FIT2086 Assignment 2

Lim Zheng Haur

32023952

Question 1

1. Calculate an estimate of the average number of daily reported cases using the provided data. Calculate a 95% confidence interval for this estimate using the t-distribution, and summarise/describe your results appropriately. Show working as required. **[4 marks]**

```
# Question 1.1
# sample 1
daily_covid <- read.csv("daily.covid.aug1to7.csv")

# estimate average = sample mean
avg_daily <- mean(daily_covid$daily.covid.cases)
# avg_daily = 7359.571

# sample size
n <- length(daily_covid$daily.covid.cases)
# n = 7

# sample variance
var_daily <- var(daily_covid$daily.covid.cases)
# var_daily = 4108400

# standard error
se <- sqrt(var_daily)/sqrt(n)
# se = 766.1033

# t-score (a = 0.05, dof = n-1)
t <- qt(p = 1-0.05/2, df = n-1)
# t = 2.446912

# 95% CI for this estimate using the t-distribution
CI <- c(avg_daily - t*se, avg_daily + t*se)
# CI = [5484.984, 9234.159]
```

The estimate of the average number of daily reported cases in our sample which is from August 1 to 7 is 7359.571. We are 95% confident that the population mean daily number of reported cases is between 5484.984 and 9234.159

2. The file daily.covid.aug8to14.csv contains data on daily reported case numbers for the second 7-day period in August. Using the provided data and the approximate method for difference in means with (different) unknown variances presented in Lecture 4, calculate the estimated mean difference in reported daily Covid-19 cases between the first 7 day block of August and the second 7 day block in August, and provide an approximate 95% confidence interval. Summarise/describe your results appropriately. Show working as required. **[3 marks]**

```
# Question 1.2
# sample 2
daily_covid_2 <- read.csv("daily.covid.aug8to14.csv")

# estimate average = sample mean
avg_daily_2 <- mean(daily_covid_2$daily.covid.cases)
# avg_daily_2 = 4879

# mean difference
mean_diff <- avg_daily - avg_daily_2
# mean_diff = 2480.571

# sample variance
var_daily_2 <- var(daily_covid_2$daily.covid.cases)
# var_daily_2 = 1286109

# sample size
n_2 <- length(daily_covid_2$daily.covid.cases)
# n_2 = 7

# standard error
se_diff <- sqrt(var_daily/n + var_daily_2/n_2)
# se_diff = 877.8634

# 95% CI of difference of means
CI <- c(mean_diff - 1.96*se_diff, mean_diff + 1.96*se_diff)
# CI = [759.9591, 4201.1837]
```

The estimated difference in mean number of daily reported cases between the August 1 to 7 and August 8 to 14 was 2480.571, i.e. the average number of daily reported cases was 2480 higher on the first week of August than the second. We are 95% confident that the population mean difference in average number of daily cases between these two groups is between 759.9904 to 4201.1521. As both ends of the confidence interval is positive, and far away from zero, the data suggest that the number of daily reported cases are on the decline.

3. It is potentially of interest to see if the daily reported case numbers are changing over time. Test the hypothesis that the population average daily reported case numbers between the two seven-day blocks is the same. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate hypothesis test for differences in means with (different) unknown variances presented in Lecture 5. What does this p-value suggest about the difference in average reported daily case numbers between the two seven-day blocks? **[3 marks]**

Sample 1 (S1) = August 1 ~ 7
Sample 2 (S2) = August 8 ~ 14

$\boldsymbol{H_0: \mu_{S1} = \mu_{S2}}$
$\boldsymbol{H_A: \mu_{S1} \neq \mu_{S2}}$

```
# Question 1.3
# z-score
# mean_diff = 2480.571 and se_diff = 877.8634
# values adapted from Question 1.2
z <- mean_diff/se_diff
# z = 2.825692

# p-value
p <- 2 * pnorm(-abs(z))
# p = 0.004717864
```

