



## 4.1 Statistics Toolkit

### Contents

- ★ 4.1.1 Sampling
- ★ 4.1.2 Data Collection
- ✤ 4.1.3 Statistical Measures
- ★ 4.1.4 Frequency Tables
- ★ 4.1.5 Linear Transformations of Data
- ★ 4.1.6 Outliers
- ✤ 4.1.7 Univariate Data
- ★ 4.1.8 Interpreting Data



### 4.1.1 Sampling

### **Types of Data**

### What are the different types of data?

- Qualitative data is data that is usually given in words not numbers to describe something
  - For example: the colour of a teacher's car
- Quantitative data is data that is given using numbers which counts or measures something
   For example: the number of pets that a student has
- Discrete data is quantitative data that needs to be counted
  - Discrete data can only take **specific values** from a set of (usually finite) values
  - For example: the number of times a coin is flipped until a 'tails' is obtained
- Continuous data is quantitative data that needs to be measured
  - Continuous data can take any value within a range of infinite values
  - For example: the height of a student
- Age can be discrete or continuous depending on the context or how it is defined
  - If you mean how many years old a person is then this is discrete
  - If you mean how long a person has been alive then this is continuous

#### What is the difference between a population and a sample?

- The **population** refers to the **whole set** of things which you are interested in
  - For example: if a vet wanted to know how long a typical French bulldog slept for in a day then the population would be all the French bulldogs in the world
- A sample refers to a subset of the population which is used to collect data from
  - For example: the vet might take a sample of French bulldogs from different cities and record how long they sleep in a day
- A sampling frame is a list of all members of the population
  - For example: a list of employees' names within a company
- Using a **sample instead of a population**:
  - Is quicker and cheaper
  - Leads to less data needing to be analysed
  - Might not fully represent the population
  - Might introduce bias



### Sampling Techniques

#### What is a random sample and a biased sample?

- A random sample is where every member of the population has an equal chance of being included in the sample
- A biased sample is one from which misleading conclusions could be drawn about the population
  - Random sampling is an attempt to minimise bias

#### What sampling techniques do I need to know?

#### Simple random sampling

- Simple random sampling is where every group of members from the population has an equal probability of being selected for the sample
- To carry this out you would...
  - uniquely number every member of a population
  - randomly select n different numbers using a random number generator or a form of lottery (where numbers are selected randomly)
- Effectiveness:
  - Useful when you have a small population or want a small sample (such as children in a class)
  - It can be time-consuming if the sample or population is large
  - This can not be used if it is not possible to number or list all the members of the population (such as fish in a lake)

#### Systematic sampling

- Systematic sampling is where a sample is formed by choosing members of a population at regular intervals using a list
- To carry this out you would...

• calculate the size of the interval  $k = \frac{\text{size of population } (N)}{\text{size of sample } (n)}$ 

- choose a random starting point between 1 and k
- select every kth member after the first one
- Effectiveness:
  - Useful when there is a natural order (such as a list of names or a conveyor belt of items)
  - Quick and easy to use
  - This can not be used if it is not possible to number or list all the members of the population (such as penguins in Antarctica)

#### Stratified sampling

 Stratified sampling is where the population is divided into disjoint groups and then a random sample is taken from each group



- The proportion of a group that is sampled is equal to the proportion of the population that belong to that group
- To carry this out you would...
  - Calculate the number of members sampled from each stratum
    - size of sample (*n*)

# size of population (N) × number of members in the group

- Take a random sample from each group
- Effectiveness:
  - Useful when there are very different groups of members within a population
  - The sample will be representative of the population structure
  - The members selected from each stratum are chosen randomly
  - This can not be used if the population can not be split into groups or if the groups overlap

#### Quota sampling

- Quota sampling is where the population is split into groups (like stratified sampling) and members of the population are selected until each quota is filled
- To carry this out you would...
  - Calculate how many people you need from each group
  - Select members from each group until that guota is filled
    - The members do not have to be selected randomly
- Effectiveness:
  - Useful when collecting data by asking people who walk past you in a public place or when a sampling frame is not available
  - This can introduce bias as some members of the population might choose not to be included in the sample

#### Convenience sampling

- Convenience sampling is where a sample is formed using available members of the population who fit the criteria
- To carry this out you would...
  - Select members that are easiest to reach
- Effectiveness:
  - Useful when a list of the population is not possible
  - This is unlikely to be representative of the population structure
  - This is likely to produce biased results

#### What are the main criticisms of sampling techniques?

- Most sampling techniques can be improved by taking a larger sample
- Sampling can introduce bias so you want to minimise the bias within a sample • To minimise bias the sample should be as close to random as possible
- A sample only gives information about those members
- Different samples may lead to different conclusions about the population



### Worked example

b)

Mike is a biologist studying mice in an open enclosure. He has access to approximately 540 field mice and 260 harvest mice. Mike wants to sample 10 mice and he wants the proportions of the two types of mice in his sample to reflect their respective proportions of the population.

a) Calculate the number of field mice and harvest mice that Mike should include in his sample.

lotal number of mice 540 + 260 = 800 - Fraction of field mice Field mice 540 × 10 = 6.75 Sample size - Fraction of harvest mice  $\frac{260}{800} \times 10 = 3.25$ Harvest mice Include 7 field mice and 3 harvest mice Given that Mike does not have a list of all mice in the enclosure, state the name of this sampling method. No list of population so can not be a random sample Quota sampling

c) Suggest one way in which Mike could improve his sampling method.

