UNIT 2: Data presentation and interpretation

Return to overview

SPECIFICATION REFERENCES

2.1 Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency

Connect to probability distributions

2.2 Interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams which include distinct sections of the population (calculations involving regression lines are excluded)

Understand informal interpretation of correlation

Understand that correlation does not imply causation

2.3 Interpret measures of central tendency and variation, extending to standard deviation
Be able to calculate standard deviation, including from summary statistics

2.4 Recognise and interpret possible outliers in data sets and statistical diagrams
Select or critique data presentation techniques in the context of a statistical problem
Be able to clean data, including dealing with missing data, errors and outliers

PRIOR KNOWLEDGE

GCSE (9–1) in Mathematics at Higher Tier

S2 Interpret and construct tables, charts and diagrams, including frequency tables, bar charts, pie charts and pictograms for categorical data, vertical line charts for ungrouped discrete numerical data and know their appropriate use

S3 Construct and interpret diagrams for grouped discrete data and continuous data, i.e. histograms with equal and unequal class intervals and cumulative frequency graphs, and know their appropriate use

S4 Interpret, analyse and compare the distributions of data sets from univariate empirical distributions through appropriate measures of central tendency (median, mean, mode and modal class) and spread (range, including consideration of outliers), quartiles and inter-quartile range

S6 Use and interpret scatter graphs of bivariate data; recognise correlation and know that it does not indicate causation; draw estimated lines of best fit; make predictions; interpolate and extrapolate apparent trends while knowing the dangers of so doing

KEYWORDS

Histogram, box plot, probability density function, cumulative distribution function, continuous random variable, scatter diagram, linear regression, explanatory (independent) variables, response (dependent) variables interpolation, extrapolation, product moment correlation coefficient (PMCC), mean, median, mode, variance, standard deviation, range, interquartile range, interpercentile range, outlier, skewness, symmetrical, positive skew, negative skew.

2a. Calculation and interpretation of measures of location; Calculation and interpretation of measures of variation; Understand and use coding (2.3) (2.4)

Teaching Time
4 Hours

OBJECTIVES

By the end of the sub-unit, students should:

- be able to calculate measures of location, mean, median and mode;
- be able to calculate measures of variation, standard deviation, variance, range and interpercentile range;
- be able to interpret and draw inferences from summary statistics.

TEACHING POINTS

The calculation of the mean, median and mode should be recapped from GCSE however the focus now is on students using calculators to do the calculations. Check understanding of the terminology and teach calculator methods.

Students require an understanding of measures of variation too and should be able to use their calculators to calculate the variance and standard deviation. They should be able to use the statistic $S_{xx}=\sum\left(x-\bar{x}\right)^2=\sum x^2-\frac{\left(\sum x\right)^2}{n}$. Students are expected to use standard deviation $=\sqrt{\frac{S_{xx}}{n-1}}$ but equivalents including spreadsheet formula ($s=\sqrt{\frac{S_{xx}}{n-1}}$) will be accepted.

The data may be discrete or continuous, grouped or ungrouped, and students need to be able to interpret these summary statistics clearly and be able to make inferences from them. Significance tests will not be expected.

Coding for both mean and standard deviation needs to be covered. Be clear that students need to be able to uncode both mean and standard deviation. Emphasise that the standard deviation is unaffected by the addition or subtraction of constants.

Students are expected to be able to use linear interpolation to calculate percentiles from grouped data.

OPPORTUNITIES FOR PROBLEM SOLVING/MODELLING

There is opportunity for further use of the large data set here. Summary statistics of elements from the data set can be calculated and then used to compare and interpret for both location and variation statistics.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

When calculating the mean, of grouped data some student may divide by the number of groups rather than the number of items of data, they may also use class widths in the calculation rather than the mid-points.

When finding the standard deviation, the most common error is forgetting to take the square root (perhaps because they are not clear about the difference between variance and standard deviation). Some students waste time by ignoring given values and recalculating Σfx and Σfx^2.

Difficulties with coding are due to a lack of understanding about how coding affects the mean and standard deviation, and poor algebraic skills. Students sometimes substitute for the wrong variable, fail to solve equations correctly or get the order of operations the wrong way around.

Students should be reminded that they must be precise in their use of language and use the correct terms such as ‘median’. ‘range’ or ‘inter-quartile range’ rather than the more general ‘average’ and ‘spread’. Students should also remember to use accurate values throughout calculations to avoid losing marks due to premature rounding.

NOTES

Students are expected to know the different notation for population summary statistics $\left(\mu, \sigma^2, \sigma\right)$ and sample summary statistics $\left(\bar{x}, s^2, s\right)$.

2b. Interpret diagrams for single-variable data; Interpret scatter diagrams and regression lines; Recognise and interpret outliers; Draw simple conclusions from statistical problems (2.1) (2.2) (2.4)

Teaching time

8 hours

OBJECTIVES

By the end of the sub-unit, students should:

- know how to interpret diagrams for single variable data;
- know how to interpret scatter diagrams and regression lines for bivariate data;
- recognise the explanatory and response variables;
- be able to make predictions using the regression line and understand its limitations;
- understand informal interpretation of correlation;
- understand that correlation does not imply causation;
- recognise and interpret possible outliers in data sets and statistical diagrams;
- be able to select or critique data presentation techniques in the context of a statistical problem;
- be able to clean data, including dealing with missing data, errors and outliers.

TEACHING POINTS

Students should be familiar with and be able to interpret histograms, frequency polygons, box and whisker plots and cumulative frequency diagrams. These should have been covered at GCSE but it is worth a recap for consistency of methods. Also cover calculating summary statistics from diagrams, including the mean and standard deviation from a histogram.

For bivariate data students should understand the terms explanatory and response variables and know where each is placed on the axes of a scatter diagram. This is particularly important as variables other than y and x could be used.

Students are not expected to know, calculate or understand the regression line formula. Students will need to understand the use of interpolation when using a regression line equation to make predictions within the range of values of the explanatory variable and they need to understand the dangers of extrapolation (predictions outside the range), again variables other than y and x could be used.

Students will be expected to describe the correlation on a scatter diagram in terms of positive, negative or no correlation and strong or weak but no calculations need to be made. Values from calculations will not be given for interpretation.

Outliers will need to be identified and interpreted from data sets and statistical diagrams. Any rules to be used will be given in the question, for example $Q_1 - 1.5 \times IQR$, $Q_3 + 1.5 \times IQR$.

Students will be expected to select an appropriate diagram or critique the choice of one which is used. They should also be able to clean data by identifying possible outliers (box plots and scatter diagrams). They may also be asked to fill in missing data using a regression line.

OPPORTUNITIES FOR PROBLEM SOLVING/MODELLING

Again all of the diagrams and techniques used in this unit could be modelled using data from the large data set.

COMMON MISCONCEPTIONS/EXAMINER REPORT QUOTES

Knowing how to interpret statistics students have calculated is sometimes found challenging, and often discriminates between students in exam questions. Full and clear reasons for interpretations and decisions need to be given for marks to be awarded.

Many students have difficulties calculating the sizes of bars in histograms, as commented on by one examiner: 'Most were able to state the correct width of the bar but few used frequency densities correctly to find the height, some finding the frequency density of but then calculating $\frac{1}{3}\times 2.5$ rather than $2.5\div\frac{1}{3}$. Some identified that 1.5 cm^2 represented 10 customers but were then unable to use this correctly to find the height … some students had an incorrect class width because they did not realize that the lower class boundary was 70 not 69.5.'