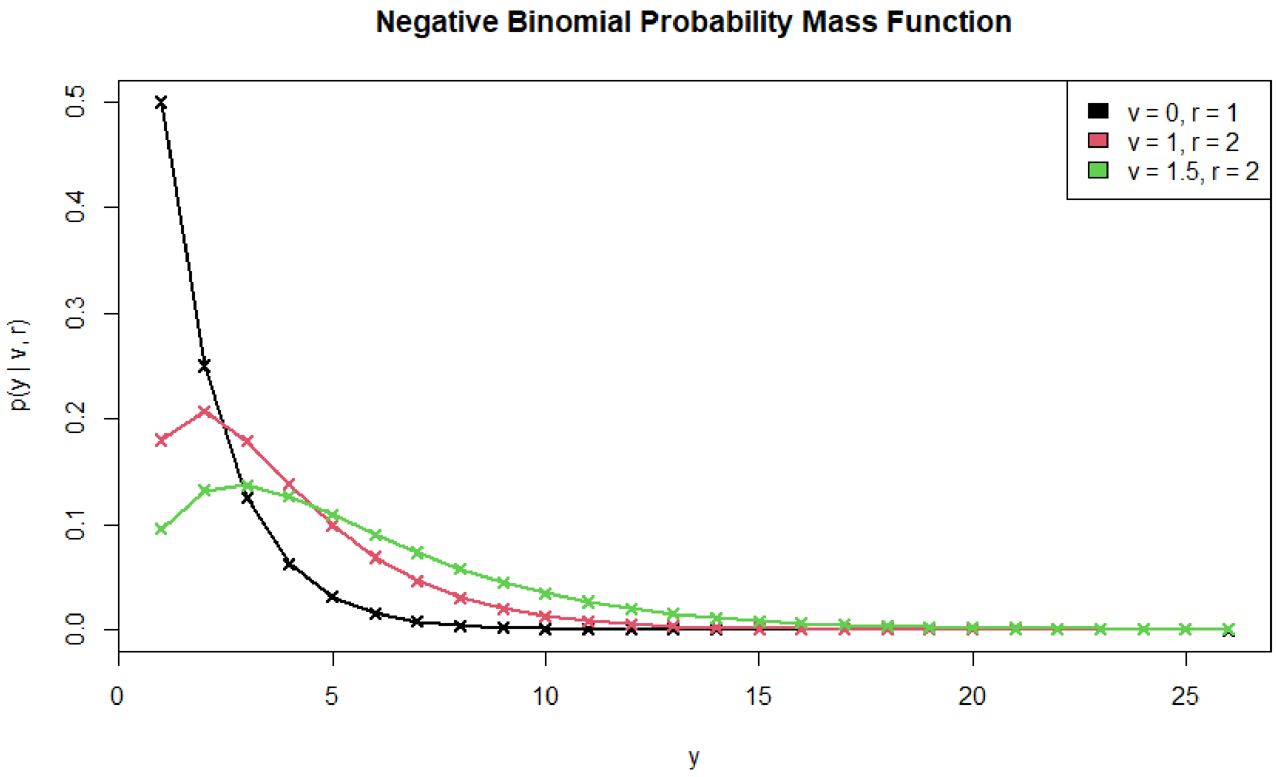
The p-value, 0.004717864, is incredibly small which suggests that we have strong evidence against the null hypothesis which states that the average daily reported case numbers between the two seven-day block is the same. Hence, we can conclude that the average daily reported case numbers is different for every seven-day block.

Question 2

1. Produce a plot of the negative binomial probability mass function (1) for the values y ∈ {0, 1, . . . , 25}, for (v = 0, r = 1), (v = 1, r = 2) and (v = 1.5, r = 2). Ensure that the graph is readable, the axis are labelled appropriately and a legend is included (hint: the choose() function in R may be useful). **[2 marks]**

```
# Question 2.1
# negative binomial probability mass function
prob_mass_func = function(y, v, r) {choose(y+r-1, y) * r^r *
(exp(v)+r)^(-r-y) * exp(y*v)}

# plot graph
plot(prob_mass_func((0:25), 0, 1), type = 'o', lwd = 2, col =
1, pch = 4, main = "Negative Binomial Probability Mass
Function", xlab = "y", ylab = "p(y | v, r)")
lines(prob_mass_func((0:25), 1, 2), type = 'o', lwd = 2, col =
2, pch = 4)
lines(prob_mass_func((0:25), 1.5, 2), type = 'o', lwd = 2, col
= 3, pch = 4)
legend(x="topright",legend = c("v = 0, r = 1","v = 1, r =
2","v = 1.5, r = 2"), fill=c(1, 2, 3))
```



2. Imagine we are given a sample of n observations y = (y1, . . . , yn). Write down the joint probability of this sample of data, under the assumption that it came from a negative binomial distribution with parameters v and r (i.e., write down the likelihood of this data). Make sure to simplify your expression, and provide working. (*hint: remember that these samples are independent and identically distributed*.) **[2 marks]**

$$p(y|v,r) = \left(\binom{y_1 + r - 1}{y_1} r^r (e^v + r)^{-r-y_1} e^{y_1 v}\right) \cdot \left(\binom{y_2 + r - 1}{y_2} r^r (e^v + r)^{-r-y_2} e^{y_2 v}\right)$$

$$\cdots \left(\binom{y_n + r - 1}{y_n} r^r (e^v + r)^{-r-y_n} e^{y_n v}\right)$$

$$p(y|v,r) = \left(\prod_{i=1}^{n} \binom{y_i + r - 1}{y_i}\right) r^{nr} (e^v + r)^{-nr-m} e^{mv}$$

