Question 1

1.

First is to calculate the sample mean and sample standard deviation,

```
covid_data_nsw <- read.csv("covid.19.ass2.2023.csv")
sample_mean_nsw <- mean(covid_data_nsw$Recovery.Time)
sample_sd_nsw <- sd(covid_data_nsw$Recovery.Time)
```

sample_mean_...	14.2579685507862
sample_sd_nsw	6.64479020420164

I found that sample mean is 14.2580, and the sample standard deviation is 6.6448.

Then, the student-t value for α = 0.05 is calculated below,

```
n_nsw <- length(covid_data_nsw$Recovery.Time)
confidence_level <- 0.95
alpha <- 1 - confidence_level
t_score <- qt(1 - alpha / 2, df = n_nsw - 1)
```

alpha	0.05
confidence_l...	0.95
mu	3
n_nsw	2353L
sample_mean_...	14.2579685507862
sample_sd_nsw	6.64479020420164
sigma	4
t_score	1.96097311367882

We can get $t_{\frac{\alpha}{2},n-1} = 1.9610$

By the formula, we can calculate the confidence interval with unknown variance,

$$(\widehat{\mu_{ML}} - t_{\frac{\alpha}{2},n-1}\frac{\widehat{\sigma_n}}{\sqrt{n}},\ \widehat{\mu_{ML}} + t_{\frac{\alpha}{2},n-1}\frac{\widehat{\sigma_n}}{\sqrt{n}})$$

```
se_nsw <- sample_sd_nsw / sqrt(n_nsw)
confidence_interval <- c(sample_mean_nsw - t_score*se_nsw, sample_mean_nsw + t_score*se_nsw)
```

```
> se_nsw
[1] 0.1369841
> confidence_interval
[1] 13.98935 14.52659
```

The approximate confidence interval is (13.9894, 14.5266).

From the data given, the average recovery time of covid taken by the patients in New South Wales is 14.25797 days. There is 95% confidence that the population mean recovery time of covid in New South Wales is between 13.9894 days and 14.5266 days.

2.

First is to calculate the sample mean and the sample standard deviation,

```
covid_data_israel <- read.csv("israeli.covid.19.ass2.2023.csv")
sample_mean_israel <- mean(covid_data_israel$Recovery.Time)
sample_sd_israel <- sd(covid_data_israel$Recovery.Time)
```

```
sample_mean_israel
.] 14.6498
sample_sd_israel
.] 5.520461
```

I found that sample mean of second dataset is 14.6498, and the sample standard deviation of second dataset is 5.5205.

There are 494 of individuals from the Israel study (n = 494), the difference in average recovery time of covid taken by patient between the two sample, which is the mean difference is valued -0.3918, and the value of difference of standard error is 0.2837.

```
n_israel <- length(covid_data_israel$Recovery.Time)
mean_difference <- sample_mean_nsw - sample_mean_israel
se_difference <- sqrt((sample_sd_nsw^2 / n_nsw) + (sample_sd_israel^2 / n_israel))
```

```
> n_israel
[1] 494
> mean_difference
[1] -0.391829
> se_difference
[1] 0.2836475
```

The 95% confidence z-value is 1.96, by the formula below we can calculate the confidence interval between 2 difference normal means.

$$(\widehat{\mu_A} - \widehat{\mu_B} - Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}, \widehat{\mu_A} - \widehat{\mu_B} + Z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}})$$

The approximate confidence interval between 2 means is (-0.9478, 0.1641)

```
z_score <- qnorm(1 - alpha / 2)
confidence_interval_difference <- c(mean_difference - z_score*se_difference, mean_difference + z_score*se_difference)
```

```
> z_score
[1] 1.959964
> confidence_interval_difference
[1] -0.9477680  0.1641099
```

In summary, the estimated difference in COVID-19 recovery time between patients in New South Wales (NSW) and Israel is -0.3918 days, suggesting slightly shorter recovery times in Israel. However, the substantial difference in sample sizes raises questions about the practical significance of this result. The relatively small magnitude of the difference may be due to chance variations and may not substantially impact clinical decisions or patient outcomes.

Additionally, the 95% confidence interval (-0.9478 days to 0.1641 days) highlights the uncertainty in the true population mean difference.

3.

We are testing for the hypothesis $H_0: \mu_x = \mu_y$, where μ_x is the average covid recovery time of patients in New South Wales, and μ_y is the average covid recovery time of patients in Israel.

First is to calculate the test statistic, $Z_{(\mu_x - \mu_y)}$ by the formula,

```
z <- (sample_mean_nsw - sample_mean_israel) / se_difference
```

The value of Z is -1.3814.

