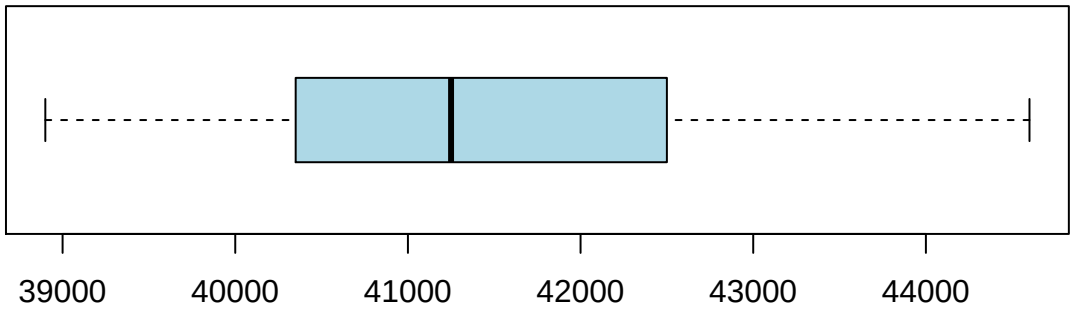
MAST20005/MAST90058: Assignment 1 Solutions

1. (a)
```
x <- c(41300, 40300, 43200, 41100, 39300, 42100, 42700, 41300,
       38900, 41200, 44600, 42300, 40700, 43500, 39800, 40400)
summary(x)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   38900   40375   41250   41419   42400   44600

sd(x)

## [1] 1569.382
```

The above provides the standard five-number summary, sample mean and sample standard deviation.

```
par(mar = c(3, 1, 1, 1))  # compact margins
boxplot(x, horizontal = TRUE, col = "lightblue")
```



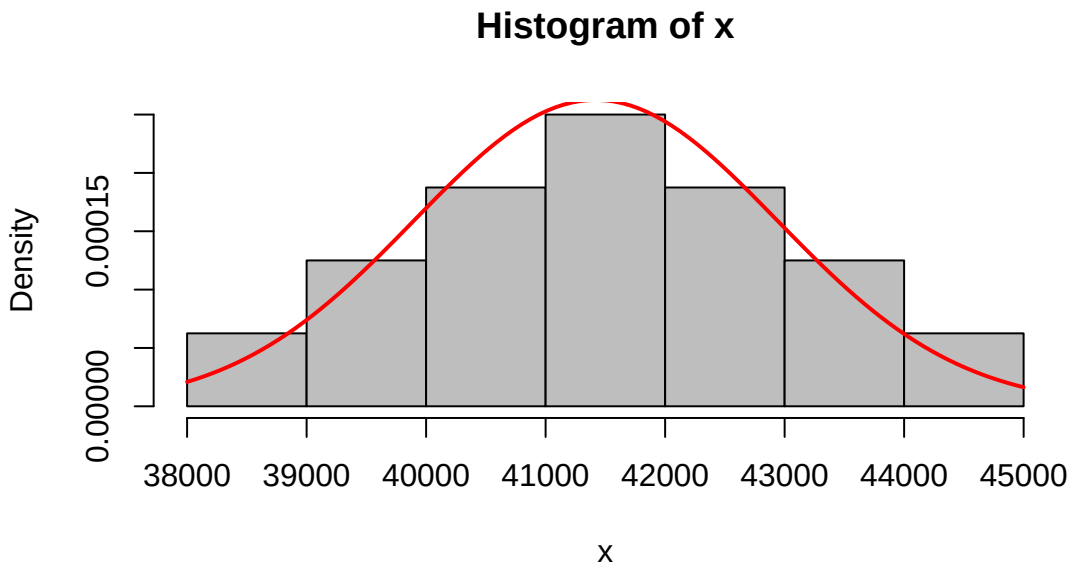
The distribution of tire distances is centred around a median value of about 41,500 km and is somewhat variable, with a sample standard deviation of around 1,500 km. The distribution seems to be mostly symmetric (with the right tail being slightly heavier).

