

Mathematics Standard Year 11

Statistical Analysis Topic Guidance

Mathematics Standard Year 11 Statistical Analysis Topic Guidance

[Topic focus 3](#_Toc496180977)

[Prior learning 3](#_Toc496180978)

[Terminology 3](#_Toc496180979)

[Use of technology 4](#_Toc496180980)

[Background information 4](#_Toc496180981)

[General comments 4](#_Toc496180982)

[Future study 5](#_Toc496180983)

[Subtopics 5](#_Toc496180984)

[MS-S1: Data Analysis Paperclip icon 5](#_Toc496180985)

[Subtopic focus 5](#_Toc496180986)

[S1.1: Classifying and representing data (grouped and ungrouped) 5](#_Toc496180987)

[Considerations and teaching strategies 5](#_Toc496180988)

[Suggested applications and exemplar questions 7](#_Toc496180989)

[S1.2: Summary statistics 8](#_Toc496180990)

[Considerations and teaching strategies 8](#_Toc496180991)

[Suggested applications and exemplar questions 8](#_Toc496180992)

[MS-S2 Relative Frequency and Probability Paperclip icon 9](#_Toc496180993)

[Subtopic focus 9](#_Toc496180994)

[Considerations and teaching strategies 9](#_Toc496180995)

[Suggested applications and exemplar questions 10](#_Toc496180996)

# Topic focus

*Statistical Analysis* involves the collection, exploration, display, analysis and interpretation of data to identify and communicate key information.

Knowledge of statistical analysis enables the careful interpretation of situations and raises awareness of contributing factors when presented with information by third parties, including the possible misrepresentation of information.

The study of statistics is important in developing students’ understanding of the contribution that statistical thinking makes to decision-making in society and in the professional and personal lives of individuals.

# Prior learning

The material in this topic builds on content from the Statistics and Probability Strand of the K–10 Mathematics syllabus, including the Stage 5.2 substrands of Single Variable Data Analysis and Probability.

# Terminology

|  |  |  |
| --- | --- | --- |
| arithmetic mean  array   back-to-back datasets  bar chart  bimodal  box-plots  categorical data  complement  continuous variable  cumulative frequency graph  data  decile  discrete  dot plot  event  expected frequency  experiment  five-number summary  frequency distribution table  grouped data  histogram  interquartile range (IQR) | mean  measure of central tendency  measure of spread  median  modality  mode  multi-stage  multimodal  negatively skewed  nominal  numerical data  ordinal  outcome  outlier  parameter  Pareto chart ****  percentile  population  positively skewed  probability  quantile  quantitative data | quartile  random sample  random variable  range  relative frequency  sample  sample space  self-selected sample  simulation  spread  standard deviation  stem and leaf plot  stratified sample  summary statistics  symmetric  systematic sample  theoretical probability  tree diagram  ungrouped data  unimodal  variable |

# Use of technology

The use of spreadsheets and/or other software is encouraged in this topic, to enable students to use real data, produce a variety of graphs and tables, create data displays and to calculate measures of location and spread. Spreadsheets are widely used in the workplace and are a suitable tool for tabulating and graphing data and for calculating summary statistics.

A spreadsheet facility allows students to investigate real-world situations and to pose and answer their own ‘what if’ questions. The element of chance can be built into spreadsheets by using random numbers. Students can create spreadsheets to record and display graphically the results of an experiment.

Large and small data sets from real-life situations can be accessed, and sourced on the internet. Students are also encouraged to use data that they have collected themselves to investigate issues that interest them.

Students can make use of contextual data, available on the internet, for example the Australian Bureau of Statistics (ABS) Yearbooks and the Australian Bureau of Meteorology are useful sources of data.

Technology should be used to create frequency tables and statistical graphs, including Pareto charts.

# Background information

This topic provides an introduction to methods of descriptive statistics for finding information in a collection of data and exploring methods for reliably generalising information found in a sample of data to the overall population from which the sample was drawn.

John Graunt, born in 1620, is credited with founding the branch of mathematics now known as statistics. He compiled the first statistical database of birthdays and deaths for 1604–1661 and published an analysis of the causes of death in a diminutive volume with the title *Natural and Political Observations Made upon the Bills of Mortality*. Graunt’s *Observations* summarised a wealth of data, gathered from parish clerks, in a series of tables augmented by brief commentaries, simplistic by the standards of modern inferential statistics.

