MAST20005/MAST90058: Assignment 1

Due date: 11am, Friday 30 August 2019

Instructions: Questions labelled with '**(R)**' require use of R. Please provide appropriate R commands and their output, along with sufficient explanation and interpretation of the output to demonstrate your understanding. **Such R output should be presented in an integrated form together with your explanations; do not attach them as separate sheets.** All other questions should be completed without reference to any R commands or output, except for looking up quantiles of distributions where necessary. Make sure you give enough explanation so your tutor can follow your reasoning if you happen to make a mistake. Please also try to be as succinct as possible. Each assignment will include marks for good presentation and for attempting all problems.

Problems:

1. **(R)** Let X be a random variable representing distance travelled (in kilometers) until a tire is worn out. The following are 16 observations of X:

   41300 40300 43200 41100 39300 42100 42700 41300
   38900 41200 44600 42300 40700 43500 39800 40400

   (a) Give basic summary statistics for these data and produce a box plot. Briefly comment on center, spread and shape of the distribution.

   (b) Assuming a normal distribution, compute maximum likelihood estimates for the parameters.

   (c) Draw a density histogram and superimpose a pdf for a normal distribution using the estimated parameters.

   (d) Draw a QQ plot to compare the data against the fitted normal distribution. Include a reference line. Comment on the fit of the model to the data.

2. A discrete random variable X has the following pmf:

x	1	2	3
$p(x)$	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

   A random sample of size $n = 20$ produced the following observations:

   1, 1, 2, 3, 1, 2, 1, 3, 2, 2, 2, 1, 3, 1, 3, 1, 1, 2, 1, 2.

   (a) i. Find $\mathbb{E}(X)$ and $\text{var}(X)$.
   ii. Find the method of moments estimator and estimate of θ.
   iii. Find the standard error of this estimate.

   (b) Let F_1, F_2 and F_3 denote the sample frequencies of 1, 2 and 3, respectively.
   i. Find the likelihood function in terms of F_1, F_2 and F_3.
   ii. Find that the maximum likelihood estimator and estimate of θ.
   iii. Find the variance of this estimator.
   (*Hint:* write the estimator in terms of the sample mean.)

3. Let $X \sim \text{Unif}(0, \theta)$, a uniform distribution with an unknown endpoint θ.

   (a) Suppose we have a single observation on X.

      i. Find the method of moments estimator (MME) for θ and derive its mean and variance.

      ii. Find the maximum likelihood estimator (MLE) for θ and derive its mean and variance.

   (b) The *mean square error* (MSE) of an estimator is defined as $\text{MSE}(\hat{\theta}) = \mathbb{E}\left[\left(\hat{\theta} - \theta\right)^2\right]$.

      i. Let $\text{bias}(\hat{\theta}) = \mathbb{E}(\hat{\theta}) - \theta$. Show that,

$$\text{MSE}(\hat{\theta}) = \text{var}(\hat{\theta}) + \text{bias}(\hat{\theta})^2.$$

      ii. Compare the MME and MLE from above in terms of their mean square errors.

      iii. Find an estimator with smaller MSE than either of the above estimators.

   (c) Suppose we have a random sample of size n from X.

      i. Find the MME and derive its mean, variance and MSE.

      ii. Find the MLE and derive its mean, variance and MSE.

      iii. Consider the estimator $a\hat{\theta}$ where $\hat{\theta}$ is the MLE. Find a that minimises the MSE.

   Some information that might be useful:

$$\mathbb{E}\left(X_{(1)}\right) = \frac{\theta}{n+1}, \quad \mathbb{E}\left(X_{(1)}^2\right) = \frac{2\theta^2}{(n+1)(n+2)}, \quad \mathbb{E}\left(X_{(n)}\right) = \frac{n\theta}{n+1}, \quad \mathbb{E}\left(X_{(n)}^2\right) = \frac{n\theta^2}{n+2}$$

4. Let $X_1, \ldots, X_n$ be a random sample from the lognormal distribution, Lognormal(μ, λ), whose pdf is:

$$f(x \mid \mu, \lambda) = \frac{1}{x\sqrt{2\pi\lambda}} \exp\left\{-\frac{(\ln x - \mu)^2}{2\lambda}\right\}, \quad x > 0.$$

   (a) Show that the MLE of μ and λ are $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n \ln X_i$ and $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^n (\ln X_i - \hat{\mu})^2$.

   (b) It is known that $\ln X_i \sim \text{N}(\mu, \lambda)$. Derive a $100 \cdot (1-\alpha)\%$ CI for λ.

   (c) **(R)** Consider the following dataset:

```
0.27, 3.30, 4.58,  2.61, 0.38, 3.77, 1.11, 1.15, 4.11, 2.10,
0.07, 1.74, 2.11, 12.79, 1.85, 0.30, 0.34, 1.31, 0.14, 0.74
```

      i. Assuming a lognormal distribution is an appropriate model for these data, compute the maximum likelihood estimate of λ and give a 95% CI.

      ii. Draw a QQ plot to compare these data to the fitted lognormal distribution, Lognormal$(\hat{\mu}, \hat{\lambda})$. Is this model appropriate for these data?

   Hint: Quantiles of the lognormal distribution can be computed using the `qlnorm()` function.

5. Let X_1, X_2, X_3, X_4 be iid rvs with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2 > 0$, for $i = 1, 2, 3, 4$. Consider the following four estimators of μ:

$$T_1 = \frac{1}{3}(X_1 + X_2) + \frac{1}{6}(X_3 + X_4) \qquad T_2 = \frac{1}{6}(X_1 + 2X_2 + 3X_3 + 4X_4)$$

$$T_3 = \frac{1}{4}(X_1 + X_2 + X_3 + X_4) \qquad T_4 = \frac{1}{3}(X_1 + X_2 + X_3) + \frac{1}{4}X_4^2$$

   (a) Which of these estimates are unbiased? Show your working.

   (b) Among the unbiased estimators, which one has the smallest variance?

Assignment1 for MAST90058

Name: Mu Tong
Student Number: 1004452

15

Problem 1:

```
#input the data
x <- c(41300,40300,43200,41100,39300,42100,42700,41300,
       38900,41200,44600,42300,40700,43500,39800,40400)
```

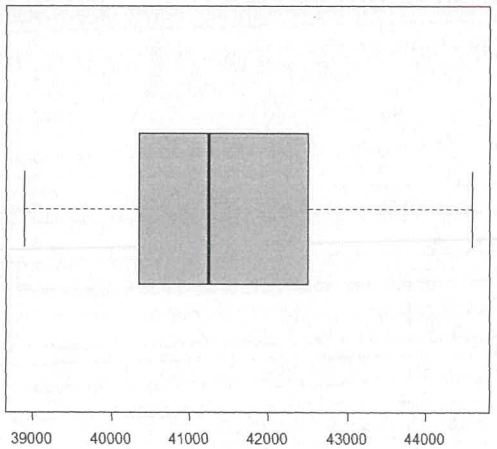
a) Show the basic summary statistics and produce a box plot.

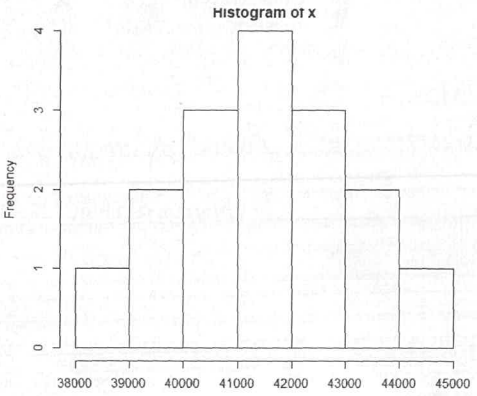
R script:

```
summary(x)
var(x)
sd(x)
IQR(x)

boxplot(x, col = 8, horizontal = TRUE)
hist(x)
```

Output:

```
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38900   40375   41250   41419   42400   44600
> boxplot(x, col = 8, horizontal = TRUE)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  38900   40375   41250   41419   42400   44600
> var(x)
[1] 2462958
> sd(x)
[1] 1569.382
> IQR(x)
[1] 2025
>
> boxplot(x, col = 8, horizontal = TRUE)
```





We can find that it is a roughly symmetric shape according to the histogram, and the center of this distribution is the mean (41419), and the spread of this distribution is the standard deviation (1569.382).

b)

R script:

```
#(b)
library(MASS)
normal.fit <- fitdistr(x, densfun = "normal")
normal.fit
```

Output:

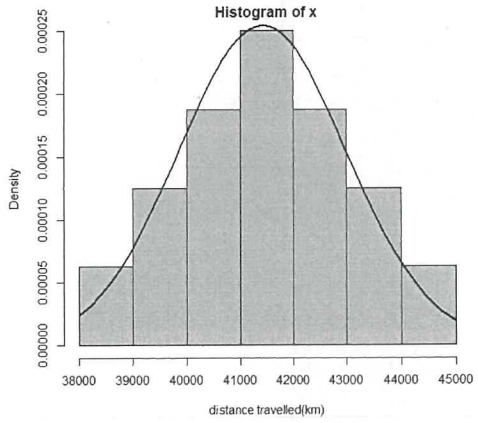
```
> library(MASS)
> normal.fit <- fitdistr(x, densfun = "normal")
> normal.fit
      mean          sd
  41418.7500    1519.5471
 (  379.8868) (  268.6205)
```

We can find that MLE for mean is 4148.7500 and MLE for standard deviation is 1519.5471.

c) R script:

```
pdf <- function(a) dnorm(a, mean = mean(x), sd = sd(x))
hist(x, freq = FALSE, col = "gray", xlab = "distance travelled(km)")
curve(pdf, col = 1, lty = 1, lwd = 2, add = TRUE)
```

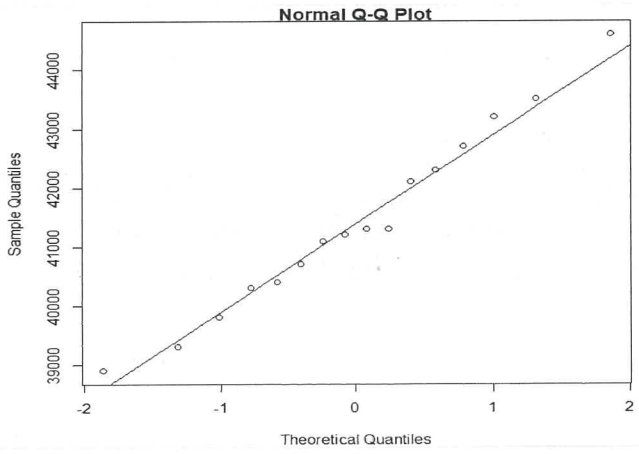
Output:



d)

R script:

```
qqnorm(x)
qqline(x)
```



We can find that the sample data roughly match the theoretical model.

2. pmf:

x	1	2	3
P(x)	θ^2	$2\theta(1-\theta)$	$(1-\theta)^2$

n = 20, 1, 1, 2, 3, 1, 2, 1, 3, 2, 2, 2, 1, 3, 1, 3, 1, 1, 2, 1, 2

x	1	2	3
Fre	9	7	4

6/7

a) [i] $E(X) = \sum_{x=1}^{3} x \cdot P(x) = 1\times\theta^2 + 2\times 2\theta(1-\theta) + 3\times(1-\theta)^2$

$= \theta^2 + 4\theta(1-\theta) + 3\times(\theta^2 - 2\theta + 1)$

$= \theta^2 + 4\theta - 4\theta^2 + 3\theta^2 - 6\theta + 3$

$= -2\theta + 3$

$Var(X) = E(X^2) - E(X)^2 = (2\theta^2 - 10\theta + 9) - (3-2\theta)^2 = -2\theta^2 + 2\theta$

$= 2\theta(-\theta + 1)$

$E(X^2) = \sum_{x=1}^{3} x^2 P(x) = 1\times\theta^2 + 4\times 2\theta(1-\theta) + 9\times(1-\theta)^2$

$= \theta^2 + 8\theta(1-\theta) + 9[\theta^2 - 2\theta + 1)$

$= \theta^2 + 8\theta - 8\theta^2 + 9\theta^2 - 18\theta + 9$

$= 2\theta^2 - 10\theta + 9$

$\therefore E(X) = -2\theta + 3$

$Var(X) = 2\theta(1-\theta) = -2\theta^2 + 2\theta$

[ii] We know that $E(\bar{X}) = -2\theta + 3$, therefore an unbiased estimator of θ based on $\bar{X}$ is $T_1 = \frac{3-\bar{X}}{2}$ $\hat{\theta} = \frac{3-\bar{X}}{2}$

$\bar{x} = \frac{1\times 9 + 2\times 7 + 3\times 4}{20} = 1.75$ $\therefore$ estimate: $t_1 = \frac{3-\bar{x}}{2} = \frac{3-1.75}{2} = 0.625$

$\hat{\theta} = 0.625$

[i-i] $Var(\hat{\theta}) = Var\left(\frac{3-\bar{X}}{2}\right) = \frac{1}{4} Var(\bar{X})$

$\because$ random sample

$\therefore Var(\bar{x}) = \frac{Var(X)}{n} = \frac{-2\theta^2 + 2\theta}{n}$

$\therefore Var(\hat{\theta}) = \frac{1}{4} Var(\bar{X}) = \frac{1}{4} \times \frac{-2\theta^2 + 2\theta}{n} = \frac{-\theta^2 + \theta}{2n}$

$se(\hat{\theta}) = \sqrt{Var(\hat{\theta})} = \sqrt{\frac{-\theta^2 + \theta}{2n}} = \sqrt{\frac{-0.625^2 + 0.625}{2\times 20}} = \frac{\sqrt{6}}{32} \approx 0.0765$

b) i. $L(\theta) = \prod_{i=1}^{3} P(x) = (\theta^2)^{F_1} \cdot [2\theta(1-\theta)]^{F_2} \cdot [(1-\theta)^2]^{F_3}$

$= 2^{F_2} \cdot \theta^{2F_1+F_2} \cdot (1-\theta)^{F_2+2F_3}$

ii. $\ln L(\theta) = F_2 \ln 2 + (2F_1+F_2)\ln\theta + (F_2+2F_3)\ln(1-\theta)$

$$\frac{\partial \ln L(\theta)}{\partial \theta} = 0 + \frac{2F_1+F_2}{\theta} - \frac{F_2+2F_3}{1-\theta}$$

letting $\frac{\partial \ln L(\theta)}{\partial \theta}$ equals to 0. $\therefore \frac{2F_1+F_2}{\theta} = \frac{F_2+2F_3}{1-\theta}$

$$\therefore \hat{\theta} = \frac{2F_1+F_2}{2(F_1+F_2+F_3)} = \frac{2F_1+F_2}{2n}$$

$$\hat{\theta} = \frac{2F_1+F_2}{2n} = \frac{2\times 9+7}{2\times 20} = 0.625$$

iii. $E(\bar{X}) = \frac{1\times F_1 + 2\times F_2 + 3\times F_3}{F_1+F_2+F_3} = \frac{F_1+2F_2+3F_3}{n}$

$$E(\bar{X}^2) = \frac{1^2\times F_1 + 2^2\times F_2 + 3^2\times F_3}{F_1+F_2+F_3} = \frac{F_1+4F_2+9F_3}{n}$$

$$Var(\bar{X}) = E(\bar{X}^2) - E(\bar{X})^2 = \frac{F_1+4F_2+9F_3}{n} - \left(\frac{F_1+2F_2+3F_3}{n}\right)^2$$

$\therefore \hat{\theta} = \frac{2F_1+F_2}{2n} = \frac{3-\bar{X}}{2} =$

$\therefore Var(\hat{\theta}) = Var\left(\frac{3-\bar{X}}{2}\right) = \frac{1}{4} Var(\bar{X}) = \frac{1}{4} \times \frac{4F_1F_3 + F_2F_3 + F_1F_2}{n^2}$

3. $X \sim Unif(0, \theta)$

a). (ii) We can know that $E(X) = \frac{0+\theta}{2} = \frac{\theta}{2}$, which is 1st moment.

∵ We only have single observation on X

∴ $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = X_1 \quad (n=1)$

∴ $\frac{\hat{\theta}}{2} = X_1 \quad \therefore \hat{\theta} = 2X_1$ is the MME for θ

∴ $E(\hat{\theta}) = E(2X_1) = 2E(\bar{X}) = 2 \times \frac{\theta}{2} = \theta$.

