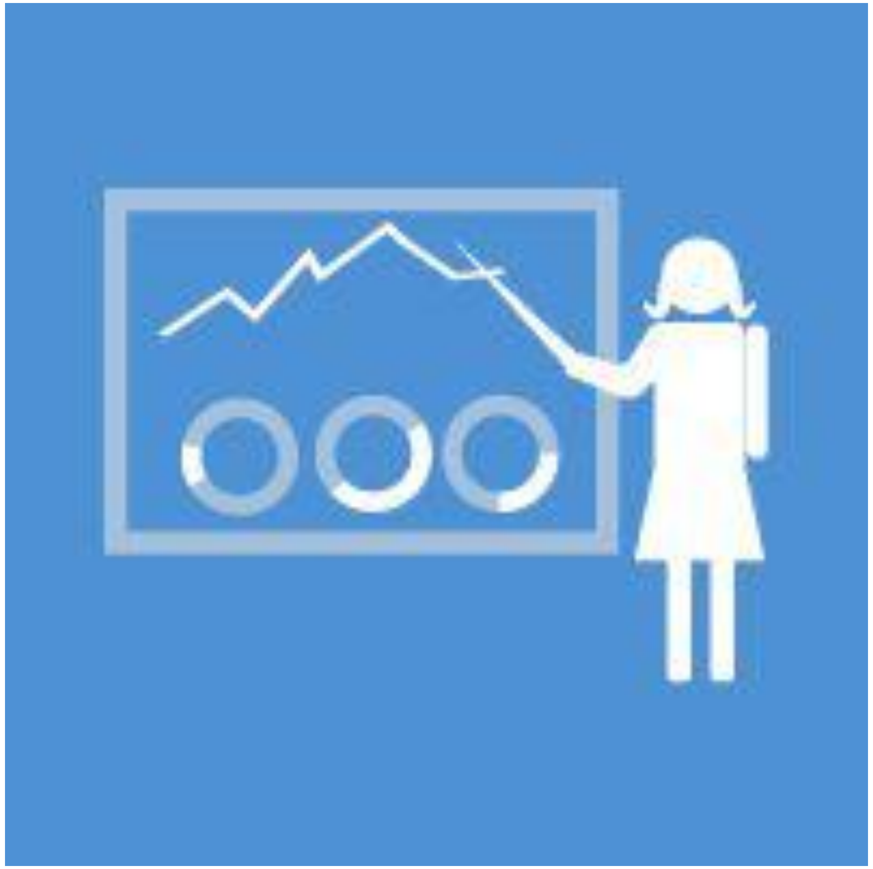
Statistics:
Data Presentation & Sampling



Material Covered

Statistical Sampling

1. Populations and Sampling.
2. Sampling Techniques.

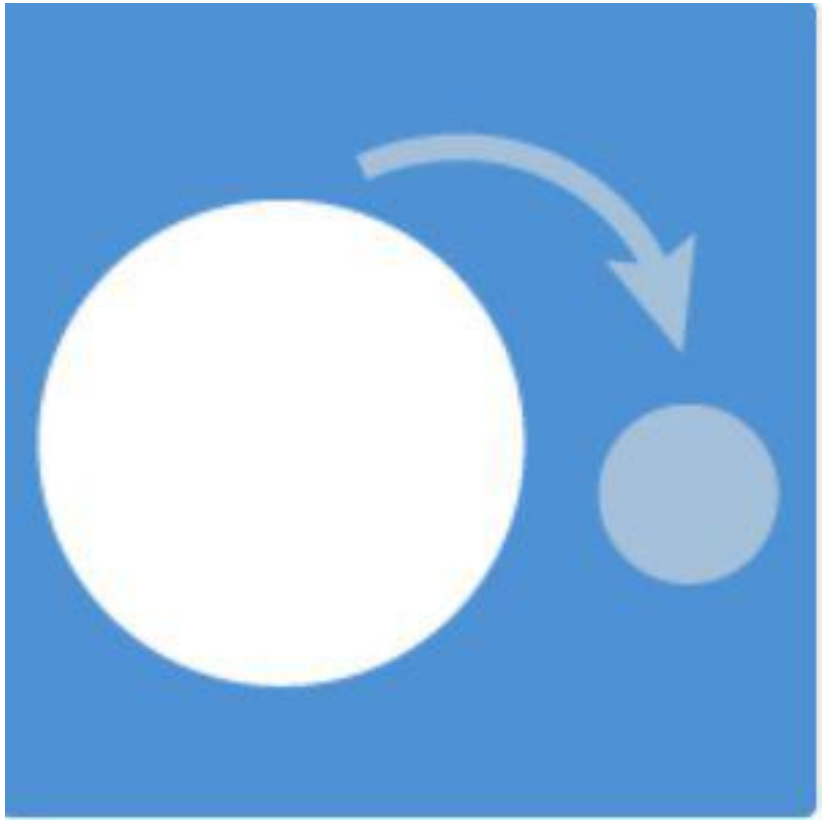
Data Analysis

1. Measures of Central Tendency
2. Measures of Variation
3. Correlation and Outliers.

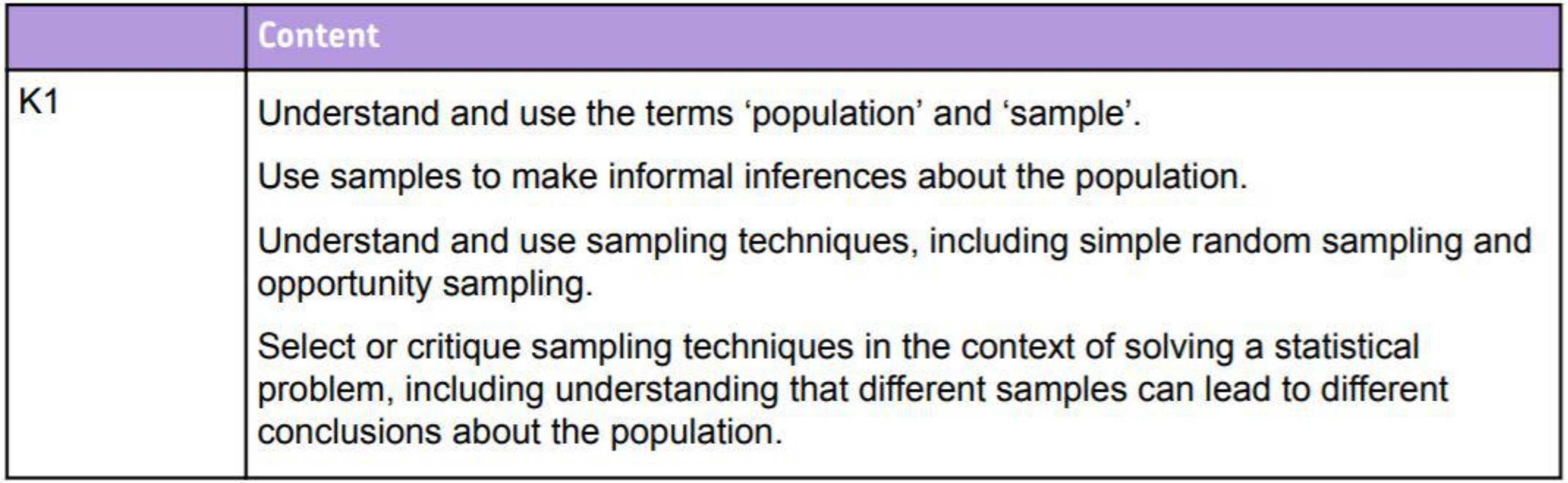
Representing Data

1. Cumulative Frequency Graphs.
2. Histograms.

Statistical Sampling

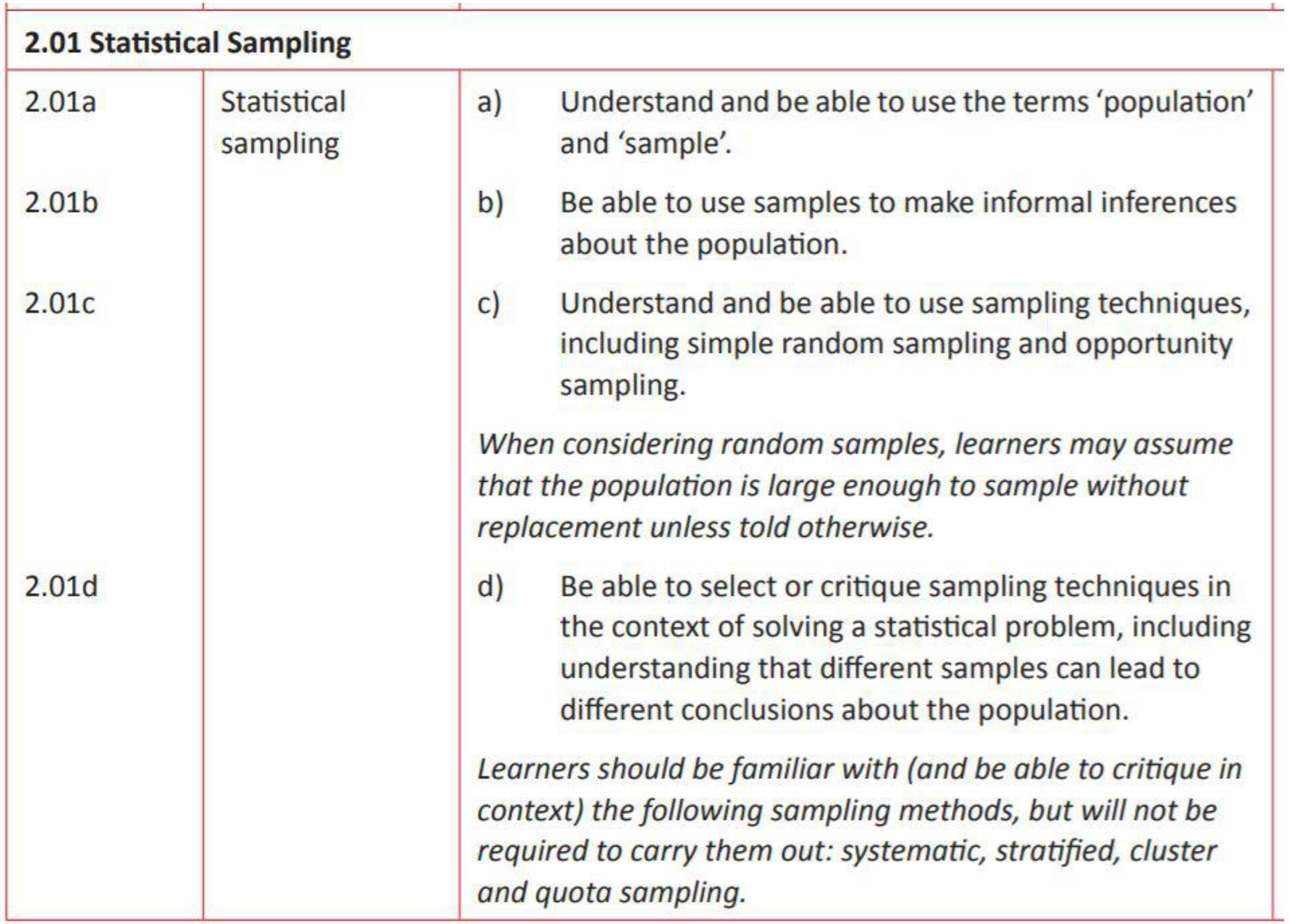


Specification Points - AQA



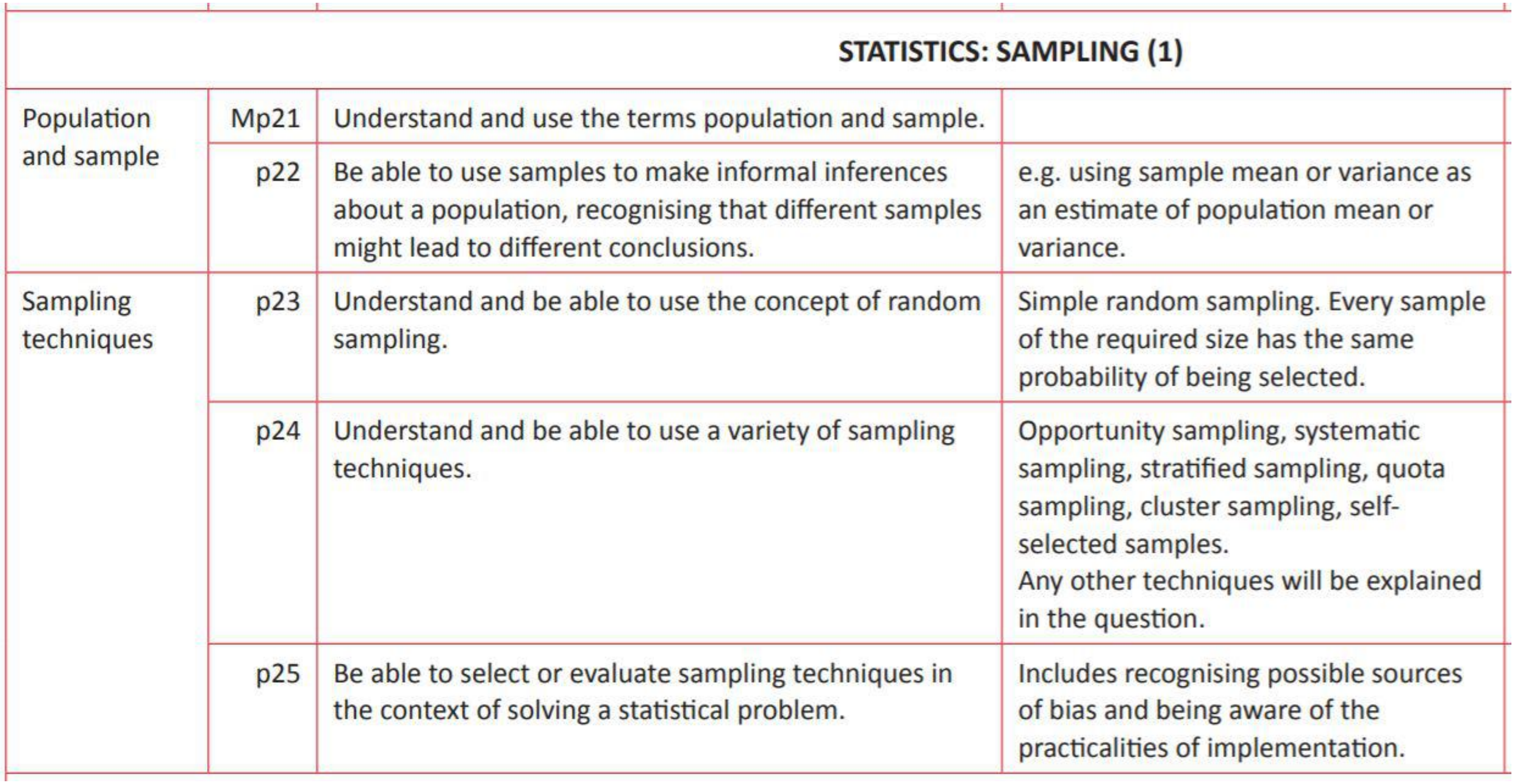
	Content
K1	Understand and use the terms ‘population’ and ‘sample’. Use samples to make informal inferences about the population. Understand and use sampling techniques, including simple random sampling and opportunity sampling. Select or critique sampling techniques in the context of solving a statistical problem, including understanding that different samples can lead to different conclusions about the population.

Specification Points – OCR A



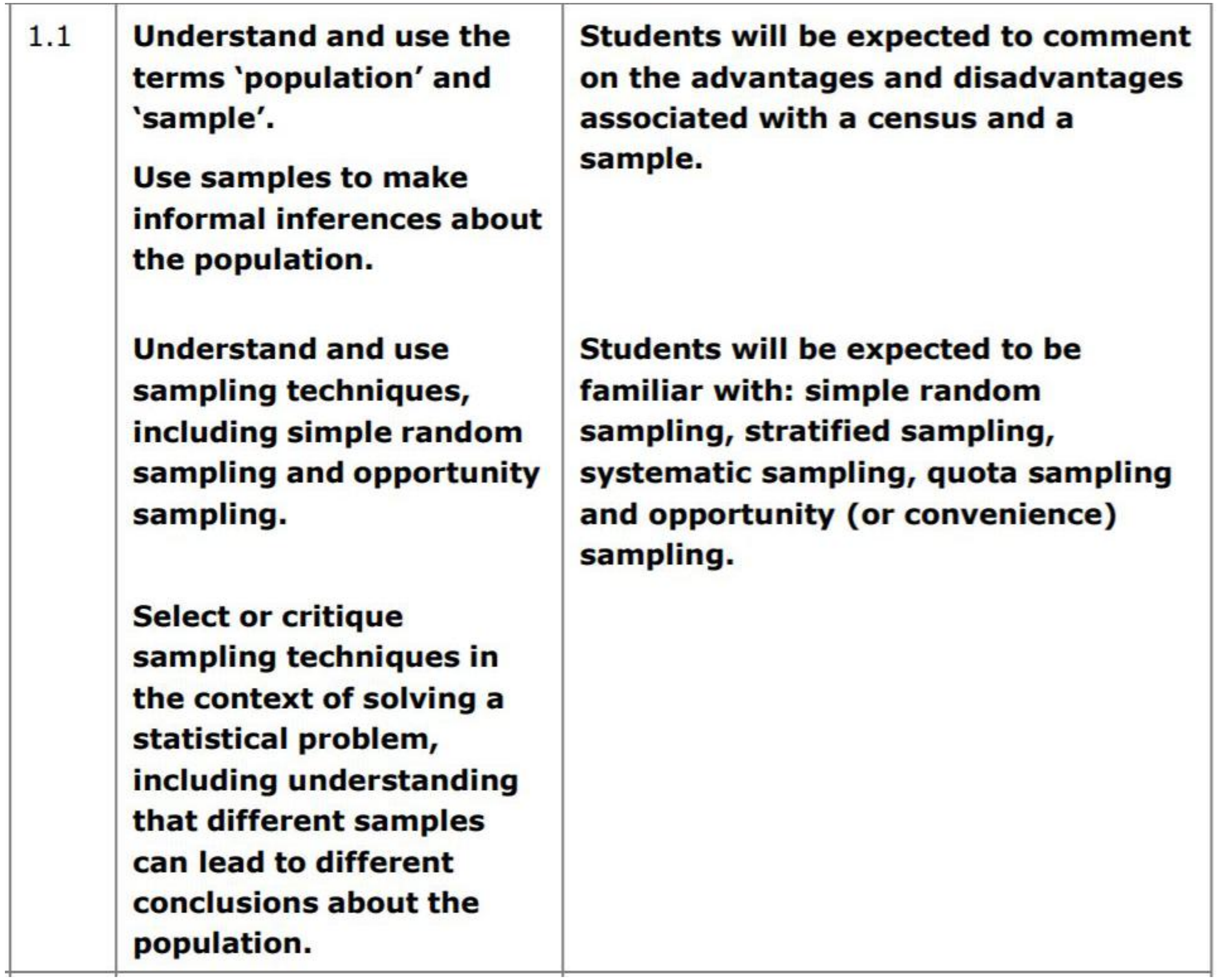
2.01 Statistical Sampling		
2.01a	Statistical sampling	a) Understand and be able to use the terms ‘population’ and ‘sample’.
2.01b		b) Be able to use samples to make informal inferences about the population.
2.01c		c) Understand and be able to use sampling techniques, including simple random sampling and opportunity sampling. *When considering random samples, learners may assume that the population is large enough to sample without replacement unless told otherwise.*
2.01d		d) Be able to 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. *Learners should be familiar with (and be able to critique in context) the following sampling methods, but will not be required to carry them out: systematic, stratified, cluster and quota sampling.*

Specification Points – OCR MEI



		STATISTICS: SAMPLING (1)	
Population and sample	Mp21	Understand and use the terms population and sample.	
	p22	Be able to use samples to make informal inferences about a population, recognising that different samples might lead to different conclusions.	e.g. using sample mean or variance as an estimate of population mean or variance.
Sampling techniques	p23	Understand and be able to use the concept of random sampling.	Simple random sampling. Every sample of the required size has the same probability of being selected.
	p24	Understand and be able to use a variety of sampling techniques.	Opportunity sampling, systematic sampling, stratified sampling, quota sampling, cluster sampling, self-selected samples. Any other techniques will be explained in the question.
	p25	Be able to select or evaluate sampling techniques in the context of solving a statistical problem.	Includes recognising possible sources of bias and being aware of the practicalities of implementation.

