**Assignment 3 Word Template**

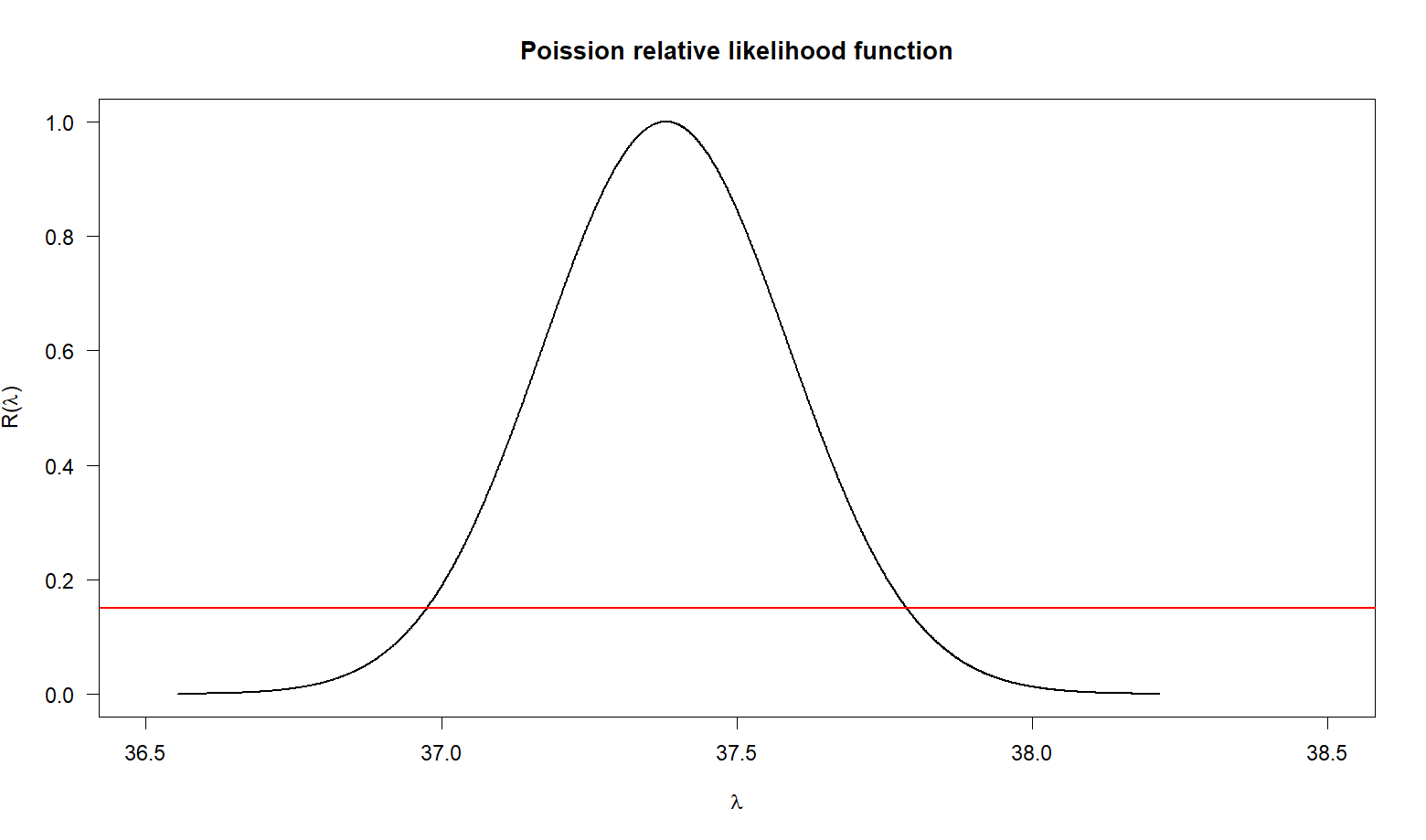
**Note**: The images in this template are to indicate where to include your plots. You may find your plots should be larger in size than these template images in order for them to be easily read.

**Analysis 1**

**1a**: My ID number is 20995558.

**1b**: The sample size is 859, the maximum likelihood estimate of lambda is 37.380.

**1c**: Relative likelihood function plot:



**1d**: The 15% likelihood interval for lambda is [36.97, 37.79].

**1e**: The approximate 15%, 90% and 95% confidence intervals for lambda are [37.34, 37.42], [37.04, 37.72], and [36.97, 37.79], respectively. These were calculated by using confidence interval level and then using CLT approximation with qnorm function (get z-score). Lastly, I put pieces together to find the approximate confidence interval:

lambda\_hat <- mean(mydata$subject.age)

n <- length(mydata$subject.age)

z\_15 <- qnorm((1 + 0.15)/2)

z\_90 <- qnorm((1 + 0.9)/2)

z\_95 <- qnorm((1 + 0.95)/2)

CI\_15 <- c(lambda\_hat - z\_15 \* sqrt(lambda\_hat/n), lambda\_hat + z\_15 \* sqrt(lambda\_hat/n))

CI\_90 <- c(lambda\_hat - z\_90 \* sqrt(lambda\_hat/n), lambda\_hat + z\_90 \* sqrt(lambda\_hat/n))

CI\_95 <- c(lambda\_hat - z\_95 \* sqrt(lambda\_hat/n), lambda\_hat + z\_95 \* sqrt(lambda\_hat/n))

**1f**: The approximate 95% confidence interval is most similar to the 15% likelihood interval. This is not what I would expect, because a 95% confidence interval should be wider than a 90% confidence interval since it contains a wider range of values. In this case, 95% confidence interval should be wider than a 15% confidence, but it does not.

