MAST20005/MAST90058: Week 2 Solutions

1. Since this is a random sample, all of the variables are iid. Therefore they all have the same mean and variance.

   (a) $\text{sd}(X_2) = \text{sd}(X_4) = \sqrt{\text{var}(X_4)} = \sqrt{4} = 2$.

   (b) By independence, $\text{var}(X_7 + X_8) = \text{var}(X_7) + \text{var}(X_8) = \text{var}(X_4) + \text{var}(X_4) = 8$.

   (c) The covariance is zero since X_3 and X_4 are independent.

   (d) By the Central Limit Theorem, $\bar{X} \approx \text{N}\left(\mathbb{E}(X_1), \frac{\text{var}(X_1)}{9}\right) = \text{N}\left(7, \frac{4}{9}\right)$.

2. (a) $\mathbb{E}(X_1) = \int_{-1}^{1} x(3/2)x^2\,dx = 0$.

   (b) First we need to know $\text{var}(X_1)$, which we calculate as follows:

$$\mathbb{E}(X_1^2) = \int_{-1}^{1} x^2(3/2)x^2\,dx = \int_{-1}^{1} (3/2)x^4\,dx = \left[\frac{3}{10}x^5\right]_{-1}^{1} = \frac{3}{5},$$
$$\text{var}(X_1) = \mathbb{E}(X_1^2) - \mathbb{E}(X_1)^2 = \frac{3}{5}.$$

   Then we do the calculations for Y,

$$\mathbb{E}(Y) = 15 \cdot \mathbb{E}(X_1) = 0,$$
$$\text{var}(Y) = 15 \cdot \text{var}(X_1) = 9.$$

   (c) The CLT implies $\bar{X}$ is approximately normally distributed. Since $Y = 15\bar{X}$ then Y will also be approximately normally distributed. Therefore, using the results from above, $Y \approx \text{N}(0, 9)$. Thus,

$$\begin{aligned}\Pr(-0.3 < Y < 1.5) &= \Pr\left(\frac{-0.3 - 0}{3} < \frac{Y - 0}{3} < \frac{1.5 - 0}{3}\right) \\ &\approx \Pr\left(-0.1 < Z < 0.5\right) \\ &= \Phi(0.5) - \Phi(-0.1) = 0.23\end{aligned}$$

3. (a) Yes. Tingjin has tried to run each experiment under the same conditions and each pot is separate to the others. Therefore it is reasonable to assume the measurements are iid. The 'population' here is the (hypothetically infinite) set of possible replicates of this particular experiment, i.e. imaginary extra pots prepared in the same way.

   (b) The family members are, presumably (since we aren't give any further information), a mixture of adults and children, which means their heights will not be identically distributed. Furthermore, the family members will be genetically related (again, a resonable assumption without being told otherwise), which means they will also not be independent. Therefore, this is not a random sample.

   The population of interest is also not clear from the given description. Was Damjan only interested in the heights of his family members? In that case, he has obtained **all** of the measurements (i.e. the whole population), so we wouldn't usually describe this as a sample. Was he interested in heights of people in general, e.g. the population of all humans? In that case, this **is** a sample from that population, but not a random one.

(c) The population of interest here is, presumably, the number of people sitting down on South Lawn at any point in time. However, this should really be specified more precisely, e.g. does Robert care about what happens at 2am?

For simplicity, suppose Robert does his sampling at the same time each day (e.g. noon), then it is better to think of the population as being just the number of people sitting down at that time of the day, for an arbitrary day. Then, for this to be a random sample, we need to assume they are iid. This is unlikely to be the case: weekends and weekdays are likely to differ, and also semester time versus non-semester time.

Suppose Robert did his sampling at different times of the day (and we therefore use the more expanded notion of the population). Then there we will be an even greater deviation from the iid assumption, since we expect this count to vary substantially across the day.

