Assignment 3

Kim CHY

October 2020

Question 1:

(a): Testing to see if the median cases is below 15.

```
#Entering 3 weeks of covid data
covid_data <- c(48,70,47,40,35,41,30,
                39,41,25,44,20,13,11,
                28,14,11,13,12,16,05)

#H0: m=m0 vs H1:m<15


#Conducting a wilcoxon to see if median cases is below 15
wilcox.test(covid_data, mu = 15, alternative = "less", exact = FALSE,
            correct = TRUE)

##
##  Wilcoxon signed rank test with continuity correction
##
## data:  covid_data
## V = 195, p-value = 0.9973
## alternative hypothesis: true location is less than 15
```

Reading from the R output from our wilconxon signed-ranked test, we can see the p-value is 0.9973, which is greater then 0.05 by a significantly margin, thus implying there are close to zero evidence to support that the median number of new cases is below 15.

(b): Testing to see if the median cases in week 2 is greater then week 3.

```
#Entering week 2 and week 3 data as a seperate vector
covid_week2 = c(39,41,25,44,20,13,11)
covid_week3 = c(28,14,11,13,12,16,05)

#H0: m2=m3 vs H1:m2>m3
#Conducting a wilcoxon sign test to compare median between week 2 and week 3
wilcox.test(covid_week2, covid_week3, alternative ="greater", exact = FALSE)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  covid_week2 and covid_week3
## W = 38, p-value = 0.04798
## alternative hypothesis: true location shift is greater than 0
```

Reading from the R output of the our wilcoxon rank-sum test, we can see the p-value is 0.04798, thus implying there is a statically significant evidence to reject the null at a significant level of 0.05, which indicate that the median number of corona virus cases in Victoria between 14th and 20th of September is higher then cases between 21st to 27th of September.
Despite the statically significant evidence we still need to be careful in our decision making based on the test as p-value is relatively close to the critical value.

Question 2:

(a): Find the p quantile, π_p.

Under the definition of population quantile

$$\begin{aligned}
&P(X \leqslant \pi_p) = F(\pi_p) = p && (0 \leq p \leq 1)\\
&F(\pi_p) = \textstyle\int_0^{\pi_p} \lambda \mathbf{e}^{-\lambda x}\, dx = [\mathbf{e}^{-\lambda x}]_0^{\pi_p} && (0 \leq p \leq 1)\\
&\implies p = F(\pi_p) = 1 - \mathbf{e}^{-\lambda \pi_p} && (0 \leq p \leq 1)\\
&\implies \mathbf{e}^{-\lambda \pi_p} = 1 - p && (0 \leq p \leq 1)\\
&\implies \pi_p = -\lambda \log(1-p) && (0 \leq p \leq 1)
\end{aligned}$$

Therefore, under the definition of population quantile, p quantitle, $\pi_p = -\lambda \log 1 - p$.

(b): Calculate the 'Type 7' sample quantile $\hat{\pi}_{0.25}$.

Under the R defintion of Type 7 quantile, $_p = x_{(k)}$, where $p = \frac{k-1}{n-1}$ and k is it position in the order statistics of the sample.

Since $n = 30$ and $p = 0.25$
$\implies k = 1 + 0.25(30-1) = \frac{33}{4} = 8.25$
$\therefore \hat{\pi}_{0.25} = x_{(8.25)} = x_{(8)} - 0.25(x_{(9)} - x_{(8)})$
$= 1.83 + 0.25(1.93 - 1.83) = 1.855$
Therefore, 25^th sample quantile under the type 7 definition is 1.855.

(c): Find the asymptotic distribution of $\hat{\pi}_{0.25}$.

$$\begin{aligned}
f(\pi_{0.25}) &= f(-\lambda \log(1 - 0.25))\\
&= \lambda \mathbf{e}^{-\lambda(-\lambda(1-0.25)}\\
&= \lambda(\tfrac{3}{4})^{\lambda^2}
\end{aligned}$$

The asymptotic distribution, of sample quantile for large sample size can be approximate as:

$$\hat{\pi}_p \approx N\left(\pi_p, \frac{p(1-p)}{n f(\pi_p)^2}\right)$$

$$\implies \pi_{0.25} \approx N\left(\hat{\pi}_{0.25}, \tfrac{0.25(1-0.25)}{n f(\pi_{0.25})^2}\right)$$
$$\implies \hat{\pi}_{0.25} \approx N\left(1.855, \tfrac{1}{160(\lambda(\frac{3}{4})^{\lambda^2})}\right)$$

(d): Calculate a standard error for $\hat{\pi}_{0.25}$.

