

Data processing, presentation and interpretation (AS)

L1	Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency Connect to probability distributions
L2	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
L3	Interpret measures of central tendency and variation, extending to standard deviation Be able to calculate standard deviation, including from summary statistics
L4	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

Commentary

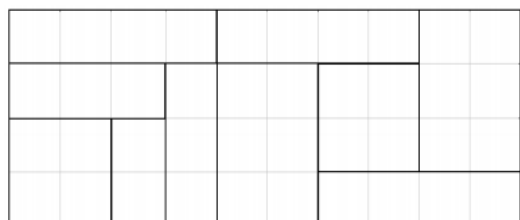
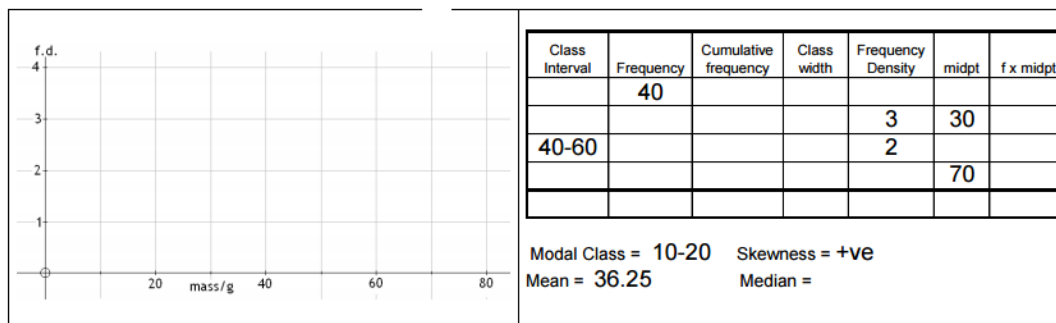
Data are collected for all kinds of reasons, perhaps as part of a statistical investigation or as part of routine data gathering by business or government. We need to organise data in ways that make it easy to see their main features; this involves using suitable data displays and summary measures where the choice is determined by the type of data you are working with.

Technology can play a large role in this unit as the calculator students use must be able to accept small data sets (in the form of a list or frequency table) and provide summary statistics. The emphasis is not on doing calculations by hand but using sample statistics and displays to draw inferences about the population.

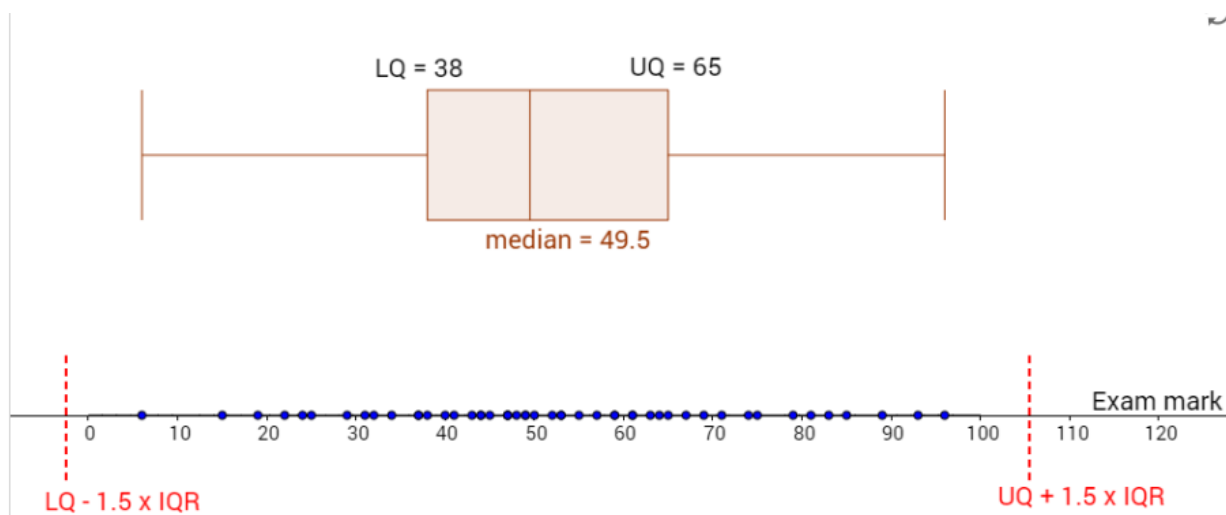
There is a lot of scope in this unit to explore large data sets (such as a pre-release data set) and to use Excel/GeoGebra to perform the calculations and produce the diagrams. It is common for there to be missing data – some software will interpret missing values as zero if they are left as blanks; other software will not work when there are blank entries. When drawing statistical diagrams for real data, this will often reveal unusually large or small values (outliers) – these can arise from errors in data entry or from people misinterpreting what data were required. It is important to deal with missing and unusual values in data when doing statistical analysis – this process is called data cleaning.

Sample MEI resource

'Histogram reconstruction' (found at <http://integralmaths.org/sow-resources.php>) is designed as a puzzle for students to use the given information and the rectangles to recreate a number of Histograms.



Effective use of technology



'Boxplots and outliers' (found at www.mei.org.uk/integrating-technology) is designed to investigate outliers. The points on the line represent the scores of 50 students in an examination. The points can be dragged along the line to investigate how the shape of the boxplot changes.

Data processing, presentation and interpretation (AS)

Time allocation:

Pre-requisites

- GCSE: Data Handling and interpreting what data tells us.
-

Links with other topics

- Probability: Linking probability with area under histogram
- Binomial(AS)/Normal: Fitting models to data sets
-

Questions and prompts for mathematical thinking

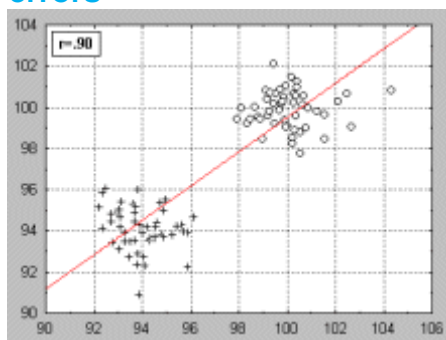
- A set of 11 data values has a mean of 8 and a standard deviation of 2. At least one of the data items is the value 8. This single value is removed from the set. What will happen to the standard deviation of the remaining values? The standard deviation:
A) Increases B) decreases C) Is unchanged D) insufficient information is given

Applications and Modelling

- Line up the boys and girls in your classroom; discuss which set has the most consistent set of heights. Check by finding the standard deviation.
- In 2013 12.3% (you could look up latest figure) of sixth form students do A level Maths – compare this to the figure for your school/college. How many would you need to have doing A level Maths in your school/college to have unusually many (or unusually few)?
-

Common errors

-



This scatter plot has $r = 0.9$. The error is to say that there is evidence of a strong positive correlation – in fact we can see there are two distinct groups each with $r \approx 0$.

- Forgetting to use frequency density for histograms with unequal widths.
- Using incorrect formulae for the standard deviation or using the root mean square deviation instead.
- Confusing 'association' with 'correlation' and misuse of the term 'data' to represent 'underlying population'
-