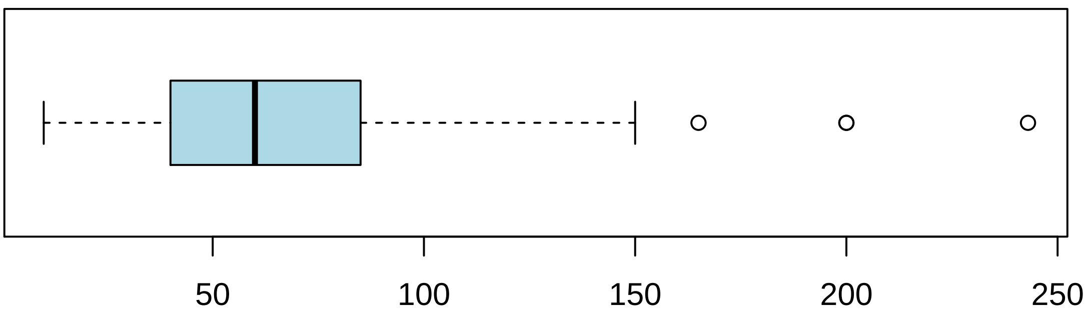
MAST20005/MAST90058: Assignment 1 Solutions

1. (a)
```r
quiz <- read.table("quiz.txt")[, 1]  # load data
summary(quiz)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   40.00   60.00   67.13   85.00  243.00

sd(quiz)

## [1] 40.54038
```

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(quiz, horizontal = TRUE, col = "lightblue")
```



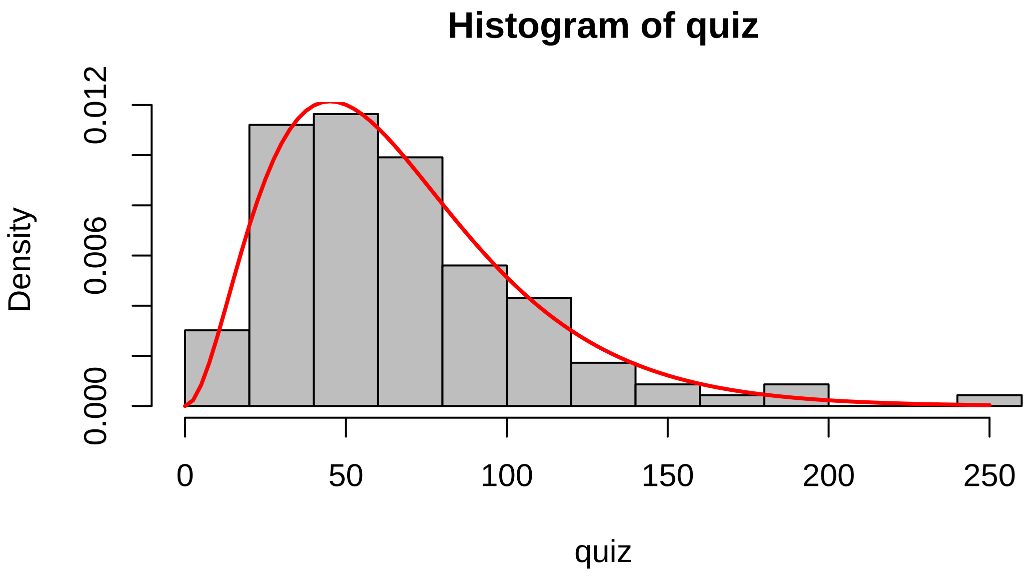
The distribution is centred around a median value of 60 and varies substantially, with sample standard deviation around 40. The distribution is asymmetric with a long right tail ('right-skewed'). Several observations are much higher than the others, as marked on the plot.

(b) Using pdf: $f(x \mid \alpha, \beta) = \beta^{\alpha} x^{\alpha-1} e^{-x\beta} / \Gamma(\alpha)$, $x \geqslant 0$, $\alpha > 0$ (shape), $\beta > 0$ (rate).

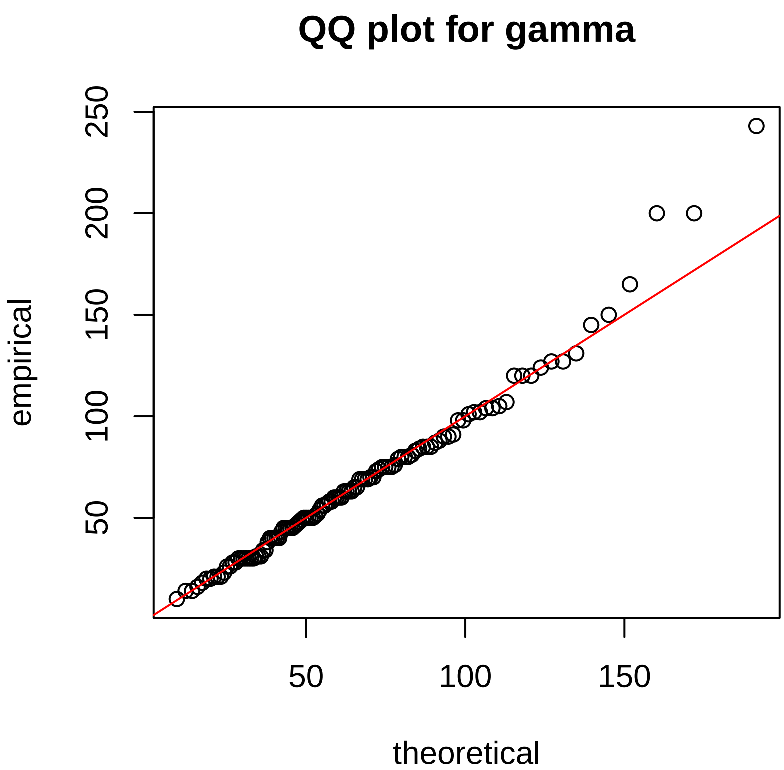
```r
library(MASS)
gammafit <- fitdistr(quiz, densfun = "gamma")
gammafit

