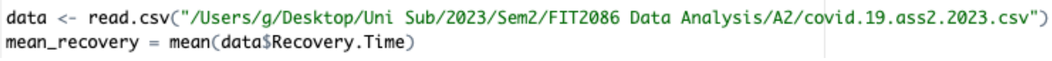
Q1)

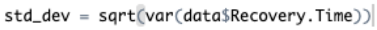
1) Average number of days to recovery for the sample data:

Mean = 14.26 days



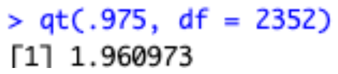
```
data <- read.csv("/Users/g/Desktop/Uni Sub/2023/Sem2/FIT2086 Data Analysis/A2/covid.19.ass2.2023.csv")
mean_recovery = mean(data$Recovery.Time)
```

Sample Standard Deviation = 6.64



```
std_dev = sqrt(var(data$Recovery.Time))
```

The 95% confidence interval of a t distribution for a population estimator falls between 1.96 standard errors on either side of the mean. 1.96 is the commonly known z score for the standard normal distribution having 2.5% of the area on either tail, but we need to test if we have a large enough sample size for the t score to converge to the population z score for the 95% confidence interval marker. This can be shown using qt function to get the t score of one tail having a 2.5% probability of the mean being past that t score value for our data set having 2353 samples, so 2353-1 = 2352 dof.



```
> qt(.975, df = 2352)
[1] 1.960973
```

, thus 95% confidence in population mean =

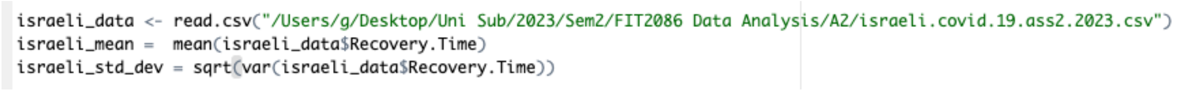
$$\left(\hat{\mu}_{\mathrm{ML}} - t_{\alpha/2,n-1}\frac{\hat{\sigma}_u}{\sqrt{n}},\ \hat{\mu}_{\mathrm{ML}} + t_{\alpha/2,n-1}\frac{\hat{\sigma}_u}{\sqrt{n}}\right)$$

[14.26 – 1.96 * 6.64/√(2353), 14.26 + 1.96 * 6.64/√(2353)]
= [13.99, 14.53]

In this context, this means we can say we are 95% confident that the true population average days of recovery is between 13.99 and 14.53 days. Meaning we are 95% sure the average amount of days for recovery for everyone who gets covid is between 13.99 and 14.53.

2) Difference in means with unknown variances calculate the estimated mean difference in recovery times.

Unknown population means and variances of the NSW and Israeli sample data, we can perform a difference of means calculation, and add a t value to get a 95% confidence interval, and see if 0 is in this interval. If 0 is in this interval, then there is not enough evidence to support that the average days of recovery for someone in NSW differs statistically significantly from the average number of days until recovery for someone in Israel.



```
israeli_data <- read.csv("/Users/g/Desktop/Uni Sub/2023/Sem2/FIT2086 Data Analysis/A2/israeli.covid.19.ass2.2023.csv")
israeli_mean =  mean(israeli_data$Recovery.Time)
israeli_std_dev = sqrt(var(israeli_data$Recovery.Time))
```

Mean Isreali recovery days: 14.65
Standard deviation of Israeli recovery days: 5.52

Where z = 1.96

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

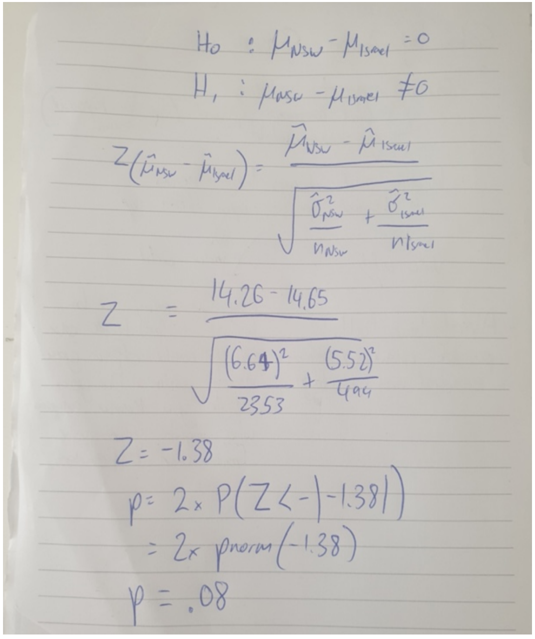
Using sample mean and sample std_dev from NSW in Q1)1) sub in the following values with z = 1.96