Specification Points - Edexcel



1.1	Understand and use the terms ‘population’ and ‘sample’. Use samples to make informal inferences about the population.	Students will be expected to comment on the advantages and disadvantages associated with a census and a sample.
	Understand and use sampling techniques, including simple random sampling and opportunity sampling.	Students will be expected to be familiar with: simple random sampling, stratified sampling, systematic sampling, quota sampling and opportunity (or convenience) 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.	

Population and Sampling

- **Population** – **Whole set** of **items** that we are **concerned** with.

- **Sample** – **Selection** of **observations** taken from a **subset** of the **population** which is used to **infer information** about the **population** as a **whole**.
- **Census** – **Observation** of **every member** of a **population**.
- **Bias** – A **cause** for the **results** of a **sample** to be **unrepresentative** of the **population**.

Population and Sampling

There are different **advantages** and **disadvantages** of a **census** and a **sample**.

	Advantages	**Disadvantages**
Sample	• Quick and easy. • Less data to process.	• Result may be inaccurate/biased if sample is not large enough.
Census	• Accurate result. • No bias.	• Time consuming and expensive. • Cannot be used when process destroys item.

Question testing definitions

Exemplar Exam Question

1) A biscuit factory produces 1,000,000 biscuits a day. The factory employs quality assurance testers who test the biscuits by eating them to ensure that they are good enough to sell. For a single day in the factory:

(i) State the population for the quality assurance test. **[1 Mark]**
(ii) State how many biscuits would have to be tasted if a census was taken for the quality assurance test. **[1 Mark]**
(iii) Suggest, with reasons, if it is sensible to for the quality assurance testers to use a census rather than a sample. **[2 Marks]**

Think about the **logistics** of this

2 simple statements and a **short explanation**

Exemplar Exam Question Answer

(i) State population

The population are the 1,000,000 biscuits produced in the factory that day.

[1 Mark]

Exemplar Exam Question Answer

(ii) State number of biscuits in census

A census is an observation of the entire sample, so there would be 1,000,000 biscuits tasted.

[1 Mark]

Exemplar Exam Question Answer

(iii) Analyse logistics of taste test

The quality assurance test requires the biscuit to be eaten

So performing a census would involve all the biscuits produced that day being eaten by the testers

[1 Mark]

Which defeats the purpose of making sure they're good enough to sell

Would also be a bad idea because of the time it would take to test every single biscuit.

Therefore it is more sensible to take a sample than to use a census.

[1 Mark]

Random Sampling Techniques

In **random sampling techniques**, each **member** of the **population** has an **equal chance** of being **selected**.

- **Simple Random Sampling** – **Sampling** where every **item** has an **equal chance** of being **selected**.

  + Easy and cheap.

  + Free of bias.

  - Selection is time-consuming for large populations.

Random Sampling Techniques

- **Stratified Sampling – Sampling** where the **population** is **divided** into **mutually exclusive strata** and a **random sample** is taken from **each**.

+ Reflects population structure.

+ Gives proportional representation of groups.

- Population is not always clearly classified into distinct groups.

$$\boxed{\textbf{Number Sampled in Stratum} = \frac{\textbf{Number in Stratum}}{\textbf{Number in Population}} \times \textbf{Sample Size}}$$

Non-Random Sampling Techniques

In **non-random sampling**, **members** of the **population** are **chosen** by a **researcher**. This is often **quicker** and **simpler** than **random sampling techniques**.

- **Quota Sampling** – **Sampling** where the **researcher** selects a **sample** that **reflects** the **characteristics** of the **entire population**.

  + No sampling frame required.

  + Can select to represent different groups.

  - Can introduce bias.

  - Sample must be selected (costly and inaccurate).

  - Non-responses not recorded as such.

Non-Random Sampling Techniques

- **Opportunity Sampling** – **Sampling** where the **sample** is **taken** from **people** who are **available** at the **time** of the **study** and fit its **criteria**.

+ Easy and cheap to conduct.

- Unlikely to provide a representative sample.

- Depends on researcher, time, location.

What is the **population**?

Exemplar Exam Question

Remember **info** about each **sample method**

1) The UK manufacturer “Broom to Breathe” are looking to get opinions from potential customers on a new product range. They are considering multiple ways of finding this information. For each of the following methods, determine what kind of sample is being taken and suggest a reason for or against using the method.

(i) Asking attendees at a housekeeping convention.

[1 Mark]

(ii) Randomly selecting 100 people with interest in brooms.

[1 Mark]

(iii) Breaking down their customer base by age range and choosing sample groups from each.

[1 Mark]

Question on **definitions**

Identify **key aspects** and **compare to population**

3 marks across **3 parts**, **not much detail needed**

Exemplar Exam Question Answer

(i) Identify type of sample

Many potential customers would be at the convention.

So going here for the survey would be <u>Opportunity Sampling</u>.

Suggest positives or negatives of method

+ Quick and cheap to do.

− Housekeeping enthusiasts don't reflect the company's entire customer base.

[1 Mark]

Exemplar Exam Question Answer

(ii) Identify type of sample

100 random people are being chosen from population.

So this method is <u>Simple Random Sampling </u>.

Suggest positives or negatives of method

+ Easy to do with no bias.

− Might not fully represent the entire population.

[1 Mark]

Exemplar Exam Question Answer

(iii) Identify type of sample

Groups are being broken down and selected from.

So this method is Stratified Random Sampling.

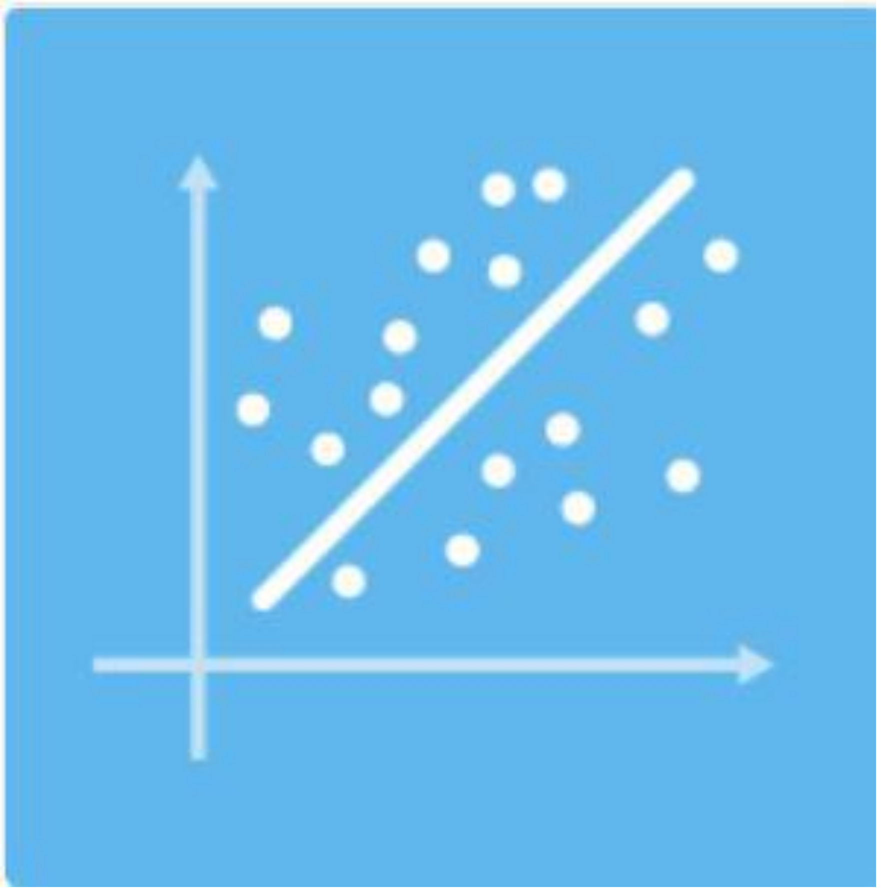
Suggest positives or negatives of method

+ Population will be clearly broken down and proportionally represented.

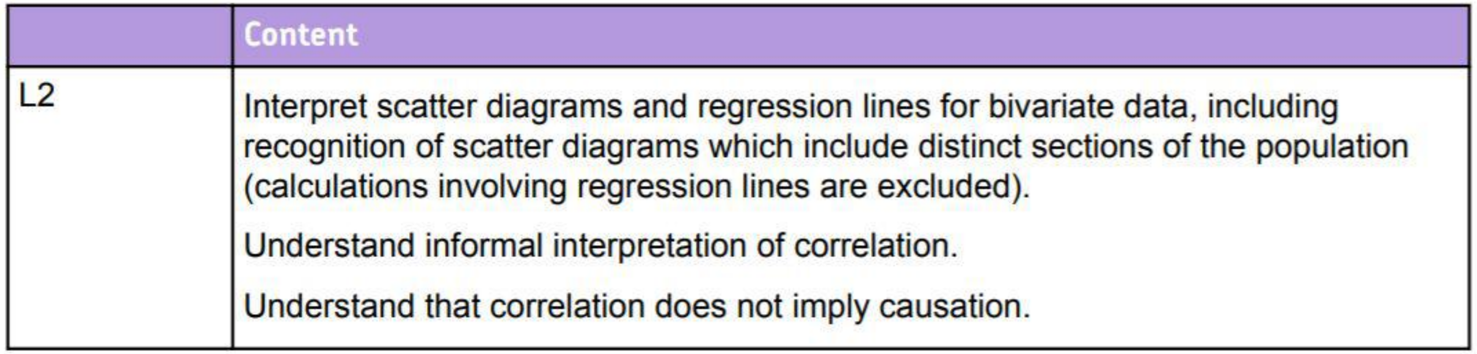
− Age isn't necessarily a useful way of breaking down broom users.

[1 Mark]

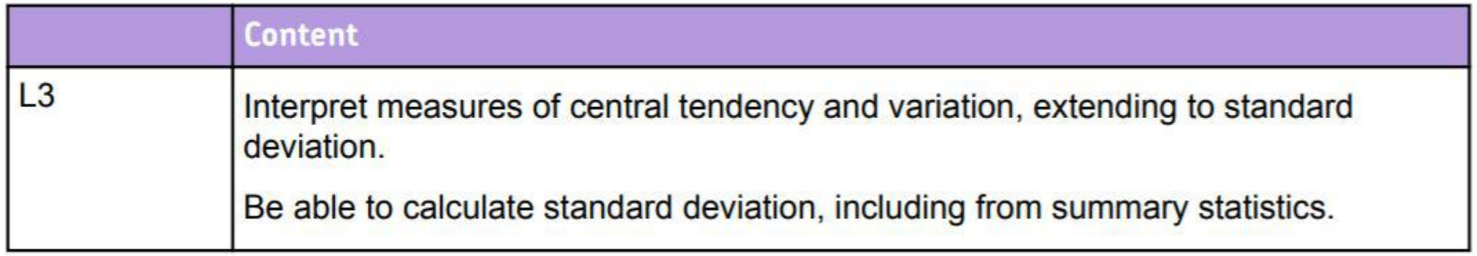
Data Analysis



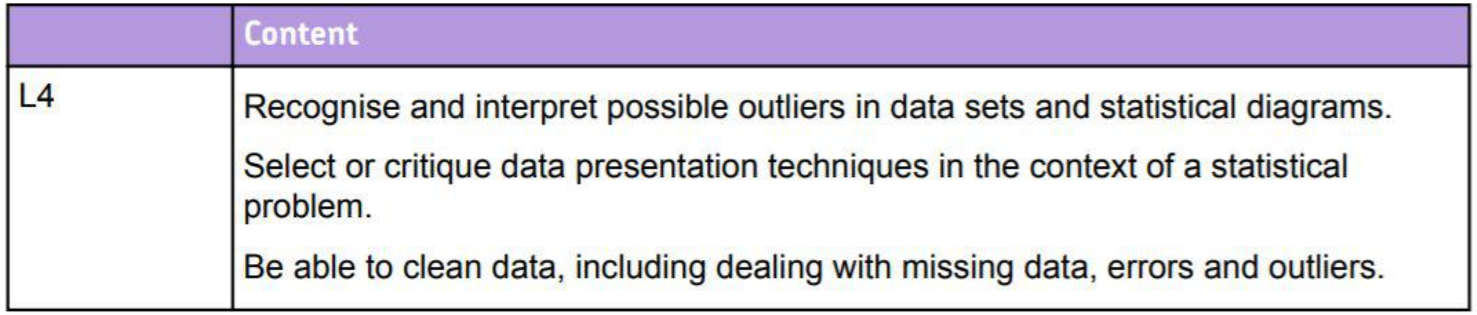
Specification Points - AQA



	Content
L2	Interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams which include distinct sections of the population (calculations involving regression lines are excluded). Understand informal interpretation of correlation. Understand that correlation does not imply causation.

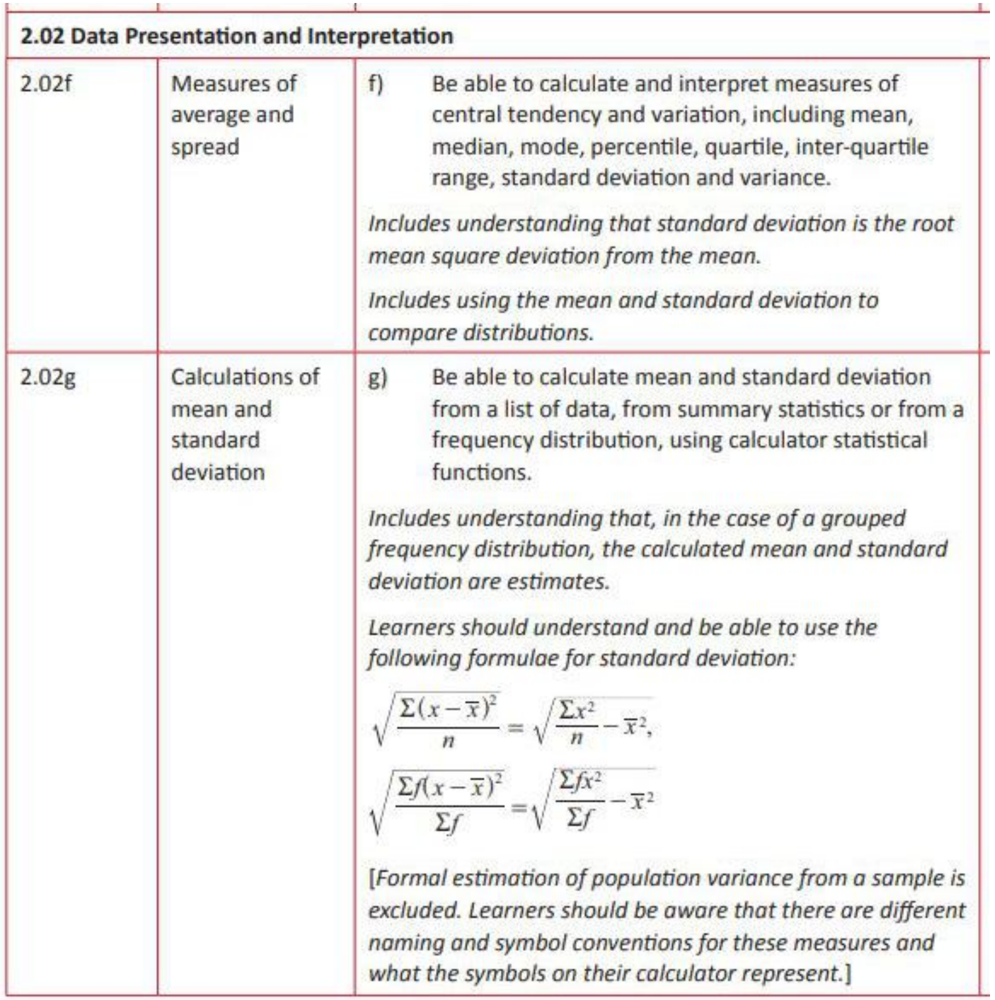


	Content
L3	Interpret measures of central tendency and variation, extending to standard deviation. Be able to calculate standard deviation, including from summary statistics.