# General comments

Teaching, learning and assessment materials may use current information from a range of sources including, but not limited to, surveys, newspapers, journals, magazines, bills and receipts, and the internet.

In their study of statistics and society, students can develop an understanding of the importance of analysing data in planning and decision-making by governments and businesses.

Students may have greater interest in probability contexts that relate to their life experiences or to data that they have generated themselves, for example, calculating the risk of injury as a probability for activities such as driving a car or playing contact sport. This data may be collected by survey, measurement or simple experiment. A group surveyed by students may represent the entire population of interest or may represent a sample of the population.

The Australian Bureau of Statistics publishes notes about graph types. Teachers may find these notes useful when giving students experience in the presentation of data displays. Students could collect, display and analyse data related to a course of study in another key learning area, for example, fitness data collected in PDHPE or altitude data in Geography, or results from a scientific experiment.

Student-generated data, obtained from activities such as rolling dice and tossing coins, provide suitable data for analysis.

# Future study

In Year 12, Mathematics Standard 1 students will study the design process for a statistical survey and explore and describe bivariate data.

In Mathematics Standard 2, students will analyse bivariate data and use the normal distribution and -scores to describe and compare datasets.

# Subtopics

* MS-S1: Data Analysis Paperclip icon
* MS-S2: Relative Frequency and Probability Paperclip icon

## MS-S1: Data Analysis Paperclip icon

### Subtopic focus

The principal focus of this topic is planning and management of data collection, classification and representation of data, calculation of summary statistics for single datasets and their use in the interpretation of data.

Through these activities students develop awareness of the importance of statistical processes and inquiry in society.

Within this subtopic, schools have the opportunity to identify areas of Stage 5 content which may need to be reviewed to meet the needs of students.

## S1.1: Classifying and representing data (grouped and ungrouped)

### Considerations and teaching strategies

* Discussions with students may include:
* the role of statistical methods in quality control, for example in manufacturing, agriculture, the pharmaceutical industry, medicine
* issues of privacy and ethics in data collection and analysis
* the role of organisations that collect and/or use statistics, including the Australian Bureau of Statistics (ABS), the United Nations (UN), and the World Health Organisation (WHO).
* Effective questionnaire design includes considerations such as:
* simple language
* unambiguous questions
* requirements for privacy
* freedom from bias
* the number of choices of answers for questions, for example if an even number of choices is given, this may force an opinion from the respondents in relation to a particular question, while for other questions it may be appropriate to allow a neutral choice.
* Students should be given a range of opportunities to determine when it is appropriate to use a sample rather than a census, and to determine the best method of sampling in a range of situations.
* Students should recognise that the purpose of a sample is to provide an estimate for a particular population characteristic when the entire population cannot be accessed.
* The generation of random numbers is fundamental to random sampling.
* Examples of classification of data could include: gender (male, female) categorical, nominal; quality (poor, average, good, excellent) categorical, ordinal; height (measured in centimetres) quantitative, continuous; school population (measured by counting individuals) quantitative, discrete.
* Consideration should be given to the selection of samples, for example for the selection of a stratified sample by age: if 18% of a population is aged under 20, a selected sample should be chosen so that 18% of the people in the sample are under 20.
* Teachers may find it necessary for students to revise:
* construction of tally charts and frequency tables for ungrouped and grouped data and
* the appropriate selection of dot plots, sector graphs, bar graphs, stem-and-leaf plots, histograms, or line graphs, to display datasets, and the selection of appropriate scales.
* A clear distinction should be made between histograms and column graphs or bar graphs, particularly when using spreadsheets to generate graphical representations.
* A histogram is a graphical display of tabulated frequencies. It shows the proportion of cases that fall into each of several class intervals (also called bins). In a histogram, the area of the bar denotes the value, rather than the height of the bar as in a regular column graph or bar graph. This difference is important when the class intervals are not of uniform width. In a histogram, the class intervals do not overlap and must be adjacent. A histogram requires a continuous scale on the horizontal axis, with the frequency plotted on the vertical axis. A histogram is usually a display of continuous data.
* The width chosen for the bins determines the number of bins. It should be noted that there is no single best number of bins and that different bin sizes can reveal different aspects of the data. It is useful to experiment with the bin width in order to best illustrate important features of the data.
* Students can be given opportunities to identify outlying values in a dataset at this stage and to suggest possible reasons for their occurrence.
* The suitability of different types of statistical displays should be compared, for example, a line graph is useful to show trends in data over equal time intervals.
* A Pareto chart is a type of chart that contains both a bar and a line graph, where individual values are represented in descending order by the bars and the cumulative total is represented by the line graph. Instructions for constructing a Pareto chart in Excel can be found at <http://www.wikihow.com/Create-a-Pareto-Chart-in-MS-Excel-2010>.

