

Things in

[-] are the guidance from the specification

#-# are from the content exemplification for enhanced content guidance (not from specification from another document)

Calculator stats playlist -

<https://www.youtube.com/playlist?list=PLd7FwnU6nvjFY9Kb3ooSpYNrk8EUiV22P>

0.0 Large data set is about 5.3% of stats marks & 0.9% overall A-level marks according to Bicen Maths YT 2023 video

Please just watch this <https://www.youtube.com/watch?v=g2GiZkWWVPBs> of Bicen Maths going through the large data set it will help you so much and then this <https://www.youtube.com/watch?v=g2GiZkWWVPBs> which is TI Maths video on the LDS

0.1 Pearson has provided a large data set, which will support the assessment of Statistics in Paper 3: Statistics and Mechanics. Students are required to become familiar with the data set in advance of the final assessment. Assessments will be designed in such a way that questions assume knowledge and understanding of the data set. The expectation is that these questions should be likely to give a material advantage to students who have studied and are familiar with the data set. They might include questions/tasks that:

- assume familiarity with the terminology and contexts of the data, and do not explain them in a way that gives students who have not studied the data set the same opportunities to access marks as students who have studied them
- use summary statistics or selected data from, or statistical diagrams based on, the data set – these might be provided in the question or task, or as stimulus materials
- are based on samples related to the contexts in the data set, where students' work with the data set will help them understand the background context and/or
- require students to interpret data in ways that would be too demanding in an unfamiliar context.

Students will not be required to have copies of the data set in the examination, nor will they be required to have detailed knowledge of the actual data within the data set.

Measurement		Units	Typical range	Examples	Details
Daily mean windspeed	How windy it is	kn (knots)	~3 kn to ~10 kn	4 kn 11 kn	Integers only
Windspeed, Beaufort conversion		Qualitative	Light to Moderate	Light Moderate Fresh Strong	Qualitative. Most days are 'light'
Daily maximum gust	The strongest gust of wind that day	kn (knots)	~8 kn to ~50 kn	17 kn 25 kn	Integers only
Wind/gust direction (bearings)	Which direction the wind is blowing from	°	10° to 360°	240° 70°	Multiples of 10 only
Wind/gust direction (cardinal)	Which direction the wind is blowing from	Compass direction	—	N SW ENE	Describes where the wind is blowing from

UK Measurements pt. 1					
Measurement		Units	Typical range	Examples	
Daily mean temperature	How hot it is	°C	~5°C to ~24°C	12.0°C 14.8°C	
Daily total rainfall	How much it rained	mm	0 to ~20mm	0 mm 10.4 mm tr	many 0 and tr values
Daily total sunshine	How many hours of sunshine	hours	0 to ~14 hours	3.3 hrs 10.4 hrs	More sunshine in the summer
Cloud cover	How much of the sky is covered in clouds	oktas	0 to 8	3 4 0	Integers, measuring what fraction of the sky is covered
Humidity	How much water vapour is in the air - above 95% is associated with fog	%	~70% to 100%	100% 77% 95%	Integers
Daily mean visibility	How far you can see	Dm 1Dm = 10m 'decametre'	~200 Dm to ~4000 Dm	1300 Dm 2200 Dm 3100 Dm	Rounded to nearest 100
Daily mean pressure	How much the atmosphere is pushing down	hPa 'hectopascals'	~990 hPa to ~1040 hPa	1017 hPa 1006 hPa 997 hPa	Integers

International Measurements			
Measurement		Units	Details
Daily mean temperature	How hot it is	°C	Warmer in summer... but Perth is colder
Daily total rainfall	How much it rained	mm	Beijing is rainy in the summer
Daily mean pressure	How much the atmosphere is pushing down	hPa 'hectopascals'	
Daily mean windspeed	How windy it is	kn (knots)	Now are rounded to 1 dp
Windspeed, Beaufort conversion		Light to Moderate	Qualitative. Most days are 'light'

What are the locations of the cities in the large data set in the UK from south to north? And how to remember the order.

- Camborne
- Hurn
- Heathrow
- Leeming
- Leuchars

Its in alphabetical order except from Hurn and Heathrow

What are the locations of the cities in the large data set outside of the UK from south to north?

- Perth
- Beijing
- Jacksonville

Think about Peanut butter and jelly

Which months are covered in the large data set for each city?

From May to October

AS Level

1.0 Statistical sampling

1.1 Understand and use the terms 'population' and 'sample'. Use samples to make informal inferences about the population.

Understand and use sampling techniques, including simple random sampling and opportunity sampling. Select or critique sampling techniques in the context of solving a statistical problem, including understanding that different samples can lead to different conclusions about the population.

[Students will be expected to comment on the advantages and disadvantages associated with a census and a sample. Students will be expected to be familiar with: simple random sampling, stratified sampling, systematic sampling, quota sampling and opportunity (or convenience) sampling.]

#Students will not be expected to be familiar with any other sampling techniques. In systematic sampling where the population is not divisible by the sample size (i.e. N/n is non-integer), the key points are to take a random starting point and a systematic selection thereafter. For example, a sample of size 50 from a population of size 498 could be taken by first selecting at random from amongst those members numbered 1 to 8 inclusive and then selecting every 10th member after that. Students should be aware that different samples could lead to different conclusions. For example, if two people each take a sample to test a hypothesis about a population, one sample may lead to a significant result whereas the other sample may not. #

What is a sampling frame	A list of all members of the population
--------------------------	---

