Stat 231 Assignment 3
MingMing Z. — 21058539

Analysis 1

1a: My ID number is 21058539.

1b: The units are individuals who could be the subject of a traffic stop in San Antonio or New Orleans, from December 2011 to June 2020 for San Antonio, and December 2009 to July 2018 for New Orleans. (Study time period is taken from Dataset Information File)

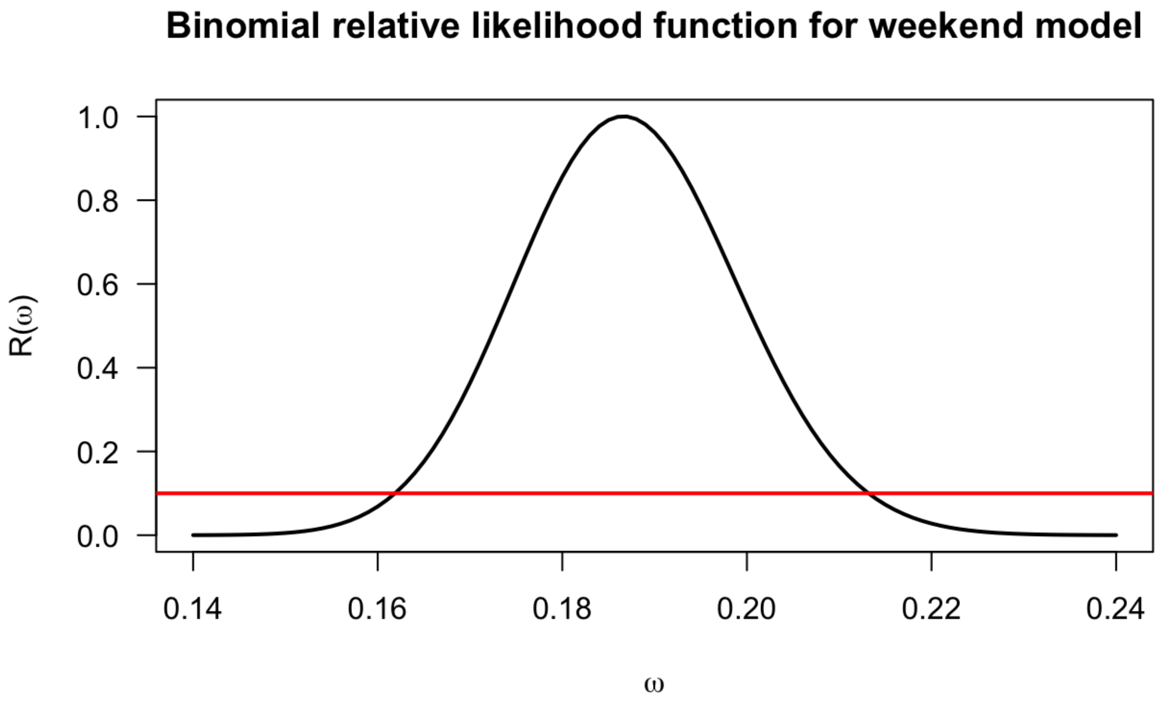
1c: Table of frequencies and percentages of stops based on `day.of.week` with sample size of 1061:

	Frequencies	Percentages(%)
Monday	137	12.9
Tuesday	190	17.9
Wednesday	195	18.4
Thursday	182	17.2
Friday	159	15.0
Saturday	111	10.5
Sunday	87	8.20

1d: A possible attribute of interest is the proportion of individuals who stopped on a weekend day compared to those who stopped on a weekday. This might help us answer if traffic stops are more common on the weekends or on weekdays.

1e: My sample contains 1061 stops, of which 198 took place on a weekend day. The maximum likelihood estimate of ω is $\frac{198}{1061}$. (Since the MLE of a binomial distribution is the observed number of successes over the total number of trials)

1f: Relative likelihood function plot.



1g: The 10% likelihood interval for ω is [0.162, 0.213].

1h: $\omega = \frac{1}{7}$ is not a plausible value for ω, because $R(\frac{1}{7})$ evaluates to 0.000460, which if we are expecting it to be a reasonable estimate, should be close to 1. However, it is far from one, and if we look at the RLF plot in 1f, the value $\frac{1}{7} \approx 0.143$ is no where close to where ω is maximised. Furthermore, $\frac{1}{7}$ does not lie in the 10% likelihood interval, and based on lecture discussed guidelines, values of outside a 10% likelihood interval are implausible values of ω in light of the observed data.

