Assignment 1

Lucas Fern
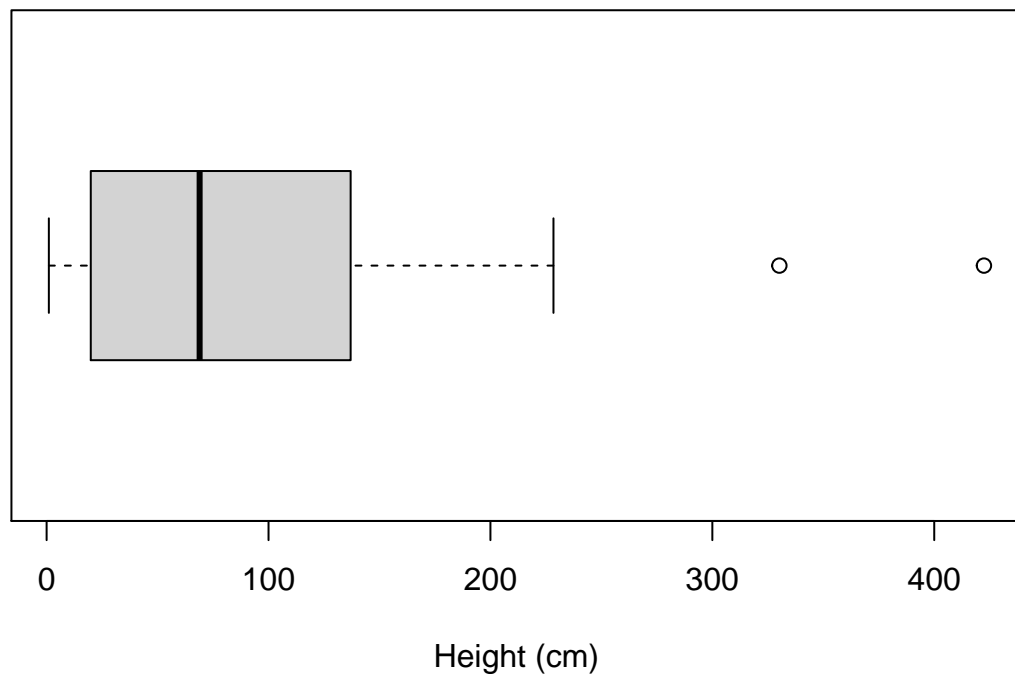
28/08/2020

Question 1

1a)

```
plant.heights <- 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(plant.heights)
```

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


Height (cm)

The distribution has a mean around 100 and a median at approximately 69cm, this indicates that the distribution is positively skewed which is confirmed by the box plot. The data points are heavily concentrated at lower values indicating that an exponential distribution could potentially fit this data closely.

1b)

```r
# No need to include a `start' for the Exponential distribution
exponential.fit <- fitdistr(plant.heights, "exponential")
exponential.fit
```

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

This estimate is using the alternative form of the exponential distribution where $\lambda = \frac{1}{\theta}$ so the estimate for the rate parameter in the form used on question 1c is $\frac{1}{0.010186757} = 98.1666687$.

