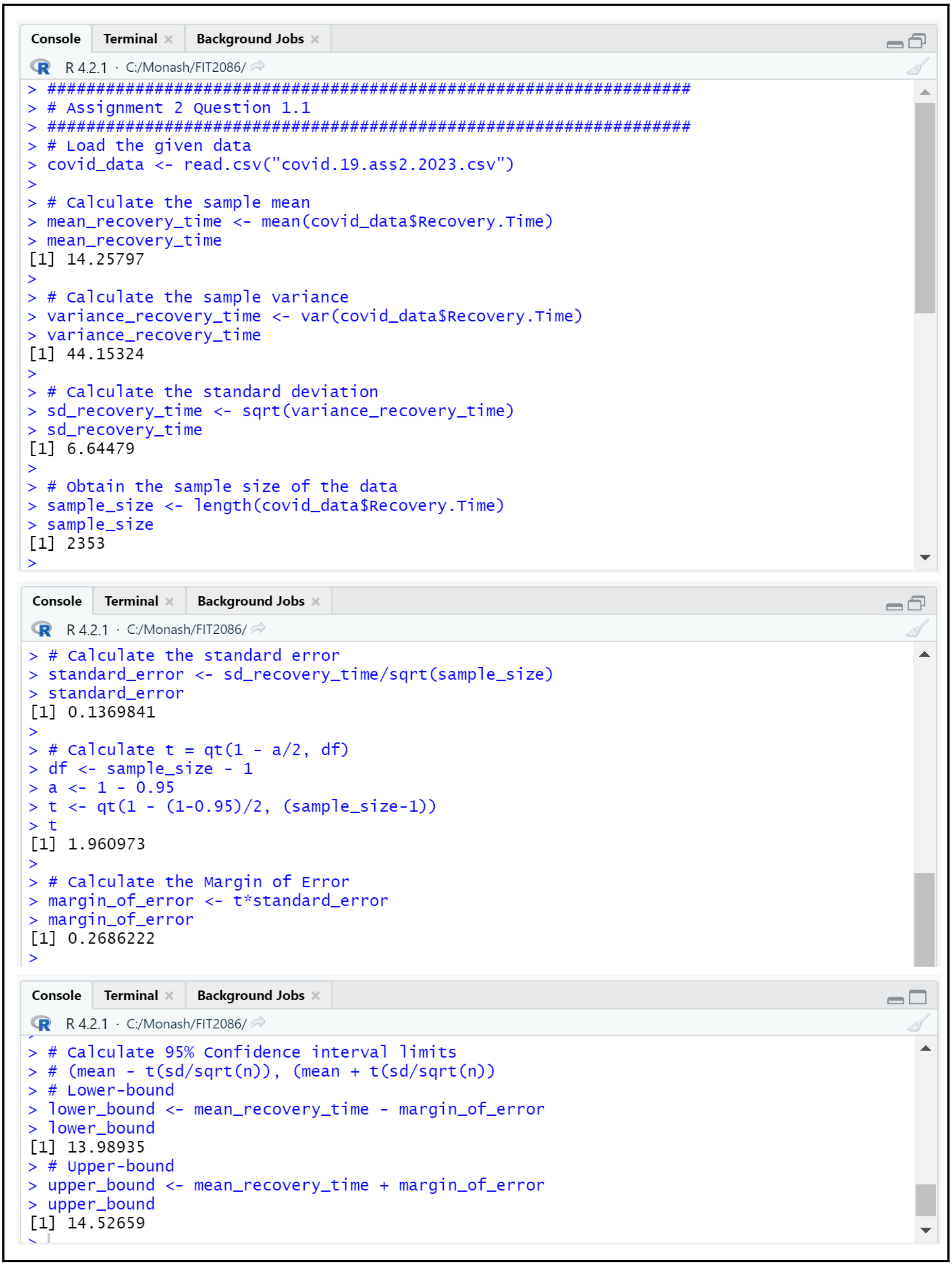
Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

FIT2086: Modelling for data analysis

Assignment 2

Question 1.1



```
> ##################################################################
> # Assignment 2 Question 1.1
> ##################################################################
> # Load the given data
> covid_data <- read.csv("covid.19.ass2.2023.csv")
>
> # Calculate the sample mean
> mean_recovery_time <- mean(covid_data$Recovery.Time)
> mean_recovery_time
[1] 14.25797
>
> # Calculate the sample variance
> variance_recovery_time <- var(covid_data$Recovery.Time)
> variance_recovery_time
[1] 44.15324
>
> # Calculate the standard deviation
> sd_recovery_time <- sqrt(variance_recovery_time)
> sd_recovery_time
[1] 6.64479
>
> # Obtain the sample size of the data
> sample_size <- length(covid_data$Recovery.Time)
> sample_size
[1] 2353
>
```

```
> # Calculate the standard error
> standard_error <- sd_recovery_time/sqrt(sample_size)
> standard_error
[1] 0.1369841
>
> # Calculate t = qt(1 - a/2, df)
> df <- sample_size - 1
> a <- 1 - 0.95
> t <- qt(1 - (1-0.95)/2, (sample_size-1))
> t
[1] 1.960973
>
> # Calculate the Margin of Error
> margin_of_error <- t*standard_error
> margin_of_error
[1] 0.2686222
>
```

```
> # Calculate 95% Confidence interval limits
> # (mean - t(sd/sqrt(n)), (mean + t(sd/sqrt(n))
> # Lower-bound
> lower_bound <- mean_recovery_time - margin_of_error
> lower_bound
[1] 13.98935
> # Upper-bound
> upper_bound <- mean_recovery_time + margin_of_error
> upper_bound
[1] 14.52659
>
```

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

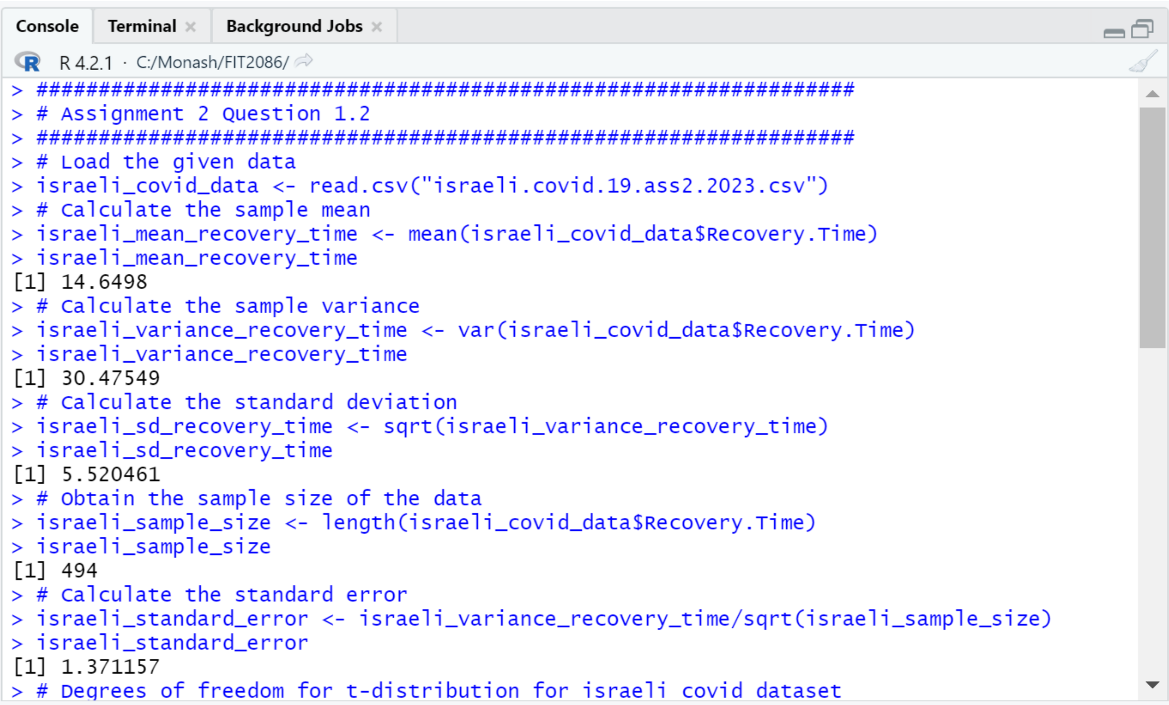
From the calculation done in R programming for Question 1.1, it can be said that:

- With 95% confidence, it can be noted that the estimated number of days to recover from Covid-19 in the New South Wales (NSW) population will fall within the interval of 13.98935 to 14.52659 days.
- It also can be noted that the estimated average number of days to recover from Covid-19 in the NSW population is around 14.257971 days.

