Note that there are FOUR questions split across TWO pages.

Question 1

Suppose that the n observations $x_1, x_2, \ldots, x_n$ are recorded, where $n = 25$ and the following summary statistics are computed:

- $x_{(1)} = -1$ (the smallest observation)
- $q_{0.25} = 1$ (the lower quartile)
- $m = q_{0.5} = 2$ (the median)
- $q_{0.75} = 4$ (the upper quartile)
- $x_{(n)} = 7$ (the largest observation)
- $\sum_{i=1}^{n} x_i = 60$
- $\sum_{i=1}^{n} x_i^2 = 264$

Showing **all working** and justifying **any formulae** used:

(i) **(1 point)** Compute the sample mean.

(ii) **(1 point)** Compute the range.

(iii) **(1 point)** Compute the interquartile range.

(iv) **(2 points)** Compute the sample variance.

Solution to Question 1

Part (i)

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} = \frac{1}{25}(60) = \frac{12}{5} = 2.4.$$

[1 point for correct answer if working is shown (no working, no mark).]

Part (ii)

$$R = x_{(n)} - x_{(1)} = 7 - (-1) = 8.$$

[1 point for correct answer if working is shown (no working, no mark).]

Part (iii)

$$\text{IQR} = q_{0.75} - q_{0.25} = 4 - 1 = 3.$$

[1 point for correct answer if working is shown (no working, no mark).]

Part (iv) Using the identity proved in Exercise 1.2.5,

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x}) = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - n(\overline{x})^2\right] = \frac{1}{24}\left[264 - 25\cdot(\frac{12}{5})^2\right] \\
&= \frac{1}{24}\left[264 - (12)^2\right] \\
&= \frac{1}{24}\left[264 - 144\right] = \frac{1}{24}\left[120\right] \\
&= 5
\end{aligned}$$

[1 point for using identity and 1 point for correct answer if working is shown (no working, no mark).]

Note that these summary statistics were not invented; they are the summary statistics of the data set:

$$\{-1, -1, 0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 4, 4, 5, 6, 6, 6, 7\}$$

Question 2

Suppose that the random variables $X_1, X_2, \ldots, X_n$ are independent and each follows the same distribution which has mean μ and variance σ^2. Recall the definitions of the sample mean and sample variance

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2,$$

where $\overline{X}$ is an estimator of μ and S^2 is an estimator of σ^2. Suppose it is known that for this distribution,

$$\mathrm{Var}\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] = 2(n-1)\sigma^4.$$

Stating any results used from the notes:

(i) **(1 point)** Show that $b_{\sigma^2}(S^2) = 0$, where $b_{\sigma^2}(S^2)$ is the bias of S^2.

(ii) **(2 point)** Prove that the mean squared error of S^2 is $\dfrac{2\sigma^4}{n-1}$.

(iii) **(1 point)** Suppose that one defines $W = \frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ as an alternative estimator of σ^2. Compute $b_{\sigma^2}(W)$, the bias of W.

(iv) **(1 point)** Compute $\mathrm{Var}\,(W)$.

(v) **(2 point)** Compute the mean squared error of W, and show that it is less than the mean squared error of S^2.

(vi) **(2 points)** Which estimator would you prefer to use to estimate σ^2? Justify your answer, stating the advantages and disadvantages of both estimators.

Solution to Question 2

Part (i) From Proposition 1.2.6 in the notes, $\mathrm{E}[S^2] = \sigma^2$. Therefore,

$$b_{\sigma^2}(S^2) = \mathrm{E}[S^2] - \sigma^2 = \sigma^2 - \sigma^2 = 0.$$

[1 point for using definition correctly AND mentioning a result in the notes.]