1c)

```r
# Adding 1 to the denominator so that we don't get an Inf when CDF = 1
x <- (1:length(plant.heights)) / (length(plant.heights) + 1)
theoretical.quantiles <- qexp(x, rate = 1/100) # Rate parameter in the format
                                               # that R uses
sorted.plant.heights <- sort(plant.heights)

plot(theoretical.quantiles, sorted.plant.heights,
     xlab = "Exponential Quantiles", ylab = "Sample Quantiles")

# Fit and plot a Line of Best Fit
fit <- lm(sorted.plant.heights ~ theoretical.quantiles)
abline(fit)
```
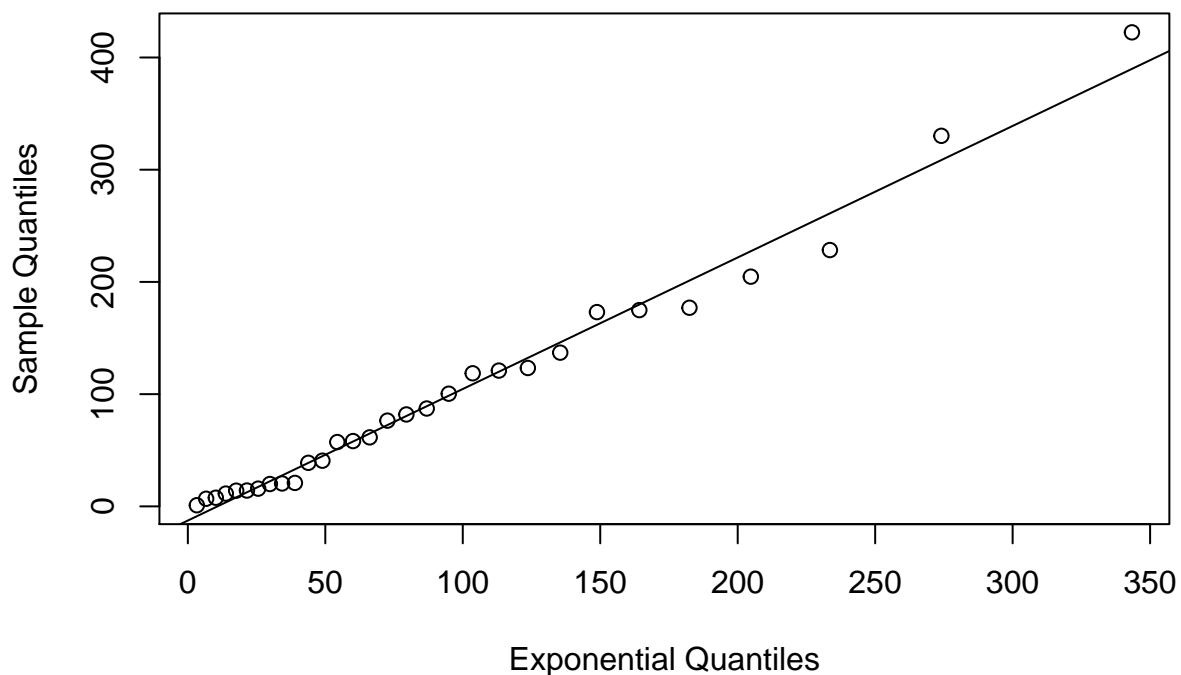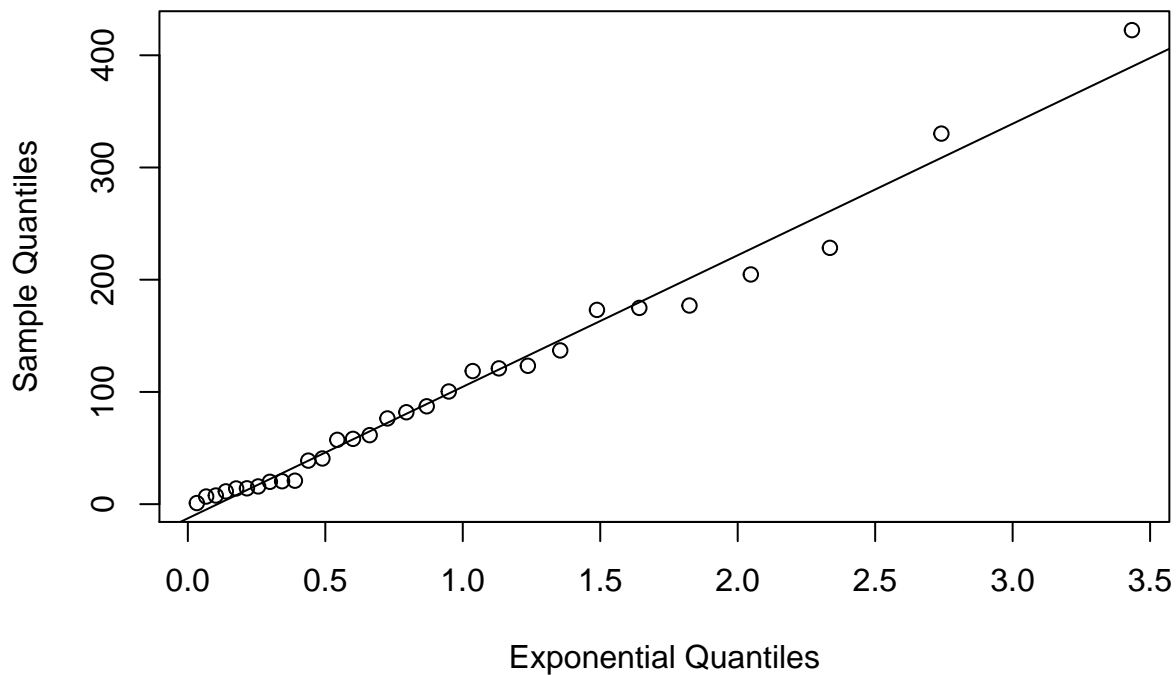
The observed data appears to fit very well to the exponential distribution.