$Var(\hat{\theta}) = E(2X_1) = 4\,Var(X_1) = 4\,Var(\bar{X}) =$

We can know that $Var(X) = \frac{(b-a)^2}{12}$ for $X \sim Unif(a, b)$

∴ $Var(X) = \frac{\theta^2}{12} \quad Var(\bar{X}) = \frac{Var(X)}{n} = \frac{Var(X)}{1} = Var(X)$

∴ $Var(\hat{\theta}) = 4\,Var(\bar{X}) = 4 \times \frac{\theta^2}{12} = \frac{\theta^2}{3}$

(iii) pdf of X_1 is $f(x_1; \theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \le x_1 \le \theta \\ 0, & \text{otherwise} \end{cases}$

∴ $L(\theta) = \prod_{i=1}^{n} f(x; \theta) = \begin{cases} \frac{1}{\theta}, & \text{if } 0 \le x_1 \le \theta \\ 0, & \text{otherwise} \end{cases}$

∵ We want to find the maximum of $L(\theta)$

⇒ θ should be the minimum in $X_1 \le \theta$

∵ ∴ when $\theta = X_1$, θ is the minimum

∴ X_1 is the minimum value of θ

∴ X_1 is the MLE of θ ∴ $\hat{\theta} = X_1$

$E(\hat{\theta}) = E(\bar{X}) = E(X_1) = \frac{\theta}{2}$

$Var(\hat{\theta}) = Var(\bar{X}) = Var(X) = \frac{\theta^2}{12}$

(b) $MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2]$

[i] Let $bias(\hat{\theta}) = E(\hat{\theta}) - \theta$. show $\boxed{MSE(\hat{\theta}) = Var(\hat{\theta}) + bias(\hat{\theta})^2}$

$$MSE(\hat{\theta}) = E[(\hat{\theta}-\theta)^2] = E[[(\hat{\theta}-E(\hat{\theta})) + (E(\hat{\theta})-\theta)]^2]$$

$$= E[\hat{\theta}-E(\hat{\theta})]^2 + 2\cdot E[\hat{\theta}-E(\hat{\theta})]\cdot E[E(\hat{\theta})-\theta] + [E(\hat{\theta})-\theta]^2$$

$\because E[\hat{\theta}-E(\hat{\theta})]^2 = Var(\hat{\theta})$, $E(\hat{\theta})-\theta) = bias(\hat{\theta})$

$\because E[\hat{\theta}-E(\hat{\theta})] = E(\hat{\theta})-E(\hat{\theta}) = 0$

$\therefore MSE(\hat{\theta}) = Var(\hat{\theta}) + bias(\hat{\theta})^2$

[iv] MME of the MSE of θ

$$MSE(\hat{\theta}) = MSE(2X_1) = Var(2X_1) + bias(2X_1)^2$$

$$= 4\,Var(X_1) + [E(2X_1) - \theta]^2 \quad E(2X_1) = \theta \text{ from part a) i)}$$

$$= 4\times\frac{\theta^2}{12} + [\theta-\theta]^2 = \frac{\theta^2}{3}$$

MLE of the MSE of θ

$$MSE(\hat{\theta}) = MSE(X_1) = Var(X_1) + bias(X_1)^2$$

$$= Var(X_1) + [E(X_1) - \theta]^2$$

$$= \frac{\theta^2}{12} + [\frac{\theta}{2}-\theta]^2 = \frac{\theta^2}{12} + \frac{\theta^2}{4} = \frac{\theta^2}{3}$$

$\therefore$ MLE and MME of θ is the same.

[iv]

(iii) Find an estimator with smaller MSE.

∴ Suppose a new estimator be aX_1.

$$\therefore MSE(aX_1) = Var(aX_1) + bias(aX_1)^2$$
$$= a^2 Var(X_1) + \left(a \cdot \frac{\theta}{2} - \theta\right)^2$$
$$= a^2 \times \frac{\theta^2}{12} + \left(\frac{a-2}{2}\theta\right)^2$$
$$= \frac{a^2\theta^2}{12} + \frac{(a-2)^2\theta^2}{4}$$

$$\because MSE(aX_1) < \frac{\theta^2}{3}$$

$$\therefore \frac{a^2\theta^2}{12} + \frac{(a-2)^2\theta^2}{4} < \frac{\theta^2}{3}$$

$$a^2 + 3(a-2)^2 < 4 \quad (\text{multiply } \frac{12}{\theta^2})$$
$$a^2 + 3(a^2 - 4a + 4) < 4$$
$$4a^2 - 12a + 12 < 4$$
$$4a^2 - 12a + 8 < 0$$
$$a^2 - 3a + 2 < 0$$
$$(a-1)(a-2) < 0$$
$$\therefore 1 < a < 2$$

∴ when $a \in (1,2)$, like $a = 1.5$

∴ estimator of $1.5X_1$ will have a smaller MSE than either of the above estimators.

(c) Random size n.

(i) MME: Find the first moment $\mu_1 = E(X) = \frac{\theta}{2}$.

the sample moment $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\therefore \frac{\theta}{2} = \bar{X} \quad \therefore \theta = 2\bar{X}$$

$\therefore \hat{\theta} = 2\bar{X}$ is the MME of θ

$$Var(\bar{X}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$
$$= \frac{1}{n^2} \cdot n \cdot Var(X_i)$$
$$= \frac{Var(X_i)}{n}$$

$$E(\hat{\theta}) = E(2\bar{X}) = 2E(\bar{X}) = 2E(X_1) = 2 \times \frac{\theta}{2} = \theta$$

$$Var(\hat{\theta}) = Var(2\bar{X}) = 4Var(\bar{X}) = 4 \times \frac{Var(X)}{n} = \frac{\theta^2}{3n}$$

$$MSE(\hat{\theta}) = MSE(2\bar{X}) = Var(2\bar{X}) + bias(2\bar{X})^2$$
$$= \frac{\theta^2}{3n} + [E(2\bar{X}) - \theta] = \frac{\theta^2}{3n} + 0 = \frac{\theta^2}{3n}$$

(ii) MLE

$$L(\theta) = \prod_{i=1}^{n} f(x_i;\theta) = \prod_{i=1}^{n} \frac{1}{\theta}, \quad (0 \le x \le \theta)$$

$$= \left(\frac{1}{\theta}\right)^n$$

∵ We want to get the maximum of $L(\theta)$

∴ θ should be the minimum in its range

∴ θ is minimum when $\theta = X_n$.

∴ $\hat{\theta} = X_n$ is the MLE of θ.

(iii) $E(X_n) = \frac{n\theta}{n+1}$

$$Var(X_n) = E(X_{(n)}^2) - E(X_n)^2$$

$$= \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2$$

$$= \frac{(n+1)^2 n\theta^2 - (n+2)n^2\theta^2}{(n+2)(n+1)^2} = \frac{n\theta^2}{(n+2)(n+1)^2}$$

$$MSE(X_n) = Var(X_n) + bias(X_n)^2$$

$$= \frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{n\theta}{n+1} - \theta\right)^2$$

$$= \frac{n\theta^2}{(n+2)(n+1)^2} + \left(\frac{-\theta}{n+1}\right)^2$$

$$= \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{\theta^2}{(n+1)^2}$$

$$= \frac{n\theta^2 + (n+2)\theta^2}{(n+2)(n+1)^2} = \frac{2n\theta^2 + 2\theta^2}{(n+2)(n+1)^2}$$

$$= \frac{2\theta^2(n+1)}{(n+2)(n+1)^2} = \frac{2\theta^2}{(n+1)(n+2)}$$

(iii) $MSE = (a\theta) = Var(a\hat{\theta}) + bias(a\hat{\theta})^2$

$= a^2 Var(\hat{\theta}) + [E(a\hat{\theta}) - \theta]^2$

$= a^2 Var(X_n) + [a E(X_n) - \theta]$

$= a^2 \times \frac{n\theta^2}{(n+2)(n+1)^2} + \left[\frac{an\theta}{n+1} - (n+1)\theta\right]^2$

$= \frac{a^2 n\theta^2}{(n+2)(n+1)^2} + \frac{(an-n-1)^2\theta^2}{(n+1)^2}$

$\frac{2MSE}{2a} = 2a \cdot \frac{n\theta^2}{(n+2)(n+1)^2} + \frac{2\theta^2 n(an-n-1)}{(n+1)^2}$

$= \frac{2n\theta^2}{(n+1)^2}\left[\frac{a}{n+2} + an - n - 1\right]$

Let $\frac{2MSE}{2a} = 0$

$\therefore \frac{a}{n+2} + an - n - 1 = 0$

$a + (n+2)(an-n-1) = 0$

$a + an^2 - n^2 - n + 2an - 2n - 2 = 0$

$a(1 + n^2 + 2n) = n^2 + n + 2n + 2 = n^2 + 3n + 2$

$a = \frac{n^2+3n+2}{n^2+2n+1} = \frac{(n+2)(n+1)}{(n+1)^2} = \frac{n+2}{n+1}$

$\therefore$ When $a = \frac{n+2}{n+1}$, which minimize the MSE.

