

Mathematics Standard 2 Year 12

Statistical Analysis Topic Guidance

Mathematics Standard 2 Year 12 Statistical Analysis Topic Guidance

[Topic focus 3](#_Toc496190839)

[Terminology 3](#_Toc496190840)

[Use of technology 3](#_Toc496190841)

[Background information 4](#_Toc496190842)

[General comments 4](#_Toc496190843)

[Future study 4](#_Toc496190844)

[Subtopics 5](#_Toc496190845)

[MS-S4: Bivariate Data Analysis  5](#_Toc496190846)

[Subtopic focus 5](#_Toc496190847)

[Considerations and teaching strategies 5](#_Toc496190848)

[Suggested applications and exemplar questions 5](#_Toc496190849)

[MS-S5: The Normal Distribution  7](#_Toc496190850)

[Subtopic focus 7](#_Toc496190851)

[Considerations and teaching strategies 7](#_Toc496190852)

[Suggested applications and exemplar questions 7](#_Toc496190853)

# Topic focus

*Statistical Analysis* involves the collection, 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 statistical analysis is important in developing students’ understanding of how conclusions drawn from data can be used to inform decisions made by groups, such as scientific investigators, business people and policy-makers.

# Terminology

|  |  |  |
| --- | --- | --- |
| bell-shapedbiasbiometric databivariatebivariate datasetcontinuous random variablecorrelationdatasetdependent variableempirical ruleeventextrapolationfrequency distributionhistogram independent variable | interceptinterpolation least-squares regression line ****line of best fit ****linearlinear associationlinear relationshipmeanmediannon-linearnormal curvenormal distributionnormally distributed | numerical variablePearson’s correlation coefficient ****probabilityprobability density functionrandom variablesamplescatterplotslopestandard deviationstandardised scorestatistical investigation processtrendline$z$-score **** |

# Use of technology

Appropriate technology should be used to construct, and determine the equation of a line of fit and least-squares line of best fit, and to calculate correlation coefficients.

Teachers should demonstrate the least-squares regression line on a spreadsheet and then have students explore the function with their own sets of data.

Graphing software can be used to fit a line of best fit to data and make predictions by interpolation or extrapolation.

Real data that is relevant to students’ experience and interest areas can be sourced online.

Online data sources include the [Australian Bureau of Statistics](http://www.abs.gov.au/) (ABS), the [Australian Bureau of Meteorology](http://www.bom.gov.au/) (BOM), the [Australian Sports Commission](http://www.ausport.gov.au/) and the [Australian Institute of Health and Welfare](http://www.aihw.gov.au/) (AIHW) websites.

Spreadsheets or other appropriate statistical software can be used to construct frequency tables and calculate mean and standard deviation.

# Background information

The concept of correlation originated in the 1880s with the works of Sir Francis Galton, Charles Darwin’s cousin. He produced the first bivariate scatterplot, which showed a correlation between children’s height and their parent’s height. A decade later, the British statistician Karl Pearson introduced a powerful idea in mathematics: that a relationship between two variables could be characterised according to its strength and expressed in numbers, leading to the development of Pearson’s correlation coefficient, $r$. This then raised the issue of how to interpret the data in a way that is helpful, rather than misleading. When correlation is mistaken for causation, we find a cause that isn’t there, which is a problem. As science grows more powerful and government relies on big data more and more, the stakes of misleading relationships grow larger. There are many humorous examples on the internet, for example the book and website entitled *Spurious connections*.

The normal distribution was developed from a model originally propounded by Abraham de Moivre, an 18th century statistician and consultant to gamblers. The normal distribution curve is sometimes called the ‘bell curve’ or the ‘Gaussian curve’ after the mathematician Gauss, who played an important role in its development. The normal distribution is the most important and widely used distribution in business, statistics and government. Indeed its importance stems primarily from the fact that the distributions of many natural phenomena are at least approximately normally distributed.

# General comments

Materials used for teaching, learning and assessment should use or include current information from a range of sources, including, but not limited to, newspapers, journals, magazines, real bills and receipts, and the internet.

Students need access to real data sets and contexts. They can also develop their own data sets for analysis or use some of the data sets available online.

Suitable data sets for statistical analysis could include, but are not limited to, home versus away sports scores, male versus female data (for example height), young people versus older people data (for example blood pressure), population pyramids of countries over time, customer waiting times at fast-food outlets at different times of the day, and monthly rainfall for different cities or regions.

This topic provides students with the opportunity to explore aspects of Mathematics involved in any area of special interest to them.

# Future study

Students may be asked to analyse data and produce a report in subjects that they are studying for the HSC or in post-school contexts and training areas. This topic will set a good baseline for knowledge, understanding and skills in statistical analysis.

The ability to analyse and critically evaluate statistical information will provide students with the confidence and skills that help them become discerning citizens.

# Subtopics

* MS-S4: Bivariate Data Analysis 
* MS-S5: The Normal Distribution 

## MS-S4: Bivariate Data Analysis Paperclip icon

### Subtopic focus

The principal focus of this subtopic is to introduce students to a variety of methods for identifying, analysing and describing associations between pairs of numerical variables.

Students develop the ability to display, interpret and analyse statistical relationships related to bivariate numerical data analysis and use this ability to make informed decisions.

