EXAM PAPERS PRACTICE

Boost your performance and confidence with these topic-based exam questions

Practice questions created by actual examiners and assessment experts

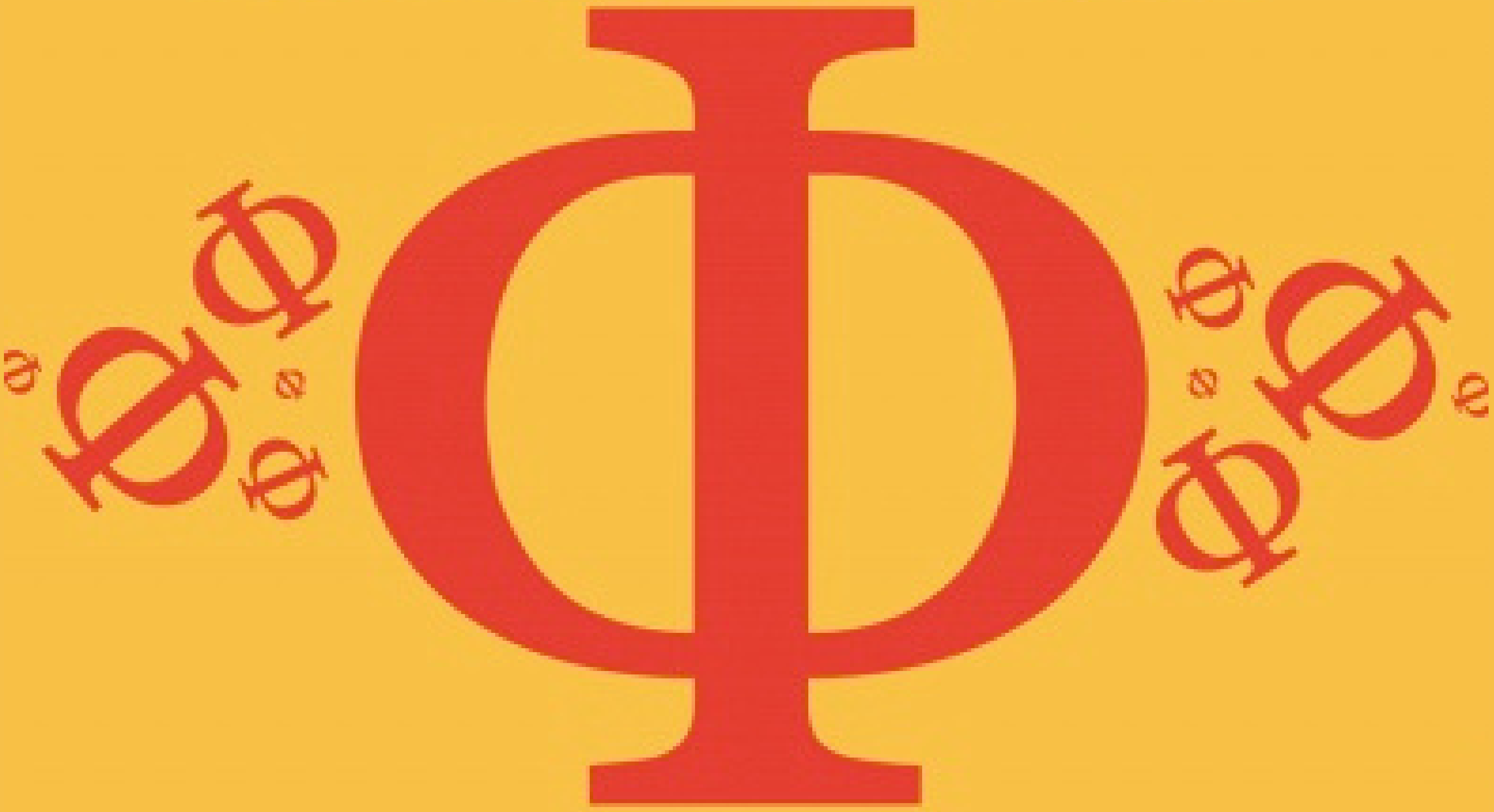
Detailed mark scheme

Suitable for all boards

Designed to test your ability and thoroughly prepare you

4. Statistics & Probability

4.2 Correlation & Regression



MATH

IB AI HL

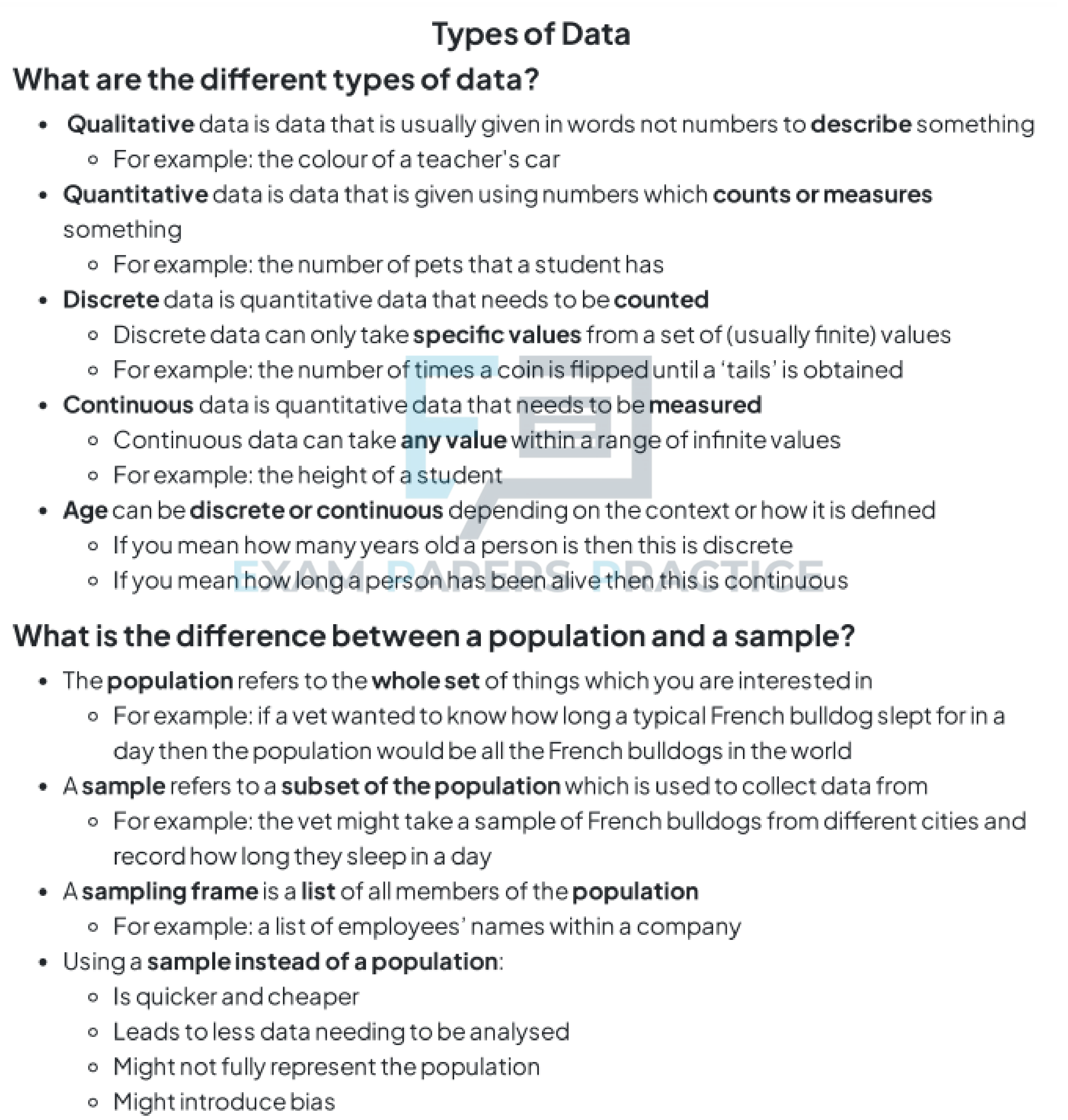
IB Maths DP

4. Statistics & Probability

CONTENTS

4.1 Statistics Toolkit

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 $k = \dfrac{\text{size of population } (N)}{\text{size of sample } (n)}$
    - choose a random starting point between 1 and *k*
    - select every *k*th 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
    - $\frac{\text{size of sample } (n)}{\text{size of population } (N)} \times \text{number of members in the group}$
  - 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 quota 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

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.

Total number of mice

$540 + 260 = 800$

Fraction of field mice

Field mice $\frac{540}{800} \times 10 = 6.75$

Sample size

Fraction of harvest mice

Harvest mice $\frac{260}{800} \times 10 = 3.25$

$\boxed{\text{Include 7 field mice and 3 harvest mice}}$

b)

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

$\boxed{\text{Quota sampling}}$

c)

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

$\boxed{\text{Mark could improve his sampling method by increasing his sample size}}$

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	1	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**

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

  - Where $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$ is the sum of the n pieces of data
  - The mean can be represented by the symbol μ
- Your **GDC** can calculate these statistical measures if you input the data using the statistics mode

Worked Example

Find the mode, median and mode for the data set given below.

43 29 70 51 64 43

Mode is the most common

$\boxed{\text{Mode} = 43}$

Median is the middle when in order

29 43 43 51 64 70

$\uparrow$

$\frac{43+51}{2} = 47$

$\boxed{\text{Median} = 47}$

$\text{Mean} = \frac{\Sigma x}{n}$

$\Sigma x = 300$ and $n = 6$ $\quad \frac{300}{6} = 50$

$\boxed{\text{Mean} = 50}$

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_1** splits the lowest 25% from the highest 75%
    - The **median, Q_2** splits the lowest 50% from the highest 50%
    - The **upper quartile, Q_3** 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

$$\text{IQR} = Q_3 - Q_1$$

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

Worked Example

Find the range and interquartile range for the data set given below.

43 29 70 51 64 43

Range = Maximum − Minimum

70 − 29

$\boxed{\text{Range} = 41}$

Find upper and lower quartiles using GDC

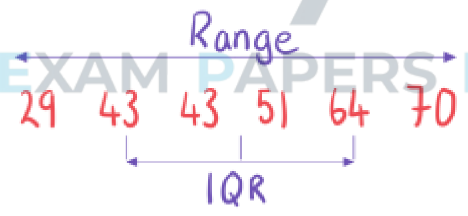
$Q_1 = 43$ and $Q_3 = 64$

$IQR = Q_3 - Q_1$

$64 - 43$

$\boxed{IQR = 21}$

By hand



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 σ^2
- The **standard deviation** is the **square-root** of the **variance**
  - Standard deviation is denoted σ
- 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
- The formula for **variance** is $\sigma^2 = \frac{\sum_{i=1}^{k} f_i(x_i-\mu)^2}{n}$
  - This can be rewritten as

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

- The formula for **standard deviation** is $\sigma = \sqrt{\frac{\sum_{i=1}^{k} f_i(x_i-\mu)^2}{n}}$
  - 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

Worked Example

Find the variance and standard deviation for the data set given below.

43 29 70 51 64 43

Find variance and standard deviation using GDC

$\sigma_x^2 = 189.333...$ and $\sigma_x = 13.759...$

Variance = 189 (3sf)

Standard deviation = 13.8 (3sf)

By hand

$$\sigma^2 = \frac{\Sigma x^2}{n} - \bar{x}^2$$

$\Sigma x^2 = 16136$ $\bar{x} = 50$ $n = 6$

$$\sigma^2 = \frac{16136}{6} - 50^2 = 189.333...$$

$$\sigma = \sqrt{189.333...} = 13.759...$$

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_i by its frequency f_i
  - Summing to get $\Sigma f_i x_i$
  - Dividing by the total frequency $n = \Sigma f_i$
  - This is given in the formula booklet

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

  - 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

$$\text{IQR} = Q_3 - Q_1$$

  - 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

Exam Tip

- Always check whether your answers make sense when using your GDC
  - The value for a measure of central tendency should be within the range of data

Worked Example

The frequency table below gives information number of pets owned by 30 students in a class.

Number of pets	0	1	2	3
Frequency	11	5	8	6

Find

a)
the mode.

Mode = value with highest frequency

$\boxed{\text{Mode} = 0}$

b)
the median.

Median = middle value

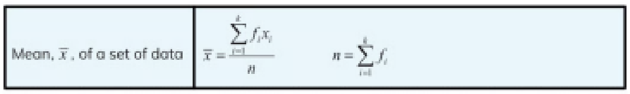
n = 30 so median is midpoint of 15th and 16th

Number of pets	0	1	2	3
Cumulative frequency	11	16	24	30

$\boxed{\text{Median} = 1}$

c)
the mean.

Formula Booklet



Mean, $\bar{x}$, of a set of data	$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$	$n = \sum_{i=1}^{k} f_i$

$$\bar{x} = \frac{\Sigma fx}{n} = \frac{11\times0 + 5\times1 + 8\times2 + 6\times3}{11 + 5 + 8 + 6} \quad \frac{39}{30}$$

$\boxed{\text{Mean} = 1.3}$

d)
the standard deviation.

