Jiayu Wang 1039580

Q1

(a)

Here, H_0 is $m_x = 15$, H_1 is $m_x < 15$. Set Y to be the number of positive numbers amongst $X_1 - 15 \ldots X_{15} - 15$, thus Y = 14.

```
> #Here, H0 is mx = 15, H1 is mx < 15.
> #Set Y to be the number of positive numbers amongst X1 - 15 ... X15 - 15, thus Y = 14
> binom.test(14, 21, alternative = "less")

        Exact binomial test

data:  14 and 21
number of successes = 14, number of trials = 21, p-value = 0.9608
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.8318242
sample estimates:
probability of success
             0.6666667
```

Since p-value is 0.9608 > 0.05, we do not reject H_0. Hence the median of daily new coronavirus cases is not below 15 at $\alpha = 0.05$.

(b)

Here, H_0 is $m_{w2} = m_{w3}$, H_1 is $m_{w2} > m_{w3}$. Set U to be the number of the times that daily case number in week2 >= that in week3.

```
> #Here, Ho is mw2 = mw3, H1 is mw2 > mw3
> #Set U to be the number of the times that daily case number in week2 >= that in week3
> w2 <- c(39,41,25,44,20,13,11)
> w3 <- c(28,14,11,13,12,16,5)
> wilcox.test(w2, w3, alternative = "greater")

        Wilcoxon rank sum test with continuity correction

data:  w2 and w3
W = 38, p-value = 0.04798
alternative hypothesis: true location shift is greater than 0
```

Since $U \approx W = 38$, p-value is 0.04798 < 0.05, we reject H_0. Hence the median number of daily new coronavirus cases in the second week is higher than that in the third week at $\alpha = 0.05$.

Q2

(a) $F(x) = \int_0^x f(x)\,dx = \int_0^x \lambda e^{-\lambda x}\,dx = \lambda\cdot(-\frac{1}{\lambda}e^{-\lambda x})\Big|_0^x = 1 - e^{-\lambda x},\ x \geq 0$

$F(\pi_p) = 1 - e^{-\lambda \pi_p} = p$

$\Rightarrow e^{-\lambda\pi_p} = 1-p \Rightarrow -\lambda\pi_p = \log(1-p) \Rightarrow \pi_p = -\frac{1}{\lambda}\log(1-p)$

(b) for type 7 quantile, $\hat{\pi}_p = X_{(k)}$, where $p = \frac{k-1}{n-1}$, $p=0.25$. $n=30$, hence $k = 0.25\times 29 + 1 = 8.25$

$\Rightarrow \hat{\pi}_{0.25} = X_{(8.25)} = X_{(8)} + 0.25\cdot[X_{(9)} - X_{(8)}] = 1.83 + 0.25[1.93-1.83] = 1.83 + 0.025 = 1.855$

(c) Since $f(x) = \lambda\cdot e^{-\lambda x}$, $\hat{\pi}_p \overset{d}{\approx} N(\pi_p, \frac{p(1-p)}{n f(\pi_p)^2})$, $p = 0.25$, $n=30$

$\Rightarrow f(\pi_p) = \lambda e^{-\lambda\pi_p} = \lambda\cdot e^{-\lambda(-\frac{1}{\lambda}\log(1-p))} = \lambda\cdot e^{\log(1-p)} = \lambda(1-p) \Rightarrow \frac{p(1-p)}{n\cdot f(\pi_p)^2} = \frac{0.25\times 0.75}{30\lambda^2\cdot 0.75^2} = \frac{1}{90\lambda^2}$

hence $\hat{\pi}_{0.25} \overset{d}{\approx} N(\pi_{0.25}, \frac{1}{90\lambda^2})$

(d) from 2(c), we knew $Var[\hat{\pi}_{0.25}] = \frac{1}{90\lambda^2} \Rightarrow sd[\hat{\pi}_{0.25}] = \frac{1}{3\sqrt{10}\lambda}$

$\bar{X} = \frac{1}{30}(0.11+0.21+0.75+1.14+1.35+1.63+1.63+1.83+1.93+2.04+2.16+2.25+2.41+2.52+2.65$
$+2.83+2.92+4.83+7.23+8.80+9.80+11.54+12.16+12.91+13.93+19.68+20.94+21.73+24.09)$

$= 6.697333$

Since $\hat{\lambda} = \frac{1}{\bar{X}}$, so $se[\hat{\pi}_{0.25}] = \frac{1}{3\sqrt{10}\hat{\lambda}} = \frac{\bar{X}}{3\sqrt{10}} = \frac{6.697333}{3\sqrt{10}} \approx 0.706$

Q3

(a) $f(x_1, \cdots x_n|\beta) = \mathcal{L}(\beta) = \prod_{i=1}^n f(x_i|\beta) = \prod_{i=1}^n \beta^2 x_i e^{-\beta x_i} = \beta^{2n}\cdot\prod_{i=1}^n x_i\cdot e^{-\beta\sum_{i=1}^n x_i} \propto \beta^{2n}\cdot e^{-\beta\sum_{i=1}^n x_i}$

$f(\beta) = e^{-\beta}$, hence $f(\beta|x_1,\cdots x_n) = \frac{f(x_1\cdots x_n|\beta)\cdot f(\beta)}{f(x_1,\cdots x_n)} \propto \beta^{2n}\cdot e^{-\beta\sum_{i=1}^n x_i}\cdot e^{-\beta} = \beta^{2n}\cdot e^{-\beta(1+\sum_{i=1}^n x_i)}$

