Higher tier content

A variety of contexts, including real-life data, will be used. (No detailed knowledge of those contexts will be expected.) Each item of content will have a code, for example **1a.01**, where **1** is 'The collection of data', **a** is 'Planning' and **01** is the numerical order of the content item.

What students need to learn:		Guidance
1. The collection of data		
(a) Planning		
1a.01	Know that a hypothesis can be tested only through the appropriate collection and analysis of data. Formal use of null hypothesis will not be required. 2H	Specifying a hypothesis is expected, e.g. a hypothesis such as 'as motor cycles get older their value is likely to go down'.
1a.02	Know the constraints that may be faced in designing an investigation to test a hypothesis including factors such as time, costs, ethical issues, confidentiality, convenience. 2H	Give examples of these factors, e.g. salaries or difficulties in finding data.
1a.03	Determine proactive strategies to mitigate issues that might arise during the statistical enquiry process.	For example dealing with difficulties in identifying the population, non-response issues or unexpected outcomes.
(b) Types of data		
1b.01 ?	Know and apply terms used to describe different types of data that can be collected for statistical analysis: raw data, quantitative, qualitative, categorical, ordinal, discrete, continuous, ungrouped, grouped, bivariate and **multivariate**. 2H Types of variable not types of data	Use of correct statistical terminology to describe given data is expected. Know that more than one term may be appropriate. Identification of variables relevant to an investigation or hypothesis is expected.
1b.02	Know the advantages and implications of merging data into more general categories, and of grouping numerical data into class intervals.	Expected to know class width, and implications of grouping data, e.g. loss of accuracy in both calculations and presentations.
1b.03	Know and apply the terms explanatory (independent) variables and response (dependent) variables. 2H	Know that on a scatter diagram the explanatory (independent) variable should be on the 'x' axis.
1b.04	Know the difference between primary and secondary data. 2H	Including advantages and disadvantages of each. Consideration of the reliability and accuracy of the data (including issues of rounding) and the recognition of possible constraints in accessing the data is expected.

Grouped data explicit on Foundation list.

Higher only explicitly on Foundation list

Higher tier content

What students need to learn:		Guidance
(c) Population and sampling		
1c.01	Know the difference between population, sample frame and sample.	Identify a population, and suggest a suitable sampling frame.
1c.02	Know that 'population' can have different meanings within a stated context.	For example, all employees in an office; all females in the UK; all items produced in a factory.
1c.03	Know reasons for employing judgement sampling or opportunity (convenience) sampling, and the associated risks of bias when these techniques are used. 1H	Including use of cluster sampling and quota sampling. Reasons including factors such as convenience, cost and time.
1c.04	a. Know appropriate sampling techniques in the context of the problem to avoid bias. b. Understand random, systematic, and quota sampling. 1H	Including advantages and disadvantages of each technique. e.g. Know that systematic and quota sampling techniques are generally non-random. Know that the period of systematic sampling may coincide with a period occurring in the data.
1c.05	Know the key features of a simple random sample and demonstrate understanding of how different techniques, both physical and electronic, are used to select random members from a population: including, but not limited to, dice, cards, random number lists, and calculator functions. 1H	Be aware that all items in the population should have the same likelihood of inclusion in a simple random sample. Selection of items for a sample may be required, including dealing with issues such as repeated random numbers and random numbers out of range.
1c.06	Use stratification and know when this is appropriate before sampling takes place. 1H	Identify suitable strata, e.g. gender or age group. Including the calculation of appropriate strata sizes. Stratifying by one **or more than one** category.

