Year 12 Mathematics Advanced

| MA-S2 Descriptive statistics and bivariate data analysis | Unit duration |
| --- | --- |
| The topic Statistical Analysis involves the exploration, display, analysis and interpretation of data to identify and communicate key information. A knowledge of statistical analysis enables careful interpretation of situations and an awareness of the contributing factors when presented with information by third parties, including its possible misrepresentation. The study of statistical analysis is important in developing students' ability to recognise, describe and apply statistical techniques in order to analyse current situations or to predict future outcomes. It also develops an awareness of how conclusions drawn from data can be used to inform decisions made by groups such as scientific investigators, business people and policy-makers. | 4 weeks |

| Subtopic focus | Outcomes |
| --- | --- |
| The principal focus of this subtopic is to introduce students to some methods for identifying, analysing and describing associations between pairs of variables (bivariate data). Students develop the ability to display, interpret and analyse statistical relationships within bivariate data. Statistical results form the basis of many decisions affecting society, and also inform individual 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. | A student:* solves problems using appropriate statistical processes MA12-8
* chooses and uses appropriate technology effectively in a range of contexts, models and applies critical thinking to recognise appropriate times for such use MA12-9
* constructs arguments to prove and justify results and provides reasoning to support conclusions which are appropriate to the context MA12-10
 |

| Prerequisite knowledge | Assessment strategies |
| --- | --- |
| This topic builds upon the statistics and data concepts explored in Stage 5, calculating with measures of location and spread, and experience with bivariate data, scatter plots and lines of best fit. | **Resource**: ma-s2\_3-how-well-can-mathematics-predict-outcomes.DOCX(includes MA-S3 Random Variables) |

All outcomes referred to in this unit come from [Mathematics Advanced](http://educationstandards.nsw.edu.au/wps/portal/nesa/11-12/stage-6-learning-areas/stage-6-mathematics/mathematics-advanced-2017) Syllabus
© NSW Education Standards Authority (NESA) for and on behalf of the Crown in right of the State of New South Wales, 2017

Glossary of terms

| Term | Description |
| --- | --- |
| **cumulative frequency** | The cumulative frequency is the accumulating total of frequencies within an ordered dataset. |
| **extrapolation** | Extrapolation occurs when the fitted model is used to make predictions using values that are outside the range of the original data upon which the fitted model was based. Extrapolation far beyond the range of the original data is not advisable as it can sometimes lead to quite erroneous predictions. |
| **interpolation** | Interpolation occurs when a fitted model is used to make predictions using values that lie within the range of the original data. |
| **least-squares regression line ** | Least-squares regression is a method for finding a straight line that best summarises the relationship between two variables, within the range of the dataset. The least-squares regression line is the line that minimises the [sum](http://www.mathwords.com/s/sum.htm) of the squares of the [residuals](http://www.mathwords.com/r/residual.htm). Also known as the least-squares line of best fit. |
| **line of best fit ** | A line of best fit is a line drawn through a scatterplot of data points that most closely represents the relationship between two variables. |
| **measures of central tendency** | Measures of central tendency are the values about which the set of data values for a particular variable are scattered. They are a measure of the centre or location of the data. The two most common measures of central tendency are the mean and the median. |
| **measures of spread** | Measures of spread describe how similar or varied the set of data values are for a particular variable. Common measures of spread include the range, combinations of quantiles (deciles, quartiles, percentiles), the interquartile range, variance and standard deviation. |
| **Pareto chart ** | 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. |
| **Pearson’s correlation coefficient ** | Pearson’s correlation coefficient is a statistic that measures the strength of the linear relationship between a pair of variables or datasets. Its value lies between -1 and 1 (inclusive). Also known as simply the correlation coefficient. For a sample, it is denoted by $r$. |
| **population** | The population in statistics is the entire dataset from which a statistical sample may be drawn. |
| **scatterplot** | A scatterplot is a two-dimensional data plot using Cartesian coordinates to display the values of two variables in a bivariate dataset. Also known as a scatter graph. |
| **standard deviation** | Standard deviation is a measure of the spread of a dataset. It gives an indication of how far, on average, individual data values are spread from the mean. |
| **variance ** | In statistics, the variance $Var\left(X\right)$ of a random variable $X$ is a measure of the spread of its distribution. $Var\left(X\right)$ is calculated differently depending on whether the random variable is discrete or continuous. |

| Lesson sequence | Content | Suggested teaching strategies and resources  | Date and initial | Comments, feedback, additional resources used |
| --- | --- | --- | --- | --- |
| Classifying data(1 lesson) | **S2.1: Data (grouped and ungrouped) and summary statistics*** classify data relating to a single random variable