95% confidence interval for the difference in means value between NSW and Israel
= [14.26-14.65 – 1.96 * √((6.64^2 / 2353) + (5.52^2 /494)) ,
14.26-14.65 + 1.96 * √((6.64^2 / 2353) + (5.52^2 /494))]

= [-.95, .17]
Since 0 is within this 95% confidence interval, there is not enough evidence to suggest there is a statistically significant different in average recovery days between NSW and Israel. This means that we do not have enough evidence to say that anyone in Israel who gets covid will have a different average recovery time than anyone in NSW.

3) Hypothesis testing population means.

We aim to hypothesis t test the difference of means for the population recovery time for NSW vs Israeli covid patients. Where $ô^2_{NSW}$ and $ô^2_{ISRAEL}$ are the unbiased estimators of population variance and $û_{NSW}$ and $û_{ISRAEL}$ are the unbiased estimates of population mean



$H_0 : \mu_{NSW} - \mu_{Israel} = 0$

$H_1 : \mu_{NSW} - \mu_{Israel} \neq 0$

$$Z(\hat{\mu}_{NSW} - \hat{\mu}_{Israel}) = \frac{\hat{\mu}_{NSW} - \hat{\mu}_{Israel}}{\sqrt{\frac{\hat{\sigma}^2_{NSW}}{n_{NSW}} + \frac{\hat{\sigma}^2_{Israel}}{n_{Israel}}}}$$

$$Z = \frac{14.26 - 14.65}{\sqrt{\frac{(6.64)^2}{2353} + \frac{(5.52)^2}{494}}}$$