Mark could improve his sampling method by increasing his sample size



### **Reliability of Data**

### How can I decide if data is reliable?

- Data from a sample is reliable if similar results would be obtained from a different sample from the same population
- The sample should be **representative** of the population
- The sample should be **big enough** 
  - Sampling a small proportion of a population is unlikely to be reliable

### What can cause data to be unreliable?

- If the sample is **biased** 
  - It is not random
- If errors are made when collecting data
  - Numbers could be recorded incorrectly, duplicated or missed out
- If the person collecting the data favours some members over others
  - They might seek out members who will lead to a desired outcome
  - They might exclude members if they would cause the sample to oppose the desired outcome
- If a significant proportion of **data is missing** 
  - Some data may be unavailable
  - Some members might decide not to be part of the sample
    - This will mean the results are not necessarily representative of the population



### 4.1.2 Data Collection

### Methods of Data Collection

### How do I choose variables to investigate?

- Keep the number of variables you investigate to a minimum
  - Too many variables at once can be **overwhelming**
  - It can be time-consuming to process unnecessary data
- You should choose variables that are linked to what you are investigating
  - If you are investigating the ability of adults to solve puzzles you might use the time it takes them as a variable
  - Consider which variables are likely to have an effect on what you are investigating
    - An adult's reading speed will affect their time to solve a puzzle
    - An adult's height is unlikely to affect their time to solve a puzzle

#### What makes a good survey?

- A survey is a method of collecting data
- Consider whether the survey needs to be **in-person** 
  - A person might be less likely to answer questions truthfully in person
  - You can quickly survey more people remotely or electronically
    - Such as postal surveys, phone surveys, internet surveys
- Consider whether the interviewer could unintentionally influence participants' responses
  - If a headteacher is asking students whether they enjoy school then they are more likely to say yes as they think that is what the headteacher wants to hear
  - This will introduce bias

#### What makes a good questionnaire?

- A questionnaire is a list of questions
- The questions should be **unbiased** 
  - Questions should **not be leading** 
    - For example: "You enjoy school, don't you?"
  - If options are given for the participant to choose from then they should cover all possible responses
- The questions should not be personal
  - This means you should not ask for **unnecessary personal information** 
    - Such as date of birth, address, etc
  - The questions should not reflect your personal opinions
    - For example: "Do you enjoy watching the boring news on TV?"
  - People can find it difficult to rate personal feeling/qualities
  - For example: "How smart do you think you are?"
- Questions can be **structured or unstructured**



- Structured questions usually ask the participants to choose from options, give a rating or rank options
  - These can be quick to analyse
  - The answer choices should **be consistent** where appropriate
- Unstructured questions let the participants to express their views in their own words
  - These tend to be more **open-ended** questions
  - These can take longer to analyse but can give more in-depth views
- Questions should be **precise and unambiguous**

- They should be phrased in a way in which the **participants understand exactly what you mean**
- For example: "Do you study French or Spanish at school" is not precise
  - Some people might reply with "Yes" or "No"
  - Some people might reply with "French" or "Spanish"



### **Reliability & Validity**

### What is reliability & validity of a data collection method?

- Reliability measures how consistent a process is at measuring a variable
  - A process is reliable if you would get the same results by repeating the process with the same sample using the same conditions
- Validity measures how accurate a process is at measuring a variable
  - A process is valid if it is accurately measuring the variable you want it to measure
- If your process is found not to be reliable or valid then:
  - Adjust the data collection process
  - Change the sampling technique
  - Use a larger sample

#### What are tests to check reliability?

- Test-retest
  - This is where you use a data collection process with a sample and then repeat the **same process** with the **same sample at a later time** 
    - The results should show positive correlation if the process is reliable
    - The results might not perfectly match due to external factors during the gap between the data collection
    - Once the sample has been through the process once they will be familiar so this could lead to different results from the second process
- Parallel forms
  - This is where you give the same sample a second set of questions (or second set of experiments) which are similar to the first set
    - The results should show positive correlation if the process is reliable
    - It can be difficult to make the two processes similar to each other

#### What are tests to check validity?

- Content-related validity checks
  - This is where you check how well the process measures all aspects of the variable
    - If the process is valid then it should cover all aspects of the variable
    - These checks require knowledge of the variable so experts are often used
    - An example of a process that is valid: A teacher wants to assess how well students understand calculus so they set questions covering differentiation, integration and applications
    - An example of a process that is not valid: A restaurant manager wants to assess how good a chef is at cooking steaks so asks the chef to make 10 medium steaks

#### Criterion-related validity checks

- This is where you check how well one variable predicts the outcome for another variable (called the criterion variable)
  - If the process is valid then the variable should be a good predictor



- An example of a process that is valid:
   Results from a mock exam being used to predict the results in the actual exam
- An example of a process that is not valid: Results from measuring the heights of meerkats being used to predict the heights of squirrels

### Worked example

Tomas is a dog trainer. Before he agrees to train a dog he assesses the dog's obedience. To do this, he first visits the dog, asks it to perform 10 basic commands and records how many the dog successfully carries out. Two days later, Tomas visits the dog a second time and asks it to do the same 10 commands. Tomas assesses 8 dogs using this process and the table below shows the number of commands performed successfully by each dog on each visit.