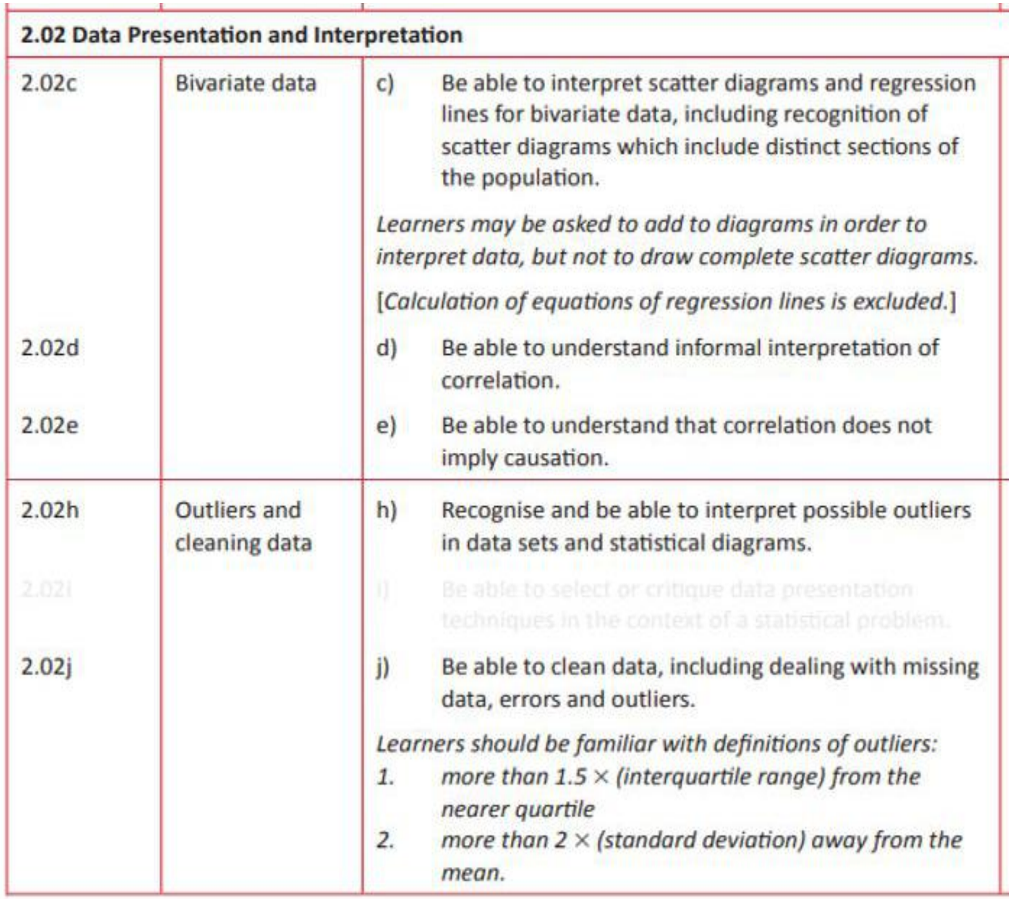


	Content
L4	Recognise and interpret possible outliers in data sets and statistical diagrams. Select or critique data presentation techniques in the context of a statistical problem. Be able to clean data, including dealing with missing data, errors and outliers.

Specification Points – OCR A

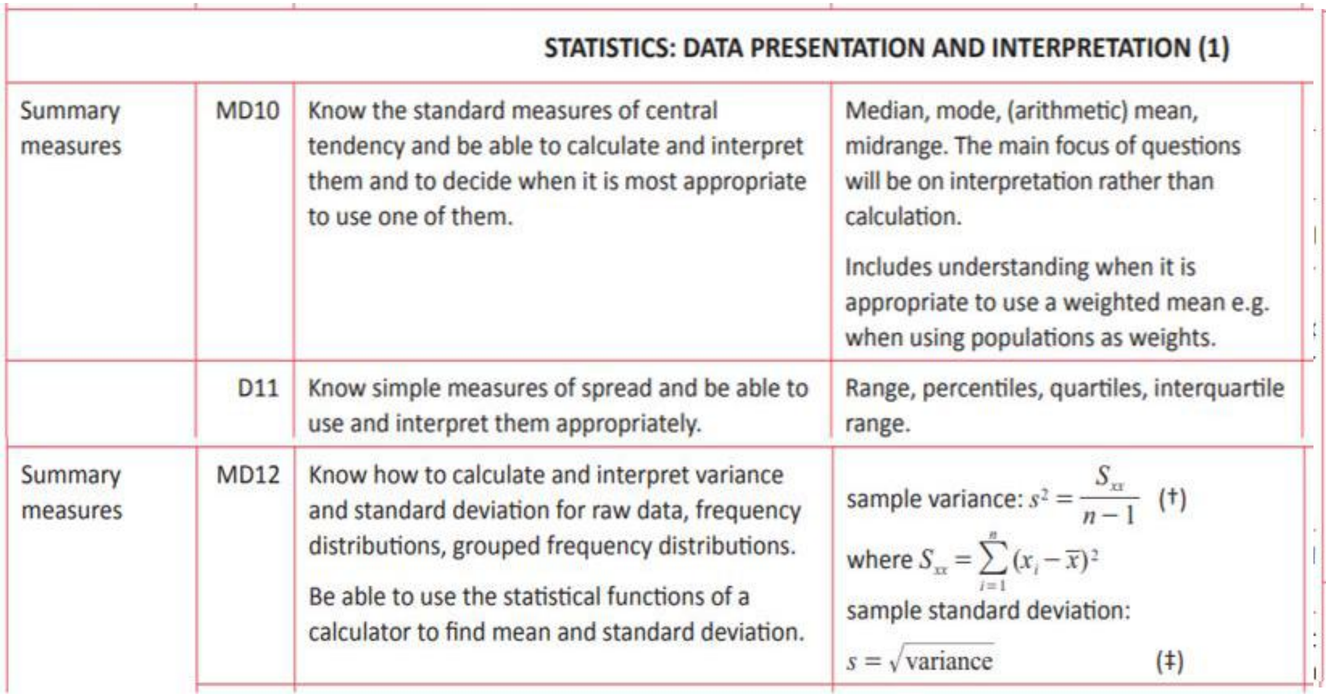


2.02 Data Presentation and Interpretation		
2.02f	Measures of average and spread	f) Be able to calculate and interpret measures of central tendency and variation, including mean, median, mode, percentile, quartile, inter-quartile range, standard deviation and variance. *Includes understanding that standard deviation is the root mean square deviation from the mean.* *Includes using the mean and standard deviation to compare distributions.*
2.02g	Calculations of mean and standard deviation	g) Be able to calculate mean and standard deviation from a list of data, from summary statistics or from a frequency distribution, using calculator statistical functions. *Includes understanding that, in the case of a grouped frequency distribution, the calculated mean and standard deviation are estimates.* *Learners should understand and be able to use the following formulae for standard deviation:* $\sqrt{\frac{\Sigma(x-\overline{x})^2}{n}} = \sqrt{\frac{\Sigma x^2}{n} - \overline{x}^2}$, $\sqrt{\frac{\Sigma f(x-\overline{x})^2}{\Sigma f}} = \sqrt{\frac{\Sigma fx^2}{\Sigma f} - \overline{x}^2}$ *[Formal estimation of population variance from a sample is excluded. Learners should be aware that there are different naming and symbol conventions for these measures and what the symbols on their calculator represent.]*

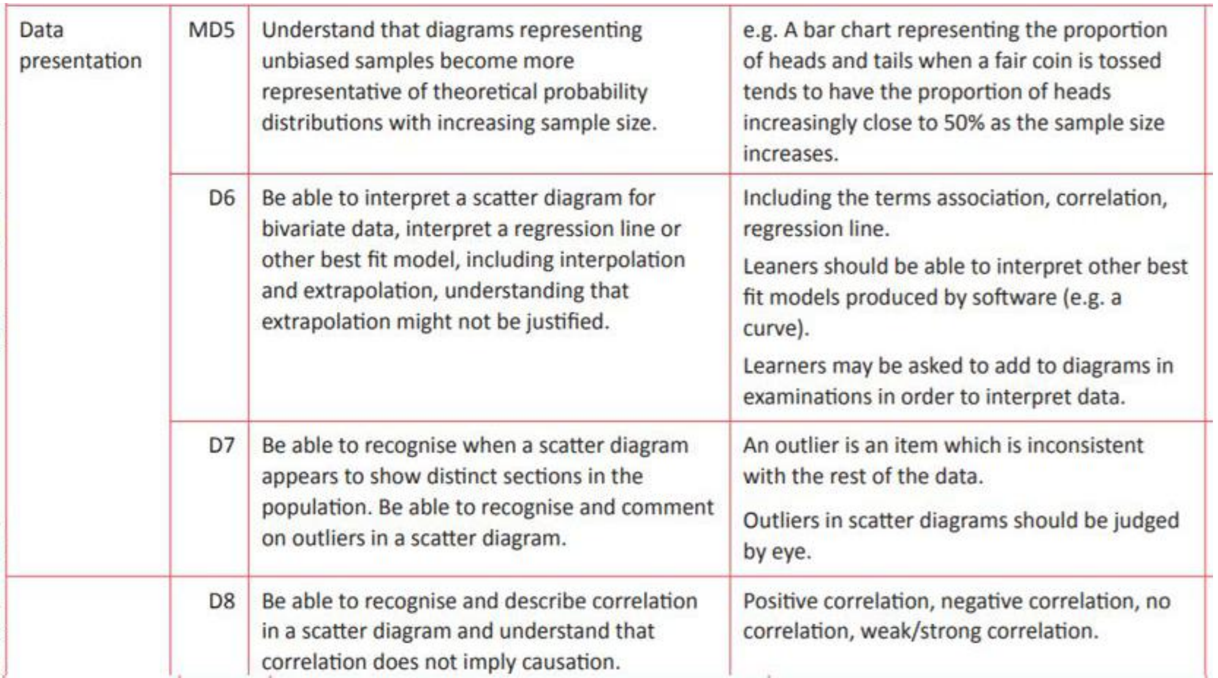


2.02 Data Presentation and Interpretation		
2.02c	Bivariate data	c) Be able to interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams which include distinct sections of the population. *Learners may be asked to add to diagrams in order to interpret data, but not to draw complete scatter diagrams.* *[Calculation of equations of regression lines is excluded.]*
2.02d		d) Be able to understand informal interpretation of correlation.
2.02e		e) Be able to understand that correlation does not imply causation.
2.02h	Outliers and cleaning data	h) Recognise and be able to interpret possible outliers in data sets and statistical diagrams.
2.02i		i) Be able to select or critique data presentation techniques in the context of a statistical problem.
2.02j		j) Be able to clean data, including dealing with missing data, errors and outliers. *Learners should be familiar with definitions of outliers:* *1. more than 1.5 × (interquartile range) from the nearer quartile* *2. more than 2 × (standard deviation) away from the mean.*

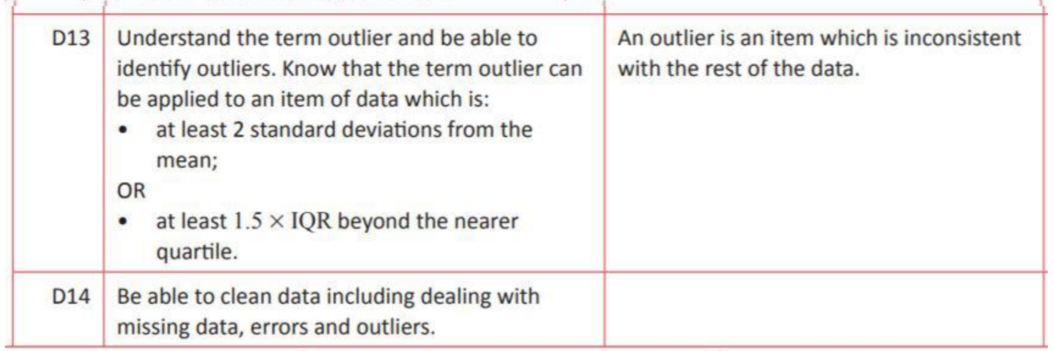
Specification Points – OCR MEI



STATISTICS: DATA PRESENTATION AND INTERPRETATION (1)			
Summary measures	MD10	Know the standard measures of central tendency and be able to calculate and interpret them and to decide when it is most appropriate to use one of them.	Median, mode, (arithmetic) mean, midrange. The main focus of questions will be on interpretation rather than calculation. Includes understanding when it is appropriate to use a weighted mean e.g. when using populations as weights.
	D11	Know simple measures of spread and be able to use and interpret them appropriately.	Range, percentiles, quartiles, interquartile range.
Summary measures	MD12	Know how to calculate and interpret variance and standard deviation for raw data, frequency distributions, grouped frequency distributions. Be able to use the statistical functions of a calculator to find mean and standard deviation.	sample variance: $s^2 = \frac{S_{xx}}{n-1}$ (†) where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$ sample standard deviation: $s = \sqrt{\text{variance}}$ (‡)

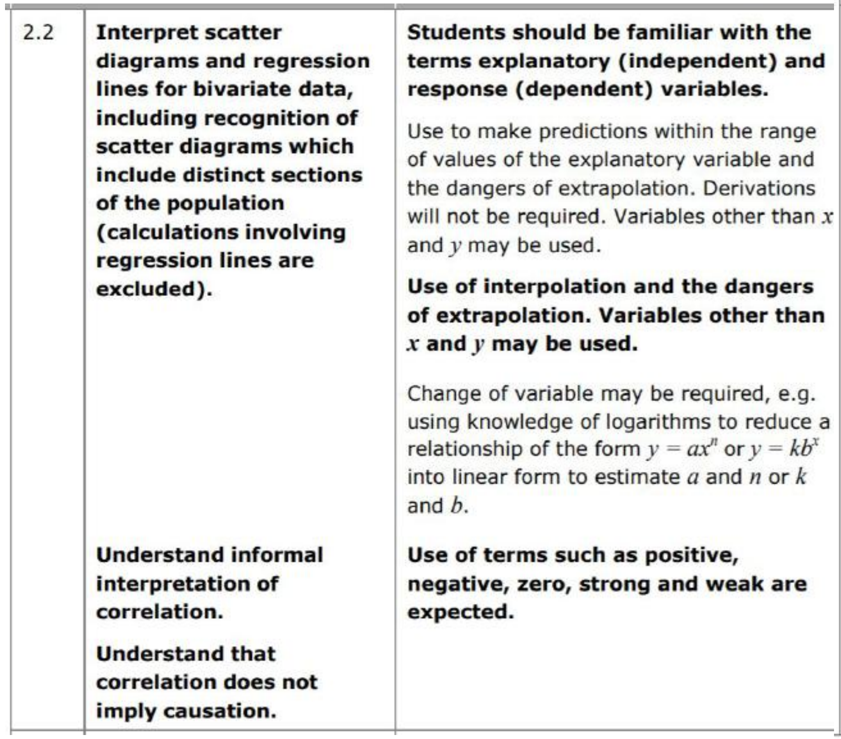


Data presentation	MD5	Understand that diagrams representing unbiased samples become more representative of theoretical probability distributions with increasing sample size.	e.g. A bar chart representing the proportion of heads and tails when a fair coin is tossed tends to have the proportion of heads increasingly close to 50% as the sample size increases.
	D6	Be able to interpret a scatter diagram for bivariate data, interpret a regression line or other best fit model, including interpolation and extrapolation, understanding that extrapolation might not be justified.	Including the terms association, correlation, regression line. Leaners should be able to interpret other best fit models produced by software (e.g. a curve). Learners may be asked to add to diagrams in examinations in order to interpret data.
	D7	Be able to recognise when a scatter diagram appears to show distinct sections in the population. Be able to recognise and comment on outliers in a scatter diagram.	An outlier is an item which is inconsistent with the rest of the data. Outliers in scatter diagrams should be judged by eye.
	D8	Be able to recognise and describe correlation in a scatter diagram and understand that correlation does not imply causation.	Positive correlation, negative correlation, no correlation, weak/strong correlation.

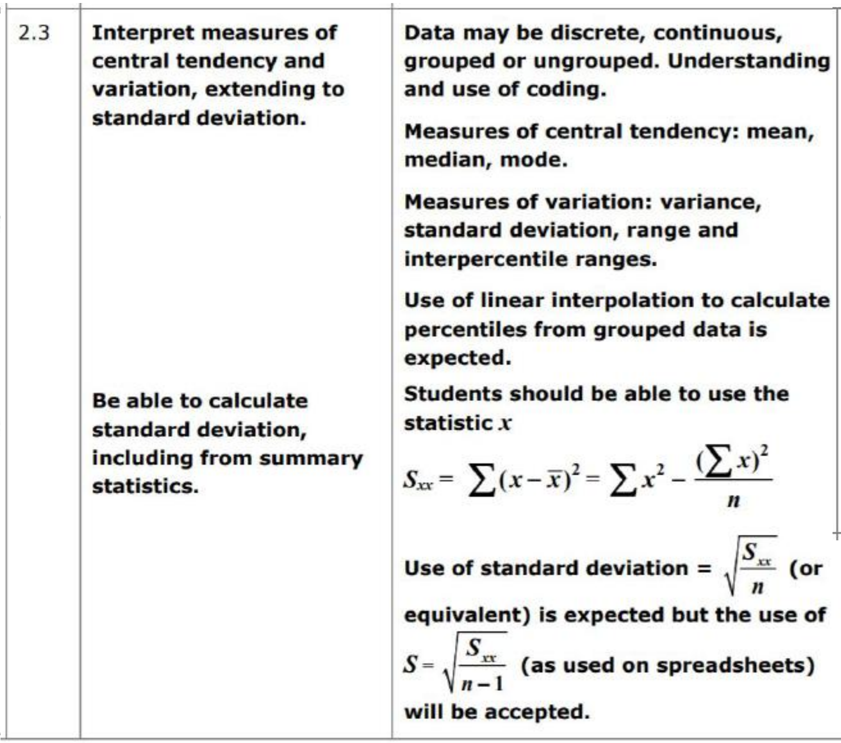


D13	Understand the term outlier and be able to identify outliers. Know that the term outlier can be applied to an item of data which is: • at least 2 standard deviations from the mean; OR • at least $1.5 \times \text{IQR}$ beyond the nearer quartile.	An outlier is an item which is inconsistent with the rest of the data.
D14	Be able to clean data including dealing with missing data, errors and outliers.	