  | **Classifying data*** Explain the criteria for each type of data classification, including displaying the decision tree showing categorical (ordinal, nominal) and quantitative (discrete, continuous).
* **Activity:** Students to work in groups to sort survey topic cards into categorical and quantitative using the template worksheet. Each group moves to the next team’s work to decide what they would move, and records any changes and reasons. Each group moves again to the next team and uses the second template worksheet to sort categorical items into nominal and ordinal and quantitative into discrete and continuous. Students move to the final team’s work and evaluate their choices, making movements and recording them. Students return to their original work to see what has changed, and anything they disagree with, using the template reflection sheet, providing justifications for why they would make any changes. **Resources:** grouping-activity-cards.DOCX, grouping-activity-categorising-data.DOCX**Reflection question –** Which survey question do you disagree with the placement of most? Why?
* Students could choose a topic and design a survey using Google Forms that could be given to the class to complete. Parameters could include that at least one question should produce data in each category. **Reflection question –** Which survey question was the most difficult to write for the topic? Why?
 |  |  |
| Organising, interpreting and displaying data(1-2 lessons) | * organise, interpret and display data into appropriate tabular and/or graphical representations including Pareto charts, cumulative frequency distribution tables or graphs, parallel box-plots and two-way tables **AAM**
	+ compare the suitability of different methods of data presentation in real-world contexts (ACMEM048)
* calculate measures of central tendency and spread and investigate their suitability in real-world contexts and use to compare large datasets
	+ investigate real-world examples from the media illustrating appropriate and inappropriate uses or misuses of measures of central tendency and spread (ACMEM056) **AAM**
* summarise and interpret grouped and ungrouped data through appropriate graphs and summary statistics **AAM**
 | **Organising, interpreting and displaying data*** Overview of data displays and review of Stage 5 as necessary, which may include bar/column charts, dot plots, stem and leaf plots, frequency histograms and polygons.
* Reflect on the differences in categorical and numerical data and how these differences change how data can be displayed. e.g. Continuous data sets can be displayed in a line graph, whereas categorical and discrete data sets should not be.

**Summary statistics*** Students review calculations of summary statistics from Stage 5, including mode, median and mean, range, upper and lower quartiles, interquartile range and standard deviation. Students should be able to calculate and interpret the relevant summary statistics for each data display. **Resource:** displaying-data.DOCX

**NESA examples*** Use a spreadsheet to examine the effect on the calculated summary statistics of changing the value of a score. The following spreadsheet 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.

A spreadsheet with two sets of data. Each set contains the heights of five students. The mean, median and standard deviation are listed beneath each set of data.* 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?

**Measures of central tendency and spread*** Discuss measures of central tendency and spread and explain why in many situations the median is preferred to the mean as a measure of centre. e.g. Outliers affect the mean, but not the median. The median is often used to represent the ‘average’ house price in an area.

**Statistics in the media*** Students identify how statistics are misrepresented or poorly represented in the media.
* Teachers may like to use, ‘A Field Guide to Lies and Statistics’ by Daniel Levitin as a source of examples.
 |  |  |
| Using boxplots to represent and compare data sets (1 lesson) | * organise, interpret and display data into appropriate tabular and/or graphical representations including but not limited to Pareto charts, cumulative frequency distribution tables or graphs, parallel box-plots and two-way tables **AAM**
 | **Boxplots*** Review of five number summary and the use of boxplots to display the summary.
* 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.
* Students could create boxplots to display the heights of male and female students in the class or year group. Questions to consider: How is the data distributed in each case? How do the results compare to one another? What sort of predictions could you make about the distributions for a different year group? Adults?
* 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**101092345678* 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. |  |  |
| Exploring Pareto charts(1 lesson) | * organise, interpret and display data into appropriate tabular and/or graphical representations including but not limited to Pareto charts, cumulative frequency distribution tables or graphs, parallel box-plots and two-way tables **AAM**
 | **Pareto charts*** 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.
* Named after Vilfredo Pareto. The left vertical axis is the frequency of occurrence, cost or other important unit of measure. The right vertical axis is the cumulative percentage of the total number of occurrences, total cost, or total of the particular unit of measure. The values are displayed in descending order, so the cumulative function is concave. Its purpose is to highlight the most important factor/s in a set of data.
* The Pareto principle (also known as the 80/20 rule) states that in many cases, roughly 80% of the effects come from 20% of the causes.

Pareto Chart of Late Arrivals by Reported Cause. Traffic, child care, public transportation, weather, overslept and emergency.Source: [Pareto\_Chart\_Example.png](https://upload.wikimedia.org/wikipedia/commons/2/2e/Pareto_Chart_Example.png) * Consider the analysis of various crimes, where concentration on certain factors associated with occurrence of crime will have the most preventative benefits. Students could explore various sites to explore the 80/20 rule (Pareto Principle): e.g. [Understanding the Pareto Principle](https://betterexplained.com/articles/understanding-the-pareto-principle-the-8020-rule/) or [Center for Problem-Oriented Policing](https://popcenter.asu.edu/content/step-18-learn-if-80-20-rule-applies)
* Students use either prepared examples or research their own examples of the 80/20 statement “80% of the \_\_\_\_\_\_\_\_\_ can be improved by addressing 20% of the \_\_\_\_\_\_\_\_” to complete their own example of a Pareto Chart.