Estimated average number of days to recover from Covid-19:

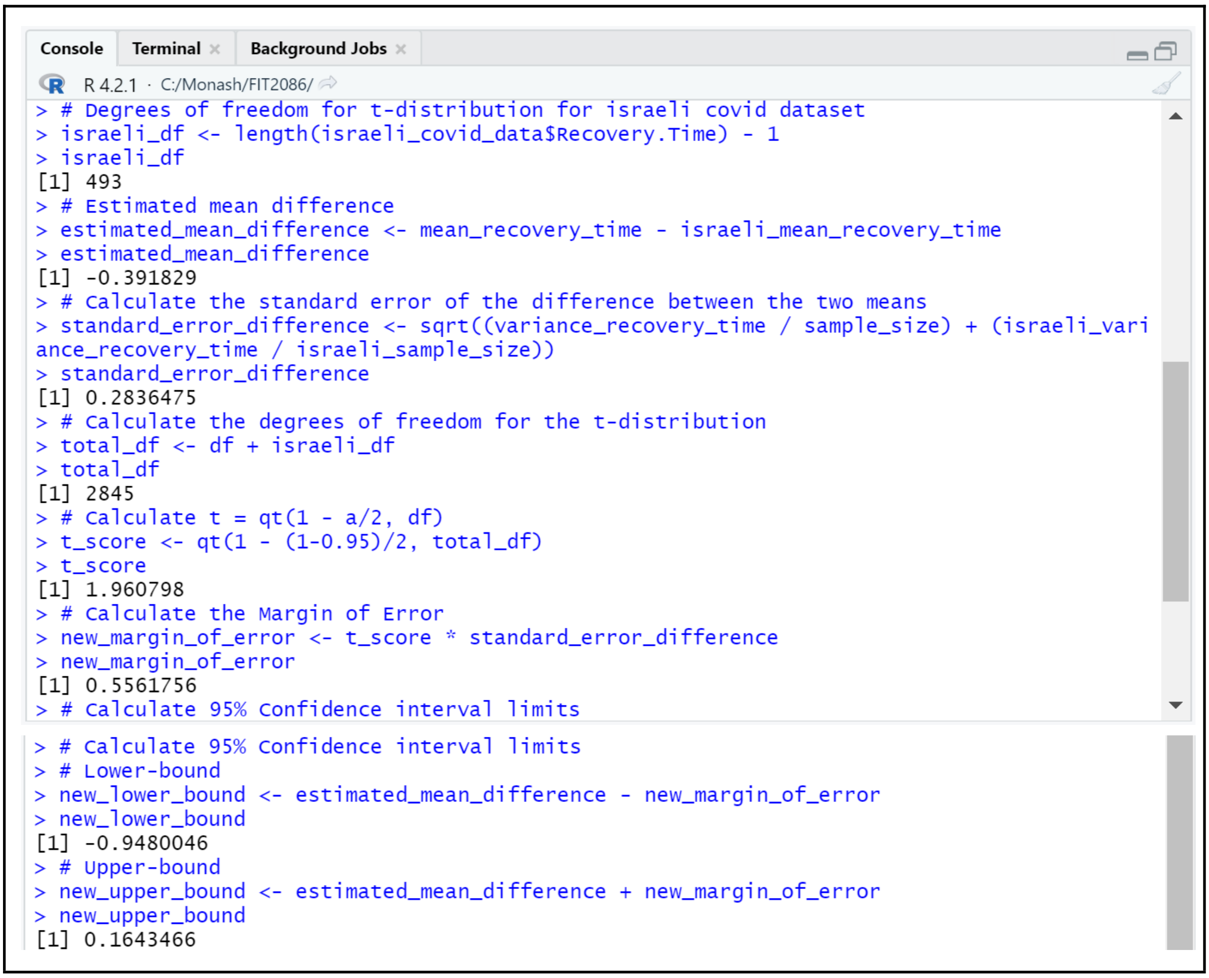
$$\frac{13.98935 + 14.52659}{2} = 14.257971$$

Question 1.2



```
> ##################################################################
> # Assignment 2 Question 1.2
> ##################################################################
> # Load the given data
> israeli_covid_data <- read.csv("israeli.covid.19.ass2.2023.csv")
> # Calculate the sample mean
> israeli_mean_recovery_time <- mean(israeli_covid_data$Recovery.Time)
> israeli_mean_recovery_time
[1] 14.6498
> # Calculate the sample variance
> israeli_variance_recovery_time <- var(israeli_covid_data$Recovery.Time)
> israeli_variance_recovery_time
[1] 30.47549
> # Calculate the standard deviation
> israeli_sd_recovery_time <- sqrt(israeli_variance_recovery_time)
> israeli_sd_recovery_time
[1] 5.520461
> # Obtain the sample size of the data
> israeli_sample_size <- length(israeli_covid_data$Recovery.Time)
> israeli_sample_size
[1] 494
> # Calculate the standard error
> israeli_standard_error <- israeli_variance_recovery_time/sqrt(israeli_sample_size)
> israeli_standard_error
[1] 1.371157
> # Degrees of freedom for t-distribution for israeli covid dataset
```

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu



```
> # Degrees of freedom for t-distribution for israeli covid dataset
> israeli_df <- length(israeli_covid_data$Recovery.Time) - 1
> israeli_df
[1] 493
> # Estimated mean difference
> estimated_mean_difference <- mean_recovery_time - israeli_mean_recovery_time
> estimated_mean_difference
[1] -0.391829
> # Calculate the standard error of the difference between the two means
> standard_error_difference <- sqrt((variance_recovery_time / sample_size) + (israeli_vari
ance_recovery_time / israeli_sample_size))
> standard_error_difference
[1] 0.2836475
> # Calculate the degrees of freedom for the t-distribution
> total_df <- df + israeli_df
> total_df
[1] 2845
> # Calculate t = qt(1 - a/2, df)
> t_score <- qt(1 - (1-0.95)/2, total_df)
> t_score
[1] 1.960798
> # Calculate the Margin of Error
> new_margin_of_error <- t_score * standard_error_difference
> new_margin_of_error
[1] 0.5561756
> # Calculate 95% Confidence interval limits
```