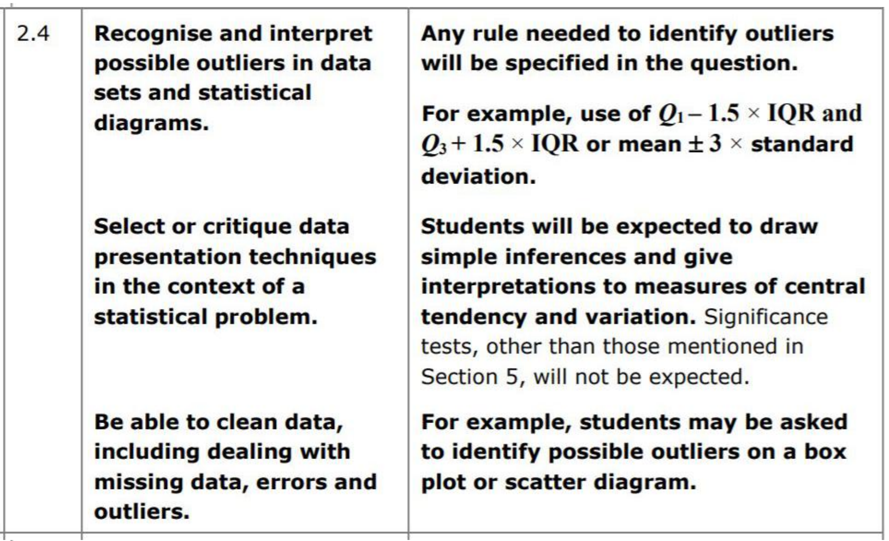
Specification Points - Edexcel



2.2	**Interpret scatter diagrams and regression lines for bivariate data, including recognition of scatter diagrams which include distinct sections of the population (calculations involving regression lines are excluded).**	**Students should be familiar with the terms explanatory (independent) and response (dependent) variables.** Use to make predictions within the range of values of the explanatory variable and the dangers of extrapolation. Derivations will not be required. Variables other than x and y may be used. **Use of interpolation and the dangers of extrapolation. Variables other than x and y may be used.** 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.
	Understand informal interpretation of correlation. **Understand that correlation does not imply causation.**	**Use of terms such as positive, negative, zero, strong and weak are expected.**



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



2.4	**Recognise and interpret possible outliers in data sets and statistical diagrams.**	**Any rule needed to identify outliers will be specified in the question.** **For example, use of $Q_1 - 1.5 \times \text{IQR}$ and $Q_3 + 1.5 \times \text{IQR}$ or mean ± 3 × standard deviation.**
	Select or critique data presentation techniques in the context of a statistical problem.	**Students will be expected to draw simple inferences and give interpretations to measures of central tendency and variation.** Significance tests, other than those mentioned in Section 5, will not be expected.
	Be able to clean data, including dealing with missing data, errors and outliers.	**For example, students may be asked to identify possible outliers on a box plot or scatter diagram.**

Measures of Central Tendency

A **measure** of **central tendency** describes the **centre** of a **group of data**.

- **Mode** – The **most common** value or class.
- **Median** – The **middle value** or class when the data is put in order from highest to lowest.

Measures of Central Tendency

- **Mean** – The **average value** or class obtained by **summing** the **values** and **dividing** by the **number of values**.

$$\boxed{\bar{x} = \frac{\sum x}{n}}$$

$$\boxed{\bar{x} = \frac{\sum xn}{\sum n}}$$

- If x is sorted into **classes** of **continuous data**, use the **midpoint** of each **class**.

Calculations for **ungrouped data**.

Exemplar Exam Question

-

No decimals. 3 **different values.** Already **in order**

Testing knowledge of **definitions. Can't just use calculator.**

4 marks in total – one value is going to be **harder** to find than others

Exemplar Exam Question Answer

9 data values total

[1 Mark]

Exemplar Exam Question Answer

Use information about mode to deduce another value

z must be either 5 or 6

Mode of data is 6

Therefore the value 6 must appear more than any other value

So z cannot be 5

$\Rightarrow z = 6$

[1 Mark]

Exemplar Exam Question Answer

Use information about mean to deduce final value

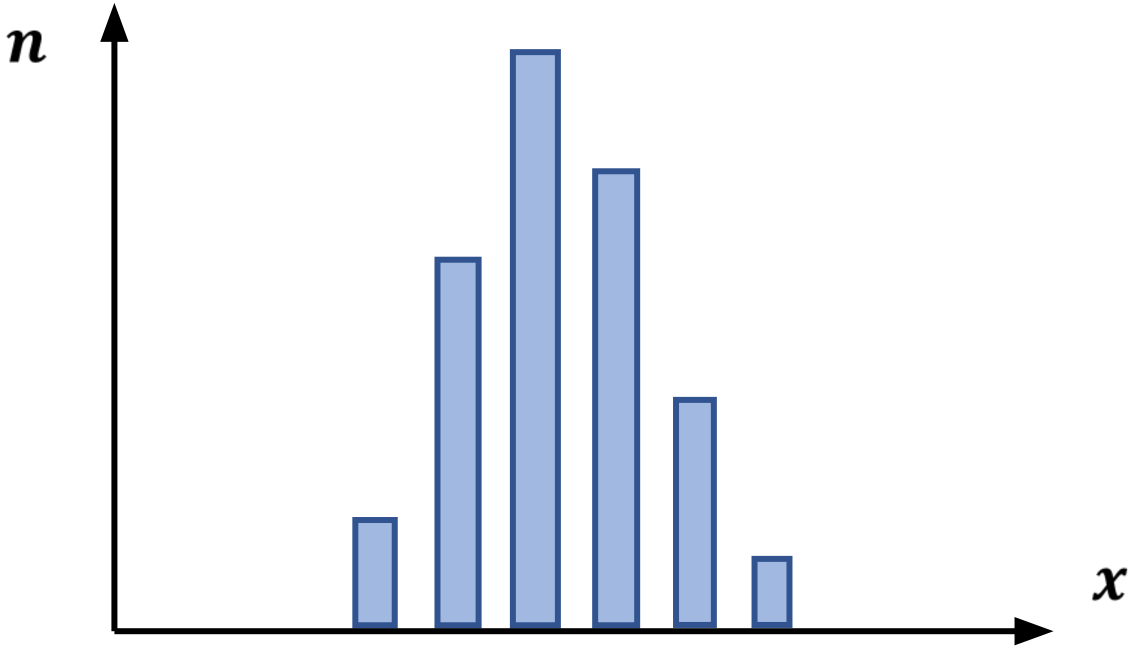
[1 Mark]

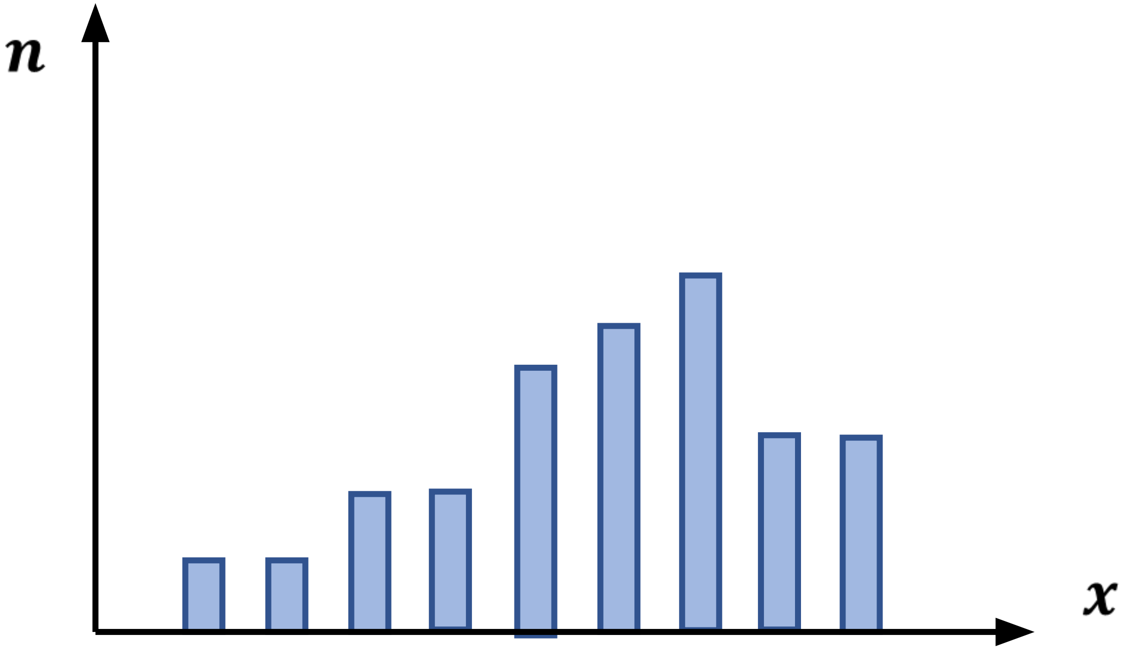
[1 Mark]

Measures of Variation

Measures of Variation

Measures of Variation





Working with grouped data

Exemplar Exam Question

Need to find **midpoints** to work with **ranged data**

1) Arnold has recently lost his 12-sided die and plans to just use two 6-sided dice instead.

His friend Beth thinks this is a bad idea and tries to show this by rolling two 6-sided dice 100 times. She records the number of times, n, that the total value of her throw, x, is in a certain range. Her results are displayed in the following table.

Calculate the variance of scores from her dice throws to 3 significant figures, and use your result to comment on why the 12-sided die cannot be substituted with two 6-sided dice. You may use the fact that the variance of 100 throws from a 12-sided die is approximately 10.

[4 marks]

Need to **analyse and interpret** results

4 mark question, **2 steps** each for **calculations and conclusion**

Exemplar Exam Question

	21
	25
	15
	22
	17

Exemplar Exam Question Answer

Calculate midpoints of ranges

Results are given in data ranges

So we need to calculate the midpoint of each range for variance calculations

	21
	25
	15
	22
	17

[1 Mark]

Exemplar Exam Question Answer

Calculate variance

Input new table to calculator and calculate

	21
	25
	15
	22
	17

[1 Mark]

Exemplar Exam Question Answer

$\sigma_{12}^2 = 10$ $\sigma_6^2 = 7.10$

Interpret data and write conclusion

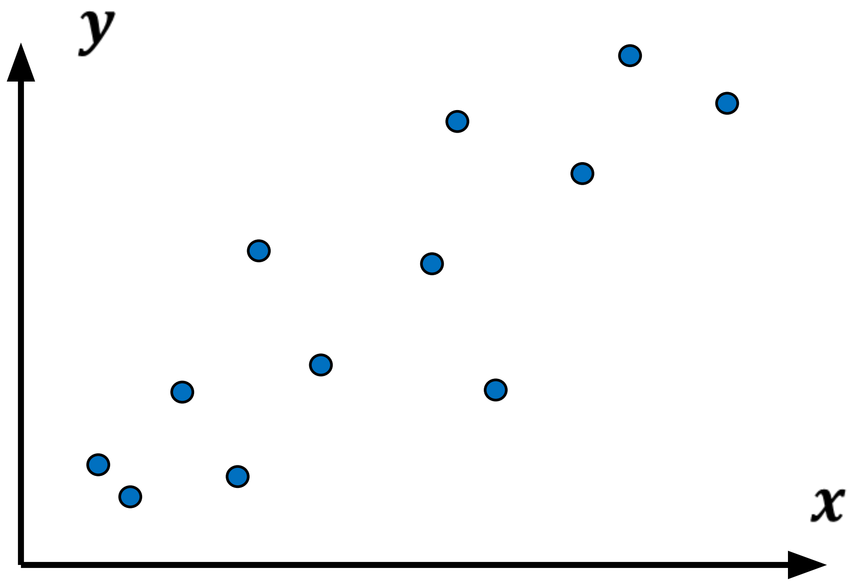
The variance for using two 6-sided dice is much lower than for using the 12-sided die

This means that the values from throwing a 12-sided die are fairly evenly spread , while throwing two 6-sided dice will more often give values close to the mean value

Therefore, the two 6-sided dice do not make a good substitute, since the values need to be more evenly spread

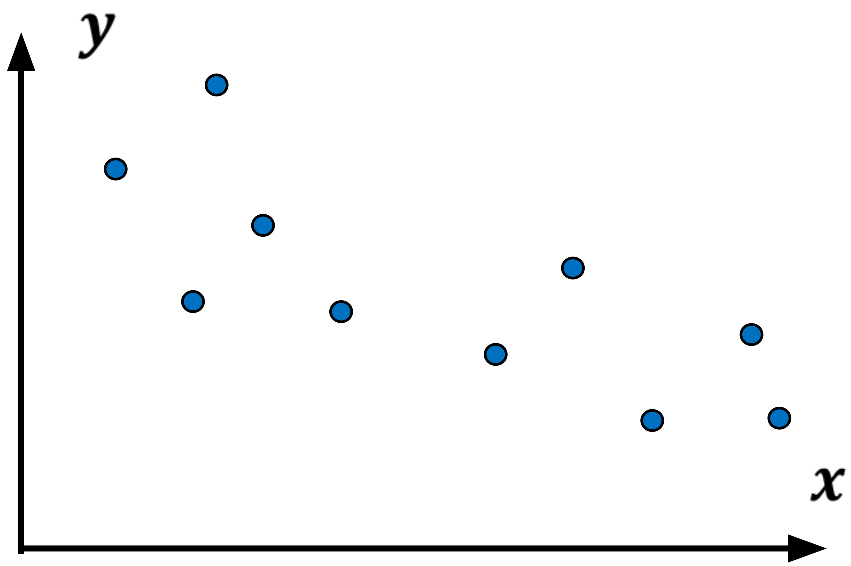
[2 Marks]

Correlation



- **2 variables** (x and y) show **correlation** if an **increase** in x usually coincides with an **increase** in y.

- x and y are **negatively correlated** if an **increase** in x usually **coincides** with a **decrease** in y.

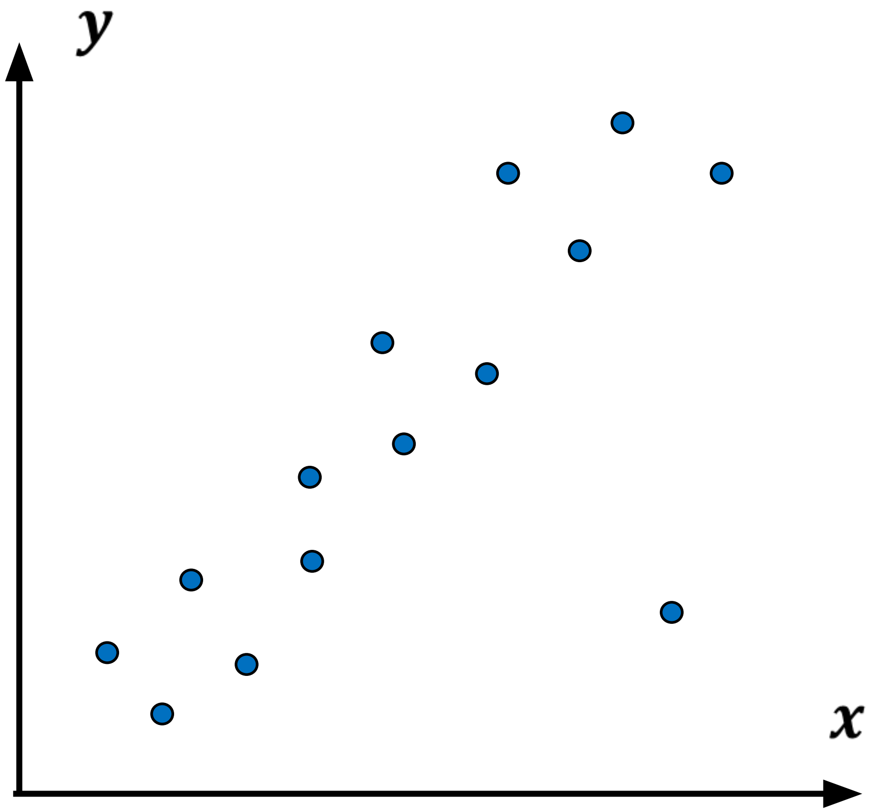


- **No correlation** implies that there is **no relationship** between the **values** of x and y.

Outliers

An **outlier** is an **extreme value** that **does not fit** with the **rest** of the **data**.

- **Removing anomalies** is called **cleaning** the **data**.
- We may want to remove an **outlier** as it may not be a **valid result** and would **lead** to **inaccurate conclusions**.
- If the **outlier** is a **true value**, **removing** it would give **false conclusions**.



Outliers

The equation to use will be specified in an exam!

There is no single **method** to **calculate** which **values** in **data** are **outliers**.

- A **value** can be considered an **outlier** if it is more than: $Q_3 + 1.5 \times IQR$

  OR less than: $Q_1 - 1.5 \times IQR$

- A **value** may also be considered an **outlier** if it lies **outside** the **range**:

AQA/**OCR**:

$$[\bar{x} - 2\sigma\,, \bar{x} + 2\sigma]$$

Edexcel:

$$[\bar{x} - 3\sigma\,, \bar{x} + 3\sigma]$$

How much **lower** does it have to be for **positive result?**

Exemplar Exam Question

Need to **input** these to **calculator**

1) A group of students were given a set of unknown alloys A to I of a metal and were tasked with measuring their melting points. Their results are recorded in the following table, with temperatures given in degrees Celsius.

The students had difficulty working with alloy I and measured a much lower melting point in comparison to the other metals. Test if the melting point for I is an outlier in comparison to the rest of the given data using the standard deviation test $\boldsymbol{x < \bar{x} - 2\sigma}$

[2 marks]

What **measures** do we need for this?