Dog		2	3	4	5	6	7	8
First visit	3	5	2	3	6	2	0	5
Second visit	3	5	2	4	5	2	1	5

a) State the reliability test that Tomas is using.

Tomas is using exactly the same process with the same sample Test-retest

b) Comment on the reliability of Tomas' process.

The number of commands that each dog successfully performed on the second visit was either the same as the first visit or very similar. Therefore the process is reliable.



### 4.1.3 Statistical Measures

### Mean, Mode, Median

### What are the mean, mode and median?

- Mean, median and mode are measures of central tendency
  - They describe where the centre of the data is
- They are all types of **averages**
- In statistics it is important to be specific about which average you are referring to
- The units for the mean, mode and median are the same as the units for the data

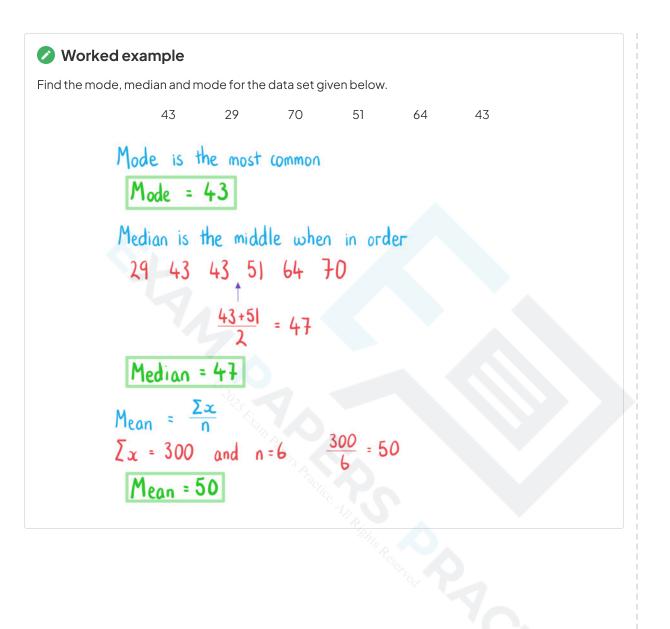
### How are the mean, mode, and median calculated for ungrouped data?

- The mode is the value that occurs most often in a data set
  - It is possible for there to be more than one mode
  - It is possible for there to be **no mode** 
    - In this case **do not** say the mode is zero
  - The **median** is the **middle** value when the data is in **order of size**
  - If there are two values in the middle then the median is the **midpoint** of the two values
- The mean is the sum of all the values divided by the number of values

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

- Where  $\sum_{i=1}^{n} X_i = X_1 + X_2 + \dots + X_n$  is the sum of the *n* pieces of data
- The mean can be represented by the symbol  $\mu$
- Your GDC can calculate these statistical measures if you input the data using the statistics mode







### **Quartiles & Range**

### What are quartiles?

- Quartiles are measures of location
- Quartiles divide a population or data set into **four equal sections** 
  - The lower quartile, Q<sub>1</sub> splits the lowest 25% from the highest 75%
  - The **median**, **Q**<sub>2</sub> splits the lowest 50% from the highest 50%
  - The **upper quartile**, **Q**<sub>3</sub> splits the lowest 75% from the highest 25%
- There are different methods for finding quartiles
  - Values obtained by hand and using technology may differ
- You will be expected to use your GDC to calculate the quartiles

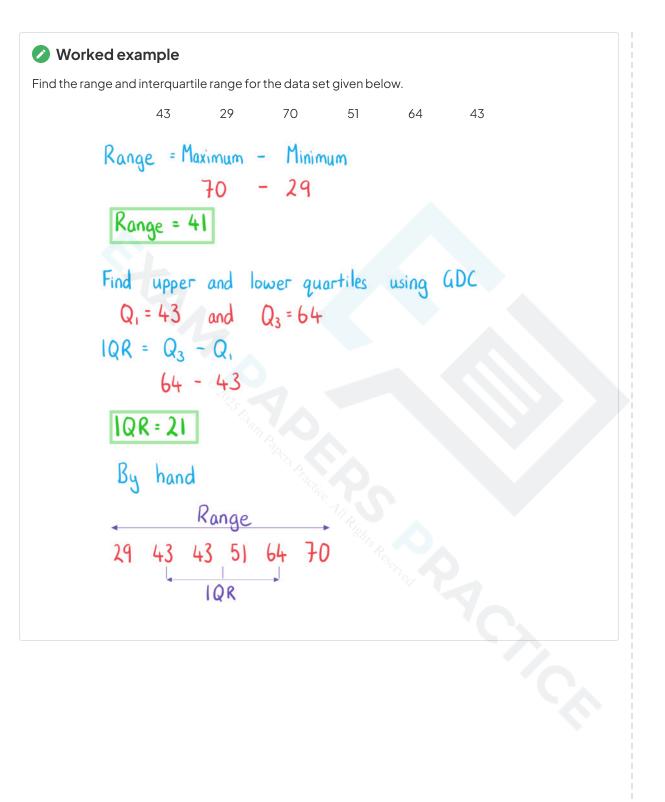
#### What are the range and interquartile range?