4. (a) i. For the sample mean:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n} x_i - n\bar{x} = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i = 0$$

For the sample variance:

$$\begin{aligned} s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - 2n\bar{x}^2 + n\bar{x}^2\right) \\ &= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right) \end{aligned}$$

ii. $\bar{x} = 219.3/5 = 43.86$, $s = \sqrt{(1/4)(9654.27 - 5 \times 43.86^2)} = 2.99$

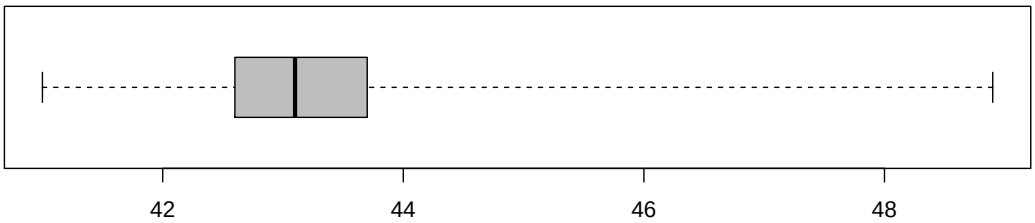
(b) i. The identity for the sample mean stays the same. The identity for the sample variance still holds, with the 'new' value of s^2 being 4 times greater than the 'old' value.

ii. Both the mean and the standard deviation will be double their previous values.

(c) i. The ordered data are: 41.0, 42.6, 43.1, 43.7, 48.9. The median is the 3rd ordered observation, 43.1. The first and third (Type 7) quartiles are the 2nd and 4th order statistics, 42.6 and 43.7 respectively. The five-number summary therefore is {41.0, 42.6, 43.1, 43.7, 48.9}.

ii. IQR = 43.7 − 42.6 = 1.1.

iii.



(d) i. The sample median is the same under this definition, but the 1st and 3rd quartiles differ:

$$(5+1)\cdot 0.25 = 1 + 0.5 \;\Rightarrow\; \widetilde{\pi}_{0.25} = x_{(1)} + 0.5\cdot(x_{(2)} - x_{(1)}) = 41.8$$
$$(5+1)\cdot 0.75 = 4 + 0.5 \;\Rightarrow\; \widetilde{\pi}_{0.75} = x_{(4)} + 0.5\cdot(x_{(5)} - x_{(4)}) = 46.3$$

Therefore, the five-number summary is $\{41.0, 41.8, 43.1, 46.3, 48.9\}$.

ii. Let $k = i + r$. Non-integer order statistics were defined by linear interpolation,

$$x_{(k)} = x_{(i+r)} = x_{(i)} + r\cdot(x_{(i+1)} - x_{(i)})$$

In other words, $x_{(k)} = \widetilde{\pi}_p$ where $(n+1)p = k$. Rearranging the latter gives $p = k/(n+1)$, which is the definition of a 'Type 6' quantile.

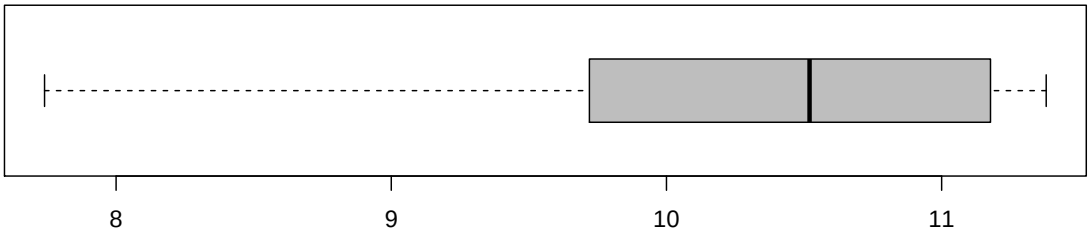
(e) Using the Type 7 quantiles: $\hat{\pi}_{0.25} - 1.5\times\text{IQR} = 40.95$ and $\hat{\pi}_{0.75} + 1.5\times\text{IQR} = 45.35$. Since the observation 48.9 lies outside of these two extremes, it would be classified as an outlier. Repeating this calculation with Type 6 quantiles gives a larger IQR and no observations end up being classified as outliers.