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

Reasons for failure of driving test.  Observation at junctions, use of mirrors, inappropriate speed, steering control, reversing around corner, and incorrect positioning. **Cumulative frequency distribution tables and graphs*** Explore examples of cumulative frequency distribution tables to graph the results for both ungrouped and grouped data. Class intervals can be restricted to equal intervals only.
* An example of each can be found at [Cumulative frequency table](https://www.onlinemathlearning.com/cumulative-frequency-table.html).

Summarise the data using appropriate summary statistics. |  |  |
| Identifying outliers(1 lesson) | * identify outliers and investigate and describe the effect of outliers on summary statistics
	+ use different approaches for identifying outliers, for example consideration of the distance from the mean or median, or the use of below $Q\_{1}-1.5 × IQR$ and above $Q\_{3} + 1.5 × IQR$ as criteria, recognising and justifying when each approach is appropriate
	+ investigate and recognise the effect of outliers on the mean, median and standard deviation
 | **Identifying outliers*** Class discussion about outliers, with consideration for the distance of the scores from the median or mean, as well as considering the effect of an outlier on the mean, median and standard deviation.
* Introduction of the syllabus definition as meeting the criteria of being below $Q\_{1}-1.5 × IQR$ or above$Q\_{3} + 1.5 × IQR$
* Complete a modelling activity such as [An outstanding cricketer](https://topdrawer.aamt.edu.au/Statistics/Downloads/An-outstanding-cricketer-Student-worksheet), using desmos, Geogebra or graphing calculator.
* Outliers need to be examined carefully in any data set, but should not be removed unless there is a strong reason to believe that they do not belong in the data set.
 |  |  |
| Describing, comparing and interpreting distributions(1 lesson) | * describe, compare and interpret the distributions of graphical displays and/or numerical datasets and report findings in a systematic and concise manner **AAM**
 | **Describing, comparing and interpreting distributions*** Display three different histograms that illustrate symmetric, positively skewed or negatively skewed and describe the features of each type of distribution.
* Examine a number of different displays and determine the skewness of each.
* The teacher explains that the type of graph you use is not as important as the interpretation that accompanies the picture. Look for these important characteristics:
	+ Location of the centre of the data
	+ Shape of the distribution of data
	+ Unusual observations in the data set
* Examine other types of graphs for students to work through each and comment on its characteristics. Other types of graphs could include dot plots, stem and leaf plots and box plots.
 |  |  |
| Investigating correlation(1 lesson) | **S2.2: Bivariate data analysis*** construct a bivariate scatterplot to identify patterns in the data that suggest the presence of an association (ACMGM052)
* use bivariate scatterplots (constructing them where needed), to describe the patterns, features and associations of bivariate datasets, justifying any conclusions **AAM**
	+ describe bivariate datasets in terms of form (linear/non-linear) and in the case of linear, also the direction (positive/negative) and strength of association (strong/moderate/weak)
	+ identify the dependent and independent variables within bivariate datasets where appropriate
	+ describe and interpret a variety of bivariate datasets involving two numerical variables using real-world examples in the media or those freely available from government or business datasets
 | **Constructing scatterplots*** Students could use the website [Plotting Scatter Graphs](https://www.transum.org/Maths/Activity/Scatter_Graphs/Plotting.asp) to plot various bivariate data sets. Similarly, students could plot these data sets in desmos. This data can be used to prompt discussions around the description of the data sets; whether it is linear/non-linear, positive/negative and the strength of the association.
* Discussion around appropriate language when investigating the relationship between two variables, including use of terms ‘correlation’ and ‘causation’. This would include discussion around independent and dependent variables. Students should also consider if there might be a third variable responsible for the relationship.
* Design a short group matching task using the examples from the [Correlation](https://www.transum.org/Maths/Activity/Correlation/Default.asp?Level=0) activity on transum.org to distinguish between examples of positive, negative and no correlation.
* Working in pairs, students complete a matching activity. Students match the definition and word to the given graph in the activity board and then give examples (ice cream sales vs temperature) and non-example (hair length vs IQ) of each case. Students compare responses in class discussion. **Resource:** correlation-matching-activity.DOCX
 |  |  |
| Calculating and interpreting Pearson’s correlation coefficient(1 lesson) | * calculate and interpret Pearson’s correlation coefficient $(r)$ using technology to quantify the strength of a linear association of a sample (ACMGM054)
 | **Pearson’s correlation coefficient, r*** Prior to the introduction of correlation coefficient definition, students work in groups to match images of graphs and descriptions of a correlation to a coefficient. Students justify their decisions to the group/class. Teacher reveals classification of values and value ranges of different correlation coefficients.
* Students can use the [Correlation Calculator](https://www.mathsisfun.com/data/correlation-calculator.html) to investigate the correlation between sets of data.
* An explanation of how the Pearson correlation coefficient can be calculated using *Excel* is located at [The Excel PEARSON Function](http://www.excelfunctions.net/Excel-Pearson-Function.html). **Resources:** pearsons-using-microsoft-excel.DOCX and pearsons-using-a-scientific-calculator.DOCX.
* Students to complete a modelling activity to analyse behavioural and social trends using correlation coefficients. **Resource**: modelling-with-excel.DOCX

