Note that there are FIVE questions split across TWO pages.

Provide justification for all solutions.

Question 1

Suppose that the random variables $X_1, X_2, \ldots, X_n$, with $n > 3$, are independent and each follows the same distribution which has mean μ and variance σ^2. We decide to define $\widehat{\Theta}$, an estimator of the mean μ, as:

$$\widehat{\Theta} = \frac{1}{n-3}\sum_{i=1}^{n} X_i.$$

Clearly stating any results or properties used:

(i) **(2 points)** Compute $b_\mu(\widehat{\Theta})$, the bias of $\widehat{\Theta}$.

(ii) **(4 points)** Compute the mean squared error of $\widehat{\Theta}$.

Solution to Question 1

Part (i)

Using the linearity of expectation,

$$\begin{aligned}
b_\mu(\widehat{\Theta}) &= \mathrm{E}[\widehat{\Theta} - \mu] \\
&= \mathrm{E}[\widehat{\Theta}] - \mu \\
&= \mathrm{E}\left[\frac{1}{n-3}\sum_{i=1}^{n} X_i\right] - \mu \\
&= \frac{1}{n-3}\sum_{i=1}^{n} \mathrm{E}\left[X_i\right] - \mu \\
&= \frac{1}{n-3}\sum_{i=1}^{n} \mu - \mu \\
&= \frac{n\mu}{n-3} - \mu \qquad (1) \\
&= \frac{n\mu - (n-3)\mu}{n-3} \\
&= \frac{n\mu - n\mu + 3\mu}{n-3} \\
&= \frac{3\mu}{n-3}.
\end{aligned}$$

[2 marks]

- **1 mark for using the correct definition of the bias,**
- **1 mark for the correct calculation up to Equation** (1)**.**

Part (ii)

By Theorem 1.5.24 in the notes, the mean squared error $\mathrm{E}\left[(\widehat{\Theta}-\mu)^2\right]$ can be expressed in terms of the bias and variance of $\widehat{\Theta}$:

$$\mathrm{E}\left[(\widehat{\Theta}-\mu)^2\right] = \left[b_\mu(\widehat{\Theta})\right]^2 + \mathrm{Var}\left[\widehat{\Theta}\right].$$

The bias was computed in Part (a) to be $b_\mu(\widehat{\Theta}) = \frac{3\mu}{n-3}$. To compute the variance:

$$\begin{aligned}
\mathrm{Var}\left[\widehat{\Theta}\right] &= \mathrm{Var}\left[\frac{1}{n-3}\sum_{i=1}^{n} X_i\right] = \frac{1}{(n-3)^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i\right] \qquad \text{(property of the variance)}\\
&= \frac{1}{(n-3)^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] \qquad \text{(since the } X_i \text{ are independent)}\\
&= \frac{1}{(n-3)^2}\sum_{i=1}^{n}\sigma^2\\
&= \frac{n\sigma^2}{(n-3)^2}.
\end{aligned}$$

Then, substituting in values for the bias and variance,

$$\mathrm{E}\left[(\widehat{\Theta}-\mu)^2\right] = \left(\frac{3\mu}{n-3}\right)^2 + \frac{n\sigma^2}{(n-3)^2} = \frac{9\mu^2+n\sigma^2}{(n-3)^2}.$$

[4 marks]

- **2 marks for computing variance correctly; if independence not mentioned, 1 mark off,**
- **2 marks for referencing theorem and substituting values in correctly,**
- **No mark is awarded for final MSE answer, so can carry over any errors from Part (a).**

Alternate Part 2(ii) solution:

A direct solution is also possible:

$$\mathrm{E}\left[(\widehat{\Theta}-\mu)^2\right] = \mathrm{E}\left[\widehat{\Theta}^2 - 2\widehat{\Theta}\mu + \mu^2\right] = \mathrm{E}\left[\widehat{\Theta}^2\right] - 2\mu\mathrm{E}[\widehat{\Theta}] + \mu^2,$$

using the linearity of expectation. Using the identity (Exercise 1.1.5 in the notes) $\mathrm{E}[X^2] = \mathrm{Var}[X] + (\mathrm{E}[X])^2$,

$$\mathrm{E}\left[(\widehat{\Theta}-\mu)^2\right] = \mathrm{Var}[\widehat{\Theta}] + \left(\mathrm{E}[\widehat{\Theta}]\right)^2 - 2\mu\mathrm{E}[\widehat{\Theta}] + \mu^2, \tag{2}$$

