Assignment 2 Rui Qin 30874157

Question 1

Question 1A

Below is the formula of the estimate using the t-distribution

$$\left(\hat{\mu}_{\mathrm{ML}} - t_{\alpha/2,n-1}\frac{\hat{\sigma}_u}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + t_{\alpha/2,n-1}\frac{\hat{\sigma}_u}{\sqrt{n}}\right)$$

Based on the formula we can calculate the result in RStudio

```
# Load the data
df <- read.csv("covid.19.ass2.2023.csv")


# Question 1A
df_mean <- mean(df$Recovery.Time)
df_sd <- sd(df$Recovery.Time)
df_var <- var(df$Recovery.Time)
se <- df_sd / sqrt(length(df$Recovery.Time))
t_critical <- qt(1 - 0.05/2, df = length(df$Recovery.Time) - 1)
margin_of_error <- t_critical * se
upper_limit <- df_mean + margin_of_error
lower_limit <- df_mean - margin_of_error
df_mean
cat("[", lower_limit, ",", upper_limit, "]\n")
> df_mean
[1] 14.25797
> cat("[", lower_limit, ",", upper_limit, "]\n")
[ 13.98935 , 14.52659 ]
```

The average duration for Covid-19 patients in New South Wales to recover is about 14.25797 days, supported by a 95% confidence interval spanning from 13.98935 to 14.52659 days.

Question 1B

Below is the formula of confidence interval with difference of means

$$\left(\hat{\mu}_A - \hat{\mu}_B - z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}},\ \ \hat{\mu}_A - \hat{\mu}_B + z_{\alpha/2}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}\right)$$

Based on the formula we can calculate the result in RStudio

```
# Question 1B
israeli_df <- read.csv("israeli.covid.19.ass2.2023.csv")
israeli_mean <- mean(israeli_df$Recovery.Time)
israeli_var <- var(israeli_df$Recovery.Time)
israeli_se <- israeli_sd / sqrt(length(israeli_df$Recovery.Time))
mean_difference <-  israeli_mean - df_mean
se_difference <- sqrt((df_var / length(df$Recovery.Time))
                      + (israeli_var / length(israeli_df$Recovery.Time)))

margin_of_error <- t_critical * se_difference
diff_upper_limit <- mean_difference + margin_of_error
diff_lower_limit <- mean_difference - margin_of_error
mean_difference
cat("[", diff_lower_limit, ",", diff_upper_limit, "]\n")
> mean_difference
[1] 0.391829
> cat("[", diff_lower_limit, ",", diff_upper_limit, "]\n")
[ -0.1643962 , 0.9480542 ]
```

The mean difference is 0.391829. And we hold a 95% confidence that the variance in mean recovery times between the Israeli patients and those in New South Wales lies within the interval from -0. 1643962 days to 0.9480542 days.

Question 1C

Based on the question we are working on testing the difference of means with unknown variances

Hypotheses:

- Null Hypothesis (H0): The average recovery time for the Israeli cohort matches that of the NSW cohort at the population level.
    - H0: μ1 = μ2
- Alternative Hypothesis (H1): The average recovery time for the Israeli cohort does not match that of the NSW cohort at the population level.

- H1: μ1 ≠ μ2

$$z_{(\hat{\mu}_x - \hat{\mu}_y)} = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\hat{\sigma}_x^2}{n_x} + \frac{\hat{\sigma}_y^2}{n_y}}}$$

```
# Question 1C
test_statistic <- (df_mean - israeli_mean) / se_difference
2 * pnorm(-abs(test_statistic))
> 2 * pnorm(-abs(test_statistic))
[1] 0.1671578
```

The p-value at 0.1675329, provides evidence against the Alternative hypothesis. This supports the assertion that there is no distinction in the average recovery times between Israeli and NSW patients.

Question 2

Question 2A

For each y value, we calculate with its v value put it into the same data set, and finally print it out with ggplot.