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

3. Take the negative logarithm of your likelihood expression and write down the negative loglikelihood of the data y under the negative binomial model with parameters v and r. Simplify this expression**. [1 mark]**

$$L(y|v,r) = -\log p\,(y|v,r)$$

$$L(y|v,r) = -\log\left(\prod_{i=1}^{n}\binom{y_i + r - 1}{y_i}\right) - nr\log r - (-nr - m)\log(e^v + r) - mv\log e$$

$$L(y|v,r) = -\log\left(\prod_{i=1}^{n}\binom{y_i + r - 1}{y_i}\right) - nr\log r - (-nr - m)\log(e^v + r) - mv$$

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

4. Derive the maximum likelihood estimator vˆ for v, under the assumption that r is fixed; that is, find the value of v that minimises the negative log-likelihood, treating r as a fixed quantity. You must provide working. **[2 marks]**

Differentiate the negative log-likelihood with respect to our parameter of interest v:

$$\frac{dL(y|v,r)}{dv} = 0 - 0 - (-nr - m)\frac{d}{dv}\log\left(e^v + r\right) - \frac{d}{dv}mv$$

$$\frac{dL(y|v,r)}{dv} = 0 - 0 - (-nr - m)\frac{e^v}{e^v + r} - m$$

$$\frac{dL(y|v,r)}{dv} = \frac{(nr + m)e^v}{e^v + r} - m$$

Set this to zero and solve for v:

$$\frac{(nr+m)e^v}{e^v+r} - m = 0$$

$$(nr + m)e^v - m(e^v + r) = 0$$

$$nre^v + me^v - me^v - mr = 0$$

$$e^v = \frac{mr}{nr}$$

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

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

5. Determine expressions for the approximate bias and variance of the maximum likelihood estimator vˆ of v for the negative binomial distribution, under the assumption that r is fixed. (hints: utilise techniques from Lecture 2, Slide 22 and the mean/variance of the sample mean) **[3 marks]**

$$E[Y] = e^v$$

$$V[Y] = \frac{e^v(e^v + r)}{r}$$

$$\hat{V}_{ML}(y) = \log\left(\frac{m}{n}\right) \ where \ m = \sum_{i=1}^{n} y_i$$

$$\hat{V}_{ML}(Y) = \log\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) = \log \bar{Y}$$

$$\frac{d\hat{V}_{ML}}{dY} = \frac{1}{\bar{Y}}$$

$$\frac{d^2\hat{V}_{ML}}{dY^2} = -\frac{1}{\bar{Y}^2}$$

$$E\left[\hat{V}_{ML}\right] = f(E[Y]) + \left[\frac{d^2\hat{V}_{ML}(E[Y])}{dY^2}\right]\frac{V[Y]}{2}$$

$$E\left[\hat{V}_{ML}\right] = f(e^v) + \left[\frac{d^2\hat{V}_{ML}(e^v)}{dY^2}\right]\frac{e^v(e^v + r)}{2r}$$

$$E\left[\hat{V}_{ML}\right] = \log e^v + \left[-\frac{1}{e^{2v}}\right]\frac{e^v(e^v + r)}{2r}$$

$$E\left[\hat{V}_{ML}\right] = v - \frac{e^v + r}{2re^v}$$

$$b_v(\hat{V}) = E\left[\hat{V}_{ML}(Y)\right] - v$$

$$b_v(\hat{V}) = -\frac{e^v + r}{2re^v}$$

$$V\left[\hat{V}_{ML}\right] = \left[\frac{d\hat{V}_{ML}(\bar{Y})}{dY}\right]^2 V[Y]$$

$$V\left[\hat{V}_{ML}\right] = \left(\frac{1}{e^v}\right)^2\left(\frac{e^v(e^v + r)}{r}\right)$$

$$V\left[\hat{V}_{ML}\right] = \frac{e^v(e^v + r)}{e^{2v}r}$$

$$Var_v\left(\hat{V}_{ML}\right) = \frac{V\left[\hat{V}_{ML}(Y)\right]}{n}$$

$$Var_v\left(\hat{V}_{ML}\right) = \frac{e^v(e^v+r)}{e^{2v}nr}$$

Question 3

1. Calculate an estimate of the preference for humans turning their heads to the right when kissing using the above data, and provide an approximate 95% confidence interval for this estimate. Summarise/describe your results appropriately. **[3 marks]**

```
# Question 3.1
# sample size
n <- 240

# observed turn right
m <- 176

# maximum likelihood estimator
theta_hat  <- m/n
# theta = 0.7333333

# 95% CI for the probability parameter theta
CI <- c(theta_hat - 1.96*sqrt((theta_hat*(1- theta_hat))/n),
theta_hat + 1.96*sqrt((theta_hat*(1-theta_hat))/n))
# CI = [0.6773852, 0.7892815]
```

