PART 2

Harshath Muruganantham

QUESTION 1

1. An estimate of the average number of days to recovery using the provided data is:

$$\hat{\mu}_{ML} = \frac{1}{n} * \sum_{i=1}^{n} y_i$$

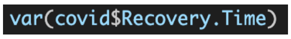
$$\hat{\mu}_{ML} \cong 14.2580$$

Using this mean, we can calculate the 95% confidence interval for this estimate using the t-distribution.

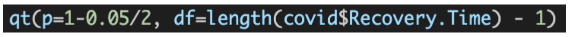
The interval will take the form:

$$(\ \hat{\mu}_{ML} - t_{\frac{\alpha}{2},n-1} \frac{\hat{\sigma}}{\sqrt{n}},\quad \hat{\mu}_{ML} + t_{\frac{\alpha}{2},n-1} \frac{\hat{\sigma}}{\sqrt{n}})$$

We have already found $\hat{\mu}_{ML}$ above. We will now find $t_{\frac{\alpha}{2},n-1}$ and $\hat{\sigma}$ (using r)



$$\hat{\sigma}^2 = \texttt{var(covid\$Recovery.Time)} \approx 44.1532$$



$$t_{\frac{\alpha}{2},n-1} = \texttt{qt(p=1-0.05/2, df=length(covid\$Recovery.Time) - 1)} \approx 1.9610$$

$$n = 2353$$

So,

$$(\ \hat{\mu}_{ML} - t_{\frac{\alpha}{2},n-1} \frac{\hat{\sigma}}{\sqrt{n}}\quad ,\quad \hat{\mu}_{ML} + t_{\frac{\alpha}{2},n-1} \frac{\hat{\sigma}}{\sqrt{n}})$$

$$= \left(14.2580 - 1.9610 * \frac{\sqrt{44.1532}}{\sqrt{2353}}, 14.2580 + 1.9610 * \frac{\sqrt{44.1532}}{\sqrt{2353}} \right)$$

$$\approx (13.9894,\quad 14.5266)$$

The estimated mean number of days to recover from COVID for the people in our sample (sample size $n = 2353$) is 14.2580 days. We are 95% confident that the population mean number of days to recover from COVID for this group is between 13.9894 days and 14.5266 days.

2. The estimated mean recovery times of patients in NSW is: $\hat{\mu}_{NSW} \cong 14.2580$

The estimated mean recovery time of patients in Israel is: $\hat{\mu}_{ISL} \cong 14.6498$

The estimated mean difference in recovery times between the Israeli patients and the patients form NSW is:

$$\hat{\mu}_{ISL} - \hat{\mu}_{NSW} = 14.6498 - 14.2580 \approx 0.3918$$

Using this difference in mean, we can calculate the 95% confidence interval for this estimate.
The interval will take the form:

$$(\hat{\mu}_{ISL} - \hat{\mu}_{NSW} - z_{a/2}\sqrt{\frac{\hat{\sigma}^2{}_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2{}_{ISL}}{n_{ISL}}},\ \hat{\mu}_{ISL} - \hat{\mu}_{NSW} + z_{a/2}\sqrt{\frac{\hat{\sigma}^2{}_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2{}_{ISL}}{n_{ISL}}})$$

We have already found $\hat{\mu}_{ISL}$ and $\hat{\mu}_{NSW}$ above. We will now find $\hat{\sigma}^2{}_{NSW}$ and $\hat{\sigma}^2{}_{ISL}$ (using r):

$\hat{\sigma}^2_{NSW} =$
`var(covid$Recovery.Time)`
≈ 44.1532

$\hat{\sigma}^2_{ISL} =$
`var(israeli_covid$Recovery.Time)`
≈ 30.4655

$$n_{NSW} = 2353$$
$$n_{ISL} = 494$$
$$z_{a/2} = 1.96$$

So,

$$(\hat{\mu}_{ISL} - \hat{\mu}_{NSW} - z_{a/2}\sqrt{\frac{\hat{\sigma}^2{}_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2{}_{ISL}}{n_{ISL}}},\ \hat{\mu}_{ISL} - \hat{\mu}_{NSW} + z_{a/2}\sqrt{\frac{\hat{\sigma}^2{}_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2{}_{ISL}}{n_{ISL}}})$$
$$= \left(0.3918 - 1.96\sqrt{\frac{44.1532}{2353} + \frac{30.4655}{494}},\ 0.3918 + 1.96\sqrt{\frac{44.1532}{2353} + \frac{30.4655}{494}}\right)$$
$$\approx (-0.1641,\ \ 0.9478)$$

The estimated difference in mean days to recover between patients in Israel (samples size $n_{ISL} = 494$) and patients in NSW (samples size $n_{NSW} = 2353$) was 0.3918 days, i.e., the average days to recover was 0.3918 days higher in Isrreal than in NSW. We are 95% confidenr that the population mean difference in days to recover from COVD between these two groups is between -0.1641 days and 0.9478 days. As this interval includes 0, we cannot rule out the possibility of there being no difference at a population level between days to recover from COVID of people in Israel and people in NSW.

