Provide justification for all solutions.

Question 1

(i) **(2 points)** State Markov's inequality.

(ii) **(2 points)** Using Markov's inequality, prove Chebyshev's inequality:

If X is a random variable with finite mean μ and finite variance σ^2, then for all $c > 0$,

$$\mathrm{P}(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

(iii) **(3 points)** Suppose Y is a random variable following an unknown distribution with unknown mean μ and unknown variance σ^2, although we will assume that μ and σ^2 are both finite. Showing all working, compute a lower bound for the probability that Y is within 3 standard deviations of its mean.

Solution to Question 1

Part (i)

Markov's inequality states: If a random variable Y can only take nonnegative values (i.e. $Y \geq 0$), then for all $a > 0$, the following is true: $\mathrm{P}(Y \geq a) \leq \dfrac{\mathrm{E}[Y]}{a}$.

- **2 points for correct statement, need to have conditions on Y and a.**

Part (ii)

Fix a value $c > 0$. Although X is not necessarily nonnegative, $(X - \mu)^2 \geq 0$ is a nonnegative random variable. Therefore, applying Markov's inequality using $a = c^2$, and $Y = (X - \mu)^2$, one has

$$\mathrm{P}((X - \mu)^2 \geq c^2) \leq \frac{\mathrm{E}[(X - \mu)^2]}{c^2}.$$

Note that the event $\{(X - \mu)^2 \geq c^2\}$ is equivalent to the event $\{|X - \mu| \geq c\}$, and so the probabilities of these events occuring is equal. Note also that $\mathrm{E}[(X - \mu)^2] = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{Var}[X] = \sigma^2$. Using these two facts, we then have

$$\mathrm{P}(|X - \mu| \geq c) = \mathrm{P}((X - \mu)^2 \geq c^2) \leq \frac{\mathrm{E}[(X - \mu)^2]}{c^2} = \frac{\sigma^2}{c^2},$$

which proves Chebyshev's inequality.

- **1 point for using Markov's inequality correctly.**
- **1 point for noting BOTH that (1) the events $\{|X - \mu| \geq c\}$ and $\{(X - \mu)^2 \geq c^2\}$ are equal AND (2) that $\mathrm{E}[(X - \mu)^2] = \sigma^2$.**

Part (iii)

We start with Chebyshev's inequality, as stated in Part (b), using Y:

$$\mathrm{P}(|Y - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

However, there is an equivalent form; if we fix any $k > 0$ and let $c = k\sigma > 0$, then

$$\begin{aligned} \mathrm{P}(|Y - \mu| \geq k\sigma) &\leq \frac{\sigma^2}{k^2\sigma^2} \\ \Rightarrow \mathrm{P}(|Y - \mu| \geq k\sigma) &\leq \frac{1}{k^2}. \end{aligned}$$

Moreover, if we let A be the event $\{|Y-\mu| \geq k\sigma\}$, then A^c is the event $\{|Y-\mu| < k\sigma\}$, and so since $1-\mathrm{P}(A) = \mathrm{P}(A^c)$,

$$\begin{aligned}
&\mathrm{P}(|Y-\mu| \geq k\sigma) \leq \frac{1}{k^2} \\
\Rightarrow &1-\mathrm{P}(|Y-\mu| \geq k\sigma) \geq 1-\frac{1}{k^2} \\
\Rightarrow &\mathrm{P}(|Y-\mu| < k\sigma) \geq 1-\frac{1}{k^2}.
\end{aligned}$$

If we let $k=3$, then

$$\mathrm{P}(|Y-\mu| < 3\sigma) \geq 1-\frac{1}{3^2} = \frac{8}{9} \approx 0.889.$$

- **2 points for providing justification for** $\mathrm{P}(|Y-\mu| < k\sigma) \geq 1-\frac{1}{k^2}$.
- **1 point for getting lower bound** $\frac{8}{9}$.

Question 2

(2 points) Suppose you are given a data set containing the ages (in years, i.e. positive integers) of the approximately $n = 1000$ students in the Department of Mathematics at Imperial. State which plot you would use to visualise the distribution of the ages, and provide motivation for your choice.

Solution to Question 2

Since the ages of students, as years, are discrete (positive integers) and likely to be mostly in the range 17 to 26, a good choice would be to use the bar chart.

If we were concerned that there are students younger than 17 or older than 26 then we could, for example, also use the categories '17 and under' and '26 and over', in addition to the categories '18', '19', '20', '21', '22', '23', '24', '25'.

- **1 point for bar chart (pie chart also acceptable).**
- **1 point for justification of/mentioning discrete data.**

Alternative answers:

- 'Histogram' would receive a maximum of **1 point**.
- 'Box plot' would receive a maximum of **1 point**.
- All other answers: **0 points**.