- The range and interquartile range are both measures of dispersion
  - They describe how spread out the data is
- The range is the largest value of the data minus the smallest value of the data
- The interquartile range is the range of the central 50% of data
  - It is the upper quartile minus the lower quartile

$$IQR = Q_3 - Q_1$$

- This is given in the **formula booklet**
- The units for the range and interquartile range are the same as the units for the data







### **Standard Deviation & Variance**

#### What are the standard deviation and variance?

- The standard deviation and variance are both measures of dispersion
  - They describe how spread out the data is in relation to the mean
- The variance is the mean of the squares of the differences between the values and the mean
  - Variance is denoted  $\sigma^2$
- The standard deviation is the square-root of the variance
  - Standard deviation is denoted  $\sigma$
- The **units** for the standard deviation are the **same** as the units for the data
- The **units** for the variance are the **square** of the units for the data

#### How are the standard deviation and variance calculated for ungrouped data?

- In the exam you will be expected to use the statistics function on your **GDC** to calculate the standard deviation and the variance
- Calculating the standard deviation and the variance by hand may deepen your understanding

$$\sum_{i=1}^{k} f_i (x_i - \mu)^2$$

n

- The formula for **variance** is  $\sigma^2 = \frac{1}{2}$ 
  - This can be rewritten as

$$\sigma^{2} = \frac{\sum_{i=1}^{k} f_{i} x_{i}^{2}}{n} - \mu^{2}$$

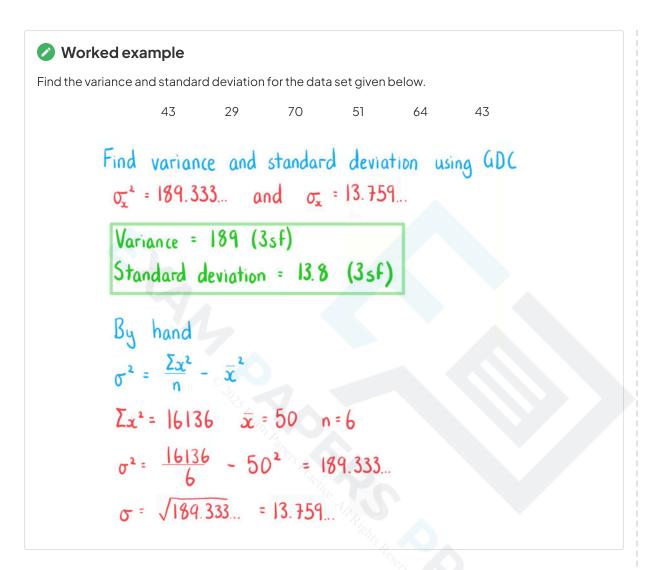
$$\sqrt{\frac{\sum_{i=1}^{k} f_i (x_i - \mu)^2}{n}}$$

- The formula for standard deviation is  $\sigma$  =
  - This can be rewritten as

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \mu^2}$$

• You **do not** need to learn these formulae as you will use your GDC to calculate these







### 4.1.4 Frequency Tables

### **Ungrouped** Data

### How are frequency tables used for ungrouped data?

- Frequency tables can be used for ungrouped data when you have lots of the same values within a data set
  - They can be used to collect and present data easily
- If the value 4 has a frequency of 3 this means that there are three 4's in the data set

## How are measures of central tendency calculated from frequency tables with ungrouped data?

- The mode is the value that has the highest frequency
- The **median** is the **middle** value
  - Use cumulative frequencies (running totals) to find the median
- The **mean** can be calculated by
  - Multiplying each value x<sub>i</sub> by its frequency f<sub>i</sub>
  - Summing to get  $\Sigma f_i x_i$
  - Dividing by the total frequency  $n = \Sigma f_i$
  - This is given in the formula booklet

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

• Your GDC can calculate these statistical measures if you input the values and their frequencies using the statistics mode

