Topic 4: Statistics and probability

Concepts

Essential understandings:

Statistics is concerned with the collection, analysis and interpretation of data and the theory of probability can be used to estimate parameters, discover empirical laws, test hypotheses and predict the occurrence of events. Statistical representations and measures allow us to represent data in many different forms to aid interpretation.

Probability enables us to quantify the likelihood of events occurring and so evaluate risk. Both statistics and probability provide important representations which enable us to make predictions, valid comparisons and informed decisions. These fields have power and limitations and should be applied with care and critically questioned to differentiate between the theoretical and the empirical/observed. Probability theory allows us to make informed choices, to evaluate risk, and to make predictions about seemingly random events.

Suggested concepts embedded in this topic:

Quantity, validity, approximation, generalization.

AHL: Change, systems.

Content-specific conceptual understandings:

- Organizing, representing, analysing and interpreting data and utilizing different statistical tools facilitates prediction and drawing of conclusions.
- Different statistical techniques require justification and the identification of their limitations and validity.
- Approximation in data can approach the truth but may not always achieve it.
- Some techniques of statistical analysis, such as regression, standardization or formulae, can be applied in a practical context to apply to general cases.
- Modelling through statistics can be reliable, but may have limitations.

AHL

- Properties of probability density functions can be used to identify measure of central tendency such as mean, mode and median.
- Probability methods such as Bayes theorem can be applied to real-world systems, such as medical studies or economics, to inform decisions and to better understand outcomes.

SL content

Recommended teaching hours: 27

The aim of the SL content in the statistics and probability topic is to introduce students to the important concepts, techniques and representations used in statistics and probability. Students should be given the opportunity to approach this topic in a practical way, to understand why certain techniques are used and to interpret the results. The use of technology such as simulations, spreadsheets, statistics software and statistics apps can greatly enhance this topic.

It is expected that most of the calculations required will be carried out using technology, but explanations of calculations by hand may enhance understanding. The emphasis is on understanding and interpreting the results obtained, in context.

In examinations students should be familiar with how to use the statistics functionality of allowed technology.

At SL the data set will be considered to be the population unless otherwise stated.

Sections SL4.1 to SL4.9 are content common to both Mathematics: analysis and approaches and Mathematics: applications and interpretation.

SL 4.1

Content	Guidance, clarification and syllabus links
Concepts of population, sample, random sample, discrete and continuous data.	This is designed to cover the key questions that students should ask when they see a data set/ analysis.
Reliability of data sources and bias in sampling.	Dealing with missing data, errors in the recording of data.
Interpretation of outliers.	Outlier is defined as a data item which is more than 1.5 × interquartile range (IQR) from the nearest quartile. Awareness that, in context, some outliers are a valid part of the sample but some outlying data items may be an error in the sample. **Link to**: box and whisker diagrams (SL4.2) and measures of dispersion (SL4.3).
Sampling techniques and their effectiveness.	Simple random, convenience, systematic, quota and stratified sampling methods.

Connections

Links to other subjects: Descriptive statistics and random samples (biology, psychology, sports exercise and health science, environmental systems and societies, geography, economics; business management); research methodologies (psychology).

Aim 8: Misleading statistics; examples of problems caused by absence of representative samples, for example Google flu predictor, US presidential elections in 1936, Literary Digest v George Gallup, Boston "pot-hole" app.

International-mindedness: The Kinsey report–famous sampling techniques.

TOK: Why have mathematics and statistics sometimes been treated as separate subjects? How easy is it to be misled by statistics? Is it ever justifiable to purposely use statistics to mislead others?

Download connections template

SL 4.2

Content	Guidance, clarification and syllabus links
Presentation of data (discrete and continuous): frequency distributions (tables).	Class intervals will be given as inequalities, without gaps.
Histograms. Cumulative frequency; cumulative frequency graphs; use to find median, quartiles, percentiles, range and interquartile range (IQR).	Frequency histograms with equal class intervals. **Not required:** Frequency density histograms.
Production and understanding of box and whisker diagrams.	Use of box and whisker diagrams to compare two distributions, using symmetry, median, interquartile range or range. Outliers should be indicated with a cross. Determining whether the data may be normally distributed by consideration of the symmetry of the box and whiskers.

Connections

Links to other subjects: Presentation of data (sciences, individuals and societies).

International-mindedness: Discussion of the different formulae for the same statistical measure (for example, variance).

TOK: What is the difference between information and data? Does "data" mean the same thing in different areas of knowledge?