4. (a) $f(x|\mu,\lambda) = \frac{1}{x\sqrt{2\pi\lambda}} e^{-\frac{(\ln x-\mu)^2}{2\lambda}}$, $x>0$.

MLE show $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i$, $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2$

$$L(\hat{\mu},\hat{\lambda}) = \prod_{i=1}^{n} f(x|\hat{\mu},\hat{\lambda}) = \prod_{i=1}^{n} \frac{1}{x_i\sqrt{2\pi\hat{\lambda}}} e^{-\frac{(\ln x_i-\hat{\mu})^2}{2\hat{\lambda}}}$$

$$= \prod_{i=1}^{n} \frac{1}{x_i\sqrt{2\pi\hat{\lambda}}} \cdot e^{-\sum_{i=1}^{n}\frac{(\ln x_i-\hat{\mu})^2}{2\hat{\lambda}}}$$

$$\ln L(\hat{\mu},\hat{\lambda}) = \sum_{i=1}^{n} -\ln[x_i\sqrt{2\pi\lambda}] - \sum_{i=1}^{n}\frac{(\ln x_i-\mu)^2}{2\lambda} = -\sum_{i=1}^{n}\ln x_i - \frac{n}{2}\ln 2\pi\hat{\lambda} - \sum_{i=1}^{n}\frac{(\ln x_i-\hat{\mu})^2}{2\hat{\lambda}}$$

$$\frac{\partial \ln L(\hat{\mu},\hat{\lambda})}{\partial\hat{\mu}} = +\sum_{i=1}^{n}\frac{2(\ln x_i-\hat{\mu})}{2\hat{\lambda}} = 0$$

$$\therefore \sum_{i=1}^{n}(\ln x_i - \hat{\mu}) = 0$$

$$\therefore n\hat{\mu} = \sum_{i=1}^{n}\ln x_i$$

$$\hat{\mu} = \frac{\sum_{i=1}^{n}\ln x_i}{n} \quad \therefore \hat{\mu} = \frac{\sum_{i=1}^{n}\ln x_i}{n}$$

$$\frac{\partial \ln L(\hat{\mu},\hat{\lambda})}{\partial\hat{\lambda}} = -\frac{n}{2\hat{\lambda}} + \sum_{i=1}^{n}\frac{(\ln x_i-\hat{\mu})^2}{2\hat{\lambda}^2} = 0$$

$$\frac{n}{2\hat{\lambda}} = \sum_{i=1}^{n}\frac{(\ln x_i-\hat{\mu})^2}{2\hat{\lambda}^2}$$

$$\therefore \hat{\lambda} = \frac{\sum_{i=1}^{n}(\ln x_i-\hat{\mu})^2}{n}$$

$$n\lambda = \sum_{i=1}^{n}(\ln x_i-\mu)^2$$

$$\lambda = \frac{\sum_{i=1}^{n}(\ln x_i-\mu)^2}{n}$$

(b) $X_i \sim N(\mu,\lambda)$, 100% (1-α)% CI for λ.

Normal distribution. We know that λ (variance) follows a χ^2 distribution.

$$\frac{(n-1)S^2}{\lambda} \sim \chi^2_{n-1}, \quad (S^2 \text{ is sample variance})$$

$$\therefore \chi^2_{1-\frac{\alpha}{2}} \le \frac{(n-1)S^2}{\lambda} \le \chi^2_{\frac{\alpha}{2}}$$

$$\therefore \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \le \lambda \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}$$

$\therefore$ 100% (1-α)% CI for λ is

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right], \text{ where } a, b \text{ is the } \frac{\alpha}{2} \text{ and } (1-\frac{\alpha}{2}) \text{ quantiles of } \chi^2_{n-1}$$

4.(c)

(i)

R script:

```
#number 4
#c
x <- c(0.27,3.30,4.58,2.61,0.38,3.77,1.11,1.15,4.11,2.10,
       0.07,1.74,2.11,12.79,1.85,0.30,0.34,1.31,0.14,0.74)

n <- length(x)
mu.hat <- mean(log(x))
lambda.hat <- mean((log(x)-mu.hat)^2)
lambda.hat
s <- sd(log(x))

a <- qchisq(0.025,n-1)
b <- qchisq(0.975,n-1)

CI <- c((n-1)* s^2 / b, (n-1)* s^2 / a)
CI
```

Output:

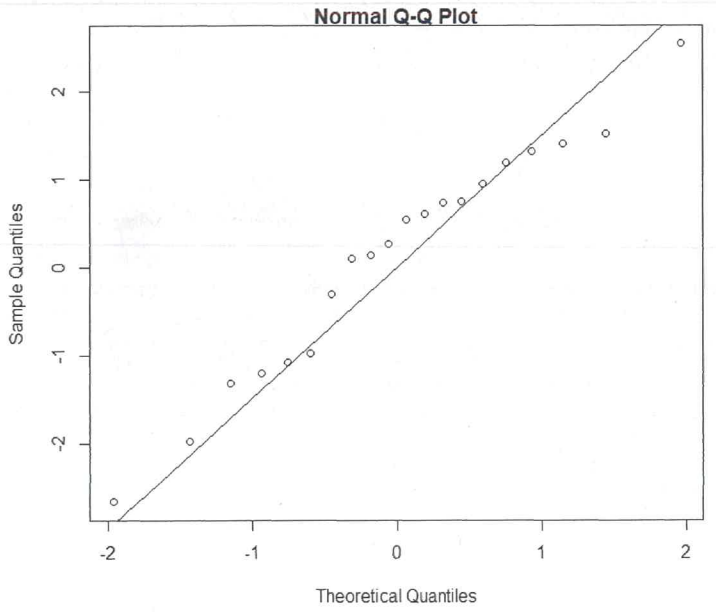
```
> CI <- c((n-1)* s^2 / b, (n-1)* s^2 / a)
> CI
[1] 0.9969874 3.6774596
```

(ii)

R script:

```
# ii
qqnorm(log(x))
qqline(log(x))
```

Output:



We can find that the model is roughly appropriate for these data.

15. X_1, X_2, X_3, X_4 iid rvs with $E(X_i)=\mu$, $\text{Var}(X_i)=\sigma^2>0$, $i=1,2,3,4$

$\therefore T_1 = \frac{1}{3}(X_1+X_2)+\frac{1}{6}(X_3+X_4)$ $T_2 = \frac{1}{6}(X_1+2X_2+3X_3+4X_4)$

$T_3 = \frac{1}{4}(X_1+X_2+X_3+X_4)$ $T_4 = \frac{1}{3}(X_1+X_2+X_3)+\frac{1}{4}X_4^2$

a) For T_1

$$E(T_1) = E\left(\frac{1}{3}(X_1+X_2)+\frac{1}{6}(X_3+X_4)\right) = \frac{1}{3}(E(X_1)+E(X_2))+\frac{1}{6}(E(X_3)+E(X_4))$$
$$= \frac{1}{3}(\mu+\mu)+\frac{1}{6}(\mu+\mu)$$
$$= \frac{2}{3}\mu+\frac{1}{3}\mu = \mu = E(X_i)$$

$\therefore E(T_1)=\mu$, which is unbiased.

For T_2 $E(T_2) = E\left(\frac{1}{6}(X_1+2X_2+3X_3+4X_4)\right) = \frac{1}{6}(E(X_1)+2E(X_2)+3E(X_3)+4E(X_4))$

$$= \frac{1}{6}(\mu+2\mu+3\mu+4\mu)$$
$$= \frac{5}{3}\mu \neq \mu$$

$E(T_2)=\frac{5}{3}\mu$, which is biased.

For T_3 $E(T_3) = E\left(\frac{1}{4}(X_1+X_2+X_3+X_4)\right) = \frac{1}{4}(E(X_1)+E(X_2)+E(X_3)+E(X_4))$

$$= \frac{1}{4}\times(\mu+\mu+\mu+\mu) = \mu = \mu$$

$E(T_3)=\mu$, which is unbiased

For T_4 $E(T_4) = E\left(\frac{1}{3}(X_1+X_2+X_3)+\frac{1}{4}X_4^2\right) = \frac{1}{3}(E(X_1)+E(X_2)+E(X_3))$

$$+\frac{1}{4}(\sigma^2+\mu^2)$$
$$= \frac{1}{3}(\mu+\mu+\mu)+\frac{1}{4}(\sigma^2+\mu^2) \neq \mu$$

$\therefore T_4$ is biased estimator.