2 mark question, 1 for **finding mean and s.d.**, other for **testing value**

Exemplar Exam Question

Alloy	Melting Point		Alloy	Melting Point
A			F	
B			G	
C			H	
D			I	
E				

Exemplar Exam Question Answer

Input data to calculate mean and standard Deviation

1	
2	
3	
4	
5	
6	
7	
8	
9	

$$\overline{x} = 954°\mathrm{C}$$

$$\sigma = 575°\mathrm{C}$$

[1 Mark]

Exemplar Exam Question Answer

Compare potential outlier to σ

$\bar{x} = 954$°C

$\sigma = 575$°C

$x_I = -38.86$°C

The temperature of I, x_I, is considered to be an outlier if

$$x_I < \bar{x} - 2\sigma$$

Using results for mean and standard deviation

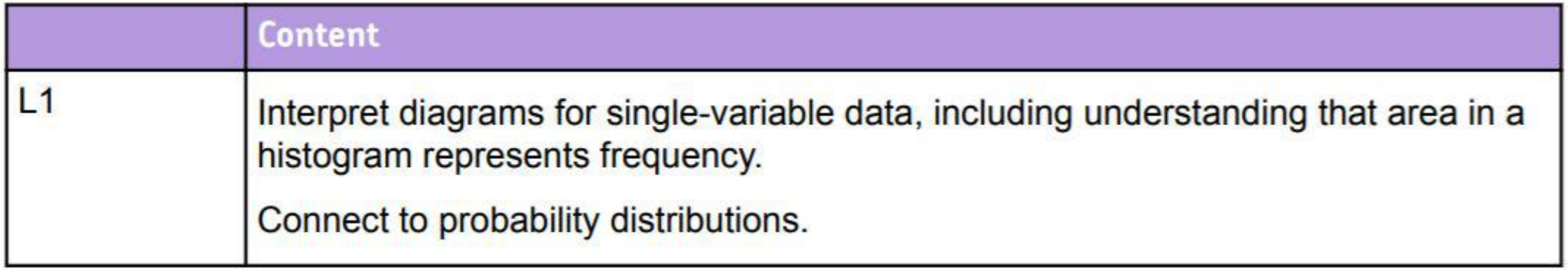
$$\bar{x} - 2\sigma = 954 - (2 \times 575) = -196 < -38.86\text{°C}$$

So melting point of metal I <u>is not</u> an outlier in terms of given standard deviation test

[1 Mark]

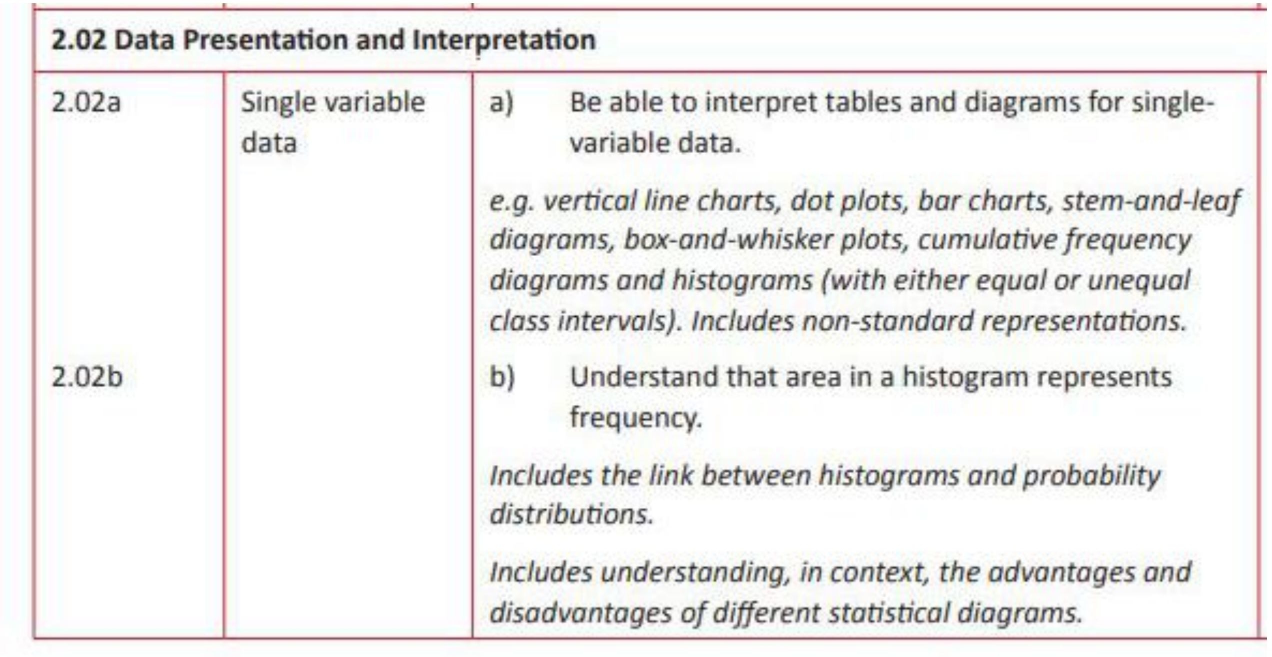
Representing Data

Specification Points - AQA



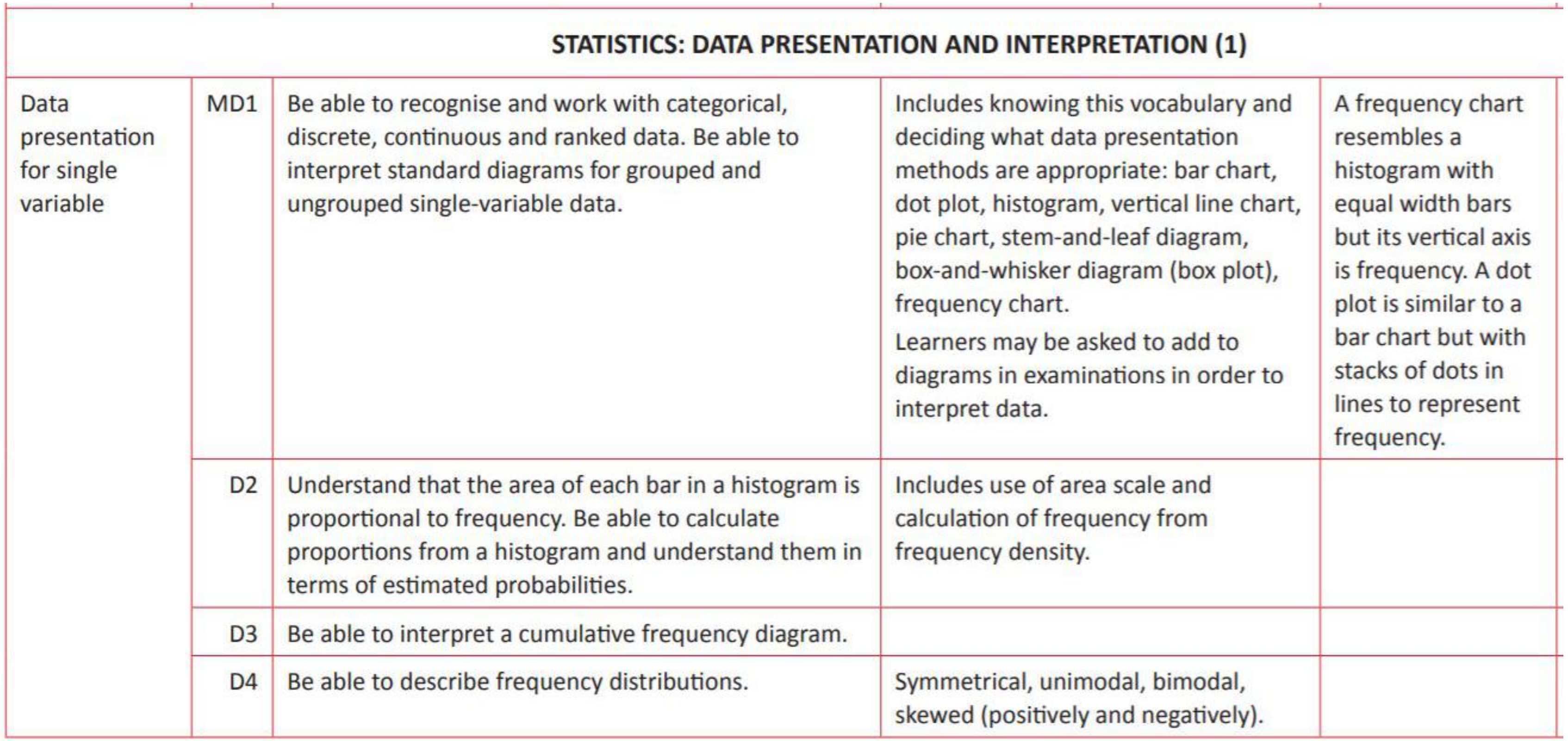
	Content
L1	Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency. Connect to probability distributions.

Specification Points – OCR A



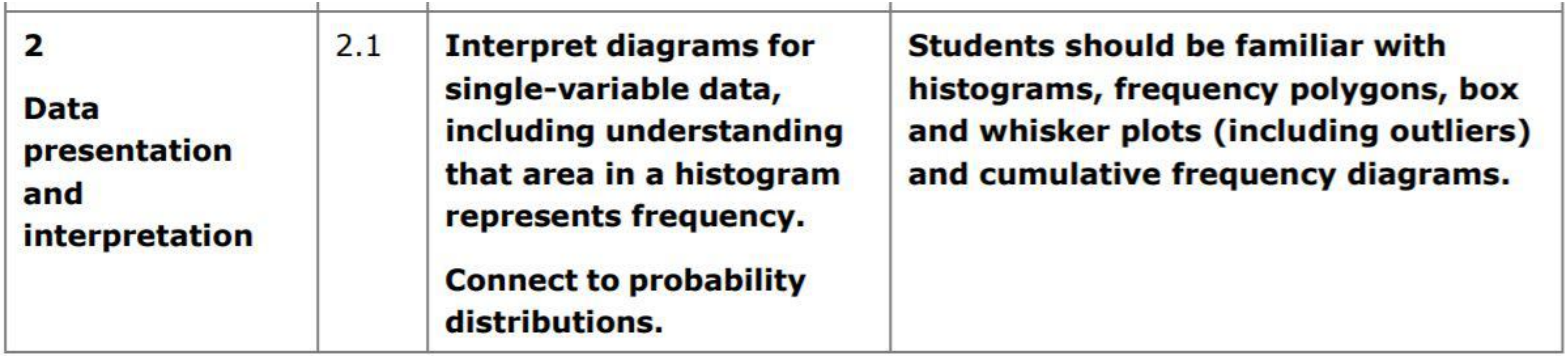
2.02 Data Presentation and Interpretation		
2.02a	Single variable data	a) Be able to interpret tables and diagrams for single-variable data. *e.g. vertical line charts, dot plots, bar charts, stem-and-leaf diagrams, box-and-whisker plots, cumulative frequency diagrams and histograms (with either equal or unequal class intervals). Includes non-standard representations.*
2.02b		b) Understand that area in a histogram represents frequency. *Includes the link between histograms and probability distributions.* *Includes understanding, in context, the advantages and disadvantages of different statistical diagrams.*

Specification Points – OCR MEI



STATISTICS: DATA PRESENTATION AND INTERPRETATION (1)				
Data presentation for single variable	MD1	Be able to recognise and work with categorical, discrete, continuous and ranked data. Be able to interpret standard diagrams for grouped and ungrouped single-variable data.	Includes knowing this vocabulary and deciding what data presentation methods are appropriate: bar chart, dot plot, histogram, vertical line chart, pie chart, stem-and-leaf diagram, box-and-whisker diagram (box plot), frequency chart. Learners may be asked to add to diagrams in examinations in order to interpret data.	A frequency chart resembles a histogram with equal width bars but its vertical axis is frequency. A dot plot is similar to a bar chart but with stacks of dots in lines to represent frequency.
	D2	Understand that the area of each bar in a histogram is proportional to frequency. Be able to calculate proportions from a histogram and understand them in terms of estimated probabilities.	Includes use of area scale and calculation of frequency from frequency density.	
	D3	Be able to interpret a cumulative frequency diagram.		
	D4	Be able to describe frequency distributions.	Symmetrical, unimodal, bimodal, skewed (positively and negatively).	

Specification Points - Edexcel



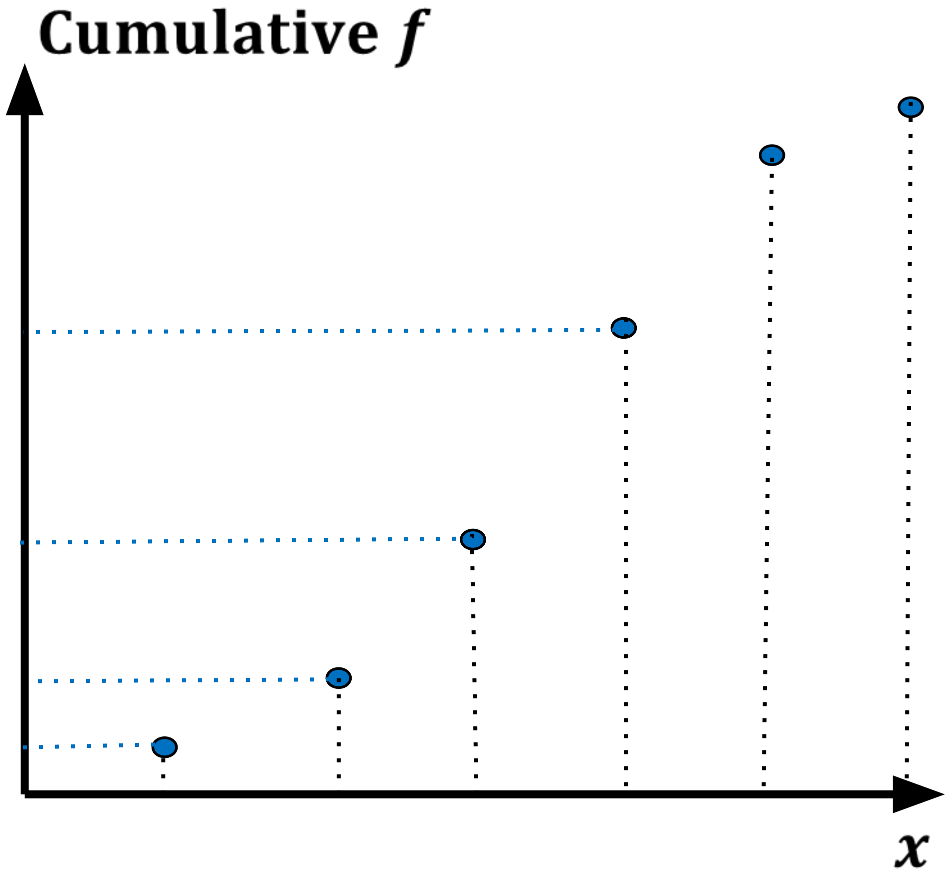
2 **Data presentation and interpretation**	2.1	**Interpret diagrams for single-variable data, including understanding that area in a histogram represents frequency.** **Connect to probability distributions.**	**Students should be familiar with histograms, frequency polygons, box and whisker plots (including outliers) and cumulative frequency diagrams.**

Cumulative Frequency Graphs

Cumulative frequency graphs plot **upper class boundary** and **cumulative frequency**.

- **Cumulative frequency** is the **sum** of **frequency** from all **previous classes**.

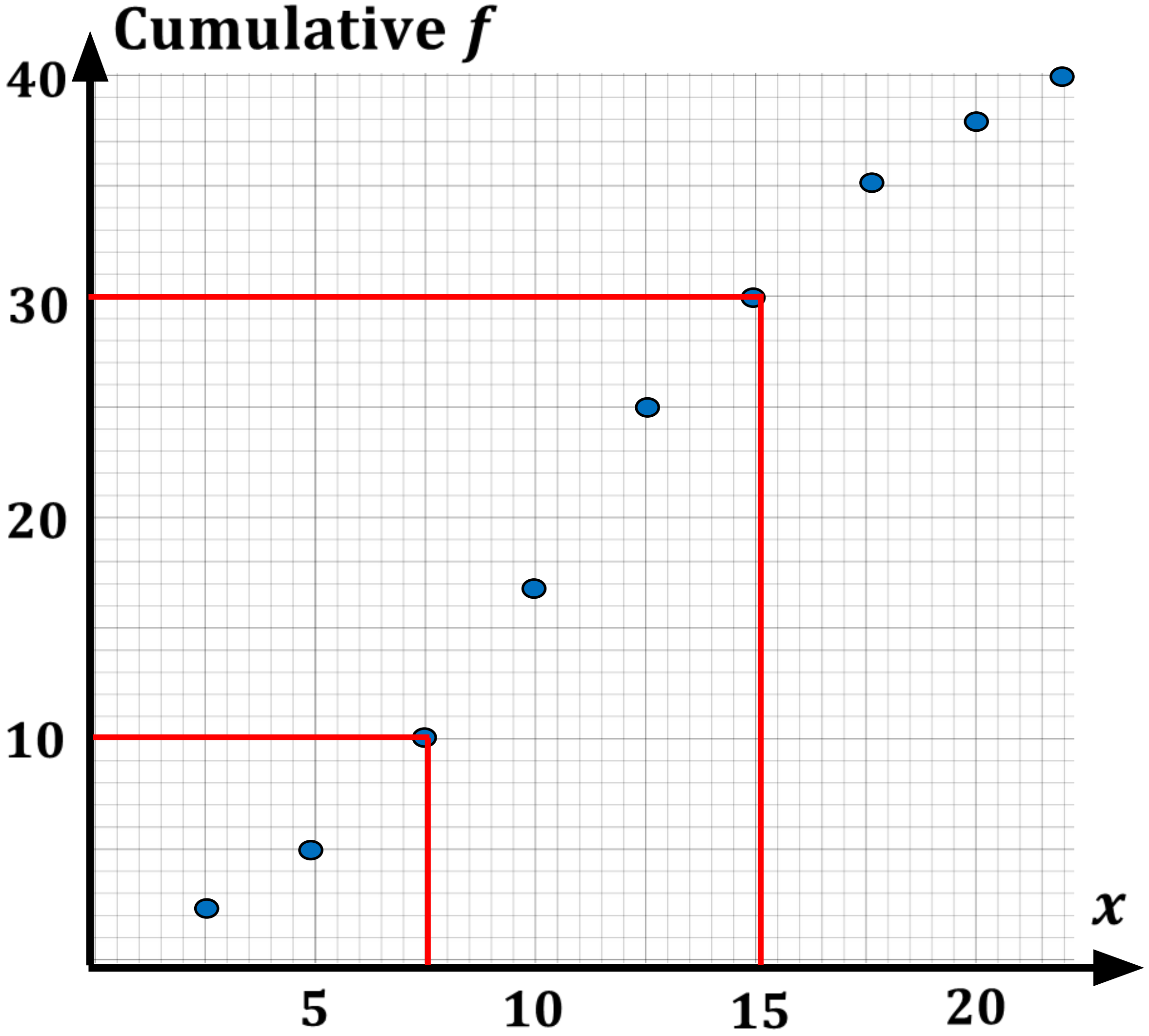
x	f	Cumulative f
$0 \leq x \leq x_1$	f_1	
$x_1 < x \leq x_2$	f_2	
$x_2 < x \leq x_3$	f_3	
$x_3 < x \leq x_4$	f_4	
...	...	



