FIT2086 Assignment2

Name: Lianzheng Xie
Student ID: 32068611

Question1 :

1. First, the mean function was used to calculate the average number of daily reported cases, and then we needed to obtain the variance and sample size.

```
data1 <- read.csv("daily.covid.aug1to7.csv", header = TRUE)
mean(data1$daily.covid.cases)
[1] 7359.571
var(data1$daily.covid.cases)
[1] 4108400
length(data1$daily.covid.cases)
[1] 7
qt(1-0.05/2,7-1)
[1] 2.446912
```

$\hat{\mu}_1 =$ 7359.571
$\hat{\sigma}_1^{\,2}$ =4108400
n = 7

$t_{\frac{\alpha}{2},n-1}$= qt(1-0.05/2,7-1) = 2.446912

According the interval:

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

$95\%\text{CI}=(7359.571 - 2.446912\sqrt{\frac{4108400}{7}}, 7359.571 + 2.446912\sqrt{\frac{4108400}{7}})$

=(5484.984, 9234.158)

The estimated average of the first 7 day block is 7359.571. We are 95% confident the population mean for this group is between 5484.984 and 9234.158.

2. The mean function was used to calculate the average number of daily reported cases, and then we needed to obtain the variance and sample size.

```
data2 <- read.csv("daily.covid.aug8to14.csv", header = TRUE)
mean(data2$daily.covid.cases)
[1] 4879
var(data2$daily.covid.cases)
[1] 1286109
length(data2$daily.covid.cases)
[1] 7
```

$\hat{\mu}_2 = 4879,\ \hat{\mu}_1 = 7359.571$
$\hat{\sigma}_1{}^2 = 4108400,\ \hat{\sigma}_2{}^2 = 1286109$
$\hat{\mu}_1 - \hat{\mu}_2 = 7359.571\text{-}4879 = 2480.571$
$n_a = n_b = 7$

According the interval:

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

$95\%\text{CI}=(2480.571 - 1.96\sqrt{\frac{4108400}{7} + \frac{1286109}{7}}, 2480.571 + 1.96\sqrt{\frac{4108400}{7} + \frac{1286109}{7}})$
=(759.9587, 4201.183)
"The estimated difference in mean between the first 7 day block and the second 7 day block is 2480.571. We are 95% confident the population mean difference is between 759.9587and 4201.183"

3. It is assumed that the average number of daily reported cases of population is the same between the two seven-day districts

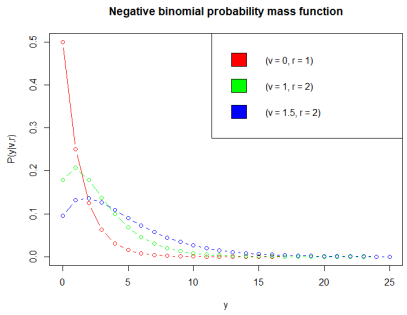
$$H_0: \mu_1 = \mu_2$$
$$vs$$
$$H_1: \mu_1 \neq \mu_2$$

The sample means are $\hat{\mu}_1 = 7359.571$ and $\hat{\mu}_2 = 4879$, and the estimates of variance are $\hat{\sigma}_1{}^2 = 4108400,\ \hat{\sigma}_2{}^2 = 1286109$

$z_{\mu_1-\mu_2} = \frac{7359.571-4879}{\sqrt{\frac{4108400}{7}+\frac{1286109}{7}}} = \frac{2480.571}{877.8634} = 2.825691$
Using the RStudio to find the p-value of P(Z<2.825691):
P = 2*pnorm(-abs(2.825691)) = 0.004717875
P <0.05
Since a p-value of 0.004717875 is less than the 0.05 level of significance, the null hypothesis should be rejected and it should be concluded that there is sufficient evidence to suggest that the average number of daily cases reported by the population differs between the two seven-day blocks

Question 2:

1. To plot a graph of the negative binomial probability mass function given for the values y∈{0, 1, ⋯ , 25},for (v = 0, r = 1), (v = 1, r = 2) and (v = 1.5, r = 2), use the R plot below



2. $P(y|v,r) = \binom{y+r-1}{y} r^r (e^v + r)^{-r-y} e^{yv} = \frac{(y+r-1)!}{y!(r-1)!} r^r (e^v + r)^{-r-y} e^{yv}$

$P(y_i|v,r) = \binom{y_i+r-1}{y_i} r^r (e^v + r)^{-r-y_i} e^{y_i v} = \frac{(y_i+r-1)!}{y_i!(r-1)!} r^r (e^v + r)^{-r-y_i} e^{y_i v}$

Joint probability

$= P((y_1 \cap y_2 \cap y_3 \cap \dots \cap y_n)|v,r)$

$= P(y_1|v,r) \times P(y_2|v,r) \times P(y_3|v,r) \times \dots \times P(y_n|v,r)$

$= \prod_{i=1}^{n} P(y_i|v,r)$

$= \left(\frac{(y_1+r-1)!}{y_1!(r-1)!} r^r (e^v + r)^{-r-y_1} e^{y_1 v}\right) * \left(\frac{(y_2+r-1)!}{y_2!(r-1)!} r^r (e^v + r)^{-r-y_2} e^{y_2 v}\right) \dots \left(\frac{(y_n+r-1)!}{y_n!(r-1)!} r^r (e^v + r)^{-r-y_n} e^{y_n v}\right)$

$= \frac{(y_1+r-1)! \times (y_2+r-1)! \times \dots \times (y_n+r-1)!}{y_1!(r-1)! \times y_2!(r-1)! \times \dots \times y_n!(r-1)!} r^{nr} (e^v + r)^{(-r-y_1)+(-r-y_2)+\cdots+(-r-y_2)} e^{(y_1 v)+(y_2 v)+\cdots+(y_n v)}$

$= \frac{\prod_{i=1}^{n}(y_i+r-1)!}{\prod_{i=1}^{n} y_i!(r-1)!} r^{nr} (e^v + r)^{-nr-\sum_{i=1}^{n} y_i} e^{v \sum_{i=1}^{n} y_i}$

The joint probability of this sample of data is $\frac{\prod_{i=1}^{n}(y_i+r-1)!}{\prod_{i=1}^{n} y_i!(r-1)!} r^{nr} (e^v + r)^{-nr-\sum_{i=1}^{n} y_i} e^{v \sum_{i=1}^{n} y_i}$

3. $\prod_{i=1}^{n} P(y_i|v,r) = \frac{\prod_{i=1}^{n}(y_i+r-1)!}{\prod_{i=1}^{n} y_i!(r-1)!} r^{nr} (e^v + r)^{-nr-\sum_{i=1}^{n} y_i} e^{v \sum_{i=1}^{n} y_i}$

Negative Log Likelihood

$= -\log\left[P\left((y_1 \cap y_2 \cap y_3 \cap \dots \cap y_n)\middle|v,r\right)\right]$

$= -\log\left[\frac{\prod_{i=1}^{n}(y_i+r-1)!}{\prod_{i=1}^{n} y_i!(r-1)!} r^{nr} (e^v + r)^{-nr-\sum_{i=1}^{n} y_i} e^{v \sum_{i=1}^{n} y_i}\right]$

$= -\left[\log[\prod_{i=1}^{n}(y_i + r - 1)!] - \log[\prod_{i=1}^{n} y_i!\,(r-1)!] + \log[r^{nr}] + \log\left[(e^v + r)^{-nr-\sum_{i=1}^{n} y_i}\right] + \log[e^{v \sum_{i=1}^{n} y_i}]\right]$

$= -\sum_{i=1}^{n} \log[(y_i + r - 1)!] + \sum_{i=1}^{n} \log[y_i!\,(r-1)!] - nr \log r + (nr + \sum_{i=1}^{n} y_i) \log[e^v + r] - v \sum_{i=1}^{n} (y_i) \log[e]$

