About data

Today's society is characterised by the proliferation of information in many forms. Every day we are bombarded with the statistical analyses of a wide range of issues — from advertising claims, opinion polls, population trends, and estimates of health risks to the performance results of students and schools. To be knowledgeable, active and responsive members in today's world, students need to have astute understandings about data.

The Data topic in the Years 1 to 10 Mathematics *Syllabus* helps students to develop understandings related to planning data collections, collecting data, handling data, exploring and presenting data, and interpreting variations in data.

Planning data collections

Students are encouraged to collect data so that they may investigate particular situations or their own and others' questions. Data collection methods include observation, surveys, experiments, and the extraction of data from existing sources.

When students are involved in planning for data collections, they develop an awareness of:

- the purpose for the data collections
- where the required data might be found or collected
- what can be observed or asked
- expected responses
- how observations or responses could be recorded.

Students may need to discuss various issues related to data collection. In surveys, for example, the students need to aim to avoid unexpected answers or lengthy responses, but they also need to be prepared to receive some. They also need to be prepared for the non-return of written forms and blank responses. Through guided reflection on experiences, students see the need to trial survey questions to test whether they provide good quality and unambiguous data, and to clarify the questions posed. For example, open-ended questions such as 'Which banks do you know about in Brisbane?' or 'Does your mother work?' are likely to produce ambiguous responses and hence unreliable data. In trialling a question for a survey, students can check for ambiguities by asking trial respondents for their interpretations of questions.

Students learn alternative ways to classify and categorise survey data for a variety of reasons. Sometimes it is to make surveys as respondent-friendly as possible because respondent-friendly surveys produce more reliable data. Sometimes it is needed to handle some responses or observations, including those where additional information beyond what was asked is provided. Sometimes in exploring a data set, categories with small frequencies are combined.

Students will need to make decisions about who or where to collect the data from, how to collect it, and how much data is to be collected. Once students are familiar with different methods of collecting data, they are able to select one that is appropriate and efficient for a given situation.

More complex data collections

As students' knowledge and understandings develop, they are able to investigate the collection of data over time (e.g. yearly rainfall records) and cross-sectional data where time is not a major focus (e.g. information about eating habits from different individuals). They consider other aspects of data collection such as sampling procedures (where data is collected from a random group to represent a population, or in an experiment or observations to represent a general situation) and whether population or census data (where data is collected from a whole population) about a particular issue is available. While students will not be calculating estimates of variation, they should engage in thoughtful discussions about generalisations, populations and samples, considering questions such as*:*

What can be said about a general situation or a whole population based on the data we are considering?

How much care do we need to take in any statements to allow for variation and the extent of the data we have?

© The State of Queensland (The Office of the Queensland Studies Authority) 2005

Collecting data

Students use a variety of practical methods to record data. It is part of the learning process to start with draft templates/methods that stem from students' personal experiences. They need to learn strategies to enable them to critically analyse and reflect on the efficiency and effectiveness of these methods. Students should be encouraged to explore a range of recording methods and will require support to trial, develop and refine *recording templates*.

Recording templates

Students need to make decisions about the purpose for the data collection and the type of recording template to be used. They will require opportunities to develop recording templates to collect their own data as well as opportunities to develop templates where data from existing sources can be combined for exploration and interpretation.

Students design and develop their own data recording templates and may trial the templates *before* conducting large data collection exercises. If students collect data separately with the intention of combining their data, they need to ensure complete agreement on the methods, criteria, and groups, *before* they collect their data. Again, a small trial of their template will assist students in ensuring this agreement amongst different collectors.

The transference of data from recording templates to spreadsheets should provide opportunities for students to use computing technology. As students become more familiar with spreadsheets, visualising their spreadsheet is of great assistance in designing and developing their data collection plans and their data recording templates.

Some decisions about grouping or categorising of data may be done during the design stage of the templates to provide a framework for recording responses or observations. Younger students will tend to use groupings and categorisations with which they are comfortable, and their main learning will be to ensure they cover the range of values or responses in the data to be collected. For example, if they are observing colours of cars, they will need to decide beforehand whether to record every different colour or how to group the colours. Students will learn that decisions about grouping can be made after the data are collected — for example, if only a small number of gold- and silver-coloured cars are observed, students could decide to group them together.

Types of data

As students' knowledge and understandings develop, the types of data being collected, and the circumstances of collection, may affect students' decisions about groups and categories at the design stage.

The broad classification of types of data is into *discrete* and *continuous*.

Discrete data — Data is discrete when the values or the names of the observations are separated or disconnected. Discrete data can be *categorical* or *count*. Each observation in a categorical data set is distinguished by a naming word — for example, gender (male/female), agreement (yes/no/don't know), religion (with a selected list of religions), or type of movie (comedy, drama, horror, etc). In a count data set, each observation is an integer that has counted something — for example, the number of children in a family, or the number of accidents at an intersection in a year.

It is important to help students distinguish between count data, and frequencies of observations. For example, when the number of children in a family is recorded, this is count data with each observation referring to a different family and taking a value — 0, 1, 2, 3 and so on. The frequency of the value 0 will refer to the number of families with 0 children; the frequency of the value 1 will refer to the number of families with 1 child; and so on.

Continuous data — These are numerical data where each observation can take a value within an interval. Continuous data can be measured and broken down into smaller parts that are meaningful. It is often associated with measurement. Heights or weights of people, animals or plants are common examples of continuous data, and the amounts of time between, or durations of, events are others. These data are continuous because our measuring instruments, if more accurate, could measure to smaller and smaller decimal fractions. Observations in continuous data are made in systems of units, and are 'rounded' 'to the nearest…', depending on either the measuring device or the desired units. For example, if we say a person's height is 1.65 metres, we mean that the height is between 1.645 and 1.655 metres. Amounts of money are continuous data because, like height, they are measured in units, and any fraction of a unit is meaningful. For example, casual wages are usually

 © The State of Queensland (The Office of the Queensland Studies Authority)

expressed in dollars per hour up to four decimal places, and total payments are obtained by the dollars per hour rate multiplied by the number of hours — which might also involve fractions of an hour. Because different countries use different units of money, exchange rates are also quoted with a number of decimal places.

Continuous data are often grouped for psychological or respondent-friendly reasons in collecting data, particularly in surveys. For example, it is often thought that in asking a person's age or weight, more reliable data will be obtained if people can choose a grouped range — for example, 30–39 years. In asking questions about lengths of time — for example, 'Approximately how much time do you spend watching television per week?' it may be respondent-friendly to provide a choice of grouped amounts such as 0–5 hours or 5–10 hours. For this example and for other similar groupings, it is important for the respondent to know in which group to place the value 5.5 hours. As students' knowledge and understandings progress, these more subtle considerations in survey design can be considered. Generally speaking, continuous data should not be grouped at the time of collection unless there are good reasons to believe that grouping will assist significantly in obtaining good quality (that is, reliable) data. For example, in asking the age of children or young adults, there is no reason not to ask for actual age in years rather than within an artificial grouping.