1i: The 96.8% approximate confidence interval for ω is [0.161, 0,212], this is an approximate confidence interval.
This interval was calculated by first finding the chi squared probability based on $P(V \leq -2log(0.1))$, using R code `p = pchisq(-2 * log(0.1), 1)`≈ 0.968, implying this should look similar to a 96.8% confidence interval.
Recall that for a Binomial(ω) model, an approximate 100q% confidence interval for ω based on the Central Limit Theorem approximation is given by $\hat{\omega} \pm a\sqrt{\frac{\hat{\omega}(1-\hat{\omega})}{n}}$ where $P(Z \leq a) = \frac{1+p}{2}$
Using this, we can find a quantile a for the confidence Interval: `a <- qnorm((1 + p) / 2)`
Finally, we use `thetahat +/- a * sqrt(thetahat * (1 - thetahat)) / n)`, which is the CLT approximation above, to find the confidence intervals.

1j: A $100\hat{\omega}$% confidence interval is approximately equivalent to a 97.3% likelihood interval, because by Theorem 4.6.3 in the course notes, we have that if a is a value such that $\hat{\omega} = 2P(Z \leq a) - 1$, where $Z \sim N(0, 1)$, then the likelihood interval, $\theta : R(\theta) \geq e^{\frac{-a^2}{2}}$ is an approximate $100\hat{\omega}$% confidence interval. So in this case, we find a using `a <- qnorm((1 + thetahat) / 2)`, and the percentage of likelihood interval using `exp(-a^2/2)`.

Analysis 2

2a: My ID number is 21058539. I will analyze the data for San Antonio.

2b: The sample size is 671. The summary statistics are as follows (rounded to 3 s.f.):

Sample Statistic	`subject.age`	`subject.age.log`
Mean	35.4	3.50
Standard deviation	13.4	0.365
5^{th} percentile	19.0	2.94
95^{th} percentile	62.0	4.13

2c: An approximate 90% confidence interval for θ is [33.623, 37.200].

2d: The result in Analysis 2c tells us that we are 90% confident that the average age of a subject in a traffic stop chosen at random falls within that interval. In other words, we are 90% confident that the interval [33.623, 37.200] contain is the true but unknown value of θ.

2e: A 90% confidence interval for μ is [3.477, 3.523]. Converted back to the original scale, this would be [32.349, 33.883].

2f: A 90% confidence interval for σ is [0.349, 0.382]. This was calculated by using the $100p\%$ confidence interval for σ which is $(\sqrt{\frac{(n-1)s^2}{b}}, \sqrt{\frac{(n-1)s^2}{a}})$, where in this case, p is 0.9, $n-1$ is the degree of freedom = 670, s^2 is the variance, and a and b are calculated using the fact that
$P(U \leq a) = \frac{1-p}{2}, P(U \leq b) = \frac{1+p}{2}$.
In R code, a and b are calculated by using `a <-qchisq(0.05, 670), b <- qchisq(0.95, 670)`
Finally, interval values are calculated using `[sqrt(s2 * 670 / b), sqrt(s2 * 670 / a)]`

2g: Assuming the sample mean remains the same, we would require a sample of size 3393 in order to have a 90% confidence interval for θ of width no greater than 2. This is because with the same sample mean of $\hat{\theta}$, we would want to find the smallest possible sample size that gives confidence interval $[\hat{\theta} - 1, \hat{\theta} + 1]$. By rearranging $\hat{\theta} + a\frac{\hat{\theta}}{\sqrt{n}} = \hat{\theta} + 1$ and $\hat{\theta} - a\frac{\hat{\theta}}{\sqrt{n}} = \hat{\theta} - 1$ (confidence interval approximation for exponential model), we get $a\frac{\hat{\theta}}{\sqrt{n}} = 1 \implies (a\hat{\theta})^2 = n$
In R, we get n = 3393 using `n <- (a * m)^2`, where a is `qnorm((1 + 0.9)/2)` and m is the mean.

Analysis 3

(a) My student number is 21058539. I enjoy a good ol plate of curry.

(b) Not exactly, the thing is, for a Binomial model like ours, a parameter estimate of 0.999 is essentially the same as 0.001 since this is like saying nearly every stop happens on the weekend or nearly every stop happens on the weekdays, which are both extremely unlikely. Now, there can be an argument that 0.001 is more plausible, since this refers to the scenario that nearly every stop happens on the weekdays, which have more days than the weekends, and is closer to the MLE $\hat{\omega}$ we found. So no, saying omega = 0.999 is more plausible than omega = 0.001 is definitely wrong.

(c) Well, if we included more counts in our weekend days count, it would, first, affect the MLE of ω to be bigger. Recall that this represents the number of weekend stops over the total sample count, where the total does not change in this scenario. For our sample, we have 159 stops on Friday, so no matter how many of these stops are after 6PM, as long 1 or more are added to the weekend count, $\sqrt{\frac{\hat{\omega}(1-\hat{\omega})}{n}}$ would increase since $\hat{\omega}(1-\hat{\omega})$ always yields a bigger value in this case. Note that if we don't add any stops to weekend, nothing changes, while if we add a lot more stops than 159 to weekend, $\hat{\omega}(1-\hat{\omega})$ may not yield a bigger value since it is maximised at $\hat{\omega} = 0.5$. Anyways, for our sample, this would mean the overall confidence interval will become wider (Recall we calculate the confidence interval using $\hat{\omega} \pm a\sqrt{\frac{\hat{\omega}(1-\hat{\omega})}{n}}$). Hope this mathematical explanation makes sense, although if you want to think of it intuitively, moving more Friday counts to weekend in our sample changes the success probability of the Binomial model closer to 0.5, so we become more unsure whether a stop happens on a weekday or a weekend. This unsureness makes us less confident, so the confidence interval widens (we need more coverage now).