3. To formally test the hypothesis that the average time taken to recover for the Israeli cohort is the same as in the NSW cohort, a hypothesis should be set up:

$$H_0 \quad : \quad \hat{\mu}_{ISL} = \hat{\mu}_{NSW}$$
$$\text{vs}$$
$$H_A \quad : \quad \hat{\mu}_{ISL} \neq \hat{\mu}_{NSW}$$

Under the assumption that the population variances in the two groups are unknown (as suggested by the question)

In order to fine the p-value using the approximate hypothesis test, we will first have to find the z-score:

$$z = \frac{\text{diff}}{\text{se}_{\text{diff}}}$$

$$z = \frac{\hat{\mu}_{ISL} - \hat{\mu}_{NSW}}{\sqrt{\frac{\hat{\sigma}^2{}_{ISL}}{n_{ISL}} + \frac{\hat{\sigma}^2{}_{NSW}}{n_{NSW}}}}$$

Using values derived in 1.2 above,

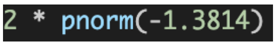
$$z = \frac{14.6498 - 14.2580}{\sqrt{\frac{44.1532}{2353} + \frac{30.4655}{494}}}$$

$$z = \frac{0.3918}{0.2836}$$

$$z \approx 1.3814$$

We can see that the observed difference (diff = 0.3918) is small compared to the standard error for the difference (se$_{\text{diff}}$ = 0.2836). Therefore, we believe this difference will offer weak evidence against the null hypothesis that the population difference is 0. (i.e., the cohort in Israel and the cohort in NSW have the same mean number of days to recover from COVID). Now, to find the p-value:

$$p = 2\,\mathbb{P}(Z < -1.3814) =$$



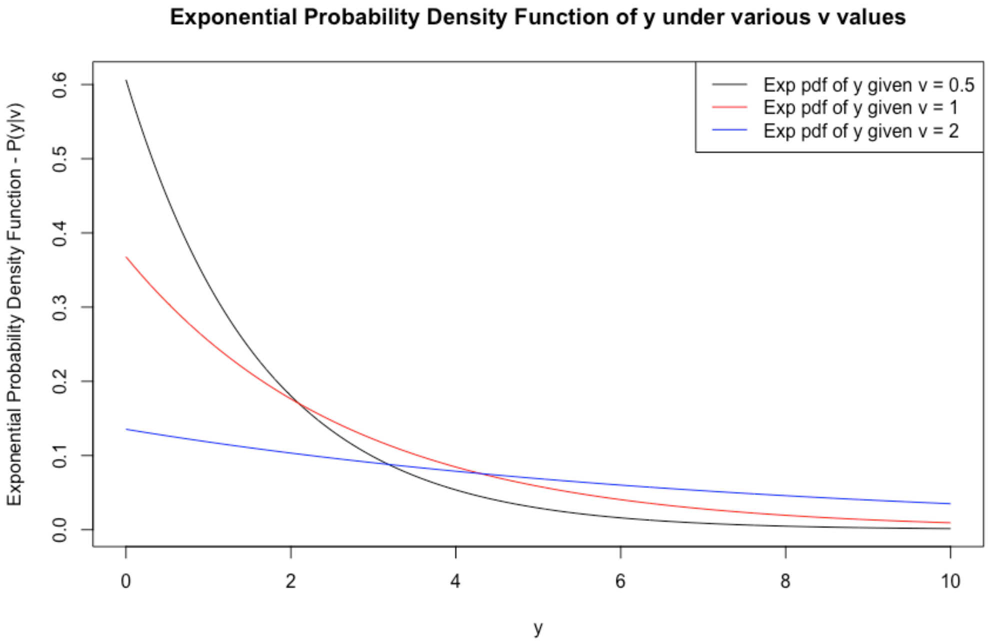
```
2 * pnorm(-1.3814)
```

$$\approx 0.1672$$

This p-value is pretty large, which suggests the observed difference is likely to have arisen just by chance under the null hypothesis that the average number of days to recover from COVID in Israel and in NSW are not equal.