**1g**: The interval [36.97, 37.79] tells us that it has 95% confident that the true average age of the subjects in the study is between approximately 36.97 and 37.79 years, which tells the age group that is most central to the study's population. This also suggests that the study predominantly involves participants within this age group, giving more information to the study’s findings.

**1h**: It is not possible for lambda = 39, because the likelihood interval calculated earlier is [36.97, 37.79] does not include 39, suggesting that based on this study's sample, λ = 39 is not a likely estimate. Meanwhile, when we plug 39 into the equation (lambda = 39), we get zero, meaning that it is impossible.

**Analysis 2**

**2a**: My ID number is 20995558. I will analyze the subject.sex variate for Chicago.

**2b**: In my sample, for Chicago, the sample size is 443, the number of stops for which subject.sex was `female' is 159, and the maximum likelihood estimate of theta\_c is 0.359.

**2c**: I will calculate a 95% likelihood interval for theta\_c. The interval is [0.305, 0.416], and was calculated by using relative likelihood function of binomial model and finding lower and upper bound using uniroot function. Based on this interval, I do not think 0.513 is a plausible value for theta\_c.

BioRLF <- function(theta, n, y, thetahat) {

(theta/thetahat)^y \* ((1-theta)/(1-thetahat))^(n-y)

}

theta\_hat <- female\_stops/n

lower\_bound <- uniroot(function(x) BioRLF(x, n, female\_stops, theta\_hat) - 0.05, lower = 0, upper = theta\_hat)$root

upper\_bound <- uniroot(function(x) BioRLF(x, n, female\_stops, theta\_hat) - 0.05, lower = theta\_hat, upper = 1)$root

**2d**: The 95% likelihood level is approximately equivalent to a 25.125% confidence level by transforming the likelihood level into a chi-squared value using a chi-squared distribution with one degree of freedom. This confidence interval is [0.205, 0.513]. The R code is provided below:

n <- length(chicago$subject.sex)

theta\_hat <- female\_stops / n

sd\_binomial <- sqrt(theta\_hat \* (1 - theta\_hat) \* n)

q <- pchisq(-2 \* log(0.95), 1)

a <- qnorm((1+q)/2)

theta\_hat - a \* sd\_binomial/sqrt(n)

theta\_hat + a \* sd\_binomial/sqrt(n)

**2e**: It is not possible for theta\_c = 0, because theta\_c represents the proportion of stops in which the subject identifies as female. If it is equal to 0, it would imply that there are no females in the city, an assertion that is unrealistic. Based on our data observations, there have indeed been stops involving females.

**Analysis 3**

**3a**: My ID number is 20995558. I will analyze the lat and lng variates for Chicago.

**3b**: The sample size is [number]. The summary statistics are as follows:

|  |  |  |
| --- | --- | --- |
| Sample Statistic | lat | lng |
| Mean | 41.845 | -87.658 |
| Standard deviation | 0.078 | 0.062 |
| 2. 5th percentile | 41.684 | -87.842 |
| 97.5th percentile | 41.972 | -87.551 |

**3c**: A 95% confidence interval for mu\_t is [41.837, 41.852], while for mu\_g is [-87.664, -87.652]. These are based on an asymptotic approximation and were calculated by getting the 95% confidence interval for the mean of [sample mean +- 1.96 \* standard error] for lat and lng respectively.

CI\_lower <- mean - 1.96 \* (sd / sqrt(n))

CI\_upper <- mean + 1.96 \* (sd / sqrt(n))

**3d**: The latitude value given by wiki.openstreetmap.org for my chosen city is 41.8781, while the longitude value is -87.6298. Based on my analyses, the latitude value is not a plausible value for mu\_t, while the longitude value is not a plausible value for mu\_g. These conclusions are because both the given latitude and longitude do not fall within their respective calculated confidence intervals.

**3e**: The probability mu\_t lies in the interval [41.837, 41.852] is 95%, which is because we get the interval as a 95% confidence interval, so that if we continue draw samples and calculate the confidence interval in the same manner, about 95% of those intervals would contain the true population mean.

**3f**: A 95% confidence interval for sigma\_t is [0.0734, 0.0838], while for sigma\_g is [0.0585, 0.0667]. These were calculated by converting it into a chi-squared value using a chi-squared distribution with one degree of freedom and then with getting z-score, and utilizing the Central Limit Theorem, we formed the confidence interval around our sample standard deviation, which serves as our point estimate for the population standard deviation. Code below:

s2 <- var(chicago$lng)

df = n - 1

a <- 0.05

lower <- qchisq(a / 2, df)

upper <- qchisq(1 - a / 2, df)

variance\_lower <- df \* s2 / upper

variance\_upper <- df \* s2 / lower

sd\_lower <- sqrt(variance\_lower)

sd\_upper <- sqrt(variance\_upper)