$\text{Var}(X) = E(X^2)-E(X)$

$\therefore E(X^2) = \text{Var}(X)+E(X)$

$= \sigma^2+\mu^2$

(b) T_1, T_3 is unbiased estimators

$\because X_1, X_2, X_3, X_4$ is iid rvs $\therefore \text{Var}(X_1+X_2) = \text{Var}(X_1)+\text{Var}(X_2)$

For T_1 $\text{Var}(T_1) = \text{Var}\left(\frac{1}{3}(X_1+X_2)+\frac{1}{6}(X_3+X_4)\right)$

$$= \frac{1}{9}(\text{Var}(X_1)+\text{Var}(X_2))+\frac{1}{36}(\text{Var}(X_3)+\text{Var}(X_4))$$
$$= \frac{1}{9}(\sigma^2+\sigma^2)+\frac{1}{36}(\sigma^2+\sigma^2) = \frac{(8+2)\sigma^2}{36} = \frac{5}{18}\sigma^2$$

For T_3 $\text{Var}(T_3) = \text{Var}\left(\frac{1}{4}(X_1+X_1+X_3+X_4)\right) = \frac{1}{16}(\text{Var}(X_1)+\text{Var}(X_2)+\text{Var}(X_3)$

$$= \frac{1}{16}(4\times\sigma^2) = \frac{1}{4}\sigma^2 < \frac{5}{18}\sigma^2 \quad +\text{Var}(X_4))$$

$\therefore \text{Var}(T_1) > \text{Var}(T_3)$

$\therefore T_3$ has the smallest variance

MAST20005/MAST90058: Assignment 2

Due date: 11am, Friday 20 September 2019

Instructions: Questions labelled with '**(R)**' require use of R. Please provide appropriate R commands and their output, along with sufficient explanation and interpretation of the output to demonstrate your understanding. **Such R output should be presented in an integrated form together with your explanations; do not attach them as separate sheets.** All other questions should be completed without reference to any R commands or output, except for looking up quantiles of distributions where necessary. Make sure you give enough explanation so your tutor can follow your reasoning if you happen to make a mistake. Please also try to be as succinct as possible. Each assignment will include marks for good presentation and for attempting all problems.

Problems:

1. Suppose that you want to know how long (in hours) it takes for a particular brand of paint to dry. Nine experiments are done and the times were measured as follows:

   6.0 5.7 5.8 6.5 7.0 6.3 5.6 6.1 5.0

   Assume these times follow a normal distribution, $N(\mu, \sigma^2)$.

   (a) Assuming $\sigma = 0.6$ based on previous experience, calculate a 95% CI for μ.

   (b) Still assuming $\sigma = 0.6$, suppose we want our estimate of μ to be within 0.2 with probability near 95%. How many experiments do we need to run?

   (c) **(R)** If σ is unknown, calculate a 95% CI for μ. Compare the width of CIs in part (a) and (c).

2. An assembly line has a target of achieving an 80% success rate when making bicycles. Long experience shows that they are never more than 10% away from that target. What sample size is required for estimating the success rate:

   (a) To within 5% with probability 0.95?

   (b) To within 2% with probability 0.95?

3. **(R)** The `pressure` dataset is available in a standard installation of R. You should be able to access it directly, for example via:

   ```
   > pres <- pressure$pressure
   > temp <- pressure$temperature
   ```

   The dataset describes the relationship between temperature in degrees Celsius and the vapor pressure of mercury in millimeters (of mercury). According to the Antoine Equation, they have the relationship:

   $$\log_{10}(\text{pressure}) = \alpha + \frac{\beta}{\text{temperature} - 10},$$

   for some constants a and b, and where temperature in the formula is measured on the Kelvin scale (K). The relationship between temperature in Kelvin (T_K) and temperature in Celsius (T_C) is $T_K = T_C + 273.15$.

(a) Suppose we want to estimate the constants by fitting the simple linear regression model,

$$y_i = \alpha + \beta x_i + \varepsilon_i.$$

Define y and x appropriately for this model to work.

(b) Fit the above model and find estimates of α and β.

(c) Assess the model fit visually using some standard diagnostic plots. Does the linear model seem appropriate?

(d) Give 95% confidence intervals for the regression coefficients. According to other sources, $\alpha = 4.86$ and $\beta = -3007$. Does your model support these two claims?

(e) Give a 95% confidence interval for the mean vapor pressure when the temperature is 70 degrees Celsius.

(f) Give a 95% prediction interval for the vapor pressure when tempeature is 70 degrees Celsius.

4. Two different types of high-speed train are manufactured. We are interested in whether there is a difference in their maximum speed. The following table summarises the maximum train speed in kilometers per hour, measured over a series of tests.

Train type	Sample size	Mean	Standard deviation
Type A	10	500	1.1
Type B	20	496	1.2

Assuming the maximum speed of both types of train follows a normal distribution, determine whether a type A train is faster than a type B train when they are running at maximum speed.

5. It was claimed that 80% of users on a particular website are male. In a random sample of 200 users, 146 of them were male. Is there evidence that the proportion p of users that are male differs from 0.8?

(a) State appropriate null and alternate hypotheses.

(b) What would you conclude if the significance level is $\alpha = 0.05$?

(c) What would you conclude if the significance level is $\alpha = 0.01$?

(d) Give a 95% confidence interval for the proportion of users that are male.

6. **(R)** Consider a Poisson random variable X with pmf

$$\Pr(X = x \mid \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

A single observation of such a variable is used to test $H_0\colon \lambda = 2$ against $H_1\colon \lambda > 2$. The null hypothesis is rejected if the observed value is greater than or equal to 4.

(a) What is the probability of committing a Type I error?

(b) What is the probability of committing a Type II error when $\lambda = 5$ in H_1?

(c) Draw a power curve for this test for alternative values of λ between 2 and 10.

(d) Find a test of these hypotheses that has an approximate significance level of 0.05. What is the *actual* significance level of your test?

Assignment 2

Mu Tong
1004452

15

1. ~~(a)~~ $n=9$. Assume Normal distribution $N(\mu, \sigma^2)$

(a) Assume $\sigma = 0.6$, 95% CI for μ.

Sample mean: $\bar{x} = \frac{6.0+5.7+5.8+6.5+7.0+6.3+5.6+6.1+5.0}{9} = \frac{54}{9} = 6$

Sample ~~std~~ Variance $\text{Var}(x) = \sum_{i=1}^{9}(x_i - \bar{x})^2 \cdot \frac{1}{n-1} = 0.33$

Sample Std $= \sqrt{\text{Var}(x)} = \sqrt{0.33} \approx 0.5745$

95% CI for μ is: $c = \Phi^{-1}(1-\frac{\alpha}{2}) = \Phi^{-1}(1-\frac{0.05}{2}) = \Phi^{-1}(0.975) = 1.96$

$$\bar{x} \pm c\cdot\frac{\sigma}{\sqrt{n}} = 6 \pm 1.96 \times \frac{0.6}{\sqrt{9}} = 6 \pm 1.96 \times 0.2 = 6 \pm 0.392$$

$\therefore$ 95% CI for μ is $(5.608, 6.392)$

(b) $n = \left(\frac{c\sigma}{\epsilon}\right)^2$ $c = 1.96$ (95%)

$\epsilon = 0.2$

$= \left(\frac{1.96 \times 0.6}{0.2}\right)^2$

$= 34.5744 \approx 35$

$\therefore$ 35 experiments need to run

(c) if σ is unknown

95% CI for μ $\bar{x} \pm c\cdot\frac{s}{\sqrt{n}} = (5.558434, 6.441566)$ from R

Comparing with part a, CI is $(5.608, 6.392)$,

95% CI for μ in part (a) is narrower.