Next is to calculate the approximate p-values using $p = 2P(Z < -\left|Z_{(\mu_x - \mu_y)}\right|)$,

```
p <- 2*pnorm(-abs(z))
```

```
z
.] -1.381394
p
.] 0.1671578
```

The value of p is 0.1672, since it is greater than the common significance level of 0.05, there is insufficient evidence to conclude that the population average recovery time for Israeli patients is different from that for NSW patients.

Question 2

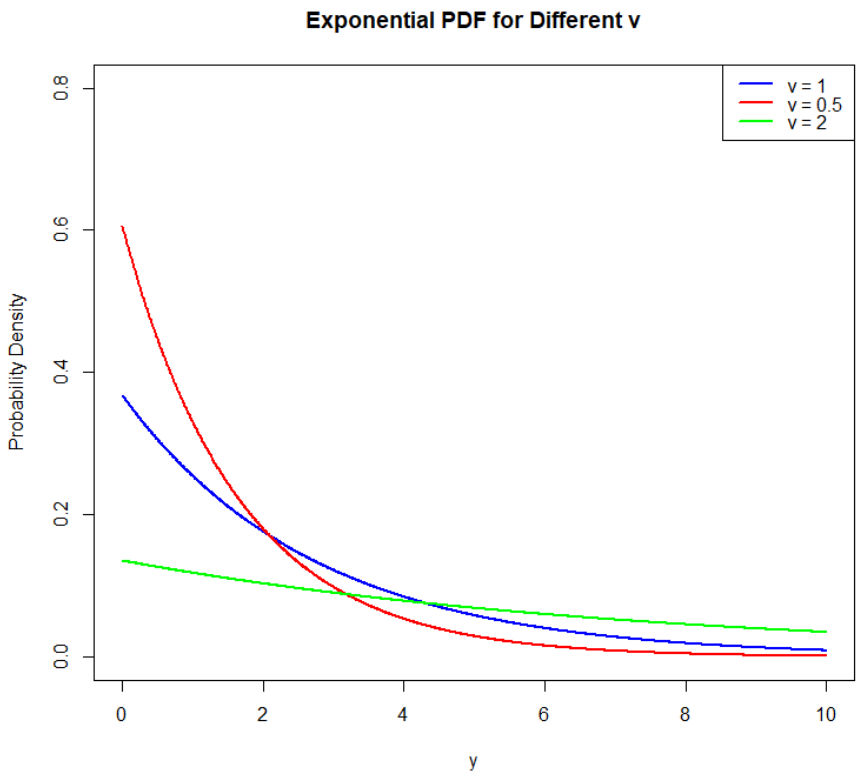
1.

```
# Q2.1
y <- seq(0, 10, by = 0.01)

# Probability density function for different values of v
pdf_v1 <- exp((-exp(-1))*y - 1)
pdf_v05 <- exp((-exp(-0.5))*y - 0.5)
pdf_v2 <- exp((-exp(-2))*y - 2)

# Create a plot
plot(y, pdf_v1, type = "l", col = "blue", lwd = 2, ylim = c(0, 0.8),
     xlab = "y", ylab = "Probability Density", main = "Exponential PDF for Different v")
lines(y, pdf_v05, type = "l", col = "red", lwd = 2)
lines(y, pdf_v2, type = "l", col = "green", lwd = 2)

# Add a legend
legend("topright", legend = c("v = 1", "v = 0.5", "v = 2"), col = c("blue", "red", "green"), lwd = 2)
```



2.

$$L = \Pi_{i=1}^{n} p(y_i|v)$$
$$= \Pi_{i=1}^{n} \exp(-e^{-v} y_i - v)$$
$$= \exp(-e^{-v} \Sigma_{i=1}^{n} y_i - nv)$$

3.

$$-\log(L) = -\log\left(\exp(-e^{-v} \Sigma_{i=1}^{n} y_i - nv)\right)$$

By the property that log(exp(x))=x:

$$-\log(L) = e^{-v} \Sigma_{i=1}^{n} y_i + nv$$
$$-\log(L) = me^{-v} + nv, \qquad m = \Sigma_{i=1}^{n} y_i$$

4.

$$d(-\log(L))/dv = \frac{d}{dv}(me^{-v} + nv)$$
$$= -me^{-v} + n$$

Let $d(-\log(L))/dv = 0$,

$$-me^{-v} + n = 0$$
$$me^{-v} = n$$

$e^{-v} = \frac{n}{m}$

$log(e)^{-v} = \log\left(\frac{n}{m}\right)$

$-vlog(e) = -\log\left(\frac{m}{n}\right)$

$-v = -\log\left(\frac{m}{n}\right)$

$\hat{v} = \log\left(\frac{m}{n}\right)$, $\quad m = \Sigma_{i=1}^{n} y_i$

5.