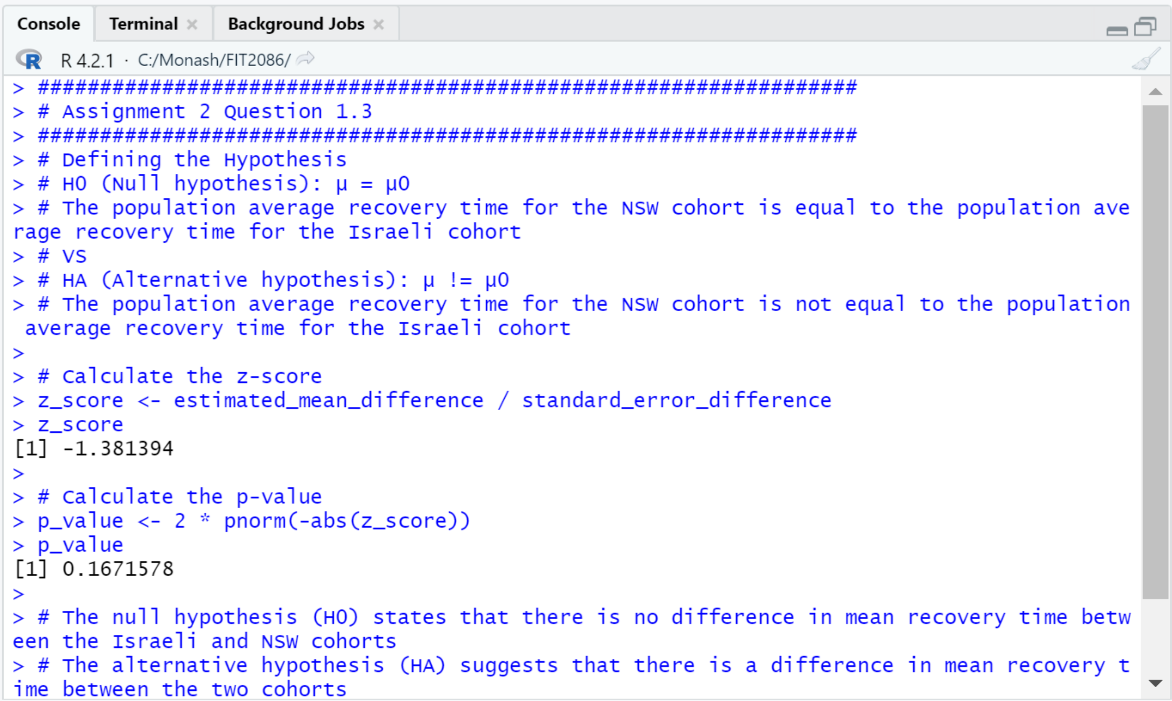
```
> # Calculate 95% Confidence interval limits
> # Lower-bound
> new_lower_bound <- estimated_mean_difference - new_margin_of_error
> new_lower_bound
[1] -0.9480046
> # Upper-bound
> new_upper_bound <- estimated_mean_difference + new_margin_of_error
> new_upper_bound
[1] 0.1643466
```

From the calculation done in R programming for Question 1.2, it can be said that:

- The estimated mean difference in recovery times from Covid-19 between Israeli and NSW patients is -0.391829 which is close to 0 which suggests that there does not exist a particularly large difference in recovery times from Covid-19 between both population groups.
- With 95% confidence, it can be noted that the estimated mean difference in recovery times from Covid-19 between Israeli and NSW patients is within the interval of -0.9480046 and 0.1643466.

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

Question 1.3



```
> ###################################################################
> # Assignment 2 Question 1.3
> ###################################################################
> # Defining the Hypothesis
> # H0 (Null hypothesis): μ = μ0
> # The population average recovery time for the NSW cohort is equal to the population ave
rage recovery time for the Israeli cohort
> # VS
> # HA (Alternative hypothesis): μ != μ0
> # The population average recovery time for the NSW cohort is not equal to the population
 average recovery time for the Israeli cohort
>
> # Calculate the z-score
> z_score <- estimated_mean_difference / standard_error_difference
> z_score
[1] -1.381394
>
> # Calculate the p-value
> p_value <- 2 * pnorm(-abs(z_score))
> p_value
[1] 0.1671578
>
> # The null hypothesis (H0) states that there is no difference in mean recovery time betw
een the Israeli and NSW cohorts
> # The alternative hypothesis (HA) suggests that there is a difference in mean recovery t
ime between the two cohorts
```

Null Hypothesis (H_0): μ_israeli = μ_nsw

- The null hypothesis assumes that there is no difference in the population average recovery time between the Israeli population group and the NSW population group.

Alternative Hypothesis (H_A): μ_israeli ≠ μ_nsw

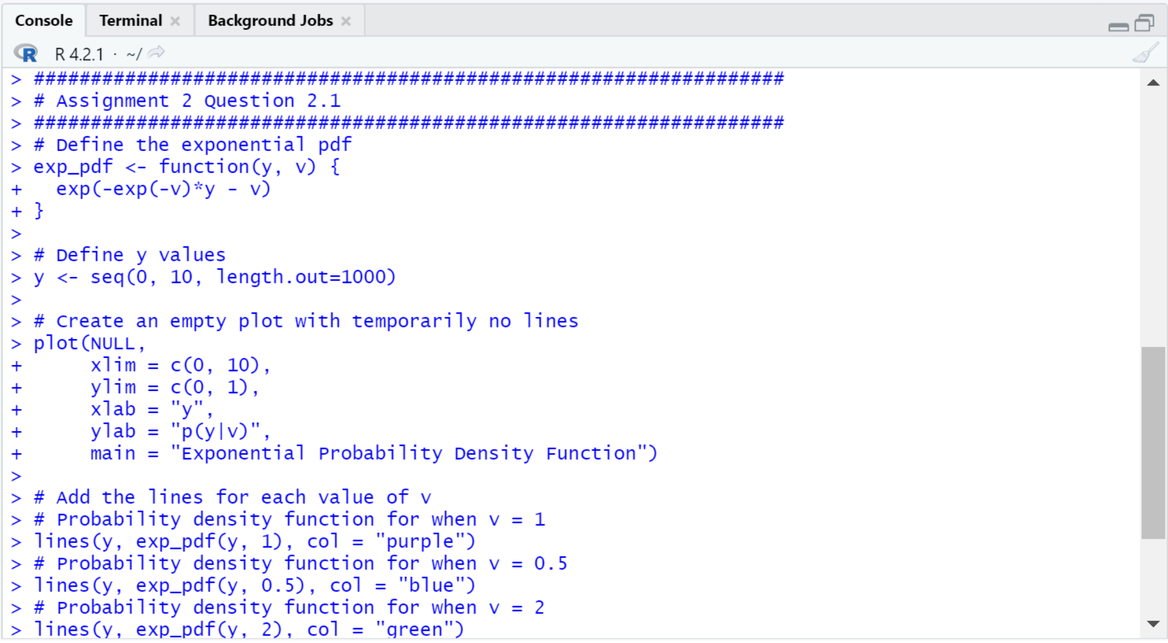
- The alternative hypothesis suggests that there is a difference in the population average recovery time between the two population groups.

From the calculation done in R programming for Question 1.3, it can be said that:

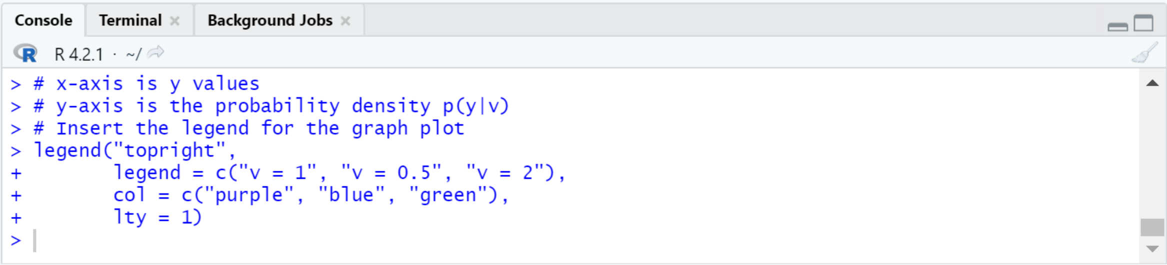
- The z-score is -1.381394 and since the z-score is a negative value, it suggests that the estimated mean difference is below the hypothesised mean difference.
- The p_value is 0.1671578 which is relatively large as it is bigger than 0.05 which strongly suggest that there exists not enough evidence to reject the null hypothesis (H_0)
- Hence, it further suggests that there exists statistical evidence that is not strong enough to prove that the NSW population average recovery time is different from the Israeli population average recovery time.