	<i>For example: a list of employees' names within a company</i>
Define population?	The whole set of items that are of interest
Define census	A measure or observation of every member in a population
What is the advantage of using a census?	It should give a completely accurate result
What is a disadvantage of using a census?	<ul style="list-style-type: none"> ● Time consuming and expensive ● Cannot be used when the testing process destroys the item ● Hard to process large quantities of data
Define sample?	A selection of observations taken from a subset of the population which is used to find out information about the population as a whole
What are the advantages of using a sample	<ul style="list-style-type: none"> ● Less time consuming and cheaper ● Fewer people have to respond ● Less data needs to be processed
What are the disadvantages of using a sample?	<ul style="list-style-type: none"> ● Data may not be as accurate ● Sample may not be large enough to give information about small subgroups of the population
Describe simple random sampling?	<ol style="list-style-type: none"> 1. Allocate numbers to objects 2. Use a random number generator 3. Ignore repeats
What are the advantages of simple random sampling?	<ul style="list-style-type: none"> ● Free of bias ● Easy and cheap to implement for small populations and small samples ● Each sampling unit has a known and equal chance of selection
What are the disadvantages of simple random sampling?	<ul style="list-style-type: none"> ● Not suitable when the population size or the sample size is large ● A sampling frame is needed
Describe stratified sampling?	Having the groups in your sample proportional to the groups in the population
What are the advantages of stratified sampling	<ul style="list-style-type: none"> ● Sample accurately reflects the population structure ● Guarantees proportional representation of groups within a population

What are the disadvantages of stratified sampling	<ul style="list-style-type: none"> ● Population must be clearly classified into distinct strata ● Selection within each stratum suffers from the same disadvantages as simple random sampling (Not suitable when the population size or the sample size is large. A sampling frame is needed)
Describe systematic sampling	Follow a system (e.g., selecting every n^{th} person along)
What are the advantages of systematic sampling	<ul style="list-style-type: none"> ● Simple and quick to use ● Suitable for larger samples and large populations
What are the disadvantages of systematic sampling	<ul style="list-style-type: none"> ● A sampling frame is needed ● It can introduce bias if the sampling frame is not random
Describe opportunity sampling (convenience sampling)	Asking people you have access to until you have a sample of desired size
What are the advantages of opportunity sampling (convenience sampling)?	<ul style="list-style-type: none"> ● Easy and Inexpensive
What are the disadvantages of opportunity sampling (convenience sampling)?	<ul style="list-style-type: none"> ● Unlikely to provide a representative sample ● Highly dependent on individual researcher
Describe quota sampling	Using opportunity sampling yet taking into account how many people of each group you want in your sample
What are the advantages of quota sampling	<ul style="list-style-type: none"> ● Allows a small sample to still be representative of the population ● No sampling frame required ● Quick, easy and inexpensive ● Allow for easy comparison between different groups within a population
What are the disadvantages of quota sampling	<ul style="list-style-type: none"> ● Non-random sampling can introduce bias ● Population must be divided into groups, which can be costly or inaccurate
Define quantitative?	Numerical data
Define qualitative?	Non numerical data

Definition of Data?	A collection of observations
Definition of Observation?	Pieces of collected information
Definition of Variable?	A factor which is possible to change
What is the equation for the number sampled in a stratum	Number sampled in a stratum = (number in stratum/number in population) x overall sample size

2.0 Data presentation and interpretation

2.1 Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency. Connect to probability distributions.

[Students should be familiar with histograms, frequency polygons, box and whisker plots (including outliers) and cumulative frequency diagrams.]

#Students will not be expected to be familiar with any other diagrams for single-variable data. If students are required to draw a cumulative frequency diagram, either a frequency polygon or a curve is acceptable. #

What type of data does histograms only work for?	Continuous grouped data
How can a histogram turn into a frequency polygon?	By joining the middle of the top (Mp) of each bar in the histogram with straight lines
What is the equation for the Area of a bar for histograms (in terms of Class width and FD)	Area of bar (frequency) = CW X FD
What is the equation for the scale factor of Area of a bar and frequency) for histograms	$K = \frac{\text{Area of bar}}{\text{frequency}}$

	<p>Histograms</p> <p>Group continuous data can be presented using histograms. Histograms show the rough location and general shape of the data, and how spread out the data is.</p> <p>The area of the bar is proportional to the frequency of each class.</p> <p>To calculate the height of each bar (frequency density):</p> <p style="text-align: center;">$\text{Area of bar} = k \times \text{frequency}$</p>
<p>What is the equation for FD using class width and frequency for histograms</p>	<p>Frequency density = $\frac{\text{frequency}}{\text{class width}}$</p> <p>Histograms</p> <p>Group continuous data can be presented using histograms. Histograms show the rough location and general shape of the data, and how spread out the data is.</p> <p>The area of the bar is proportional to the frequency of each class.</p> <p>To calculate the height of each bar (frequency density):</p> <p style="text-align: center;">$\text{Area of bar} = k \times \text{frequency}$</p> <p>When $k = 1$,</p> <p style="text-align: center;">$\text{Frequency density} = \frac{\text{frequency}}{\text{class width}}$</p>
<p>What is the equation to find the Scale factor of a width of a bar for a histogram?</p>	<p>$k = \text{Width of a bar} / \text{Class width}$</p> <p>Width of bar is most likely given in a question unless k (SF) is given</p>
<p>What is the equation for the height of a bar for a histogram?</p>	<p>$h = \text{frequency} / (k \times \text{width of bar})$</p> <p>Where k = the scale factor of Area of a bar to frequency</p>
<p>How do you find Q1 from a CF graph? (After finding the Upper bound of each class whether its discrete/continuous and plotting the graph)</p>	<ul style="list-style-type: none"> ● Read from the y axis (CF) $\frac{n}{4}$th value or (sum of the CF/4)th value ● To the x axis (classes upper bound)
<p>How do you find Q2 from a CF graph? (After finding the Upper bound of each class whether its discrete/continuous and plotting the graph)</p>	<ul style="list-style-type: none"> ● Read from the y axis (CF) $\frac{n}{2}$th value or (sum of the CF/2)th value ● To the x axis (classes upper bound)
<p>How do you find Q3 from a CF graph (After finding the Upper bound of each class whether its discrete/continuous and plotting the graph)</p>	<ul style="list-style-type: none"> ● Read from the y axis (CF) $\frac{3n}{4}$th value or (3 x sum of the CF/4)th value ● To the x axis (classes upper bound)

2.2 Interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams that include distinct sections of the population (calculations involving regression lines are excluded). Understand informal interpretation of correlation. Understand that correlation does not imply causation.