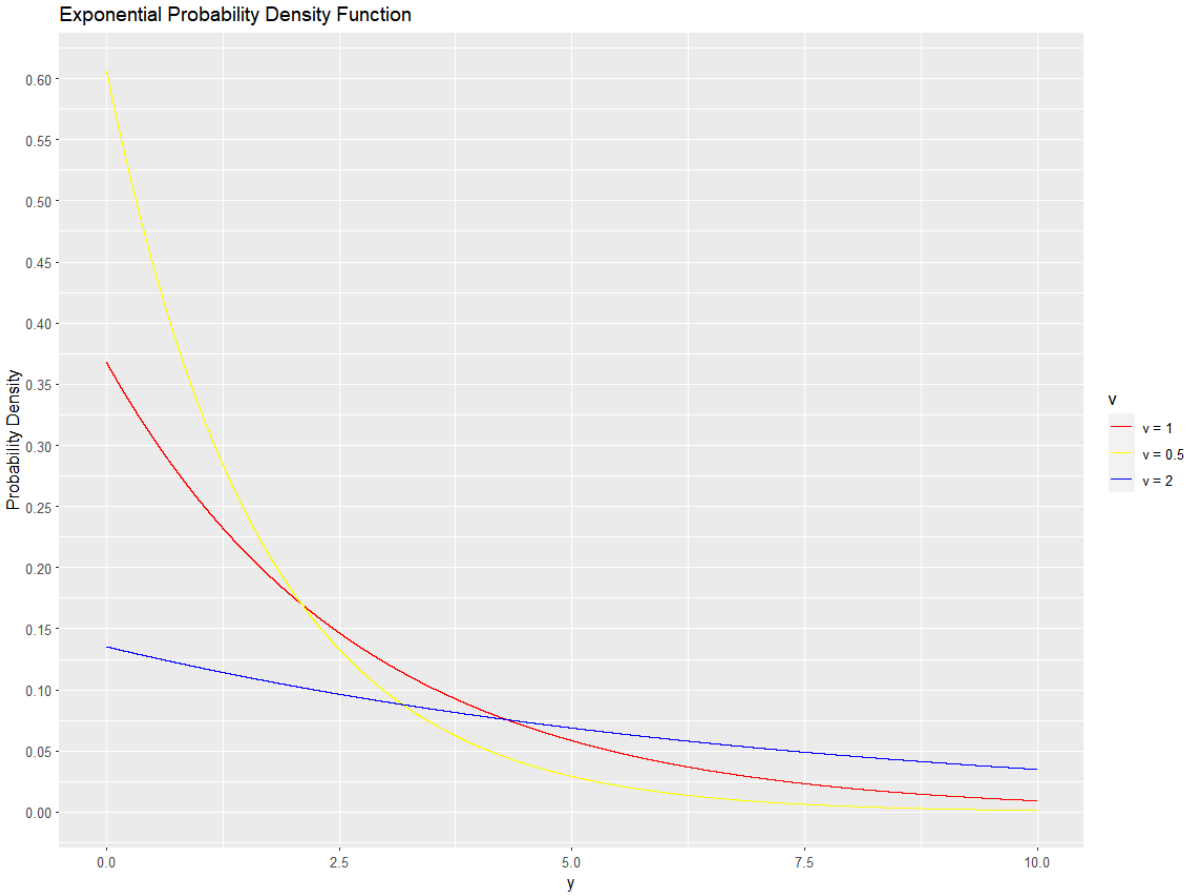
```
# Question 2A
library(ggplot2)

y <- seq(0, 10, by = 0.001)
v_values <- c(1, 0.5, 2)
data <- data.frame()

for (v in v_values) {
density <- exp((-exp(-v) * y) - v)
data <- rbind(data, data.frame(y = y, density = density, v = as.factor(v)))
}

ggplot(data, aes(x = y, y = density, color = v)) +
geom_line() +
labs(
x = "y",
y = "Probability Density",
title = "Exponential Probability Density Function",
```

```
) +
scale_color_manual(
values = c("1" = "red", "0.5" = "yellow", "2" = "blue"),
labels = c("v = 1", "v = 0.5", "v = 2")
)+
scale_y_continuous(breaks = seq(0, 0.7, by = 0.05))
```



Question 2B

L(v ; y) = ∏(i=1, n) p(yi|v) = ∏(i=1, n) exp ((−e ^ −v) yi − v)

= exp [((− e ^ −v) y1 − v) + ((− e ^ −v) y2 − v) + … + ((− e ^ −v) yn − v)]

= exp [(− e ^ −v) y1 + (− e ^ −v) y2 + … + (− e ^ −v) yn − nv]

= exp [(− e ^ −v) (y1 + y2 + … + yn) − nv]

= exp [-∑(i=1, n) (e ^ −v) yi − nv]

= exp [- ((e ^ −v) ∑(i=1, n) yi + nv)]

Question 2C

− logL(v ; y)

= - log { exp [- ((e ^ −v) ∑(i=1, n) yi + nv)] }

= - (- ((e ^ −v) ∑(i=1, n) yi + nv))

= (e ^ −v) ∑(i=1, n) yi + nv

Question 2D

NL (v ; y) = − logL(v ; y) = (e ^ −v) ∑(i=1, n) yi + nv

d/dv NL (v ; y) = - e ^ −v ∑(i=1, n) yi + n = 0

- e ^ −v ∑(i=1, n) yi = n
- e ^ −v = - n / ∑(i=1, n) yi
- - v = ln (n / ∑(i=1, n) yi)
- V_estimator = - ln (n / ∑(i=1, n) yi)