### Suggested applications and exemplar questions

* Prepare questionnaires and discuss consistency of presentation and possible different interpretations of the questions. Teachers can briefly address the issues of non-response or unexpected response.
* Students can investigate the shuffle mode on an MP3 player as an example of random selection.
* The random number generator on a calculator can be used to generate random numbers, for example, to randomly select a student from the class roll or to play calculator cricket.

Students can collect examples of misleading statistical displays and prepare accurate versions of each. They can describe the inaccuracies, and their corrected version, in appropriate mathematical language.

* The following table gives the six most common reasons for candidates failing their driving test. Display the information in a Pareto chart:

|  |  |
| --- | --- |
| **Reason for failure** | Percentage of candidates |
| Observation at junctions | 11.9% |
| Use of mirrors | 8.2% |
| Inappropriate speed | 5.1% |
| Steering control | 4.7% |
| Reversing around corner | 4.3% |
| Incorrect positioning | 4.2% |

## S1.2: Summary statistics

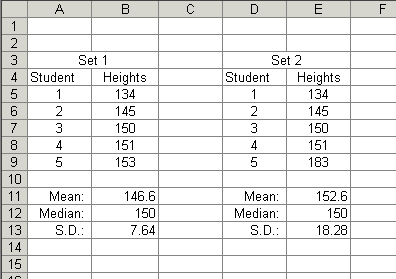
### Considerations and teaching strategies

* Teachers may find it necessary for students to revise calculation of the mean (or arithmetic mean) for small data sets, using the formula

* Teachers can demonstrate the effect of changing value(s) in a data set on the summary statistics. Outlying values could be investigated in this way.
* Class intervals can be restricted to equal intervals only.
* Quartiles can be determined for data sets containing odd and even numbers of data values. In calculating the first and third quartiles, the median scores are excluded. Students should be aware that the second quartile is the median.

### Suggested applications and exemplar questions

* Interpret and evaluate data from students’ own data sets and draw conclusions that can be justified.
* Use a spreadsheet to examine the effect on the calculated summary statistics of changing the value of a score. The spreadsheet below provides such an example. There is only one difference between the two sets of data: for the fifth student in Set 2, the outlying value of 183 cm has the effect of increasing the mean and standard deviation, while leaving the median unchanged.



* Use the following spreadsheet functions on a range of cells containing numerical data to calculate summary statistics: sum, minimum, maximum, mean, mode, median, quartile, standard deviation. In the spreadsheet above the mean in cell B11 is calculated by using the formula: =average(B5:B9).
* A data set of nine scores has a median of 7. The scores 6, 6, 12 and 17 are added to the data set. What is the median of the data set now?
* Using the box-plot, what percentage of drivers in this sample have reaction times of three or more seconds? What percentage of drivers in this sample have reaction times between four and nine seconds? What is the interquartile range for this data set?

Reaction time in seconds prior to braking − drivers over 55

1

0

10

9

2

3

4

5

6

7

8

* The box-plots show the distribution of the ages of children in Numbertown in 2002 and 2012.

There are two box-plots drawn above each other, above a number line indicating age in years. The number line begins at 0 and ends at 18.

The higher box-plot is labelled 2012. The whiskers begin at 0 and end at 10. The box begins at 2 and ends at 12 with a dividing line at 6.

The lower box-plot is labelled 2002. The whiskers begin at 0 and end at 10. The box begins at 8 and ends at 16 with a dividing line at 12.

The number of children aged 12–18 years was the same in both 2002 and 2012. By considering the data, provide advice to town planners about recreational facilities that should be offered, giving statistical reasons.

* Describe and explain any differences expected between the two data sets containing the finishing times in the Olympic Marathon (elite athletes) and in the Sydney Marathon (open entry) in a particular year.