(d) Don't get too excited, the thing is, our confidence level (like 90%) remains constant regardless of whether the ages are in years or months, you are just changing a sense of scale. This means, even if the interval appears wider, it won't necessarily have more coverage. In other words, the coverage is tied to our chosen confidence level and is not affected by scaling, and speaking of scaling, I would guess the width of your interval was also multiplied by 12 approximately?

(e) Well, we would worry about the validity of the Central Limit Theorem approximation under usual circumstances with other models, but in 2f, we are using an assumed Gaussian model, which is what the CLT relies on. So luckily, we don't have to be concerned.

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))

# beginning of analysis
# 1c.

freqt<-table(mydata$day.of.week)
freqt
sum(freqt)
(freqt/1061) * 100

# 1f.
theta <- seq(from = 0.14, to = 0.24, by = 0.001)
n <- 1061
y <- 198
BinLF <- function(theta, n, y) {
  (theta^y) * (1 - theta)^(n-y)
}
RLFVals <- BinLF(theta, n, y) / max(BinLF(theta, n, y))

plot(theta, RLFVals, main = "Binomial relative likelihood function for weekend model",
     xlab = expression(omega), ylab = expression(paste("R(", omega, ")")), type = "l",
     lwd = 2, las = 1)
abline(h = 0.1, col = "red", lwd = 2)

# 1g.
thetahat <- 198/1061
BinRLF <- function(x) {
  BinLF(x, n, y) / BinLF(thetahat, n, y)
}

uniroot(function(x) BinRLF(x) - 0.1, lower = 0.1, upper = thetahat)$root
uniroot(function(x) BinRLF(x) - 0.1, lower = thetahat, upper = 0.22)$root

# 1h.
BinRLF(1/7)

# 1i.
p <- pchisq(-2 * log(0.1), 1)
a <- qnorm((1 + p) / 2)
thetahat - a * sqrt((thetahat * (1 - thetahat)) / n)
thetahat + a * sqrt((thetahat * (1 - thetahat)) / n)
```

```
# 1j.
a <- qnorm((1 + thetahat) / 2)
thetahat - a * sqrt((thetahat * (1 - thetahat)) / n)
thetahat + a * sqrt((thetahat * (1 - thetahat)) / n)
approx <- exp(-a^2/2)
approx
uniroot(function(x) BinRLF(x) - approx, lower = 0.1, upper = thetahat)$root
uniroot(function(x) BinRLF(x) - approx, lower = thetahat, upper = 0.22)$root

# 2b.
SanAntonio <- subset(mydata, city=="sa")
sum(table(SanAntonio$subject.age))
mean(SanAntonio$subject.age)
mean(log(SanAntonio$subject.age))
sd(SanAntonio$subject.age)
sd(log(SanAntonio$subject.age))
quantile(SanAntonio$subject.age, c(0.05, 0.95))
quantile(log(SanAntonio$subject.age), c(0.05, 0.95))

# 2c.
newth <- mean(SanAntonio$subject.age)
c <- qnorm((1 + 0.9)/2)
newth - c * newth/sqrt(n)
newth + c * newth/sqrt(n)

# 2e.
length(log(SanAntonio$subject.age))
x_bar <- mean(log(SanAntonio$subject.age))
s <- sd(log(SanAntonio$subject.age))
lower <- x_bar - c * (s / sqrt(671))
higher <- x_bar + c * (s / sqrt(671))
lower
higher
exp(lower)
exp(higher)

# 2f.
s2 <- s^2

(1 - 0.9)/2
(1 + 0.9)/2
chia <- qchisq(0.05, 670)
chib <- qchisq(0.95, 670)

sqrt(s2 * 670 / chib)
sqrt(s2 * 670 / chia)

# 2g.
m <- mean(SanAntonio$subject.age)
a <- qnorm((1 + 0.9)/2)
```

```
n <- (a * m)^2
n
m - a * m/sqrt(n)
m + a * m/sqrt(n)

# 3c.
thetahat <- 198/1061
thetahat
(thetahat * (1 - thetahat))
sqrt((thetahat * (1 - thetahat)) / 1061)
thetahat <- (198 + 159)/1061
thetahat
(thetahat * (1 - thetahat))
sqrt((thetahat * (1 - thetahat)) / 1061)

thetahat <- (198 + 500)/1061
thetahat
(thetahat * (1 - thetahat))
sqrt((thetahat * (1 - thetahat)) / 1061)
```