Part (ii)

Method 1: First use the assumption above to compute the variance of S^2 as

$$\begin{aligned}\mathrm{Var}\left(S^2\right) &= \mathrm{Var}\left(\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)\\ &= \frac{1}{(n-1)^2}\mathrm{Var}\left(\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right)\\ &= \frac{1}{(n-1)^2}\left(2(n-1)\sigma^4\right)\\ &= \frac{2\sigma^4}{n-1}.\end{aligned}$$

Now, using Theorem 1.5.24 from lectures, and the fact given in (i) that $b_{\sigma^2}(S^2)=0$, the mean squared error of S^2 is computed as

$$\left(b_{\sigma^2}(S^2)\right)^2+\mathrm{Var}\left(S^2\right)=(0)^2+\frac{2\sigma^4}{n-1}=\frac{2\sigma^4}{n-1}.$$

Method 2: Recalling that $\mathrm{E}[S^2]=\sigma^2$, the mean squared error is computed directly as

$$\mathrm{E}[(S^2-\sigma^2)^2]=\mathrm{Var}\left(S^2\right),$$

and one computes $\mathrm{Var}\left(S^2\right)=\frac{2\sigma^4}{n-1}$ as in Method 1, and so the mean squared error is

$$\mathrm{E}[(S^2-\sigma^2)^2]=\frac{2\sigma^4}{n-1}.$$

Method 3: One expands the definition of the mean squared error (using the linearity of expectation)

$$\begin{aligned}\mathrm{E}[(S^2-\sigma^2)^2] &= \mathrm{E}[(S^2)^2-2\sigma^2S^2+\sigma^4]=\mathrm{E}[(S^2)^2]-2\sigma^2\mathrm{E}[S^2]+\mathrm{E}[\sigma^4]\\ &= \mathrm{E}[(S^2)^2]-\sigma^4\end{aligned}$$

and then one computes $\mathrm{E}[(S^2)^2]$ using

$$\mathrm{E}[(S^2)^2]=\mathrm{Var}\left(S^2\right)+(\mathrm{E}[S^2])^2=\mathrm{Var}\left(S^2\right)+(\sigma^2)^2,$$

and then this becomes the same as Method 2.

[Method 1:1 point for computing the variance correctly and 1 point for using formula for mean squared error from the notes correctly.]

[Method 2/3: 1 point for applying the mean squared error definition correctly and 1 point for computing $\mathrm{Var}\left(S^2\right)$ or $\mathrm{E}[(S^2)^2]$ correctly.]

Part (iii)

Recalling from Proposition 1.2.6 in the notes that $\mathrm{E}[S^2] = \sigma^2$, and noticing that $W = \dfrac{n-1}{n+1}S^2$,

$$\begin{aligned}
b_{\sigma^2}(W) &= \mathrm{E}[W] - \sigma^2 = \mathrm{E}[\tfrac{n-1}{n+1}S^2] - \sigma^2 = \tfrac{n-1}{n+1}\mathrm{E}[S^2] - \sigma^2 = \tfrac{n-1}{n+1}\sigma^2 - \sigma^2 \\
&= \left(\tfrac{n-1}{n+1} - 1\right)\sigma^2 \\
&= \left(\tfrac{n-1-(n+1)}{n+1}\right)\sigma^2 \\
\Rightarrow b_{\sigma^2}(W) &= \frac{-2}{n+1}\sigma^2
\end{aligned}$$

[1 point for computing the bias correctly.]

Part (iv)

Method 1:

$$\begin{aligned}
\mathrm{Var}\,(W) &= \mathrm{Var}\left(\frac{1}{n+1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\
&= \frac{1}{(n+1)^2}\mathrm{Var}\left(\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right) \\
&= \frac{1}{(n+1)^2}2(n-1)\sigma^4 \\
&= \frac{2(n-1)}{(n+1)^2}\sigma^4
\end{aligned}$$

Method 2 (very similar to Method 1):

$$\begin{aligned}
\mathrm{Var}\,(W) &= \mathrm{Var}\left(\tfrac{n-1}{n+1}S^2\right) \\
&= \left(\frac{n-1}{n+1}\right)^2 \mathrm{Var}\left(S^2\right) \\
&= \frac{(n-1)^2}{(n+1)^2}\frac{2\sigma^4}{n-1} \\
&= \frac{2(n-1)}{(n+1)^2}\sigma^4
\end{aligned}$$

[1 point for computing the variance correctly (either method).]