[Students should be familiar with the terms explanatory (independent) and response (dependent) variables. Use of interpolation and the dangers of extrapolation. Variables other than x and y may be used. Use of terms such as positive, negative, zero, strong and weak are expected.]

#Where students are required to interpret a graph or a result, they are expected to comment on values within the context of the question. Where students are asked to interpret a correlation, they are expected to give an answer using the context of the question. Where students are asked to describe a correlation, they are not expected to give an answer using the context of the question.#

How should you describe correlation on a scatter graph?	<ul style="list-style-type: none"> ● Positive, negative, or zero correlation <p><i>Don't say strong, moderate or weak for positive or negative correlation as the graph may be ambiguous and you would lose the mark - Ms Malkin</i></p>
What are the 2 reasons about why values that are extrapolated using a regression line are inaccurate?	<ul style="list-style-type: none"> ● The result is unreliable as it has been extrapolated n (units) outside the data range ● We cannot assume the same trend continues indefinitely
What does the gradient of a line of regression mean? (i.e. the script for regression questions)	<ul style="list-style-type: none"> ● For every one (unit) increase in (context)x, the (context)y increases/decreases by b(units) <p><i>x = variable, y = variable,</i></p>
Why can the equation $y = a + bx$ be used to find	<ul style="list-style-type: none"> ● x might be inaccurate

estimates for y but should be treated with caution for making estimates for x? And what should be done

- You should calculate x using the regression line x on y ($x = a + by$)

Interpret what the describe the correlation question means for the scatter graph of population density (people hectares) on y axis and Distance from centre (km) on x axis

Example 1: In the study of a city, the population density, in people/hectare, and the distance from the city centre, in km, was investigated by picking a number of sample areas with the following results.

Area	A	B	C	D	E	F	G	H	I	J
Distance (km)	0.6	3.8	2.4	3.0	2.0	1.5	1.8	3.4	4.0	0.9
Population density (people/hectare)	50	22	14	20	33	47	25	8	16	38

a. Draw a scatter diagram to represent this data.

Population density (people/hectare)

Distance from centre (km)

Remember to label your axis and include units

As distance from the centre increases the population density decreases

(As x variable increases the y variable increases/decreases))

b. Describe the correlation between distance and population density.

There is a weak negative correlation.

Describe the strength of correlation and whether it is positive or negative

c. Interpret your answer to part b.

As distance from the centre increases, the population density decreases.

Interpret results in context to the question

2.3 Interpret measures of central tendency and variation, extending to standard deviation. Be able to calculate standard deviation, including from summary statistics.

[Data may be discrete, continuous, grouped or ungrouped. Understanding and use of coding. Measures of central tendency: mean, median, mode. Measures of variation: variance, standard deviation, range and interpercentile ranges. Use of linear interpolation to calculate percentiles from grouped data is expected. Students should be able to use the statistic x

$$S_{xx} = \sum (x - \bar{x})^2 = \sum x^2 - \frac{(\sum x)^2}{n}$$
 Use of standard deviation = $\sqrt{\frac{S_{xx}}{n}}$ (or equivalent) is expected but the use of $S = \sqrt{\frac{S_{xx}}{n-1}}$ (as used on spreadsheets) will be accepted.]

#For data in a list or a (non-grouped) frequency table, use the following to find the position: Q1: $n/4$ and • round up if a decimal • + 0.5 if an integer Median: $n/2$ and • round up if a decimal • + 0.5 if an integer Q3: $3n/4$ and • round up if decimal • + 0.5 if integer This is considered the “main” method. However, other sensible approaches are accepted and, because this is a contentious area, data sets are typically chosen so that all sensible methods give the same answer. For data in a grouped frequency table, a histogram or cumulative frequency graph: Q1: $n/4$ th position Median: $n/2$ th position Q3: $3n/4$ th position (Using linear interpolation if necessary.) Grouped data is always regarded as continuous no matter what the nature of the underlying variable may be, so we would usually ask students to "use interpolation to estimate..." in these circumstances. Similarly, a list of individual data values or data in a (non-grouped) frequency table is always regarded as discrete. The alternative approach using $(n+1)/4$ etc. will also be allowed.

<p>Definition of continuous data?</p> <div><p>Continuous vs Discrete Data</p><table><tr><td>Any Value</td><td></td><td>Specific Values</td></tr><tr><td>"Measured"</td><td></td><td>"Counted"</td></tr><tr><td>5.6, 2.489</td><td>vs</td><td>1, 2, 3, 4, 5, 6</td></tr><tr><td>Temperature</td><td></td><td># of cats</td></tr></table></div>	Any Value		Specific Values	"Measured"		"Counted"	5.6, 2.489	vs	1, 2, 3, 4, 5, 6	Temperature		# of cats	<ul style="list-style-type: none">● Data which can take any numerical value● (Measurements that are rounded) <p><i>E.g, Temperature 5.5, 10.8, 1.0 , and data in which a lower/upper bound needs to be found</i></p>
Any Value		Specific Values											
"Measured"		"Counted"											
5.6, 2.489	vs	1, 2, 3, 4, 5, 6											
Temperature		# of cats											
<p>Definition of discrete data?</p> <div><p>Continuous vs Discrete Data</p><table><tr><td>Any Value</td><td></td><td>Specific Values</td></tr><tr><td>"Measured"</td><td></td><td>"Counted"</td></tr><tr><td>5.6, 2.489</td><td>vs</td><td>1, 2, 3, 4, 5, 6</td></tr><tr><td>Temperature</td><td></td><td># of cats</td></tr></table></div>	Any Value		Specific Values	"Measured"		"Counted"	5.6, 2.489	vs	1, 2, 3, 4, 5, 6	Temperature		# of cats	<ul style="list-style-type: none">● Data which can only take certain numerical values● (Exact values or whole numbers that are not rounded) <p><i>E.g. number of cats 1, 2, 3, 4, 5, 6</i></p>
Any Value		Specific Values											
"Measured"		"Counted"											
5.6, 2.489	vs	1, 2, 3, 4, 5, 6											
Temperature		# of cats											