Cumulative Frequency Graphs

We can **find** the $I.Q.R$ of a **set** of **continuous grouped data** from a **cumulative frequency graph**.

- **Plot horizontal lines** at $\frac{n}{4}$ and $\frac{3n}{4}$ to find where they **intersect** the **curve of best fit**.



Exemplar Exam Question

Use graph to find useful information

This way of recording data **results** in a **cumulative frequency graph**

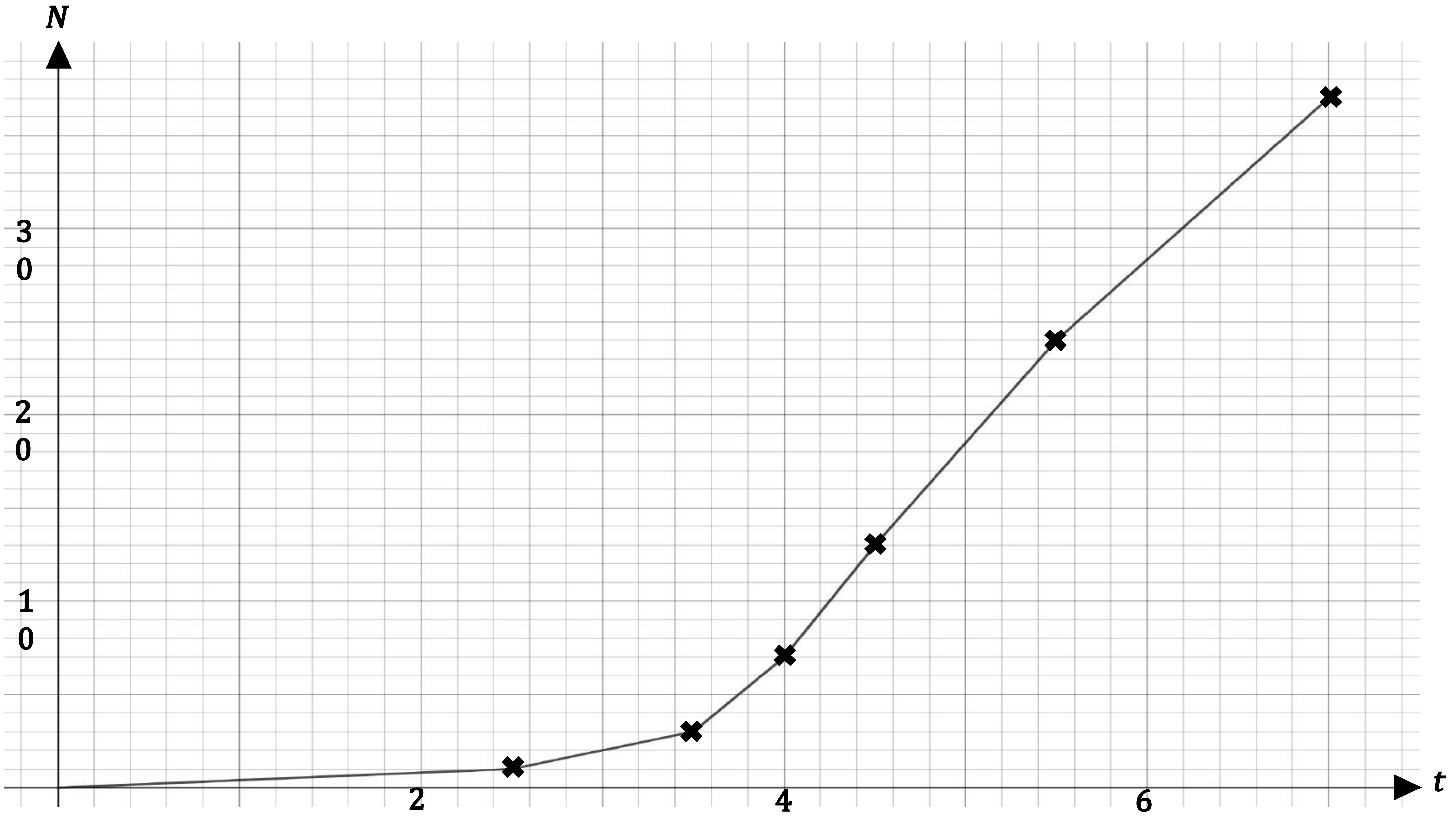
1) A group of 37 office workers take part in a charity marathon. After certain periods of time the number of workers who have finished the race, $\boldsymbol{N}$, is recorded. The data of $\boldsymbol{N}$ against time $\boldsymbol{t}$ in hours is shown on the following graph.

From the graph calculate the Interquartile Range of the times it took the workers to finish the marathon. Give your answer in hours to 1 d.p.

Check **units** and **rounding**

[2 Marks]

2 mark question, fairly simple calculation


N
3
0
2
0
1
0
t
2
4
6

Exemplar Exam Question Answer

Calculate quartiles of N

Start by calculating the quartiles of the cumulative value N

The maximum value of N is 37

First quartile of N is:

$$\frac{N_{max}}{4} = \frac{37}{4}$$

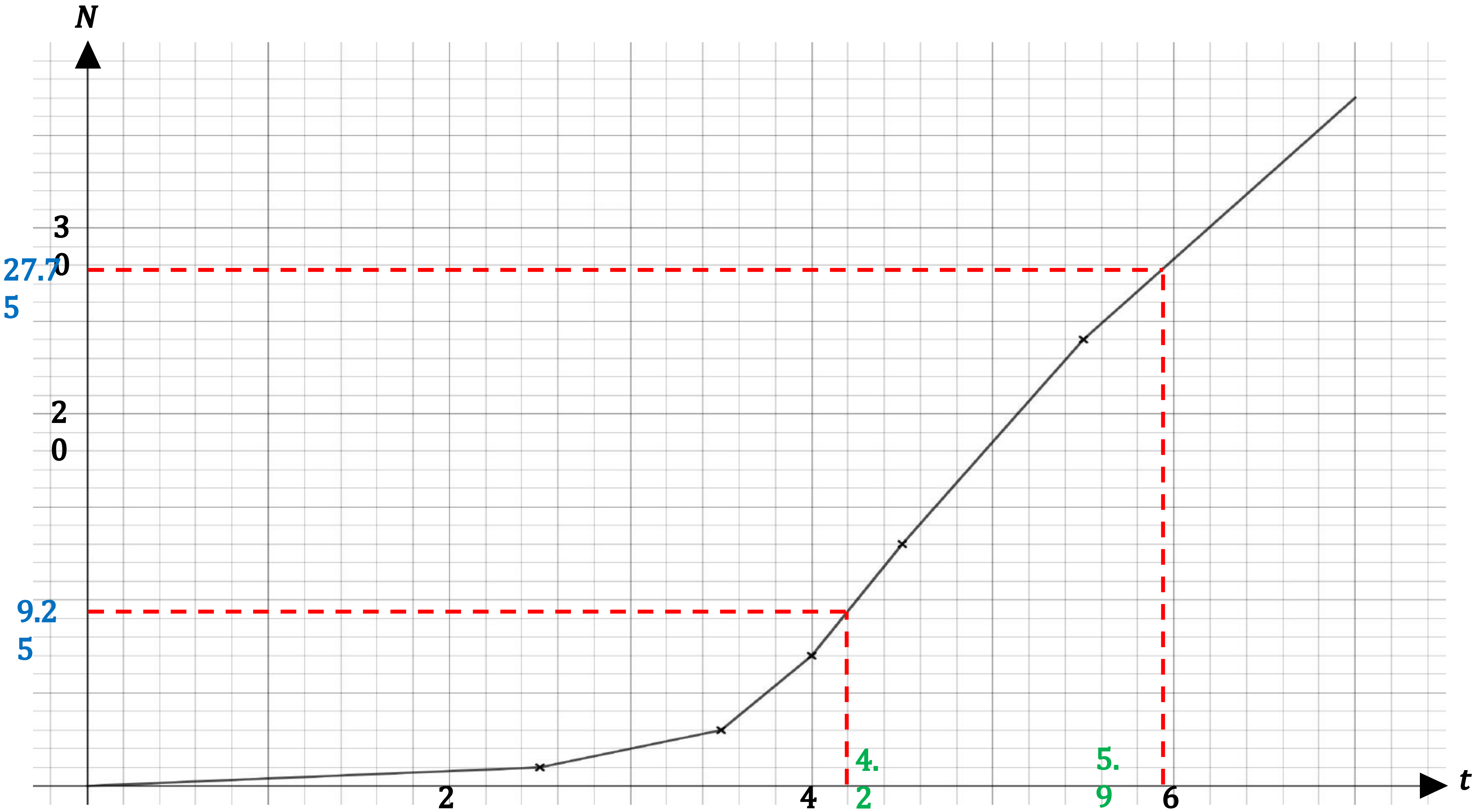
$$\frac{N_{max}}{4} = 9.25$$

Third quartile of N is:

$$\frac{3N_{max}}{4} = \frac{3\times 37}{4}$$

$$\frac{3N_{max}}{4} = 27.75$$

[1 Mark]


N
3
0
27.7
5
2
0
9.2
5
4.
2
5.
9
2
4
6
t

Exemplar Exam Question Answer

So Inter Quartile Range is

$$IQR = 5.9 - 4.2$$

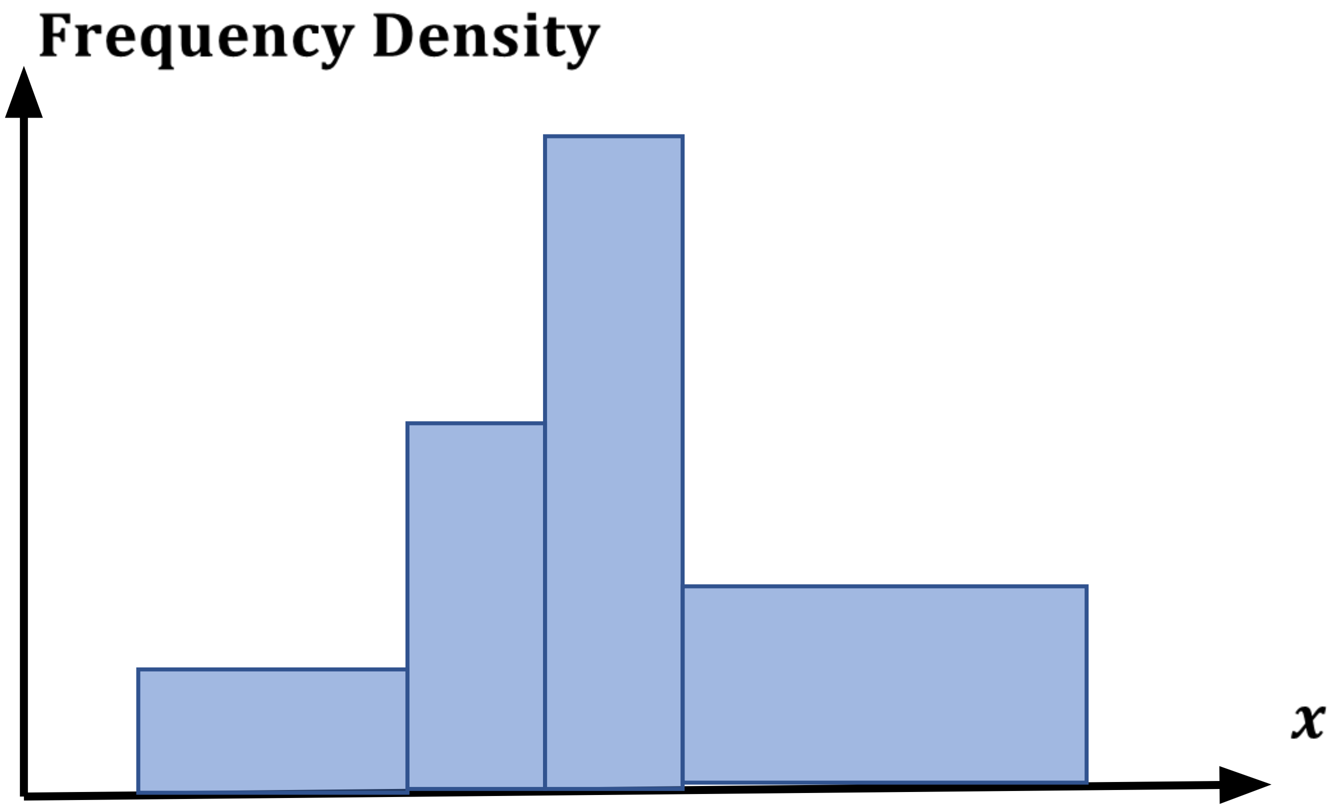
$$IQR = 1.7 \text{ hours}$$

[1 Mark]

Histograms

Histograms are used to show **grouped continuous data** with **unequal class intervals**.

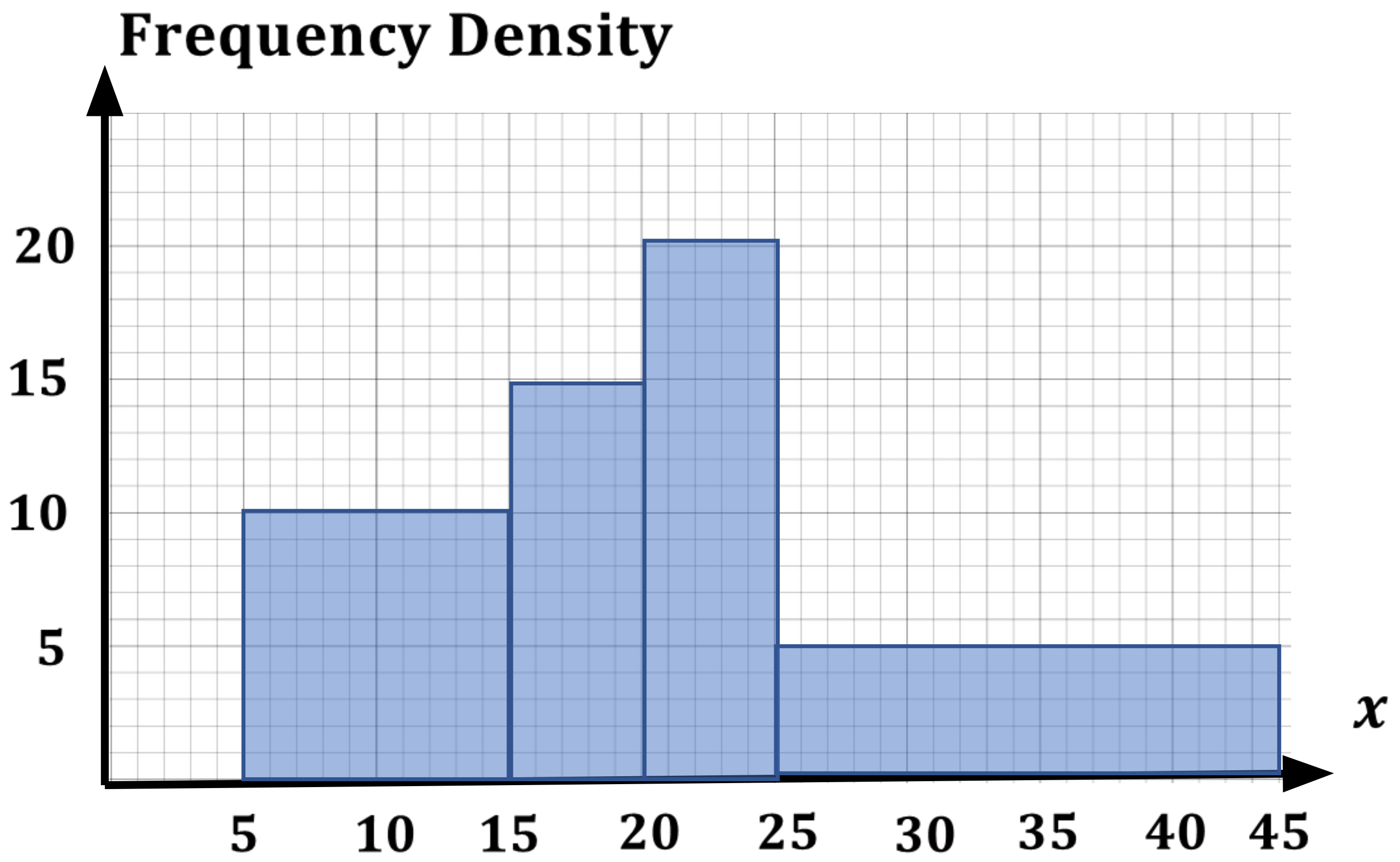
- A **histogram** plots **frequency density** against x.
- In a **histogram** the **area** of each **bar** is **proportional** to the **frequency** in each **class**.
- There are **no gaps** between the **bars**.



Histograms

Data can be used to **predict** the **outcome** of a **random selection** from a **group**.

- The **ratio** of the **areas** of each **bar** can be used to **calculate** the **probability** of a **particular result** of a **selection**.
- The **total area** of all **bars** gives **probability** $= 1$.
- We **assume** that **data** is **evenly distributed** within **each class**.