Question 3

(5 points) Suppose $X_1, X_2, \ldots, X_n$, where $n = 100$, are independent and identically distributed random variables following a normal distribution with mean μ and variance σ^2. The value of μ is unknown, but σ^2 is known to be $\sigma^2 = 10$. We will also assume that the value of μ is finite. Suppose we observe $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ as $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and that we compute the sample mean of these observations to be $\overline{x} = 4.2$. Using Table 1 below, construct a 95% confidence interval for μ, providing full justification.

Solution to Question 3

Since the random variables are independent and normally distributed, from a result in lectures $\overline{X} \sim \mathrm{N}\left(\theta, \frac{\sigma^2}{n}\right)$. If we define the random variable

$$Z = \frac{\theta - \overline{X}}{\sigma/\sqrt{n}},$$

then $Z \sim \mathrm{N}(0, 1)$.

For any significance level α, if we define z_α to be the value such that $\mathrm{P}(Z < z_\alpha) = \alpha$, then

$$\begin{aligned}
&\mathrm{P}\left(Z < z_{1-\alpha/2}\right) = 1 - \alpha/2, \\
&\mathrm{P}\left(Z < z_{\alpha/2}\right) = \alpha/2, \\
\Rightarrow &\mathrm{P}\left(z_{\alpha/2} < Z < z_{1-\alpha/2}\right) = 1 - \alpha. \\
\Rightarrow &\mathrm{P}\left(z_{\alpha/2} < \frac{\theta - \overline{X}}{\sigma/\sqrt{n}} < z_{1-\alpha/2}\right) = 1 - \alpha \\
\Rightarrow &\mathrm{P}\left(\overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.
\end{aligned}$$

To construct a 95% confidence interval, $1 - \alpha = 0.95 \Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025 \Rightarrow 1 - \alpha/2 = 0.975$.

Using Table 1, we find $z_{0.975} = 1.96$, and therefore by symmetry of the normal distribution, $z_{0.025} = -1.96$. Since $\sigma^2 = 10$, then $\sigma = \sqrt{10}$. Since $\mathbf{X}$ is observed as $\mathbf{x}$ and $\overline{x} = 4.2$, a 95% confidence interval is therefore

$$\begin{aligned}
&\left(\overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) \\
&= \left(4.2 - 1.96 \cdot \frac{\sqrt{10}}{\sqrt{100}}, 4.2 + 1.96 \cdot \frac{\sqrt{10}}{\sqrt{100}}\right) \\
&= \left(4.2 - \frac{1.96}{\sqrt{10}}, 4.2 + \frac{1.96}{\sqrt{10}}\right).
\end{aligned}$$

The interval can be computed to two decimal places as $(3.58, 4.82)$, but it is fine to leave it in the above form.

- **1 point for defining Z and showing it is $\mathrm{N}(0, 1)$.**
- **1 point for getting the correct format of the confidence interval $\mathrm{P}\left(\overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + z_{1-\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right)$, and 1 point for some justification (only directly stating the equation will lost a point).**
- **1 point for getting the correct threshold from the table.**
- **1 point for substituting in correctly (remembering to use $\sqrt{n}$ and σ, not σ^2, etc.)**

Alternative solution to Question 3

One could use Chebyshev's inequality to obtain, for all $k > 0$,

$$\mathrm{P}\left(\overline{X} + k \cdot \frac{\sigma}{\sqrt{n}} < \theta < \overline{X} + k \cdot \frac{\sigma}{\sqrt{n}}\right) \geq 1 - \frac{1}{k^2}.$$

Solving for k:

$$\begin{aligned} 0.95 &= 1 - \frac{1}{k^2} \\ \Rightarrow \frac{1}{k^2} &= 0.05 = \frac{1}{20} \\ \Rightarrow k &= \sqrt{20} = 2\sqrt{5}. \end{aligned}$$

Since $\sqrt{n} = 10$ and $\sigma = \sqrt{10} = \sqrt{2}\sqrt{5}$, then the realised confidence interval is

$$\begin{aligned} &\left(\overline{x} + k \cdot \frac{\sigma}{\sqrt{n}}, \overline{x} + k \cdot \frac{\sigma}{\sqrt{n}}\right) \\ &= \left(4.2 - 2\sqrt{5} \cdot \frac{\sqrt{2}\sqrt{5}}{10}, 4.2 + 2\sqrt{5} \cdot \frac{\sqrt{2}\sqrt{5}}{10}\right) \\ &= \left(4.2 - 2\sqrt{2} \cdot \frac{5}{10}, 4.2 + 2\sqrt{2} \cdot \frac{5}{10}\right) \\ &= \left(4.2 - \sqrt{2}, 4.2 + \sqrt{2}\right). \end{aligned}$$

This can be computed to 2 decimal places as $(2.79, 5.61)$, but it is fine to leave it in the above form.