1d)

Regenerating the plot with

```
theoretical.quantiles <- qexp(x, rate = 1)
```

yields the following plot:

From this we can observe that Jen will effectively see the same plot as Tom, since changing the rate parameter of the Exponential distribution simply has the effect of linearly dilating the quantiles from the line $x = 0$. The difference is that the slope of Jen's best fitting line will be off from Toms by a factor of 100.

```
lm(sorted.plant.heights ~ theoretical.quantiles)
```

```
##
## Call:
## lm(formula = sorted.plant.heights ~ theoretical.quantiles)
##
## Coefficients:
##           (Intercept)  theoretical.quantiles
##                 -12.64                 117.21
```

Tells us that Jen's line of best fit will have a y intercept of ~-12.642 and a slope of ~1.172.

Question 2

2a)

$$f(x|\mu, \lambda) = \frac{1}{x\sqrt{2\pi\lambda}} \exp\left(-\frac{(\ln x - \mu)^2}{2\lambda}\right)$$

$$\implies \Pr(X_1 = x_1, ..., X_n = x_n|\mu, \lambda) = \prod_{i=1}^{n}\left[\frac{1}{x_i}\right]\frac{1}{\sqrt{2\pi\lambda}^n}\exp\left(-\frac{\sum_{i=1}^{n}(\ln x_i - \mu)^2}{2\lambda}\right)$$

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

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

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

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

$$\implies \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln X_i \quad \blacksquare$$

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

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

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

$$\implies \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(\ln X_i - \hat{\mu})^2 \quad \blacksquare$$

2bi)

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

$Y_i = \ln(X_i) \sim N(\mu, \lambda)$

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$

$\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - (\frac{1}{n}\sum_{i=1}^{n}Y_i))$
$= \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})$

$= S^2 \cdot \frac{n-1}{n}$

$Var(\hat{\lambda}) = \left(\frac{n-1}{n}\right)^2 Var(S^2)$
$= \left(\frac{n-1}{n}\right)^2 (2\lambda^2)/(n-1)$
$= \frac{n-1}{n^2}(2\lambda^2)$

2bii)

2bii) A confidence interval for λ is:

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right] = \left[\frac{\sum(y_i - \bar{y})}{b}, \frac{\sum(y_i - \bar{y})}{a}\right]$$

Where $a = F^{-1}(\frac{\alpha}{2})$, $b = F^{-1}(1 - \frac{\alpha}{2})$
and F is the Chi-Squared CDF with n-1 degrees of freedom.

2ci)

```r
data <- c(12.9, 2.3, 2.4, 65.0, 6.7, 1.8, 1.5, 1.7,
          248.7, 1.0, 2.0, 4.9, 3.6, 4.1, 6.8)

mu.hat <- sum(log(data))/length(data)
lambda.hat <- sum((log(data)-mu.hat)^2)/length(data)

variance <- ((length(data) - 1) / length(data)^2) * 2 * lambda.hat^2
standard.error <- sqrt(variance)

standard.error
```

```
## [1] 0.7360108
```

2cii)

```r
lambda.hat
```

```
## [1] 2.086394
```

```r
# 0.975 and 0.025 for the 95% CI
conf.interval <- c(((length(data) - 1) * standard.error^2) /
                      qchisq(0.975, (length(data) - 1)),
                   ((length(data) - 1) * standard.error^2) /
                      qchisq(0.025, (length(data) - 1)))
conf.interval
```

```
## [1] 0.2903626 1.3473682
```

Question 3

3ai)

$$3ai)\ E(X) = \frac{\theta}{2}$$
$$\Rightarrow \hat{\theta} = 2\bar{X} = 2x$$

$$\bar{X} = X \sim R(0, \theta) \quad \text{has } E(X) = \frac{\theta}{2}$$
$$\text{and } Var(X) = \frac{1}{12}\theta^2$$
$$E(\hat{\theta}) = 2\frac{\theta}{2} = \theta$$

$$\Rightarrow Var(\hat{\theta}) = 2^2\left(\frac{\theta^2}{12}\right)$$
$$= \frac{1}{3}(\theta^2)$$

3aii)

aii) $Pr(X = x \mid \theta) = f_X(x \mid \theta)$
$$= \tfrac{1}{\theta}$$
$$= L(\theta)$$