What does σ^2 stand for?	Variance
What is the equation for variance (σ^2) for discrete (ungrouped) data?	<ul style="list-style-type: none"> ● $\frac{\Sigma x^2}{n} - (\bar{x})^2 = \sigma^2$ ● n = the number of terms ● Σx^2 = the sum of the frequencies squared ● \bar{x} = the mean of the frequencies <p>(Another variation is $(\frac{\Sigma (x-\bar{x})^2}{n}) = \sigma^2$)</p> <p>(Im assuming you know $\bar{x} = \frac{\Sigma x}{n}$ Σx = the sum of the frequencies)</p>
What is the equation for variance (σ^2) for continuous (grouped) data?	<ul style="list-style-type: none"> ● $A \frac{\Sigma f x^2}{\Sigma f} - (\frac{\Sigma f x}{\Sigma f})^2 = \sigma^2$ ● $\Sigma f x$ = the sum of the frequencies x the mid point (should be calculated from the fx column) ● $\Sigma f x^2$ = the sum of the frequencies x the midpoint squared+ (should be a separate column for fx^2)
What does σ stand for?	Standard deviance
What is the equation for standard deviance (σ) for discrete (ungrouped) data?	<ul style="list-style-type: none"> ● $\sqrt{\frac{\Sigma x^2}{n} - (\bar{x})^2} = \sigma$ ● n = the number of terms ● Σx^2 = the sum of the frequencies squared ● \bar{x} = the mean of the frequencies <p>(Another variation is $\sqrt{(\frac{\Sigma (x-\bar{x})^2}{n})} = \sigma$, Im assuming you know $\bar{x} = \frac{\Sigma x}{n}$)</p>
What is the equation for standard deviation (σ) for discrete/continuous (grouped) data	<ul style="list-style-type: none"> ● $\sqrt{\frac{\Sigma f x^2}{\Sigma f} - (\frac{\Sigma f x}{\Sigma f})^2} = \sigma$ ● $\Sigma f x$ = the sum of the frequencies x the midpoint (should be calculated from the fx column) ● $\Sigma f x^2$ = the sum of the frequencies x the midpoint squared (should be a separate column for fx^2) <p>($\bar{x} = \frac{\Sigma f x}{\Sigma f}$ \bar{x} = the mean)</p>
How would you calculate the lower quartile (Q_1) for discrete (ungrouped) data?	<ul style="list-style-type: none"> ● $Q_1 = \frac{n}{4}^{th}$ data value
How would you estimate the median for	<ul style="list-style-type: none"> ● $(n+1)/2$

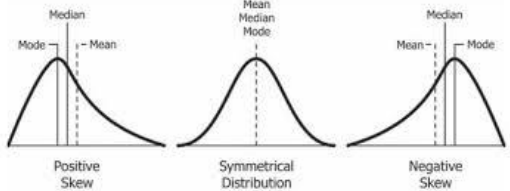
discrete (ungrouped) data?	<ul style="list-style-type: none"> ● n = number of observations ● (Do not round answers)
How would you calculate the upper quartile (Q_3) for discrete (ungrouped) data?	<ul style="list-style-type: none"> ● $Q_3 = \frac{3n}{4}th$ data value
What is the LB,UB and CW for discrete (grouped) data of the class $175 \leq x < 225$	<ul style="list-style-type: none"> ● LB: 175 ● UB: 225 ● CW:50
What is the LB,UB and CW for continuous (grouped) data of the class 175 - 225	<ul style="list-style-type: none"> ● LB: 174.5 ● UB: 225.5 ● CW: 51
How would you estimate the Median/ Q_2 for discrete/continuous (grouped) data?	<ul style="list-style-type: none"> ● $n/2=Q_2$ then doing linear interpolation to estimate (Using CF) ● $\frac{Q_2 - LBofCB}{UB - LBofCB} = \frac{UBofCF - n/2}{UB - LBofCF}$ ● CB = Class boundary <p><i>Always helps to draw a number line - Ms Malkin</i></p>
What is the equation for the mean for discrete (ungrouped) data	<p>The mean can be calculated using:</p> $\bar{x} = \frac{\Sigma x}{n}$ <p>Where \bar{x} (x bar) is the mean, Σx is the sum of the data values, n is the number of data values</p> <p>For data given in a cumulative frequency table, the mean can be calculated using:</p> $\bar{x} = \frac{\Sigma xf}{\Sigma f}$ <p>Where Σfx is the sum of the products of the data values and their frequencies, Σf is the sum of frequencies</p> <ul style="list-style-type: none"> ● $\underline{x} = \frac{\Sigma x}{n}$
What is the equation for the mean for discrete/continuous (grouped) data?	<p>The mean can be calculated using:</p> $\bar{x} = \frac{\Sigma x}{n}$ <p>Where \bar{x} (x bar) is the mean, Σx is the sum of the data values, n is the number of data values</p> <p>For data given in a cumulative frequency table, the mean can be calculated using:</p> $\bar{x} = \frac{\Sigma xf}{\Sigma f}$ <p>Where Σfx is the sum of the products of the data values and their frequencies, Σf is the sum of frequencies</p> <ul style="list-style-type: none"> ● $\underline{x} = \frac{\Sigma fx}{\Sigma f}$ ● Σfx = the sum of the frequencies x the midpoint (should be calculated from the fx column)
What is the equation for coding for the mean	<ul style="list-style-type: none"> ● $\underline{y} = a + b\underline{x}$ ● The mean is effected by addition and multiplication vice versa
What is the equation for coding for standard deviation	<ul style="list-style-type: none"> ● $\sigma_y = b\sigma_x$ ● Standard deviation is only affected by mutiplication