(b) Using pdf: $f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$.

```
library(MASS)
normfit <- fitdistr(x, densfun = "normal")
normfit

##       mean          sd
##   41418.7500    1519.5471
##  (  379.8868) (  268.6205)
```

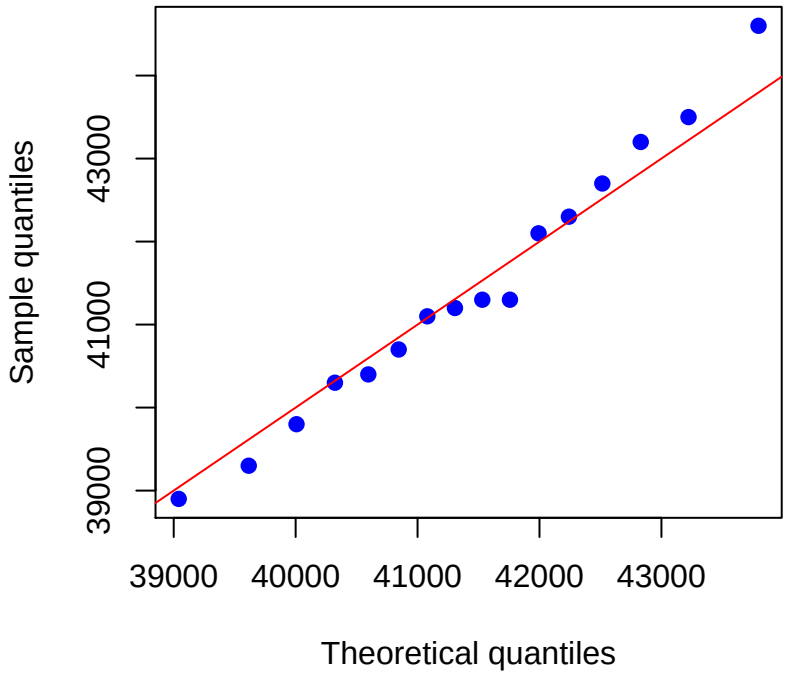
This gives $\hat{\mu} = 41400$ and $\hat{\sigma} = 1520$.

(c)
```
hist(x, freq = FALSE, col = "grey")
curve(dnorm(x, mean = normfit$estimate["mean"],
               sd   = normfit$estimate["sd"]),
      from = 38000, to = 45000, lwd = 2, col = "red", add = TRUE)
```



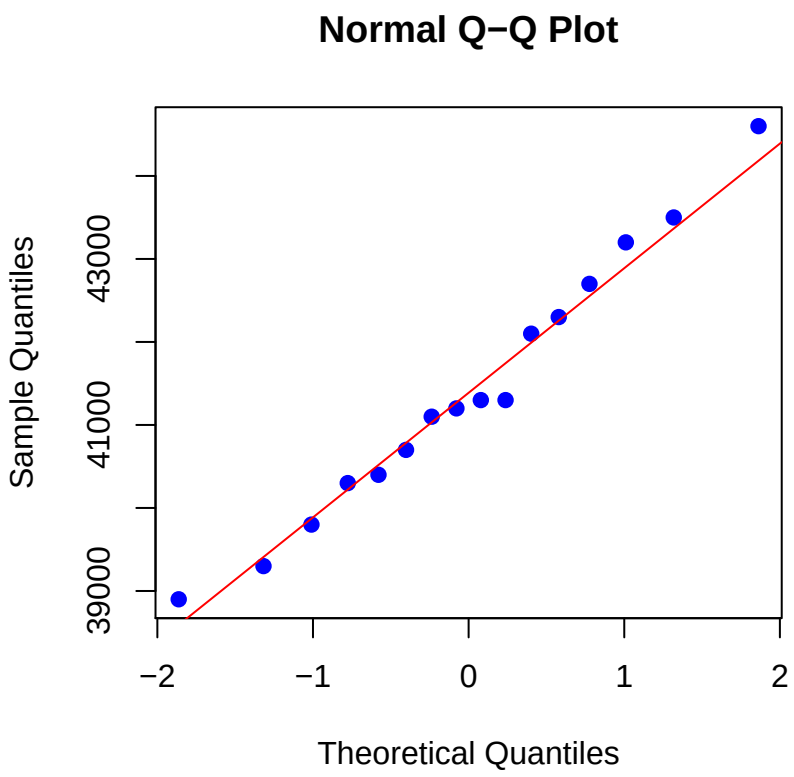
(d) We could plot directly against the fitted distribution:

```
n <- length(x)
p <- 1:n / (1 + n)
plot(qnorm(p, mean = normfit$estimate["mean"],
               sd   = normfit$estimate["sd"]),
     sort(x), pch = 19, col = 4,
     xlab = "Theoretical quantiles",
     ylab = "Sample quantiles")
abline(0, 1, col = 2)
```



It also suffices to compare against the standard normal (very similar output):

```
qqnorm(x, pch = 19, col = 4)  # normal QQ plot
qqline(x, col = 2)  # add reference line
```



The model looks like a very good fit to the data.