#### How are measures of dispersion calculated from frequency tables with ungrouped data?

- The range is the largest value of the data minus the smallest value of the data
- The interquartile range is calculated by

$$IQR = Q_3 - Q_1$$

- The quartiles can be found by using your GDC and inputting the values and their frequencies
- The standard deviation and variance can be calculated by hand using the formulae
  - Variance

$$\sigma^{2} = \frac{\sum_{i=1}^{k} f_{i} x_{i}^{2}}{n} - \mu^{2}$$

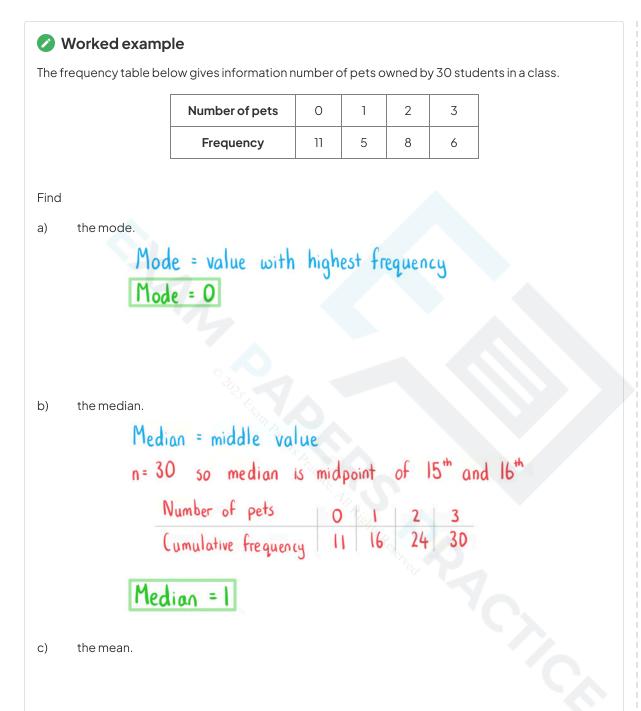


Standard deviation

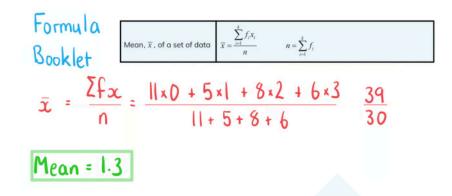
$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \mu^2}$$

- You **do not need to learn** these formulae as you will be expected to use your GDC to find the standard deviation and variance
  - You may want to see these formulae to deepen your understanding









d) the standard deviation.

Use GDC	σ <sub>x</sub> = 1.159
Standard	deviation = 1.16 (3sf)



### **Grouped Data**

#### How are frequency tables used for grouped data?

- Frequency tables can be used for grouped data when you have lots of the same values within the same interval
  - Class intervals will be written using inequalities and without gaps
    - $10 \le x \le 20$  and  $20 \le x \le 30$
  - If the class interval  $10 \le x \le 20$  has a frequency of 3 this means there are three values in that interval
    - You do not know the exact data values when you are given grouped data

## How are measures of central tendency calculated from frequency tables with grouped data?

- The modal class is the class that has the highest frequency
  - This is for equal class intervals only
- The **median** is the **middle** value
  - The exact value can not be calculated but it can be estimated by using a cumulative frequency graph
- The **exact mean** can not be calculated as you do not have the raw data
- The mean can be estimated by
  - Identifying the mid-interval value (midpoint) x<sub>i</sub> for each class
  - Multiplying each value by the class frequency f<sub>i</sub>
  - Summing to get  $\Sigma f_i x_i$
  - Dividing by the total frequency  $n = \Sigma f_i$
  - This is given in the formula booklet

$$\overline{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$$

• Your **GDC** can estimate the mean if you input the mid-interval values and the class frequencies using the statistics mode

#### How are measures of dispersion calculated from frequency tables with grouped data?

- The exact range can not be calculated as the largest and smallest values are unknown
- The interquartile range can be estimated by

$$IQR = Q_3 - Q_1$$

- Estimates of the quartiles can be found by using a cumulative frequency graph
- The standard deviation and variance can be estimated using the mid-interval values x<sub>i</sub> in the formulae
  - Variance



$$\sigma^2 = \frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \mu^2$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i x_i^2}{n} - \mu^2}$$

- You **do not need to learn** these formulae as you will be expected to use your GDC to estimate the standard deviation and variance using the mid-interval values
  - You may want to use these formulae to deepen your understanding



### Worked example

The table below shows the heights in cm of a group of 25 students.

Height, $h$	Frequency		
150 ≤ <i>h</i> < 155	3		
$155 \le h < 160$	5		
$160 \le h < 165$	9		
$165 \le h < 170$	7		
$170 \le h < 175$	1		

#### a) Write down the modal class.

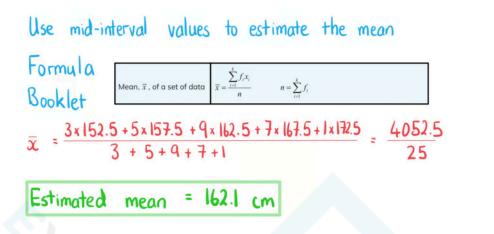
Modal class = class with highest frequency Modal class = 160 < h < 165

b) Write down the mid-interval value of the modal class.

Mid-interval value = Upper boundary + lower boundary 2 160 + 165 2 Mid-interval value = 162.5 cm

c) Calculate an estimate for the mean height.