Calculating probabilities from **histograms**

Exemplar Exam Question

Referring between data sets

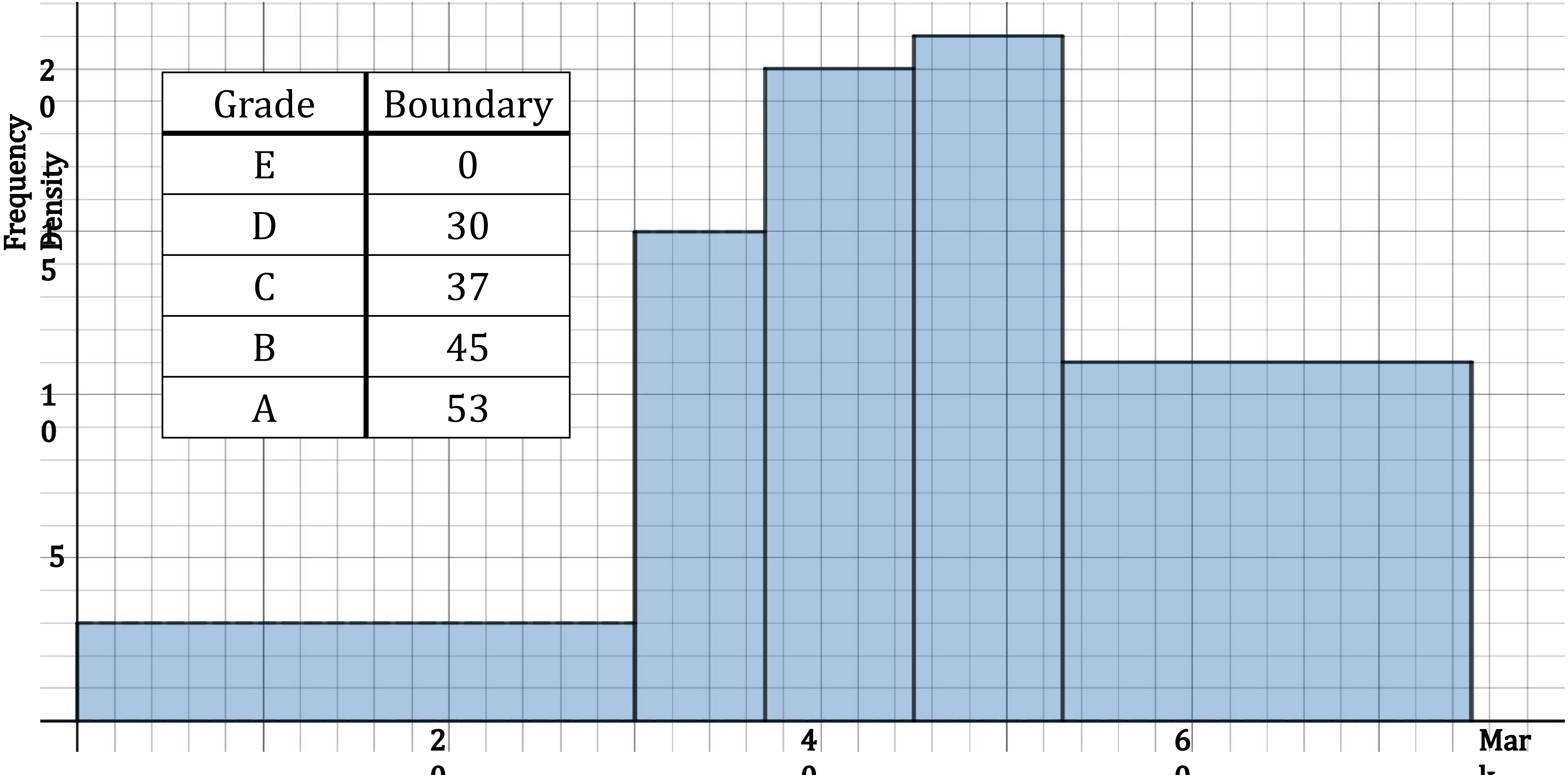
1) A school records the grades achieved by students in a recent exam, where the maximum mark is 75, to produce the following histogram. The table given shows the grade boundaries for the paper.

   A student's paper is selected at random to evaluate the difficulty of the paper. Calculate the probability that the selected student got a B or higher in the exam.

[2 marks]

May need to calculate **area of multiple sections**

2 marks, **2 steps** to calculation

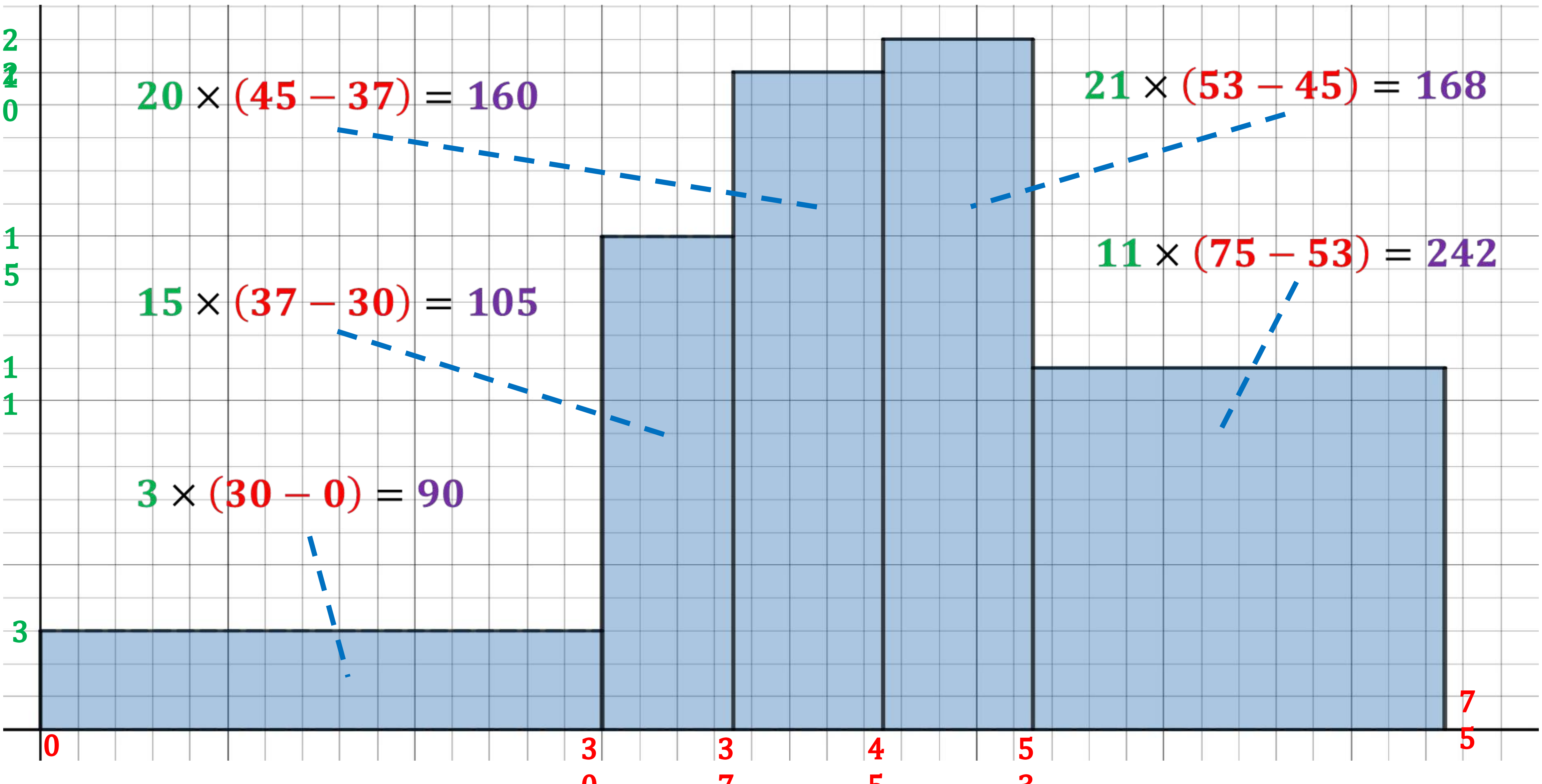


Grade	Boundary
E	0
D	30
C	37
B	45
A	53

Exemplar Exam Question Answer

Calculate total area of histogram

To find total area of the histogram, find area of each section of graph and sum.



$20 \times (45 - 37) = 160$

$21 \times (53 - 45) = 168$

$15 \times (37 - 30) = 105$

$11 \times (75 - 53) = 242$

$3 \times (30 - 0) = 90$

Exemplar Exam Question Answer

Calculate total area of graph

So the total area is

$$90 + 105 + 160 + 168 + 242 = 765$$

[1 Mark]

Exemplar Exam Question Answer

Calculate probability of B or higher

Grade	Boundary
E	0
D	30
C	37
B	45
A	53

Probability of segment on histogram given by

$$\mathbf{P(segment)} = \frac{\textit{Area of Segment}}{\textit{Total Area of Graph}}$$

From reference table, can deduce that a B grade or higher is awarded for marks above 45.

Using area of those segments of histogram and calculated total area

$$\mathbf{P(B)} = \frac{168+242}{765}$$

$$\mathbf{P(B) = 0.5359}$$

[1 Mark]

MINI MOCK PAPER

Exam Question

1. A computer has a simple system for generating random integers between 0 and 9. It multiplies together two large integers $\boldsymbol{X}$ and $\boldsymbol{Y}$ and then each time the system is used, it reads the next digit along in their product $\boldsymbol{Z} = \boldsymbol{XY}$.

(i) Calculate the mean and interquartile range of the digits of $\boldsymbol{Z}$ when $\boldsymbol{Z} = \mathbf{121932631135269}$.

[2 Marks]

An issue with this method is that the first few digits of $\boldsymbol{Z}$ are much more likely to have a low value than a high value.

(ii) Suggest how omitting the first few digits from $\boldsymbol{Z}$ this would affect the mean and of the digits, and discus why this would improve the effectiveness of the method

[2 Marks]

Exam Question Answer

$$Z = 121932631135269$$

(i) Input digits of Z to calculator

1	1
2	2
3	1
4	9
5	3
6	2

or

		Freq
1	0	0
2	1	4
3	2	3
4	3	3
5	4	0
6	5	1

Exam Question Answer

$$Z = 121932631135269$$

Calculate Mean and Interquartile Range

From calculator:

$$\text{Mean} = 3.6$$

[1 Mark]

$$\text{IQR} = Q_3 - Q_1 = 6 - 1 = 0$$

[1 Mark]

Exam Question Answer

$$Z = 121932631135269$$

(ii) Predict affect on Mean

Removing low digits will likely raise mean

[1 Mark]

Discuss desired Mean

Want digits to be evenly spread

$\Rightarrow$ desired mean is 5

Removing digits raises current low mean towards this value

[1 Mark]

Exam Question

2. A football club breaks down their 500 season ticket holders by age, A, to produce the following histogram. Using the data from this:

(i) Complete the table of the number of season ticket holders, N, per age range.

[3 Marks]

(ii) Plot a cumulative frequency graph of ages against number of ticket holders on the axes provided and use it to estimate the interquartile range of ages.

[4 Marks]

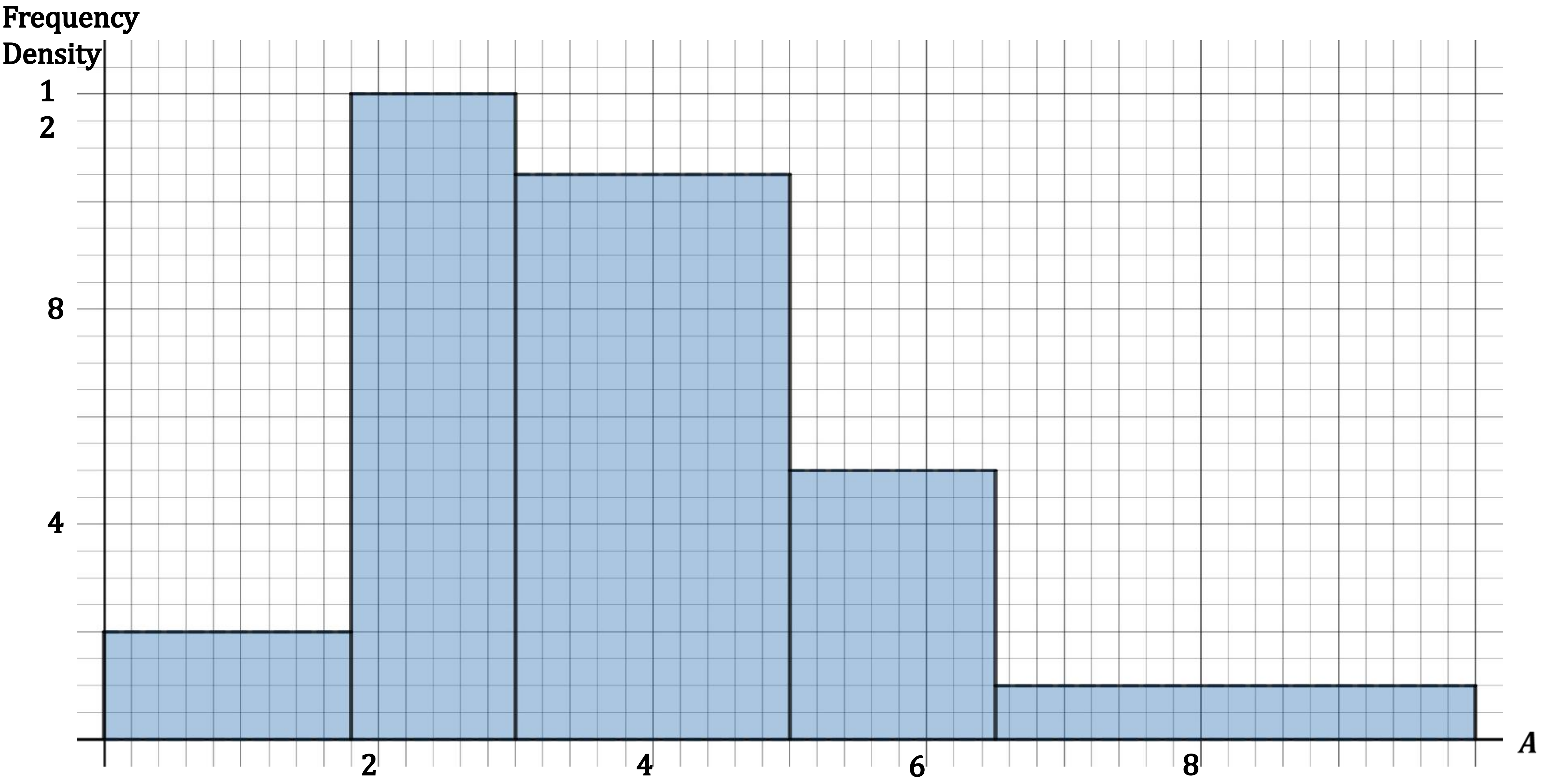
Exam Question

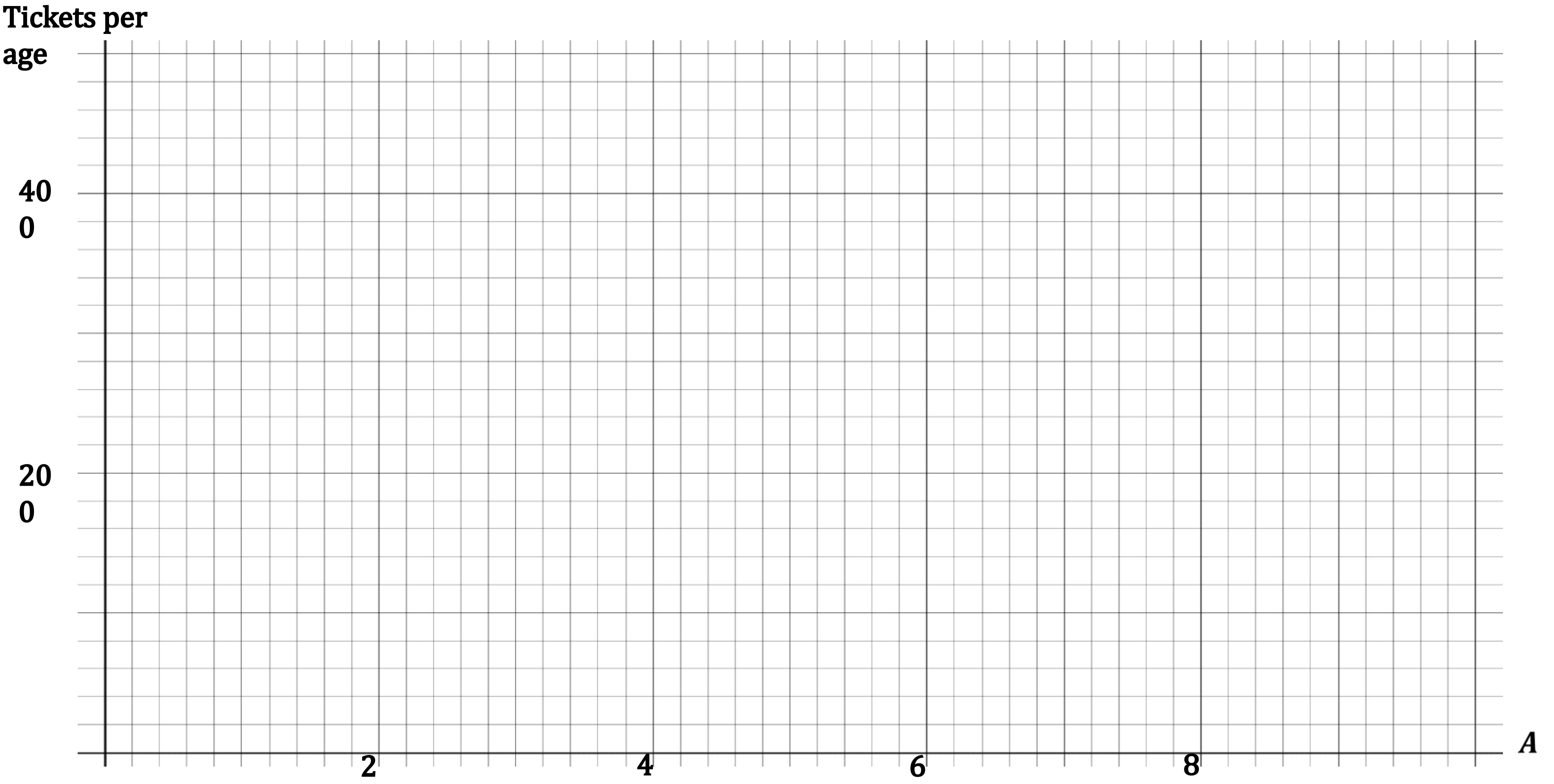
2. (cont.)