Question 2E

Bias and variance of an estimator

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}\right] - \theta, \quad \mathrm{Var}_\theta(\hat{\theta}) = \mathbb{V}\left[\hat{\theta}\right]$$

In this case, v is the true parameter, and V_estimator is the estimator

E[V_estimator]

= E[- ln (n / ∑(i=1, n) yi)]

= - E[ln (n / ∑(i=1, n) yi)]

= - E[ln(n) – ln (∑(i=1, n) yi)]

= - ln(n) + E [ln (∑(i=1, n) yi)]

∵ E [Y] = e ^ v

∴ - E[ln (n / ∑(i=1, n) yi)] = - ln (n/(e^v)) = - ln(n) + v

∴ Bias(V_estimator) = - ln(n) + v – v = - ln(n)

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta}(Y) - \mathbb{E}\left[\hat{\theta}(Y)\right]\right)^2\right] = \mathbb{V}\left[\hat{\theta}(Y)\right]$$

Var(V_estimator) = E[(V_estimator - E[V_estimator] ^ 2)]

= E[(v – (- ln(n)))^2]

= E[(ln(n) - ln (n / ∑(i=1, n) yi)))^2]

= E[(ln(∑(i=1, n) yi))^2]

= ln^2(E[∑(i=1, n) yi])

= ln^2 (e^v)

= (ln(e^v))^2

= v^2

Question 3

Question 3A

We can calculate based on the question so below is the formula we going to use.

$$\left(\hat{\mu}_{\mathrm{ML}} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

```
# Question 3A
# 124 volunteer, 80 to the right
n <- 124
x <- 80

p <- x / n
# Calculate the standard error
se <- sqrt((p * (1 - x / n)) / n)

# Calculate the confidence interval
lower <- p - 1.96 * se
upper <- p + 1.96 * se
p
cat("[", lower, ",", upper, "]\n")
> p
[1] 0.6451613
> cat("[", lower, ",", upper, "]\n")
[ 0.5609452 , 0.7293773 ]
```

The estimated proportion of humans turning their heads to the right is 0.6451613. We are 95% confident that the true population mean within this group falls between 0.5609452 and 0.7293773.

Question 3B

Hypotheses:

- Null Hypothesis (H0): The proportion of couples turning their heads to the right when kissing is equal to 0.5.

- Alternative Hypothesis (H1): The proportion of couples turning their heads to the right when kissing is not equal to 0.5.

$$z_{\hat{\theta}} = \frac{\hat{\theta} - \theta_0}{\sqrt{\theta_0(1-\theta_0)/n}}$$

Based on the z-score formula we can calculate the P value

```
# Question 3B

p_0 <- 0.5

z <- (p - p_0) / sqrt((p_0 * (1 - p_0)) / n)

p_value <- 2 * (1 - pnorm(abs(z)))

z

p_value

> z

[1] 3.232895

> p_value

[1] 0.001225424
```

We accept the alternative hypothesis over the null hypothesis, indicating that humans do not turn their heads to a specific side when kissing.

Question 3C

```
# Question 3C

binom.test(x, n, p = 0.5)$p.value

> binom.test(x, n, p = 0.5)$p.value

[1] 0.001564734
```

The p-value calculated for the preference for head tilting when kissing is approximately 0.0012 (approximate p-value) or 0.0016 (exact p-value). These p-values suggest that the chance of observing a preference stronger if the sample is larger, but it is still very low which gives the same conclusion.

Question 3D

To test whether the proportion of people who prefer using their right-hand matches the proportion of people who tilt their head to the right when kissing, we consider individuals who are right-handed in equal proportion when tilting their head to the right during a kiss as the null hypothesis.

Hypotheses:

- Null Hypothesis (H0): θ right_hand = θ right_head
- Alternative Hypothesis (H1): θ right_hand != θ right_head

H0 : θright_hand = θright_head vs H1 : θA != θB

We can use the below formula to calculate the z-score

$$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{\hat{\theta}_x - \hat{\theta}_y}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(1/n_x + 1/n_y)}}$$

```
# Question 3D
p_head <- x/n
p_hand <- 83/(83+17)
p_hat <- (x+83)/(n+83+17)
z <- (p_hand - p_head) / sqrt(p_hat*(1 - p_hat)*(1/n + 1/(83+17)))
z
2 * (pnorm(-abs(z)))
 > z
[1] 3.089364
> 2 * (pnorm(-abs(z)))
[1] 0.002005856
```

This p-value (0.002) is smaller than 0.05. So, we would reject the null hypothesis positing that the rate of right-handedness in the population is equivalent to the preference for turning heads to the right during kissing. The results suggest that there is a relationship between right-handedness and the preference for head-turning direction when kissing.