2. (a) i. $\mathbb{E}(X) = 1 \times \theta^2 + 2 \times 2\theta(1-\theta) + 3 \times (1-\theta)^2 = -2\theta + 3$.
$\mathbb{E}(X^2) = 1^2 \times \theta^2 + 2^2 \times 2\theta(1-\theta) + 3^2 \times (1-\theta)^2 = 2\theta^2 - 10\theta + 9$.
$\text{var}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2 = 2\theta - 2\theta^2 = 2\theta(1-\theta)$.

ii. The MM estimator is obtained by solving $-2\theta + 3 = \bar{X}$, which gives $\tilde{\theta} = \frac{3-\bar{X}}{2}$.
Since $\bar{x} = 1.75$, we can calculate the estimate as $\tilde{\theta} = \frac{3-1.75}{2} = 0.625$.

iii. $\text{var}(\bar{X}) = \frac{1}{n}\text{var}(X) = \frac{2\theta-2\theta^2}{n}$, and $\text{var}(\tilde{\theta}) = \left(\frac{1}{2}\right)^2 \text{var}(\bar{X}) = \frac{\theta-\theta^2}{2n}$, so we have $\text{se}(\bar{\theta}) = \sqrt{\frac{\tilde{\theta}-\tilde{\theta}^2}{2n}} = \sqrt{\frac{0.625-0.625^2}{2\times 20}} = 0.0765$.
Alternatively, we could use $\text{se}(\bar{\theta}) = \frac{1}{2}\frac{s}{\sqrt{20}} = 0.0879$, although this is less precise.

(b) i. The likelihood function is,

$$L(\theta) = \prod_{i=1}^{n} p(X_i) = \{\theta^2\}^{F_1}\{2\theta(1-\theta)\}^{F_2}\{(1-\theta)^2\}^{F_3} = 2^{F_2}\theta^{2F_1+F_2}(1-\theta)^{F_2+2F_3}.$$

ii. The log-likelihood function is,

$$\ln L = (2F_1 + F_2)\ln\theta + (F_2 + 2F_3)\ln(1-\theta) + \text{const.}$$

Taking the first derivative,

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

Setting this to zero and solving gives the maximum likelihood estimator,

$$\hat{\theta} = \frac{2F_1 + F_2}{2n}.$$

For the given sample, the maximum likelihood estimate is $\frac{2f_1+f_2}{2n} = 0.625$.

iii. Since $F_1 + F_2 + F_3 = n$ and $n\bar{X} = \sum X_i = F_1 + 2F_2 + 3F_3$, we can obtain $2F_1 + F_2 = 3n - n\bar{X}$. Therefore, $\hat{\theta} = \frac{2F_1+F_2}{2n} = \frac{3-\bar{X}}{2} = \tilde{\theta}$, i.e. the MLE is the same as the method of moments estimator. So we have $\text{var}(\hat{\theta}) = \text{var}(\tilde{\theta}) = \frac{\theta-\theta^2}{2n}$.

3. **Only the final answers are given here. For more details, please see the video consultation *Mean square error* on the LMS.**

(a) i. $\tilde{\theta} = 2X, \quad \mathbb{E}(\tilde{\theta}) = \theta, \quad \text{var}(\tilde{\theta}) = \frac{1}{3}\theta^2$.

ii. $\hat{\theta} = X, \quad \mathbb{E}(\hat{\theta}) = \frac{1}{2}\theta, \quad \text{var}(\hat{\theta}) = \frac{1}{12}\theta^2$.

(b) i. (See the video consultation)

ii. $\text{MSE}(\tilde{\theta}) = \text{MSE}(\hat{\theta}) = \frac{1}{3}\theta^2$.

iii. $\text{MSE}(\frac{3}{2}X) = \frac{1}{4}\theta^2$.