QUESTION 2

1.



2. The joint probability of this sample of data given that these samples are independent and identically distributed is:

$$p(\boldsymbol{y}|v) = (e^{-e^{-v}*y_1-v}) * (e^{-e^{-v}*y_2-v}) * (e^{-e^{-v}*y_3-v}) * \dots * (e^{-e^{-v}*y_n-v})$$
$$= \prod_{i=1}^{n} e^{-e^{-v}*y_i-v}$$
$$= e^{\sum_{i=1}^{n}(-e^{-v}*y_i-v)}$$
$$= e^{-e^{-v}*\sum_{i=1}^{n} y_i-nv}$$
$$= e^{-e^{-v}*m-nv}$$

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

3. The negative loglikelihood of the data y under the exponential model with log-scale v is:

$$L(\boldsymbol{y}|v) = -\log\left(e^{-e^{-v}*m-nv}\right)$$
$$L(\boldsymbol{y}|v) = e^{-v} * m + nv$$

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

4. To find the maximum likelihood estimator $\hat{v}$ for v, we can find the value of v that minimises the negative log-likelihood, that is find $\frac{dL(y|v)}{dv} = 0$:

First let's find $\frac{dL(\boldsymbol{y}|v)}{dv}$:

$$\frac{dL(\boldsymbol{y}|v)}{dv} = -e^{-v} * m + n$$

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

Now solve for v when $\frac{dL(y|v)}{dv} = 0$:

$$\frac{dL(\boldsymbol{y}|v)}{dv} = 0$$
$$-e^{-v} * m + n = 0$$
$$e^{-v} * m = n$$
$$e^{-v} = \frac{n}{m}$$
$$-v = \ln\left(\frac{n}{m}\right)$$
$$v = -\ln\left(\frac{n}{m}\right) = \ln\left(\left(\frac{n}{m}\right)^{-1}\right)$$
$$v = \ln\left(\frac{m}{n}\right)$$

Therefore, the maximum likelihood estimate $\hat{v}$ is:

$$\hat{v}_{MLE} = \ln\left(\frac{m}{n}\right)$$

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

$$\hat{v}_{MLE} = \ln(\bar{y})$$

$\bar{y}$ is the sample mean of our population.

5. To find the approximate bias and variance of the maximum likelihood estimator $\hat{v}$ of v for the exponential distribution, we shall be using these two equations:

$$b(\hat{v}) \approx \left(\frac{\sigma^2}{2n}\right)\left[\frac{d^2f(y)}{dy^2}\,|_{y=\mu}\right]$$

$$Var(\hat{v}) \approx \left(\frac{\sigma^2}{n}\right)\left[\frac{df(y)}{dy}\,|_{y=\mu}\right]^2$$

First, Let's find σ^2 and μ:

We know $\mathbb{V}[\bar{Y}] = \frac{\sigma^2}{n}$ and $\mathbb{V}[Y] = e^{2v}$, so:

$$\sigma^2 = n * \mathbb{V}[\bar{Y}]$$

$$\sigma^2 = n * e^{2v}$$

We also know $\mathbb{E}[\bar{Y}] = \mu$ and $\mathbb{E}[Y] = e^{v}$, so:

$$\mu = e^{v}$$

Now let's substitute these values into the above equations:

$$b(\hat{v}) \approx \left(\frac{n * e^{2v}}{2n}\right)\left[\frac{d^2f(y)}{dy^2}\,|_{y=e^v}\right]$$

$$Var(\hat{v}) \approx \left(\frac{n * e^{2v}}{n}\right)\left[\frac{df(y)}{dy}\,|_{y=e^v}\right]^2$$

Now let's find $\frac{df(x)}{dx}$ and $\frac{d^2f(x)}{dx^2}$.