$= -\sum_{i=1}^{n} \log[(y_i + r - 1)!] + \sum_{i=1}^{n} \log[y_i!\,(r-1)!] - nr \log r + (nr + \sum_{i=1}^{n} y_i) \log[e^v + r] - v \sum_{i=1}^{n} y_i$

The negative log-likelihood of the data is

$$-\sum_{i=1}^{n} \log[(y_i + r - 1)!] + \sum_{i=1}^{n} \log[y_i!\,(r-1)!] - nr \log r + (nr + \sum_{i=1}^{n} y_i) \log[e^v + r] - v \sum_{i=1}^{n} y_i$$

4. Maximum Likelihood estimator for v:

0 = $\frac{d}{dv}(-\log[P((y_1 \cap y_2 \cap y_3 \cap \dots \cap y_n)|v,r)])$

0 =$\frac{d}{dv}(-\sum_{i=1}^{n}\log[(y_i+\mathrm{r}-1)!]+\sum_{i=1}^{n}\log[y_i!\,(r-1)!]-nr\log r+r\sum_{i=1}^{n}y_i\log[e^v+r]-v\sum_{i=1}^{n}y_i)$

0=$\frac{d}{dv}(\sum_{i=1}^{n}\log[(y_i+\mathrm{r}-1)!])$+$\frac{d}{dv}(\sum_{i=1}^{n}\log[y_i!\,(r-1)!])$-$\frac{d}{dv}(nr\log r)$+$\frac{d}{dv}((nr+\sum_{i=1}^{n}y_i)\log[e^v+r])$-

$\frac{d}{dv}(v\sum_{i=1}^{n}y_i)$

0 = 0+0-0+$\frac{e^v}{e^v+r}(nr+\sum_{i=1}^{n}y_i)-\sum_{i=1}^{n}y_i$

So, $\frac{e^v}{e^v+r}(nr+\sum_{i=1}^{n}y_i)=\sum_{i=1}^{n}y_i$

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

$$nre^v+e^v\sum_{i=1}^{n}y_i=e^v\sum_{i=1}^{n}y_i+r\sum_{i=1}^{n}y_i$$

$$nre^v=r\sum_{i=1}^{n}y_i$$

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

$$e^v=\frac{n\bar{Y}}{n}$$

$$e^v=\bar{Y}$$

$\hat{v}$ =ln($\bar{Y}$)

5. Approximate bias and variance of the maximum likelihood estimator v^ for v

$\hat{v}$ =ln($\bar{Y}$)

Var($\bar{Y}$)= Var($\frac{\sum_{i=1}^{n}y_i}{n}$) =$\frac{1}{n}\sigma^2$

$E[\bar{Y}]$=E($\frac{1}{n}\sum_{i=1}^{n}y_i$)=$\mu$

$E[\hat{v}(y)]$=$E[\ln(\bar{Y})]\approx \ln(\mu)\ -\frac{1}{2n\mu^2}\sigma^2$

$b_v(\hat{v})=\ E[\hat{v}(y)]-\hat{v}=\ln(\mu)\ -\frac{1}{2n\mu^2}\sigma^2-\ln(\bar{Y})=-\frac{1}{2nre^v}(e^v+r)$

$Var[\hat{v}(y)]=\ V[\ln(\bar{Y})]\approx\sigma^2\left(\frac{1}{\bar{Y}}\right)^2=\frac{1}{nre^{3v}}(e^v+r)$

Question3:

1. 240 volunteer students, 176 to right, 64 to left.