$L(\theta) = \tfrac{1}{\theta}$ is strictly decreasing for $\theta > 0$
and since $\theta \geqslant x$, $\theta = x$ maximises $L(\theta)$.
$$\Rightarrow \tilde{\theta} = x = \bar{x}$$

$\tilde{\theta}$ then has the same mean and variance as x.
$$E(\tilde{\theta}) = E(X) = \tfrac{\theta}{2}$$
$$Var(\tilde{\theta}) = Var(X) = \tfrac{1}{12}\theta^2$$

3bi)

bi) $MSE(\hat{\theta}) = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$
$$= E(\hat{\theta}^2) - E(\hat{\theta})^2 + E(\hat{\theta})^2 - 2\theta E(\hat{\theta}) + \theta^2$$
$$= E(\hat{\theta}^2) - 2\theta E(\hat{\theta}) + \theta^2$$

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

7

3bii)

bii) $MSE(\hat{\theta}) = MSE(2\bar{X})$
$$= \tfrac{1}{3}\theta^2 + 0^2 = \tfrac{1}{3}\theta^2$$

$MSE(\tilde{\theta}) = MSE(\bar{X})$
$$= \tfrac{1}{12}\theta^2 + \tfrac{\theta^2}{4}$$
$$= \tfrac{1}{3}\theta^2$$

They are the same.

3biii)

biii) $\hat{\hat{\theta}} = \tfrac{3}{2}\bar{X}$
$E(\hat{\hat{\theta}}) = \tfrac{3}{2}E(\bar{X}) = \tfrac{3}{4}\theta$
$Var(\hat{\hat{\theta}}) = \tfrac{9}{4}Var(\bar{X}) = \tfrac{9\theta^2}{48}$

$MSE(\hat{\hat{\theta}}) = \tfrac{9\theta^2}{48} + \tfrac{1}{16}\theta^2$
$$= \tfrac{12}{48}\theta^2$$
$$= \tfrac{1}{4}\theta^2$$

3ci)

ci) $E(X) = \frac{\theta}{2}$

$\bar{X} \approx N(\frac{\theta}{2}, \frac{\theta^2}{12n})$

$\Rightarrow \frac{\hat{\theta}}{2} = \bar{X}$

$\hat{\theta} = 2\bar{X}$

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

$Var(\hat{\theta}) = 2^2 \frac{\theta^2}{12n} = \frac{1}{3n}\theta^2$

$MSE(\hat{\theta}) = \frac{\theta^2}{3n}$

3cii)

cii) $Pr(X_1 = x_1, \ldots, X_n = x_n | \theta) = \prod_{i=1}^{n} \left(\frac{1}{\theta}\right)$

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

$\frac{1}{\theta^n}$ is decreasing for $\theta > 0$ so the smallest θ possible maximises $L(\theta)$

$\therefore$ the MLE is $\max(X_1, \ldots, X_n) = X_{(n)}$

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

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

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

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

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

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

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

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

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

3ciii)

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

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

$$MSE(a\hat{\theta}) = a^2 Var(\hat{\theta}) + \left(aE(\hat{\theta}) - \theta\right)^2$$

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

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

setting equal to 0:

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

Question 4

```
sample.size <- 20
ntrials <- 1000
sample.means <- 1:ntrials
sample.midranges <- 1:ntrials

for (i in 1:ntrials) {
  x <- rnorm(sample.size)
  sample.means[i] <- mean(x)
  sample.midranges[i] <- (max(x) + min(x)) / 2
}

mean(sample.means)
```

```
## [1] -0.009691173
```

```
mean(sample.midranges)
```

```
## [1] 0.0038145
```

```
sd(sample.means)
```

```
## [1] 0.2271058
```

```
sd(sample.midranges)
```

[1] 0.3822993

These values show that the mean of both estimators for the population mean are equal to the population mean, indicating that they are unbiased. The variance of the sample mean is, however, much lower than the variance of the average of the minimum and maximum values. This box plot confirms these claims (the magenta line is plotted at the true population mean).