$Z = -1.38$

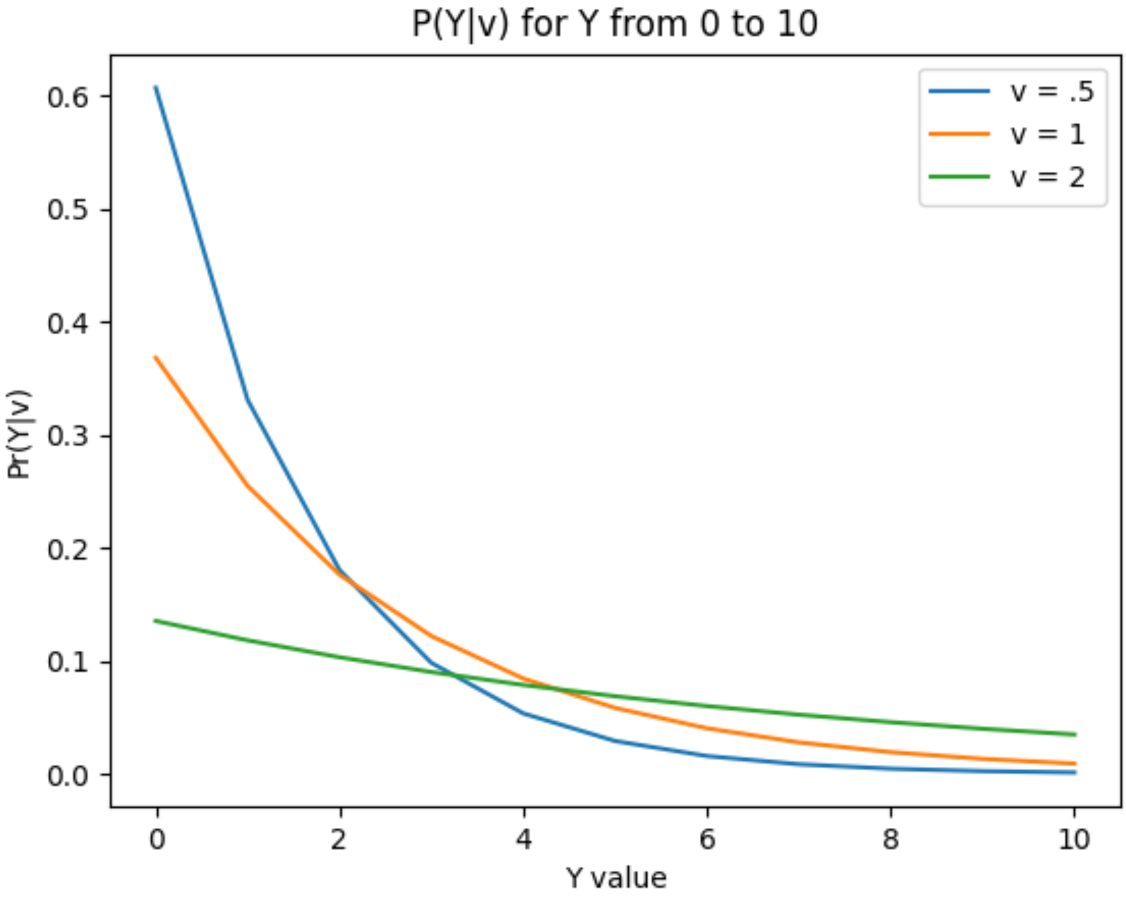
$p = 2 \times P(Z < -|-1.38|)$

$= 2 \times \text{pnorm}(-1.38)$

$p = .08$

Since p > .05, we do not have enough evidence to reject the null hypothesis that there is a difference in population means between the average recovery time for a patient with COVID in NSW compared to Israel.

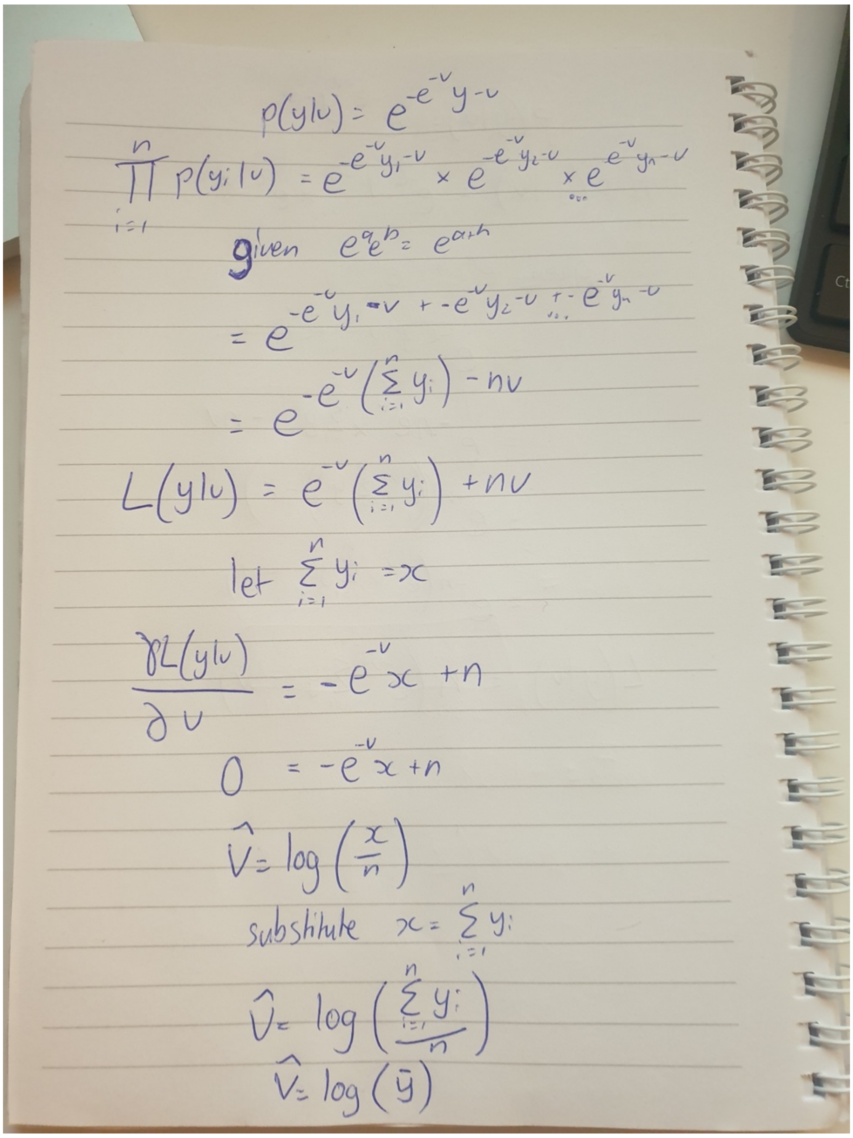
2) 1)



2) Find the likelihood of the p(y|v) pdf

Given the y random variables are i.i.d, The likelihood is equal to the joint probability of the pdf function we are given which is just the product of marginal probabilities. Then simplify before taking the negative log. Solve for the differential of the negative log = 0.