Handling data

Checking for errors and unusual responses

Students should always check their data for recording or observing errors, and in the data collections of older students, to ensure units and observational circumstances are consistent. If a recording error is detected and it is clear what the error is (for example, a person's height recorded as 16.5 cm) the error is corrected. If it is not clear what the error is, the person who recorded that particular observation may be able to recall the circumstances. If it is clear there is an error but the correction is not obvious, that observation is discarded. If it is not clear that there is an error, the observation is not discarded.

As students' gain more experience in data collection, they learn that where circumstances change during the data collection (whether through error or chance), the data set is split into two data sets if it is not possible to reconcile the inconsistencies.

As students' knowledge and experience develop, they also learn to examine responses in survey data. Decisions about the classification of unusual or ambiguous responses are made, as are decisions about combining categories for those with small frequencies.

Using computer software

Students with access to, and ability to handle, spreadsheet software, enter their 'cleaned' data into a single spreadsheet, with each row corresponding to the person or entity on which observations were made, and each column corresponding to data collected or observed on a characteristic of the person or entity.

For example, consider a situation where data are collected on cars passing the school in 9.00–9.15 am and 2.00–2.15 pm timeslots on a Monday and Tuesday, the colour of the car, the direction of its travel (left/right), and the number of occupants are noted. The data are entered on a spreadsheet, the first few lines of which are reproduced below. Time of day (morning or afternoon), day, colour and direction are all categorical, while the number of occupants is count data. The observations are taken per car so each row corresponds to a car.

Time of day	Day	Colour	Direction	No. of occupants
morning	Monday	grey	left	1
morning	Monday	blue	left	1
morning	Monday	white	right	3
morning	Monday	white	left	2

© The State of Queensland (The Office of the Queensland Studies Authority) 2005

In a further example, the first few lines of the spreadsheet for data gathered from a survey of students in which their school grade, gender, height and number of siblings were recorded is shown below.

Grade	Gender	Height in cm	No. of siblings
6	M	140	1
5	F	135	2
5	M	130	0
7	M	136	4

Grade and gender are categorical, height is continuous and number of siblings is count data. The observations are recorded per student so each row corresponds to a student.

Exploring and presenting data

Once students are satisfied with the quality and 'cleaning' of the data set they have collected, choices can be made about arranging and presenting the data in a way that responds to the original questions or issues. In the early years, students may create their own displays of data and use these to describe what the data illustrates.

Discrete data

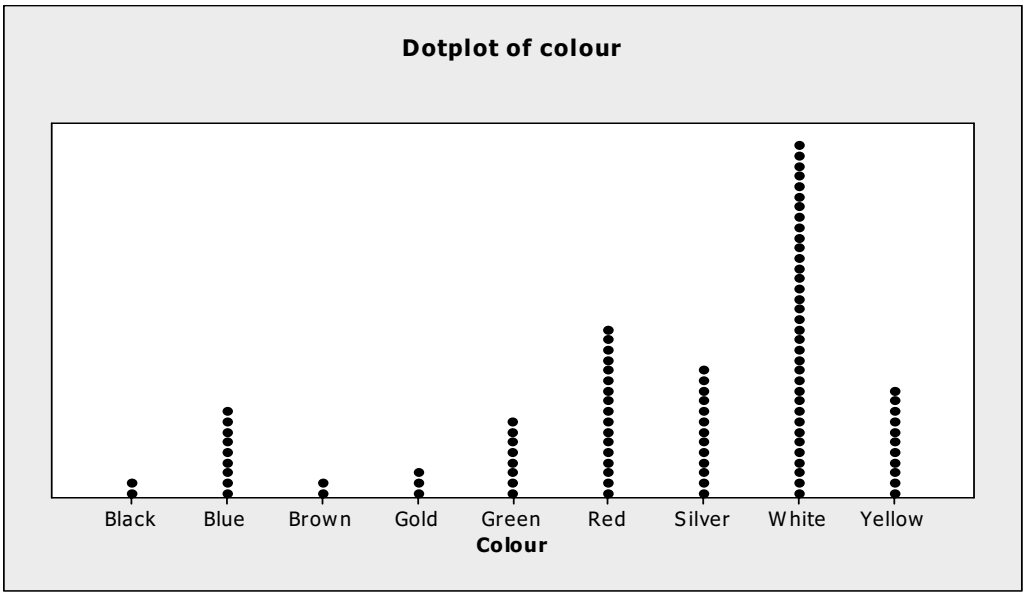
Students may use *lists* or drawings as an initial display of collected data. Generally lists are re-arranged to make the data easier to interpret. More refined lists or *tables* are used to organise available data into categories or groups that present a particular perspective on the data. As students develop understandings of data displays and the effects of grouping data, they can experiment with arrangements to explore the differences in interpretations of the same data set, and see if different groupings or categories provide alternative viewpoints.

	Has a DVD	No DVD
Has a computer	6	4
No computer	15	5

The table shows that six students have both a computer and a DVD, four students have a computer but not a DVD, 15 students have a DVD but no computer, and five students have neither a computer nor a DVD.

In the early years of schooling, students produce a range of graphs that use *objects*, *pictures* and *people* to illustrate available data. These displays are meaningful to the students and easier for them to interpret than other types of displays — students are able to see how a response is linked to the display. As students develop, they replace the objects, pictures or people in graphs with a symbol such as a dot. They have then created a dotplot (see page 9). Below is a dotplot of data on the colour of cars travelling through a certain intersection (Ballard, Cox, Verran & Wilson, reported in MacGillivray 2005, p. 103). Each dot counts a car.

© The State of Queensland (The Office of the Queensland Studies Authority)



Students' developing understandings enable them to use bar graphs, pie charts, dotplots, histograms or scatter graphs when representing data and to choose the graph that best suits the data they have collected or accessed.

Bar graphs and pie charts — These suit discrete data that can be arranged into clearly defined categories or have distinct, separated values. Either a count (frequency of occurrence) or a percentage (relative frequency) of the number of observations in each category or for each distinct value (for count data) is given. Bar graphs (also called bar charts) can be created with bars running vertically or horizontally. Comparisons between categories shown on a bar graph are made using the height (or length) of the bars. Areas of the pieces of pie represent relative frequencies and are used to compare categories shown on a pie chart. A strip chart is an alternative to a pie chart, in which areas again represent the relative frequencies of the categories.

Students use collected data to create bar graphs that include titles and names for the axes. One of the axes has labels for the categories or values for a count variable and the other has frequencies or relative frequencies. The frequency of the values should be clearly indicated so that the graph is easily read.