(c) i. $\tilde{\theta} = 2\bar{X}, \quad \mathbb{E}(\tilde{\theta}) = \theta, \quad \text{var}(\tilde{\theta}) = \frac{1}{3n}\theta^2, \quad \text{MSE}(\tilde{\theta}) = \frac{1}{3n}\theta^2$.

ii. $\hat{\theta} = X_{(n)}, \quad \mathbb{E}(\hat{\theta}) = \frac{n}{n+1}\theta, \quad \text{var}(\hat{\theta}) = \frac{n}{(n+1)^2(n+2)}\theta^2, \quad \text{MSE}(\hat{\theta}) = \frac{2}{(n+1)(n+2)}\theta^2$.

iii. $a = \frac{n+2}{n+1}$.

4. (a) The likelihood function is

$$L(\mu, \lambda) = \frac{1}{(2\pi\lambda)^{n/2}} \exp\left\{-\frac{1}{2\lambda}\sum_{i=1}^{n}(\ln x_i - \mu)^2\right\} \prod_{i=1}^{n}\frac{1}{x_i}.$$

The log-likelihood function is of the form,

$$\ell(\mu, \lambda) = -\frac{n}{2}\ln(2\pi\lambda) - \frac{1}{2\lambda}\sum_{i=1}^{n}(\ln x_i - \mu)^2 - \ln\left(\prod_{i=1}^{n} x_i\right).$$

Differentiating with respect to μ and setting equal to zero gives

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

which implies the MLE of μ is $\hat{\mu} = \frac{1}{n}\sum \ln X_i$. Differentiating with respect to λ gives

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

Solving this shows that the MLE of λ is $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(\ln X_i - \hat{\mu})^2$.

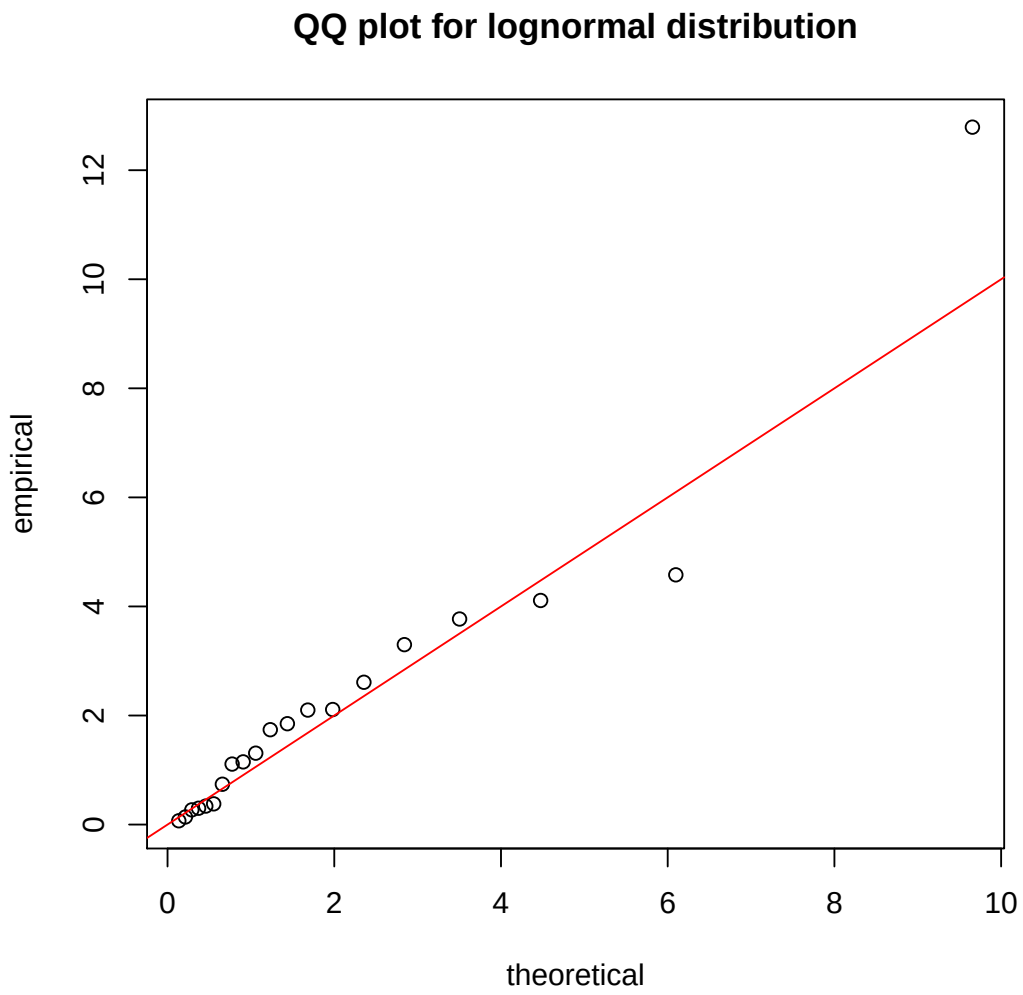
(b) Since $\ln X_i \sim \text{N}(\mu, \lambda)$, we have