From question 2 part c we know:

$$\hat{\pi}_{0.25} \approx N\left(1.855, \frac{1}{160(\lambda(\frac{3}{4})^{\lambda^2})}\right)$$

Since we know the population distribution of our sample, we can MLE to approximate. From MLE of an exponential distribution, we know $\hat{\lambda} = (\bar{X})^{-1} = \frac{n}{\sum_{i=1}^{n} x_i}$, and for our sample above:

$\hat{\lambda} = \frac{30}{200.1} = 0.1493$
$\implies \mathbf{SE}(\hat{\pi}_{0.25}) = \sqrt{\frac{1}{160(\lambda(\frac{3}{4})^{\lambda^2})}}$
$= \sqrt{\frac{1}{160(0.1493(\frac{3}{4})^{0.1493^2})}} = 0.2053$

Therefore, the standard error for the $\hat{\pi}_{0.25}$ is 0.2053.

Question 3:

(a): Derive the posterior distribution of β.

Let $L(\beta)$ be the likelihood function of $f(x)$:
$L(\beta) = \prod_{i=1}^{n} f(x_i)$

$$= \prod_{i=1}^{n} \beta^2 x_i e^{-\beta x_i}$$

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

From Bayesian theory,the posterior $f(\beta | \prod_{i=1}^{n} x_i) \propto L(\beta) f(\beta)$:

$$\therefore f(\beta | \prod_{i=1}^{n} x_i) \propto (\beta^{2n} (\prod_{i=1}^{n} x_i) e^{-\beta \sum_{i=1}^{n} x_i})(e^{-\beta})$$

Since, $\prod_{i=1}^{n} x_i$ is does not depend on β, we can factorise it out, which implies:

$$\begin{aligned} f(\beta | \prod_{i=1}^{n} x_i) &\propto (\beta^{2n} e^{-\beta \sum_{i=1}^{n} x_i})(e^{-\beta}) \\ &\propto \beta^{2n} e^{-\beta(\sum_{i=1}^{n} x_i + 1)} \end{aligned}$$

Reading from the above, it is in the form of a gamma distribution with parameter 2n+1 and $(\sum_{i=1}^{n} x_i + 1)$, which implies our posterior distribution is:
$\beta | \prod_{i=1}^{n} x_i \sim Gamma(2n+1, \sum_{i=1}^{n} x_i + 1)$.

(b): Derive the posterior mean and the posterior standard deviation of β.

From question 3 part a we know our posterior distribution follow: $\beta | \prod_{i=1}^{n} x_i \sim Gamma(2n+1, \sum_{i=1}^{n} x_i + 1)$. Henceh, we can use the known population of Gamma distribution standard deviation.
$\implies Std(\beta | \prod_{i=1}^{n} x_i) = \sqrt{Var(\beta | \prod_{i=1}^{n} x_i)} = \sqrt{\frac{2n+1}{(\sum_{i=1}^{n} x_i + 1)^2}}$

Question 4:

(a):If μ is unknown and σ^2 is known, find a sufficient statistic for μ.

Let $L(\mu)$ be the likelihood function of $f(x)$:

$$\begin{aligned} L(\mu) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(x_i-\mu)^2}{2\sigma^2}\right) \\ &= (2\pi)^{\frac{-n}{2}} (\sigma)^{-n} \exp\left(\frac{\sum_{i=1}^{n} -(x_i-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

Since we know σ^2, this mean we can ignored $(2\pi)^{\frac{-n}{2}}(\sigma)^{-n}$ as they are just constant.

$\exp\left(\frac{\sum_{i=1}^{n} -(x_i-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{-\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right) \exp\left(\frac{\mu\sum_{i=1}^{n} x_i}{2\sigma^2} - \frac{n\mu}{2\sigma^2}\right)$

Since, we can factorise the likelihood into two part with one of them does not depend on μ, this $\sum_{i=1}^{n} x_i$ is a sufficient statistics for μ.

(b):If μ is know and σ^2 is unknown, find a sufficient statistic for σ^2.

$$\begin{aligned} L(\sigma) &= (2\pi)^{\frac{-n}{2}} (\sigma)^{-n} \exp\left(\frac{\sum_{i=1}^{n} -(x_i-\mu)^2}{2\sigma^2}\right) \\ &= \exp(-\frac{n}{2}\log(2\pi)) \exp\left(\frac{\sum_{i=1}^{n} -(x_i-\mu)^2}{2\sigma^2} - n\log(\sigma)\right) \end{aligned}$$

Similar to question 4 part a, since we can factorise them into the form of an exponential family, this implied $\sum_{i=1}^{n} -(x_i-\mu)^2$ is a sufficient statistics for σ^2.

(c):If μ is know and σ^2 is unknown, find a sufficient statistic for σ.

Since $\sigma = \sqrt{\sigma^2}$, this implied $\sum_{i=1}^{n} -(x_i-\mu)^2$ is also a sufficient statistics for σ.

