Data Representation and Interpretation

8

Free and always will be!

Focus: A set of questions and solutions for Year 8 students on Data Representation and Interpretation, tailored to the Australian Curriculum under the strand 'Statistics and Probability':

1. Data Collection and Organisation:

a) Explain the difference between primary and secondary data.

b) How can you organise the heights of students in your class for analysis?

2. Measures of Central Tendency:

a) Given the data set of test scores: $\{ 75, 82, 88, 75, 90, 92, 85 \}$ **, calculate the mean, median, and mode.**

b) Discuss when the median might be a better measure of central tendency than the mean.

3. Measures of Spread:

a) Calculate the range and interquartile range (IQR) for the test scores above.

b) What does the standard deviation tell us about a data set?

4. Data Representation:

a) Describe how to create a box plot for the test scores given (Draw one if needed).

b) When would you choose to use a histogram over a bar graph?

5. Interpretation of Data:

a) From the test scores, interpret what the skewness of the distribution might suggest about student performance.

b) If you notice a bimodal distribution in the test scores, what could this indicate?

6. Practical Application:

a) How might you use data analysis to improve school sports day activities?

b) Suggest how analysing data on book borrowing from the library could help in managing resources.

7. Measures of Central Tendency:

a) Calculate the mean, median, and mode for the following data set: $\{12, 15, 18, 12, 20, 15\}$.

b) Explain why the median might be preferred over the mean when dealing with salary data.

8. Measures of Spread:

a) Determine the range and interquartile range (IQR) for the data set in question 7a.

b) Calculate the variance and standard deviation for the data set: $\{4, 7, 10, 13, 16\}$.

9. Comparing Datasets:

a) You have two sets of test scores: Set A { $80, 82, 85, 87, 90$ }, and Set B { $70, 75, 85, 95, 100$ }. Compare these sets using measures of central tendency and spread.

b) What can you infer about the distributions from the measures calculated in part a?

10. Real-World Application:

a) How could a teacher use measures of central tendency and spread to assess class performance on a test?

b) In sports, if you're looking at the performance of athletes, why might you use the median rather than the mean for some statistics?

11. Data Analysis:

If you were to plot these data sets on a box plot, what would you expect to see in terms of the shape and spread of the data for Set A and Set B from above?

Solutions

1a.

Primary data is data collected directly by the researcher for a specific purpose, like a survey you conduct in your classroom.
Secondary data is data that has already been collected by someone else for another purpose, like using census data.

b.

You could organise the data into a frequency table or a stem-and-leaf plot. For example, if heights range from $140\,cm$ to $170\,cm$, you could use stems for tens $(14, 15, 16, 17)$ and leaves for units.

2a.

$$\begin{aligned}\text{Mean: Sum of scores} &= 75+82+88+75+90+92+85\\ &= 587,\\ \text{divided by 7 (number of scores)} &= \frac{587}{7}\\ &\approx 83.86.\end{aligned}$$

Median: Arrange in order: 75, 75, 82, 85, 88, 90, 92. The middle value is 85 .

Mode: The number that appears most frequently is 75 .

b.

The median is better when the data set contains outliers or is skewed significantly. It gives a better representation of the 'middle' of the data without being affected by extreme values.

3a.

$$\begin{aligned}\text{Range: Highest - Lowest} &= 92-75\\ &= 17.\end{aligned}$$

IQR:

Q1 (First Quartile) - after ordering, it's the median of the lower half: 75 .

Q3 (Third Quartile) - median of the upper half: 90 .

$$\begin{aligned}\text{IQR} &= \text{Q3 - Q1}\\ &= 90-75\\ &= 15.\end{aligned}$$

b.

Standard deviation measures the dispersion of data points relative to the mean. A small standard deviation indicates that data points tend to be close to the mean, while a large standard deviation suggests a wider spread from the mean.

4a.

Steps:
Order the data: $75,\ 75,\ 82,\ 85,\ 88,\ 90,\ 92$.
Find the median (85), Q1 (75), Q3 (90), and identify the minimum (75) and maximum (92) .
Draw a number line that includes these values.
Mark the median, Q1, Q3, minimum, and maximum.
Draw a box from Q1 to Q3, with a line for the median inside. Extend lines (whiskers) from the box to the minimum and maximum.

b.

Use a histogram for continuous data where you want to show the distribution of data in intervals (bins). It's ideal for numerical data like heights or ages, whereas bar graphs are better for categorical data or discrete data points.

5a.

If the distribution is skewed right (tail to the right), it suggests many students scored around or below the median with fewer scoring much higher, indicating potentially a tougher test or a class with varied abilities.

b.

A bimodal distribution might indicate two distinct groups within the data set, perhaps due to different levels of preparation, understanding, or even different teaching styles encountered by students.