Use GDC $\sigma_x = 1.159...$

$\boxed{\text{Standard deviation} = 1.16 \text{ (3sf)}}$

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
    - $10 \le x < 20$ and $20 \le x < 30$
  - If the class interval $10 \le x < 20$ 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_i for each class
  - Multiplying each value by the class frequency f_i
  - Summing to get $\Sigma f_i x_i$
  - Dividing by the total frequency $n = \Sigma f_i$
  - This is given in the formula booklet

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

  - 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

$$\text{IQR} = Q_3 - Q_1$$

  - **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_i 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

Exam Tip

- As you can only estimate statistical measures from a grouped frequency table it is good practice to indicate that the values are not exact
  - You can do this by rounding values rather than leaving as surds and fractions
  - $\overline{x} = 0.333$ (3sf) rather than $\overline{x} = \frac{1}{3}$

Worked Example

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

Height, h	Frequency
$150 \le h < 155$	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

$\boxed{\text{Modal class} = 160 \le h < 165}$

b)

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

Mid-interval value = $\dfrac{\text{Upper boundary} + \text{lower boundary}}{2}$

$\dfrac{160 + 165}{2}$

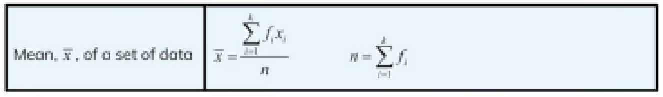
$\boxed{\text{Mid-interval value} = 162.5 \text{ cm}}$

c)

Calculate an estimate for the mean height.

Use mid-interval values to estimate the mean

Formula Booklet



Mean, $\bar{x}$, of a set of data	$\bar{x} = \dfrac{\sum_{i=1}^{k} f_i x_i}{n}$ $\quad n = \sum_{i=1}^{k} f_i$

$$\bar{x} = \frac{3 \times 152.5 + 5 \times 157.5 + 9 \times 162.5 + 7 \times 167.5 + 1 \times 172.5}{3 + 5 + 9 + 7 + 1} = \frac{4052.5}{25}$$

$\boxed{\text{Estimated mean} = 162.1 \text{ cm}}$

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 σ^2 be the **variance** of some data
  - σ is the **standard deviation**
- If you **multiply** each value by a constant ***k*** then you will need to **multiply** the **variance by *k*²**
  - 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 σ^2
    - Standard deviation is σ

Exam Tip

- If you forget these results in an exam then you can look in the HL section of the formula booklet to see them written in a more algebraic way
  - Linear transformation of a single variable

$$\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$$
$$\mathrm{Var}(aX + b) = a^2\,\mathrm{Var}(X)$$

  - where E(...) means the mean and Var(...) means the variance

Worked Example

A teacher marks his students' tests. The raw mean score is 31 marks and the standard deviation is 5 marks. The teacher standardises the score by doubling the raw score and then adding 10.

a)

Calculate the mean standardised score.

If data is multiplied by k then mean is multiplied by k

If k is added to data then k is added to the mean

$31 \times 2 + 10$

$\boxed{\text{Mean of standardised scores} = 72}$

b)

Calculate the standard deviation of the standardised scores.

If data is multiplied by k then standard deviation is multiplied by $|k|$

If k is added to data then standard deviation is unchanged

5×2

$\boxed{\text{Standard deviation of standardised scores} = 10}$

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 $\mathbf{x < Q_1 - 1.5 \times IQR}$ or $\mathbf{x > Q_3 + 1.5 \times IQR}$
- 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

Worked Example

The ages, in years, of a number of children attending a birthday party are given below.

2, 7, 5, 4, 8, 4, 6, 5, 5, 29, 2, 5, 13

a)

Identify any outliers within the data set.

x is an outlier if $x < Q_1 - 1.5 \times IQR$ or $x > Q_3 + 1.5 \times IQR$

Using GDC

$Q_1 = 4$ and $Q_3 = 7.5$ $\therefore IQR = 3.5$

$Q_1 - 1.5 \times IQR = 4 - 1.5 \times 3.5 = -1.25$

$Q_3 + 1.5 \times IQR = 7.5 + 1.5 \times 3.5 = 12.75$

Outliers are 13 and 29

b)

Suggest which value(s) should be removed. Justify your answer.

13 should not be removed as it is a valid age of a child.

29 should be removed as this is an age of an adult.

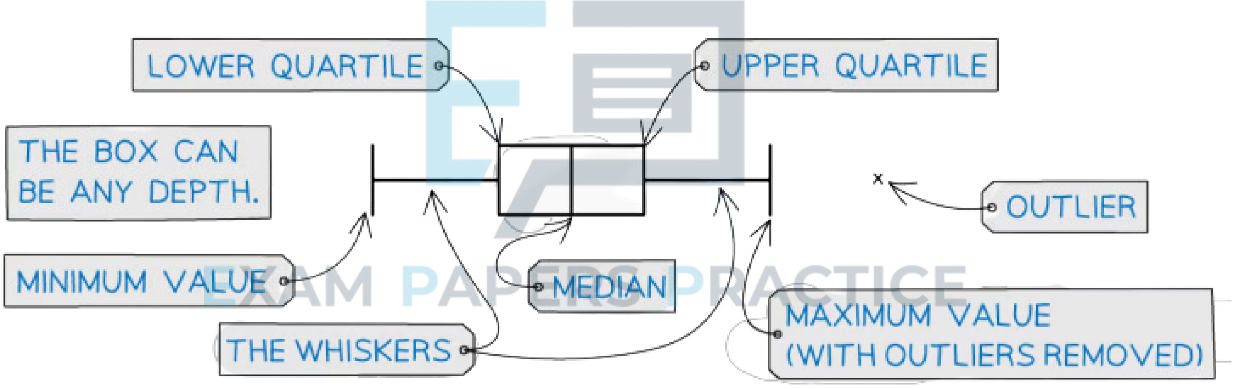
4.1.7 Univariate Data

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

Exam Tip

- In an exam you can use your GDC to draw a box plot if you have the raw data
  - You calculator's box plot can also include outliers so this is a good way to check

Worked Example

The distances, in metres, travelled by 15 snails in a one-minute period are recorded and shown below:

0.5, 0.7, 1.0, 1.1, 1.2, 1.2, 1.2, 1.3, 1.4, 1.4, 1.4, 1.4, 1.5, 1.5, 1.5

a)

i)

Find the values of Q_1, Q_2 and Q_3.

ii)

Find the interquartile range.

iii)

Identify any outliers.

Using GDC

$\boxed{Q_1 = 1.1 \text{ m} \quad Q_2 = 1.3 \text{ m} \quad Q_3 = 1.4 \text{ m}}$

$IQR = Q_3 - Q_1 = 1.4 - 1.1$

$\boxed{IQR = 0.3 \text{ m}}$

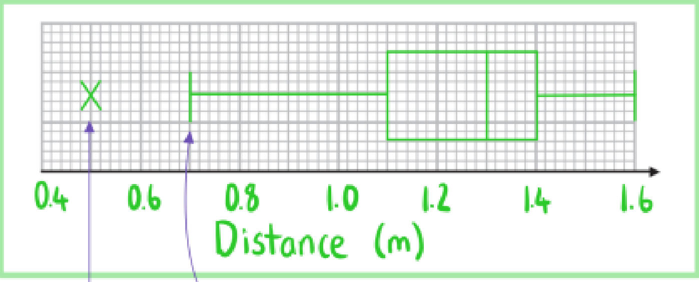
$Q_1 - 1.5 \times IQR = 1.1 - 1.5 \times 0.3 = 0.65$

$Q_3 + 1.5 \times IQR = 1.4 + 1.5 \times 0.3 = 1.85$

$\boxed{0.5 \text{ m is an outlier}}$

b)

Draw a box plot for the data.



Label outlier with a cross

Use next smallest after outlier

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 pth 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** $\mathrm{IQR} = Q_3 - Q_1$
  - They can be used to construct a **box plot** for grouped data

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**

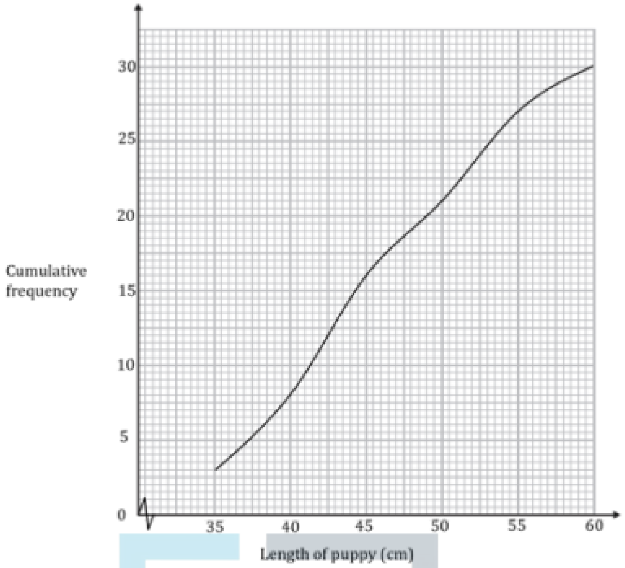
What are cumulative frequency graphs useful for?

- 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^{th} 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

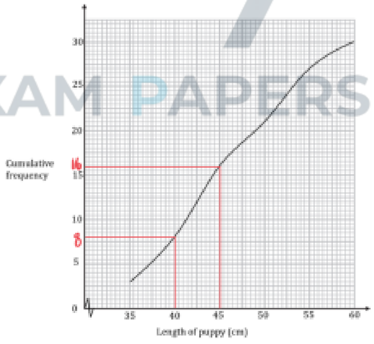
- They can used to estimate the number of values that are bigger/small than a given value
    - Draw a **vertical line** from the given value on the *x*-axis to the curve
    - Draw a **horizontal line** from the curve to the *y*-axis
    - This value is an estimate for how many values are less than 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** $\text{IQR} = Q_3 - Q_1$
    - They can be used to construct a **box plot** for grouped data

Worked Example

The cumulative frequency graph below shows the lengths in cm, l, of 30 puppies in a training group.



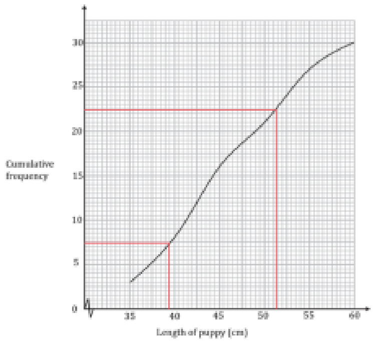
a)
Given that the interval $40 \le l < 45$ was used when collecting data, find the frequency of this class.



$16 - 8$

$\boxed{\text{Frequency} = 8}$

b)
Use the graph to find an estimate for the interquartile range of the lengths.



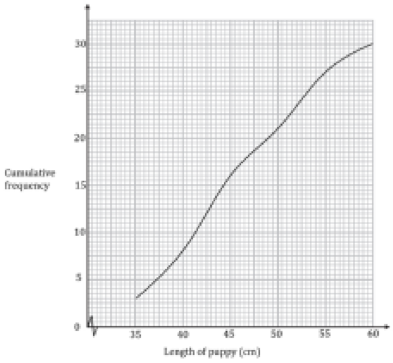
$\frac{1}{4} \times 30 = 7.5$ $\quad Q_1 = 39.5$

$\frac{3}{4} \times 30 = 22.5$ $\quad Q_3 = 51.4$

$IQR = Q_3 - Q_1 = 51.4 - 39.5$

$\boxed{IQR = 11.9 \text{ cm}}$

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



30 - 22 = 8 puppies longer than 51 cm

$\frac{8}{30} \times 100\% = 26.666...\%$

$\boxed{26.7\% \ (3sf)}$

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

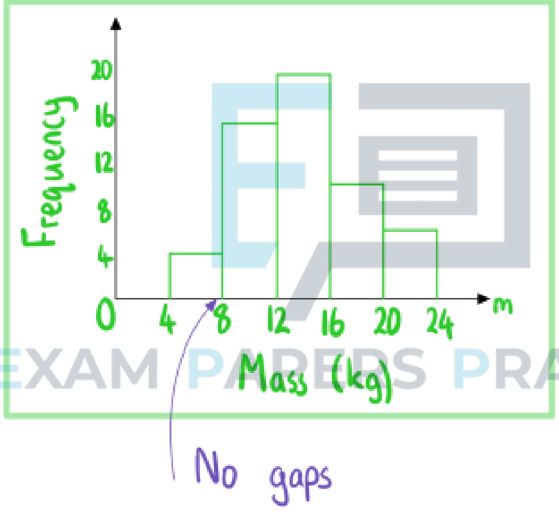
Worked Example

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

Mass, m kg	Frequency
$4 \le m < 8$	4
$8 \le m < 12$	15
$12 \le m < 16$	19
$16 \le m < 20$	10
$20 \le m < 24$	6

a)

Draw a frequency histogram to represent the data.



b)

Write down the modal class.