Download connections template

SL 4.3

Content	Guidance, clarification and syllabus links
Measures of central tendency (mean, median and mode). Estimation of mean from grouped data.	Calculation of mean using formula and technology. Students should use mid-interval values to estimate the mean of grouped data.
Modal class.	For equal class intervals only.
Measures of dispersion (interquartile range, standard deviation and variance).	Calculation of standard deviation and variance of the sample using only technology, however hand calculations may enhance understanding. Variance is the square of the standard deviation.
Effect of constant changes on the original data.	**Examples:** If three is subtracted from the data items, then the mean is decreased by three, but the standard deviation is unchanged. If all the data items are doubled, the mean is doubled and the standard deviation is also doubled.
Quartiles of discrete data.	Using technology. Awareness that different methods for finding quartiles exist and therefore the values obtained using technology and by hand may differ.

Connections

Other contexts: Comparing variation and spread in populations, human or natural, for example agricultural crop data, social indicators, reliability and maintenance.

Links to other subjects: Descriptive statistics (sciences and individuals and societies); consumer price index (economics).

International-mindedness: The benefits of sharing and analysing data from different countries; discussion of the different formulae for variance.

TOK: Could mathematics make alternative, equally true, formulae? What does this tell us about mathematical truths? Does the use of statistics lead to an over-emphasis on attributes that can be easily measured over those that cannot?

Download connections template

SL 4.4

Content	Guidance, clarification and syllabus links
Linear correlation of bivariate data. Pearson's product-moment correlation coefficient, *r*.	Technology should be used to calculate *r*. However, hand calculations of *r* may enhance understanding. Critical values of *r* will be given where appropriate. Students should be aware that Pearson's product moment correlation coefficient (*r*) is only meaningful for linear relationships.
Scatter diagrams; lines of best fit, by eye, passing through the mean point.	Positive, zero, negative; strong, weak, no correlation. Students should be able to make the distinction between correlation and causation and know that correlation does not imply causation.
Equation of the regression line of y on x. Use of the equation of the regression line for prediction purposes. Interpret the meaning of the parameters, a and b, in a linear regression $y = ax + b$.	Technology should be used to find the equation. Students should be aware: • of the dangers of extrapolation • that they cannot always reliably make a prediction of x from a value of y, when using a y on x line.

Connections

Other contexts: Linear regressions where correlation exists between two variables. Exploring cause and dependence for categorical variables, for example, on what factors might political persuasion depend?

Links to other subjects: Curves of best fit, correlation and causation (sciences group subjects); scatter graphs (geography).

Aim 8: The correlation between smoking and lung cancer was "discovered" using mathematics. Science had to justify the cause.

TOK: Correlation and causation–can we have knowledge of cause and effect relationships given that we can only observe correlation? What factors affect the reliability and validity of mathematical models in describing real-life phenomena?

Download connections template

SL 4.5

Content	Guidance, clarification and syllabus links
Concepts of trial, outcome, equally likely outcomes, relative frequency, sample space (U) and event. The probability of an event A is $\mathrm{P}(A) = \frac{n(A)}{n(U)}$. The complementary events A and A' (not A).	Sample spaces can be represented in many ways, for example as a table or a list. Experiments using coins, dice, cards and so on, can enhance understanding of the distinction between experimental (relative frequency) and theoretical probability. Simulations may be used to enhance this topic.
Expected number of occurrences.	**Example:** If there are 128 students in a class and the probability of being absent is 0.1, the expected number of absent students is 12.8.

Connections

Other contexts: Actuarial studies and the link between probability of life spans and insurance premiums, government planning based on likely projected figures, Monte Carlo methods.

Links to other subjects: Theoretical genetics and Punnett squares (biology); the position of a particle (physics).

Aim 8: The ethics of gambling.

International-mindedness: The St Petersburg paradox; Chebyshev and Pavlovsky (Russian).

TOK: To what extent are theoretical and experimental probabilities linked? What is the role of emotion in our perception of risk, for example in business, medicine and travel safety?

Use of technology: Computer simulations may be useful to enhance this topic.

Download connections template

SL 4.6

Content	Guidance, clarification and syllabus links
Use of Venn diagrams, tree diagrams, sample space diagrams and tables of outcomes to calculate probabilities.	
Combined events: $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$. Mutually exclusive events: $\mathrm{P}(A \cap B) = 0$.	The non-exclusivity of "or".
Conditional probability: $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$.	An alternate form of this is: $\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A \mid B)$. Problems can be solved with the aid of a Venn diagram, tree diagram, sample space diagram or table of outcomes without explicit use of formulae. Probabilities with and without replacement.
Independent events: $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$.	