From 3.4, we know $f(\bar{y}) = \ln(\bar{y})$, so:

$$\frac{df(\bar{y})}{d\bar{y}} = \frac{1}{\bar{y}}$$

$$\frac{df(y)}{dy} = \frac{1}{\frac{\sum_{i=1}^{n} y_i}{n}}$$

and

$$\frac{d^2f(\bar{y})}{dy^2} = -\frac{1}{\bar{y}^2}$$

$$\frac{d^2f(y)}{dy^2} = -\frac{1}{(\frac{\sum_{i=1}^{n} y_i}{n})^2}$$

Now let's use these values to find the approximate bias and variance:

$$b(\hat{v}) \approx \left(\frac{n * e^{2v}}{2n}\right)\left[\frac{d^2f(y)}{dy^2}\,|_{y=e^v}\right]$$

$$b(\hat{v}) \approx \left(\frac{e^{2v}}{2}\right)\left[-\frac{1}{\left(\frac{\sum_{i=1}^{n} y_i}{n}\right)^2}\,|_{y=e^v}\right]$$

$$b(\hat{v}) \approx \left(\frac{e^{2v}}{2}\right)\left[-\frac{1}{\left(\frac{\sum_{i=1}^{n} e_i^v}{n}\right)^2}\right]$$
$$b(\hat{v}) \approx \left(\frac{e^{2v}}{2}\right)\left[-\frac{1}{(\frac{n * e^v}{n})^2}\right]$$
$$b(\hat{v}) \approx \left(\frac{e^{2v}}{2}\right)\left[-\frac{1}{e^{2v}}\right]$$
$$b(\hat{v}) \approx -\frac{1}{2}$$

and,

$$Var(\hat{v}) \approx \left(\frac{n * e^{2v}}{n}\right)\left[\frac{df(y)}{dy}|_{y=e^v}\right]^2$$
$$Var(\hat{v}) \approx (e^{2v})\left[\frac{1}{\frac{\sum_{i=1}^{n} y_i}{n}}|_{y=e^v}\right]^2$$
$$Var(\hat{v}) \approx (e^{2v})\left[\frac{1}{\frac{\sum_{i=1}^{n} e_i^v}{n}}\right]^2$$
$$Var(\hat{v}) \approx (e^{2v})\left[\frac{1}{e^{2v}}\right]$$
$$Var(\hat{v}) = \frac{e^{2v}}{e^{2v}}$$
$$Var(\hat{v}) = 1$$

The approximate bias of the maximum likelihood estimator $\hat{v}$ of v for the exponential distribution was found to be $-\frac{1}{2}$.
The approximate variance of the maximum likelihood estimator $\hat{v}$ of v for the exponential distribution was found to be 1.

QUESTION 3

1. An estimate of the preference for humans turning their heads to the right when kissing is:

$$\hat{\theta} = \frac{80}{124} \approx 0.6452$$

Using this sample mean, we can calculate the 95% confidence interval for this estimate.

The interval will take the form:

$$\left(\hat{\theta} - 1.96\sqrt{\frac{v(\hat{\theta})}{n}},\ \hat{\theta} + 1.96\sqrt{\frac{v(\hat{\theta})}{n}} \right)$$

We have already found $\hat{\theta}$ above. We will now find $v(\hat{\theta})$

$$v(\hat{\theta}) = p(1-p)\,, p = \hat{\theta}$$
$$v(\hat{\theta}) = \hat{\theta}(1-\hat{\theta})$$
$$v(\hat{\theta}) = \frac{80}{124}\left(1 - \frac{80}{124}\right)$$
$$v(\hat{\theta}) \approx 0.2289$$
$$\text{and } n = 124$$

So,

$$(\ \hat{\theta} - 1.96\sqrt{\frac{v(\hat{\theta})}{n}},\ \hat{\theta} + 1.96\sqrt{\frac{v(\hat{\theta})}{n}})$$
$$= (\ 0.6452 - 1.96\sqrt{\frac{0.2289}{124}},\ 0.6452 + 1.96\sqrt{\frac{0.2289}{124}})$$
$$\approx (0.5609,\ \ 0.7294)$$

The estimated preference for humans turning their heads to the right when kissing in our sample (sample size $n = 124$) is 0.6452 percent. We are 95% confident that the preference for humans turning their heads to the right when kissing for this group is between 0.5609 percent and 0.7294 percent.