What students need to learn:		Guidance
(d) Collecting data		
1d.01	a. Know that data can be collected from different sources: experimental (laboratory, field and natural), simulation, questionnaires, observation, reference, census, population and sampling. b. Know that sources of secondary data should be acknowledged.	The design of data collection sheets is expected. Simulations may include use of random numbers.
1d.02	Know the importance of reliability and validity with regards to collected data.	Reliability is the extent to which repeated measurements yield similar results. Validity is the extent to which a test measures what was intended.
1d.03	Determine factors that may lead to bias, including issues of sensitivity of the content matter, **level of control** and know how to minimise data distortion.	**Know the 'random response' technique for sensitive questions.**
1d.04	Know the key features to be considered when planning data collection: leading questions, avoiding biased sources, time factors, open/closed questions, different types of interview technique.	The design of suitable questions and data collection sheets is expected. Awareness of the advantages and disadvantages of data collection techniques. The rationale behind pilots for questionnaires and pre-tests for experiments should be known.
1d.05	Know and demonstrate understanding of techniques used to deal with problems that may arise with collected data.	For example, missing data, incorrect formats, non-responses, incomplete responses, etc.
1d.06	Know why data may need to be 'cleaned' before further processing, including issues that arise on spreadsheets and apply techniques to clean data in context.	In the pre-processing stage: consideration of genuine and other outliers and anomalies, or removal of extraneous symbols or notation when using technology (e.g. spreadsheets, statistical software). See also 2c.03.
1d.07	Know the importance of identifying and controlling extraneous variables **and the use of control groups**.	**Understand the advantage of using matched pairs when using control groups.**

Higher tier content

Choropleth - Foundation list only.

What students need to learn:		Guidance
2. Processing, representing and analysing data		
(a) Tabulation, diagrams and representation		
2a.01	Represent data sets pictorially using calculated key values as necessary, and interpret and compare data sets displayed pictorially: tabulation, tally, pictogram, pie chart, stem and leaf diagram, Venn diagram.	Use of two-way tables is expected. Diagrams should have a key where appropriate. Stem and leaf diagrams need to be ordered to allow identification of key values.
2a.02	Interpret and compare data sets displayed pictorially: population pyramid, choropleth map, **comparative pie chart, comparative 2D representations, comparative 3D representations**.	Interpretation of data sets in tabular form is expected. **The relationship between area and frequency, and calculations of radius, for comparative pie charts is expected.**
2a.03	Represent data sets graphically using calculated key values as necessary, and interpret and compare data sets displayed graphically: bar charts, line graphs, time series, scatter diagrams, bar line (vertical line) charts, frequency polygons, cumulative frequency (discrete and grouped) charts, histograms (equal class width), and box plots.	Use of multiple and composite (including percentage composite) bar charts is expected. No distinction will be made between cumulative frequency polygons (other than step polygons) and curves, while frequency polygons could be open or closed. Note: the 'y' axis of histograms may be labelled 'frequency' where equal class widths are used.
2a.04	**Calculate and use frequency density to draw histograms (unequal class width), and interpret and compare data sets displayed in histograms (unequal class width).**	**Students are required to know the formula for frequency density (see *Appendix 2*).** **Correct labelling of frequency density axis or use of an appropriate key) will be expected. (But see note in 2a.03)** **Use of a standard class width with appropriate units will be acceptable.**

What students need to learn:		Guidance
(a) Tabulation, diagrams and representation *continued*		
2a.05	Justify the appropriate format and produce accurate visualisation of data. 1H	Be familiar with the capabilities and advantages of using statistical software and spreadsheets to produce suitable diagrams and graphs, and know to avoid the inappropriate use of such technology. Appropriate format could take account of target audience. e.g. realising when a simple visualisation of data is appropriate, and when a more technical visualisation is appropriate.
2a.06	Recognise where errors in construction lead to graphical misrepresentation, including but not limited to incorrect scales, truncated axis, distorted sizing **or the misuse of formula when calculating the frequency densities of histograms**. 2H	Correct use of class boundaries is required**, including in the calculation of frequency densities**. Understand the possible distortion when interpreting 3D representations.
2a.07	Extract and calculate corresponding values in order to compare data sets that have been presented in different formats and be able to present the same information in multiple formats. 1H 2H	Including extracting information from spreadsheets, lists of statistics or graphs produced by statistical software.
2a.08	Select and justify appropriate form of representation with regard to the nature of data. 1H	e.g. scatter diagrams for bivariate data, histograms for grouped data, etc.
2a.09	Determine skewness from data by inspection **and by calculation.** **Use of:** $\textbf{Skew} = \frac{\textbf{3(mean} - \textbf{median)}}{\textbf{standard deviation}}$ 1H **Formula will be given in the formulae sheet.**	For example, know that, for positive skew could be indicated by: • mean > median > mode • median − LQ < UQ − median
2a.10	Interpret a distribution of data in terms of skewness identified from inspection **or calculation.** 1H	For example, with positive skew know that values above the median have a greater spread than values below the median.