Question 2 done in pictures below



$$p(y|v) = e^{-e^{-v}y-v}$$

$$\prod_{i=1}^{n} p(y_i|v) = e^{-e^{-v}y_1-v} \times e^{-e^{-v}y_2-v} \times e^{e^{-v}y_n-v}$$

given $e^a e^b = e^{a+b}$

$$= e^{-e^{-v}y_1-v + -e^{-v}y_2-v + \dots -e^{-v}y_n-v}$$

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

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

let $\sum_{i=1}^{n} y_i = x$

$$\frac{\partial L(y|v)}{\partial v} = -e^{-v}x + n$$

$$0 = -e^{-v}x + n$$

$$\hat{v} = \log\left(\frac{x}{n}\right)$$

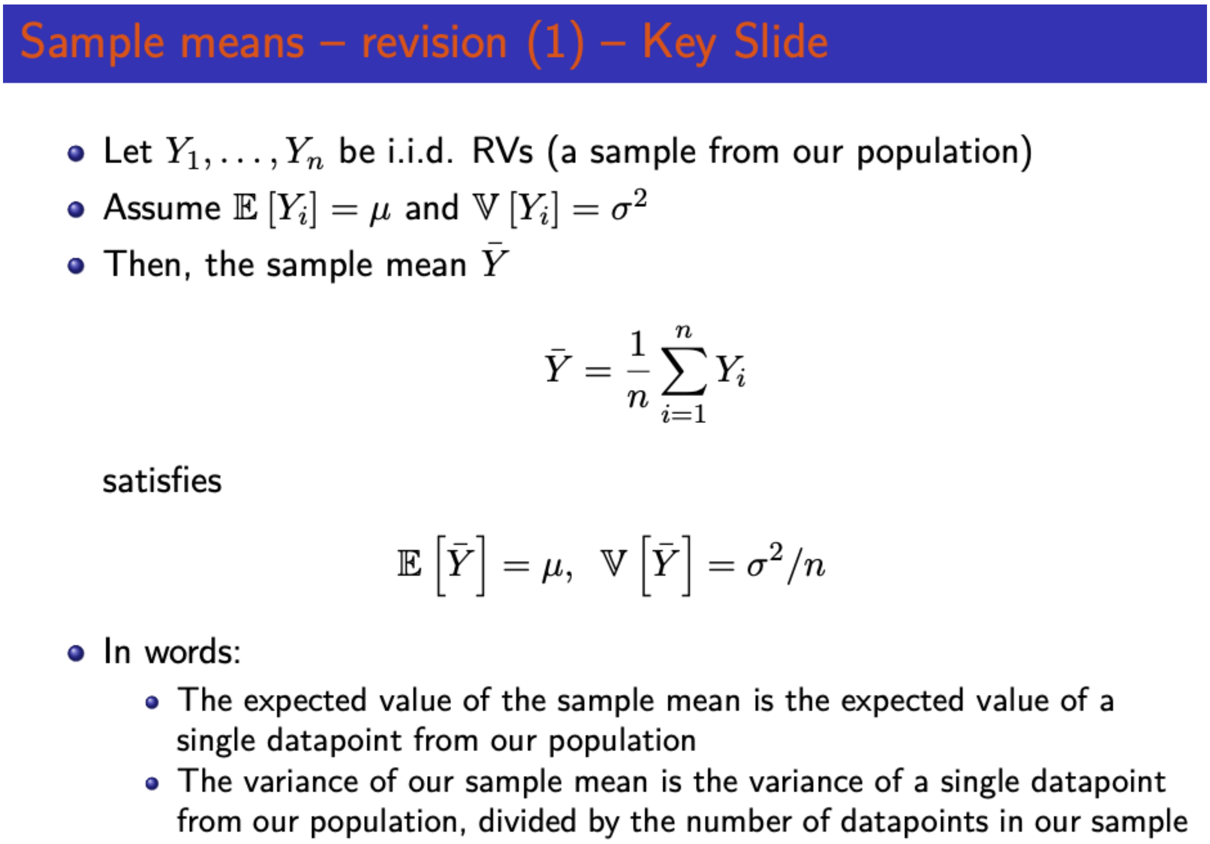
substitute $x = \sum_{i=1}^{n} y_i$

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

$$\hat{v} = \log(\bar{y})$$

- If $\hat{\theta}(Y)$ is an estimator of a parameter θ its bias is

$$b_\theta(\hat{\theta}) = \mathbb{E}\left[\hat{\theta}(Y)\right] - \theta$$



Sample means – revision (1) – Key Slide

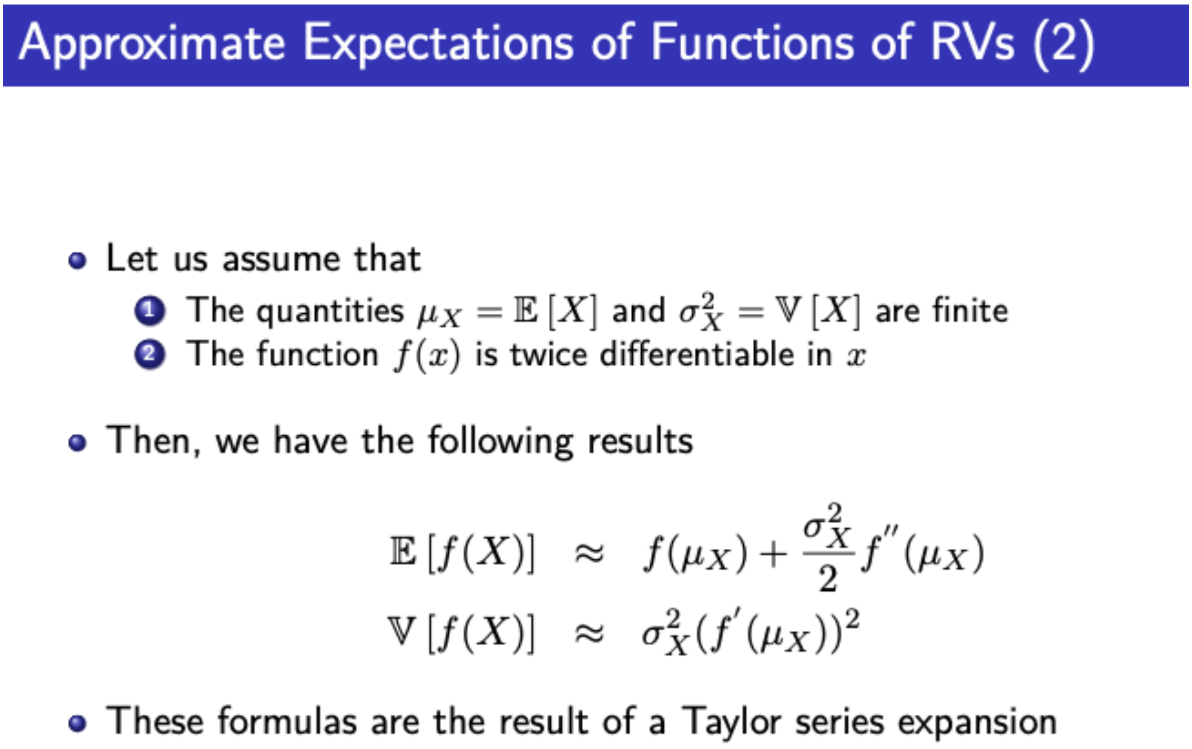
- Let $Y_1, \dots, Y_n$ be i.i.d. RVs (a sample from our population)
- Assume $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2$
- Then, the sample mean $\bar{Y}$

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

  satisfies

$$\mathbb{E}\left[\bar{Y}\right] = \mu, \quad \mathbb{V}\left[\bar{Y}\right] = \sigma^2/n$$

- In words:
  - The expected value of the sample mean is the expected value of a single datapoint from our population
  - The variance of our sample mean is the variance of a single datapoint from our population, divided by the number of datapoints in our sample



Approximate Expectations of Functions of RVs (2)

- Let us assume that
  1. The quantities $\mu_X = \mathbb{E}[X]$ and $\sigma_X^2 = \mathbb{V}[X]$ are finite
  2. The function $f(x)$ is twice differentiable in x
- Then, we have the following results

$$\begin{aligned}\mathbb{E}[f(X)] &\approx f(\mu_X) + \frac{\sigma_X^2}{2} f''(\mu_X) \\ \mathbb{V}[f(X)] &\approx \sigma_X^2 (f'(\mu_X))^2\end{aligned}$$

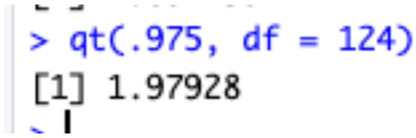
- These formulas are the result of a Taylor series expansion

3. 1. The distribution is Bernoulli, as they either tilt to the right, which we will call success or left to fail.