Modal class = class with highest frequency

$\boxed{\text{Modal class} = 12 \le m < 16}$

4.1.8 Interpreting Data

Interpreting Data

How do I 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

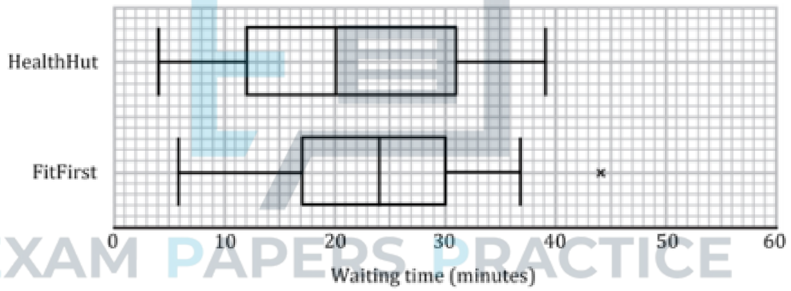
How do I compare two or more data sets?

- 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

Worked Example

The box plots below show the waiting times for the two doctor surgeries, HealthHut and FitFirst.



Compare the two distributions of waiting times in context.

Compare :

- a measure of central tendency
- a measure of dispersion

HealthHut's median waiting time is smaller than FitFirst's (20 < 24). On average patients get seen quicker at HealthHut.

FitFirst's interquartile range is smaller than HealthHut's (13 < 19). There is less variability of waiting times at FitFirst.

4.2 Correlation & Regression

4.2.1 Bivariate Data

Scatter Diagrams

What does bivariate data mean?

- **Bivariate data** is data which is collected on **two variables** and looks at how one of the factors affects the other
  - Each data value from one variable will be **paired** with a data value from the other variable
  - The two variables are often related, but do not have to be

What is a scatter diagram?

- A **scatter diagram** is a way of graphing bivariate data
  - One variable will be on the *x*-axis and the other will be on the y-axis
  - The variable that can be **controlled** in the data collection is known as the **independent** or **explanatory variable** and is plotted on the *x*-axis
  - The variable that is **measured** or discovered in the data collection is known as the **dependent** or **response variable** and is plotted on the y-axis
- Scatter diagrams can contain **outliers** that do not follow the trend of the data

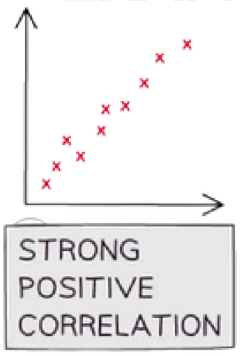
Exam Tip

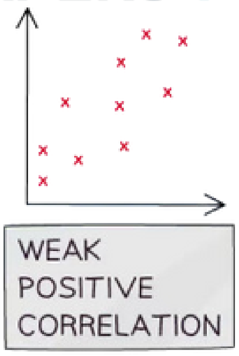
- If you use scatter diagrams in your Internal Assessment then be aware that finding outliers for bivariate data is different to finding outliers for univariate data
  - (x, y) could be an outlier for the bivariate data even if x and y are not outliers for their separate univariate data

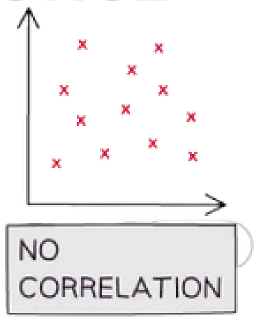
Correlation

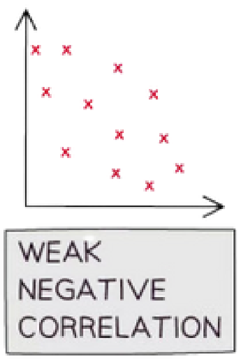
What is correlation?

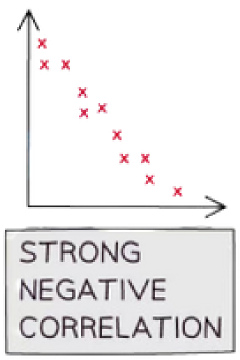
- **Correlation** is how the **two variables change in relation to each other**
    - Correlation could be the result of a **causal relationship** but this is not always the case
- **Linear correlation** is when the changes are proportional to each other
- **Perfect linear correlation** means that the bivariate data will all lie on a straight line on a scatter diagram
- When describing correlation mention
    - The type of the correlation
        - **Positive correlation** is when an **increase** in one variable results in the other variable **increasing**
        - **Negative correlation** is when an **increase** in one variable results in the other variable **decreasing**
        - **No linear correlation** is when the data points don't appear to follow a trend
    - The strength of the correlation
        - **Strong linear correlation** is when the data points lie **close** to a **straight line**
        - **Weak linear correlation** is when the data points are **not close** to a **straight line**
- If there is **strong linear correlation** you can draw a **line of best fit** (by eye)
    - The line of best fit will pass through the mean point $(\bar{x}, \bar{y})$
    - If you are asked to draw a line of best fit
        - Plot the mean point
        - Draw a line going through it that follows the trend of the data











What is the difference between correlation and causation?

- It is important to be aware that just because correlation exists, it does not mean that the change in one of the variables is **causing** the change in the other variable
    - **Correlation does not imply causation!**
- If a change in one variable **causes** a change in the other then the two variables are said to have a **causal relationship**
    - Observing correlation between two variables does **not always** mean that there is a causal relationship
        - There could be **underlying factors** which is causing the correlation
    - Look at the two variables in question and consider the context of the question to decide if there could be a causal relationship
        - If the two variables are temperature and number of ice creams sold at a park then it is likely to be a causal relationship
        - Correlation may exist between global temperatures and the number of monkeys kept as pets in the UK but they are unlikely to have a causal relationship

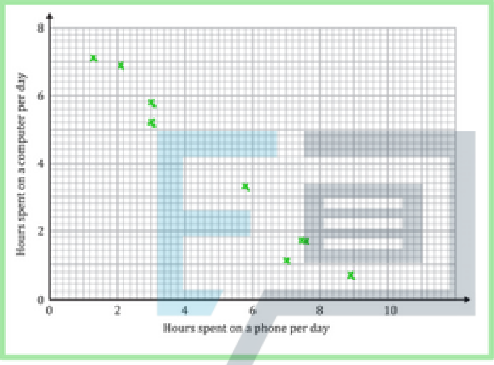
Worked Example

A teacher is interested in the relationship between the number of hours her students spend on a phone per day and the number of hours they spend on a computer. She takes a sample of nine students and records the results in the table below.

Hours spent on a phone per day	7.6	7.0	8.9	3.0	3.0	7.5	2.1	1.3	5.8
Hours spent on a computer per day	1.7	1.1	0.7	5.8	5.2	1.7	6.9	7.1	3.3

a)

Draw a scatter diagram for the data.



b)

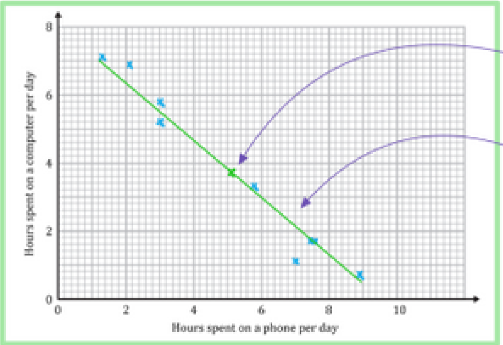
Describe the correlation.

Strong negative linear correlation

c)

Draw a line of best fit.

Mean point $(\bar{x}, \bar{y}) = (5.133..., 3.722...)$



Plot the mean point

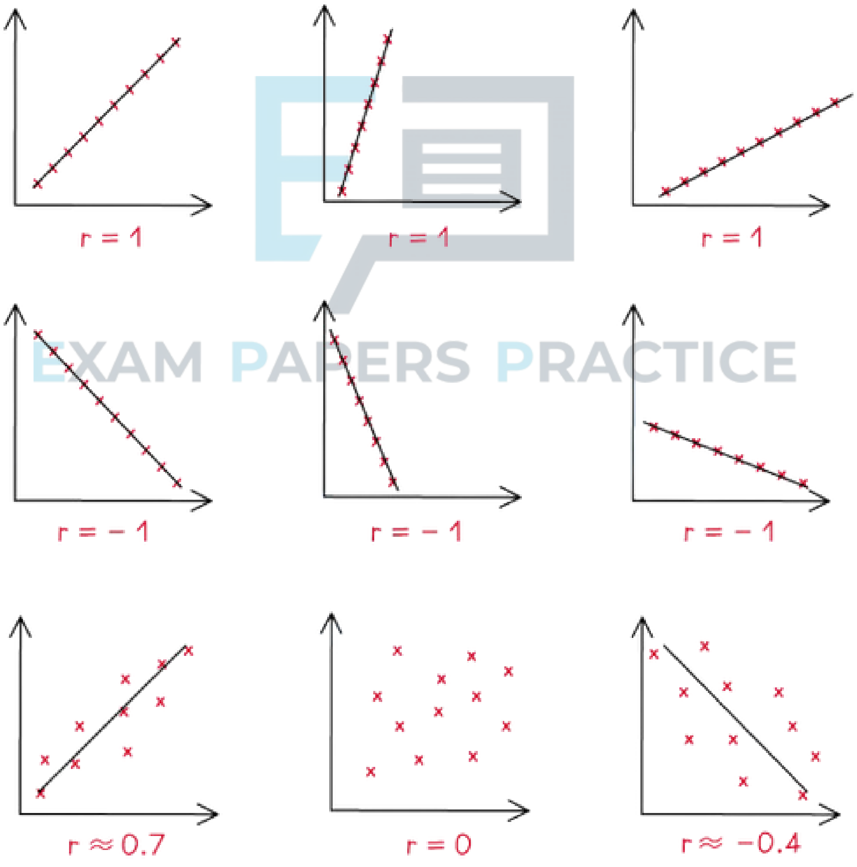
Draw it by eye

4.2.2 Correlation Coefficients

PMCC

What is Pearson's product-moment correlation coefficient?

- Pearson's product-moment correlation coefficient (PMCC) is a way of giving a numerical value to a **linear relationship** of bivariate data
- The PMCC of a sample is denoted by the letter r
  - r can take any value such that $-1 \le r \le 1$
  - A **positive value** of r describes **positive correlation**
  - A **negative value** of r describes **negative correlation**
  - $r = 0$ means there is **no linear correlation**
  - $r = 1$ means **perfect positive linear** correlation
  - $r = -1$ means **perfect negative linear** correlation
  - The closer to 1 or -1 the stronger the correlation



How do I calculate Pearson's product-moment correlation coefficient (PMCC)?

- You will be expected to use the statistics mode on your GDC to calculate the PMCC
- The formula can be useful to deepen your understanding

$$r = \frac{S_{xy}}{S_x S_y}$$

- $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$ is linked to the **covariance**
- $S_x = \sqrt{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}$ and $S_y = \sqrt{\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2}$ are linked to the **variances**
  - You **do not need to learn this** as using your GDC will be expected

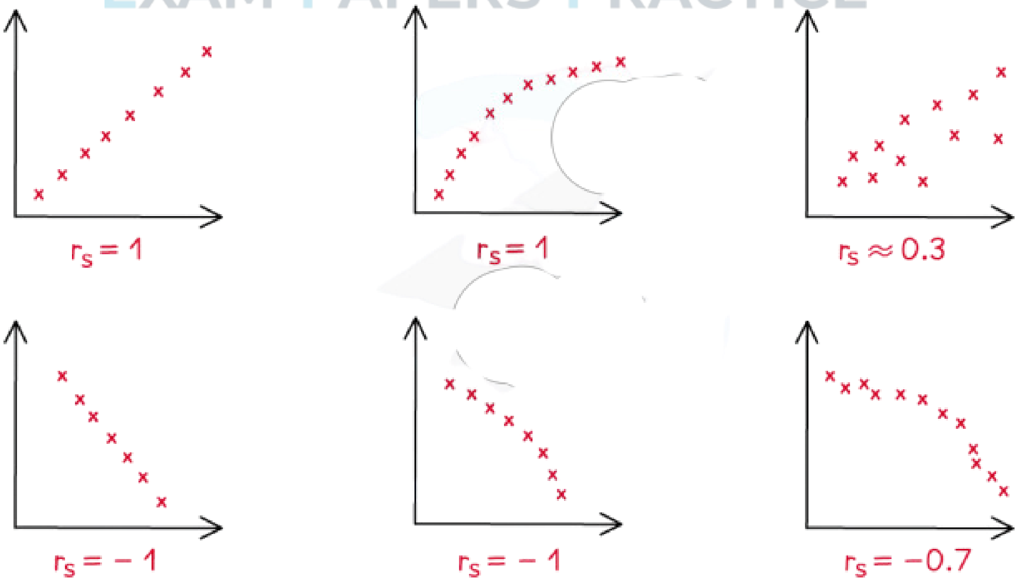
When does the PMCC suggest there is a linear relationship?

- **Critical values** of *r* indicate when the PMCC would suggest there is a linear relationship
  - In your exam you will be given critical values where appropriate
  - Critical values will depend on the size of the sample
- If the **absolute value** of the **PMCC** is **bigger** than the **critical value** then this suggests a linear model is appropriate

Spearman's Rank

What is Spearman's rank correlation coefficient?