One computes $\mathrm{E}[\widehat{\Theta}]$ as in Part (a) to be

$$\mathrm{E}[\widehat{\Theta}] = \mathrm{E}\left[\frac{1}{n-3}\sum_{i=1}^{n} X_i\right] = \frac{1}{n-3}\sum_{i=1}^{n}\mathrm{E}\left[X_i\right] = \frac{n\mu}{n-3},$$

and $\mathrm{Var}[\widehat{\Theta}]$ is computed as above (using the independence of the X_i) to be

$$\mathrm{Var}[\widehat{\Theta}] = \frac{n\sigma^2}{(n-3)^2}.$$

Then

$$\begin{aligned}
\mathrm{E}\left[(\widehat{\Theta}-\mu)^2\right] &= \frac{n\sigma^2}{(n-3)^2} + \left(\frac{n\mu}{n-3}\right)^2 - 2\mu\left(\frac{n\mu}{n-3}\right) + \mu^2 \\
&= \frac{n\sigma^2}{(n-3)^2} + \frac{n^2\mu^2 - 2n(n-3)\mu^2 + (n-3)^2\mu^2}{(n-3)^2} \\
&= \frac{n\sigma^2}{(n-3)^2} + \frac{n^2\mu^2 - 2n^2\mu^2 + 6n\mu^2 + (n^2-6n+9)\mu^2}{(n-3)^2} \\
&= \frac{n\sigma^2}{(n-3)^2} + \frac{9\mu^2}{(n-3)^2},
\end{aligned}$$

as in the above solution.

[4 marks]

- **2 marks for computing variance correctly (if independence not mentioned, 1 mark off),**
- **1 mark for expanding formula up to Equation (2) correctly,**
- **1 mark computing expectation correctly. Incorrect final algebraic manipulation not penalised.**

Question 2

Suppose the random variable X follows an unknown distribution but is known to only take values in the range $[2, 7]$.

(2 points) Find values a and b such that $a \leq \mathrm{Var}[X] \leq b$ and the interval width, $b - a$, is minimised.

Solution to Question 2

First, we have $a = 0$ since $\mathrm{Var}[X] \geq 0$; $(X - \mathrm{E}[X])^2 \geq 0$ so by definition $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2] \geq 0$.

Next, Proposition 1.1.6 in the notes states that for a random variable X bounded on the interval $[c, d]$, $\mathrm{Var}[X] \leq \frac{(d-c)^2}{4}$; this bound is tight by Problem Sheet 1, Question 2(d).

Therefore $\mathrm{Var}[X] \leq \frac{(7-2)^2}{4} = \frac{25}{4}$, and so $b = \frac{25}{4}$.

[2 marks]

- **1 mark for correct a,**
- **1 mark for correct b if the proposition has clearly been used.**

Question 3

Suppose that $X_1, X_2, \ldots, X_n$ are random variables that represent a sample of transaction values for a particular bank account. These transaction values are assumed each independently follow the same distribution with unknown mean μ and variance σ^2.

Suppose further that we observe these random variables as $x_1, x_2, \ldots, x_n$ (positive values represent money received, and negative values represent payments), for $n = 18$, and their sample mean is computed to be $\bar{x} = £4,895$. We decide to assume the standard deviation in transaction values is less than £100.

(i) **(4 points)** Given the data above, construct a 90% confidence interval for the unknown mean μ.

(ii) **(1 point)** Suppose we are now told that the data follow a Cauchy distribution. Would you still be able to follow the same approach as in (a)? Justify your answer.

Solution to Question 3

Part (i):

Chebyshev's inequality gives us, for a random variable X with $\mathrm{E}[X] = \mu$ and $\mathrm{Var}[X] = \sigma^2$, and any $k > 0$,

$$\mathrm{P}\left(|X-\mu| \geq k\sigma\right) \leq \frac{1}{k^2} \qquad \Rightarrow \mathrm{P}\left(|X-\mu| < k\sigma\right) \geq 1 - \frac{1}{k^2} \qquad \Rightarrow \mathrm{P}\left(X - k\sigma < \mu < X + k\sigma\right) \geq 1 - \frac{1}{k^2}.$$

Setting $1 - \frac{1}{k^2} = 0.9$, solving for $k > 0$ one obtains $\frac{1}{k^2} = 0.1 = \frac{1}{10} \Rightarrow k = \sqrt{10}$.