What students need to learn:		Guidance
(b) Measures of central tendency		
2b.01	Calculate averages for discrete and grouped data: mode, median, arithmetic mean, **weighted mean, geometric mean, mean seasonal variation**. The term 'mean' should be understood to be 'arithmetic mean' **unless 'geometric mean' is stated**. 1H 2H	Calculations of mean and median for grouped data will include equal **or unequal** class widths. Linear interpolation for median is expected. 1H Use of class midpoints (mid-interval values) to estimate mean of grouped data is expected. Understand the effect on the mean, mode and median of changes in the data, including the addition or withdrawal of a population or sample member. Understand the effect of transformations of the data on the mean, mode and median. (Transformations will be restricted to simple scaling and translations.)
2b.02	Justify the rationale for selecting appropriate types of average in context. 1H 2H	e.g. mode is appropriate when considering demand for items of clothing in different sizes, or when data is non-numeric; e.g. median more appropriate than mean if data is skewed; etc e.g. mean is appropriate to take account of all data **and allows calculation of standard deviation**

What students need to learn:		Guidance
(b) Measures of central tendency *continued*		
2b.03	Compare different data sets using appropriate calculated or given measure of central tendency: mode, modal class, median and mean. 1H 2H	An awareness of which measure is more appropriate to use is expected, e.g. selecting the appropriate values from those produced by statistical software.
(c) Measures of dispersion		
2c.01	Calculate different measures of spread: range, quartiles, interquartile range (IQR), percentiles, **interpercentile range, interdecile range and standard deviation**. 1H 2H	For example, 10th to 90th interpercentile range. Any value of n may be expected, so that required bounds (e.g. quartiles) may or may not be values in the data set. Alternative methods will be acceptable provided that the method used is clear from the working. (e.g. if the median lies between two data values the arithmetic mean of these two values may be used.) **For standard deviation only the formulae for a set of values are given. Students will need to know how to apply these to grouped data, i.e.** **Standard deviation =** $\sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$ **or** $\sqrt{\frac{\sum fx^2}{\sum f}-\left(\frac{\sum fx}{\sum f}\right)^2}$
2c.02	Identify outliers by inspection **and using appropriate calculations.** 2H	**Calculations are expected to be known:** **Small outlier is** **< LQ – 1.5 × IQR** **Large outlier is** **> UQ + 1.5 × IQR** **Or outlier is outside** $\mu \pm 3\sigma$

What students need to learn:		Guidance
(c) Measures of dispersion *continued*		
2c.03	Comment on outliers with reference to the original data.	Know that outliers may be genuine unusual values or may be the result of errors in recording data. **Outlier boundaries may need to be calculated.**
2c.04	Compare different data sets using appropriate calculated or given measure of spread: range, interquartile range (IQR), percentiles **and standard deviation**.	An awareness of which measure is more appropriate to use is expected, e.g. selecting the appropriate values from those produced by statistical software.
2c.05	Use calculated or given median and interquartile range (IQR) **or interpercentile range or interdecile range or mean and standard deviation** to compare data samples and to compare sample data with population data.	The appropriate pairing of a measure of central tendency and a measure of dispersion is expected. (e.g. use of mean with IQR is not appropriate.)
2c.06	**Use calculated or given means and standard deviation to standardise and interpret data collected in two comparable samples.** **Formulae for standard deviation will be given in the formulae sheet.**	**Know how to standardise using these values:** **standardised score =** $\frac{x-\mu}{\sigma}$ **(Formula will not be given.)**
(d) Further summary statistics		
2d.01	Use different types of index **and weighted index** numbers in context, including but not limited to retail price index (RPI), consumer price index (CPI) and gross domestic product (GDP).	Calculation and interpretation of simple **and chain based** index numbers is expected.

