Reasoning for Statistics

1. **Appropriate Use of Average**
   - The mean is preferred when there are no extreme values because the mean takes into account all the data values in its computation.
   - The median is preferred when there are extreme values that affect the mean.
   - The mode is preferred when we are interested in the data value that is the most common or popular.

2. **How is average affected by systemic measurement error?**

   Ms Woo weighed seven oranges. The modal mass of the oranges was 148 g. The median mass of the oranges was 153 g. The mean mass of the oranges was 153.7 g. The weighing machine used by Ms Woo was found to be inaccurate. The correct mass of each orange was 18 g more than what Ms Woo has recorded. Write down the correct values of the modal, median and mean masses of the oranges.

   The distribution and range of the masses remains unchanged.

   Correct modal mass = (148 + 18) g = 166 g

   Correct median mass = (153 + 18) g= 171 g

   Correct mean mass = (153.7 + 18) g = 171.7 g

3. Statistical Diagrams – Advantages and Disadvantages

Type	Advantage	Disadvantage
Pictogram	• Colourful and appealing	• Difficult to draw • Difficult to draw a fraction of a picture • Actual frequency may be distorted because of the different sizes of pictures
Bar Graph	• Easier to draw than pictogram • Do not have to draw a fraction of a picture • Not distorted by different sizes of pictures	• Less colourful and appealing than pictogram • More abstract because frequency is represented by length of each bar • If frequency axis does not start from zero, the data may be distorted/misinterpreted
Pie Chart	• Easier to compare relative size of each category with the whole, e.g. one quarter, or more (or less) than half	• More abstract than pictogram because frequency is represented by sector area (or angle) • Harder to construct than a bar chart • Looks very cluttered if there are many categories (or sectors)
Line Graph	• Used to observe the rising or falling trend of a set of numerical data over a period of time	• Must not read in between the plotted points, unlike graphs of linear functions • If vertical axis does not start from zero, the data may be distorted/misinterpreted
Dot Diagram	• Suitable to display a small set of numerical data	• Not suitable to display a large set of numerical data • Not suitable to display a data set with many different values or a large range

Stem-and-leaf Diagram	• Individual data values are retained • Suitable to display a data set with many different values by grouping the values into equal class intervals	• Tedious to construct if there are too many data values
Histogram	• Suitable to display a data set with many different values by grouping the values into equal class intervals	• Individual data values are lost

4. Possible reasons why a graph is misleading

Feature of the graph	Why misleading?
Inconsistent scale on the vertical axis	Exaggerates the difference between the years
No defined scale on the vertical axis	Unable to determine the increase/ decrease without a scale
Not all years are shown	Attendance, sales, number of cases, etc could have been higher or lower
Unequal spacing of years	Misrepresents the trend
Vertical axis not in percentages	Number of students, cars, babies, etc likely to be different in each year
Title is biased e.g. use of words like "best , great, worst, etc"	Does not allow readers to make own judgement
Different sizes of pictures	Not clear whether the area or the height of each picture is used in comparing (e.g. 2015 looks four times costlier than 2012, but the height makes it look just twice as much)

5. Comparison between two groups (Must provide clear, concise responses that interpret the context of the question)

Measure	How?
Average - Median	Since median speed of the cars in the morning (70 km/h) is higher than median speed of the cars in the afternoon (65 km/h), on average, the cars are faster in the morning.
Spread - Interquartile range	Since the interquartile range of the speeds of cars in the morning (30 km/h) is higher than the interquartile range of the speeds of cars in the afternoon (10 km/h), the speeds of the cars are more spread out in the morning.
Spread - Standard deviation	Since the standard deviation of the speeds of cars in the morning (2.5 km/h) is higher than the standard deviation of the speeds of cars in the afternoon (1.6 km/h), the speeds of cars in the morning is less consistent. Higher SD – Data is less consistent Lower SD – Data is more consistent

6. Comparison between two cumulative frequency curves

Median The times taken by 120 women to complete the cycle race had the same interquartile range as the men's times but a higher median. Describe how the cumulative frequency curve for the women may differ from the curve for the men.	The cumulative frequency curve representing the times taken by women will be on the right side of the cumulative frequency curve representing the times taken by the men.
Interquartile Range The times taken by 120 women to complete the cycle race had a greater interquartile range then the men's times but same median. Describe how the cumulative frequency curve for the women may differ from the curve for the men.	The cumulative frequency curve representing the times taken by women will be less steep (gentler) than the cumulative frequency curve representing the times taken by the men.

7. Standard Deviation (measures the spread of a set of data about the mean)

Effect of data rearrangement on SD

Player 1	Player 2
13, 14, 14, 14, 15	13, 15, 14, 14, 14

No change in SD.

Effect of translation on SD

Player 1	Player 2
13, 14, 14, 14, 15	14, 15, 15, 15, 16

No change in SD.

Effect of multiplier on SD

Player 1	Player 2
13, 14, 14, 14, 15	130, 140, 140, 140, 150

New SD = 10×original SD

Effect of greater spread on SD

Player 1	Player 2
13, 14, 14, 14, 15	11, 14, 14, 14, 17

Greater spread of data set implies larger SD.