Students note that the shape of the bar graph or pie chart is the same whether frequencies or relative frequencies (percentages) are plotted.

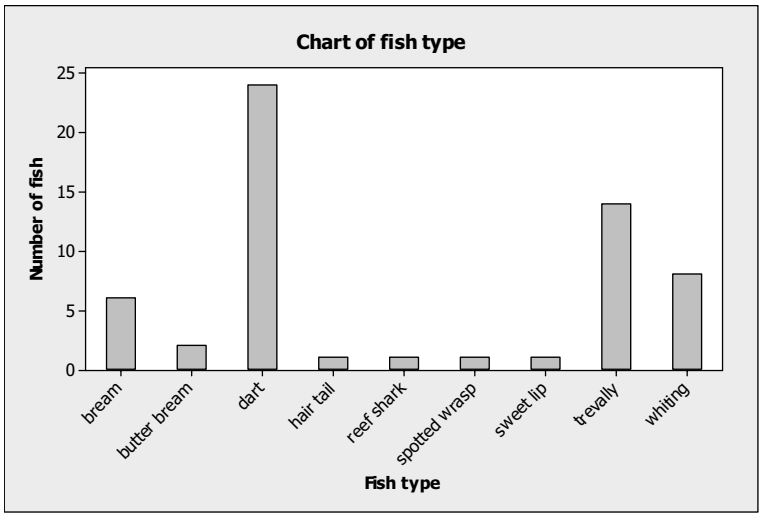
The first bar graph below is of categorical data, showing the number (frequency) of fish of different types caught during a group fishing expedition to North Stradbroke Island (Carr & Salzburger, reported in MacGillivray 2005, p. 3).

The second bar graph below is of count data, showing the number (frequency) of houses with 1, 2, 3, 4, 5 or 6 bedrooms in a real estate investigation of a region.

For categorical and count data, the *mode* is the category or count value that occurs most often — that is, it has the highest frequency. The mode is readily displayed in a bar graph.

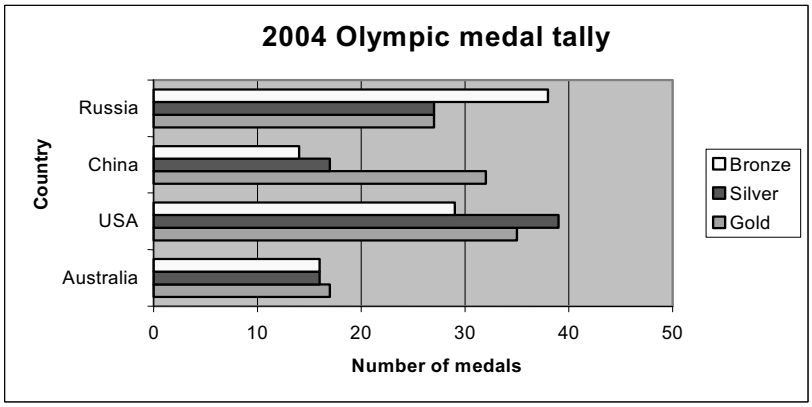
In bar graphs or pie charts, the only scale is the one that refers to the frequencies or relative frequencies of categories.

© The State of Queensland (The Office of the Queensland Studies Authority) 2005



Compound bar graphs — These are graphs where two or more sets of data appear together and can be compared and contrasted. For example, a compound bar graph could be used to represent the spread of medals won by some countries at the Olympic games using the following data table:

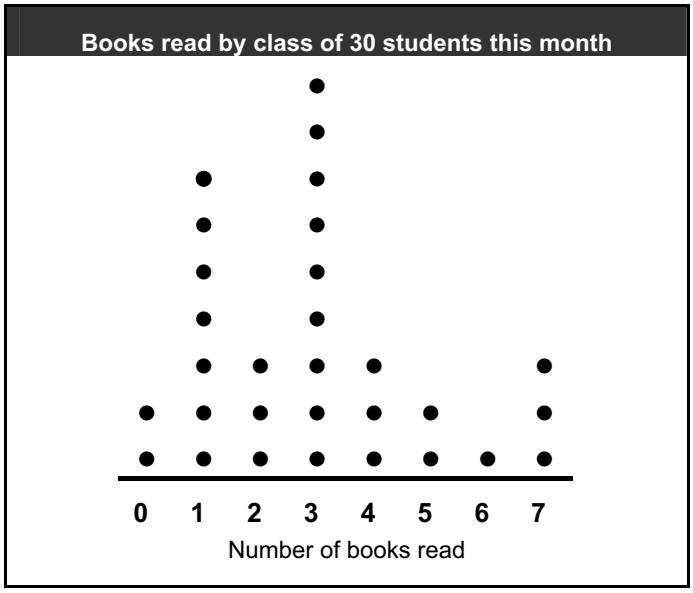
	Australia	USA	China	Russia
Gold	17	35	32	27
Silver	16	39	17	27
Bronze	16	29	14	38



© The State of Queensland (The Office of the Queensland Studies Authority)

In the early years of schooling, students are unlikely to come across, or to collect, count data with large values of counts. As students progress, however, they may be involved in considering count data with large counts such as attendances at football matches, or the number of vehicles per minute passing a certain point on a freeway. For data with large values of counts, it is more appropriate to use the presentations and graphs that are used for continuous data, than to use bar graphs and pie charts.

Dotplots: These are plots that can be used for any type of variable. They provide an easy way of presenting the frequencies of either small or large amounts of discrete or continuous data. A dotplot is created by placing one dot for each observation above the appropriate name, number or value of the observation. For continuous and count data, dotplots make it easy to identify the greatest value, the least value, the range (difference between greatest and least) and the general shape of the 'raw' data. The following dotplot represents the number of books read in a month by students in a class of 30. It shows that nine students each read three books in the month (three is the mode), that the highest number of books read was seven, and that only three students read seven books in the month.



Continuous data

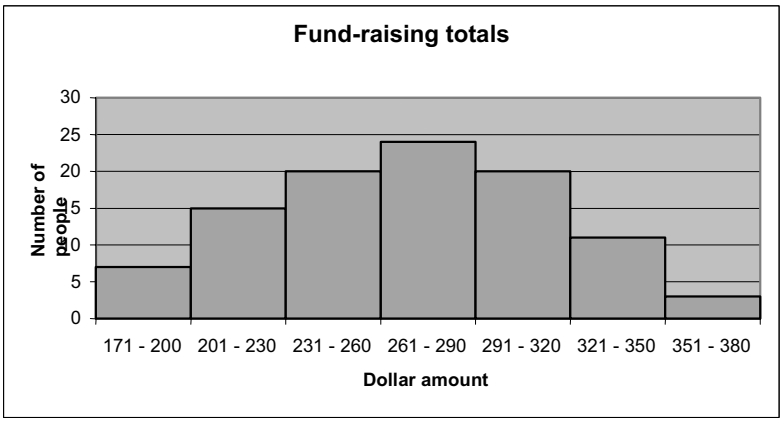
Dotplots can also be used for continuous data, and often provide a good way for students to first 'eyeball' data if computer technology is available to them. For example, a dotplot of prices will clearly show the most popular prices such as those ending in 99 cents. If a very large amount of continuous data is available, such as 500–1000 observations, a dotplot will give a good picture of the data.