What students need to learn:		Guidance
(d) Further summary statistics *continued*		
2d.02	Interpret data related to rates of change over time (including, but not limited to, percentage change, births, deaths, house prices, and unemployment) when given in graphical form. Calculate and interpret rates of change over time from tables using context specific formula. 2H 1H	Making predictions using rates of change formulae is expected, e.g. crude birth rate = $\frac{\text{number of births} \times 1000}{\text{total population}}$ **standardised birth rate** = $\frac{\text{crude rate}}{1000} \times \text{standard population}$ **Formulae will be given**.
(e) Scatter diagrams and correlation		
2e.01	Know and apply vocabulary of correlation: positive, negative, zero, causation, association, interpolation and extrapolation. 2H	Know that a dependent variable should be plotted on the 'y' axis.
2e.02	Describe and make comparisons of correlation by inspection: strong or weak. 2H	e.g. Informal interpretation using scatter diagrams.
2e.03	Know that correlation does not necessarily imply causation **and multiple factors may interact.** 2H	Be aware of spurious correlation. e.g. car ownership and birth rate in a number of cities may show correlation as both variables are likely to be affected by population size of the cities.
2e.04	Determine line of best fit by eye, by drawing through a calculated double mean point $(\bar{x}, \bar{y})$ **and by using the equation of the regression line.** 2H	**The linear regression line of the form** $y = a + bx$ Awareness of issues relating to interpolation and extrapolation, and the interpretation of gradient and intercept are expected. Non-linear models will not be tested.

What students need to learn:		Guidance
(e) Scatter diagrams and correlation *continued*		
2e.05	**Apply formula to determine Spearman's rank correlation coefficient. Values found using calculator functions will be permissible** 1H	**Formula will be given in the formulae sheet.** **Tied ranks will not be tested. *(Scientific calculator functions are sufficient).***
2e.06	Interpret **calculated or** given Spearman's rank correlation coefficient in the context of the problem. 1H	Be aware that values range on a scale from -1 to +1. Know that values closer to these limits indicate 'stronger' correlation, but no formal interpretation of strength of correlation is expected. e.g. in comparing ranks given by two judges in a competition know that +1 means perfect agreement, -1 means complete opposite ranks, and 0 means no agreement between ranks given.
2e.07	**Interpret given Pearson's product moment correlation coefficient (PMCC) in the context of the problem.** 1H	**Be aware that values range on a scale from -1 to +1. Know that values closer to these limits indicate 'stronger' linear correlation, but no formal interpretation of strength of correlation is expected.** **Know that +1 means perfect linear positive correlation, -1 means perfect linear negative correlation, and 0 means no linear correlation.** **The calculation of PMCC will not be required.**
2e.08	**Understand the distinction between Spearman's rank correlation coefficient and Pearson's product moment correlation coefficient (PMCC).** 1H	**e.g. recognise the relative strengths of rank correlation and product moment correlation on a scatter graph.** **The PMCC measures the strength of linear correlation.** **The calculation of PMCC will not be required.** **e.g. if there is positive non-linear correlation both coefficients will be positive but Spearman's coefficient will be greater than PMCC.**

What students need to learn:		Guidance
(f) Time series		
2f.01	Identify trends in data through inspection and by calculation of 4 **or other determined appropriate** point moving averages. 1H	Drawing a trend line either by eye or by using averages. Interpretation of the gradient of trend lines is expected.
2f.02	Interpret seasonal and cyclic trends in context. **Use such trends to make predictions.** 1H	**Demonstrating the calculation of predictions, using average seasonal effect, is expected. Awareness of the dangers of extrapolation when making predictions is expected.**
(g) Quality assurance		
2g.01	**Know that a set of sample means are more closely distributed than individual values from the same population.** 2H	**e.g. the set of mean heights from each class in a school will show less variation than the set of heights of all students in the school.**
2g.02	**Use action and warning lines in quality assurance sampling applications.** 2H	**Control charts used for sample mean, median or range is expected.** **For example, in a manufacturing process to test that certain measurements are within allowable limits.** **Understand that almost all means, medians or ranges fall inside the action lines (action limits), and only 1 in 20 fall outside the warning lines (warning limits).** **Know that warning lines are set at ±2 standard deviations of the sample mean from the expected value, and action lines are set at ±3 standard deviations of the sample mean from the expected value.** **Know the action to be taken if a sample value falls outside each type of limit.**