In our sample of n = 240 preference for humans turning their heads, the observed probability of turning their heads to the right is 176/240. We are 95% confident that the true population probability of a head lies between 0.6773852 (biased to the right) and 0.7892815 (heavily biased to the right). The interval rules out the possibility of the population probability of success being 1/2 (i.e. unbiased turning left or right).

2. Test the hypothesis that there is no preference in humans for tilting their head to one particular side when kissing. Write down explicitly the hypothesis you are testing, and then calculate a p-value using the approximate approach for testing a Bernoulli population discussed in Lecture 5. What does this p-value suggest? **[2 marks]**

$\boldsymbol{H_0}: \widehat{\boldsymbol{\theta}}_{\boldsymbol{ML}} = \boldsymbol{\theta_0}$
$\boldsymbol{H_A}: \widehat{\boldsymbol{\theta}}_{\boldsymbol{ML}} \neq \boldsymbol{\theta_0}$

```
# Question 3.2
# sample size
n <- 240

# observed turn right
m <- 176

# maximum likelihood estimator
theta_hat  <- m/n
# theta = 0.7333333

# theta null (no preference)
theta_0 <- 1/2

# z-score
z <- (theta_hat-theta_0) / sqrt(theta_0*(1-theta_0)/n)
# z = 7.229569

# p-value
p <- 2 * pnorm(-abs(z))
# p = 4.845296e-13
```

The p-value, 4.845296e-13, is incredibly small which suggests that we have strong evidence against the null hypothesis which states that there is no preference in humans for tilting their head to one particular side when kissing. Hence, we can conclude that humans have a preference for tilting their head to one particular side when kissing.

3. Using R, calculate an exact p-value to test the above hypothesis. What does this p-value suggest? Please provide the appropriate R command that you used to calculate your p-value. **[1 mark]**

```
# Question 3.3
# sample size
n <- 240

# observed turn right
m <- 176
binom.test(x = m, n = n, p = 1/2)
# p-value of 2.854e-13

     Exact binomial test

data:  m and n
number of successes = 176, number of trials = 240, p-value =
2.854e-13
alternative hypothesis: true probability of success is not
equal to 0.5
95 percent confidence interval:
 0.6726355 0.7881673
sample estimates:
probability of success
             0.7333333
```

The exact p-value is 2.854e-13, which is a bit smaller than our approximate procedure, but gives the same overall conclusion. The two p-values might be more similar if the sample size n is larger as normal approximation method would provide a better result with a larger n.

4. It is entirely possible that any preference for head turning to the right/left could be simply a product of right/left-handedness. To test this we the handedness of the 240 volunteers was also recorded. It was found that 210 of the participants were right-handed and 30 were left handed. Using the approximate hypothesis testing procedure for testing two Bernoulli populations from Lecture 5, test the hypothesis that the rate of right-handedness in the population from which the participants was drawn is the same as the preference for turning heads to the right when kissing. Summarise your findings. What does the p-value suggest? **[2 marks]**

$\boldsymbol{H_0: \theta_x = \theta_y}$
$\boldsymbol{H_A: \theta_x \neq \theta_y}$

θ_x is the population probability of turning to the right when kissing and θ_y is the population probability of right-handed.

```
# Question 3.4
# observed turn right when kissing
mx <- 176

# observed right handed
my <- 210

# sample size
n <- 240

# pooled estimate of the success probability
theta_hat_p <- (mx + my)/(n + n)
# theta_hat_p = 0.8041667

#estimate of population probability of success
theta_hat_x <- mx/n
theta_hat_y <- my/n
# theta_hat_x = 0.7333333
# theta_hat_y = 0.875

# z-score
z <- (theta_hat_x - theta_hat_y)/sqrt(thetha_hat_p * (1-
theta_hat_p) * (1/n + 1/n))
# z = 7.229569

# p-value
p <- 2*pnorm(-abs(z))
# p = 4.845296e-13
```

The p-value, 4.845296e-13, is incredibly small which suggests that we have strong evidence against the null hypothesis which states that rate of right handedness is the same as the preference for turning the heads to the right when kissing. Hence, we can conclude that right handedness do not infer the participants preference for tilting their head to the right when kissing.