Question 5

5a)
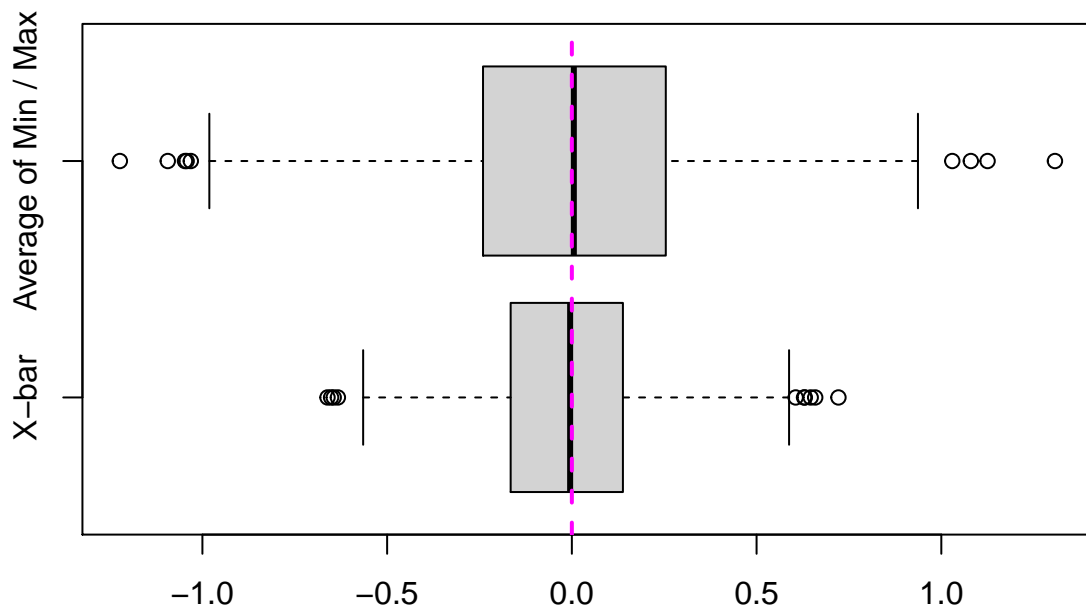
$$5a) \quad E(T_1) = E(\tfrac{1}{4}X_1 + \tfrac{1}{4}X_2 + \tfrac{1}{2}X_3)$$
$$= \tfrac{1}{4}\mu + \tfrac{1}{4}\mu + \tfrac{1}{2}\mu$$
$$= \mu$$
$$E(T_2) = E(\tfrac{1}{3}X_1 + \tfrac{2}{3}X_2 + X_3)$$
$$= \tfrac{1}{3}\mu + \tfrac{2}{3}\mu + \mu$$
$$= 2\mu$$
$$E(T_3) = E(\tfrac{1}{3}X_1 + \tfrac{1}{3}X_2 + \tfrac{1}{3}X_3)$$
$$= \tfrac{1}{3}\mu + \tfrac{1}{3}\mu + \tfrac{1}{3}\mu$$
$$= \mu$$
$$E(T_4) = E(\tfrac{1}{2}X_1 + \tfrac{1}{2}X_2 + \tfrac{1}{4}X_3^2)$$
$$= \tfrac{1}{2}\mu + \tfrac{1}{2}\mu + E(X_3^2)$$
$$E(X_3^2) = Var(X_3) + E(X_3)^2$$
$$= \tfrac{\sigma^2}{9} + \mu^2$$
$$\Rightarrow E(T_4) = \mu + \tfrac{1}{4}\left(\tfrac{\sigma^2}{9} + \mu^2\right)$$

So T_1 and T_3 are unbiased.

13

5b)

b) $\text{Var}(T_1) = \text{Var}(\frac{1}{4}X_1 + \frac{1}{4}X_2 + \frac{1}{2}X_3)$

$= \frac{1}{16}\text{Var}(X_1) + \frac{1}{16}\text{Var}(X_2) + \frac{1}{4}\text{Var}(X_3)$ (since IID)

$= \frac{1}{16}(\sigma^2 + \frac{\sigma^2}{4}) + \frac{1}{4}(\frac{\sigma^2}{9})$

$= 61\sigma^2/576$

$\text{Var}(T_3) = \text{Var}(\frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3)$

$= \frac{1}{9}(\sigma^2 + \frac{\sigma^2}{4} + \frac{\sigma^2}{9})$ (IID)

$= 49\sigma^2/324$

So T_1 has the smallest variance.

5c)

c) $T_5 = \frac{1}{2}X_2 + \frac{1}{2}X_3$

$E(T_5) = \frac{1}{2}\mu + \frac{1}{2}\mu = \mu$

$\text{Var}(T_5) = \frac{1}{4}(\frac{\sigma^2}{4} + \frac{\sigma^2}{9})$

$= 13\sigma^2/144$