The random variables $X_1, X_2, \ldots, X_n$ are assumed to be independent with unknown mean μ and (unknown) variance σ^2, so defining $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, following Proposition 1.2.6 in the notes, $\mathrm{E}[\overline{X}] = \mu$ and $\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$. Applying Chebyshev's inequality to $\overline{X}$ with $k = \sqrt{10}$ and $n = 18$,

$$\mathrm{P}\left(\overline{X} - k\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + k\frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2} \qquad \Rightarrow \mathrm{P}\left(\overline{X} - \sqrt{10}\frac{\sigma}{\sqrt{18}} < \mu < \overline{X} + \sqrt{10}\frac{\sigma}{\sqrt{18}}\right) \geq 0.9.$$

Now, we are given that $\overline{x} = 4895$, and we decide to assume that $\sigma < 100$, which implies $-100 < -\sigma$, so

$$\left(4895 - \sqrt{10}\frac{\sigma}{\sqrt{18}}, 4895 + \sqrt{10}\frac{\sigma}{\sqrt{18}}\right) \subset \left(4895 - \sqrt{10}\frac{100}{\sqrt{18}}, 4895 + \sqrt{10}\frac{100}{\sqrt{18}}\right),$$

and so $\left(4895 - \sqrt{10}\frac{100}{\sqrt{18}}, 4895 + \sqrt{10}\frac{100}{\sqrt{18}}\right)$ is a 90% confidence interval for μ. This can also be written as $\left(4895 - \frac{100\sqrt{5}}{3}, 4895 + \frac{100\sqrt{5}}{3}\right)$.

[4 marks]

- **1 mark for quoting Chebyshev's inequality (need to quote, but do not need to state),**
- **1 mark for finding $k = \sqrt{10}$,**
- **1 mark for using $\overline{X}$ in Chebyshev's inequality with correct mean and variance,**
- **1 mark for correctly using inequality $\sigma < 100$ to construct interval.**

Part (ii):

We know from Problem Sheet 8, Question 1 that the mean and variance of the Cauchy distribution are undefined.

Therefore, we cannot use the same approach as in Part (a) to obtain a confidence interval for the mean μ, since the mean does not exist. Moreover, we would not be able to assume that $\sigma < 100 \Rightarrow \sigma^2 < 10000$, since the variance is undefined, and so could not use a value for σ in Chebyshev's inequality.

[1 mark]

- **1 mark for a justification that uses the fact that the mean and/or variance of Cauchy distribution does not exist.**

Question 4

Consider the sample of the following 11 values:

$$\{5, 12, 7, 4, 10, 8, 11, 15, 17, 16, 18\}.$$

Showing **all working**:

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

(ii) **(2 point)** Compute the upper and lower quartiles.

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

Note: values provided without working will receive 0 points.

Solution to Question 4

Before doing the individual parts, it will be useful to order the values in non-decreasing order:

$$\{4, 5, 7, 8, 10, 11, 12, 15, 16, 17, 18\}.$$

Part (i):

There are $n = 11$ values, so the median is the $(11 + 1)/2 = 6$th smallest value, which is 11.

[1 mark]

- **1 mark for correct value and justification**

Part (ii):

From Part (i), the median is the 6th smallest value. Therefore the lower quartile is the $(1 + 6)/2 = 3.5$th smallest value. The 3rd smallest value is 7, and the 4th smallest value is 8, so any value in the interval $[7, 8]$ could be chosen as the lower quartile. By convention, the average 7.5 is usually chosen.

Similarly, the upper quartile is the $(6 + 11)/2 = 8.5$th smallest value. The 8th smallest value is 15, and the 9th smallest value is 16, so any value in the interval $[15, 16]$ could be chosen as the upper quartile. By convention, the average 15.5 is usually chosen.

[2 marks]

- **1 mark for correct value and justification for the lower quartile. Any value in the interval $[7, 8]$ can be given.**
- **1 mark for correct value and justification for the upper quartile. Any value in the interval $[15, 16]$ can be given.**

Part (iii):

The interquartile range (IQR) is the difference between the upper quartile value and the lower quartile value. If these values in Part (ii) were computed as 7.5 and 15.5, then the IQR is $15.5 - 7.5 = 8$.

[1 mark]

- **1 mark for correct value and justification based on values in Part (ii).**

Question 5

For each of the following scenarios, which plot would be the best choice for an exploratory data analysis? In this question you can provide the answer without a full explanation.

(i) **(1 point)** The data is categorical and one wishes to display the proportions of each category recorded.

(ii) **(1 point)** The data is continuous and one wishes to see if it follows a normal distribution.

(iii) **(1 point)** The data is continuous and one wishes to visualise the spread of the data and identify any outliers.

Solution to Question 5

Part (i)

The pie chart.

[1 mark]

- **1 mark for correct answer.**

Part (ii)

The quantile-quantile (Q-Q) plot.

[1 mark]

- **1 mark for correct answer.**

Part (iii)

The box plot.

[1 mark]

- **1 mark for correct answer.**

Total: 20 points