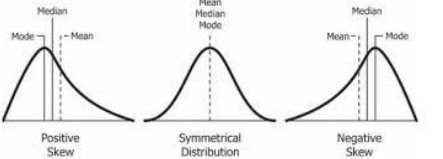
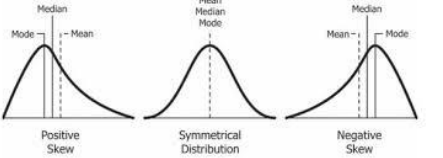
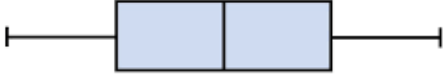


<p>State the assumption involved with using class midpoints to calculate an estimate of a mean from a grouped frequency table?</p> <p>(d) (i) State the assumption involved with using class midpoints to calculate an estimate of a mean from a grouped frequency table.</p>	<p>● Using class midpoints to estimate the mean assumes that the values are evenly distributed within the classes</p> <table border="1"><tr><td>(d)(i)</td><td>Using class midpoints to estimate the mean assumes that the values are uniformly distributed within the class(es).</td><td>B1</td><td>2.4</td></tr></table>	(d)(i)	Using class midpoints to estimate the mean assumes that the values are uniformly distributed within the class(es) .	B1	2.4
(d)(i)	Using class midpoints to estimate the mean assumes that the values are uniformly distributed within the class(es) .	B1	2.4		
<p>What is the range?</p>	<p>The difference between the largest and smallest values in a data set</p> <p><i>All data points in the set will be included in the range including extreme values LIKE OUTLIERS</i></p>				
<p>State giving a reason whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data?</p> <p>Taruni decided to ask every member of the company the time, x minutes, it takes them to travel to the office.</p> <p>(c) State the data selection process Taruni used.</p> <p>Taruni's results are summarised by the box plot and summary statistics below.</p> <div data-bbox="261 952 671 1064"><p style="text-align: center;">Journey time (minutes)</p></div> <div data-bbox="359 1075 576 1099">$n = 95 \quad \sum x = 4133 \quad \sum x^2 = 202294$</div> <p>(f) State, giving a reason, whether you would recommend using the mean and standard deviation or the median and interquartile range to describe these data.</p>	<p>● There are outliers in the data (or data is skew) which will affect mean and sd -> use median and IQR</p> <table border="1"><tr><td>(f)</td><td>There are outliers in the data (or data is skew) which will affect mean and sd Therefore use median and IQR</td><td>(2)</td><td>2.4</td></tr></table> <p>1st B1 for acknowledging outliers or skewness are a problem for mean and sd "extreme values"/"anomalies" OK. May be implied by saying median and IQR not affected by. We need to see mention of "outliers", "skewness" and the problem so "data is skewed so use median and IQR" is B0 unless mention that they are not affected by extreme values or mean and standard deviation can be "inflated" by the positive skew etc</p> <p>2nd dB1 dep on 1st B1 for therefore choosing median and IQR</p>	(f)	There are outliers in the data (or data is skew) which will affect mean and sd Therefore use median and IQR	(2)	2.4
(f)	There are outliers in the data (or data is skew) which will affect mean and sd Therefore use median and IQR	(2)	2.4		

2.4 Recognise and interpret possible outliers in data sets and statistical diagrams. Select or critique data presentation techniques in the context of a statistical problem. Be able to clean data, including dealing with missing data, errors and outliers

[Any rule needed to identify outliers will be specified in the question. For example, use of $Q_1 - 1.5 \times IQR$ and $Q_3 + 1.5 \times IQR$ or $\text{mean} \pm 3 \times \text{standard deviation}$. Students will be expected to draw simple inferences and give interpretations to measures of central tendency and variation. For example, students may be asked to identify possible outliers on a box plot or scatter diagram.]

#Students may be asked to consider possible reasons for outliers or missing data. They should also be aware of the possible implications of omitting or including outliers from a data set.

How are outliers commonly found?	<ul style="list-style-type: none"> ● Outside the interval ($Q_1 - 1.5 \times IQR$) and ($Q_3 + 1.5 \times IQR$). ● or ± 3 standard deviation <p><i>There may be other rules which you are told to apply</i></p>
What does a positive skew box plot & frequency diagram look like	 <p>The diagrams show three frequency curves:</p> <ul style="list-style-type: none"> Positive Skew: The curve is skewed to the right. The Mode is at the peak on the left, followed by the Median, and then the Mean is furthest to the right. Symmetrical Distribution: The curve is a symmetric bell shape. The Mean, Median, and Mode are all at the same central point. Negative Skew: The curve is skewed to the left. The Mean is furthest to the left, followed by the Median, and then the Mode is at the peak on the right.