The club wants to conduct a survey with season ticket holders about refurbishments to their ground, and plans to take a stratified sample of 75 people by age range.

(iii) Calculate the number of people under 30 in the sample

[1 Mark]

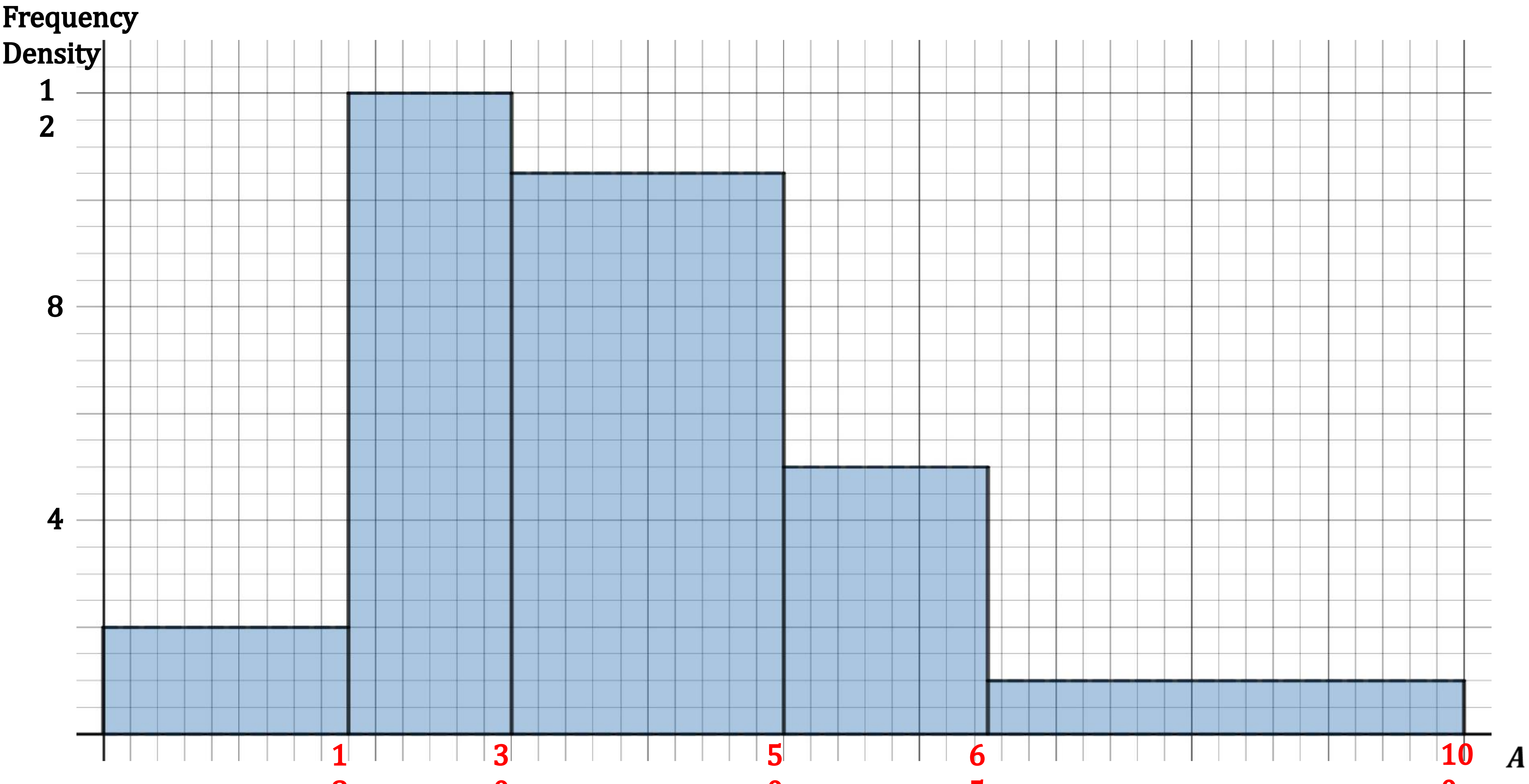

Frequency
Density
1
2
8
4
2
4
6
8
A


Tickets per age
400
200
2
4
6
8
A

Exam Question Answer

(i) Recognise information required to complete table

Age ranges can be found by reading off where each segment in the histogram begins and ends


Frequency Density
1
2
8
4
1
3
5
6
10
A

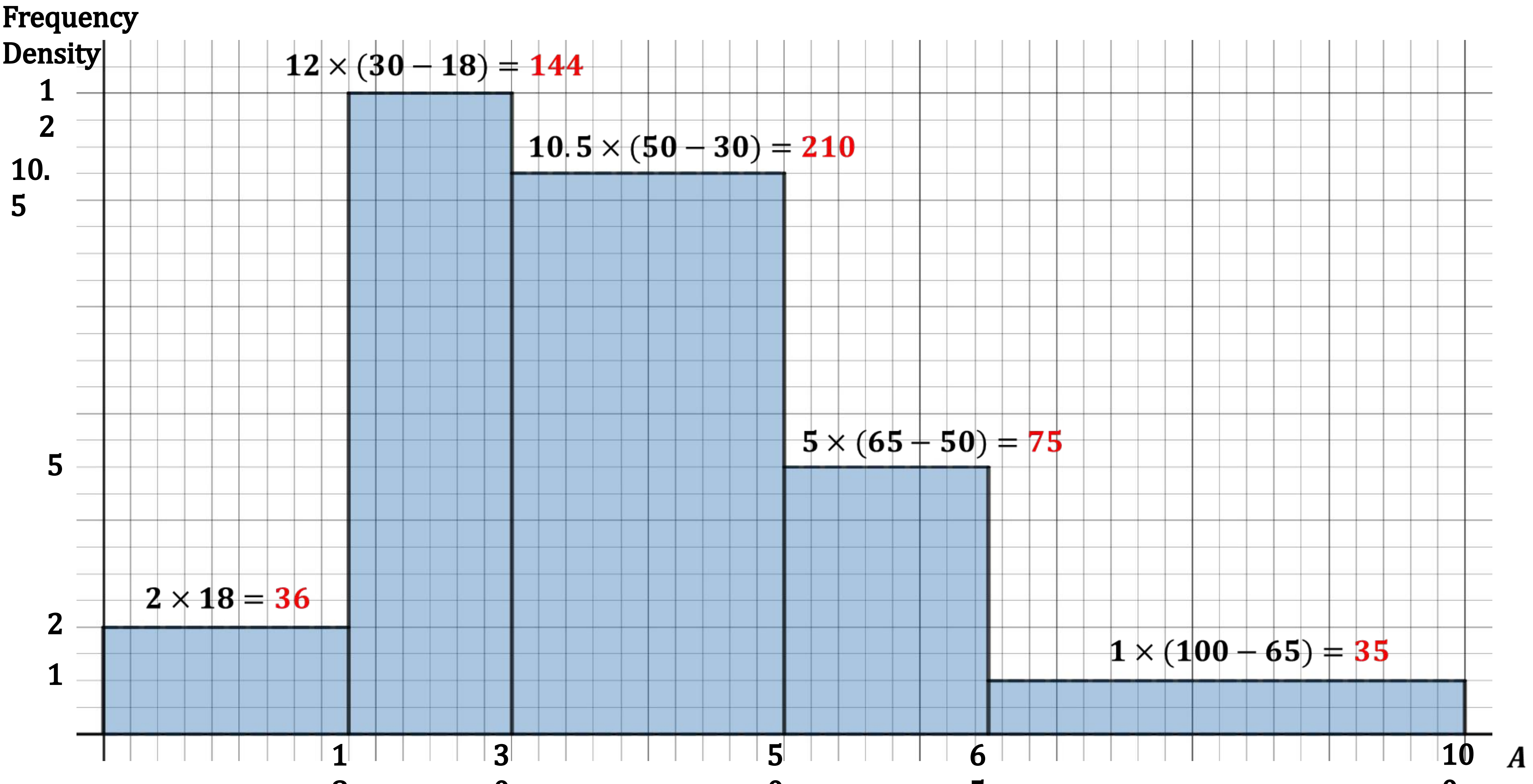
[1 Mark]

Exam Question Answer

Recognise information required to complete table

Number of ticket holders in each range is proportional to area of segment of histogram

Can deduce constant of proportionality using the fact that the total area of the graph represents 500 ticket holders


Frequency Density
1
2
10.
5
5
2
1
12 × (30 − 18) = 144
10.5 × (50 − 30) = 210
5 × (65 − 50) = 75
2 × 18 = 36
1 × (100 − 65) = 35
1
3
5
6
10
A

Exam Question Answer

Recognise information required to complete table

$36 + 144 + 210 + 75 + 35 = 500$

So the number of ticket holders is each age range is equal to the area of the respective segment in the histogram

[1 Mark]

	36
	144
	210
	75
	35

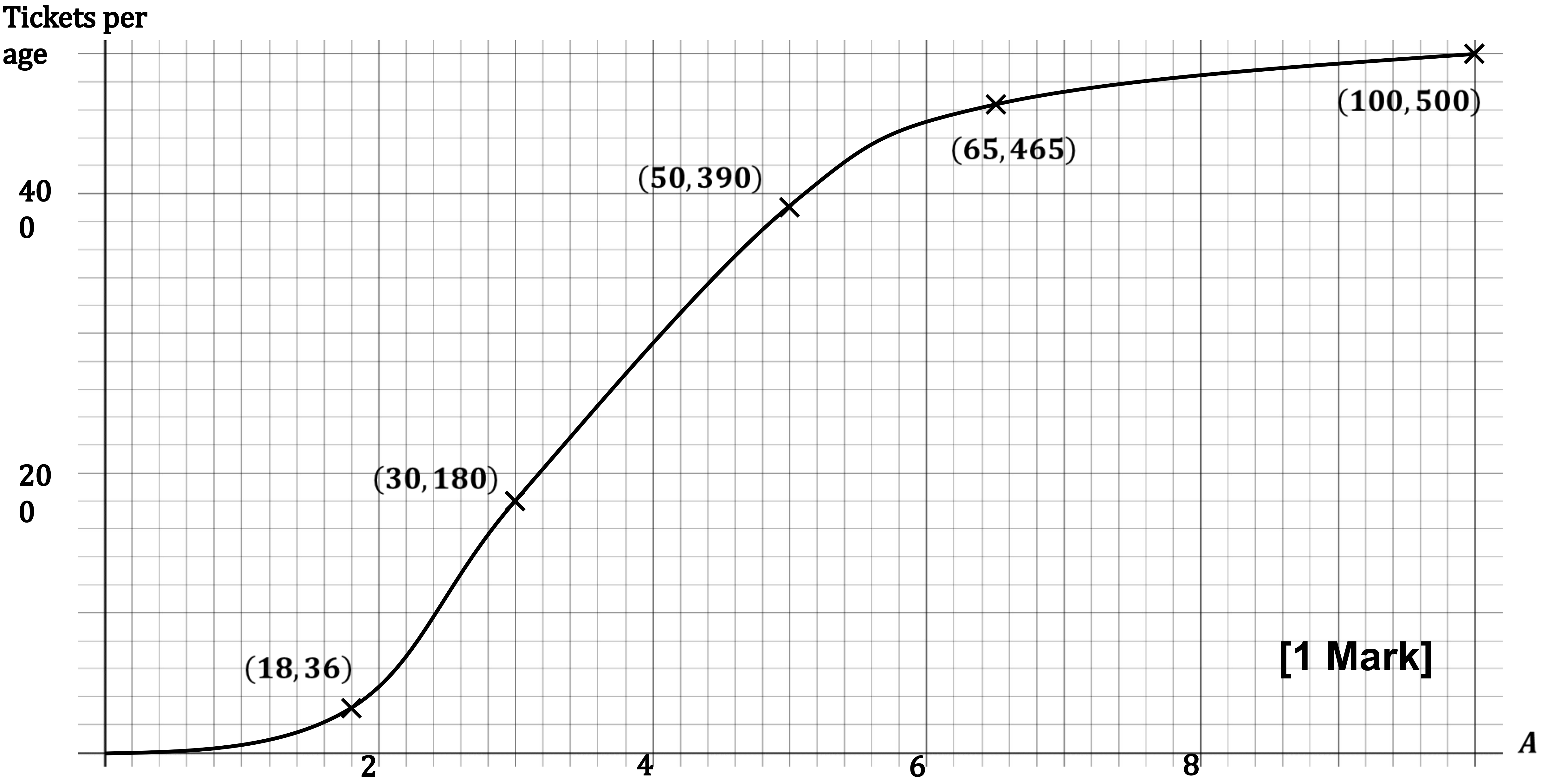
[1 Mark]

Exam Question Answer

(ii) Create cumulative frequency table

		Cumulative Frequency
	36	
	144	
	210	
	75	
	35	

This can now be used to plot graph **[1 Mark]**



[1 Mark]

Exam Question Answer

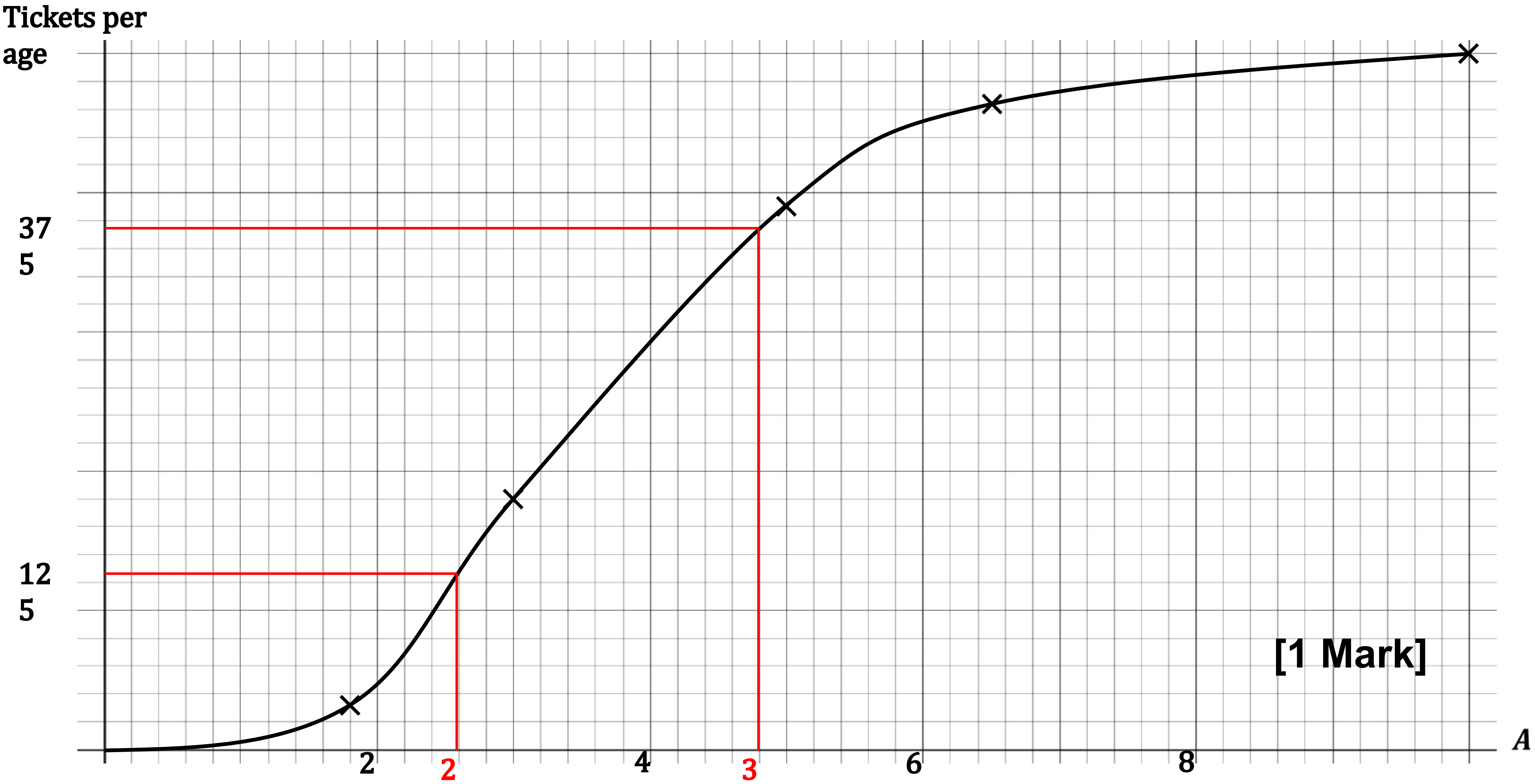
Determine N values corresponding to 1st and 3rd quartiles

Total number of ticket holders is 500

$$\frac{N}{4} = \frac{500}{4} = 125$$

$$\frac{3N}{4} = \frac{3\times500}{4} = 375$$

Find corresponding A values from graph



[1 Mark]

Exam Question Answer

$Q_1 = 26, \; Q_3 = 30$

Calculate interquartile range

$$IQR = Q_3 - Q_1 = 30 - 26$$

$$= 4 \qquad \text{(allow values between 2 and 6)}$$

[1 Mark]

Exam Question Answer

(iii) Calculate number of people in stratum

Stratum consists of age ranges $0 \le A < 18$ and $18 \le A < 30$

From table, number of people in stratum is $36 + 144 = 180$

Calculate number of people in sample

$$\text{No. in sample} = \frac{\text{No. in stratum}}{\text{No. in population}} \times \text{Sample Size}$$

$$= \frac{180}{500} \times 75 = 27$$

[1 Mark]