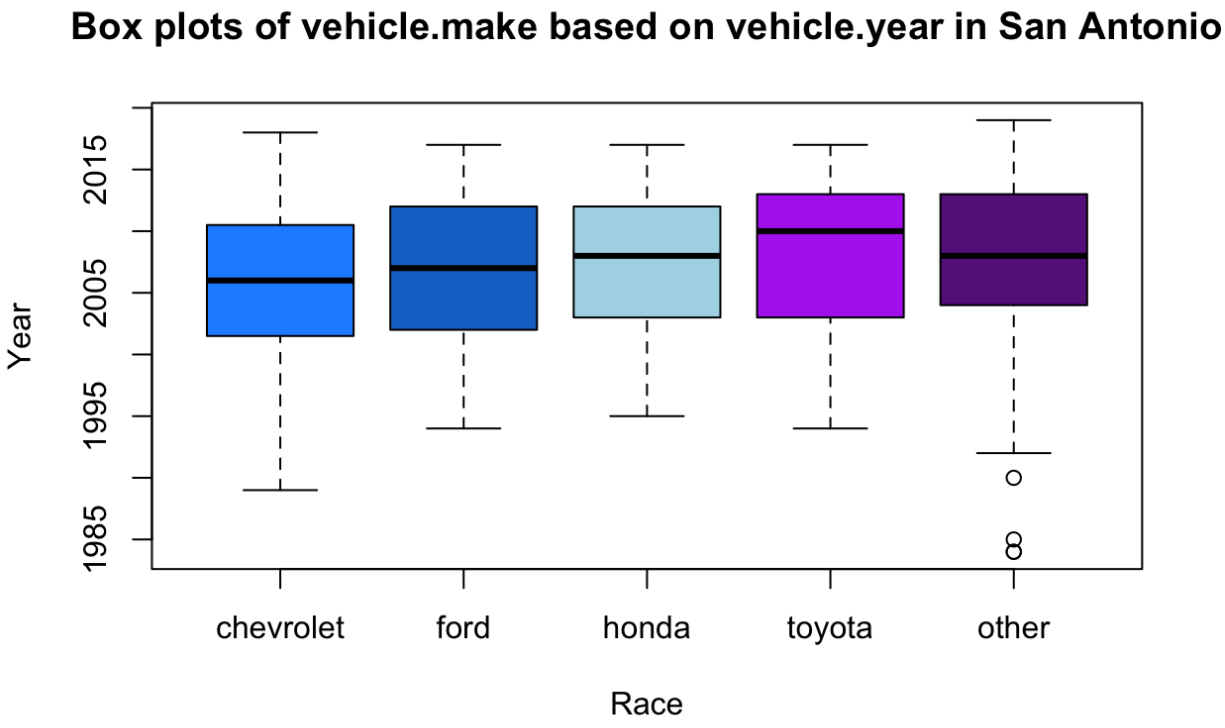
Analysis 3

3a: My student number is 21058539 and I analysed San Antonio for analysis 1. My favourite number is 42 since it's the answer to the universe!

3b: To begin this discussion, let's define some terms! First, a study error is a **systematic** difference between a target population and a study population. So you are right about it being a difference between a target population and a study population, but let's clarify this systematic part. Basically, differences between the two populations don't necessarily mean there's study error, as long as the study population accurately reflects the broader target population. But if bias are introduced that **consistently** affect the **representativeness** of the study population, we might classify that as a study error. Hope this helps!

3c: We have five 16 year olds and four 17 year olds, so we have a total of nine children. If we remove them, we would have $671 - 9 = 662$ individuals in our sample. Unlike median, mean is not robust, meaning it is more likely to be affected by the inclusion or exclusion of extreme values. But since the number of children is relatively small compared to the total sample size, I suspect the removal of these values won't affect the sample mean that much.
Our new mean is 35.7, which doesn't differ from the original mean of 35.4 that much.

3d: Hey Melissa, tell your friend to check this box plot out!



From this figure, we can see that there isn't too much of a difference between the different manufactures of a car when it comes to comparing vehicle year. Particularly, the median falls around the same range of 2005 - 2010, and the 1st and 3rd quartiles of all groups also align relatively close. Now of course, this is only for my sample, so I don't know the distribution for the entire dataset, but I don't think saying that vehicle.year should vary depending on the vehicle make is right.