Part (v)

The mean squared error can be computed using Theorem 1.5.24 and the bias and variance from Parts (iii) and (iv):

$$\begin{aligned}(b_{\sigma^2}(W))^2 + \mathrm{Var}\,(W) \\ &= \left(\frac{-2}{n+1}\sigma^2\right)^2 + \frac{2(n-1)}{(n+1)^2}\sigma^4 \\ &= \frac{4}{(n+1)^2}\sigma^4 + \frac{2(n-1)}{(n+1)^2}\sigma^4 \\ &= \frac{4+2n-2}{(n+1)^2}\sigma^4 \\ &= \frac{2(n+1)}{(n+1)^2}\sigma^4 \\ &= \frac{2}{n+1}\sigma^4\end{aligned}$$

Since for any value of n the inequality $\frac{2}{n+1} < \frac{2}{n-1}$ is true, then

$$\mathrm{MSE}(W) = \frac{2\sigma^4}{n+1} < \frac{2\sigma^4}{n-1} = \mathrm{MSE}(S^2)$$

and so the mean squared error of W is less than the mean squared error of S^2.

[1 point for using the theorem for the mean squared error correctly and subsituting in values from Parts (iii) and (iv); 1 point for getting the value correct AND these values is incorrect.]

Part (vi)

Does not matter which estimator is preferred, both can be justified.

The advantage of S^2 is that it is unbiased, but the disadvantage is that it has a higher mean squared error.

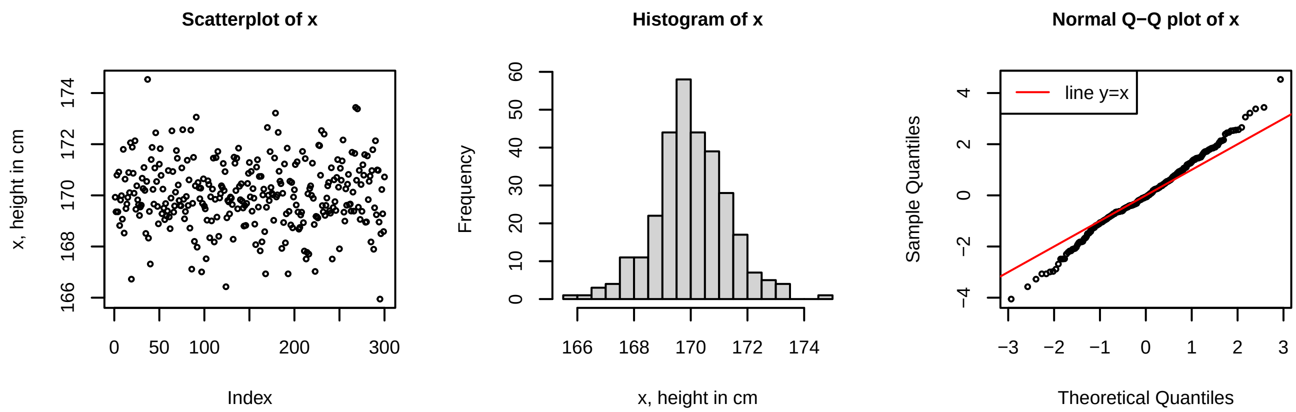
The advantage of W is that it has a lower mean squared error but the disadvantage is that it is biased.

[1 point for the advantages/disadvantages of S^2 and 1 point for the advantages/disadvantages of W.]

Question 3

Suppose $X_1, X_2, \ldots, X_n$ are the random variables representing the heights of the $n = 300$ students in a particular module, measured in cm. These random variables are observed as $x_1, x_2, \ldots, x_n$, which are plotted below in (a) a scatterplot of the data, (b) a histogram of the data, (c) a Q-Q plot of the data after being standardised by the sample mean and variance.