**NESA example*** The height and length of the right foot of 10 high school students were measured. The results were tabulated as follows:

Height and right foot in cm. 165 and 26, 153 and 21, 146 and 20, 138 and 19, 149 and 22, 172 and 24, 170 and 25, 18 and 23, 163 and 22, 154 and 25.* + Using technology, calculate the Pearson correlation coefficient for the data.
	+ Describe the strength of the association between height and length of the right foot for this dataset.
 |  |  |
| Developing and interpreting linear models(2 lessons) | * model a linear relationship by fitting an appropriate line of best fit to a scatterplot and using it to describe and quantify associations **AAM**
	+ fit a line of best fit to the data by eye and using technology (ACMEM141, ACMEM142)
	+ fit a least-squares regression line to the data using technology (ACMGM057)
	+ interpret the intercept and gradient of the fitted line (ACMGM059)
 | **Modelling linear relationships** * Explain that we collect data to identify trends and make predictions based on the model which can have commercial or environmental implications. Some examples could include: age and height; examination scores and study time; years of education and income; amount of traffic and travel time; size of ship and cargo capacity.
* 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.
* Potential data sources include information we collect ourselves or data from other sources such as the [World Bank](https://www.worldbank.org/) or [Google Trends](https://trends.google.com/trends/?geo=US).
* Student can go to the desmos activity [Line of best fit](https://www.desmos.com/calculator/qn6dwxsji9). Teacher to guide students through a set of activities exploring the line of best fit by eye using technology and compare it with the least-squares regression line. **Resource**: line-of-best-fit.DOCX
* The least squares regression line is the line that minimises the sum of the squares of the residuals. The residuals are defined as being the vertical distances between each point and the line. The sum of the points above the line is always equal to the sum of the points below the line.
 |  |  |
| Analysing and predicting with a linear model(2 lessons) | * use the appropriate line of best fit, both found by eye and by applying the equation of the fitted line, to make predictions by either interpolation or extrapolation **AAM**
	+ distinguish between interpolation and extrapolation, recognising the limitations of using the fitted line to make predictions, and interpolate from plotted data to make predictions where appropriate
* solve problems that involve identifying, analysing and describing associations between two numeric variables **AAM**
* construct, interpret and analyse scatterplots for bivariate numerical data in practical contexts **AAM**
	+ demonstrate an awareness of issues of privacy and bias, ethics, and responsiveness to diverse groups and cultures when collecting and using data
 | **Using a line of best fit to interpolate and extrapolate*** For each example where a student is required to construct a line of best fit for a model, discuss the limitations of the model for the given context. e.g. If the data is comparing arm span and foot size noting that the model cannot be extended indefinitely in both directions.
* Interpolation and extrapolation occur when the linear model is used to answer questions or make predictions. Interpolating occurs when reading from the graph **between** the given points in the data set. Extrapolating occurs when reading from the graph **outside** of the given points in the data set.
* Students could conduct a more rigorous investigation around [Alcohol consumption vs. GDP per capita](https://ourworldindata.org/alcohol-consumption#alcohol-consumption-vs-income). The data can be downloaded as an Excel spreadsheet so that it could be explored in more detail for subsets of the data, for example.

**NESA example*** 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.

A graph showing the age and height of males and a line of best fit for the data.* + Determine the gradient of the line of best fit shown on the graph.
	+ Explain the meaning of the gradient in the context of the data.
	+ Determine the equation of the line of best fit shown on the graph.
	+ Use the line of best fit to predict the height of a typical 17-year-old male.
	+ Why would this model not be useful for predicting the height of a typical 45-year-old male?
 |  |  |

Reflection and evaluation

Please include feedback about the engagement of the students and the difficulty of the content included in this section. You may also refer to the sequencing of the lessons and the placement of the topic within the scope and sequence. All ICT, literacy, numeracy and group activities should be recorded in the ‘Comments, feedback, additional resources used’ section.