Appendix: R Code

Include your R code here.

```
# setting up
mydata <- read.csv("~/Desktop/School/Stat 231/stat231f24dataset21058539.csv")
dim(mydata)
colnames(mydata)
sum(is.na(mydata))

skewness <- function(x) {
  (sum((x - mean(x))^3)/length(x))/(sum((x - mean(x))^2)/length(x))^(3/2)
}
kurtosis <- function(x) {
  (sum((x - mean(x))^4)/length(x))/(sum((x - mean(x))^2)/length(x))^2
}

# beginning of analysis
# 1c.

library(MASS)
SanAntonio <- subset(mydata, city=="sa")
NewOrleans <- subset(mydata, city=="no")
summary(mydata$subject.age)
summary(SanAntonio$subject.age)
summary(NewOrleans$subject.age)

dim(mydata)
dim(SanAntonio)
dim(NewOrleans)

sd(SanAntonio$subject.age)
skewness(SanAntonio$subject.age)

# 1d.

truehist(SanAntonio$subject.age, xlab = "Age",
         ylab = "Relative Frequency", main = "Relative frequency histogram of
         subject.age", xlim = c(0, 100), ylim = c(0, 0.04),
         las = 1, col = "dodgerblue3", density = 50)

agem <- mean(SanAntonio$subject.age)
curve(dexp(x, rate = 1 / agem), col = "red", add = TRUE, lwd = 2)

# 1e.

plot(ecdf(SanAntonio$subject.age), xlab = "Age",
     main = "e.c.d.f. of subject.age", las = 1, lwd = 2, pch = NA)

curve(pexp(x, rate = 1 / agem), col = "red", add = TRUE, lwd = 2)
```

```
# 1h.

pexp(25, rate = 1 / agem, lower.tail = FALSE)

# 1i.

ExpRLF <- function(theta, n, thetahat) {
  (thetahat/theta)^n * exp(n * (1 - thetahat/theta))
}

theta <- seq(25, 45, by = 0.01)

plot(theta, ExpRLF(theta, 671, agem), xlab = expression(theta),
     ylab = expression(paste("R(", theta, ")")), type = "l", lwd = 2,
     , main = "Exponential relative likelihood function
     for subject.age model", las = 1)

# 2b.
table(NewOrleans$subject.race)
nrow(NewOrleans)

# 2c.
exptd_freq <- c(10.92, 209.04, 31.59,123.24,15.21)
obsvd_freq <- table(NewOrleans$subject.race)

barplot(rbind(obsvd_freq, exptd_freq), beside = T, xlab = "Races",
        ylab = "Frequency", las = 1, main = "Bar plot of observed and
        expected frequencies of races in New Orleans",
        col = c("orange", "dodgerblue3"), ylim = c(0, 300),
        density = 50, angle = c(45, 135))

legend("topright", legend = c("Observed frequency", "Expected frequency"),
        fill = c("orange", "dodgerblue3"), density = 50, angle = c(45, 135))

# 2e.

boxplot(NewOrleans$vehicle.year ~ factor(NewOrleans$subject.race, levels =
        c("asian/pacific islander", "black", "hispanic", "white", "other")),
        col = c("dodgerblue1", "dodgerblue3", "darkorchid2", "darkorchid4"),
        xlab = "Race", ylab = "Year", main =
        "Box plots of subject.race based on vehicle.year in New Orleans")

# 2g.

qqnorm(NewOrleans$vehicle.year, pch = 1, cex = 0.5, ylab =
         "Sample Quantiles (Vehicle Year)", main =
         "Gaussian Q-Q Plot based on vehicle.year in New Orleans")
qqline(NewOrleans$vehicle.year, col = "red", lwd = 2, lty = 3)
```

```
# 3c.

table(SanAntonio$subject.age)
nrow(SanAntonio)
adults <- subset(SanAntonio$subject.age, SanAntonio$subject.age >= 18)
table(adults)
mean(adults)

# 3d.

table(SanAntonio$vehicle.make)

boxplot(SanAntonio$vehicle.year ~ factor(SanAntonio$vehicle.make, levels =
  c("chevrolet", "ford", "honda", "toyota", "other")), col =
  c("dodgerblue1", "dodgerblue3", "lightblue", "darkorchid2", "darkorchid4"),
  xlab = "Race", ylab = "Year", main = "Box plots of
  vehicle.make based on vehicle.year in San Antonio")
```