For smaller sets of continuous data, students will note that there are too many different values for a dotplot to give a good picture representation of the data set. This is why most graphical representations of a set of continuous data group the data into intervals. Groups used with continuous data are called '*bins*' and alterations to the groups or bins can change the shape of a data display. Students should be given opportunities to experiment with different bins to see the changes that occur in data displays.

A *histogram* is a picture of the frequencies or relative frequencies of intervals of values. Histograms are used only for continuous data (in the same way that bar graphs are used only for discrete data). The bins give the intervals in which the continuous data are grouped. When the bin intervals are equal, as with a bar graph, the heights of the histogram boxes give frequencies or relative frequencies, but now of the observations that lie in the range of each bin.

For example, the histogram below is of the continuous data set of the amount of funds raised by each person in a charity drive. The bins are intervals of equal length. Students will need to understand that, for example, $260.40 will be placed in the $231–$260 bin, and that they need to decide in which bin to place $260.50.

© The State of Queensland (The Office of the Queensland Studies Authority) 2005



When creating histograms, students need to make choices about bin width (the range of values in each bin) and/or the number of bins. Students should understand that the larger the data set, the better the picture provided by a histogram, and that the original data are not visible because of the grouping of data into bins — the histogram is a summary graphical presentation of the data.

In most instances, histograms have bins of equal widths although they can be drawn with bins with unequal widths, that is, representing unequal intervals of values. Great care is required in drawing histograms with bins of unequal widths as the area of the histogram boxes must represent the frequencies or relative frequencies, that is, the height of the box no longer represents the frequency or relative frequency.

Another plot that shows much the same information as a histogram is a *stem and leaf plot* (often called a stem plot). Stem and leaf plots are like a horizontal histogram, however, the bin widths — stems — in stem and leaf plots are restricted by the number system. The stem is always the first digit or digits of the values in the data set and the leaf is the next digit.

The following data set of test scores for 28 students has been used to create a stem and leaf plot that gives a picture of the data.

55 90 62 75 80 90 54 95 80 95 62 70 55 84 57 65 74 63 68 72 98 75 91 87 76 65 61 77

As the smallest value in this data set is 54 and the highest is 98, the stem is the tens digit of the numbers. (The first digit of each value is the stem and the other digits are the leaves.) The first plot (Plot A) is an interim plot created by grouping the observations; the second plot (Plot B) is the final stem and leaf plot, obtained by ordering the leaves.

Stem	Leaves
5	5 4 5 7
6	2 2 5 3 8 5 1
7	5 0 4 2 5 6 7
8	0 0 4 7
9	0 0 5 5 8 1

Plot A

Stem	Leaves
5	4 5 5 7
6	1 2 2 3 5 5 8
7	0 2 4 5 5 6 7
8	0 0 4 7
9	0 0 1 5 5 8

Plot B

The plots show that there were four numbers in the fifties — 54, 55, 55 and 57, seven numbers in the sixties — 61, 62, 62, 63, 65, 65, 68 and so on.

Stem and leaf plots can be created very quickly by hand to inspect data when the data set is reasonably small. Like histograms, they show the shape and distribution of the data. One of the advantages of the stem and leaf plot is that although the data are grouped, all the data are shown. This is particularly useful in calculating quantities such as the data median. Stem and leaf plots sometimes become unwieldy when used for larger data sets.

© The State of Queensland (The Office of the Queensland Studies Authority)

Two and more variables

The two-way tables and compound bar graphs discussed earlier in this document are presentations of two categorical variables.

Two continuous variables — Students learn about and use *scatterplots* to investigate how one continuous variable varies with another. For example, a scatterplot of the amount of time spent watching television per week against the age of the watcher illustrates how viewing varies with age. Each point on the scatterplot corresponds to a pair of observations — age and time per week watching television. In a spreadsheet, each row would correspond to a person, with one column corresponding to age and another to time per week watching television. Each point on a scatterplot of time per week watching television against age corresponds to one person.

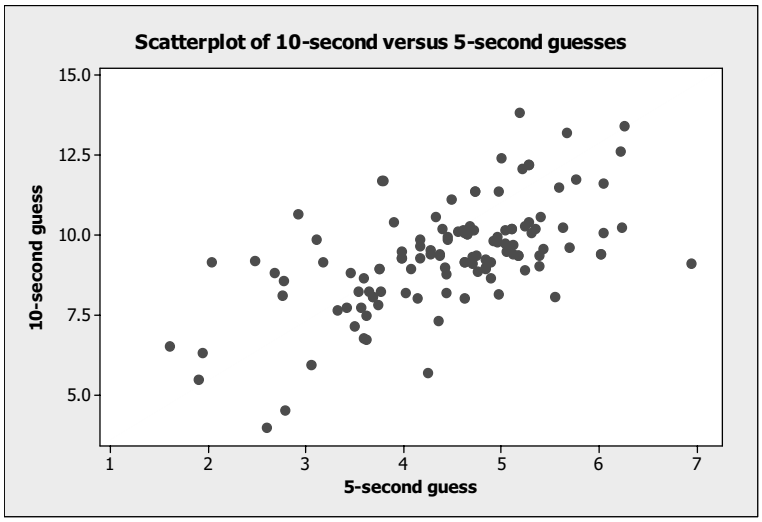
Scatterplots are good visual indications of the relationships between variables. For example, scatterplots can display students' height and weight (or age and height) and students are able to use the display to make comments about the possible forms of the relationship between weight and height. If weight generally increases as height increases, students may comment that this appears to be a positive relationship. If weight tends to decrease as height increases, this indicates a negative relationship. If an increase in height brings no average change in weight, there is no apparent relationship.

In the graph below of height versus weight for 12-year-old students, students could investigate the advantages and disadvantages of using an origin other than (0, 0) for such a scatterplot. They could also discuss which is the better variable for the vertical axis. In the case of the height–weight graph below, a little thought and discussion should lead students to agree that it would be more appropriate to put weight on the vertical scale and height on the horizontal, because the main interest tends to be how weight depends on, or changes with, height. In this case, it can also be readily seen that starting the graph at (0,0) makes it difficult to see how the two variables are related. On the other hand, care must be taken in interpreting such graphs when the origin of the vertical axis is not zero. For example, if the scale on the vertical axis in this graph was from 120 to 180, care must be taken not to interpret a height of 160 cm as double a height of 140 cm.