The expected value of a binomial distribution is just the probability of success.

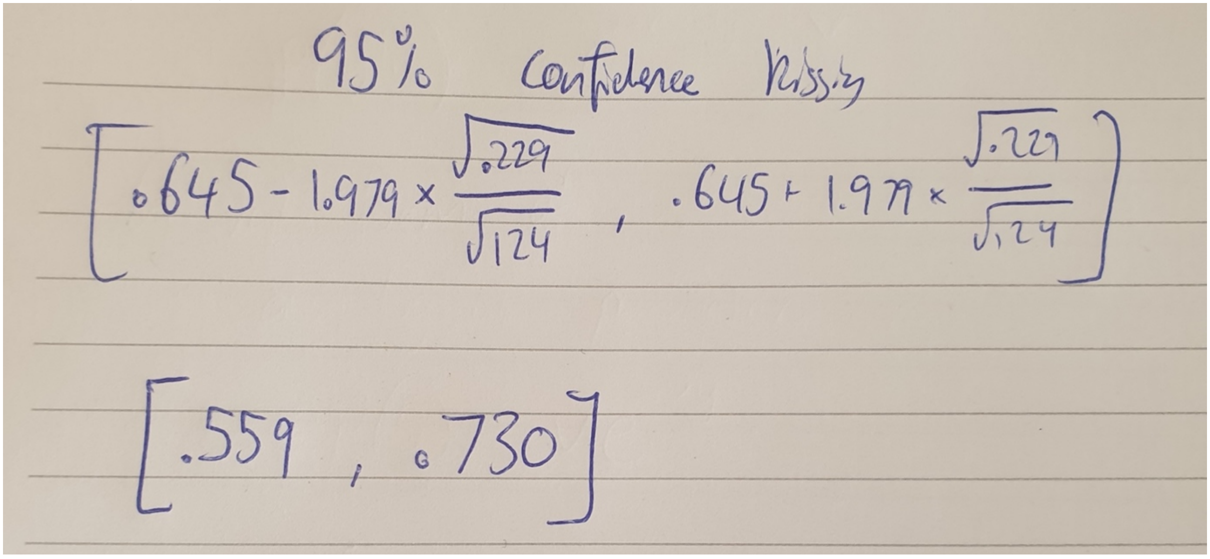
We are told there are 80/124 results are successes so the preference is our sample average that p = 64.5% of the time, people turn their head to the right.

The 95% confidence interval is.1.979 standard errors in either side of the sample mean



```
> qt(.975, df = 124)
[1] 1.97928
```

The variance in a Bernoulli distribution = p(1-p)
V = .645*(1-.645) = .229



95% Confidence Kissing

$$\left[.645 - 1.979 \times \frac{\sqrt{.229}}{\sqrt{124}},\ .645 + 1.979 \times \frac{\sqrt{.229}}{\sqrt{124}}\right]$$