$\propto \beta^{(2n+1)-1}\cdot e^{-(1+\sum_{i=1}^n x_i)\beta}$

hence $\beta|X \sim Gamma(2n+1, 1+\sum_{i=1}^n x_i)$

(b) for $\beta|X \sim Gamma(a,b)$, $E[\beta|X] = \frac{a}{b}$, $Var[\beta|X] = \frac{a}{b^2}$, where $a = 2n+1$, $b = 1+\sum_{i=1}^n x_i$

$\Rightarrow E[\beta|X] = \frac{a}{b} = \frac{2n+1}{1+\sum_{i=1}^n x_i}$

$Var[\beta|X] = \frac{a}{b^2}$, $sd[\beta|X] = \sqrt{Var[\beta|X]} = \frac{\sqrt{a}}{b} = \frac{\sqrt{2n+1}}{1+\sum_{i=1}^n x_i}$

Q4

$f(x_1,\cdots,x_n|\mu,\sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \prod_{i=1}^n \exp[-\frac{1}{2}\cdot\ln(2\pi\sigma^2) - (\frac{x_i^2}{2\sigma^2} - \frac{2\mu x_i}{2\sigma^2} + \frac{\mu^2}{2\sigma^2})]$

$= \prod_{i=1}^n \exp[-\frac{x_i^2}{2\sigma^2} + \frac{2\mu x_i}{2\sigma^2} - (\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\cdot\ln(2\pi\sigma^2))] = \exp[-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n x_i - n(\frac{\mu^2}{2\sigma^2} + \frac{1}{2}\cdot\ln(2\pi\sigma^2))]$

(a) If μ is unknown, σ^2 is know, then according to the expansion above, $\sum_{i=1}^n x_i$ is a sufficient statistic for μ.

(b) If μ is known and σ^2 is unknown, $(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$ are jointly sufficient for σ^2.

(c) If μ is known and σ^2 is unknown, $(\sum_{i=1}^n x_i^2, \sum_{i=1}^n x_i)$ are jointly sufficient for σ.

Q6

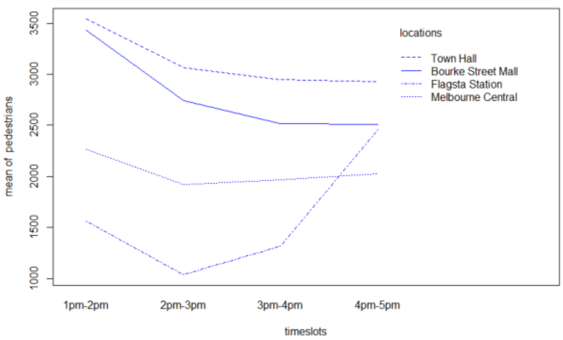
```
> pedestrians <- c(1706,1636,1339,2387,2284,2116,3715,3541,3369,3715,3689,2884,
+                  1063,1065,977,2062,1885,1819,3209,2907,3077,2940,2753,2525,
+                  1380,1306,1261,2108,1896,1893,3030,2837,2978,2751,2508,2288,
+                  2539,2544,2297,1980,2025,2064,2964,2824,2987,2687,2423,2429)
> timeslots <- gl(4,12,48,labels= c("1pm-2pm","2pm-3pm","3pm-4pm","4pm-5pm"))
> locations <- gl(4,3,48,labels=c("Flagsta Station","Melbourne Central","Town Hall","Bourke Street Mall"))
> record <- data.frame(timeslots,locations,pedestrians)
> head(record)
  timeslots         locations pedestrians
1   1pm-2pm   Flagsta Station        1706
2   1pm-2pm   Flagsta Station        1636
3   1pm-2pm   Flagsta Station        1339
4   1pm-2pm Melbourne Central        2387
5   1pm-2pm Melbourne Central        2284
6   1pm-2pm Melbourne Central        2116
```

```
> model <- lm(pedestrians ~ factor(timeslots)+factor(locations),data = record)
> anova(model)
Analysis of Variance Table

Response: pedestrians
                  Df   Sum Sq Mean Sq F value    Pr(>F)
factor(timeslots)  3  2216082  738694  6.1731  0.001462 **
factor(locations)  3 17472573 5824191 48.6717 1.425e-13 ***
Residuals         41  4906178  119663
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Here, H_0 is pedestrian numbers do not vary by time. H_1 is pedestrian numbers vary by time. From this, we know p-value = 0.001462 (boxed in red) < 0.05. F statistic used is 6.1731, which is greater than $F_{3,9}^{-1}$ = 3.862548. Hence, we conclude there is a clear difference in pedestrian numbers between time slots, we reject H_0 at 5% significance level.

(Also, for whether pedestrian numbers vary between different locations, H_0 is pedestrian numbers do not vary between different locations, H_1 is pedestrian numbers vary between different locations. From R results, we know p-value for this is 1.425e-13 < 0.05. Hence, we conclude there is a difference in pedestrian numbers between different locations, we reject H_0 at 5% significance level.)

```
> model2 <- lm(pedestrians~factor(timeslots)*factor(locations), data = record)
> anova(model2)
Analysis of Variance Table

Response: pedestrians
                                    Df   Sum Sq Mean Sq F value    Pr(>F)
factor(timeslots)                    3  2216082  738694  22.348 5.431e-08 ***
factor(locations)                    3 17472573 5824191 176.202 < 2.2e-16 ***
factor(timeslots):factor(locations)  9  3848452  427606  12.937 2.190e-08 ***
Residuals                           32  1057727   33054
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> with(record, interaction.plot(timeslots, locations, pedestrians, col = "blue"))
```



Yes, it is possible to test for interaction (the codes and plot are provided above). Here we set H_0 is the interaction between the factors is 0, H_1 is there is interaction between the factors.

From the R results, we conclude the interaction is significant (p-value = 2.190e-08 < 0.1%) at a 0.1% level.