- Spearman's rank correlation coefficient is a measure of how well the relationship between two variables can be described using a **monotonic** function
  - **Monotonic** means the points are either always increasing or always decreasing
  - This can be used as a way to **measure correlation in linear models**
  - Though Spearman's Rank correlation coefficient can also be used to assess a non-linear relationship
- Each data is ranked, from biggest to smallest or from smallest to biggest
  - For *n* data values, they are ranked from 1 to *n*
  - It doesn't matter whether variables are ranked from biggest to smallest or smallest to biggest, but they must be ranked in the **same order for both variables**
- Spearman's rank of a sample is denoted by r_s
  - r_s can take any value such that $-1 \le r_s \le 1$
  - A **positive value** of r_s describes a **degree of agreement** between the rankings
  - A **negative value** of r_s describes a **degree of disagreement** between the rankings
  - $r_s = 0$ means the data shows **no monotonic behaviour**
  - $r_s = 1$ means the rankings are in complete agreement: the data is **strictly increasing**
    - An increase in one variable means an increase in the other
  - $r_s = -1$ means the rankings are in complete disagreement: the data is **strictly decreasing**
    - An increase in one variable means a decrease in the other
  - The **closer to 1 or -1** the **stronger the correlation** of the rankings



Copyright © Save My Exams. All Rights Reserved

How do I calculate Spearman's rank correlation coefficient (PMCC)?

- Rank each set of data independently
  - 1 to *n* for the *x*-values
  - 1 to *n* for the *y*-values
- If some values are equal then give each the average of the ranks they would occupy

  - For example: if the 3rd, 4th and 5th highest values are equal then give each the ranking of 4
    - $\frac{3+4+5}{3}=4$
- Calculate the PMCC of the **rankings** using your GDC
  - This value is **Spearman's rank correlation coefficient**

Appropriateness & Limitations

Which correlation coefficient should I use?

- **Pearson's PMCC** tests for a **linear relationship** between two variables
  - It will not tell you if the variables have a non-linear relationship
    - Such as exponential growth
  - Use this if you are interested in a linear relationship
- **Spearman's rank** tests for a **monotonic relationship** (always increasing or always decreasing) between two variables
  - It will not tell you what function can be used to model the relationship
    - Both linear relationships and exponential relationships can be monotonic
  - Use this if you think there is a non-linear monotonic relationship

How are Pearson's and Spearman's correlation coefficients connected?

- If there is **linear correlation** then the relationship is also **monotonic**
  - $r = 1 \Rightarrow r_s = 1$
  - $r = -1 \Rightarrow r_s = -1$
  - However the **converse is not true**
- It is possible for Spearman's rank to be 1 (or -1) but for the PMCC to be different
  - For example: data that follows an **exponential growth model**
    - $r_s = 1$ as the points are always increasing
    - $r < 1$ as the points do not lie on a straight line

Are Pearson's and Spearman's correlation coefficients affected by outliers?

- Pearson's PMCC **is** affected by outliers
  - as it uses the numerical value of each data point
- Spearman's rank is **not usually** affected by outliers
  - as it only uses the ranks of each data point

Exam Tip

- You can use your GDC to plot the scatter diagram to help you visualise the data

Worked Example

The table below shows the scores of eight students for a maths test and an English test.

Maths (x)	7	18	37	52	61	68	75	82
English (y)	5	3	9	12	17	41	49	97

a)
Write down the value of Pearson's product-moment correlation coefficient, r.

Enter data into GDC.

$r = 0.79433...$

$\boxed{r = 0.794 \ \text{(3sf)}}$

b)
Find the value of Spearman's rank correlation coefficient, r_s.

Rank the data

x rank	8	7	6	5	4	3	2	1
y rank	7	8	6	5	4	3	2	1

Find PMCC of ranks

$r_s = 0.97619...$

$\boxed{r_s = 0.976 \ \text{(3sf)}}$

c)
Comment on the values of the two correlation coefficients.

The value of r suggests there is strong positive linear correlation. The value of r_s suggests strong positive correlation, which is not necessarily linear.

4.2.3 Linear Regression

Linear Regression

What is linear regression?

- If **strong linear correlation** exists on a scatter diagram then the data can be modelled by a **linear model**
  - Drawing lines of best fit by eye is not the best method as it can be difficult to judge the best position for the line
- The **least squares regression line** is the line of best fit that minimises the **sum of the squares** of the gap between the line and each data value
  - This is usually called the **regression line of *y* on *x***
  - It can be calculated by looking at the vertical distances between the line and the data values
- The **regression line of y on x** is written in the form $y = ax + b$
- *a* is the **gradient** of the line
  - It represents the change in *y* for each individual unit change in *x*
    - If *a* is **positive** this means *y* **increases** by *a* for a unit increase in *x*
    - If *a* is **negative** this means *y* **decreases** by |a| for a unit increase in *x*
- *b* is the ***y* – intercept**
  - It shows the value of *y* when *x* is zero
- You are expected to use your **GDC** to find the equation of the regression line
  - Enter the bivariate data and choose the **model "*ax* + *b*"**
  - Remember the **mean point** $(\overline{x},\ \overline{y})$ will lie on the regression line

How do I use a regression line?

- The equation of the regression line can be used to decide what type of correlation there is if there is no scatter diagram
  - If *a* is **positive** then the data set has **positive correlation**
  - If *a* is **negative** then the data set has **negative correlation**
- The equation of the regression line can also be used to **predict** the value of a **dependent variable (*y*)** from an **independent variable (*x*)**
  - The equation should **only be used** to make **predictions for *y***
    - Using a *y* on *x* line to **predict *x* is not always reliable**
  - Making a prediction **within the range** of the given data is called **interpolation**
    - This is usually reliable
    - The stronger the correlation the more reliable the prediction
  - Making a prediction **outside of the range** of the given data is called **extrapolation**
    - This is much less reliable
  - The prediction will be more reliable if the number of data values in the original sample set is bigger

Exam Tip

- Once you calculate the values of *a* and *b* store then in your GDC
  - This means you can use the full display values rather than the rounded values when using the linear regression equation to predict values
  - This avoids rounding errors

Worked Example

Barry is a music teacher. For 7 students, he records the time they spend practising per week (x hours) and their score in a test (y %).

Time (x)	2	5	6	7	10	11	12
Score (y)	11	49	55	75	63	68	82

a)

Write down the equation of the regression line of y on x, giving your answer in the form $y = ax + b$ where a and b are constants to be found.

Enter data into GDC

a is the coefficient of x $\quad a = 5.5680...$

b is the constant term $\quad b = 15.4136...$

$$\boxed{y = 5.57x + 15.4}$$

b)

Give an interpretation of the value of a.

$a = 5.57$ means that the model suggests that the score increases by 5.57 % for every extra hour of practice.

c)

Another of Barry's students practises for 15 hours a week, estimate their score. Comment on the validity of this prediction.

Substitute $x = 15$

$y = (5.5680...) \times 15 + (15.4136...) = 98.93..$

The model predicts a score of 98.9% but this is unreliable as $x = 15$ is outside the range of data. Therefore extrapolation is being used.

4.3 Further Correlation & Regression

4.3.1 Non-linear Regression

Non-linear Regression

What is non-linear regression?

- You have already seen that **linear regression** is when you can use a straight line to fit bivariate data
- **Non-linear regression** is when you can use a **curve** (rather than a straight line) to fit bivariate data
- In your exam the regression could be:
  - Linear: $y = ax + b$
  - Quadratic: $y = ax^2 + bx + c$
  - Cubic: $y = ax^3 + bx^2 + cx + d$
  - Exponential: $y = ab^x$ or $y = ae^{bx}$
  - Power: $y = ax^b$
  - Sine: $y = a\sin(bx + c) + d$

How do I find the equation of the non-linear regression model?

- Using your GDC:
  - Type the **two sets** of the data into your GDC
  - Select the **relevant model**
    - The exam question will tell you which model to use
  - Your GDC will calculate the **constants**
- You can use **logarithms** to **linearise exponential and power** relationships
  - Power: $y = ax^b$ then $\ln y = \ln a + b\ln x$
    - $\ln y$ and $\ln x$ will have a linear relationship
  - Exponential: $y = ab^x$ then $\ln y = \ln a + x\ln b$
    - $\ln y$ and x will have a linear relationship

Exam Tip

- You can use your GDC to plot the scatter diagram and include the graph of a regression model
  - This will allow you to get a sense of how well the model fits the data

Worked Example

Scarlett and Violet collect data on the length of a film (x minutes) and the audience rating (y %).

x	75	93	101	107	115	124	132	140	171
y	83	75	51	38	47	56	76	91	70

a)

Scarlett claims that there is a cubic relationship. Find the equation of a cubic regression model of the form $y = ax^3 + bx^2 + cx + d$.

Type the data into GDC and choose the cubic regression model

$a = -0.0005291...$ $b = 0.2030...$ $c = -24.89...$ $d = 1037.7...$

$$\boxed{y = -0.000529x^3 + 0.203x^2 - 24.9x + 1040}$$

b)

Violet claims that there is a sine relationship. Find the equation of a sine regression model of the form $y = a\sin(bx + c) + d$.

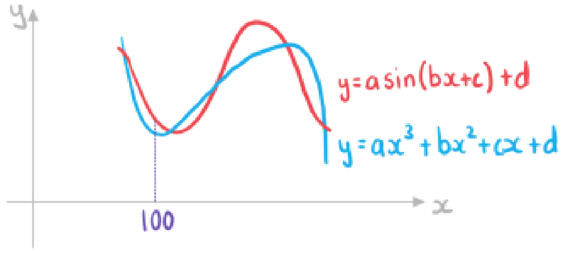
Type the data into GDC and choose the sine regression model

$a = 24.74...$ $b = 0.08030...$ $c = 2.086...$ $d = 69.49...$

$$\boxed{y = 24.7\sin(0.0803x + 2.09) + 69.5}$$

c)

Whose model predicts a higher audience rating for a film which is 100 minutes long?



Using the cubic model $y = 49.640...$

Using the sine model $y = 53.690...$

$\boxed{\text{Violet's model predicts a higher rating.}}$

Least Squares Regression Curves

What is a residual?

- Given a set of *n* pairs of data and a **regression model** *y* = *f*(*x*)
- A **residual** is the **actual *y*-value** (from the data) **minus** the **predicted *y*-value** (using the regression model)
    - $y_i - f(x_i)$
- The **sum of the square residuals** is denoted by SS_{res}
    - $SS_{res} = \sum_{i=1}^{n}(y_i - f(x_i))^2$
- If you have two regression models using the **same data** then the one with the **smaller** SS_{res} **fits the data better**

What is a least squares regression curve?

- The **least squares regression curve** can be thought of as a "**curve of best fit**" *y* = *f*(*x*)
- For a **given type of model** the least squares regression curve **minimises the sum of the square residuals**
    - Your GDC calculates the constants for the least squares regression curves

Why is the sum of the square residuals not always a good measure of fit?

- If two models are formed using the **same number of pairs** of data then the sum of the square residuals is a **good measure of fit**
- If two models use **different number of pairs** of data then SS_{res} is **not always a good measure of fit**
    - The sum will increase with more pairs of data and so can no longer be compared against a data set with a different number of pairs
    - Compare the two scenarios
        - 10 pairs of data and the absolute value of each residual is 15 then $SS_{res} = 10 \times 15^2 = 2250$
        - 2250 pairs of data and the absolute value of each residual is 1 then $SS_{res} = 2250 \times 1^2 = 2250$
    - They have the same value of SS_{res} but the residuals in the second scenario are much smaller
- Your GDC may give you the **mean squared error**
    - $MSe = \frac{1}{n}SS_{res} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$
    - This is a **better measure of fit**
    - You **do not need to know this** for your exam but it might help with your understanding

Worked Example

Jet is the owner of a gym and he is testing different prices options. The table below shows the number of new members per month (M) and the price of a monthly membership (£p).

p	10	20	30
M	97	68	55

Jet believes that he can fit the data with either the model $M_1(p) = \dfrac{2700}{p+20}$ or the model $M_2(p) = \dfrac{2100}{p+10}$.

Jet wants to choose the model with the smallest value for the sum of square residuals.

Determine which model Jet should choose.

Calculate the predicted values

p	M	M_1	M_2
10	97	90	105
20	68	67.5	70
30	55	54	52.5

For M_1 : $SS_{res} = (97-90)^2 + (68-67.5)^2 + (55-54)^2 = 50.25$

For M_2 : $SS_{res} = (97-105)^2 + (68-70)^2 + (55-52.5)^2 = 74.25$

$\boxed{\text{Jet should choose model } M_1}$

The Coefficient of Determination

What is the coefficient of determination?

- The **coefficient of determination** is a **measure of fit** for a model
  - If the coefficient of determination is 0.57 this means 57% of the variation of the *y*-variable can be explained by the variation in the *x*-variable
  - The other 43% can be explained by other factors
  - The higher this proportion the more the model fits the data
- The coefficient of determination is **denoted by** R^2
  - $R^2 \le 1$
  - $R^2 = 1$ means the model is a **perfect fit** for the data
  - The closer to 1 the better the fit
  - R^2 is usually greater than or equal to zero
    - R^2 can be negative but this is outside the scope of this course
- If the regression model is linear then the coefficient of determination is **equal to square of the PMCC**
  - $R^2 = r^2$ for linear models
  - Some GDCs will simply denote R^2 as r^2 due to its connection to the PMCC for linear models

How do I calculate the coefficient of determination?