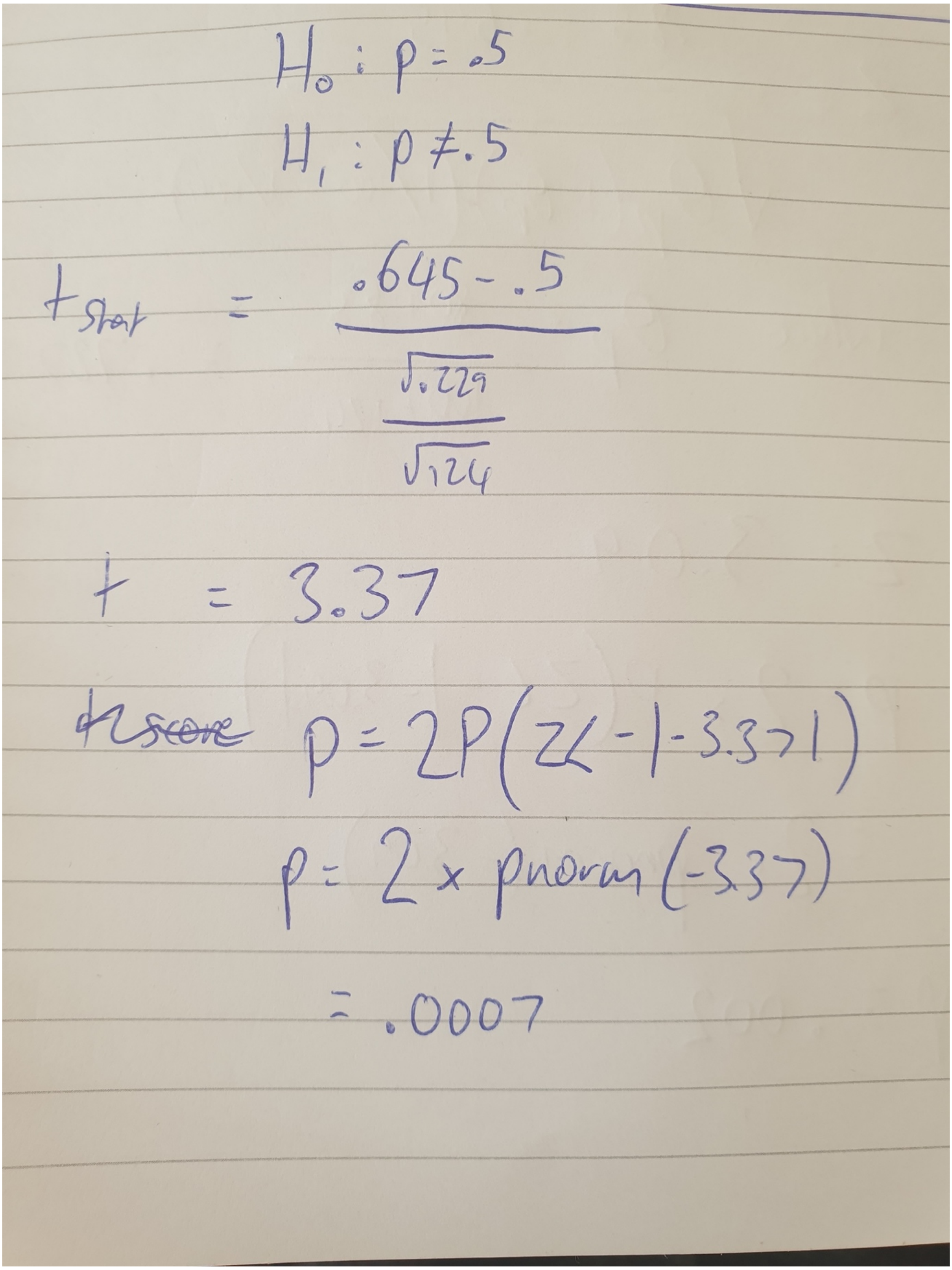
$$[.559\ ,\ .730]$$

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

We are 95% confident that the population average for the amount of people with preference of turning their head to the right when making out lies within 55.9% and 73.0%

3.2 .Test the hypothesis that there is no preference in humans for tilting their head to one particular side when kissing.

If there was no preference than p = .5 since you would be indifferent about which side to turn, this conduct a hypothesis test on the



$H_0 : p = .5$

$H_1 : p \neq .5$

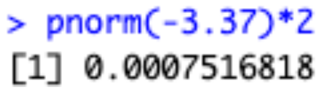
$$t_{stat} = \frac{.645 - .5}{\frac{\sqrt{.229}}{\sqrt{124}}}$$

$t = 3.37$

~~t score~~ $p = 2P(Z < -|-3.37|)$

$p = 2 \times \text{pnorm}(-3.37)$

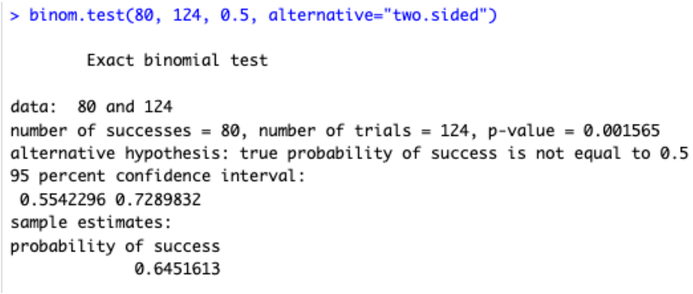
$= .0007$



```
> pnorm(-3.37)*2
[1] 0.0007516818
```

This p value < .05 therefore there is enough evidence to reject the null hypothesis that there is no preference between turning your head to a particular side when kissing in favour of the alternative that there is a preference in humans to turn their head to one side when kissing.