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

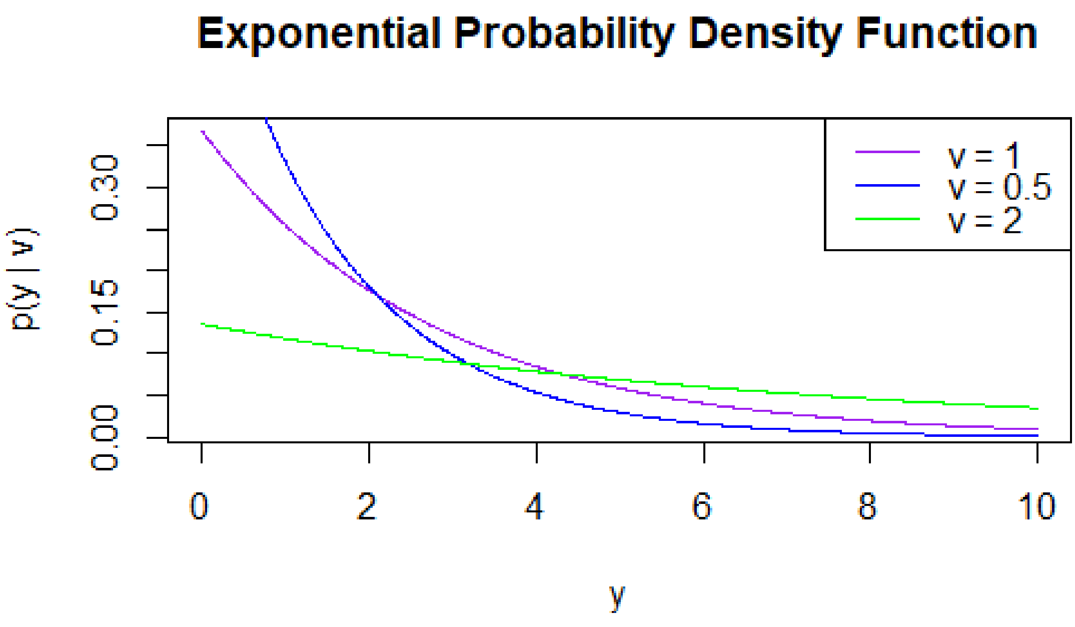
Question 2.1



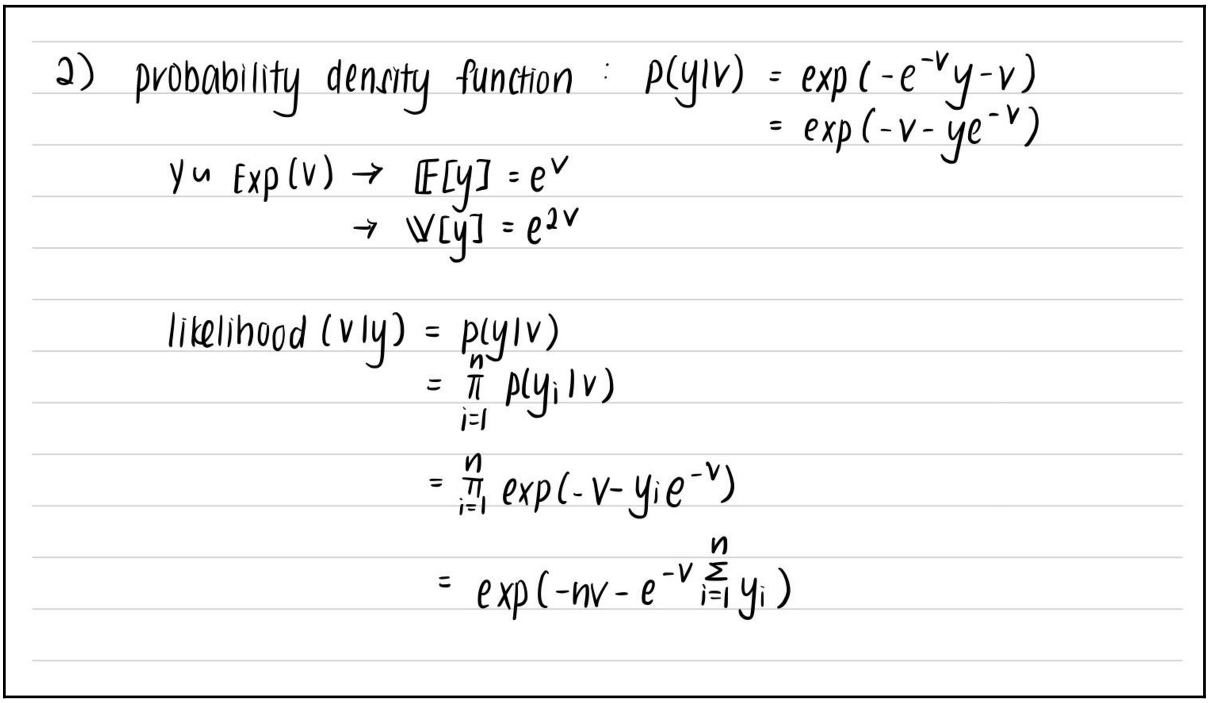
```
> ##################################################################
> # Assignment 2 Question 2.1
> ##################################################################
> # Define the exponential pdf
> exp_pdf <- function(y, v) {
+   exp(-exp(-v)*y - v)
+ }
>
> # Define y values
> y <- seq(0, 10, length.out=1000)
>
> # Create an empty plot with temporarily no lines
> plot(NULL,
+      xlim = c(0, 10),
+      ylim = c(0, 1),
+      xlab = "y",
+      ylab = "p(y|v)",
+      main = "Exponential Probability Density Function")
>
> # Add the lines for each value of v
> # Probability density function for when v = 1
> lines(y, exp_pdf(y, 1), col = "purple")
> # Probability density function for when v = 0.5
> lines(y, exp_pdf(y, 0.5), col = "blue")
> # Probability density function for when v = 2
> lines(y, exp_pdf(y, 2), col = "green")
```



```
> # x-axis is y values
> # y-axis is the probability density p(y|v)
> # Insert the legend for the graph plot
> legend("topright",
+        legend = c("v = 1", "v = 0.5", "v = 2"),
+        col = c("purple", "blue", "green"),
+        lty = 1)
> |
```



Question 2.2



2) probability density function : $p(y|v) = \exp(-e^{-v}y - v)$

$$= \exp(-v - ye^{-v})$$

$$y \sim \text{Exp}(v) \rightarrow \mathbb{E}[y] = e^{v}$$

$$\rightarrow \mathbb{V}[y] = e^{2v}$$

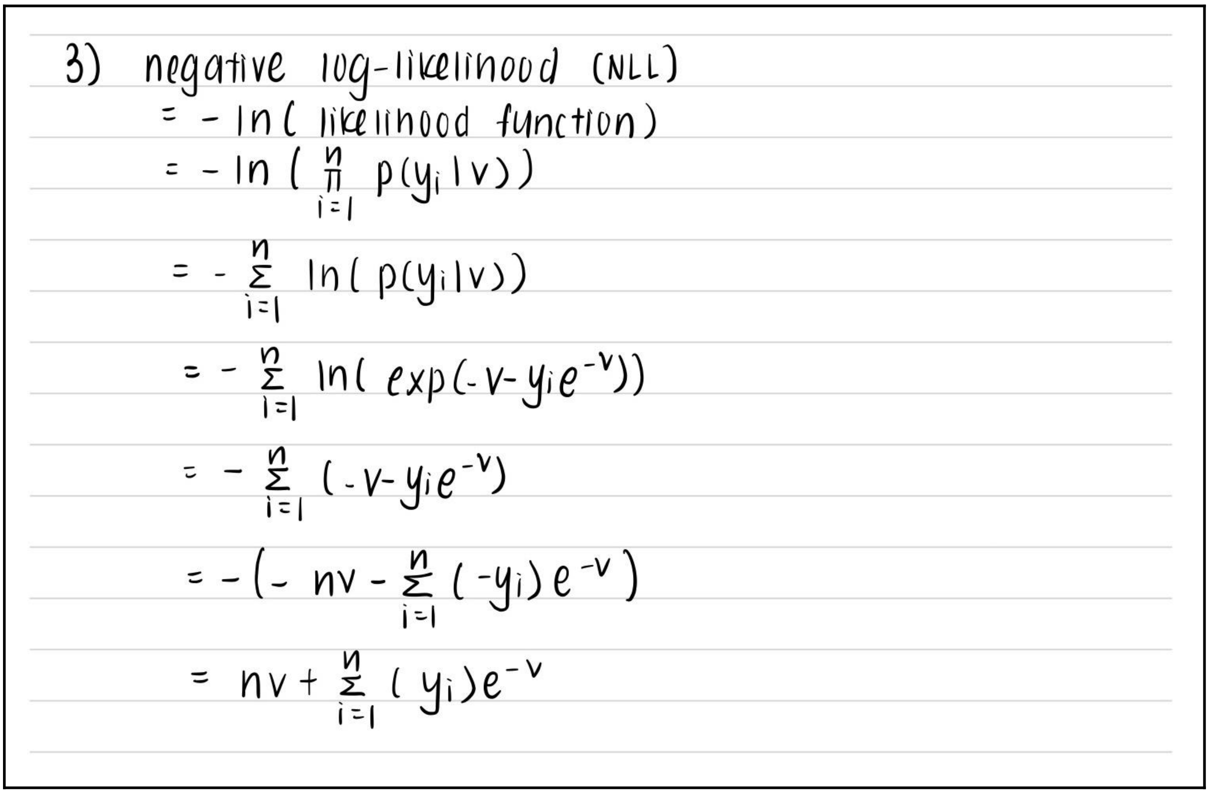
$$\text{likelihood}(v|y) = p(y|v)$$

$$= \prod_{i=1}^{n} p(y_i|v)$$

$$= \prod_{i=1}^{n} \exp(-v - y_i e^{-v})$$

$$= \exp(-nv - e^{-v}\sum_{i=1}^{n} y_i)$$

Question 2.3



3) negative log-likelihood (NLL)