## MS-S2 Relative Frequency and Probability Paperclip icon

### Subtopic focus

The principal focus of this subtopic is to draw conclusions related to the chance that an event will occur.

Students develop awareness of the broad range of applications of probability concepts in everyday life and their use in decision-making.

Within this subtopic, schools have the opportunity to identify areas of Stage 5 content which may need to be reviewed to meet the needs of students.

### Considerations and teaching strategies

* Factorial notation is not required in the Mathematics Standard courses as the number of outcomes for a simple multi-stage experiment is determined by systematic listing.
* Statements involving the language of probability could be collected from various media (newspapers, magazines, radio, television, internet, etc) and discussed.
* Data could be generated from simple experiments, and also obtained from other sources, for example weather and sporting statistics from newspapers. Other data is available from Australian Bureau of Statistics (ABS) Yearbooks, and various websites.
* Practical experiments could involve tossing coins, rolling dice or selecting cards from a pack of cards.
* If an experiment is repeated times, and on each of those times the probability that the event occurs is , then the expected frequency of the event is For example, with ten tosses of a coin , each toss sees the probability of a tail appearing as half , so on average we may see 5 tails appear in ten tosses, but we may actually see 6, or 8, or 4, or… any number from 0 to 10 inclusive.

### Suggested applications and exemplar questions

* Comment critically on statements involving probability, for example ‘Since it either rains or is fine, the probability of a fine day is 50–50’.
* Investigate expressions used in other disciplines and in everyday life to describe likely or unlikely events, for example ‘once in a blue moon’, ‘a one in 300-year flood’, or ‘a 75% chance of recovery following a medical operation’.
* Determine the number of combinations of raised dots that are possible in the Braille system for reading and writing. Investigate whether or not all the possible combinations are used. Students could undertake a similar activity for Morse code.
* Investigate limitations on the number of
* postcodes or telephone numbers that can be used
* possible car number plates available, given selected styles of number plates, for example two letters – two digits – two letters.
* Experiments could be carried out in which the probability is not intuitively obvious, for example probability of a drawing pin landing point up.
* Students may investigate the relative frequency of each different-coloured lolly when selecting a lolly from a packet of lollies of various colours.
* Examine the birth notices on a particular day in a major daily newspaper. Record the number of boys and the number of girls. On this basis, estimate the probability that a child born is (a) male, (b) female. Compare these results with those published by the Australian Bureau of Statistics (ABS).
* In a collection of DVDs, five are rated ‘PG’, three are rated ‘G’, and two are rated ‘M’. If a DVD is selected at random from the collection, what is the probability that it is not rated ‘M’?
* Jo and Lee each have a spinner. Jo’s spinner can land on 1, 2, 3, 4 or 5 and Lee’s spinner can land on A, B, C, or D. They spin their spinners simultaneously.

1. Use an array to list all possible outcomes.
2. What is the probability that the spinners show a D and an even number?

*Solution*

(a)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **A** | **B** | **C** | **D** |
| 1 | A1 | B1 | C1 | D1 |
| 2 | A2 | B2 | C2 | D2 |
| 3 | A3 | B3 | C3 | D3 |
| 4 | A4 | B4 | C4 | D4 |
| 5 | A5 | B5 | C5 | D5 |

(b) There are 20 possible outcomes. Two of these are D with an even number

as highlighted in the array.

P(D with an even number)

* Lou and Ali are on a fitness program for one month. The probability that Lou will finish the program successfully is 0.7, while the probability that Ali will finish it successfully is 0.6. The probability tree diagram shows this information.

The first stage of the tree diagram tree diagram divides into two branches. The first branch leads to Lou successful and there is a label of 0.7 on the branch. The second branch leads to Lou not successful and there is a label of 0.3 on the branch.

The second stage of the tree diagram divides into two branches from each of the previous branches. In each case, one branch indicates Ali successful and is labelled 0.6, and the other is labelled Ali not successful and is labelled 0.4

What is the probability that only one of Lou and Ali will be successful?

* Collect and tabulate rainfall data to predict the probability of rain, for example by tabulating rainfall data from the Bureau of Meteorology, the number of days with rain in Mudgee for the past three summers can be found. This data can be used to predict the probability of rain in summer in Mudgee. This could be extended to predict the probability of two rainy days in succession in summer in Mudgee.