The second scatterplot below is of a sample of people each estimating 10 seconds and 5 seconds (Gemmell, Grandelis, Mawhinney & Picco, reported in MacGillivray 2005, p. 46). In this example, readers can perceive that it does not matter whether the 10-second guess or the 5-second guess is on the vertical axis as the emphasis is on the relationship between the two guesses, although it will matter more if it is known that every person in the group was asked to guess 5 seconds first and then 10 seconds. This illustrates the great importance of knowing and reporting the details of the collection of a data set. In this particular experiment, the participants were in fact all asked to guess 5 seconds first and then 10 seconds, so plotting the 10-second guess on the vertical axis is the more sensible choice here.

This second plot shows that there is a reasonably positive relationship between the two guesses, that is, people who guess longer for 5 seconds also tend to guess longer for 10 seconds. There is, however, considerable variation in guesses, with more variation for those participants who tend to guess the shortest or the longest, particularly those who guess short.

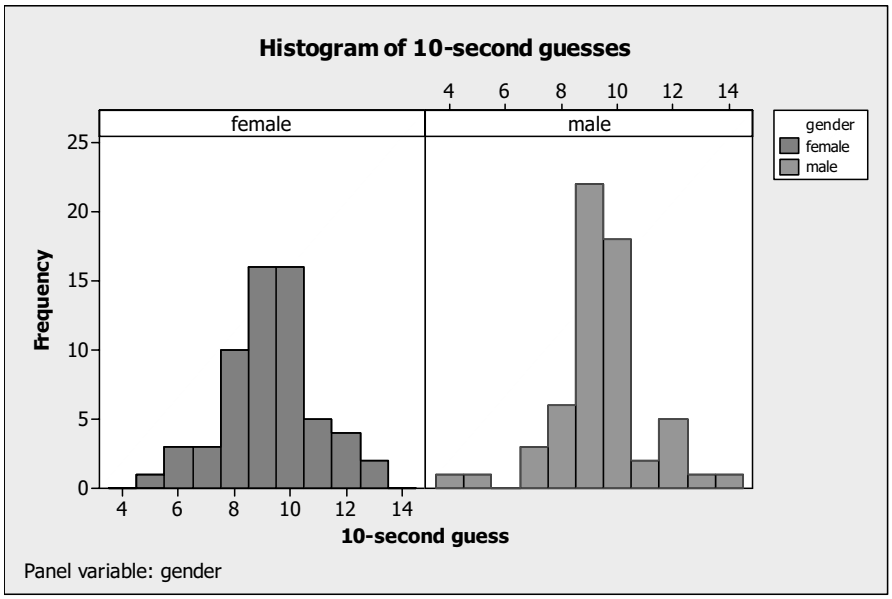
Where a value lies well away from the majority of the other values in a data set, it is often called an *outlier*. Outliers are of interest because of the circumstances that cause them to be outliers.

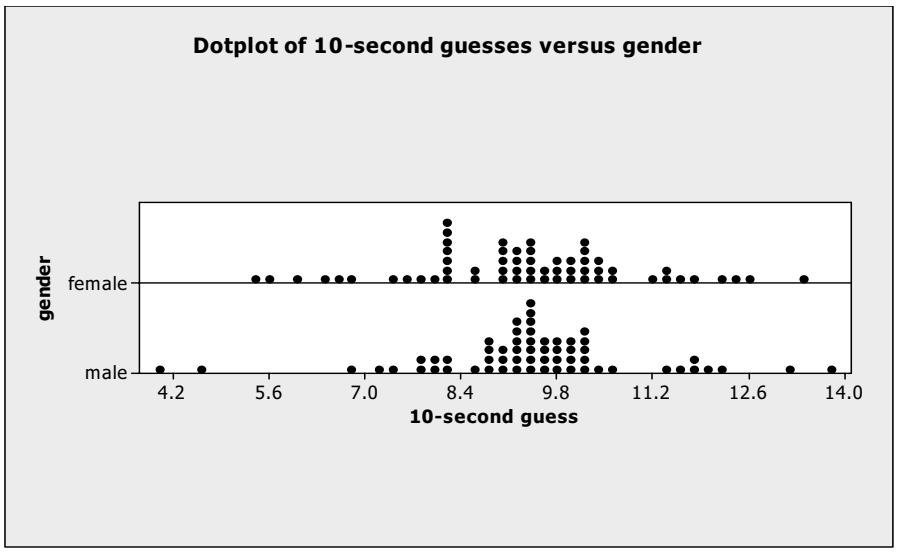


© The State of Queensland (The Office of the Queensland Studies Authority) 2005

A continuous variable and a categorical variable — As students' familiarity with data grows through handling different types of data sets, they will find that some of the most commonly used graphs are those which compare continuous data over two or more categories — for example, heights of males and females, maximum temperatures in spring and autumn, or travel time to school over different days of the week. To help students develop familiarity and confidence with data, it is important that they be given opportunities to compare continuous data over three and more than three categories, not just over two.

Such comparisons can be made using dotplots or histograms, but the most important point for students to discover and remember is that they must be on the same scale. Comparisons of graphs of continuous data are simply not possible unless they are on the same scale. Below are dotplots and histograms of the 10-second guesses in the above experiment, but this time comparing males and females. Students will see that both males and females tended to underestimate rather than over-estimate 10 seconds. Also, generally speaking, there was less variation in guessed time for males than females, but that there was a tendency for a few males to have extreme guesses. Students could discuss which of the two presentations — histogram or dotplot — is their personal preference.