$$= -\ln(\text{likelihood function})$$

$$= -\ln\left(\prod_{i=1}^{n} p(y_i|v)\right)$$

$$= -\sum_{i=1}^{n} \ln(p(y_i|v))$$

$$= -\sum_{i=1}^{n} \ln(\exp(-v - y_i e^{-v}))$$

$$= -\sum_{i=1}^{n} (-v - y_i e^{-v})$$

$$= -\left(-nv - \sum_{i=1}^{n} (-y_i) e^{-v}\right)$$

$$= nv + \sum_{i=1}^{n} (y_i) e^{-v}$$

Question 2.4

4) likelihood function $(y|v) = \exp(-nv - e^{-v}\sum_{i=1}^{n} y_i)$

negative log-likelihood : $nv + \sum_{i=1}^{n} y_i e^{-v}$

minimise negative log-likelihood function :

$$\frac{d}{dv} = n + \sum_{i=1}^{n} y_i \frac{e^{-v}}{-1}$$

$$= n - \sum_{i=1}^{n} y_i e^{-v}$$

let $d/dv = 0$

$$0 = n - \sum_{i=1}^{n} y_i e^{-v}$$

$$n = \sum_{i=1}^{n} y_i e^{-v}$$

$$\frac{n}{\sum_{i=1}^{n} y_i} = e^{-v}$$

$$\log_e\left(\frac{n}{\sum_{i=1}^{n} y_i}\right) = \log_e(e^{-v})$$

$$\log_e\left(\frac{n}{\sum_{i=1}^{n} y_i}\right) = -v \log_e(e)$$

$$\log_e\left(\frac{n}{\sum_{i=1}^{n} y_i}\right) = -v \longrightarrow v = -\ln\left(\frac{n}{\sum_{i=1}^{n} y_i}\right)$$

$$v = \ln\left(\frac{n}{\sum_{i=1}^{n} y_i}\right)^{-1}$$

$$v = \ln\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \#$$

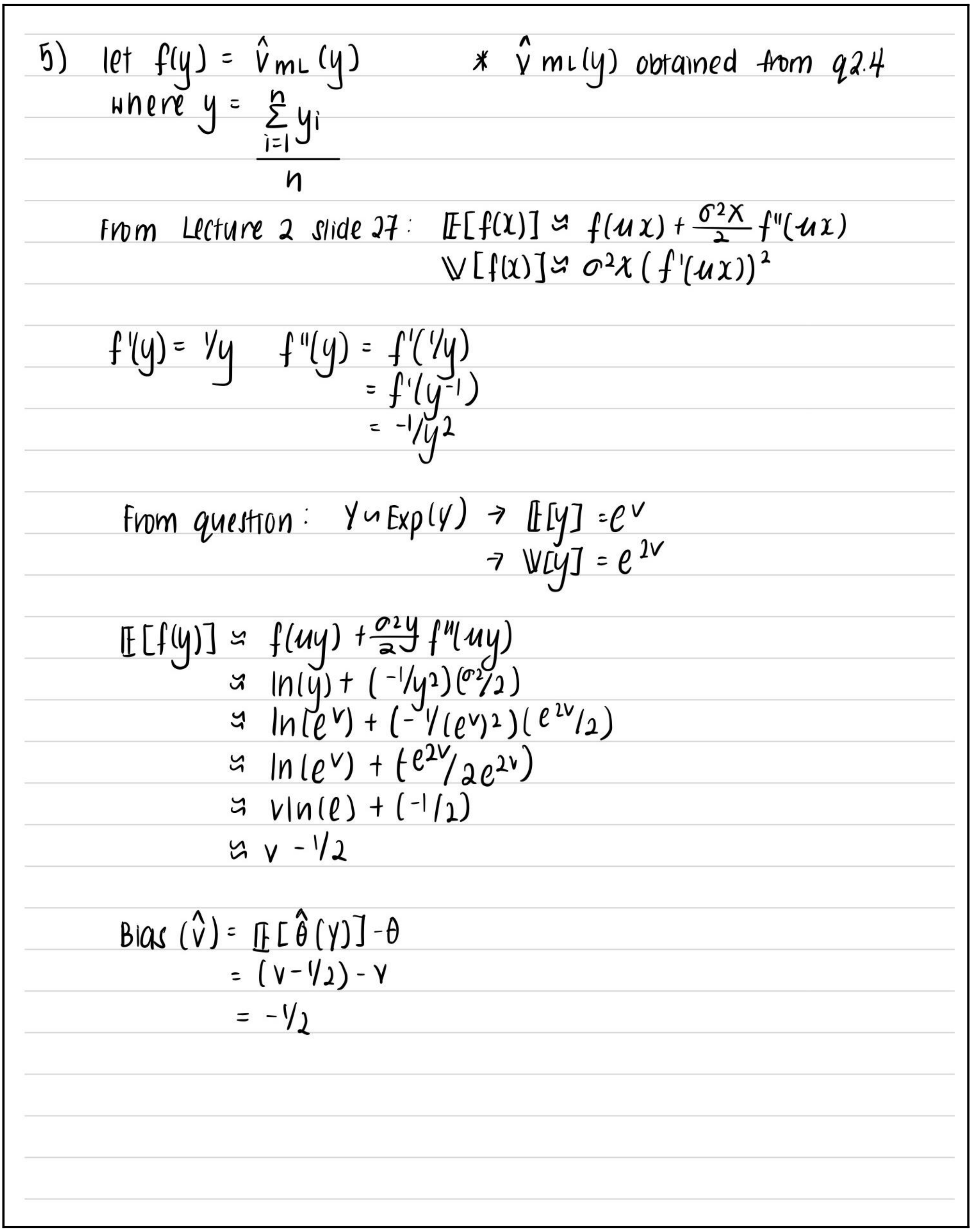
$$\therefore \hat{v}_{ML}(y) = \ln\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)$$

Student Name: Foo Kai Yan

Student ID: 33085625

Student Email: kfoo0012@student.monash.edu

Question 2.5



5) let $f(y) = \hat{v}_{mL}(y)$ $\quad$ * $\hat{v}_{mL}(y)$ obtained from q2.4

where $y = \dfrac{\sum_{i=1}^{n} y_i}{n}$

From Lecture 2 slide 27: $\mathbb{E}[f(x)] \approx f(\mu x) + \frac{\sigma^2 x}{2} f''(\mu x)$