$$\frac{n\hat{\lambda}}{\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n}\left(\ln X_i - \frac{1}{n}\sum_{i=1}^{n}\ln X_i\right)^2 \sim \chi^2_{n-1}.$$

Therefore we know that $1 - \alpha = \Pr\left(a < \frac{n\hat{\lambda}}{\lambda} < b\right)$ where a and b are the $\alpha/2$ and $1 - \alpha/2$ quantiles of χ^2_{n-1}. Rearranging the inequality shows that a $100 \cdot (1 - \alpha)\%$ CI for λ is $\left(\frac{n\hat{\lambda}}{b}, \frac{n\hat{\lambda}}{a}\right)$.

(c) i.
```r
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)  # sample size
lambda.hat <- (1 / n) * (n - 1) * var(log(x))  # MLE
lambda.hat
## [1] 1.637668
a <- qchisq(0.025, n - 1)  # quantiles
b <- qchisq(0.975, n - 1)
lambda.hat * c(1 / b, 1 / a) * n  # 95% CI
## [1] 0.9969874 3.6774596
```

ii.
```r
p = (1:n) / (n + 1)  # probabilities
theoretical <- qlnorm(p, meanlog = mean(log(x)),
                         sdlog   = sqrt(lambda.hat))
empirical   <- sort(x)
plot(theoretical, empirical,
     main = "QQ plot for lognormal distribution")
abline(0, 1, col = "red")  # add reference line
```



The QQ plot shows that the model fits the data fairly well. The majority of points lie close to the straight line. The few more extreme points (in the top-

right) deviate a little from the line, which might indicate that the right tail of the distribution might not describe the data quite as well. However, this is actually fairly common to see in QQ plots (the tails are 'noisier' because they only have a few points and the variance of the extreme quantiles is larger than the central ones), so we shouldn't overemphasise them. Therefore, we are happy to use the lognormal as a model for these data.

5. (a) Calculating the expectations:

$$\begin{aligned}
\mathbb{E}(T_1) &= \frac{1}{3}\{\mathbb{E}(X_1) + \mathbb{E}(X_2)\} + \frac{1}{6}\{\mathbb{E}(X_3) + \mathbb{E}(X_4)\} = \mu \\
\mathbb{E}(T_2) &= \frac{1}{6}\{\mathbb{E}(X_1) + 2\,\mathbb{E}(X_2) + 3\,\mathbb{E}(X_3) + 4\,\mathbb{E}(X_4)\} = \frac{5}{3}\mu \\
\mathbb{E}(T_3) &= \frac{1}{4}\{\mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_3) + \mathbb{E}(X_4)\} = \mu \\
\mathbb{E}(T_4) &= \frac{1}{3}\{\mathbb{E}(X_1) + \mathbb{E}(X_2) + \mathbb{E}(X_3)\} + \frac{1}{4}\,\mathbb{E}(X_4^2) = \mu + \frac{1}{4}\sigma^2 > \mu
\end{aligned}$$

Therefore, T_1 and T_3 are unbiased.

(b) The variances of T_1 and T_3 can be calculated by:

$$\begin{aligned}
\mathrm{var}(T_1) &= \frac{1}{9}\{\mathrm{var}(X_1) + \mathrm{var}(X_2)\} + \frac{1}{36}\{\mathrm{var}(X_3) + \mathrm{var}(X_4)\} = \frac{5}{18}\sigma^2 \\
\mathrm{var}(T_3) &= \frac{1}{16}\{\mathrm{var}(X_1) + \mathrm{var}(X_2) + \mathrm{var}(X_3) + \mathrm{var}(X_4)\} = \frac{1}{4}\sigma^2
\end{aligned}$$

Since $\frac{1}{4} < \frac{5}{18}$, T_3 has a smaller variance than T_1.