- When finding the constants for regression models your **GDC might give you the value** of R^2
  - You will only be asked to calculate the coefficient of determination for models for which GDCs give the value of R^2
- The coefficient of determination can be calculated by
  - $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$
    - Where $SS_{tot} = \sum_{i=1}^{n} (y_i - \bar{y})^2$
  - You **do not need to know this** formula but it might help with your understanding

Does the coefficient of determination determine the validity of a model?

- If R^2 is close to 1 then the model fits the data well
  - However this alone **does not guarantee** that it is a **good model for the relationship** between the two variables
- Consider the scenario where there are big gaps between data points and a model which fits the data well
  - The model only fits the data at the data points
  - As there are gaps between the data points the model might not be a good fit for these areas
- Different types of models have **different number of parameters**
  - Therefore using different types of models to fit the same data will have **different levels of accuracy**
  - Linear models need **at least two pairs** of data

- Quadratic models need **at least three pairs** of data
- Cubic models need **at least four pairs** of data
  - Using four pairs of data will mean the cubic model will have $R^2 = 1$
    This is because the cubic graph will go through all four pieces of data – the value is likely to decrease as extra pairs of data are included
  - However this does not mean it is a better fit than the quadratic model
  - The quadratic model could be more accurate as it has one more pair of data than is needed

Worked Example

Data is collected on the lengths of cheetahs (x metres) and their average running speeds (y ms^{-1}).

x	1.21	1.33	1.12	1.45	1.42	1.39	1.24	1.19	1.32
y	24.3	25.1	22.2	35.1	35.1	33.4	27.1	23.1	24.8

a)

Find the equation of the least squares regression curve using:

(i)

a quadratic model $y = ax^2 + bx + c$.

(ii)

an exponential model $y = ab^x$.

Type the data into GDC and choose the :

quadratic regression model

$a = 140.9...$

$b = -322.6...$

$c = 207.5...$

$\boxed{y = 141x^2 - 323x + 208}$

exponential regression model

$a = 4.193...$

$b = 4.250...$

$\boxed{y = 4.19 \times 4.25^x}$

b)

Based solely on the coefficients of determination, suggest which model is better fit for the data.

Find the coefficients of determination using GDC

Quadratic $R^2 = 0.86429...$

Exponential $R^2 = 0.80157...$

Based on the coefficients of determination, the quadratic regression model as its R^2 value is bigger.

4.3.2 Logarithmic Scales

Logarithmic Scales

What are logarithmic scales?

- **Logarithmic scales** are scales where intervals **increase exponentially**
  - A normal scale might go 1, 2, 3, 4, ...
  - A logarithmic scale might go 1, 10, 100, 1000, ...
- Sometimes we can keep the scales with **constant intervals** by **changing the variables**
  - If the values of x increase exponentially: 1, 10, 100, 1000, ...
  - Then you can use the variable **log *x*** instead which will have the scale: 1, 2, 3, 4, ...
  - This will change the shape of the graph
    - If the graph transforms to a straight line then it is easier to analyse
- **Any base** can be used for logarithmic scales
  - The most common bases are 10 and e

Why do we use logarithmic scales?

- For variables that have a **large range** it can be difficult to plot on one graph
  - Especially when a lot of the values are **clustered in one region**
  - For example: populations of countries
    - This can range from 800 to 1 450 000 000
- If we are interested in the **rate of growth** of a variable rather than the actual values then a logarithmic scale is useful

log-log & semi-log Graphs

What is a log-log graph?

- A **log-log graph** is used when **both scales** of the original graph are logarithmic
    - You transform both variables by taking logarithms of the values
- log *y* & log *x* will be used instead of *y* & *x*
- **Power graphs** ($y = ax^b$) look like **straight lines** on log-log graphs

What is a semi-log graph?

- A **semi-log graph** is used when **only one scale** (the *y*-axis) of the original graph are logarithmic
    - You transform only the *y*-variable by taking logarithms of those values
- log *y* will be used instead of *y*
- **Exponential graphs** ($y = ab^x$) look like **straight lines** on semi-log graphs

How can I estimate values using log-log and semi-log graphs?

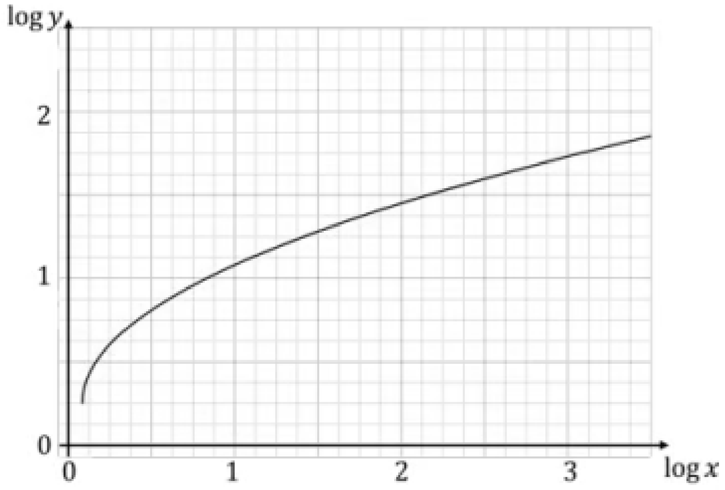
- Identify whether **one or both of the scales** are logarithmic
- Identify the variable so that the scales have **equal intervals**
    - *x* : 1, 10, 100, 1000, ... use log *x*
    - For *x* : 1, e, e², e^3, ... use ln *x*
- If you are asked to estimate a value:
    - First find the value of any logarithms
        - For example: log *y*, ln *x*, etc
    - Use the graph to read off the value
    - If it is a value for a logarithm find the actual value using:
        - $\log x = k \Rightarrow x = 10^k$
        - $\ln x = k \Rightarrow x = \mathrm{e}^k$

Exam Tip

- Pay close attention to which base is being used (log or ln)

Worked Example

The function $y = f(x)$ is drawn below using a log-log graph.

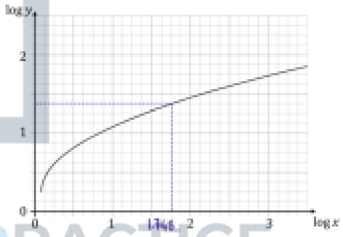


Show that when $x = 56$ the value of y is approximately 24.

Find $\log x$

$x = 56 \Rightarrow \log 56 = 1.7481...$

Use graph to find $\log y$



$\log y \approx 1.375$

Find y

Exponents & logarithms	$a^x = b \Leftrightarrow x = \log_a b$

$$\boxed{y = 10^{1.375} = 23.71... \approx 24}$$

4.3.3 Linearising using Logarithms

Exponential Relationships

How do I use logarithms to linearise exponential relationships?

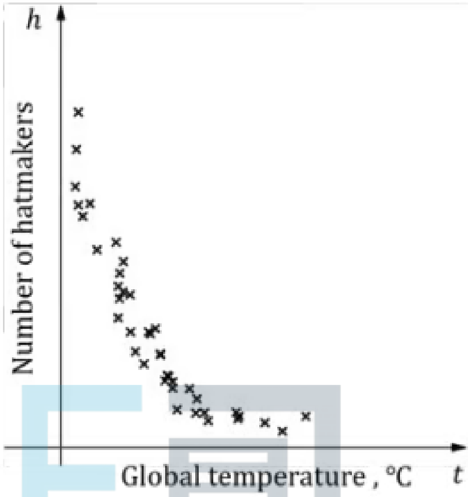
- Graphs of **exponential functions** appear as straight lines on **semi-log graphs**
- Suppose $y = ab^x$
  - You can take logarithms of both sides
    - $\ln y = \ln(ab^x)$
  - You can split the right hand side into the sum of two logarithms
    - $\ln y = \ln a + \ln(b^x)$
  - You can bring down the power in the final term
    - $\ln y = \ln a + x\ln b$
- $\ln y = \ln a + x\ln b$ is in linear form $Y = mX + c$
  - $Y = \ln y$
  - $X = x$
  - $m = \ln b$
  - $c = \ln a$

How can I use linearised data to find the values of the parameters in an exponential model *y* = *ab*^{x}?

- **STEP 1**: **Linearise** the data using $Y = \ln y$ and $X = x$
- **STEP 2**: Find the equation of the **regression line** of *Y* on *X*: $Y = mX + c$
- **STEP 3**: **Equate coefficients** between $Y = mX + c$ and $\ln y = \ln a + x\ln b$
  - $m = \ln b$
  - $c = \ln a$
- **STEP 4**: **Solve** to find *a* and *b*
  - $a = e^c$
  - $b = e^m$

Worked Example

Hatter has noticed that over the past 50 years there seems to be fewer hatmakers in London. He also knows that global temperatures have been rising over the same time period. He decides to see if there could be any correlation, so he collects data on the number of hatmakers and the global mean temperatures from the past 50 years and records the information in the graph below.



Hatter suggests that the equation for h in terms of t can be written in the form
$h = ab^t$
. He linearises the data using $x = t$ and $y = \ln h$ and calculates the regression line of y on x to be $y = 4.382 - 1.005x$.

Find the values of a and b.

Write $h = ab^t$ in linearised form

$\ln(h) = \ln(ab^t) \Rightarrow \ln h = \boxed{\ln a} + t\ln b$

Compare coefficients

$y = 4.382 - 1.005x \Rightarrow \ln h = \boxed{4.382} - 1.005t$

$\ln a = 4.382 \Rightarrow a = e^{4.382} = 79.997\ldots$ $\boxed{a = 80.0 \ (3sf)}$

$\ln b = -1.005 \Rightarrow b = e^{-1.005} = 0.36604\ldots$ $\boxed{b = 0.366 \ (3sf)}$

Power Relationships

How do I use logarithms to linearise power relationships?

- Graphs of **power functions** appear as straight lines on **log-log graphs**
- Suppose $y = ax^b$
  - You can take logarithms of both sides
    - $\ln y = \ln(ax^b)$
  - You can split the right hand side into the sum of two logarithms
    - $\ln y = \ln a + \ln(x^b)$
  - You can bring down the power in the final term
    - $\ln y = \ln a + b\ln x$
- $\ln y = \ln a + b\ln x$ is in linear form $Y = mX + c$
  - $Y = \ln y$
  - $X = \ln x$
  - $m = b$
  - $c = \ln a$

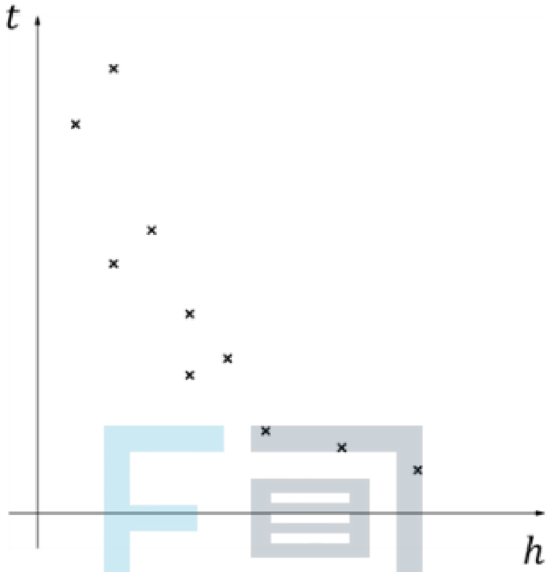
How can I use linearised data to find the values of the parameters in an power model *y* = *ax*^{b}?

- **STEP 1**: **Linearise** the data using $Y = \ln y$ and $X = \ln x$
- **STEP 2**: Find the equation of the **regression line** of *Y* on *X*: $Y = mX + c$
- **STEP 3**: **Equate coefficients** between $Y = mX + c$ and $\ln y = \ln a + b\ln x$
  - $m = b$
  - $c = \ln a$
- **STEP 4**: **Solve** to find *a* and *b*
  - $a = e^c$
  - $b = m$

Worked Example

The graph below shows the heights, h metres, and the amount of time spent sleeping, t hours, of a group of young giraffes. It is believed the data can be modelled using $t = ah^b$

.



The data are coded using the changes of variables $x = \ln h$ and $y = \ln t$. The regression line of y on x is found to be $y = 0.3 - 1.2x$.

Find the values of a and b.

Write $t = ah^b$ in linearised form

$\ln(t) = \ln(ah^b) \Rightarrow \ln t = \boxed{\ln a} + b \ln h$

Compare coefficients

$y = 0.3 - 1.2x \Rightarrow \ln t = \boxed{0.3} - 1.2 \ln h$

$\ln a = 0.3 \Rightarrow a = e^{0.3} = 1.3498...$ $\boxed{a = 1.35 \ (3sf)}$

$\boxed{b = -1.2}$

4.4 Probability

4.4.1 Probability & Types of Events

Probability Basics

What key words and terminology are used with probability?

- An **experiment** is a repeatable activity that has a result that can be observed or recorded
  - **Trials** are what we call the repeats of the experiment
- An **outcome** is a possible result of a trial
- An **event** is an outcome or a collection of outcomes
  - Events are usually denoted with capital letters: *A*, *B*, etc
  - *n*(*A*) is the number of outcomes that are included in event *A*
  - An event can have one or more than one outcome
- A **sample space** is the set of all possible outcomes of an experiment
  - This is denoted by *U*
  - *n*(*U*) is the total number of outcomes
  - It can be represented as a **list** or a **table**

How do I calculate basic probabilities?