$\mathbb{V}[f(x)] \approx \sigma^2 x \, (f'(\mu x))^2$

$f'(y) = 1/y \quad f''(y) = f'(1/y)$

$= f'(y^{-1})$

$= -1/y^2$

From question: $Y \sim Exp(v) \rightarrow \mathbb{E}[y] = e^v$

$\rightarrow \mathbb{V}[y] = e^{2v}$

$\mathbb{E}[f(y)] \approx f(\mu y) + \frac{\sigma^2 y}{2} f''(\mu y)$

$\approx \ln(y) + (-1/y^2)(\sigma^2/2)$

$\approx \ln(e^v) + (-1/(e^v)^2)(e^{2v}/2)$

$\approx \ln(e^v) + (-e^{2v}/2e^{2v})$

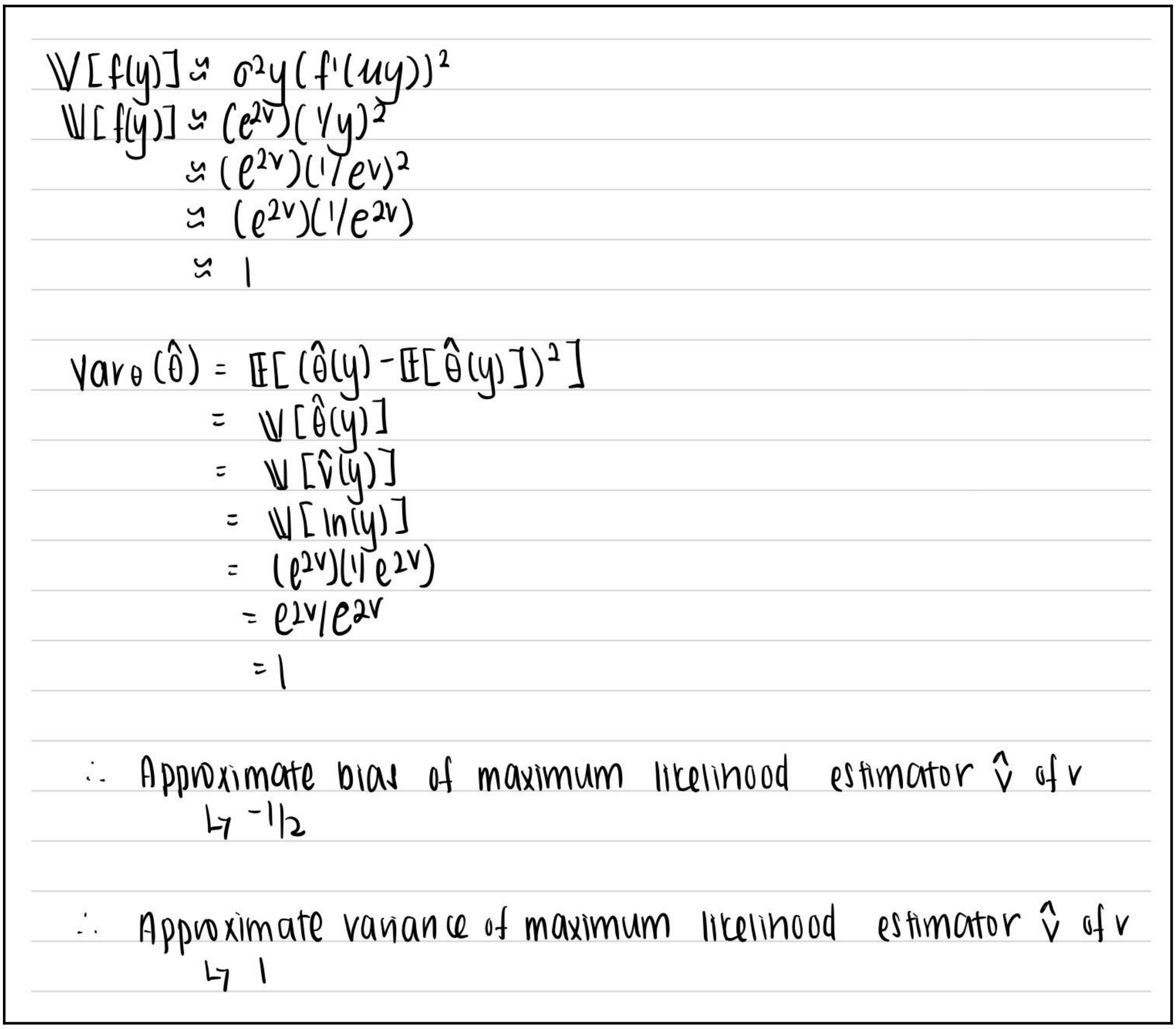
$\approx v\ln(e) + (-1/2)$

$\approx v - 1/2$

$\text{Bias}(\hat{v}) = \mathbb{E}[\hat{\theta}(Y)] - \theta$

$= (v - 1/2) - v$

$= -1/2$



$$\mathbb{V}[f(y)] \simeq \sigma^2 y(f'(\mu y))^2$$

$$\mathbb{V}[f(y)] \simeq (e^{2v})(1/y)^2$$

$$\simeq (e^{2v})(1/e^{v})^2$$

$$\simeq (e^{2v})(1/e^{2v})$$

$$\simeq 1$$

$$\mathrm{Var}_\theta(\hat{\theta}) = \mathbb{E}[(\hat{\theta}(y) - \mathbb{E}[\hat{\theta}(y)])^2]$$

$$= \mathbb{V}[\hat{\theta}(y)]$$

$$= \mathbb{V}[\hat{v}(y)]$$

$$= \mathbb{V}[\ln(y)]$$

$$= (e^{2v})(1/e^{2v})$$

$$= e^{2v}/e^{2v}$$

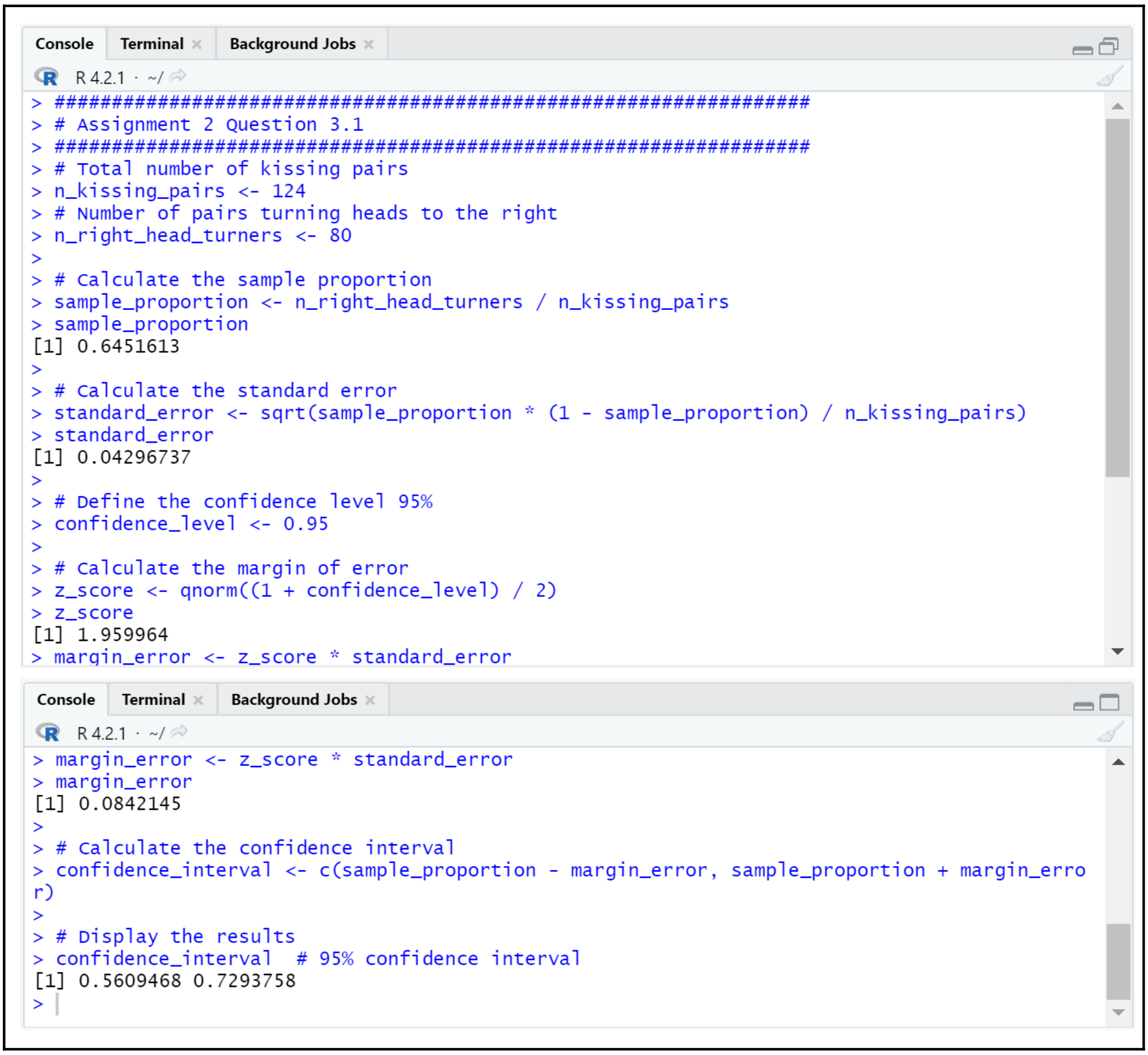
$$= 1$$

$\therefore$ Approximate bias of maximum likelihood estimator $\hat{v}$ of v
↳ $-1/2$