Question 1

Part 3

R script:

```
# Question 1
x <- c(6,5.7,5.8,6.5,7.0,6.3,5.6,6.1,5)

var(x)
sd(x)

t.test(x,conf.level = 0.95)
```

R output:

```
> var(x)
[1] 0.33
> sd(x)
[1] 0.5744563
>
> t.test(x,conf.level = 0.95)

        One Sample t-test

data:  x
t = 31.334, df = 8, p-value = 1.171e-09
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.558434 6.441566
sample estimates:
mean of x
        6
```

2. (a) Sample size for proportions

$$n = \frac{c^2 \hat{p}(1-\hat{p})}{\epsilon^2}$$

$c = \Phi^{-1}(1-\frac{\alpha}{2}) = \Phi^{-1}(0.975) = 1.96$

$\hat{p} = 0.8 \quad \epsilon = 0.05$

$$= \frac{1.96^2}{0.05^2} \times 0.8 \times 0.2 = 245.86 \approx 246$$

∴ require 246

(b) $$n = \frac{c^2 \hat{p}(1-\hat{p})}{\epsilon^2}$$

$c = \Phi^{-1}(0.975) = 1.96$

$\hat{p} = 0.8$

$\epsilon = 0.02$

$$= \frac{1.96^2}{0.02^2} \times 0.8 \times 0.2 = 1536.64$$

$$\approx 1537$$

∴ required sample size is 1537.

Question 3

a) R script:

3/8

```
#a) define x and y
y <- log(pres)
x <- 1 / (temp - 10)
```

0/1

b) R script

```
#b) fit the model
fit <- lm(y ~ x)
summary(fit)
```

R output:

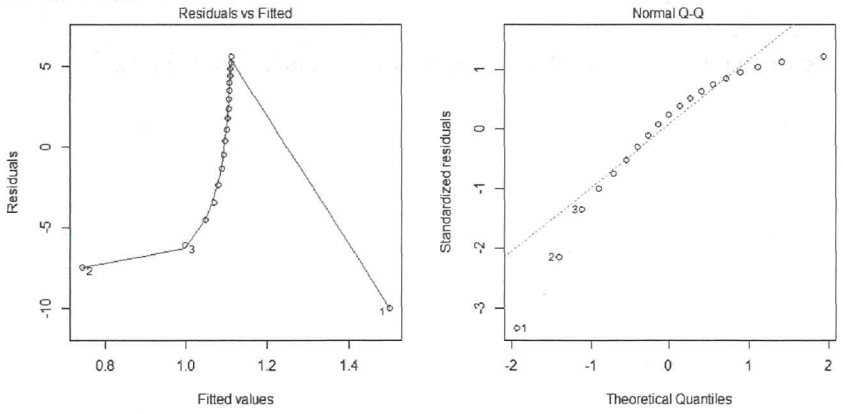
```
> fit <- lm(y ~ x)
> summary(fit)

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-10.017  -2.933   1.074   3.702   5.580

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.123      1.114   1.008    0.327
x             -3.769     32.577  -0.116    0.909

Residual standard error: 4.735 on 17 degrees of freedom
Multiple R-squared:  0.0007869, Adjusted R-squared:  -0.05799
F-statistic: 0.01339 on 1 and 17 DF,  p-value: 0.9092
```

We can get the estimates of alpha is 1.123, and the estimates of beta is -3.769.

c) R script:

```
#c) plot diagnostic plots
par(mfrow = c(1, 2))
plot(fit, 1:2)
```

R output:



The linear model looks approximately appropriate from the QQ plot, because most of points are on the straight line except some points.

d) R script:

```
#d) 95% CI
confint(fit)
```

R output:

```
> confint(fit)
                  2.5 %    97.5 %
(Intercept)  -1.226584  3.471995
x           -72.500778 64.961925
```

We can find that 4.86 is not in the 95% CI for alpha (-1,22,3.47), and -3007 is not in the 95% CI for beta (-72.50,64.96); therefore, this model doesn't support these two claims.

e) R script:

```
#e) CI
newdata = data.frame(x = 1 / (70 - 10))
predict(fit,newdata,interval = "confidence", level = 0.95)
```

R output:

```
> newdata = data.frame(x = 1 / (70 - 10))
>
> predict(fit,newdata,interval = "confidence", level = 0.95)
       fit       lwr      upr
1 1.059882 -1.316822 3.436586
```

We can find 95% CI for pressure when temperature is 70 degrees Celsius is (-1.32,3.43).

f) R script:

```
#f) PI
predict(fit,newdata,interval = "predict", level = 0.95)
```

R output:

```
> predict(fit,newdata,interval = "predict", level = 0.95)
       fit       lwr      upr
1 1.059882 -9.208992 11.32876
```

We can find 95% PI for pressure when temperature is 70 degrees Celsius is (-9.21,11.33).

4.

Train type	Sample size	Mean	Standard deviation
Type A	10	500	1.1
Type B	20	496	1.2

Type A: $n_1 = 10$ $\mu_1 = 500$ $s_1 = 1.1$

Type B: $n_2 = 20$ $\mu_2 = 496$ $s_2 = 1.2$

Assuming both follows Normal distribution, and $\alpha = 0.05$ and assumly they have common variance.

$H_0: \mu_1 - \mu_2 = 0$ $H_1: \mu_1 - \mu_2 > 0$

We use t-distribution with $df = 28$, since $(df = 10+20-2 = 28)$

$\because \alpha = 0.05$

$\therefore \Phi^{-1}(0.95) = 1.701131$ from R script qt(0.95, 28)

Pooled variance estimate:

$$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\frac{(10-1)1.1^2 + (20-1)\times 1.2^2}{10+20-2}}$$

$$\approx 1.16879$$

We use two-sample t-test but assumly H_0

$$T = \frac{\bar{X} - \bar{Y}}{s_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \sim t_{n+m-2}$$

$$\therefore t = \frac{\mu_1 - \mu_2}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} = \frac{500-496}{1.16879\times\sqrt{\frac{1}{10}+\frac{1}{20}}} \approx 8.83645 > 1.6879$$

$\therefore t > \Phi^{-1}(0.95)$

Critical region: We reject H_0 if $t > \Phi^{-1}(0.95)$

$\therefore$ We ~~fail~~ reject H_0, there is sufficient evidence that Type A train is faster than train B at the 5% level of significance.

5.

$N = 200 \quad X = 146 \quad \hat{P} = \frac{X}{N} = \frac{146}{200} = 0.73$

(a) $H_0: P = 0.8$

$H_1: P \neq 0.8$

(b) $\alpha = 0.05 \quad \therefore \Phi^{-1}(1-\frac{\alpha}{2}) = \Phi^{-1}(0.975) = 1.96$

$\therefore$ critical region: $\{z : |z| > 1.96\}$

$\therefore$ We reject H_0 if the value for $|z|$ is greater than 1.96.

$$z = \frac{\hat{P} - P_0}{\sqrt{P_0(1-P_0)/n}} = \frac{0.73 - 0.80}{\sqrt{0.80 \times 0.2/200}} \approx -2.47$$

$|z| = |-2.47| = 2.47 > 1.96$

$\therefore$ We reject H_0 if $\alpha = 0.05$

$\therefore$ the proportion p of users that are male differs.

(c) $\alpha = 0.01 \quad \Phi^{-1}(1-\frac{\alpha}{2}) = \Phi^{-1}(0.995) = 2.5758$ from R(qnorm(0.995))

$\therefore$ critical region: $\{z : |z| > 2.5758\}$

$$z = \frac{\hat{P} - P_0}{\sqrt{P_0(1-P_0)/n}} = -2.47$$

$|z| = |-2.47| = 2.47 < 2.5758$

$\therefore z < \Phi^{-1}(0.995)$

$\therefore$ We do not reject H_0

$\therefore$ the proportion p of users that are male doesn't differ.