6a.

Collect data on which activities students enjoy most, how long they can sustain physical activity, or their performance in trials. Use measures like mode for popular events, mean for average performance times, and consider variability for event planning.

b.

Analyse patterns in book borrowing to see which genres or authors are popular, at what times of the year borrowing peaks, or which books are rarely checked out. This can guide purchasing decisions, stock management, and organising library events or displays.

7a.

Mean: $\sum = \text{Sum} = 12 + 15 + 18 + 12 + 20 + 15$

$= 92$, divided by 6

$= \dfrac{92}{6}$

$\approx 15.33.$

Median: Arrange in order: $12,\ 12,\ [15,\ 15],\ 18,\ 20$.

The median is the average of the $3rd$ and $4th$ numbers:

$\rightarrow \dfrac{15 + 15}{2}$

$= 15$.

Mode: The number that appears most frequently is 12 and 15,

making this set bimodal with modes 12 and 15.

b.

Salaries can have significant outliers (like very high executive salaries), which skew the mean upwards. The median, being the middle value, provides a better representation of what a 'typical' salary might be, not being as influenced by these outliers.

8a.

Range: Highest value - Lowest value

$= 20 - 12$

$= 8\,.$

IQR:

Q1 (First Quartile) - after ordering, it's the median of the lower half: $12\,.$

Q3 (Third Quartile) - median of the upper half: $18\,.$

IQR $=$ Q3 $-$ Q1

$= 18 - 12$

$= 6\,.$

b.

Mean: Sum $= 4 + 7 + 10 + 13 + 16$

$= 50$, divided by 5

$= 10\,.$

Variance:

Deviations from mean: $(4-10)^2, (7-10)^2, (10-10)^2, (13-10)^2, (16-10)^2$

$= 36,\ 9,\ 0,\ 9,\ 36\,.$

Sum of squared deviations $= 36 + 9 + 0 + 9 + 36$

$= 90.$

$$\text{Variance } = \sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

$$\text{Variance } = \frac{\text{Sum of squared deviations}}{\text{Number of Scores}}$$

$$= \frac{90}{5}$$

$= 18\,.$

Standard Deviation: Square root of variance $= \sqrt{18}$

$\approx 4.24\,.$

9a.

Mean:

Set A: $\dfrac{80+82+85+87+90}{5}$

$= 84.8$

Set B: $\dfrac{70+75+85+95+100}{5}$

$= 85$.

Median:

Set A: 85.

Set B: 85.

Range:

Set A: $90 - 80$

$= 10$.

Set B: $100 - 70$

$= 30$.

IQR:

Set A: Q1 $= 82$, Q3 $= 87$,

IQR $= 5$.

Set B: Q1 $= 75$, Q3 $= 95$,

IQR $= 20$.

Conclusion: Set B has a higher mean and a much wider range and IQR, indicating more variability in scores.

b.

Set A's scores are more consistent (lower range and IQR), suggesting a narrower distribution of scores. Set B's larger range and IQR indicate a wider spread, possibly with some students performing significantly better or worse than others.

10a.

The mean gives an overall average performance, the median tells where the middle student stands, and the mode indicates the most common score. Range and IQR show how dispersed the scores are, which can indicate the level of difficulty or consistency in understanding among students.

b.

In sports, performances can be affected by outliers (exceptionally good or bad performances). The median can provide a better insight into typical performance without being skewed by one or two extreme results, especially in individual sports.

11.

Set A would have a compact box plot with a short box (small IQR) and short whiskers (small range), indicating closely grouped data. Set B would have a longer box and whiskers, showing a wider spread of scores with possible outliers at the high or low end.

Additional Notes for Teachers:

Learning Outcomes:
Students should be able to calculate, interpret, and compare measures of central tendency and spread to understand data distributions. Students should be able to collect, organise, describe, and interpret data using various statistical tools and representations.

Teaching Strategies:
Use real data from class activities or sports to make learning relevant. Discuss the implications of each measure in different contexts. Engage students with real-life data collection projects or use existing school data for analysis. Encourage discussion on how data influences decision-making.

Assessment:
Provide datasets for students to analyse, asking them to interpret results and explain why certain measures are more appropriate in given scenarios. Assess through projects or tasks where students must analyse data, choose appropriate representations, and draw conclusions from their findings.

Resources:
Use statistical software or spreadsheets for calculations, and encourage manual computation for understanding. Data visualisation tools can help in interpreting these measures graphically. Use software like Excel or Google Sheets for data analysis, or traditional methods like graph paper for manual plotting. Online resources or educational games can make learning data analysis interactive.

This set of questions aligns with the Australian Curriculum for Year 8, focusing on deepening students' understanding of data representation and interpretation.