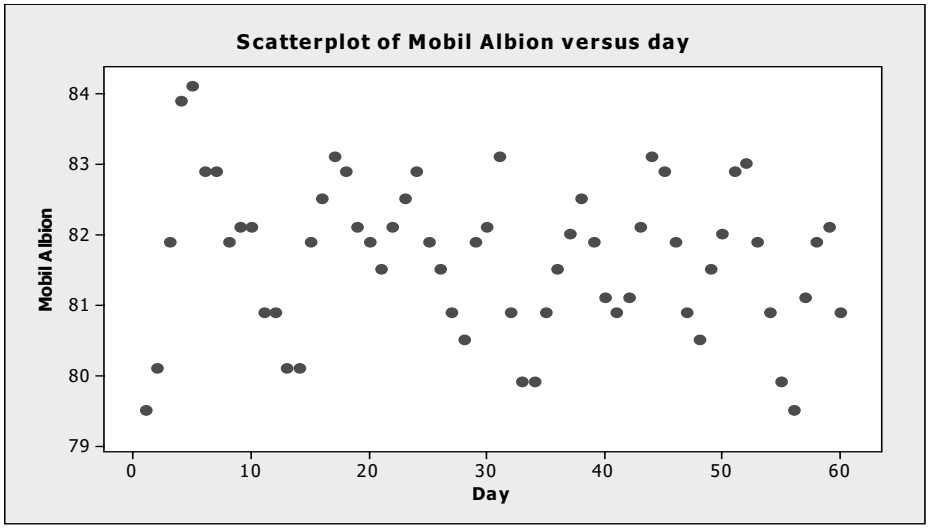


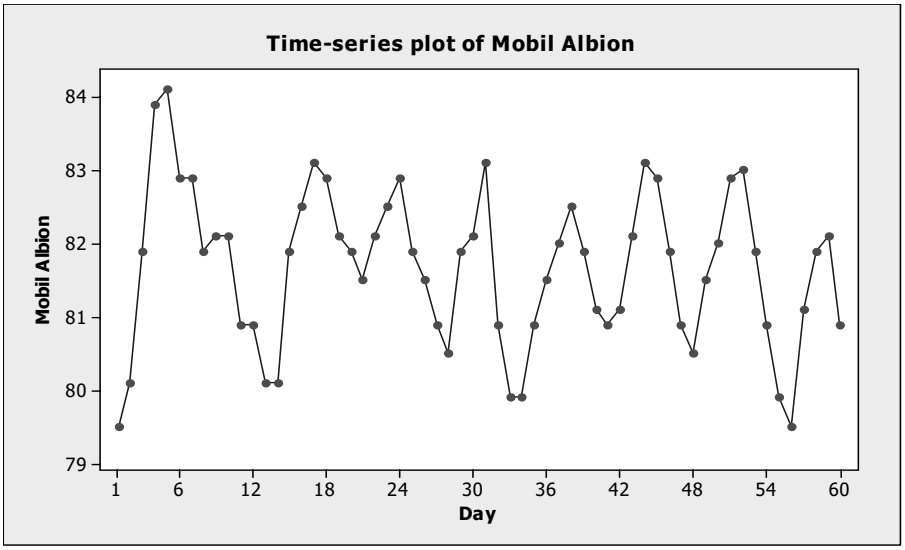


© The State of Queensland (The Office of the Queensland Studies Authority)

Data over time — As students' familiarity with data and types of data sets grows, they will find that in some data sets, the variation over time is a particular focus or of particular interest. Graphs to present such data are essentially scatterplots with time on the horizontal axis; however, in order to focus on changes over time, the points are often joined. For example, the two plots below are of petrol prices at a service station over 60 days (MacGillivray 2005, pp. 13–16, 25–26). The only difference between the two is that in the second one, the points are joined to emphasise that the prices are consecutive over days. The second one is often called a *time-series plot*. Data that are deliberately collected over time with the intent of examining the variation or patterns over time are often called a time series.

In a time-series plot, it may be easier to detect any patterns that may emerge by joining the points as this emphasises the consecutive nature of the data.

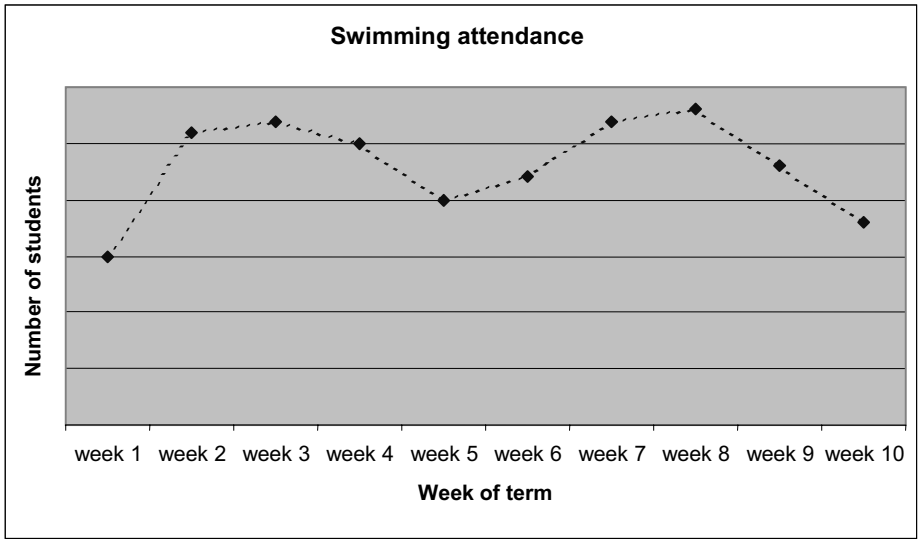


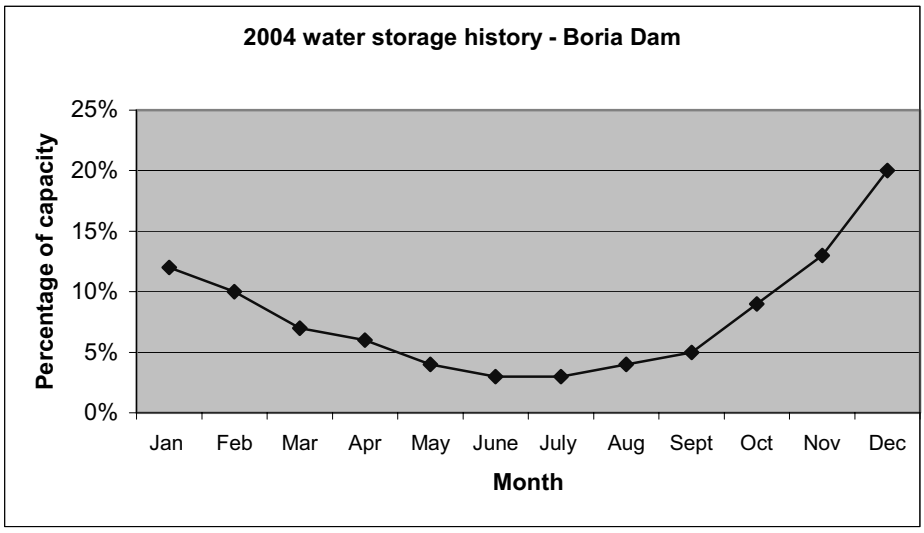


© The State of Queensland (The Office of the Queensland Studies Authority) 2005

In non-statistical computer spreadsheet software, such plots may be called *line graphs*. In the two line graphs below, it does not matter whether the joining lines are dotted or solid. These two graphs do not contain enough essential information for interpretation. In the first graph, is the swimming attendance a total for the week, or on a particular day in each week, or an average over the week? In the second graph, is the capacity an average over the month or is it recorded at the same stage of each month? Discussion of questions such as these in relation to similar data displays would help students learn to provide all the information needed for a reader to be able to interpret a graph.