(d) 95% CI is $\hat{P} = 0.73$

$$\hat{p} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \qquad c = \Phi^{-1}(0.975) = 1.96$$

$$= 0.73 \pm 1.96 \times \sqrt{\frac{0.73 \times 0.27}{200}}$$

$\therefore$ 95% CI is $(0.6685, 0.7915)$

Question 6

a) R:

```
# a) Type I error
ppois(3,2,lower.tail = FALSE)
```

R output:

```
> ppois(3,2,lower.tail = FALSE)
[1] 0.1428765
```

b) R:

```
# b) Type II error when lambda = 5
ppois(3,5,lower.tail = TRUE)
```

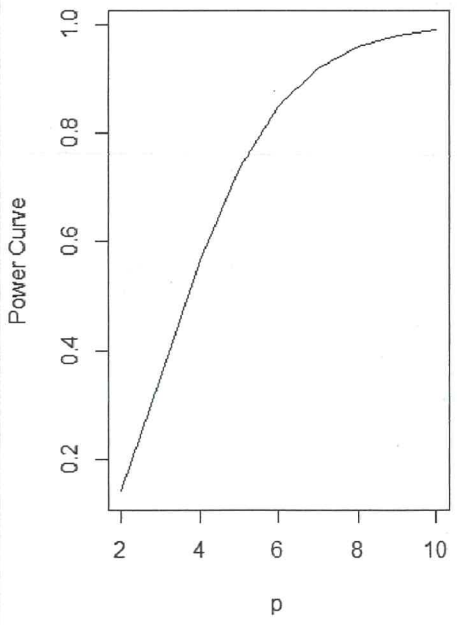
R output:

```
> ppois(3,5,lower.tail = TRUE)
[1] 0.2650259
```

c) R:

```
# c) Draw power curve of lambda between 2 to 10

K1 <- function(p)
  1 - ppois(3,p)

p <- seq(2,10,0.1)
K <- K1(p)
plot(p,K,type = "l", ylab = "Power Curve")
```

R output:



d) R:

```
K1(3)
K1(4)
K1(2)
K1(1)
K1(1.5)
K1(1.6)
K1(1.4)
K1(1.3)
K1(1.35)
K1(1.36)
K1(1.37)
```

R output:

```
> K1(3)
[1] 0.3527681
> K1(4)
[1] 0.5665299
> K1(2)
[1] 0.1428765
> K1(1)
[1] 0.01898816
> K1(1.5)
[1] 0.06564245
> K1(1.6)
[1] 0.07881349
> K1(1.4)
[1] 0.05372525
> K1(1.3)
[1] 0.04309545
> K1(1.35)
[1] 0.04824799
> K1(1.36)
[1] 0.04931753
> K1(1.37)
[1] 0.05040005
```

The actual significance level is 0.05040005 when lambda is 1.37

MAST20005/MAST90058: Assignment 3

Due date: 11am, Friday 18 October 2019

Instructions: Questions labelled with '(R)' require use of R. Please provide appropriate R commands and their output, along with sufficient explanation and interpretation of the output to demonstrate your understanding. **Such R output should be presented in an integrated form together with your explanations; do not attach them as separate sheets.** All other questions should be completed without reference to any R commands or output, except for looking up quantiles of distributions where necessary. Make sure you give enough explanation so your tutor can follow your reasoning if you happen to make a mistake. Please also try to be as succinct as possible. Each assignment will include marks for good presentation and for attempting all problems.

Problems:

1. **(R)** A study measured the weight gain after 8 weeks of two groups of mice. The first group was fed a high-protein diet, and the second group a low-protein diet. The data are given below:

High-protein:	134	146	104	119	124	161	112	83	113	129	97	123
Low-protein:	70	118	101	85	107	132	94					

   (a) Use the sign test with $\alpha = 0.05$ to test if the median weight gain in the first group is smaller than 110. Clearly state your hypotheses.

   (b) Use the Wilcoxon rank-sum test with $\alpha = 0.05$ to test if the median weight gain in the first group is larger than that in the second group. Clearly state your hypotheses.

   (c) Use a t-test with $\alpha = 0.05$ to test if the mean weight gain in the first group is larger than that in the second group. Clearly state your hypotheses.

2. **(R)** A survey was conducted that asked people of different ages how much they exercised (in hours per week). The responses received were as follows:

	0 hours	1 hour	2 hours	3 hours	4+ hours
Younger than 40 years	10	24	10	6	3
40 years or older	7	22	18	10	5

   Using a significance level of $\alpha = 0.05$, test whether:

   (a) The hours of exercise for the younger age group follows a Poisson distribution.

   (b) Age is independent of hours of exercise.

3. Let X have a shifted exponential distribution with pdf,

$$f(x) = e^{-(x-\theta)}, \quad x \geqslant \theta.$$

Suppose we have a random sample of n observations on X.

(a) Find the cdf of the sample minimum, $X_{(1)}$.

(b) Find the p quantile, π_p.

(c) Find the asymptotic variance of the sample median, $\hat{M}$.

4. **(MAST20005 students only)** Consider the one-way analysis of variance model,

$$X_{ij} = \alpha_i + \epsilon_{ij}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, n,$$

where $\epsilon_{ij} \sim \mathrm{N}(0, \sigma_j^2)$ are independent but not identically distributed (σ_j^2 varies by j). Let,

$$\bar{X}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} X_{ij}.$$

(a) Find the distribution of $\bar{X}_{i\cdot}$.

(b) Find $\mathbb{E}\left\{\sum_{j=1}^{n} (X_{ij} - \bar{X}_{i\cdot})^2\right\}$.

5. **(MAST90058 students only) (R)** Retail sales are affected by store locations and the number of competitors nearby. In a study, three different type of locations are considered: outer surburb, inner surburb, and CBD. For each type of location and number of competitors, retail sales of three stores were reported (in thousands of dollars). The data obtained were:

	Number of competitors			
Locations	3	2	1	0
Outer surburb	270	290	446	440
	310	350	487	428
	220	305	500	530
Inner surburb	410	382	598	470
	305	320	480	415
	450	380	510	400
CBD	180	220	290	246
	290	170	283	275
	330	260	260	330

Perform a two-way analysis of variance to examine whether these data suggest that retail sales are affected by store locations. State and test appropriate hypotheses at a 5% significance level. You should report the value of the appropriate statistic, the p-value, the assumptions you have made and your conclusions. Is it possible to test for interaction? If yes, then perform the test and draw an interaction plot; otherwise, explain why it is not possible.

19/20

Assignment3 for MAST90058

Name: Mu Tong
Student Number: 1004452

Question 1:

(a) R script:

```
high_protein <- c(134,146,104,119,124,161,112,83,113,129,97,123)
low_protein <- c(80,118,101,85,107,132,94)

## part a: sigh test with alpha = 0.05

binom.test(sum(high_protein > 110),length(high_protein),alternative = "less")
```

Output:

```
        Exact binomial test

data:  sum(high_protein > 110) and length(high_protein)
number of successes = 9, number of trials = 12, p-value = 0.9807
alternative hypothesis: true probability of success is less than 0.5
95 percent confidence interval:
 0.0000000 0.9281297
sample estimates:
probability of success
                  0.75
```

Null hypothesis: the median weight gain in the first group is 110 (m = 110)

Alternative hypothesis: the median weight gain in the first group is smaller than 110 (m < 110)

We cannot reject null hypothesis because p-value (0.98) is bigger than 0.05.

(b) R script:

```
## part b: Wilcoxon rank-sum test with alpha = 0.05

wilcox.test(high_protein,low_protein,alternative = "greater")
```

R output:

```
        Wilcoxon rank sum test

data:  high_protein and low_protein
W = 63, p-value = 0.04156
alternative hypothesis: true location shift is greater than 0
```

Null hypothesis: the median weight gain in the first group is the same as that in the second group.

Alternative hypothesis: the median weight gain in the first group is larger than that in the second group.

We can reject null hypothesis because the p-value (0.04156) is smaller than 0.05.

(c) R script:

```
## part c: t test with alpha = 0.05

t.test(high_protein,low_protein,var.equal = TRUE,alternative = "greater")
```

Output:

```
> t.test(high_protein,low_protein,var.equal = TRUE,alternative = "greater")

        Two Sample t-test

data:  high_protein and low_protein
t = 1.8716, df = 17, p-value = 0.03929
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 1.268498      Inf
sample estimates:
mean of x mean of y
 120.4167  102.4286
```

Null hypothesis: the mean weight gain in the first group is the same as that in the second group.

Alternative hypothesis: the mean weight gain in the first group is larger than that in the second group.

We can reject null hypothesis because the p-value (0.039) is smaller than 0.05.

To check the assumption of the same variance.

```
> var.test(high_protein,low_protein)

        F test to compare two variances

data:  high_protein and low_protein
F = 1.3314, num df = 11, denom df = 6, p-value = 0.7569
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.2461046 5.1665708
sample estimates:
ratio of variances
          1.331367
```

We can get the high p-value is 0.7569, so we cannot reject null hypothesis of the same variance.

Question 2

(a) R script:

```
# (a)
x <- rep(0:4, c(10,24,10,6,3))
table(x)
x.bar <- mean(x)
x.bar

p0 <- dpois(0,x.bar)
p1 <- dpois(1,x.bar)
p2 <- dpois(2,x.bar)
p3 <- 1 - (p0 + p1 +p2)

p <- c(p0,p1,p2,p3)
y <- c(10,24,10,9)

chisq.test(y,p = p)
# X-squared = 3.118

1 - pchisq(3.118,2)
```

3/3

Output:

```
> chisq.test(y,p = p)

        Chi-squared test for given probabilities

data:  y
X-squared = 3.118, df = 3, p-value = 0.3738

> # X-squared = 3.118
>
> 1 - pchisq(3.118,2)
[1] 0.2103463
```

Null hypothesis: the hours of exercise for the younger age group follows a Poisson distribution.
Alternative hypothesis: the hours of exercise for the younger age group doesn't follow a Poisson distribution.