### 4.1.5 Linear Transformations of Data

### Linear Transformations of Data

### Why are linear transformations of data used?

- Sometimes data might be very large or very small
- You can apply a linear transformation to the data to make the values more manageable
  - You may have heard this referred to as:
    - Effects of constant changes
    - Linear coding
- Linear transformations of data can affect the statistical measures

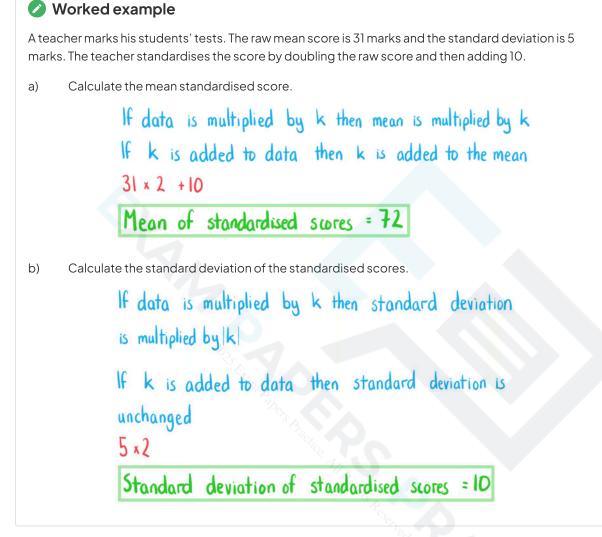
#### How is the mean affected by a linear transformation of data?

- Let  $\overline{X}$  be the **mean** of some data
- If you multiply each value by a constant k then you will need to multiply the mean by k
  - Mean is  $k\overline{x}$
- If you add or subtract a constant a from all the values then you will need to add or subtract the constant a to the mean
  - Mean is  $\overline{X} \pm a$

#### How is the variance and standard deviation affected by a linear transformation of data?

- Let  $\sigma^2$  be the **variance** of some data
  - $\sigma$  is the standard deviation
- If you multiply each value by a constant *k* then you will need to multiply the variance by k<sup>2</sup>
  - Variance is  $k^2 \sigma^2$
  - You will need to **multiply** the **standard deviation** by the **absolute value** of k
    - Standard deviation is  $|k|\sigma$
  - If you add or subtract a constant a from all the values then the variance and the standard deviation stay the same
    - Variance is  $\sigma^2$
    - Standard deviation is  $\sigma$







### 4.1.6 Outliers

### Outliers

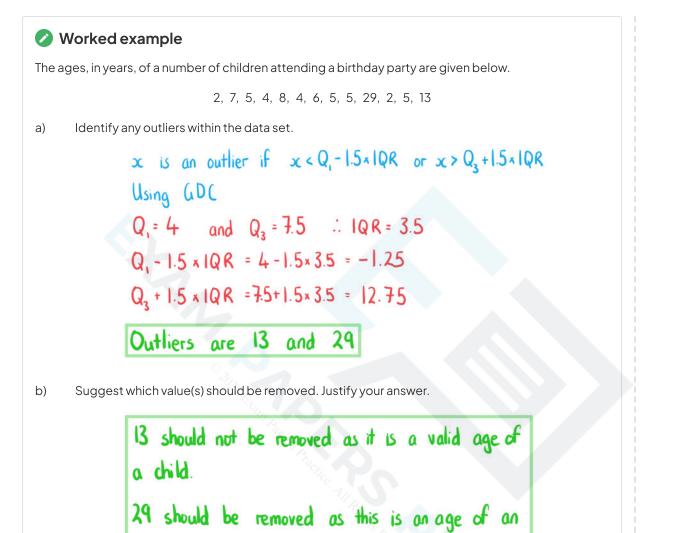
### What are outliers?

- Outliers are extreme data values that do not fit with the rest of the data
  - They are either a lot bigger or a lot smaller than the rest of the data
- Outliers are defined as values that are more than 1.5 × IQR from the nearest quartile
  - x is an outlier if x < Q<sub>1</sub> 1.5 × IQR or x > Q<sub>3</sub> + 1.5 × IQR
- Outliers can have a big effect on some statistical measures

### Should I remove outliers?

- The decision to remove outliers will depend on the context
- Outliers should be removed if they are found to be errors
  - The data may have been recorded incorrectly
  - For example: The number 17 may have been recorded as 71 by mistake
- Outliers **should not be removed** if they are a **valid part of the sample** 
  - The data may need to be checked to verify that it is not an error
  - For example: The annual salaries of employees of a business might appear to have an outlier but this could be the director's salary





adult.

For more help, please visit www.exampaperspractice.co.uk



### 4.1.7 Univariate Data

### **Box Plots**

Univariate data is data that is in **one variable**.