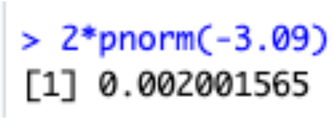
3.3



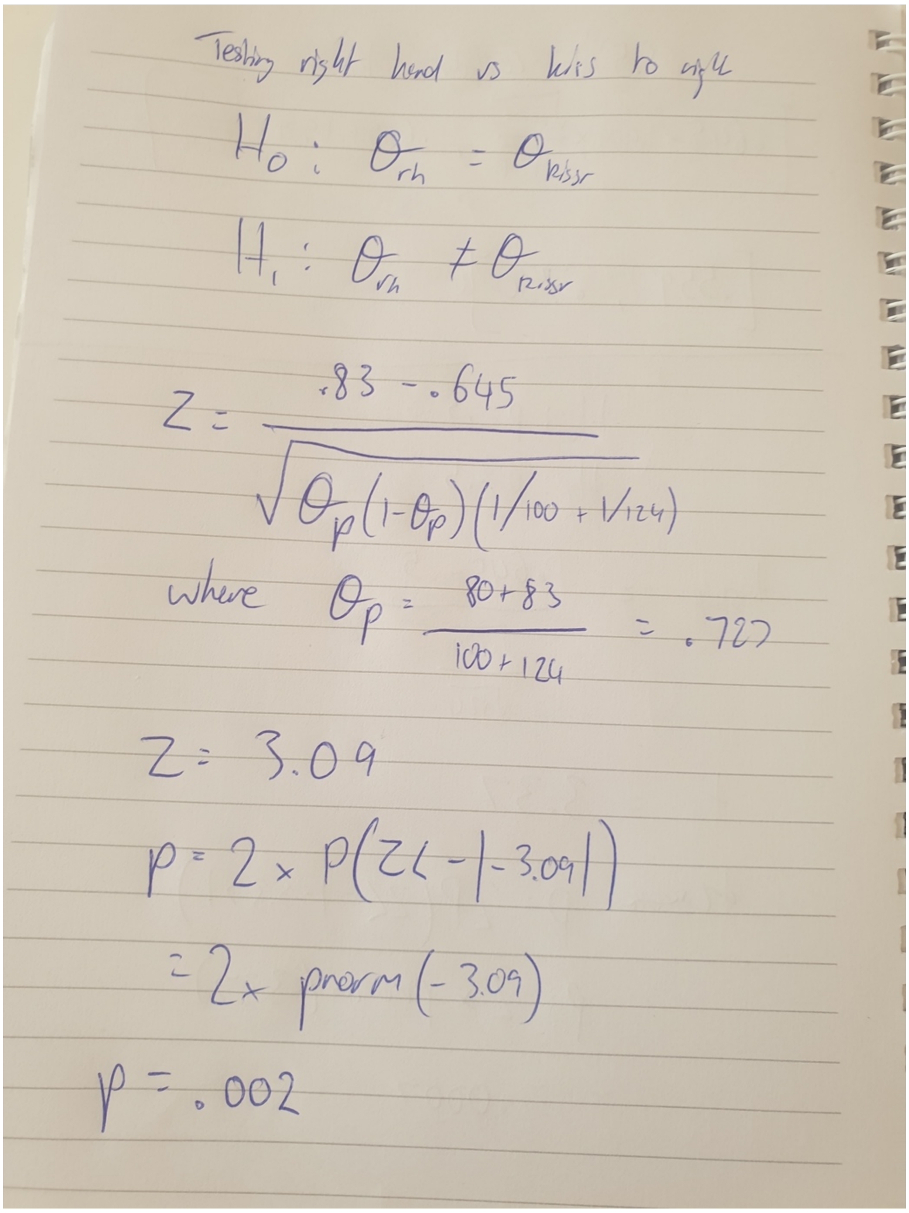
```
> binom.test(80, 124, 0.5, alternative="two.sided")

        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
```

3.4

Classify right handed as success and left as failure. P = successes/total = 83/100 = 83



```
> 2*pnorm(-3.09)
[1] 0.002001565
```



Testing right hand vs kiss to right

$H_0 : \theta_{rh} = \theta_{kissr}$

$H_1 : \theta_{rh} \neq \theta_{kissr}$

$$Z = \frac{.83 - .645}{\sqrt{\theta_p(1-\theta_p)(1/100 + 1/124)}}$$

where $\theta_p = \frac{80+83}{100+124} = .727$

$Z = 3.09$

$p = 2 \times P(Z < -|-3.09|)$

$= 2 \times \text{pnorm}(-3.09)$

$p = .002$

The p value = .002 > .05 suggests there is enough evidence to reject the null hypothesis that the rate of kissing to the right is due to the rate of right handedness in favour of the alternative that kissing to the right preference is not caused by right handedness.