(2 points) Do these plots suggest that $X_1, X_2, \ldots, X_n$ follow a normal distribution? Provide justification for your answer.



Solution to Question 3

Although the scatterplot and histogram may suggest the data is normal, it is not possible to tell conclusively from these plots.

The Q-Q plot shows that the sample quantiles do not agree with the theoretical quantiles (do not lie along the line $y = x$) for a large proportion of the quantiles. Therefore, this suggests that the random variables which have been observed do not follow a normal distribution.

[1 point for correct conclusion and 1 point for correct justification. No need to mention scatterplot/histogram, just need to cite evidence from Q-Q plot.]

Note: if it is argued that most of the points in the Q-Q plot lie along the line $y = x$, and so the data seems to be normally distributed, the answer can be accepted.

Question 4

Suppose $X_1, X_2, \ldots, X_n$, where $n = 20$, are independent and identically distributed random variables representing the heights of n students measured in cm. Suppose that for $i = 1, 2, \ldots, n$, each X_i is assumed to follow a normal distribution with $\mathrm{E}(X_i) = \theta$ and $\mathrm{Var}(X_i) = \sigma^2$, where θ is unknown but σ^2 is known to be $\sigma^2 = 15$.

Now suppose that the heights of the students are measured as $x_1, x_2, \ldots, x_n$, and from these measurements it is computed that $\overline{x} = 182$cm.

(i) **(3 points)** Given the assumptions and the data above, construct a 99% confidence interval for the unknown mean θ. You may find Table 1 below to be helpful.

(ii) **(1 point)** If the variance σ^2 were unknown, how else could you construct the confidence interval for θ?

Table 1: Selected values of z for $\mathrm{P}(Z < z)$, where Z has a standard normal distribution

z	$\mathrm{P}(Z < z)$
1.281	0.900
1.645	0.950
1.960	0.975
2.326	0.990
2.576	0.995

Solution to Question 4

Part (i) Since each $X_i \sim \mathrm{N}\left(\theta, \sigma^2\right)$, by Corollary 3.1.3 in the notes,

$$\overline{X} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right).$$

Defining $Z = \dfrac{\theta - \overline{X}}{\sigma/\sqrt{n}} \Rightarrow Z \sim \mathrm{N}(0,1)$, then from Table 1,

$$\begin{aligned}
&\mathrm{P}(Z < 2.576) = 0.995\\
&\mathrm{P}(Z < -2.576) = 0.005\\
\Rightarrow &\mathrm{P}(-2.576 < Z < 2.576) = 0.99, \qquad \text{(using fact that } \mathrm{P}(Z = -2.576) = 0\text{).}
\end{aligned}$$

Then we have

$$\begin{aligned}
&\mathrm{P}(-2.576 < Z < 2.576) = 0.99\\
\Rightarrow &\mathrm{P}\left(-2.576 < \frac{\theta - \overline{X}}{\sigma/\sqrt{n}} < 2.576\right) = 0.99\\
\Rightarrow &\mathrm{P}\left(\overline{X} - 2.576\frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + 2.576 \cdot \frac{\sigma}{\sqrt{n}}\right) = 0.99.
\end{aligned}$$

Therefore, since $n = 20$, $\sigma^2 = 15$ and $\overline{x} = 182$, the 99% confidence interval for θ is

$$\left(182 - 2.576 \cdot \frac{\sqrt{15}}{\sqrt{20}}, 182 + 2.576 \cdot \frac{\sqrt{15}}{\sqrt{20}}\right) = \left(182 - 2.576 \cdot \frac{\sqrt{3}}{2}, 182 + 2.576 \cdot \frac{\sqrt{3}}{2}\right) = (179.77, 184.23).$$

[1 point for critical value, 1 point for quoting/deriving correct formula for interval, 1 point for substituting in correct values (no need to work out final answer with calculator).]

Part (ii) If one had the sample variance of the data s^2, one could use Student's t distribution to construct the confidence interval.

[1 point for mentioning Student's t-distribution.]

Total: 20 points