5. The set $\{1, 1, 4, 4\}$ maximises the sample variance.

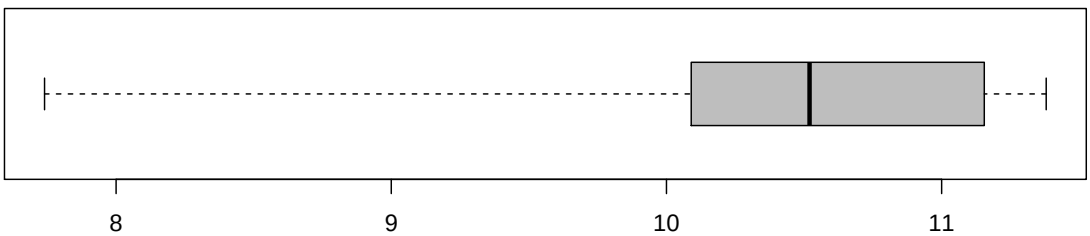
6. (a)
```
x <- c(10.39, 10.43, 9.99, 11.17,  8.91,
       11.20, 11.38, 7.74, 10.61, 11.11)
quantile(x, type = 6)

##      0%     25%     50%     75%    100%
##  7.7400  9.7200 10.5200 11.1775 11.3800

quantile(x, type = 7)

##     0%    25%    50%    75%   100%
##  7.740 10.090 10.520 11.155 11.380
```

(b) If using Type 6 quantiles, there are no outliers. If using Type 7 quantiles, then the observation 7.74 is an outlier.

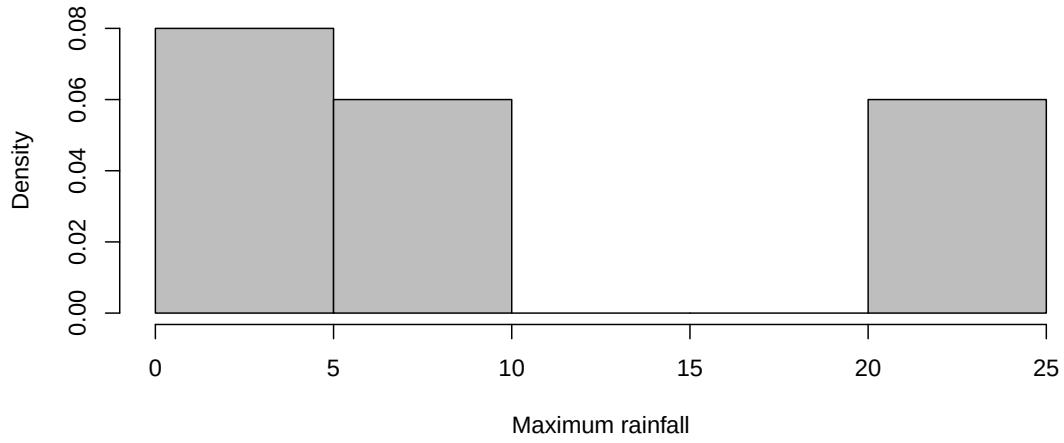
(c) With Type 6 quantiles:



With Type 7 quantiles:



7. (a)
```r
x <- c(9.9, 4.7, 20.5, 1.8, 4.7, 9.8, 20.5, 20.2, 6.5, 3.0)
hist(x, xlab = "Maximum rainfall", freq = FALSE, main = NULL, col = 8)
```



(b) Define $p_k = k/(n+1)$, for $k = 1, \ldots, n$. These have values:

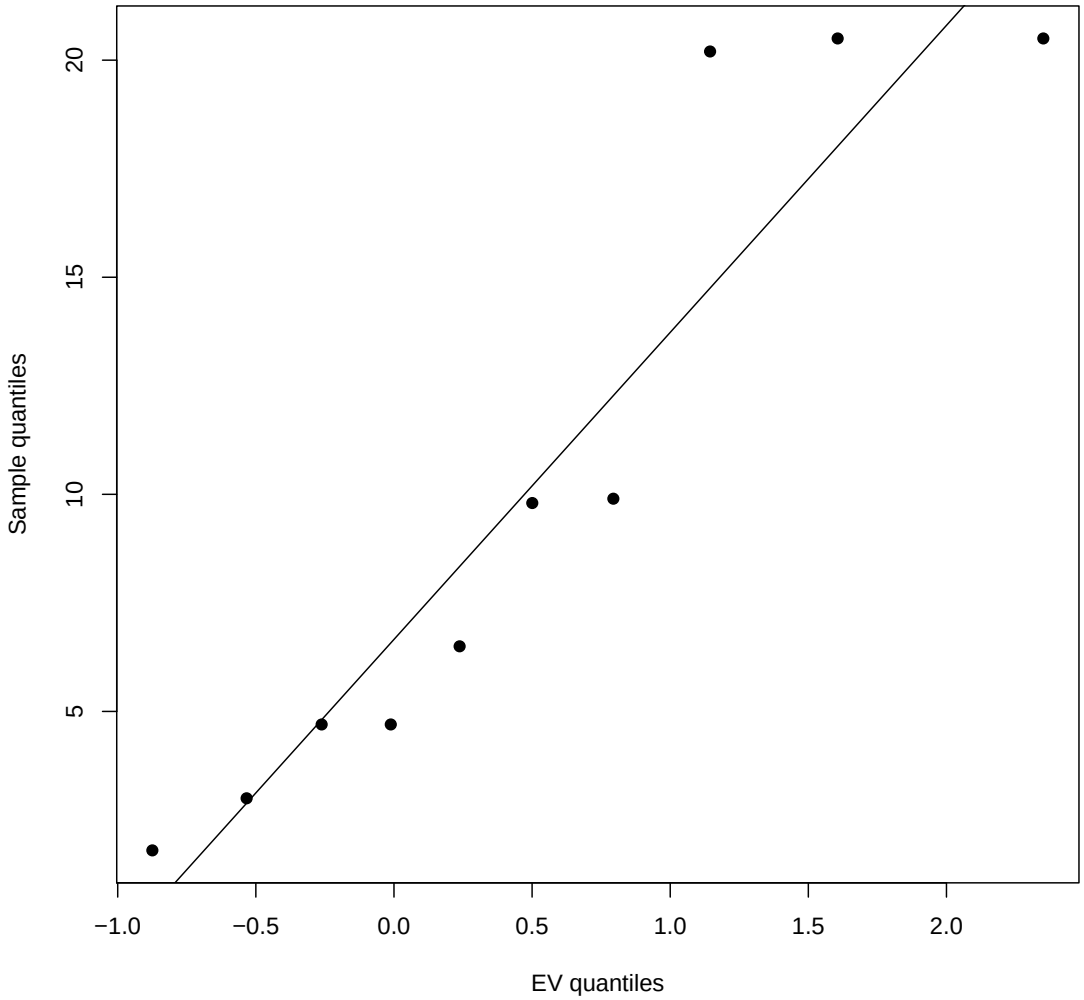
$$p_1 = 1/11 = 0.09, \ldots, p_{10} = 10/11 = 0.91$$

The theoretical quantiles are:

$$F^{-1}(p_1) = -0.9, \ldots, F^{-1}(p_{10}) = 2.4$$

We plot these against the corresponding sample quantiles, which are just the order statistics $(1.8, \ldots, 20.5)$. Some R code to make the plot:

```r
p <- 1:10/11
Finv <- -log(-log(p))
x.sorted <- sort(x)
plot(Finv, x.sorted, pch = 19,
     xlab = "EV quantiles",
     ylab = "Sample quantiles")
# Adds best fitting line
fit <- lm(x.sorted ~ Finv)
abline(fit)
```




```
coef(fit)

## (Intercept)        Finv
##    6.658273    7.071246
```

To estimate θ and ξ we can use the intercept and slope of the best fitting line computed according to some appropriate method. Based on the analysis above we obtain $\hat{\theta} = 6.66$ and $\hat{\xi} = 7.07$.