$\hat{\theta}_{right}$= 176/240

$Z_{\frac{\alpha}{2}}$= $Z_{\frac{0.05}{2}}$=1.96

V= $\hat{\theta}_{right}(1-\hat{\theta}_{right})$

According the interval:

$$\left(\hat{\theta}_{\mathrm{ML}}-1.96\sqrt{\frac{\hat{\theta}_{\mathrm{ML}}(1-\hat{\theta}_{\mathrm{ML}})}{n}},\ \hat{\theta}_{\mathrm{ML}}+1.96\sqrt{\frac{\hat{\theta}_{\mathrm{ML}}(1-\hat{\theta}_{\mathrm{ML}})}{n}}\right)$$

95% CI of $\hat{\theta}_{right}=(\frac{176}{240}-1.96\sqrt{\frac{\frac{176}{240}\left(1-\frac{176}{240}\right)}{240}},\frac{176}{240}+1.96\sqrt{\frac{\frac{176}{240}\left(1-\frac{176}{240}\right)}{240}})$

= (0.6773852, 0.7892815)

The estimated of the preference for humans turning their heads to the right is 0.733. We are 95% confident the population mean for this group is between 0.6773852 and 0.7892815.

2. To test the hypothesis that humans do not prefer to tilt their heads to one side when kissing. That is, we set the null hypothesis to be that humans are equally likely to tilt their heads to the left or right when kissing.

$$H_0: \hat{\theta}_{right} = \hat{\theta}_{left}$$

$$vs$$

$$H_1: \hat{\theta}_{right} \neq \hat{\theta}_{left}$$

Using all of the data collected, 240 volunteer students, 176 to right, 64 to left, we can obtain an approximate z-score to use as our test statistic

$$\hat{\theta}_p = 0.5$$

Our estimate of $\hat{\theta}$ is 0.5; this gives us an approximate z-score of

$z_{\hat{\theta}}=\frac{\hat{\theta}_{right}-\hat{\theta}_p}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)\frac{1}{n}}}=\frac{\frac{176}{240}-0.5}{\sqrt{0.5(1-0.5)(\frac{1}{240})}}$ ≈7.229569

And a p-value of

P= 2*pnorm(-|$z_{\hat{\theta}}$|) = 2*pnorm(-7.229569) ≈4.845293× 10^{-13}

This p-value is incredibly small, suggesting that under the null hypothesis that humans have no preference to tilt their heads to one side or the other when kissing, the observed differences are unlikely to have arisen by chance.

3. Use binom.test()function:

binom.test(x=176,n=240,p=0.5)= 2.854× 10^{-13}

The exact p-value is 2.854e-13, which is a bit than our approximate procedure, but gives the same overall conclusion. If the sample size was larger we would expect the two p-values to be closer, as the normal approximation on which our approximate method is based would be better.

4. To test the hypothesis that the proportion of participants who were right-handed was the same as the proportion who tended to turn their heads to the right when kissing. That is, we set the null hypothesis that the same proportion of right-handed people tilt their heads to the right when kissing.

$$H_0: \hat{\theta}_{rhand} = \hat{\theta}_{right}$$
$$vs$$
$$H_1: \hat{\theta}_{rhand} \neq \hat{\theta}_{right}$$

For out two samples, we have

$$\hat{\theta}_{rhand} = \frac{210}{240}$$
$$\hat{\theta}_{right} = \frac{176}{240}$$
$$\hat{\theta}_p = \frac{m_{rhand} + m_{right}}{n_{rhand} + n_{right}} = \frac{210 + 176}{240 + 240} = \frac{193}{240}$$

$z_{\hat{\theta}} = \frac{\hat{\theta}_{rhand} - \hat{\theta}_{right}}{\sqrt{\hat{\theta}_p(1-\hat{\theta}_p)(\frac{1}{n}+\frac{1}{n})}} = \frac{\frac{210}{240} - \frac{176}{240}}{\sqrt{\frac{193}{240}(1-\frac{193}{240})(\frac{1}{240}+\frac{1}{240})}}$=3.910587

P = 2*pnorm(-3.910587) = 9.207207e-05

This p-value is a very small number and < 0.001, indicating that under the null hypothesis that the proportion of participants who were right-handed was the same as the proportion who tended to turn their heads to the right when kissing, the observed differences were unlikely to have arisen by chance.