Question 5:

(a): Derive the likelihood ratio test and show it is based on the statistic $Y = \sum_{i=1}^{n} X_i$.

$H_0 : \lambda = \lambda_0$ vs $H_1 : \lambda : \lambda \neq \lambda_0$

For likelihood test of an exponential distribution: $\lambda = \frac{L_0}{L_1}$

To maximise L under H_0, we need to show $\hat{\lambda} = \lambda_0$, which give:

$$\begin{aligned} L_0 = L(\lambda_0) &= \prod_{i=1}^{n} \lambda_0 \mathbf{e}^{-\lambda_0 x_i} \\ &= \lambda_0^n \mathbf{e}^{-\lambda_0 \sum_{i=1}^{n} x_i} \end{aligned}$$

Now, to maximise L under H_1, we need to show $\lambda = (\bar{X})^{-1}$, which give:

$$\begin{aligned} L_1 = L((\bar{X})^{-1}) &= \prod_{i=1}^{n} \frac{1}{\bar{X}} \mathbf{e}^{\frac{-x_i}{\bar{X}}} \\ &= \bar{X}^{-n} \mathbf{e}^{\frac{-\sum_{i=1}^{n} x_i}{\bar{X}}} \end{aligned}$$

$$\begin{aligned} \lambda &= \frac{L_0}{L_1} \\ &= \frac{\lambda_0^n \mathbf{e}^{-\lambda_0 \sum_{i=1}^{n} x_i}}{\bar{X}^{-n} \mathbf{e}^{\frac{-\sum_{i=1}^{n} x_i}{\bar{X}}}} \\ &= \left(\frac{\bar{X}}{\lambda_0}\right)^n e^{\frac{\sum_{i=1}^{n} x_i}{\bar{x}} - \lambda \sum_{i=1}^{n} x_i} \end{aligned}$$

Now we need to rearrange $\lambda \leqslant k$
$\implies \left(\frac{\bar{X}}{\lambda_0}\right)^n \exp\left(\frac{\sum_{i=1}^{n} x_i}{\bar{x}} - \lambda \sum_{i=1}^{n} x_i\right) \leqslant k \implies \exp\left(\frac{\sum_{i=1}^{n} x_i}{\bar{x}} - \lambda_0 \sum_{i=1}^{n} x_i\right) \leqslant \left(\frac{\lambda_0}{\bar{X}}\right)^n k$
$\implies \sum_{i=1}^{n} x_i \left(\bar{x}^{-1} - \lambda_0\right) \leqslant \log\left(\left(\frac{\lambda_0}{\bar{X}}\right)^n k\right) = \log(k) + n\log\left(\lambda_0\right) - n\log\left(\bar{X}\right)$
$\implies \sum_{i=1}^{n} x_i \leqslant \frac{\log(k) + n\log\left(\lambda_0\right) - n\log\left(\bar{X}\right)}{\left(\bar{x}^{-1} - \lambda_0\right)}$
$\implies Y \leqslant k^*$
where $Y = \sum_{i=1}^{n} x_i$ and $k^* = \frac{\log(k) + n\log\left(\lambda_0\right) - n\log\left(\bar{X}\right)}{\left(\bar{x}^{-1} - \lambda_0\right)}$

Following the same process but for the lower tail we would get: $c^* \leqslant$ Y
where $Y = \sum_{i=1}^{n} x_i$ and $c^* = \frac{\log(c) + n\log\left(\lambda_0\right) - n\log\left(\bar{X}\right)}{\left(\bar{x}^{-1} - \lambda_0\right)}$

From the above we can see, the likelihood ratio test of an exponential distribution is based on Y.

(b): What is the distribution of Y when H_0 is true?

We know that if H_0 is true then the set of our exponential distribution will follow $2\lambda n\bar{X}$ will follow a chi-squared distribution with 2n degree of freedom:
$2\lambda_0 n\bar{X} = 2\lambda_0 Y \sim \chi^2_{2n}$
$\implies Y \sim \frac{1}{2\lambda_0}\chi 2_{2n}$

(c): For $n = 50$ and $\lambda_0 = 1$, find a test based on Y with significance level 0.05.

From question 5 part A we can see that the test exist $c^* \leqslant Y \leqslant k^*$, this would implies we reject the null if and only if Y is less than the $1 - \frac{\alpha}{2}$ or greater than $\frac{\alpha}{2}$. Therefore, a test on Y with significance level of 0.05 is:
$\implies \frac{1}{2}\Psi^{-1}(0.025) \leqslant Y \leqslant \frac{1}{2}\Psi^{-1}(0.975)$
where $\Psi^{-1}(x)$ is the inverse qunatile distribution of χ^2_{100}
$\implies 37.11096 \leqslant Y \leqslant 64.7806$