	<p>Normal Distribution</p>  <p>Positive Skew</p>  <p>Negative Skew</p> 
What does a symmetrical skew box plot & frequency diagram look like	<p>Normal Distribution</p>  <p>Positive Skew</p>  <p>Negative Skew</p>  <p>Frequency Diagrams:</p> 
What does a negative skew box plot & frequency diagram look like	<p>Frequency Diagrams:</p>  <p>Box Plots:</p> <p>Normal Distribution</p>  <p>Positive Skew</p>  <p>Negative Skew</p> 

4. Sara was studying the relationship between rainfall, r mm, and humidity, $h\%$, in the UK. She takes a random sample of 11 days from May 1987 for Leuchars from the large data set.

She obtained the following results.

h	93	86	95	97	86	94	97	97	87	97	86
r	1.1	0.3	3.7	20.6	0	0	2.4	1.1	0.1	0.9	0.1

Sara examined the rainfall figures and found

$$Q_1 = 0.1 \quad Q_2 = 0.9 \quad Q_3 = 2.4$$

A value that is more than 1.5 times the interquartile range (IQR) above Q_3 is called an outlier.

(a) Show that $r = 20.6$ is an outlier.

(1)

(b) Give a reason why Sara might:

(i) include

(ii) exclude

this day's reading.

(2)

Give a reason why Sara might include AND exclude this day's reading

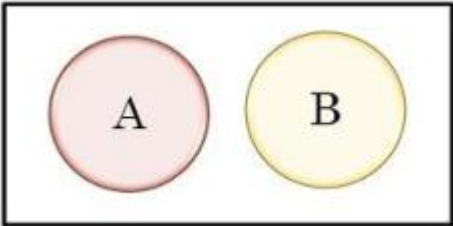
- Include outlier - It is a piece of data and we should consider all the data
- Exclude outlier - It is an extreme value and could influence the analysis/ It could be a mistake


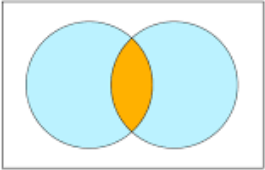
(b)(i)	e.g. It is a piece of data and we should consider all the data o.e.	B1	2.4
(ii)	e.g. It is an extreme value and could unduly influence the analysis or It could be a mistake	B1	2.4
		(2)	

3.0 Probability

3.1 Understand and use mutually exclusive and independent events when calculating probabilities. Link to discrete and continuous distributions.

[Venn diagrams or tree diagrams may be used. No formal knowledge of probability density functions is required but students should understand that area under the curve represents probability in the case of a continuous distribution.]

Give an example of a dependent event	Picking sweets out of a bag without replacement
What are mutually exclusive events? (With example and formula)	<p>Events that cannot occur at the same time. E.g.,</p> <div style="text-align: center;">  </div> <p>Or flipping a coin TWICE, you cannot have HH and TT at the same time meaning these events are also mutually exclusive. This can be given by:</p>

	$P(A \cap B) = 0$ $P(A \cap B) = 0$
What is the addition rule for the probability of event A OR event B?	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ For mutually exclusive, $P(A) + P(B)$
What is the addition rule for mutually exclusive events	<ul style="list-style-type: none"> ● $P(A \cup B) = P(A) + P(B)$ ● Probability of A OR B as the OR is shaped like the Hull of a wooden boat [U]. $P(A \text{ or } B) = P(A) + P(B)$ <ul style="list-style-type: none"> ●
What is the multiplication rule and what is it used for?	<ul style="list-style-type: none"> ● $P(A \cap B) = P(A) \times P(B)$ $P(A \text{ and } B) = P(A) \times P(B)$ <ul style="list-style-type: none"> ● ● To prove that two events are independent <div> $P(A \cap B)$ Formula  </div> $P(A \cap B) = P(A) \times P(B)$ 

4.0 Statistical distributions

4.1 Understand and use simple, discrete probability distributions (calculation of mean and variance of discrete random variables is excluded), including the binomial distribution, as a model; calculate probabilities using the binomial distribution.

[Students will be expected to use distributions to model a real-world situation and to comment critically on the appropriateness. Students should know and be able to identify the discrete uniform distribution. The notation $X \sim B(n, p)$ may be used. Use of a calculator to find individual or cumulative binomial probabilities.]

#Students are expected to have an understanding of the assumptions required for a binomial distribution to be applicable and be able to comment on these within the context of the question. #

<p>What 4 conditions/assumptions that must be satisfied for a Binomial Distribution to be valid? With e.g.</p> <p>In an experiment a group of children each repeatedly throw a dart at a target. For each child, the random variable H represents the number of times the dart hits the target in the first 10 throws.</p> <p>Peta models H as $B(10, 0.1)$</p> <p>(a) State two assumptions Peta needs to make to use her model.</p> <p>(2)</p>	<ol style="list-style-type: none">Two possible outcomes in each trialA fixed number of trialsIndependent trialsIdentical trials (same probability) <ul style="list-style-type: none">The <u>probability</u> of a dart hitting the target is <u>constant</u> (from child to child and for each throw by each child)The <u>throws</u> of each of the darts are <u>independent</u> <table><tr><td>(a)</td><td>1st B1 for stating that the <u>probability</u> (or possibility or chance) is <u>constant</u> (or fixed or same)</td><td></td></tr><tr><td></td><td>2nd B1 for stating that <u>throws</u> are <u>independent</u> ["trials" are independent is B0]</td><td></td></tr></table> <table><tr><th>Qu</th><th>Scheme</th><th>Marks</th><th>AO</th></tr><tr><td>(a)</td><td>The <u>probability</u> of a dart hitting the target is <u>constant</u> (from child to child and for each throw by each child) (o.e.)</td><td>B1</td><td>1.2</td></tr><tr><td></td><td>The <u>throws</u> of each of the darts are <u>independent</u> (o.e.)</td><td>B1</td><td>1.2</td></tr><tr><td></td><td></td><td>(2)</td><td></td></tr></table>	(a)	1 st B1 for stating that the <u>probability</u> (or possibility or chance) is <u>constant</u> (or fixed or same)			2 nd B1 for stating that <u>throws</u> are <u>independent</u> ["trials" are independent is B0]		Qu	Scheme	Marks	AO	(a)	The <u>probability</u> of a dart hitting the target is <u>constant</u> (from child to child and for each throw by each child) (o.e.)	B1	1.2		The <u>throws</u> of each of the darts are <u>independent</u> (o.e.)	B1	1.2			(2)	
(a)	1 st B1 for stating that the <u>probability</u> (or possibility or chance) is <u>constant</u> (or fixed or same)																						
	2 nd B1 for stating that <u>throws</u> are <u>independent</u> ["trials" are independent is B0]																						
Qu	Scheme	Marks	AO																				
(a)	The <u>probability</u> of a dart hitting the target is <u>constant</u> (from child to child and for each throw by each child) (o.e.)	B1	1.2																				
	The <u>throws</u> of each of the darts are <u>independent</u> (o.e.)	B1	1.2																				
		(2)																					
<p>Draw the notation for a binomial distribution?</p>	<p>$X \sim B(n, p)$</p> <ul style="list-style-type: none">n = number of trials																						