### Considerations and teaching strategies

* Outliers need to be examined carefully, but should not be removed unless there is a strong reason to believe that they do not belong in the data set.
* The least-squares line of best fit is also called the regression line. This is the line that lies closer to the data points than any other possible line (according to a standard measure of closeness).
* It should be noted that the predictions made using a line of best fit:
	+ are more accurate when the correlation is stronger and there are many data points
	+ should not be used to make predictions beyond the bounds of the data points to which it was fitted
	+ should not be used to make predictions about a population that is different from the population from which the sample was drawn.
* The ‘trendline’ feature of a spreadsheet graph of a scatterplot should be explored, including the display of the trendline equation (which uses the least-squares method). This feature also allows lines of fit that are non-linear. The ‘forecast’ function on a spreadsheet could then be used to make predictions in relation to bivariate data.
* A Pearson correlation coefficient calculator can be found at <http://www.socscistatistics.com/tests/pearson/>.
* An explanation of how the Pearson correlation coefficient can be calculated using *Excel* is located at <http://www.excelfunctions.net/Excel-Pearson-Function.html>.

### Suggested applications and exemplar questions

* Students could measure body dimensions such as arm-span, height and hip-height, as well as length of stride. It is recommended that students have access to published biometric data to provide suitable and realistic learning contexts. Comparisons could be made using parameters such as age or gender.
* The biometric data obtained should be used to construct lines of fit by hand. The work in relation to lines of fit is extended to include the least-squares lines of best fit and the determination of its equation using appropriate technology.
* Biometric data could be extended to include the results of sporting events, for example the progression of world-record times for the men’s 100-metre freestyle swimming event.
* Predictions could be made using the line of fit. Students should assess the accuracy of the predictions by measurement and calculation in relation to additional data not in the original dataset, and by the value of the correlation coefficient.

For example:

Ahmed collected data on the age ($a$) and height ($h$) of males aged 11 to 16 years.

He created a scatterplot of the data and constructed a line of best fit to model the

relationship between the age and height of males.



1. Determine the gradient of the line of best fit shown on the graph.
2. Explain the meaning of the gradient in the context of the data.
3. Determine the equation of the line of best fit shown on the graph.
4. Use the line of best fit to predict the height of a typical 17-year-old male.
5. Why would this model not be useful for predicting the height of a typical 45-year-old male?
* The height and length of the right foot of 10 high school students were measured. The results were tabulated as follows:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Height (cm) | 165 | 153 | 146 | 138 | 149 | 172 | 170 | 158 | 163 | 154 |
| Right Foot (cm) | 26 | 21 | 20 | 19 | 22 | 24 | 25 | 23 | 22 | 25 |

1. Using technology, calculate the Pearson correlation coefficient for the data.
2. Describe the strength of the association between height and length of the right foot for this dataset.

## MS-S5: The Normal Distribution Paperclip icon

### Subtopic focus

The principal focus of this subtopic is to introduce students to a variety of methods for identifying, analysing and describing associations between pairs of numerical variables.

Students develop the ability to display, interpret and analyse statistical relationships related to bivariate numerical data analysis and use this ability to make informed decisions.

### Considerations and teaching strategies

* Initially, students should explore $z$-scores from a pictorial perspective, investigating only whole number multiples of the standard deviation.
* Students should see a diagram that illustrates the empirical rule in terms of areas under the bell curve. These diagrams are freely available on the internet.
* Teachers explain the importance of $z$-scores as:
	+ raw scores by themselves do not necessarily provide information about the position in a distribution
	+ standardizing different distributions allows us to make comparisons between these distributions.
* Teachers should compare two or more sets of scores before and after conversion to $z$-scores in order to assist explanation of the advantages of using standardised scores.
* Graphical representations of datasets before and after standardisation should be explored.
* Students should extend their understanding of the area under the bell curve to the probability of obtaining any value of $z$-score, as represented in a normal distribution table as illustrated in the following diagram:



* Students should understand that the normal distribution table lists areas under the bell curve to the left of different values of $z$-score, as illustrated in the previous diagram, and should know how to find the probability for a given or calculated $z$-score using the normal distribution table.
* Teachers should briefly explain the application of the normal distribution to quality control and the benefits to consumers of goods and services. Reference should be made to situations where quality control guidelines need to be very accurate, for example the manufacturing of medications.

### Suggested applications and exemplar questions

* Given the means and standard deviations of each set of test scores, compare student performances in the different tests to establish which is the ‘better’ performance.
* Packets of rice are each labelled as having a mass of 1 kg. The mass of these packets is normally distributed with a mean of 1.02 kg and a standard deviation of 0.01 kg. Complete the following table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Mass in kg | 1.00 | 1.01 | 1.02 | 1.03 | 1.04 |
| *z*-score |  |  | 0 | 1 |  |

1. What percentage of packets will have a mass less than 1.02 kg?
2. What percentage of packets will have a mass between 1.00 and 1.04 kg?
3. What percentage of packets will have a mass between 1.00 and 1.02 kg?
4. What percentage of packets will have a mass less than the labelled mass?
* A machine is set for the production of cylinders of mean diameter 5.00 cm, with standard deviation 0.020 cm. Assuming a normal distribution, between which values will 95% of the diameters lie? If a cylinder, randomly selected from this production, has a diameter of 5.070 cm, what conclusion could be drawn?
* Students could investigate whether the results of a particular experiment are normally distributed.
* Find the probability that a person selected at random from a pool of people that took a test on which the mean was 100 and the standard deviation was 15 will have a score:
1. between 100 and 120
2. of a least 120
3. of greater than 120

(normal distribution tables would be used to answer this question)

* The lifetime of a particular make of lightbulb is normally distributed with mean 1020 hours and standard deviation 85 hours. Find the probability that a lightbulb of the same make chosen at random has a lifetime between 1003 and 1088 hours. Normal distribution tables would be used to answer this question.