Sample Questions: Data Analysis

Question 1: Comparing Data Sets using Measures of Spread and Center (Multiple Choice)

The box plots below represent the test scores of two different Year 10 classes, Class A and Class B. [Image of two horizontal box plots, one labeled Class A and one labeled Class B.

- **Class A:** Min=50, $Q_1 = 65$, Median=75, $Q_3 = 85$, Max=100.
- **Class B:** Min=60, $Q_1 = 70$, Median=80, $Q_3 = 90$, Max=100.]

Which of the following statements is correct based on the data displayed in the box plots?

A. Class B is more consistent because it has a smaller Interquartile Range (IQR).
B. Class A performed better overall because its minimum score is lower than Class B's minimum score.
C. Class B performed better overall because its median score is higher than Class A's median score.
D. Class A has a greater spread of scores because its range is larger than Class B's range.

Solution:

1. **Calculate Medians:** Median (A) = 75; Median (B) = 80.
2. **Calculate IQRs:** $\mathrm{IQR(A)} = Q_3 - Q_1 = 85 - 65 = 20$; $\mathrm{IQR(B)} = 90 - 70 = 20$.
3. **Calculate Ranges:** Range (A) = $100 - 50 = 50$; Range (B) = $100 - 60 = 40$.

- Statement A is incorrect because both classes have the same IQR (20).
- Statement B is incorrect because a lower minimum score does not indicate a better overall performance (the median is a better measure).
- Statement C is **correct** because the median of Class B (80) is higher than the median of Class A (75), meaning 50% of Class B scored above 80, while 50% of Class A scored above 75.
- Statement D is incorrect; the IQR is the standard measure of spread, and they are equal. While the range of Class A (50) is larger than Class B (40), the overall distribution of the middle 50% is the same.

Answer: C

Concept: Interpreting and comparing the measures of center (median) and spread (IQR) from box plots.

next

Question 2: Standard Deviation and Data Interpretation (Short Answer/Calculation)

The times (in seconds) taken by five students to complete a task are: $10, 11, 13, 16, 20$.

a) Calculate the mean ($\bar{x}$) of the data set.

b) Explain what the standard deviation would represent for this data set and predict (without calculating) whether the standard deviation would be large or small.

Solution:

a) Calculate the mean ($\bar{x}$):

$$\bar{x} = \frac{\sum x}{n} = \frac{10 + 11 + 13 + 16 + 20}{5} = \frac{70}{5} = 14 \text{ seconds}$$

b) Explanation and Prediction:

- **Explanation:** The standard deviation represents the **average distance** or **typical spread** of each student's time away from the mean time of 14 seconds.
- **Prediction:** The standard deviation would be **relatively large**. The values (10 and 20) are quite far from the mean (14), suggesting a wide spread of performance times. (Actual calculation yields $\text{SD} \approx 3.9$ seconds, which is large relative to the mean of 14 seconds).

Final Answer: a) $\mathbf{14}$ seconds. b) The standard deviation represents the **typical distance of the scores from the mean**. It would be **relatively large** because the data points are widely spread (e.g., 10 and 20) around the mean of 14.

Concept: Calculating the mean and understanding the standard deviation as a measure of the typical deviation from the mean, and interpreting it contextually.

next

Question 3: Bivariate Data and Line of Best Fit (Extended Response/Application)

A study recorded the time spent studying mathematics (x hours) and the resulting exam mark (y out of 100) for several students. The data is plotted on a scatter plot, and the line of best fit is determined to have the equation $\mathbf{y = 6x + 55}$.

a) Describe the type, strength, and direction of the correlation you would expect to see in this scatter plot.

b) Use the line of best fit equation to predict the mark of a student who studied for 4.5 hours.

c) Explain whether it would be appropriate to use the equation to predict the mark of a student who studied for 15 hours.

Solution:

a) Describe the correlation:

- **Type:** Linear (since the relationship is modeled by a straight line).
- **Direction:** Positive (The gradient $m = 6$ is positive, meaning as study time increases, the exam mark tends to increase).
- **Strength:** Strong (A reasonable expectation for this type of data, suggesting the study time is a reliable predictor of the mark).

b) Predict the mark for $x = 4.5$ hours:

1. **Substitute $x = 4.5$ into the equation:**

$$y = 6(4.5) + 55$$
$$y = 27 + 55$$
$$y = 82$$

c) Appropriateness of predicting for $x = 15$ hours:

- **No, it would likely be inappropriate.** Predicting outside the range of the original data (extrapolation) can be unreliable. While the linear trend holds for typical study times, predicting a mark for 15 hours of study might lead to an unrealistic score above 100 or assume the student maintains the same productivity for an extended period, which is unlikely.

next

Final Answer:

a) **Linear, positive, and strong correlation.**

b) **82** marks.

c) **Inappropriate (Extrapolation): 15** hours is likely outside the range of the study data (extrapolation). Predicting far beyond the collected data range assumes the linear relationship continues indefinitely, which is often not true for real-world scenarios (e.g., marks are capped at 100).

Concept: Interpreting the characteristics of bivariate data, describing correlation, and applying a line of best fit for prediction, while understanding the limitations of interpolation versus extrapolation.

next