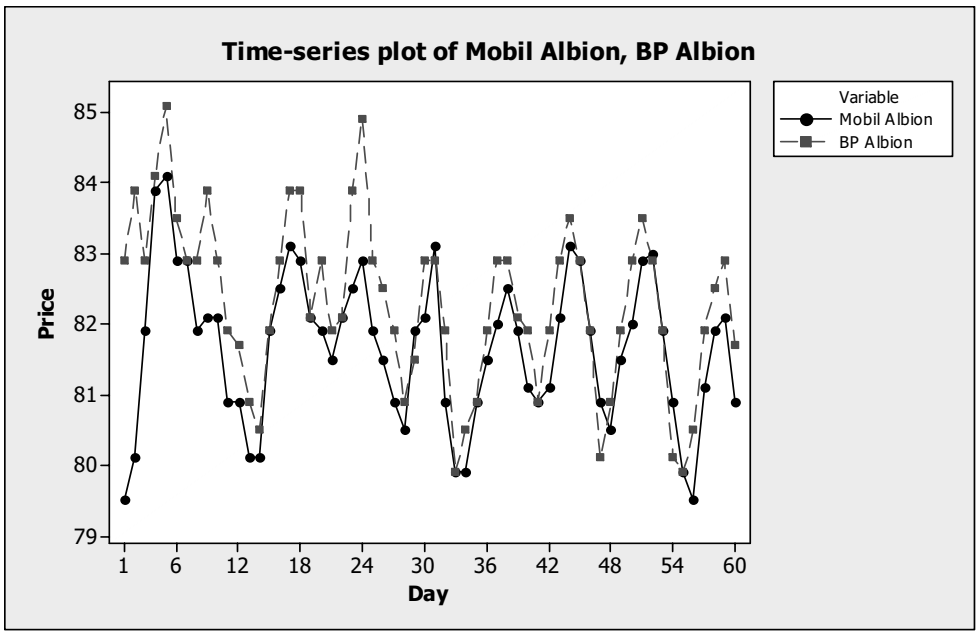
It is important for students to realise that the connecting lines in such graphs do not mean that they can calculate values in between the recording time points. Provided they know the circumstances of the data collection, students could *estimate* an intermediate value. For example, in the graph of the percentage storage capacity in the Boria Dam, if it is known that the measurements were recorded on the 15th of each month, the graph could be used to estimate the storage capacity at an intermediate point. Because nothing is known about how the storage capacity varies between the recording points, the simplest way to estimate the capacity on a day in between the middle of two months is to use the straight line. (This is called linear interpolation.)



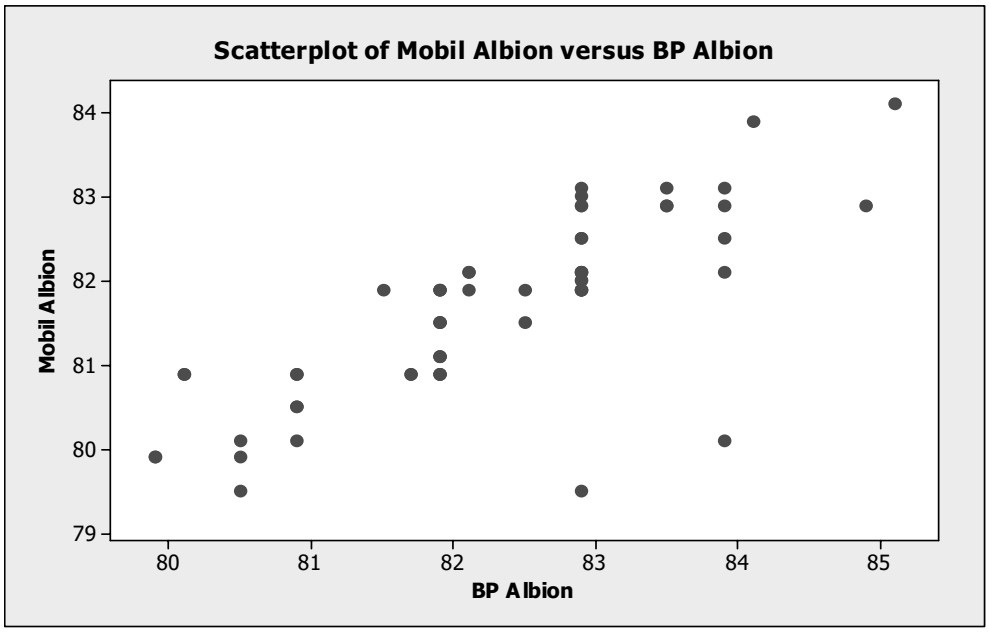


© The State of Queensland (The Office of the Queensland Studies Authority)

A *multiple time-series plot* that includes more than one set of continuous data collected over the *same* time period is often used for making comparisons between similar data over the same period of time. For example, the plot below shows petrol prices at two service stations over the same 60-day period. If the intent is to focus on how the prices at the two service stations relate to each other, this is very difficult to see from the time-series plot and much easier to see in the scatterplot directly comparing the two sets of prices.



The first point on the scatterplot below shows the price on a day when the petrol at both service stations was the same price — 79.9 cents. The second point relates to a day when the price at BP Albion was 80.1 cents and 80.9 cents at Mobil Albion. The scatterplot does not tell us which days these were.



© The State of Queensland (The Office of the Queensland Studies Authority) 2005

Discussing and interpreting variation

Variation or variability in data is its changeability or unpredictability. Variation or variability can occur naturally or because of effects of other variables or experimental or observational conditions. Some variation may be due also to measurement errors or inconsistent collection conditions. By measurement errors, we do not mean the natural variation inherent in taking measurements and using instruments, but errors owing to carelessness or incorrect use of instruments.