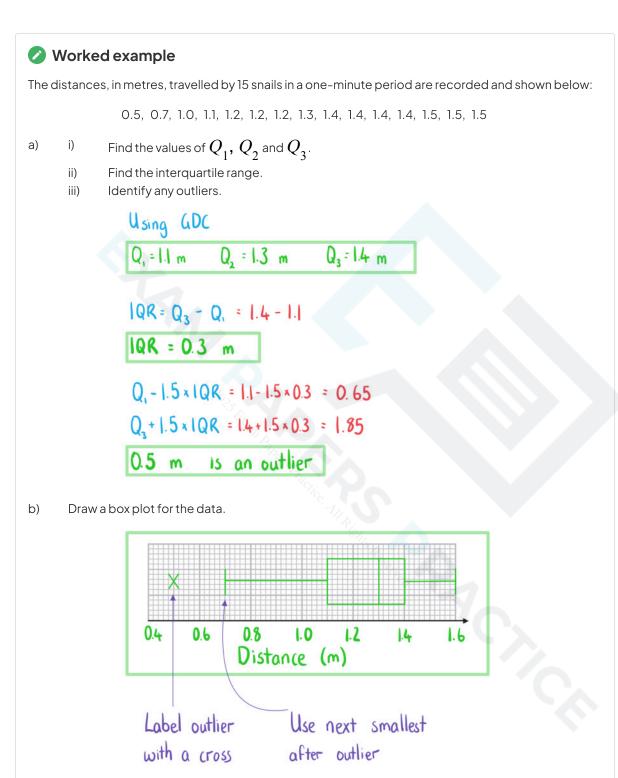
#### What is a box plot (box and whisker diagram)?

- A box plot is a graph that clearly shows key statistics from a data set
  - It shows the median, quartiles, minimum and maximum values and outliers
  - It does not show any other individual data items
- The middle 50% of the data will be represented by the box section of the graph and the lower and upper 25% of the data will be represented by each of the whiskers
- Any outliers are represented with a cross on the outside of the whiskers
   If there is an outlier then the whisker will end at the value before the outlier
- Only one axis is used when graphing a box plot
- It is still important to make sure the axis has a clear, even scale and is labelled with units

#### What are box plots useful for?

- Box plots can clearly show the shape of the distribution
  If a box plot is symmetrical about the median then the data could be normally distributed
- Box plots are often used for **comparing two sets of data** 
  - Two box plots will be drawn next to each other using the same axis
  - They are useful for **comparing data** because it is easy to see the main shape of the distribution of the data from a box plot
    - You can easily compare the medians and interquartile ranges







### **Cumulative Frequency Graphs**

#### What is cumulative frequency?

- The cumulative frequency of x is the running total of the frequencies for the values that are less than or equal to x
- For grouped data you use the upper boundary of a class interval to find the cumulative frequency of that class

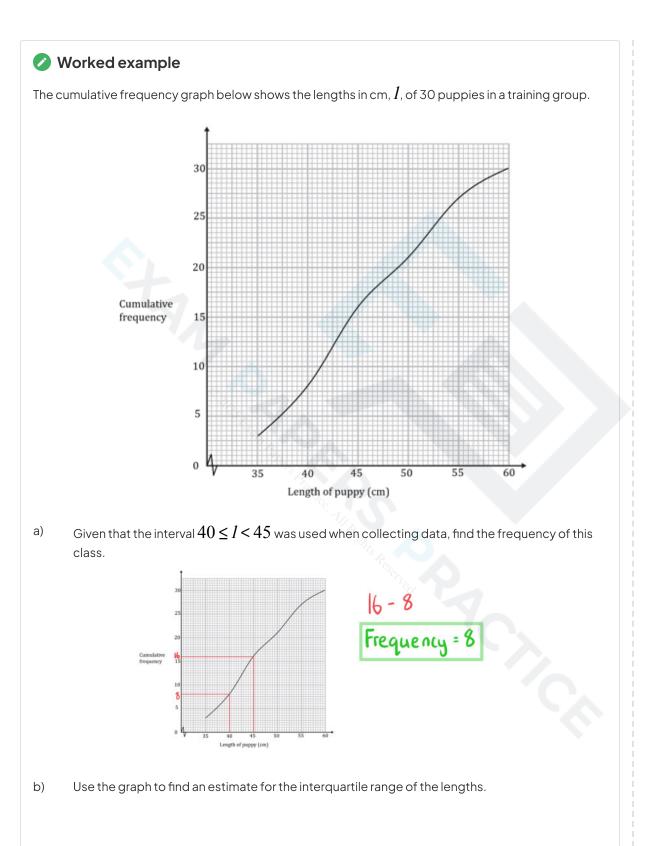
#### What is a cumulative frequency graph?

- A cumulative frequency graph is used with data that has been organised into a **grouped frequency** table
- Some coordinates are plotted
  - The x-coordinates are the upper boundaries of the class intervals
  - The y-coordinates are the cumulative frequencies of that class interval
- The coordinates are then joined together by hand using a **smooth increasing curve**

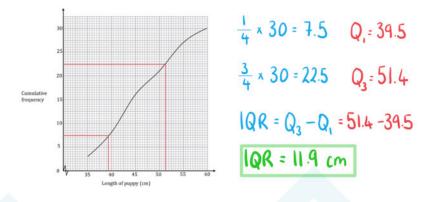
### What are cumulative frequency graphs useful for?