We used goodness of fit test to test, and we find the p-value is greater than 0.05, therefore, we fail to reject null hypothesis, which means that the data follows a Poisson distribution.

(b) R script:

4/4

```
# (b)

x <- rbind(younger = c(10,24,10,9),
           older = c(7,22,18,15))
x

chisq.test(x)
```

Output:

```
> x
        [,1] [,2] [,3] [,4]
younger   10   24   10    9
older      7   22   18   15
> chisq.test(x)

        Pearson's Chi-squared test

data:  x
X-squared = 3.7205, df = 3, p-value = 0.2933
```

Null hypothesis: age is independent of hours of exercise.

Alternative hypothesis: age is not independent of hours of exercise.

We use contingency tables to test the independence, we find the p-value is 0.2933 which is much greater than 0.05, therefore, we fail to reject null hypothesis at 5% significance level, which means that age is independent of hours of exercise.

3. pdf : $f(x) = e^{-(x-\theta)}$, $x \geq \theta$

(a) Find cdf of sample minimum, $X_{(1)}$

$$F_1(x) = P_r(X_{(1)} \leq x) = 1 - P_r(X_{(1)} > x)$$
$$= 1 - (1 - F(x))^n$$

$$\therefore F(x) = \int_\theta^x e^{-(x-\theta)} dx = \int_\theta^x e^{-x} \cdot e^{\theta} dx = e^{\theta} \int_\theta^x e^{-x}$$
$$= e^{\theta} \cdot (-e^{-x})\Big|_\theta^x = e^{\theta} \cdot (-e^{-x} + e^{-\theta})$$
$$= 1 - e^{-x+\theta} = 1 - e^{\theta - x}$$

$$\therefore F_1(x) = 1 - (1 - F(x))^n$$
$$= 1 - (1 - (1 - e^{\theta - x}))^n$$
$$= 1 - (1 - 1 + e^{\theta - x})^n = \underline{1 - e^{n(\theta - x)}}, \quad x \geq \theta$$

(b) $\therefore p = F(\pi_p) = P_r(X \leq \pi_p)$

$$\therefore 1 - e^{\theta - \pi_p} = p$$

$$\therefore e^{\theta - \pi_p} = 1 - p$$

$$\ln(e^{\theta - \pi_p}) = \ln(1 - p)$$

$$\theta - \pi_p = \ln(1 - p)$$

$$\therefore \pi_p = \underline{\theta - \ln(1 - p)}$$

(c) Find asymptotic variance of sample median, $\hat{m}$.

We already now that

$$\hat{M} \approx N\left(m, \frac{1}{4n f(m)^2}\right)$$

$\therefore$ we need to get the population median m.

$$\int_{\theta}^{m} f(x)\,dx = \frac{1}{2}$$

$$\therefore 1 - e^{\theta - m} = \frac{1}{2}$$

$$\therefore e^{\theta - m} = \frac{1}{2}$$

$$\Rightarrow \ln(e^{\theta - m}) = \ln\frac{1}{2}$$

$$\therefore \theta - m = \ln\frac{1}{2}$$

$$\therefore m = \theta - \ln\frac{1}{2} = \theta + \ln 2$$

$$\therefore f(m)^2 = (e^{-(\theta + \ln 2 - \theta)})^2 = (e^{-\ln 2})^2 = e^{\ln(\frac{1}{2})^2} = (\frac{1}{2})^2 = \frac{1}{4}$$

$$\therefore \hat{M} \approx N\left(m, \frac{1}{4n f(m)^2}\right) = N\left(m, \frac{1}{4n \times \frac{1}{4}}\right) = N\left(m, \frac{1}{n}\right)$$

$\therefore$ the variance of $\hat{M} = \frac{1}{n}$.

6/6

Question 5:

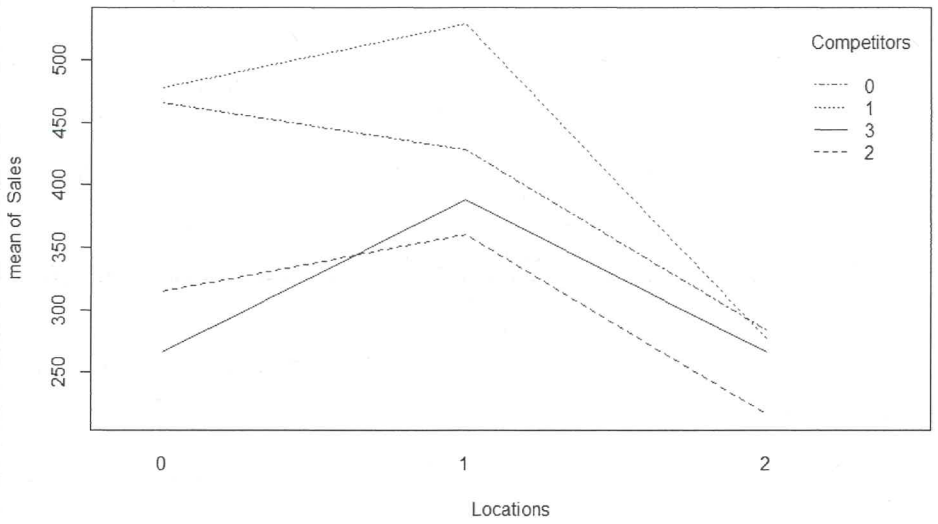
R script:

```
## question 3

##create the table

## 0 in Location column means outer surburb
## 1 in Location column means inner surburb
## 2 in Location column means CBD

retail <- data.frame("Competitors" = c(3,3,3,2,2,2,1,1,1,0,0,0,
                                       3,3,3,2,2,2,1,1,1,0,0,0,
                                       3,3,3,2,2,2,1,1,1,0,0,0),
                     "Locations" = c(0,0,0,0,0,0,0,0,0,0,0,0,
                                     1,1,1,1,1,1,1,1,1,1,1,1,
                                     2,2,2,2,2,2,2,2,2,2,2,2),
                     "Sales" = c(270,310,220,290,350,305,
                                 446,487,500,440,428,530,
                                 410,305,450,382,320,380,
                                 598,480,510,470,415,400,
                                 180,290,330,220,170,260,
                                 290,283,260,246,275,330))
retail
model <- lm(Sales ~ factor(Competitors) + factor(Locations),retail)
anova(model)

model2 <- lm(Sales ~ factor(Competitors) * factor(Locations),retail)
anova(model2)

with(retail,interaction.plot(Locations,Competitors,Sales,col = "blue"))
```

5/6

Output:

```
> anova(model)
Analysis of Variance Table

Response: Sales
                   Df Sum Sq Mean Sq F value    Pr(>F)
factor(Competitors)  3 111311   37104  10.483 7.032e-05 ***
factor(Locations)    2 175542   87771  24.799 4.399e-07 ***
Residuals           30 106180    3539
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> model2 <- lm(Sales ~ factor(Competitors) * factor(Locations),retail)
> anova(model2)
Analysis of Variance Table

Response: Sales
                                   Df Sum Sq Mean Sq F value    Pr(>F)
factor(Competitors)                 3 111311   37104 15.3462 8.732e-06 ***
factor(Locations)                   2 175542   87771 36.3023 5.528e-08 ***
factor(Competitors):factor(Locations)  6  48153    8026  3.3194   0.01596 *
Residuals                          24  58027    2418
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
> with(retail,interaction.plot(Locations,Competitors,Sales,col = "blue"))
```



In the first model, we make some assumptions. The first one is that they both follow normal distribution, and the second one is that factors are independent, and the last one is that all populations have the same variance.

we find that the F-values for number of competitors and locations factors are 10.483 and 24.799 respectively, and the p-value for these two factors are 0.00007032 and 0.0000004399 respectively, which are much less than 0.05. Therefore, we get the conclusion about there is a clear difference in retail sales between the number of competitors (We reject null hypothesis of factor competitor) and also between the locations (we reject null hypothesis of factor location).

It is possible to test for interaction, and the interaction plot is showed above. From the interaction plot, we can find that the interaction p-value is 0.01596 which is smaller than 0.05, therefore, we can say that there is no interaction between these two factors.