Given $E[Y] = e^{v}, V[Y] = e^{2v}$

$\hat{v}(Y) = log((\Sigma_{i=1}^{n} y_i)/\text{n})$

$= \log\left(\frac{nY}{n}\right)$

$= \log(Y)$

$\hat{v}'(Y) = \frac{1}{y}$

$\hat{v}''(Y) = -\frac{1}{y^2}$

$E[\,\hat{v}(Y)] = \hat{v}(E[Y]) + \frac{v[y]}{2} * \hat{v}''(E[Y])$

$= \log(e^{v}) + \frac{e^{2v}}{2} * (-\frac{1}{e^{2v}})$

$= v \log e - \frac{1}{2}$

$= v - \frac{1}{2}$

Bias approximation:

$b(\hat{v}) = E[\,\hat{v}(Y)] - v$

$= v - \frac{1}{2} - v$

$= -\frac{1}{2}$

Variance approximation:

$V[\hat{v}(Y)] = \left(\frac{1}{E[Y]}\right)^2 * V[Y]$

$= \left(\frac{1}{e^{v}}\right)^2 * e^{2v}$

$= 1$

Question 3

1.

First is to calculate the mean and standard error of the data:

```
total_pairs <- 124
right_head_pairs <- 80
p_hat <- right_head_pairs/total_pairs
se <- sqrt(p_hat*(1-p_hat)/total_pairs)
```

```
> p_hat
[1] 0.6451613
> se
[1] 0.04296737
```

We get a mean of 0.6452 and a standard error of 0.043.

Next is to calculate the t-value, then use the t-value to calculate the lower bound and upper bound of the 95% confidence interval.

```
student_t <- qt(1-0.05/2, df = total_pairs)
lower_CI <- (p_hat - student_t*se)
upper_CI <- (p_hat + student_t*se)
```

```
> lower_CI
[1] 0.5601168
> upper_CI
[1] 0.7302058
```

We get a confidence interval of (0.5601, 0.7302).

We have estimated that approximately 64.52% of people prefer to turn their heads to the right when kissing. With 95% confidence, we can say that the true population proportion of individuals who tilt their heads to the right while kissing falls within the range of 56.01% to 73.02%. This suggests a strong preference for this direction during kissing among the observed individuals.

2.

We are conducting the hypothesis test $H_0: \theta = \theta_0$, where θ_0 is 0.5. In other words, we are testing whether there is no particular preference among humans for tilting their heads to one side when kissing.

First is to calculate the estimate of the population success probability which is p-hat we calculated just now, z-score and p-value.

```
mu0 <- 0.5
z <- (p_hat - mu0)/sqrt(mu0*(1-mu0)/total_pairs)
p <- 2*pnorm(-abs(z))
```

```
> p
[1] 0.001225424
```

We get a p-value of 0.0012.

With a p-value less than 0.01, we have strong evidence to reject the null hypothesis. This indicates that there is indeed a strong preference among humans for tilting their heads to a specific side when kissing.

3.

By using the binom.test() function in R, we can calculate the exact p-value.

```
res <- binom.test(x=right_head_pairs, n=total_pairs)
res$p.value
```

```
> res$p.value
[1] 0.001564734
```

The exact p-value calculated is 0.0016, which provides strong evidence against the null hypothesis. This indicates a clear preference among humans for tilting their heads to a specific side when kissing.

4.

We are conducting the hypothesis test H_0: $\theta_x = \theta_y$, where θ_x is the population success probability of people turning their heads to the right when kissing and θ_y is the population success probability of people being right-handed. In other words, we are testing whether there is a relation of right-handed people tend to tilt their heads to one side when kissing.

First is to calculate both means,

θ_x = 80/124, θ_y = 83/100

A pooled estimate of θ, θ_p is calculated:

$$\theta_p = (m_x + m_y)/(n_x + n_y) = \frac{80+83}{124+100} = 0.7277$$

```
total_handedness <- 100
right_handed <- 83
p_right_handed <- right_handed / total_handedness
pooled <- (right_head_pairs+right_handed)/(total_pairs+total_handedness)
```

```
> pooled
[1] 0.7276786
```

The test statistic is:

$$z = (\theta_x - \theta_y)/\sqrt{\theta_p(1-\theta_p)\left(\frac{1}{n_x}+\frac{1}{n_y}\right)}$$

```
z <- (p_hat - p_right_handed) / sqrt(
  pooled*(1-pooled)*(1/total_handedness+1/total_pairs))
```

```
> z
[1] -3.089364
```

We get a z-value of -3.0894.

Then calculate the p-value,

```
p <- 2*pnorm(-abs(z))
```

```
> p
[1] 0.002005856
```

We get a p-value of 0.002.

With a p-value less than 0.01, we have strong evidence to reject the null hypothesis. This implies that the preference for head turning to the right or left when kissing is not simply a result of individuals being right-handed or left-handed.