- They can be used to estimate statistical measures
  - Draw a horizontal line from the y-axis to the curve
    - For the median: draw the line at 50% of the total frequency
    - For the lower quartile: draw the line at 25% of the total frequency
    - For the upper quartile: draw the line at 75% of the total frequency
    - For the p<sup>th</sup> percentile: draw the line at p% of the total frequency
  - Draw a vertical line down from the curve to the x-axis
  - This **x-value** is the relevant statistical measure
- They can used to estimate the number of values that are bigger/small than a given value
  - Draw a vertical line from the given value on the x-axis to the curve
  - Draw a horizontal line from the curve to the y-axis
  - This value is an estimate for how many values are less than or equal to the given value
    - To estimate the number that is greater than the value subtract this number from the total frequency
  - They can be used to estimate the interquartile range  $IQR = Q_3 Q_1$
  - They can be used to construct a **box plot** for grouped data

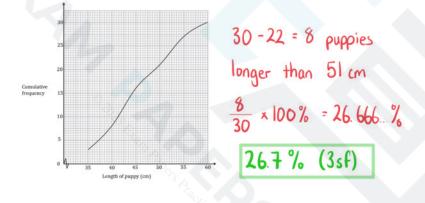








c) Estimate the percentage of puppies with length more than 51 cm.





### **Histograms**

### What is a (frequency) histogram?

- A frequency histogram clearly shows the frequency of class intervals
  - The classes will have **equal class intervals**
  - The **frequency** will be on the *y*-axis
  - The bar for a class interval will begin at the lower boundary and end at the upper boundary
- A frequency histogram is **similar to a bar chart** 
  - A bar chart is used for qualitative or discrete data and has gaps between the bars
  - A frequency histogram is used for continuous data and has no gaps between bars

### What are (frequency) histograms useful for?

- They show the **modal class** clearly
- They show the shape of the distribution
  - It is important the class intervals are of equal width
- They can show whether the variable can be modelled by a normal distribution
  - If the shape is symmetrical and bell-shaped

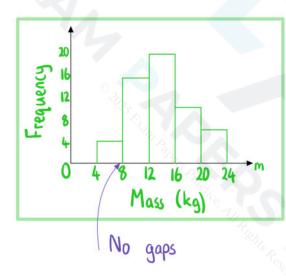


### Worked example

The table below and its corresponding histogram show the mass, in kg, of some new born bottlenose dolphins.

Mass, <i>M</i> kg	Frequency		
$4 \le m < 8$	4		
8 ≤ <i>m</i> < 12	15		
12 ≤ <i>m</i> < 16	19		
$16 \le m < 20$	10		
$20 \le m < 24$	6		

Draw a frequency histogram to represent the data. a)



Write down the modal class. b)



### 4.1.8 Interpreting Data

### **Interpreting Data**

### How do l interpret statistical measures?

- The mode is useful for qualitative data
  - It is not as useful for quantitative data as there is not always a unique mode
- The mean includes all values
  - It is affected by outliers
  - A smaller/larger mean is preferable depending on the scenario
    - A smaller mean time for completing a puzzle is better
    - A bigger mean score on a test is better
- The median is not affected by outliers
  - It does not use all the values
- The range gives the full spread of the all of the data
  - It is affected by outliers
- The interquartile range gives the spread of the middle 50% about the median and is not affected by outliers
  - It does not use all the values
  - A bigger IQR means the data is more spread out about the median
  - A smaller IQR means the data is more centred about the median
- The **standard deviation** and **variance** use all the values to give a measure of the **average spread** of the data about the mean
  - They are affected by outliers
  - A bigger standard deviation means the data is more spread out about the mean
  - A smaller standard deviation means the data is more centred about the mean

### How do I choose which diagram to use to represent data?

- Box plots
  - Can be used with ungrouped **univariate** data
  - Shows the range, interquartile range and quartiles clearly
  - Very useful for comparing data patterns quickly
- Cumulative frequency graphs
  - Can be used with continuous grouped univariate data
  - Shows the running total of the frequencies that fall below the upper bound of each class
- Histograms
  - Can be used with continuous grouped univariate data
  - Used with equal class intervals
  - Shows the frequencies of the group
- Scatter diagrams
  - Can be used with ungrouped **bivariate** data
  - Shows the graphical relationship between the variables



#### How do I compare two or more data sets?

- Compare a **measure of central tendency** 
  - If the data contains outliers use the median
  - If the data is **roughly symmetrical use the mean**
- Compare a **measure of dispersion** 
  - If the data contains outliers use the interquartile range
  - If the data is roughly symmetrical use the standard deviation
- Consider whether it is better to have a smaller or bigger average
  - This will depend on the context
    - A smaller average time for completing a puzzle is better
    - A bigger average score on a test is better
- Consider whether it is better to have a smaller or bigger spread
  - Usually a smaller spread means it is more consistent
- Always relate the comparisons to the context and consider reasons
  - Consider the sampling technique and the data collection method