- If all outcomes are **equally likely** then probability for each outcome is the same
  - Probability for each outcome is $\frac{1}{n(U)}$
- **Theoretical probability** of an event can be calculated without using an experiment by dividing the number of outcomes of that event by the total number of outcomes

$$P(A)=\frac{n(A)}{n(U)}$$

  - This is given in the **formula booklet**
  - Identifying all possible outcomes either as a list or a table can help
- **Experimental probability** (also known as **relative frequency**) of an outcome can be calculated using results from an experiment by dividing its frequency by the number of trials
  - **Relative frequency** of an outcome is $\frac{\text{Frequency of that outcome from the trials}}{\text{Total number of trials }(n)}$

How do I calculate the expected number of occurrences of an outcome?

- **Theoretical probability** can be used to calculate the **expected number of occurrences** of an outcome from *n* trials
- If the probability of an outcome is *p* and there are *n* trials then:
  - The expected number of occurrences is ***np***
  - This **does not mean** that there will **exactly *np* occurrences**
  - If the experiment is repeated multiple times then we expect the number of occurrences to average out to be *np*

What is the complement of an event?

- The probabilities of all the outcomes **add up to 1**
- Complementary events are when there are **two events** and **exactly one** of them will occur
  - One event has to occur but both events can not occur at the same time
- The **complement of event *A*** is the event where event ***A* does not happen**

  - This can be thought of as **not *A***
  - This is denoted *A'*

$$\mathrm{P}(A)+\mathrm{P}(A')=1$$

    - This is in the **formula booklet**
    - It is commonly written as $\mathrm{P}(A')=1-\mathrm{P}(A)$

What are different types of combined events?

- The **intersection** of two events (*A* and *B*) is the event where **both *A* and *B*** occur
  - This can be thought of as ***A* and *B***
  - This is denoted as $A\cap B$
- The **union** of two events (*A* and *B*) is the event where ***A* or *B* or both occur**
  - This can be thought of as ***A* or *B***
  - This is denoted $A\cup B$
- The event where *A* occurs given that event *B* has occurred is called **conditional probability**
  - This can be thought as ***A* given *B***
  - This is denoted $A|B$

How do I find the probability of combined events?

- The probability of *A* **or** *B* (or both) occurring can be found using the formula

$$\mathrm{P}(A\cup B)=\mathrm{P}(A)+\mathrm{P}(B)-\mathrm{P}(A\cap B)$$

    - This is given in the **formula booklet**
    - You subtract the probability of *A* and *B* both occurring because it has been included twice (once in P(*A*) and once in P(*B*))
- The probability of *A* **and** *B* occurring can be found using the formula

$$\mathrm{P}(A\cap B)=\mathrm{P}(A)\mathrm{P}(B|A)$$

    - A rearranged version is given in the **formula booklet**
    - Basically you multiply the probability of *A* by the probability of *B* then happening

Exam Tip

- In an exam drawing a Venn diagram or tree diagram can help even if the question does not ask you to

Worked Example

Dave has two fair spinners, A and B. Spinner A has three sides numbered 1, 4, 9 and spinner B has four sides numbered 2, 3, 5, 7. Dave spins both spinners and forms a two-digit number by using the spinner A for the first digit and spinner B for the second digit.

T is the event that the two-digit number is a multiple of 3.

a)

List all the possible two-digit numbers.

A two-way table would be a systematic way to list all the outcomes

	2	3	5	7
1	12	13	15	17
4	42	43	45	47
9	92	93	95	97

b)

Find $\mathrm{P}(T)$.

$$P(T) = \frac{n(T) \leftarrow \text{Number of multiples of 3}}{n(U) \leftarrow \text{Total number of outcomes}}$$

{12, 15, 42, 45, 93} are the multiples of 3

$$\boxed{P(T) = \frac{5}{12}}$$

c)

Find $\mathrm{P}(T')$.

$$P(T) + P(T') = 1 \Rightarrow P(T') = 1 - P(T)$$

$$P(T') = 1 - \frac{5}{12}$$

$$\boxed{P(T') = \frac{7}{12}}$$

Independent & Mutually Exclusive Events

What are mutually exclusive events?

- Two events are **mutually exclusive** if they **cannot both occur**
  - For example: when rolling a dice the events "getting a prime number" and "getting a 6" are mutually exclusive
- If *A* and *B* are mutually exclusive events then:
  - $\mathrm{P}(A \cap B) = 0$

What are independent events?

- Two events are **independent** if **one occurring does not affect the probability of the other occurring**
  - For example: when flipping a coin twice the events "getting a tails on the first flip" and "getting a tails on the second flip" are independent
- If *A* and *B* are independent events then:
  - $\mathrm{P}(A | B) = \mathrm{P}(A)$ and $\mathrm{P}(B|A) = \mathrm{P}(B)$
- If *A* and *B* are independent events then:
  - $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$
    - This is given in the **formula booklet**
    - This is a useful formula to test whether two events are statistically independent

How do I find the probability of combined mutually exclusive events?

- If *A* and *B* are **mutually exclusive** events then

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B)$$

  - This is given in the **formula booklet**
  - This occurs because $\mathrm{P}(A \cap B) = 0$
- For any two events *A* and *B* the events $A \cap B$ and $A \cap B'$ are **mutually exclusive** and *A* is the **union** of these two events
  - $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B')$
    - This works for any two events *A* and *B*

Worked Example

a)

A student is chosen at random from a class. The probability that they have a dog is 0.8, the probability they have a cat is 0.6 and the probability that they have a cat or a dog is 0.9.

Find the probability that the student has both a dog and a cat.

Let D be event "has a dog" and C be "has a cat"

$$P(D \cup C) = P(D) + P(C) - P(D \cap C)$$

$$0.9 = 0.8 + 0.6 - P(D \cap C)$$

$$\boxed{P(D \cap C) = 0.5}$$

b)

Two events, Q and R, are such that $\mathrm{P}(Q) = 0.8$ and $\mathrm{P}(Q \cap R) = 0.1$.

Given that Q and R are independent, find $\mathrm{P}(R)$.

Q and R independent $\Rightarrow P(Q \cap R) = P(Q)P(R)$

$$0.1 = 0.8 \times P(R) \quad \therefore \quad P(R) = \frac{0.1}{0.8}$$

$$\boxed{P(R) = 0.125 \quad \text{or} \quad \frac{1}{8}}$$

c)

Two events, S and T, are such that $\mathrm{P}(S) = 2\mathrm{P}(T)$.

Given that S and T are mutually exclusive and that $\mathrm{P}(S \cup T) = 0.6$ find $\mathrm{P}(S)$ and $\mathrm{P}(T)$.

S and T mutually exclusive $\Rightarrow P(S \cup T) = P(S) + P(T)$

$$0.6 = P(S) + P(T)$$

$$0.6 = 2P(T) + P(T) \qquad P(S) = 2P(T)$$

$$0.6 = 3P(T)$$

$$\boxed{P(T) = 0.2 \quad \text{and} \quad P(S) = 0.4}$$

4.4.2 Conditional Probability

Conditional Probability

What is conditional probability?

- **Conditional probability** is where the probability of an **event** happening can vary depending on the outcome of a prior event
- The event A happening **given that** event B has happened is denoted $A|B$
- A common example of conditional probability involves selecting multiple objects from a bag **without replacement**
  - The probability of selecting a certain item changes depending on what was selected before
    - This is because the total number of items will change as they are not replaced once they have been selected

How do I calculate conditional probabilities?

- Some conditional probabilities can be calculated by using counting outcomes
  - Probabilities without replacement can be calculated like this
  - For example: There are 10 balls in a bag, 6 of them are red, two of them are selected without replacement
    - To find the probability that the second ball selected is red given that the first one is red count how many balls are left:
    - A red one has already been selected so there are 9 balls left and 5 are red so the probability is $\frac{5}{9}$
- You can use sample space diagrams to find the probability of A given B:
  - reduce your sample space to just include outcomes for event B
  - find the proportion that also contains outcomes for event A
- There is a formula for conditional probability that you can use
  - $\mathrm{P}(A|B)=\dfrac{\mathrm{P}(A\cap B)}{\mathrm{P}(B)}$
  - This is given in the **formula booklet**
  - This can be rearranged to give $\mathrm{P}(A\cap B)=\mathrm{P}(B)\mathrm{P}(A|B)$

Worked Example

In a class of 30 students: 19 students have a dog, 17 students have a cat and 11 have both a dog and a cat. One student is selected at random.

a)

Find the probability that the student has a dog.

Let D be event "has a dog" and C be "has a cat"

$$P(D) = \frac{n(D) \leftarrow \text{Number who have dogs}}{n(U) \leftarrow \text{Total number of students}}$$

$$\boxed{P(D) = \frac{19}{30}}$$

b)

Find the probability that the student has a dog given that they have a cat.

17 have a cat of which 11 also have a dog

$$\boxed{P(D|C) = \frac{11}{17}} \quad \text{Could also use} \quad P(D|C) = \frac{P(D \cap C)}{P(C)}$$

c)

Find the probability that the student has a cat given that they have a dog.

19 have a dog of which 11 also have a cat

$$\boxed{P(C|D) = \frac{11}{19}} \quad \text{Could also use} \quad P(C|D) = \frac{P(C \cap D)}{P(D)}$$

4.4.3 Sample Space Diagrams

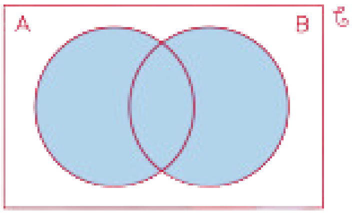
Venn Diagrams

What is a Venn diagram?

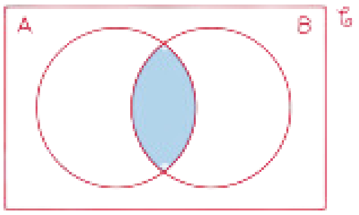
- A Venn diagram is a way to illustrate **events** from an **experiment** and are particularly useful when there is an overlap between possible **outcomes**
- A Venn diagram consists of
  - a **rectangle** representing the **sample space (*U*)**
    - The rectangle is labelled *U*
    - Some mathematicians instead use S or ξ
  - a **circle** for each **event**
    - Circles may or may not overlap depending on which **outcomes** are shared between **events**
- The numbers in the circles represent either the **frequency** of that event or the **probability** of that event
  - If the **frequencies** are used then they should **add up to the total frequency**
  - If the **probabilities** are used then they should **add up to 1**

What do the different regions mean on a Venn diagram?

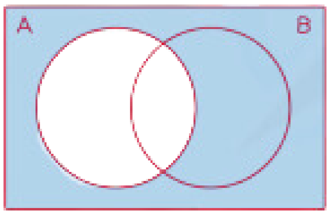
- A' is represented by the regions that are **not in** the *A* circle
- $A \cap B$ is represented by the region where the *A* and *B* circles **overlap**
- $A \cup B$ is represented by the regions that **are in** *A* or *B* or both
- Venn diagrams show '**AND**' and '**OR**' statements easily
- Venn diagrams also instantly show **mutually exclusive** events as these circles will **not overlap**
- **Independent** events can not be instantly seen
  - You need to use probabilities to deduce if two events are independent



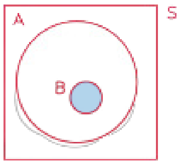
A ∪ B (UNION)
"A OR B OR BOTH"



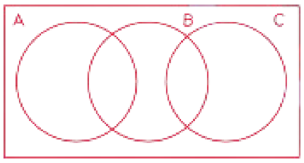
A ∩ B (INTERSECTION)
"A AND B"



A' (COMPLEMENT)
"NOT A"



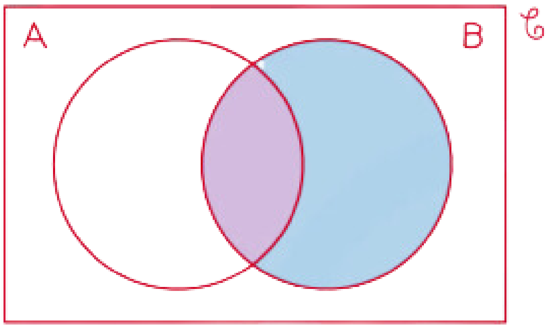
THE BUBBLE FOR EVENT B LIES ENTIRELY IN THE BUBBLE FOR EVENT A IF EVENT B OCCURS, SO DOES EVENT A (BUT NOT NECESSARILY VICE VERSA)



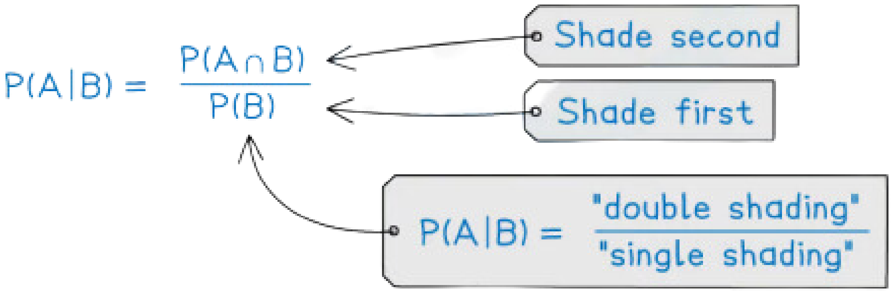
THE BUBBLES FOR EVENTS A AND C DO NOT OVERLAP: THEY ARE MUTUALLY EXCLUSIVE

How do I solve probability problems involving Venn diagrams?