	<ul style="list-style-type: none"> ● p = probability of single success ● \sim = "is distributed"
How would you calculate $P(X \geq 6)$?	<ul style="list-style-type: none"> ● $1 - P(X \leq 5)$ ● Use calc CD (with knowing the probability and number in sample) ● To 4 dp (according to Mrs Cope its better I think to) can use 3dp tho

5.0 Statistical hypothesis testing

5.1 Understand and apply the language of statistical hypothesis testing, developed through a binomial model: null hypothesis, alternative hypothesis, significance level, test statistic, 1-tail test, 2-tail test, critical value, critical region, acceptance region, p-value.

[An informal appreciation that the expected value of a binomial distribution is given by np may be required for a 2-tail test.]

#There are two possible approaches to finding critical regions: selecting a region where the probability will be less than the significance level (or half of this for a 2-tail test) or selecting a region where the probability is as close as possible to the required significance level. Unless both approaches lead to the same answer, it will be made clear in the question which approach students should use. Questions will be worded to give suitable indication as to whether a 1-tail or 2-tail test is required. For example, using phrases such as: • evidence of an increase in the proportion (1-tail) • evidence of a change in the proportion (2-tail) It cannot be assumed that all students have a calculator with the facility to find critical regions and so tables are provided for this purpose. However, where they do have a calculator with this function, students may use it to find the critical region. The p-value for a 1-tail test is the probability in the tail of interest. For a 2-tail test that probability is simply doubled. For example, when calculating the appropriate probability for a 2-tailed test using a 5% significance level, you would compare this probability to 2.5% and if you double the probability you get the p-value which should be compared with 5%.The language used in the conclusion to hypothesis tests should be non-assertive (e.g. “There is insufficient evidence to reject H_0 ”) and contextual (e.g. “This supports the manager’s belief that...”).#

5.2 Conduct a statistical hypothesis test for the proportion in the binomial distribution and interpret the results in context.

Understand that a sample is being used to make an inference about the population and appreciate that the significance level is the probability of incorrectly rejecting the null hypothesis.

[Hypotheses should be expressed in terms of the population parameter p . A formal understanding of Type I errors is not expected.]

#Hypothesis tests can be answered using either a probability approach or by finding the critical region. If finding the critical region is required, it will be possible to use the table of values to do this. If finding the critical region has not been specified, then it may not be possible using the table of values, but the question can still be answered in this way by students who have access to these values using a calculator.#

A LEVEL

2.0 Data presentation and interpretation

2.2 [Use to make predictions within the range of values of the explanatory variable. Change of variable may be required, e.g. using knowledge of logarithms to reduce a relationship of the form $y = ax^n$ or $y = kb^x$ into linear form to estimate a and n or k and b .]

--	--

2.4 [Significance tests, other than those mentioned in Section 5, will not be expected.]

--	--

3.0 Probability

3.1 [Set notation to describe events may be used. Use of $P(B|A) = P(B)$, $P(A|B) = P(A)$, $P(A \cap B) = P(A)P(B)$ in connection with independent events.]

--	--

3.2 Understand and use conditional probability, including the use of tree diagrams, Venn diagrams, two-way tables.

Understand and use the conditional probability formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$

[Understanding and use of $P(A') = 1 - P(A)$, $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, $P(A \cap B) = P(A)P(B|A)$.]

--	--

3.3 Modelling with probability, including critiquing assumptions made and the likely effect of more realistic assumptions.

[For example, questioning the assumption that a die or coin is fair.]

--	--

4.0 Statistical distributions

4.2 Understand and use the Normal distribution as a model; find probabilities using the Normal distribution. Link to histograms, mean, standard deviation, points of inflection and the binomial distribution.