##       shape          rate
##   3.041098377   0.045302228
##  (0.379119742) (0.006138772)
```

This gives $\hat{\alpha} = 3.0$ and $\hat{\beta} = 0.045$.
Alternate pdf: $f(x \mid k, \theta) = \theta^{-k} x^{k-1} e^{-x/\theta} / \Gamma(k)$, $x \geqslant 0$, $k > 0$ (shape), $\theta > 0$ (scale).
With some relevant R code, should get $\hat{k} = \hat{\alpha} = 3.0$ and $\hat{\theta} = 1/\hat{\beta} = 22$.

(c)
```r
hist(quiz, breaks = 15, freq = FALSE, col = "grey")
curve(dgamma(x, shape = gammafit$estimate["shape"],
                rate  = gammafit$estimate["rate"]),
      from = 0, to = 250, lwd = 2, col = "red", add = TRUE)
```



(d)
```
n <- length(quiz)
p <- (1:n) / (n + 1)  # probabilities
theoretical <- qgamma(p, shape = gammafit$estimate["shape"],
                         rate  = gammafit$estimate["rate"])
empirical   <- sort(quiz)
plot(theoretical, empirical, main = "QQ plot for gamma")
abline(0, 1, col = "red")  # add reference line
```



The model looks like a very good fit to the data, except possibly for the very end of the right tail.

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

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

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

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

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

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

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

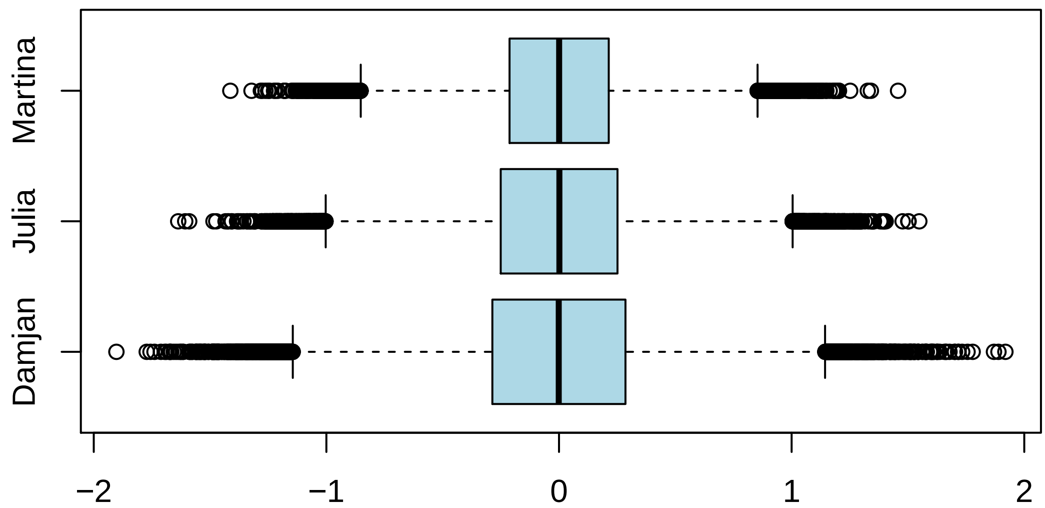
iii. $a = \frac{n+2}{n+1}$.

4. Simulating from a standard normal distribution:

```
B <- 100000  # simulation runs
t1 <- numeric(B)
t2 <- numeric(B)
t3 <- numeric(B)
for (i in 1:B) {
    x <- rnorm(10)
    t1[i] <- 0.5 * (min(x) + max(x))  # Damjan's estimator
    t2[i] <- median(x)                # Julia's  estimator
    t3[i] <- mean(x)                  # Martina's estimator
}
mean(t1)

## [1] -0.001219032

mean(t2)

## [1] 0.0004520544

mean(t3)

## [1] 0.0001132827

sd(t1)

## [1] 0.4304497

sd(t2)

## [1] 0.3721903

sd(t3)

## [1] 0.3165136

sd(t1) / sd(t3)

## [1] 1.359972

sd(t2) / sd(t3)

## [1] 1.175906
```

All of the estimators appear to be unbiased, but Martina's estimator looks to be the most efficient (smallest variance). Compared to Martina's estimator, Damjan's has a standard deviation that is about 36% greater, and Julia's is about 18% greater.

```r
par(mar = c(3, 4, 1, 1))  # compact margins
boxplot(t1, t2, t3, names = c("Damjan", "Julia", "Martina"),
        horizontal = TRUE, col = "lightblue")
```



Repeating the simulations with different normal distributions (other than a standard normal) leads to the same conclusions.

5. (a) Calculating the expectations:

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

Therefore, only 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.