- Draw, or add to a given Venn diagram, filling in as many values as possible from the information provided in the question
- It is usually helpful to work from the centre outwards
  - Fill in **intersections** (overlaps) first
- If two events are independent you can use the formula
  - $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$
- To find the conditional probability $\mathrm{P}(A|B)$
  - Add together the frequencies/probabilities in the *B* circle
    - This is your denominator
  - Out of those frequencies/probabilities add together the ones that are also in the *A* circle
    - This is your numerator
  - Evaluate the fraction



Event A|B
"A given B"



$$P(A|B) = \frac{P(A\cap B)}{P(B)}$$

Shade second

Shade first

$$P(A|B) = \frac{\text{"double shading"}}{\text{"single shading"}}$$

Exam Tip

- If you struggle to fill in a Venn diagram in an exam:
    - Label the missing parts using algebra
    - Form equations using known facts such as:
        - the sum of the probabilities should be 1
        - $P(A \cap B) = P(A)P(B)$ if A and B are independent events

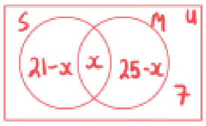
Worked Example

40 people are asked if they have sugar and/or milk in their coffee. 21 people have sugar, 25 people have milk and 7 people have neither.

a)

Draw a Venn diagram to represent the information.

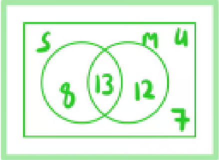
Find the centre first



Total should be 40

$(21-x)+x+(25-x)+7=40$

$53-x=40 \quad \therefore \; x=13$



b)

One of the 40 people are randomly selected, find the probability that they have sugar but not milk with their coffee.

S and not M is the part of S circle that does not include M

$P(S \cap M') = \frac{8}{40}$

Remember to write as a fraction of the total

$\boxed{P(S \cap M') = \frac{1}{5}}$

c)

Given that a person who has sugar is selected at random, find the probability that they have milk with their coffee.

Given that sugar has been selected we only want the S circle as our total.

Out of the S circle 13 also have milk

$\boxed{P(M|S) = \frac{13}{21}}$

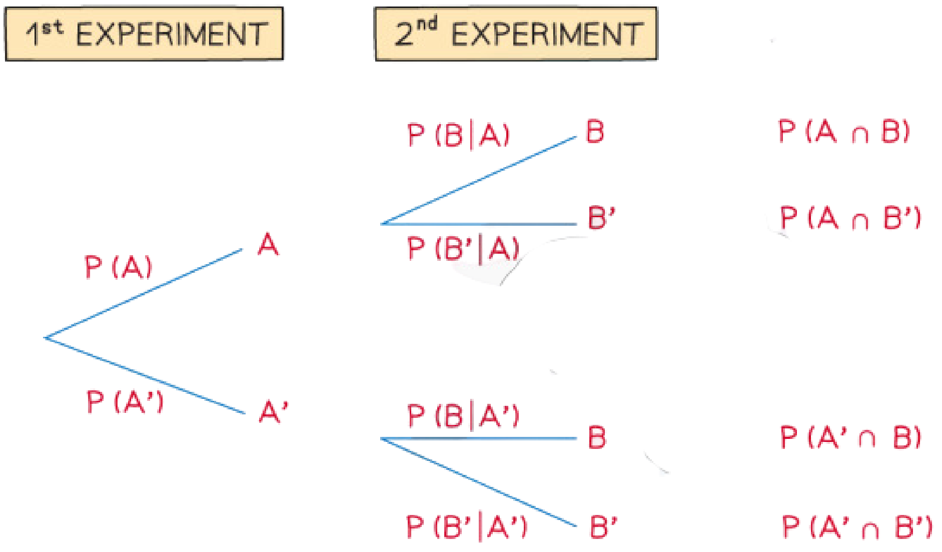
Tree Diagrams

What is a tree diagram?

- A **tree diagram** is another way to show the outcomes of combined events
  - They are very useful for intersections of events
- The events on the branches must be **mutually exclusive**
  - Usually they are an event and its complement
- The probabilities on the second sets of branches **can depend** on the outcome of the first event
  - These are **conditional probabilities**
- When selecting the items from a bag:
  - The second set of branches will be the **same** as the first if the items **are replaced**
  - The second set of branches will be the **different** to the first if the items **are not replaced**

How are probabilities calculated using a tree diagram?

- To find the probability that two events happen together you **multiply** the corresponding probabilities on their branches
  - It is helpful to find the probability of all combined outcomes once you have drawn the tree
- To find the probability of an event you can:
  - **add together** the probabilities of the **combined outcomes** that are part of that event
    - For example: $\mathrm{P}(A \cup B) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B') + \mathrm{P}(A' \cap B)$
  - **subtract** the probabilities of the combined outcomes that are not part of that event from 1
    - For example: $\mathrm{P}(A \cup B) = 1 - \mathrm{P}(A' \cap B')$



Do I have to use a tree diagram?

- If there are **multiple events** or trials then a tree diagram can get big
- You can break down the problem by using the words **AND/OR/NOT** to help you find probabilities without a tree

- You can speed up the process by only drawing parts of the tree that you are interested in

Which events do I put on the first branch?

- If the events *A* and *B* are **independent** then the **order does not matter**
- If the events *A* and *B* are **not independent** then the **order does matter**
  - If you have the probability of ***A* given *B*** then put ***B* on the first set** of branches
  - If you have the probability of ***B* given *A*** then put ***A* on the first set** of branches

Exam Tip

- In an exam do not waste time drawing a full tree diagram for scenarios with lots of events unless the question asks you to
  - Only draw the parts that you are interested in

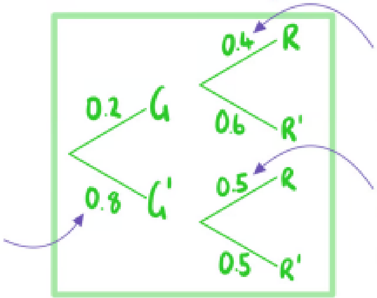
Worked Example

20% of people in a company wear glasses. 40% of people in the company who wear glasses are right-handed. 50% of people in the company who don't wear glasses are right-handed.

a)

Draw a tree diagram to represent the information.

Let G be the event "wears glasses" and R be "is right-handed"



40% of people who wear glasses are right-handed

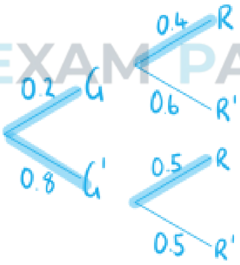
50% of people who don't wear glasses are right-handed

Branches add to 1

b)

One of the people in the company are randomly selected, find the probability that they are right-handed.

Find options that contain R



$P(G \cap R) = 0.2 \times 0.4 = 0.08$

Multiply along branches

$P(G' \cap R) = 0.8 \times 0.5 = 0.4$

$$P(R) = P(G \cap R) + P(G \cap R') = 0.08 + 0.4$$

$$\boxed{P(R) = 0.48}$$

c)

Given that a person who is right-handed is selected at random, find the probability that they wear glasses.

$$P(G|R) = \frac{P(G \cap R)}{P(R)} = \frac{0.08}{0.48}$$

$$\boxed{P(G|R) = \frac{1}{6}}$$

4.5 Probability Distributions

4.5.1 Discrete Probability Distributions

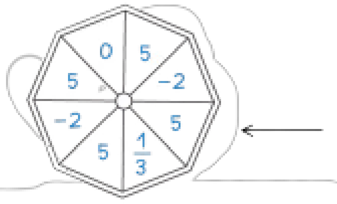
Discrete Probability Distributions

What is a discrete random variable?

- A **random variable** is a variable whose value depends on the outcome of a **random event**
  - The value of the random variable is not known until the event is carried out (this is what is meant by 'random' in this case)
- **Random variables** are denoted using **upper case letters** (X, Y, etc)
- **Particular outcomes** of the event are denoted using **lower case letters** (x, y, etc)
- $\mathrm{P}(X=x)$ means "the probability of the random variable X taking the value x"
- A **discrete** random variable (often abbreviated to DRV) can only take **certain values** within a set
  - Discrete random variables **usually count** something
  - Discrete random variables usually can only take a finite number of values but it is possible that it can take an infinite number of values (see the examples below)
- **Examples** of discrete random variables include:
  - The number of times a coin lands on heads when flipped 20 times
    - this has a finite number of outcomes: {0,1,2,...,20}
  - The number of emails a manager receives within an hour
    - this has an infinite number of outcomes: {1,2,3,...}
  - The number of times a dice is rolled until it lands on a 6
    - this has an infinite number of outcomes: {1,2,3,...}
  - The number that a dice lands on when rolled once
    - this has a finite number of outcomes: {1,2,3,4,5,6}

What is a probability distribution of a discrete random variable?

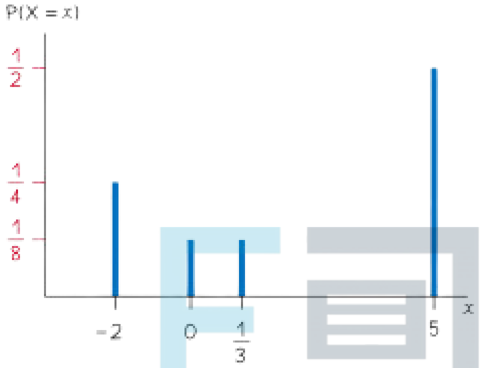
- A **discrete probability distribution** fully describes **all the values** that a discrete random variable can take along with their **associated probabilities**
  - This can be given in a **table**
  - Or it can be given as a **function** (called a discrete probability distribution function or "pdf")
  - They can be represented by **vertical line graphs** (the possible values for along the horizontal axis and the probability on the vertical axis)
- The **sum of the probabilities** of **all the values** of a discrete random variable is **1**
  - This is usually written $\sum \mathrm{P}(X=x)=1$
- A **discrete uniform distribution** is one where the random variable takes a finite number of values each with an **equal probability**
  - If there are n values then the probability of each one is $\frac{1}{n}$



LET x BE THE NUMBER THAT THE SPINNER LANDS ON

x	-2	0	$\frac{1}{3}$	5
$P(X=x)$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{8}$	$\frac{1}{2}$

$$P(X=x) = \begin{cases} \frac{1}{8} & x = 0, \frac{1}{3} \\ \frac{1}{4} & x = -2 \\ \frac{1}{2} & x = 5 \\ 0 & \text{OTHERWISE} \end{cases}$$



How do I calculate probabilities using a discrete probability distribution?

- First **draw a table** to represent the probability distribution
  - If it is given as a function then find each probability
  - If any probabilities are unknown then use algebra to represent them
- **Form an equation** using $\sum P(X=x)=1$
  - Add together all the probabilities and make the sum equal to 1
- To find $P(X=k)$
  - If k is a possible value of the random variable X then $P(X=k)$ will be given in the table
  - If k is not a possible value then $P(X=k)=0$
- To find $P(X\le k)$
  - Identify all possible values, x_i, that X can take which satisfy $x_i \le k$
  - Add together all their corresponding probabilities
  - $P(X\le k)=\sum_{x_i\le k} P(X=x_i)$
  - Some mathematicians use the notation $F(x)$ to represent the cumulative distribution
    - $F(x)=P(X\le x)$
- Using a similar method you can find $P(X<k)$, $P(X>k)$ and $P(X\ge k)$
- As all the probabilities add up to 1 you can form the following equivalent equations:
  - $P(X<k)+P(X=k)+P(X>k)=1$
  - $P(X>k)=1-P(X\le k)$
  - $P(X\ge k)=1-P(X<k)$

How do I know which inequality to use?

- $P(X\le k)$ would be used for phrases such as:
  - At most , no greater than , etc

- $\mathrm{P}(X < k)$ would be used for phrases such as:
  - Fewer than
- $\mathrm{P}(X \geq k)$ would be used for phrases such as:
  - At least , no fewer than , etc
- $\mathrm{P}(X > k)$ would be used for phrases such as:
  - Greater than , etc

Worked Example

The probability distribution of the discrete random variable X is given by the function

$$\mathrm{P}(X = x) = \begin{cases} kx^2 & x = -3,\ -1,\ 2,\ 4 \\ 0 & \text{otherwise.} \end{cases}$$

a)

Show that $k = \dfrac{1}{30}$.

Construct a table

x	-3	-1	2	4
$\mathrm{P}(X=x)$	$9k$	k	$4k$	$16k$

Substitute in the values of x

e.g. $\mathrm{P}(X=-3) = k(-3)^2 = 9k$

The probabilities add up to 1

$9k + k + 4k + 16k = 1$

$30k = 1$

$$\boxed{k = \frac{1}{30}}$$

b)

Calculate $\mathrm{P}(X \leq 3)$.