$\therefore$ Approximate variance of maximum likelihood estimator $\hat{v}$ of v
↳ 1

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

Question 3.1



```
> #################################################################
> # Assignment 2 Question 3.1
> #################################################################
> # Total number of kissing pairs
> n_kissing_pairs <- 124
> # Number of pairs turning heads to the right
> n_right_head_turners <- 80
>
> # Calculate the sample proportion
> sample_proportion <- n_right_head_turners / n_kissing_pairs
> sample_proportion
[1] 0.6451613
>
> # Calculate the standard error
> standard_error <- sqrt(sample_proportion * (1 - sample_proportion) / n_kissing_pairs)
> standard_error
[1] 0.04296737
>
> # Define the confidence level 95%
> confidence_level <- 0.95
>
> # Calculate the margin of error
> z_score <- qnorm((1 + confidence_level) / 2)
> z_score
[1] 1.959964
> margin_error <- z_score * standard_error
```

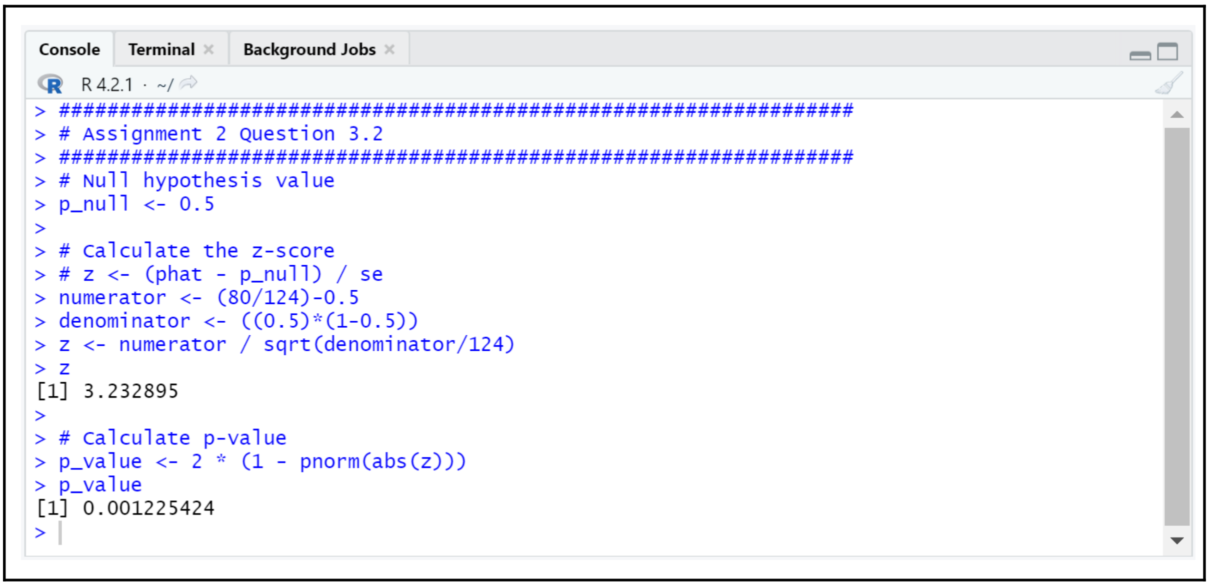
```
> margin_error <- z_score * standard_error
> margin_error
[1] 0.0842145
>
> # Calculate the confidence interval
> confidence_interval <- c(sample_proportion - margin_error, sample_proportion + margin_erro
r)
>
> # Display the results
> confidence_interval  # 95% confidence interval
[1] 0.5609468 0.7293758
>
```

From the calculation done in R programming for Question 3.1, it can be said that:

- The sample proportion of kissing pairs that turned their heads to the right is 0.6451613 which shows that the kissing pairs in the sample population that turned their heads to the right when kissing is 64.5%
- The standard error of the sample proportion is 0.04296737 which is the variability in the sample proportion due to random sampling
- With 95% confidence, it can be noted that the true proportion of kissing pairs from the population turns their head to the right when kissing falls within the interval of 0.5609468 and 0.7293758

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

Question 3.2



```
> ####################################################################
> # Assignment 2 Question 3.2
> ####################################################################
> # Null hypothesis value
> p_null <- 0.5
>
> # Calculate the z-score
> # z <- (phat - p_null) / se
> numerator <- (80/124)-0.5
> denominator <- ((0.5)*(1-0.5))
> z <- numerator / sqrt(denominator/124)
> z
[1] 3.232895
>
> # Calculate p-value
> p_value <- 2 * (1 - pnorm(abs(z)))
> p_value
[1] 0.001225424
>
```

Null Hypothesis (H_0): p = 0.5

- The null hypothesis assumes that there is no preference which means that all kissing pairs have equally likely chances, 50%, to tilt their heads to either left or right side when kissing

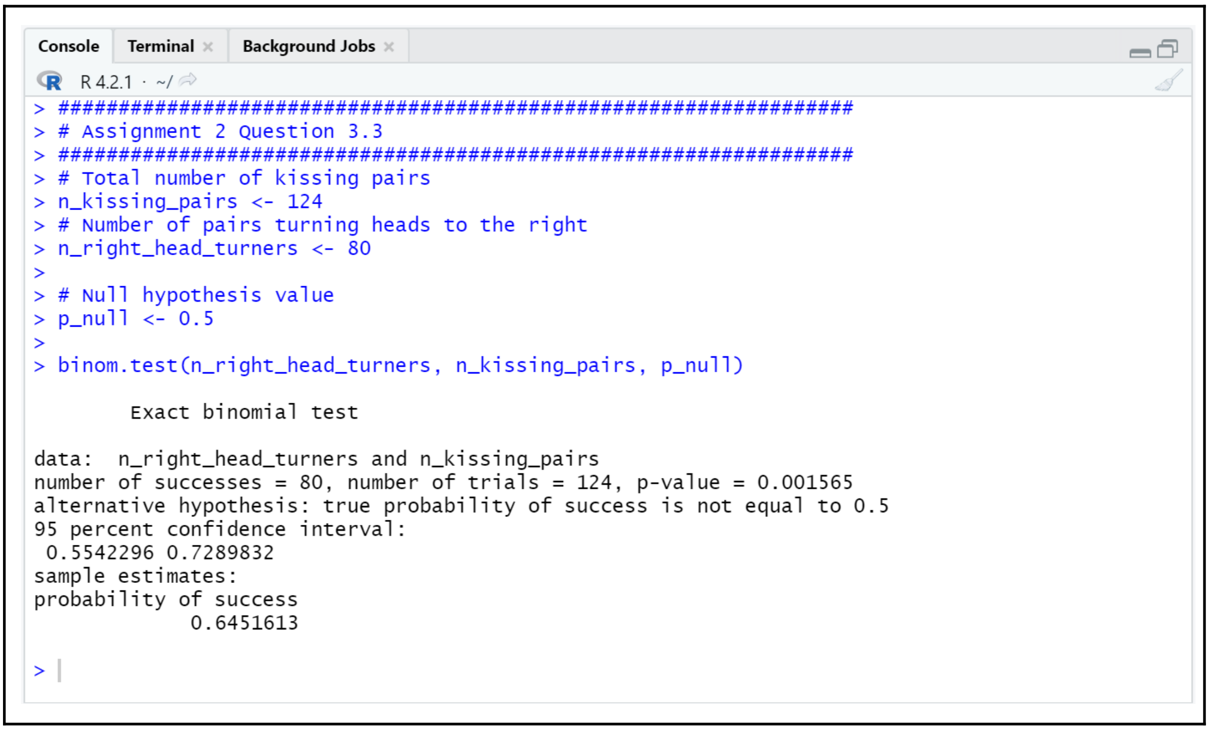
Alternative Hypothesis (H_A): p ≠ 0.5

- The alternative hypothesis suggests that there is a preference which means that all kissing pairs does not have equally likely chances to tilt their heads to either left or right side when kissing