2. To formally test the hypothesis that there is no preference in humans for tilting their head to one particular side when kissing, a hypothesis should be set up ($\theta_0 = 0.5$):

$$H_0 \quad : \quad \theta = \theta_0$$
$$\text{vs}$$
$$H_A \quad : \quad \theta \neq \theta_0$$

In order to fine the p-value using the approximate hypothesis test, we will first have to find the z-score:

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

Using values derived in 3.1 above and the fact that $\theta_0 = 0.5$ if there is no preference,

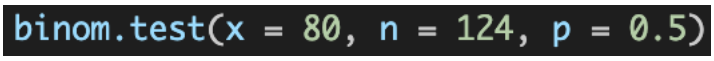
$$z = \frac{0.6452 - 0.5}{\sqrt{0.5(1-0.5)/124}}$$
$$z = \frac{0.1452}{0.0449}$$
$$z \approx 3.2329$$

Now, to find the p-value:

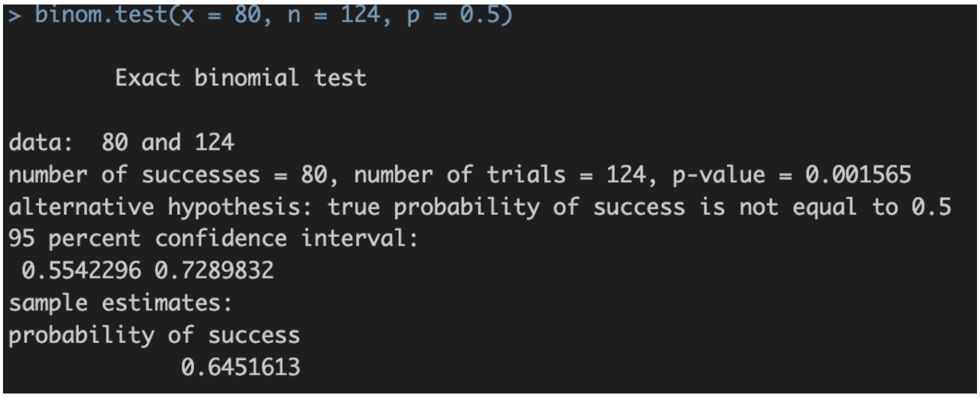
$p = 2\,\mathbb{P}(Z < -3.2329) =$ `2*pnorm(-3.2329)` ≈ 0.0012

This p-value is very small, which suggests the observed difference is likely to not have arisen just by chance under the null hypothesis that there is no preference in humans tilting their head to one particular side when kissing. This is strong evidence against the null hypothesis.

3. We can use the R function binom.test with the parameters $x = 80$ (preference to tilt head right), $n = 124$ (sample size) and $p = 0.5$ (our null hypothesis):



```
binom.test(x = 80, n = 124, p = 0.5)
```



```
> binom.test(x = 80, n = 124, p = 0.5)

	Exact binomial test

data:  80 and 124
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
```

The exact p-value found is 0.001565 which is a bit larger than the approximate derived above, but still gives the same conclusion that there is strong evidence against the null hypothesis. If the sample size was larger, it can be expected the two p-values will be closer as the normal approximation on which the approximate method above is based on would be better.

4. To formally test the hypothesis that the rate of right-handedness in the population (x) is the same as the preference for turning heads to the right when kissing (y) according to the given data, we can set up the following hypothesis:

$$H_0 \quad : \quad \theta_x = \theta_y$$

vs

$$H_A \quad : \quad \theta_x \neq \theta_y$$

In order to fine the p-value using the approximate hypothesis test, we will first have to find the z-score and $\hat{\theta}_p$:

$$\hat{\theta}_p = \frac{m_x + m_y}{n_x + n_y}$$

Where m_x is the number of right-handed people in our sample, m_y is the number of people who turn their heads to the right whilst kissing in our sample and $n_x + n_y$ are our respective size of the two samples.

$$\hat{\theta}_p = \frac{83 + 80}{100 + 124}$$
$$\hat{\theta}_p \approx 0.7277$$

Now, let's find the z-score:

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

$$\hat{\theta}_x = \frac{83}{100} = 0.83, \qquad \hat{\theta}_y = \frac{80}{124} \approx 0.6452$$

$$z_{(\hat{\theta}_x - \hat{\theta}_y)} = \frac{0.83 - 0.6452}{\sqrt{0.7277 * (1 - 0.7277)\left(\frac{1}{100} + \frac{1}{124}\right)}}$$
$$z_{(\hat{\theta}_x - \hat{\theta}_y)} \approx 3.0888$$

Now, to find the p-value:

$p = 2\,\mathbb{P}(Z < -3.0888) =$

```
2 * pnorm(-3.0888)
```

≈ 0.0020

This p-value is very small, which suggests the observed difference is likely to not have arisen just by chance under the null hypothesis that the rate of right-handedness in the population is the same as the preference for turning heads to the right when kissing. This is strong evidence against the null hypothesis.