(Note how the interval is wider than the first solution using the normality properties.)

- **1 point for getting the correct form of Chebyshev's inequality.**
- **1 point for solving for $k = 2\sqrt{5}$.**
- **1 point for correctly substituting in the values.**
- **Maximum of 3 marks for this Chebyshev solution; this is not the best confidence interval using all the information provided.**

Question 4

Suppose that the $n > 2$ random variables $X_1, X_2, \ldots, X_n$ are independent and each follows the same distribution which has finite mean μ and finite variance σ^2. Furthermore, assume it is known that the fourth (raw) moment is $\mathrm{E}[X_i^4] = \mu^4 + 4\mu^2\sigma^2 + 2\sigma^4$ for $i = 1, 2, \ldots, n$. We decide to define $\widehat{\Theta}$, an estimator of the variance σ^2, as:

$$\widehat{\Theta} = \frac{1}{n}\sum_{i=1}^{n} X_i^2.$$

Clearly stating any results or properties used:

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

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

Solution to Question 4

Part (i)

Since $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2$, from an exercise in the notes, $\mathrm{E}[X_i^2] = \mu^2 + \sigma^2$.

The bias is defined as $b_{\sigma^2}(\widehat{\Theta}) = \mathrm{E}\left[\widehat{\Theta}\right] - \sigma^2$.

Using the linearity of expectation,

$$\begin{aligned} b_{\sigma^2}(\widehat{\Theta}) &= \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\left[X_i^2\right] - \sigma^2 \\ &= \frac{1}{n}\sum_{i=1}^{n} (\mu^2 + \sigma^2) - \sigma^2 \\ &= \frac{1}{n} \cdot n(\mu^2 + \sigma^2) - \sigma^2 \\ &= (\mu^2 + \sigma^2) - \sigma^2 \\ \Rightarrow b_{\sigma^2}(\widehat{\Theta}) &= \mu^2. \end{aligned}$$

[2 marks]

- **2 marks for correct calculation.**

Part (ii)

From a theorem in the notes,

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

An exercise in the notes gives us that for any random variable Y,

$$\begin{aligned} \mathrm{Var}[Y] &= \mathrm{E}[Y^2] - (\mathrm{E}[Y])^2 \\ \Rightarrow \mathrm{Var}[Y^2] &= \mathrm{E}[Y^4] - (\mathrm{E}[Y^2])^2. \end{aligned}$$

Computing the variance, using the fact that the X_i are independent, and so the X_i^2 are independent,

$$\begin{aligned}
\mathrm{Var}[\widehat{\Theta}] &= \mathrm{Var}\left[\frac{1}{n}\sum_{i=1}^{n} X_i^2\right] \\
&= \frac{1}{n^2}\mathrm{Var}\left[\sum_{i=1}^{n} X_i^2\right] \qquad \text{(property of the variance)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[X_i^2\right] \qquad \text{(property of the variance, since } X_i^2 \text{ are independent)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left(\mathrm{E}[X_i^4] - (\mathrm{E}[X_i^2])^2\right) \qquad \text{(applying exercise in notes)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left(\mu^4 + 4\mu^2\sigma^2 + 2\sigma^4 - (\mu^2+\sigma^2)^2\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left(\mu^4 + 4\mu^2\sigma^2 + 2\sigma^4 - (\mu^4 + 2\mu^2\sigma^2 + \sigma^4)\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \left(2\mu^2\sigma^2 + \sigma^4\right) \\
&= \frac{1}{n^2}\cdot n\left(2\mu^2\sigma^2 + \sigma^4\right) \\
\Rightarrow \mathrm{Var}[\widehat{\Theta}] &= \frac{1}{n}\left(2\mu^2\sigma^2 + \sigma^4\right).
\end{aligned}$$

Therefore,

$$\mathrm{E}\left[(\widehat{\Theta} - \sigma^2)^2\right] = (\mu^2)^2 + \frac{1}{n}\left(2\mu^2\sigma^2 + \sigma^4\right) = \mu^4 + \frac{1}{n}\left(2\mu^2\sigma^2 + \sigma^4\right).$$

[4 marks]

- **1 mark for stating and using theorem relating mean squared error to bias and variance.**
- **1 mark for using exercise to show** $\mathrm{Var}[Y^2] = \mathrm{E}[Y^4] - (\mathrm{E}[Y^2])^2$.
- **2 marks for remainder of computation of** $\mathrm{Var}[\widehat{\Theta}]$.

Table 1: Partial table showing 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