The p-value is 0.001225424 which very small as it is smaller than 0.05 which strongly suggests that there exists a strong evidence against the null hypothesis (H_0). The small p-value suggests that the observed population of kissing pairs have a preference to tilt their head to either left or right when kissings. Hence the hypothesis on kissing pairs having no preference to tilt their heads during kissing is rejected as the data shows that there exists strong evidence that there is indeed a preference amongst the kissing pairs to tilt their heads to a particular direction during kissing.

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

Question 3.3



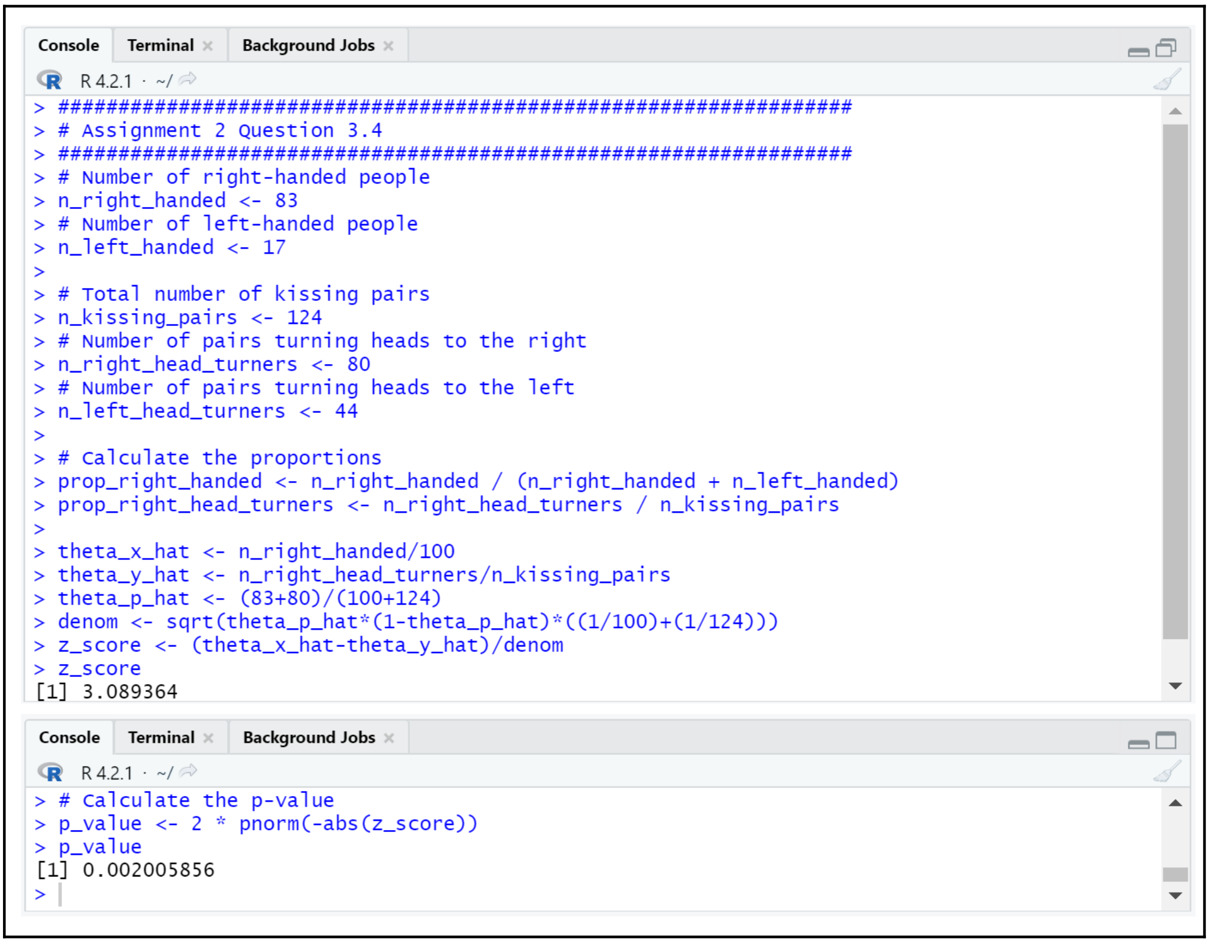
```
> ##################################################################
> # Assignment 2 Question 3.3
> ##################################################################
> # Total number of kissing pairs
> n_kissing_pairs <- 124
> # Number of pairs turning heads to the right
> n_right_head_turners <- 80
> 
> # Null hypothesis value
> p_null <- 0.5
> 
> binom.test(n_right_head_turners, n_kissing_pairs, p_null)

        Exact binomial test

data:  n_right_head_turners and n_kissing_pairs
number of successes = 80, number of trials = 124, p-value = 0.001565
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.5542296 0.7289832
sample estimates:
probability of success 
             0.6451613 

> 
```

n_right_head_turners represents 80 which is the number of successes
n_kissing_pairs represents 124 which is the total number of trials
p_null represents 0.5 which is the probability of success when kissing pairs have no preferences

binom.test is a measure of the probability of observing a result where there exists no preference for the kissings pairs to tilt their heads to either side during kissing. Since the exact p-value is typically used to determine the strength of evidence against the null hypothesis, this p-value is 0.001565 which is a relatively small as it is less than 0.05 so this p-value suggest that there exist a strong evidence against the null hypothesis which indicates that there exist a preference amongst the kissing pairs to tilt their heads to a particular direction during kissing. Hence, the null hypothesis is rejected and people will most likely tilt their heads to a particular direction one would prefer during kissing due to the existence of a clear preference.

Student Name: Foo Kai Yan
Student ID: 33085625
Student Email: kfoo0012@student.monash.edu

Question 3.4



```
> ####################################################################
> # Assignment 2 Question 3.4
> ####################################################################
> # Number of right-handed people
> n_right_handed <- 83
> # Number of left-handed people
> n_left_handed <- 17
> 
> # Total number of kissing pairs
> n_kissing_pairs <- 124
> # Number of pairs turning heads to the right
> n_right_head_turners <- 80
> # Number of pairs turning heads to the left
> n_left_head_turners <- 44
> 
> # Calculate the proportions
> prop_right_handed <- n_right_handed / (n_right_handed + n_left_handed)
> prop_right_head_turners <- n_right_head_turners / n_kissing_pairs
> 
> theta_x_hat <- n_right_handed/100
> theta_y_hat <- n_right_head_turners/n_kissing_pairs
> theta_p_hat <- (83+80)/(100+124)
> denom <- sqrt(theta_p_hat*(1-theta_p_hat)*((1/100)+(1/124)))
> z_score <- (theta_x_hat-theta_y_hat)/denom
> z_score
[1] 3.089364
```

```
> # Calculate the p-value
> p_value <- 2 * pnorm(-abs(z_score))
> p_value
[1] 0.002005856
> 
```

Null Hypothesis (H_0):

The null hypothesis assumes that there is no significant difference between the rate of right-handedness in the population and the preference for turning heads to the right when kissing.

The z-score is 3.089364 which is a positive value and with the p-value being 0.002005856, it indicates that this is a small p-value that is less than 0.05 which strongly suggests that there exists strong evidence against the null hypothesis. This also shows that there is a significant difference between the rate of right-handedness in the population and the preference for turning heads to the right when kissing. Hence, the null hypothesis is rejected.