MAST20005/MAST90058: Assignment 1 Solutions

1. (a)
```r
x <- c(173.1,  61.5, 123.3, 100.4,  20.4,  20.9,
       228.4,   1.0,   6.8,  11.4,   7.7,  40.7,
        15.8, 422.4,  58.2,  19.9,  38.8, 121.0,
       118.6, 174.9,  87.2,  14.0, 204.7,  81.9,
        57.3, 177.0,  14.1, 137.0,  76.4, 330.2)
summary(x)
```

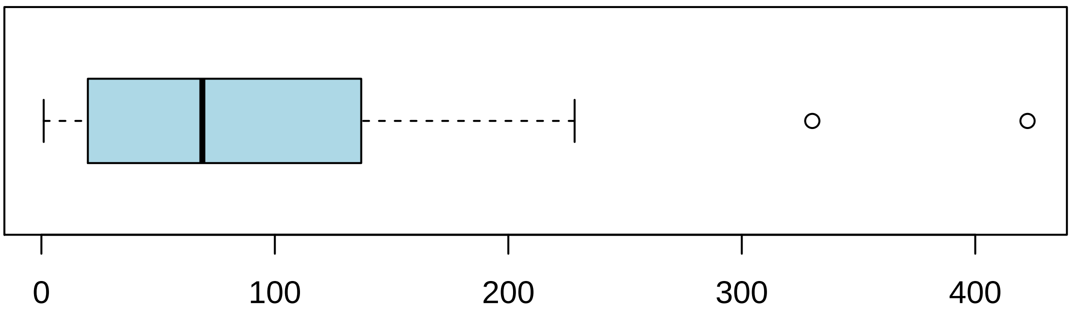
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   20.02   68.95   98.17  133.57  422.40
```

```r
sd(x)
```

```
## [1] 100.5084
```

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

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



The distribution of claims is centred around median value of 98 and has pronounced variability with sample standard deviation also around 100.5. The distribution is asymmetric (right-skewed).

(b) Using pdf: $f(x \mid \theta) = \frac{1}{\theta} e^{-\frac{x}{\theta}}$.

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

```
##       rate
##   0.010186757
##  (0.001859839)
```

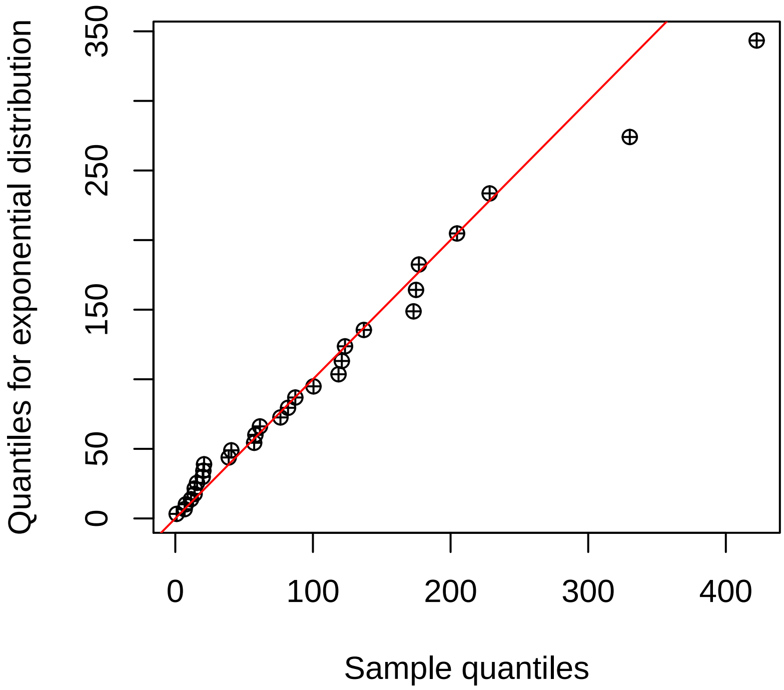
```r
1 / normfit$estimate
```

```
##     rate
## 98.16667
```

This gives $\hat{\theta} = 98.17$.

Alternate pdf: $f(x \mid \lambda) = \lambda e^{-\lambda x}$. This gives $\hat{\lambda} = 0.01$.

(c)
```
n <- length(x)
p <- 1:n / (1 + n)
x.sorted <- sort(x)          # sample quantiles
Finv <- -100 * log(1 - p)    # theoretical quantiles, for Exp(100)
plot(x.sorted, Finv, pch = 10, xlab = "Sample quantiles",
     ylab = "Quantiles for exponential distribution")
abline(0, 1, col = 2)
```



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

(d) The approach will work. The only difference will be that the best fitting line will have a different slope.

The points in Jen's plot are $\{x_{(k)}, G^{-1}(k/(n+1))\}$, where $G^{-1}(p) = -\log(1-p)$ is the quantile function of Exp(1). However, if the hypothesis is correct then the data follow $X \sim \text{Exp}(100)$, with theoretical quantiles given by $F^{-1}(p) = -100 \log(1-p)$. Since $x_{(k)} \approx F^{-1}(k/(n+1)) = \frac{1}{100} \times G^{-1}(k/(n+1))$, Jen's best fitting line will have approximately an intercept of 0 and a slope of 1/100.

If Jen orients the QQ plot the other way around, with the sample quantiles on the y-axis and the theoretical quantiles on the x-axis, the only change is that the slope would be approximately 100 rather than 1/100.

2. (a) The likelihood function is

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

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

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

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

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

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

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

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

i. Since $\mathrm{var}(n\hat{\lambda}/\lambda) = 2(n-1)$, we have $n^2/\lambda^2\,\mathrm{var}(\hat{\lambda}) = 2(n-1)$ and therefore $\mathrm{var}(\hat{\lambda}) = 2(n-1)\lambda^2/n^2$. The standard deviation is $\mathrm{sd}(\hat{\lambda}) = \sqrt{\mathrm{var}(\hat{\lambda})} = \sqrt{2(n-1)\lambda^2/n^2} = \sqrt{2(n-1)}\,\lambda/n$.

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

(c) i.
```
x <- c(12.9, 2.3, 2.4, 65.0, 6.7, 248.7, 1.0, 2.0,
       4.9, 3.6, 1.8,  1.5, 1.7,   4.1, 6.8)
n <- length(x)  # sample size
lambda.hat <- (1 / n) * (n - 1) * var(log(x))  # MLE
lambda.hat
## [1] 2.086394
sqrt(2 * (n - 1)) * lambda.hat / n
## [1] 0.7360108
```

The standard error is 0.74.

ii. The MLE is given above. The CI is calculated as follows:

```
a <- qchisq(0.025, n - 1)  # quantiles
b <- qchisq(0.975, n - 1)
n * lambda.hat / c(b, a)  # 95% CI
## [1] 1.198207 5.560036
```

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

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

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

(b) i. (See the video consultation)

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

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

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

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

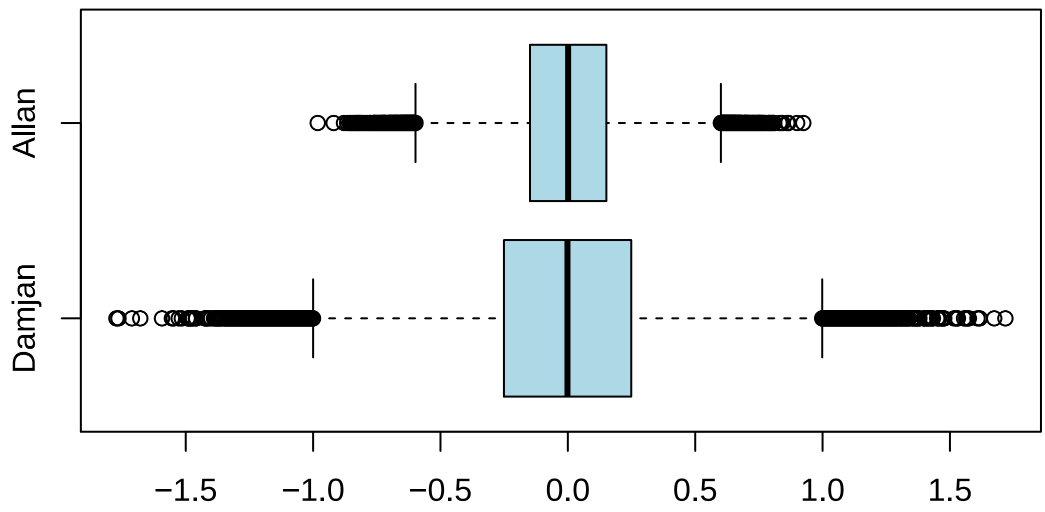
iii. $a = \frac{n+2}{n+1}$.

4. Simulating from a standard normal distribution:

```
B <- 100000  # simulation runs
t1 <- numeric(B)
t2 <- numeric(B)
for (i in 1:B) {
    x <- rnorm(20)
    t1[i] <- 0.5 * (min(x) + max(x))  # Damjan's estimator
    t2[i] <- mean(x)                  # Allan's estimator
}
mean(t1)

## [1] -0.001254656

mean(t2)

## [1] 0.001158698

sd(t1)

## [1] 0.3777801

sd(t2)

## [1] 0.2234677

sd(t1) / sd(t2)

## [1] 1.690535
```

Both estimators appear to be unbiased, but Damjan's estimator has much greater standard deviation (about 69% greater).

```
par(mar = c(3, 4, 1, 1))  # compact margins
boxplot(t1, t2, names = c("Damjan", "Allan"), horizontal = TRUE,
        col = "lightblue")
```



Simulating from a normal distribution with different parameters will not change any of the above conclusions (working not shown).

5. (a) The expectations can be calculated by

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

Therefore, T_1 and T_3 are unbiased.

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

$$\begin{aligned}
\mathrm{var}(T_1) &= \frac{1}{16}\{\mathrm{var}(X_1) + \mathrm{var}(X_2)\} + \frac{1}{4}\,\mathrm{var}(X_3) = \frac{61}{576}\sigma^2 = 0.106\,\sigma^2, \\
\mathrm{var}(T_3) &= \frac{1}{9}\{\mathrm{var}(X_1) + \mathrm{var}(X_2) + \mathrm{var}(X_3)\} = \frac{49}{324}\sigma^2 = 0.151\,\sigma^2.
\end{aligned}$$

Therefore, T_1 has a smaller variance than T_3.

(c) Let $T_5 = \frac{1}{6}\{\mathbb{E}(X_1) + 2\,\mathbb{E}(X_2) + 3\,\mathbb{E}(X_3)\}$,

$$\begin{aligned}
\mathbb{E}(T_5) &= \frac{1}{6}\{\mathbb{E}(X_1) + 2\,\mathbb{E}(X_2) + 3\,\mathbb{E}(X_3)\} = \mu, \\
\mathrm{var}(T_5) &= \frac{1}{36}\{\mathrm{var}(X_1) + 4\,\mathrm{var}(X_2) + 9\,\mathrm{var}(X_3)\} = \frac{1}{12}\sigma^2 = 0.083\,\sigma^2.
\end{aligned}$$