[The notation $X \sim N(\mu, \sigma^2)$ may be used. Knowledge of the shape and the symmetry of the distribution is required. Knowledge of the probability density function is not required. Derivation of the mean, variance and cumulative distribution function is not required

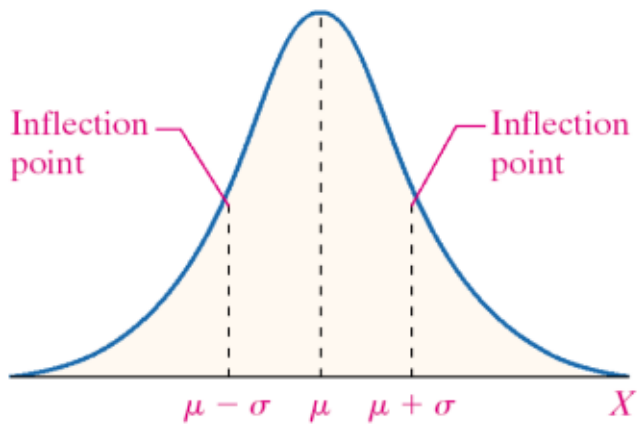
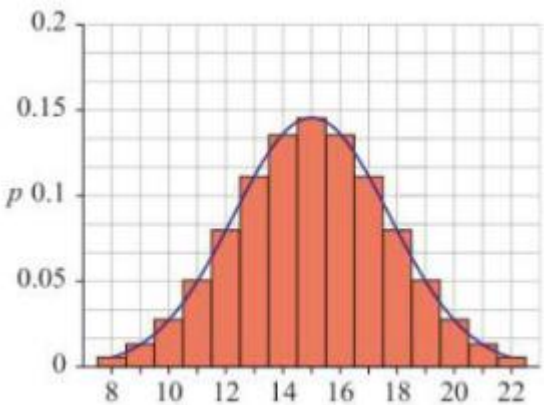
Questions may involve the solution of simultaneous equations. Students will be expected to use their calculator to find probabilities connected with the normal distribution.

Students should know that the points of inflection on the normal curve are at $x = \mu \pm \sigma$ The derivation of this result is not expected.

Students should know that when n is large and p is close to 0.5 the distribution $B(n, p)$ can be approximated by $N(np, np[1 - p])$ The application of a continuity correction is expected.]

#A formal understanding of skewness is not required, but students should be aware that if a distribution is not symmetric it may not be suitable to be modelled by a normal distribution. Students are expected to be able to standardise the Normal distribution. For example, if asked to find σ given that $X \sim N(25, \sigma^2)$ and $P(X \geq 34) = 0.35$, they would need to standardise and look up or work out the z-value corresponding to $P(Z \leq \frac{34 - 25}{\sigma}) = 0.65$ and solve to find σ . Students may use either tables or calculators to work out z-values, but they should work to at least the same degree of accuracy as can be achieved using tables. Inverse Normal values can be found using either a calculator or table of values. When asked for a condition for this approximation, students are expected to state “n is large and p is close to 0.5”. A response of “np > 5 and nq > 5” will only be given credit if it is accompanied by a statement that “n should be large”. Now that students have calculators that will calculate binomial probabilities, questions may also require students to find the error when using the approximation. #

What proportion lies within 1 and 3 σ 's of the mean in a Normal Distribution?	<ul style="list-style-type: none"> ● $\pm 1 \sigma$ is around 68% ● $\pm 3 \sigma$ is around 99.8%
What is the z-score? How is it calculated?	<p>A measure of how many standard deviations a value is to the right of the mean which is calculated by:</p> $z = \frac{x - \mu}{\sigma}$ <p><i>This is sometimes referred to as a test statistic in context</i></p>
What is the condition for approximations	N is large and p is close to 0.5

when find normal values using a calculator or a table of values?	Where N is the number of terms
Where are the points of inflection of a Normal Distribution?	
What is the continuity correction? How can it be used?	 <p>Say you need to work out $P(X < 8)$ on the Binomial Distribution, You can calculate $P(Y \leq 7.5)$ on the Normal Distribution</p>

4.3 Select an appropriate probability distribution for a context, with appropriate reasoning, including recognising when the binomial or Normal model may not be appropriate.

[Students should know under what conditions a binomial distribution or a Normal distribution might be a suitable model.]

--	--

5.0 Statistical hypothesis testing

5.1 Extend to correlation coefficients as measures of how close data points lie to a straight line. And be able to interpret a given correlation coefficient using a given p-value or critical value (calculation of correlation coefficients is excluded).

[Students should know that the product-moment correlation coefficient r satisfies $r \leq 1$ and that a value of $r = \pm 1$ means the data points all lie on a straight line. Students will be expected to calculate a value of r using their calculator but use of the formula is not required. Hypotheses should be stated in terms of ρ with a null hypothesis of $\rho = 0$ where ρ represents the population correlation coefficient. Tables of critical values or a p-value will be given.]

#Students are expected to be able to find summary statistics using a calculator, so given a table of x and y values they should know how to enter these into their calculator to find the value of r. If p-values are required in connection with testing a correlation coefficient then these would have to be given in the question. However, since students are expected to be able to calculate probabilities connected with a binomial distribution or a normal distribution, then they could be asked to calculate the p-value in these cases.#

What does the Product Moment Correlation Coefficient (PMCC) describe? And how?	<ul style="list-style-type: none">● It describes how correlated 2 variables are● It can take on any values between -1
--	--

	and +1 inclusively, where $r = +1$ means perfect positive correlation, $r = -1$ means perfect negative correlation, and $r = 0$ means no correlation
--	--

5.3 Conduct a statistical hypothesis test for the mean of a Normal distribution with known, given or assumed variance and interpret the results in context.

[Students should know that: If $X \sim N(\mu, \sigma^2)$ then $\underline{X} \sim N(\mu, \frac{\sigma^2}{n})$ and that a test for μ can be carried out using: $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1^2)$. No proofs required. Hypotheses should be stated in terms of the population mean μ . Knowledge of the Central Limit Theorem or other large sample approximations is not required.]

How is a sample normally distributed? What is required for the population?	<p>If, $X \sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$</p> <p>This requires the population to also be normally distributed</p>
What is the test statistic for a normally distributed sample that is used during a hypothesis test?	<p>$z = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, \bar{x} is the mean of the sample, μ_0 is the hypothesised mean of the distribution, σ^2 is the variance of the distribution and n is the sample size.</p> <p>Key point</p>