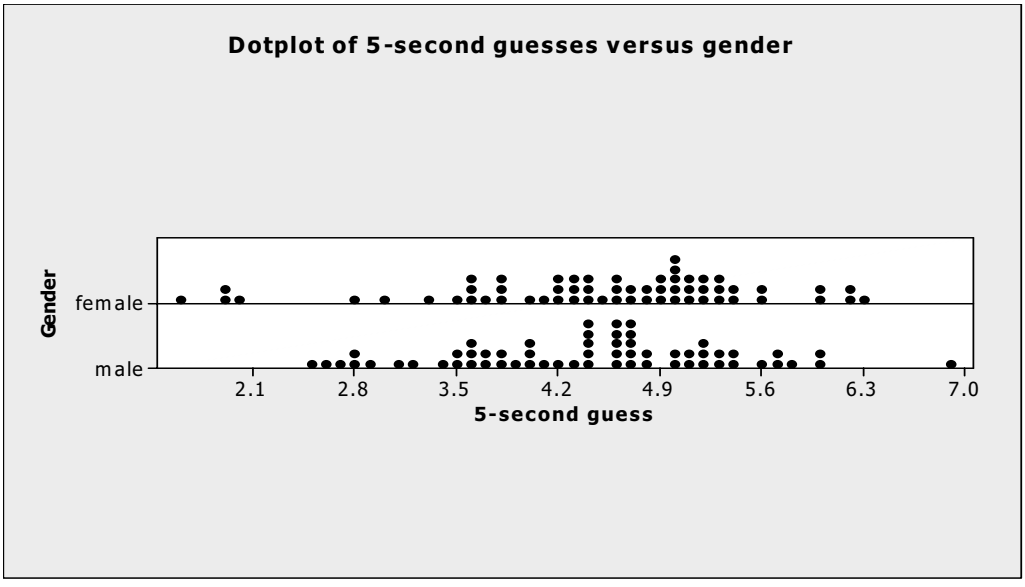
From the early years, students' attention is drawn to variation in data and its possible causes. They engage in discussions about the quality of the data set and what the display of data might indicate about possible patterns or causes, allowing for variation, and the extent of the variation.

Natural variation

To develop understandings about variation, students in the early years examine their data and the methods used to collect it to check that data are free from errors. Students need to understand that most variation occurs naturally — *natural variation* — so that even after some causes might be identified, there will still be variation in the data. For example, natural variation can be expected to occur in data collections about the height of students of the same age. Some of this variation could occur because of gender, or could be explained by different nationalities. In later years, students might think to find out parents' heights. Even after allowing for causes such as gender and family history, however, there will still be natural, unexplained and unexplainable variation. Data analysis aims to describe variation and to allow for it in investigating possible explanations for part of the variation.

Students need to consider the use of different data displays to illustrate variation within data sets, being aware that a suitable graph usually illustrates variation better than a list or a table because of the visual element of the graph.

As students' knowledge and understandings about data develop, their attention should be drawn to the realisation that data analysis and statistical decisions cannot be made just on the basis of graphs or tables.



© The State of Queensland (The Office of the Queensland Studies Authority)

Statistics that represent certain features of data

As students' familiarity and development with data displays grow, they will see that commenting on comparisons in data (as above in comparing males' and females' 10-second guesses), involves commenting on comparisons in location and spread, and then on any unusual features. Students can then be introduced to *measures of location* — median and mean. These measures provide ways of gaining an 'impression' about a set of data and, depending on the type and shape of data in a set, either the median or the mean might be a preferable summary statistic to quote as a measure of location; sometimes both are needed, particularly if there is a considerable difference between them.

Mean and median — Adding all the values in a set of data and dividing by the number of values is the way to calculate the *mean* (often called the average). The *median* of a data set is the value in the middle of the ordered set of data. The mean is more sensitive to extreme values than the median. Outliers within the data do affect the median, although changing the value of a single extreme observation does not. For example, where a cricketer's last five scores when batting are 35; 38; 28; 139 and 30, the median score is 35 and the mean is 54. The median may be a better indicator of this cricketer's usual performances than the mean, but the mean is related to the cricketer's aggregate performance. Both aspects are important in judging a cricketer's record. Students demonstrate their understandings of the mean and median by commenting on their individual choice of the most appropriate measure in a particular situation.

Mode — For discrete data, the *mode* is the category or number with the most observations, and this can be a useful quantity. Students will see that in many real-world examples, the mode occurs at or near the outermost categories. For example, in a survey most respondents might strongly agree with a proposal; in recording the number of accidents at an intersection in a year, the most commonly occurring number might be 0. If two categories or values of discrete data have equal frequencies, or if the second most commonly occurring category or value is separated from the most common, then both should be quoted. For example, if a question in a survey has the most responses in the 'strongly agree' category, and the second most is a non-negligible number in the 'strongly disagree' category, then clearly this is a significant feature of the data.

The concept of 'mode' for continuous data is much more difficult and requires understanding of functions and other calculus-based concepts; that is, understandings beyond those identified in the Years 1 to 10 syllabus. As students experiment with choices of bins in histograms, they will see how choice of bin widths and positions can alter the appearance of a histogram, including which interval has the greatest frequency. As with categorical or count data, modes in continuous data can often occur at the extreme values. For example, in a histogram of the time between phone calls in a call centre, it would not be surprising to see the greatest frequency occurring in the bin starting at 0.

Range and spread — They also learn the terms range and spread. The *range* is simply the difference between the highest and lowest values within the data set, while the *spread* refers to the dispersion of the data — for example, how spread out are the data. Comment on spread will occur mostly in relation to graphs and plots that compare continuous data across categories as in comparing the 10-second time guess for males and females.

Teaching about data

To emphasise the use of data to investigate questions and problems, students should be provided with opportunities to interpret a variety of data displays to determine what the data tells them and what it does not tell them. As their knowledge and understandings develop, students should be challenged to use comparative and quantitative language when working within each data set and between two and more than two data sets (Harrradine, 2004).

For students to critically understand the numerical information presented in a statistical form, they need to have an appreciation of the processes of data collection and presentation. This understanding is of value to them as citizens and will also provide them with knowledge, procedures and strategies that may be applied in areas other than mathematics.

© The State of Queensland (The Office of the Queensland Studies Authority) 2005

Resources

Harradine, A. 2004, 'Distribution division: Making it possible for more students to make reasoned decision using data', unpublished paper prepared for conference in Sweden.

MacGillivray, H.L. 2005, *Data Analysis: Introductory Methods in Context*, 2nd edn, Pearson Education, Australia.

Moore, D.S. 1999, *The Basic Practice of Statistics*, 2nd edition, Freeman New York.

Moore, D.S. & McCabe, G.P. 1993, *An Introduction to the Practice of Statistics*, Freeman, New York.

Salsburg, D. 2002, *The Lady Tasting Tea: How Statistics Revolutionised Science in the Twentieth Century*, Freeman/Owl, New York.

Smith, P.J. 1993, *Into Statistics*, Nelson, Melbourne.

Statistics Canada (website), *Statistics: Power from Data http://www.statcan.ca/english/edu/power/toc/contents.htm*, (accessed 22 February 2004).

Utts, J.M. & Heckard, R.F. 2000, *Mind on Statistics*, Duxbury, Pacific Grove.

Vardeman, S. B. & Jobe, J. M. 2001, *Basic Engineering Data Collection and Analysis*, Duxbury, Pacific Grove.

Acknowledgments

Grateful acknowledgment is made to Professor Helen MacGillivray (Queensland University of Technology) for her valuable contribution to the development of this paper.

The contribution of Anthony Harradine (Noel Baker Centre for School Mathematics, Prince Alfred College, South Australia) is also acknowledged.

 © The State of Queensland (The Office of the Queensland Studies Authority)