Connections

Aim 8: The gambling issue: use of probability in casinos. Could or should mathematics help increase incomes in gambling?

TOK: Can calculation of gambling probabilities be considered an ethical application of mathematics? Should mathematicians be held responsible for unethical applications of their work?

Download connections template

SL 4.7

Content	Guidance, clarification and syllabus links
Concept of discrete random variables and their probability distributions. Expected value (mean), for discrete data. Applications.	Probability distributions will be given in the following ways: X 1 2 3 4 5 $P(X = x)$ 0.1 0.2 0.15 0.05 0.5 $P(X = x) = \frac{1}{18}(4 + x)$ for $x \in \{1, 2, 3\}$ $E(X) = 0$ indicates a fair game where X represents the gain of a player.

Connections

Other contexts: Games of chance.

Aim 8: Why has it been argued that theories based on the calculable probabilities found in casinos are pernicious when applied to everyday life (for example, economics)?

TOK: What do we mean by a "fair" game? Is it fair that casinos should make a profit?

Download connections template

SL 4.8

Content	Guidance, clarification and syllabus links
Binomial distribution. Mean and variance of the binomial distribution.	Situations where the binomial distribution is an appropriate model. In examinations, binomial probabilities should be found using available technology. **Not required:** Formal proof of mean and variance. **Link to:** expected number of occurrences (SL4.5).

Connections

Aim 8: Pascal's triangle, attributing the origin of a mathematical discovery to the wrong mathematician.

International-mindedness: The so-called "Pascal's triangle" was known to the Chinese mathematician Yang Hui much earlier than Pascal.

TOK: What criteria can we use to decide between different models?

Enrichment: Hypothesis testing using the binomial distribution.

Download connections template

SL 4.9

Content	Guidance, clarification and syllabus links
The normal distribution and curve. Properties of the normal distribution. Diagrammatic representation.	Awareness of the natural occurrence of the normal distribution. Students should be aware that approximately 68% of the data lies between $\mu \pm \sigma$, 95% lies between $\mu \pm 2\sigma$ and 99.7% of the data lies between $\mu \pm 3\sigma$.
Normal probability calculations.	Probabilities and values of the variable must be found using technology.
Inverse normal calculations	For inverse normal calculations mean and standard deviation will be given. This does not involve transformation to the standardized normal variable *z*.

Connections

Links to other subjects: Normally distributed real-life measurements and descriptive statistics (sciences group subjects, psychology, environmental systems and societies)

Aim 8: Why might the misuse of the normal distribution lead to dangerous inferences and conclusions?

International-mindedness: De Moivre's derivation of the normal distribution and Quetelet's use of it to describe *l'homme moyen*.

TOK: To what extent can we trust mathematical models such as the normal distribution? How can we know what to include, and what to exclude, in a model?

Download connections template

SL 4.10

Content	Guidance, clarification and syllabus links
Equation of the regression line of x on y.	
Use of the equation for prediction purposes.	Students should be aware that they cannot always reliably make a prediction of y from a value of x, when using an x on y line.

Connections

TOK: Is it possible to have knowledge of the future?

Download connections template

SL 4.11

Content	Guidance, clarification and syllabus links
Formal definition and use of the formulae: $P(A\|B) = \frac{P(A \cap B)}{P(B)}$ for conditional probabilities, and $P(A\|B) = P(A) = P(A\|B')$ for independent events.	An alternate form of this is: $P(A \cap B) = P(B)P(A\|B)$. Testing for independence.

Connections

Other contexts: Use of probability methods in medical studies to assess risk factors for certain diseases.

TOK: Given the interdisciplinary nature of many real-world applications of probability, is the division of knowledge into discrete disciplines or areas of knowledge artificial and/or useful?

Download connections template

SL 4.12

Content	Guidance, clarification and syllabus links
Standardization of normal variables (z- values).	Probabilities and values of the variable must be found using technology. The standardized value (z) gives the number of standard deviations from the mean.
Inverse normal calculations where mean and standard deviation are unknown.	Use of z-values to calculate unknown means and standard deviations.

Connections

Links to other subjects: The normal distribution (biology); descriptive statistics (psychology).

Download connections template

AHL content

Recommended teaching hours: 6

The aim of the AHL content in the statistics and probability topic is to extend and build upon the aims, concepts and skills from the SL content. Students are introduced to further conditional probability theory in the form of Bayes Theorem and properties of discrete and continuous random variables are further explored.