Higher tier content

What students need to learn:		Guidance
(h) Estimation		
2h.01	Use calculated or given summary statistical data to make estimates of population characteristics. Use samples to estimate population mean. Use sample data to predict population proportions. 2H	e.g. predict that approximately half the population will be above the sample median.
2h.02	**Apply Petersen capture recapture formula to calculate an estimate of the size of a population.** 2H	**Know the assumptions needed and be familiar with their appropriateness in practice.**
2h.03	Know that sample size has an impact on reliability and replication. 2H	e.g. know that results/conclusions are likely to be more reliable if based on larger samples.
3. Probability		
3p.01	Use collected data to calculate estimates of probabilities. 2H	Use of relative frequency.
3p.02	Compare the probability of different possible outcomes using the 0–1 or 0–100% scale and statements of likelihood. 2H	Locate events on a probability scale and use the language of likelihood (e.g. certain, impossible, evens, likely, very unlikely, etc.).
3p.03	Use probability values to calculate expected frequency of a specified characteristic within a sample or population. 2H	Given total frequency, use probability as a proportion to find expected frequency.
3p.04	Use collected data and calculated probabilities to determine and interpret relative risks and absolute risks, and express in terms of expected frequencies in groups. 2H	e.g. use driving test pass rate data with Instructor A and Instructor B to determine the probability (absolute risk) of passing with A, or determine the relative probability (relative risk) of passing with A compared with B. $\text{Relative risk} = \frac{\text{risk of passing with A}}{\text{risk of passing with B}}$
3p.05	Compare experimental data with theoretical predictions to identify possible bias within the experimental design. 2H	e.g. consider whether a set of dice rolls suggests that the dice is fair.

What students need to learn:		Guidance
3. Probability *continued*		
3p.06	Recognise that experimental probability will tend towards theoretical probability as the number of trials increases when all variables are random. 2H	Understand that increasing sample size generally leads to better estimates of probability and population parameters. Students may be expected to estimate probabilities from relative frequency diagrams and frequency tables.
3p.07	Use two-way tables, sample space diagrams, tree diagrams and Venn diagrams to represent all the different outcomes possible for at most three events. 1H Only mentions Venn diagrams. Tree diagrams on Foundation list.	Use of these for conditional probability is expected. (See 3p.09.) Sample space diagrams may include listing or tabulating all outcomes of single events, or successive events, in a systematic way. Understand the terms mutually exclusive and exhaustive. Know the addition law for two mutually exclusive events: $P(A \text{ or } B) = P(A) + P(B)$ **Know the general addition law:** $\mathbf{P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)}$
3p.08	Know and apply the formal notation for independent events. 2H 1H	Understand the difference between independent and conditional events. The multiplication law for independent events must be known: $P(A \text{ and } B) = P(A) \times P(B)$ Know that for independent events: $P(A \mid B) = P(A)$ and $P(B \mid A) = P(B)$
3p.09	Know and apply the formal notation for conditional probability. 1H	The formula for conditional probability must be known: $P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$

What students need to learn:		Guidance
3. Probability *continued*		
3p.10	**Comment on the differences between experimental and theoretical values in terms of possible bias. Formal tests of significance will not be required.** 1H? ←	**e.g. compare observed outcomes with expected frequencies from a binomial model.** Mentions Binomial in list.
3p.11	**Know and interpret the characteristics of a binomial distribution.** 1H	**The notation $\mathrm{B}(n, p)$ may be used.** **Be familiar with mean of a binomial distribution (np) and with the conditions which make a binomial model suitable.** **Calculate probabilities or use given probabilities, which may be found using any standard method, e.g. use of functions on a calculator, spreadsheets, Pascal's triangle. Questions will not be set with n larger than 10.**
3p.12	**Know and interpret the characteristics of a normal distribution.** 2H	**The notation $\mathrm{N}(\mu, \sigma^2)$ may be used.** **Use of normal distribution tables will not be expected.** **Know the distribution is symmetrical with a 'bell' shape, and that median, mean and mode are equal.**
3p.13	**Know that, for a normal distribution, values more than three standard deviations from the mean are very unusual; know that approximately 95% of the data lie within two standard deviations of the mean and that 68% (just over two thirds) lie within one standard deviation of the mean** 2H	**Be familiar with the conditions which make a normal model suitable.** **e.g. that data are continuous, the distribution is symmetrical and bell-shaped, and that mean, median and mode are approximately equal.**