Substitute k into the probabilities

x	-3	-1	2	4
$\mathrm{P}(X=x)$	$\frac{3}{10}$	$\frac{1}{30}$	$\frac{2}{15}$	$\frac{8}{15}$

$X \leq 3 : X = -3, -1, 2$

$\mathrm{P}(X \leq 3) = \mathrm{P}(X=-3) + \mathrm{P}(X=-1) + \mathrm{P}(X=2)$

$= \frac{3}{10} + \frac{1}{30} + \frac{2}{15}$

$$\boxed{\mathrm{P}(X \leq 3) = \frac{7}{15}}$$

4.5.2 Expected Values

Expected Values E(X)

What does E(X) mean and how do I calculate E(X)?

- **E(X)** means the **expected value** or the **mean** of a **random variable X**
  - The expected value does not need to be an obtainable value of *X*
  - For example: the expected value number of times a coin will land on tails when flipped 5 times is 2.5
- For a **discrete** random variable, it is calculated by:
  - **Multiplying each value** of X with its corresponding **probability**
  - **Adding** all these terms together

$$\mathrm{E}(X) = \sum x\mathrm{P}(X = x)$$

    - This is given in the **formula booklet**
- Look out for **symmetrical** distributions (where the values of X are symmetrical and their probabilities are symmetrical) as the mean of these is the same as the median
  - For example: if X can take the values 1, 5, 9 with probabilities 0.3, 0.4, 0.3 respectively then by symmetry the mean would be 5

How can I decide if a game is fair?

- Let *X* be the random variable that represents the **gain/loss** of a player in a game
  - *X* will be **negative** if there is a **loss**
- Normally the expected gain or loss is calculated by **subtracting** the **cost to play** the game from the **expected value** of the **prize**
- If E(*X*) is **positive** then it means the player can **expect to make a gain**
- If E(*X*) is **negative** then it means the player can **expect to make a loss**
- The game is called **fair** if the **expected gain is 0**
  - E(*X*) = 0

Worked Example

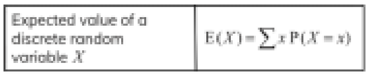
Daphne pays \$5 to play a game where she wins a prize of \$1, \$5, \$10 or \$100. The random variable W represents the amount she wins and has the probability distribution shown in the following table:

w	1	5	10	100
$P(W=w)$	0.35	0.5	0.05	0.01

a)

Calculate the expected value of Daphne's prize.

Formula booklet



Expected value of a discrete random variable X	$E(X)=\sum xP(X=x)$

$E(W) = \sum \omega P(W = \omega)$

$= 1 \times 0.35 + 5 \times 0.5 + 10 \times 0.05 + 100 \times 0.01$

$\boxed{\text{Expected value} = \$4.35}$

b)

Determine whether the game is fair.

A game is fair is expected gain/loss is 0

Prize - cost

$4.35 - 5 = -0.65$

$\boxed{\text{Expected loss is } \$0.65 \text{ so game is not fair}}$

4.6 Random Variables

4.6.1 Linear Combinations of Random Variables

Transformation of a Single Variable

What is Var(*X*)?

- Var(*X*) represents the variance of the random variable *X*
- Var(*X*) can be calculated by the formula
  - $\text{Var}(X) = \text{E}(X^2) - [\text{E}(X)]^2$
    - where $\text{E}(X^2) = \sum x^2 \text{P}(X = x)$
  - You will **not be required** to use this formula in the exam

What are the formulae for E(*aX* ± *b*) and Var(*aX* ± *b*)?

- If *a* and *b* are constants then the following formulae are true:
  - **E(*aX* ± b) = aE(*X*) ± b**
  - **Var(*aX* ± b) = a² Var(*X*)**
    - These are given in the **formula booklet**
- This is the same as linear transformations of data
  - The mean is affected by multiplication and addition/subtraction
  - The variance is affected by multiplication but not addition/subtraction
- Remember division can be written as a multiplication
  - $\frac{X}{a} = \frac{1}{a}X$

Worked Example

X is a random variable such that $\mathrm{E}(X) = 5$ and $\mathrm{Var}(X) = 4$.

Find the value of:

(i)

$\mathrm{E}(3X + 5)$

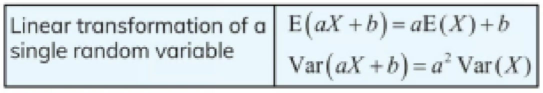
(ii)

$\mathrm{Var}(3X + 5)$

(iii)

$\mathrm{Var}(2 - X)$.

Formula booklet



Linear transformation of a single random variable	$\mathrm{E}(aX+b) = a\mathrm{E}(X)+b$ $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$

$\mathrm{E}(3X+5) = 3\mathrm{E}(X) + 5 = 3(5) + 5$ $\boxed{\mathrm{E}(3X+5) = 20}$

$\mathrm{Var}(3X+5) = 3^2\mathrm{Var}(X) = 9(4)$ $\boxed{\mathrm{Var}(3X+5) = 36}$

$\mathrm{Var}(2-X) = (-1)^2\,\mathrm{Var}(X) = 1(4)$ $\boxed{\mathrm{Var}(2-X) = 4}$

Transformation of Multiple Variables

What is the mean and variance of $aX + bY$?

- Let X and Y be two random variables and let a and b be two constants
- $\mathrm{E}(aX + bY) = a\mathrm{E}(X) + b\mathrm{E}(Y)$
  - This is true for **any random variables** X and Y
- $\mathrm{Var}(aX + bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$
  - This is true if X and Y are **independent**
- $\mathrm{E}(aX - bY) = a\mathrm{E}(X) - b\mathrm{E}(Y)$
- $\mathrm{Var}(aX - bY) = a^2\,\mathrm{Var}(X) + b^2\,\mathrm{Var}(Y)$
  - Notice that you still add the two terms together on the right hand side
    - This is because b^2 is positive even if b is negative
  - Therefore the variances of $aX + bY$ and $aX - bY$ are the same

What is the mean and variance of a linear combination of n random variables?

- Let $X_1, X_2, \ldots, X_n$ be n random variables and $a_1, a_2, \ldots, a_n$ be n constants

$$\mathrm{E}\left(a_1X_1 \pm a_2X_2 \pm \ \ldots \pm a_nX_n\right) = a_1\mathrm{E}\left(X_1\right) \pm a_2\mathrm{E}\left(X_2\right) \pm \ \ldots \pm a_n\mathrm{E}\left(X_n\right)$$

  - This is given in the **formula booklet**
  - This can be written as $\mathrm{E}\left(\sum a_iX_i\right) = \sum a_i\mathrm{E}\left(X_i\right)$
  - This is true for **any random variable**

$$\mathrm{Var}\left(a_1X_1 \pm a_2X_2 \pm \ \ldots \pm a_nX_n\right) = a_1{}^2\,\mathrm{Var}\left(X_1\right) + a_2{}^2\,\mathrm{Var}\left(X_2\right) + \ \ldots + a_n{}^2\,\mathrm{Var}\left(X_n\right)$$

  - This is given in the **formula booklet**
  - This can be written as $\mathrm{Var}\left(\sum a_iX_i\right) = \sum a_i^2\mathrm{Var}\left(X_i\right)$
  - This is true if the random variables are **independent**
    - Notice that the constants get squared so the terms on the right-hand side will always be positive

For a given random variable X, what is the difference between $2X$ and $X_1 + X_2$?

- **$2X$** means **one observation** of X is taken and **then doubled**
- **$X_1 + X_2$** means **two observations** of X are taken and then **added together**
- **$2X$** and **$X_1 + X_2$** have the **same expected values**
  - **$\mathrm{E}(2X) = 2\mathrm{E}(X)$**
  - **$\mathrm{E}(X_1 + X_2) = \mathrm{E}(X_1) + \mathrm{E}(X_2) = 2\mathrm{E}(X)$**
- **$2X$** and **$X_1 + X_2$** have **different variances**
  - **$\mathrm{Var}(2X) = 2^2\mathrm{Var}(X) = 4\mathrm{Var}(X)$**
  - **$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = 2\mathrm{Var}(X)$**
- To see the distinction:
  - Suppose X could take the values 0 and 1
    - $2X$ could then take the values 0 and 2
    - $X_1 + X_2$ could then take the values 0, 1 and 2

- Questions are likely to describe the variables in content
    - For example: The mass of a carton containing 6 eggs is the mass of the carton plus the mass of the 6 **individual** eggs
    - This can be modelled by $M = C + E_1 + E_2 + E_3 + E_4 + E_5 + E_6$ where
        - C is the mass of a carton
        - E is the mass of an egg
    - It is **not** $C + 6E$ because the masses of the 6 eggs could be **different**

Exam Tip

- In an exam when dealing with multiple variables ask yourself which of the two cases is true
    - You are adding together **different observations** using the same variable: $X_1 + X_2 + \ldots + X_n$
    - You are taking a **single observation** of a variable and multiplying it by a constant: nX

Worked Example

X and Y are independent random variables such that

$$\mathrm{E}(X) = 5 \text{ \& } \mathrm{Var}(X) = 3,$$

$$\mathrm{E}(Y) = -2 \text{ \& } \mathrm{Var}(Y) = 4.$$

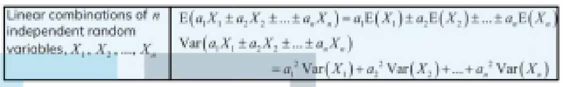
Find the value of:

(i)
$\mathrm{E}(2X + 5Y)$,

(ii)
$\mathrm{Var}(2X + 5Y)$,

(iii)
$\mathrm{Var}(4X - Y)$.

Formula booklet



Linear combinations of n independent random variables, $X_1, X_2, \ldots, X_n$	$\mathrm{E}(a_1X_1 \pm a_2X_2 \pm \ldots \pm a_nX_n) = a_1\mathrm{E}(X_1) \pm a_2\mathrm{E}(X_2) \pm \ldots \pm a_n\mathrm{E}(X_n$ $\mathrm{Var}(a_1X_1 \pm a_2X_2 \pm \ldots \pm a_nX_n)$ $= a_1^2\,\mathrm{Var}(X_1) + a_2^2\,\mathrm{Var}(X_2) + \ldots + a_n^2\,\mathrm{Var}(X_n)$

$E(2X+5Y) = 2E(X) + 5E(Y) = 2(5) + 5(-2)$ $\boxed{E(2X+5Y) = 0}$

$Var(2X+5Y) = 2^2 Var(X) + 5^2 Var(Y) = 4(3) + 25(4)$ $\boxed{Var(2X+5Y) = 112}$

$Var(4X-Y) = 4^2 Var(X) + Var(Y) = 16(3) + 4$ $\boxed{Var(4X-Y) = 52}$

4.6.2 Unbiased Estimates

Unbiased Estimates

What is an unbiased estimator of a population parameter?

- An **estimator** is a **random variable** that is used to **estimate a population parameter**
    - An **estimate** is the value produced by the estimator when a sample is used
- An estimator is called unbiased if its expected value is equal to the population parameter
    - An estimate from an unbiased estimator is called an **unbiased estimate**
    - This means that the **mean** of the **unbiased estimates** will get **closer** to the **population parameter** as **more samples** are taken

- The **sample mean** is an **unbiased estimate** for the **population mean**
- The **sample variance** is **not an unbiased estimate** for the **population variance**
    - On average the sample variance will **underestimate** the population variance
    - As the **sample size increases** the sample variance gets **closer to the unbiased estimate**

What are the formulae for unbiased estimates of the mean and variance of a population?

- A sample of *n* data values (x_1, x_2, ... etc) can be used to find unbiased estimates for the mean and variance of the population
- An unbiased estimate for the mean μ of a population can be calculated using
    - $\bar{x} = \dfrac{\sum x}{n}$

- An unbiased estimate for the variance σ^2 of a population can be calculated using
    - $s^2_{n-1} = \dfrac{n}{n-1} s^2_n$
    - This is given in the **formula booklet**
    - s^2_n is the variance of the sample data
        - $s^2_n = \dfrac{\sum (x-\bar{x})^2}{n} = \dfrac{\sum x^2}{n} - (\bar{x})^2$

- Different calculators can use different notations for s^2_{n-1}
    - σ^2_{n-1}, s^2, $\hat{s}^2$ are notations you might see
    - You may also see the square roots of these

Is s_{n-1} an unbiased estimate for the standard deviation?

- Unfortunately s_{n-1} is not an unbiased estimate for the standard deviation of the population
- It is better to work with the unbiased variance rather than standard deviation
- There is not a formula for an unbiased estimate for the standard deviation that works for all populations
    - Therefore you will not be asked to find one in your exam

How do I show the sample mean is an unbiased estimate for the population mean?

- You **do not need to learn this proof**
  - It is simply here to help with your understanding
- Suppose the population of X has mean μ and variance σ^2
- Take a sample of *n* observations
  - $X_1, X_2, \dots, X_n$
  - $\mathrm{E}(X_i) = \mu$
- Using the formula for a linear combination of *n* independent variables:

$$\begin{aligned}\mathrm{E}\left(\overline{X}\right) &= \mathrm{E}\left(\frac{X_1 + X_2 + \ \dots + X_n}{n}\right)\\ &= \frac{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ \dots + \mathrm{E}(X_n)}{n}\\ &= \frac{\mu + \mu + \ \dots + \mu}{n}\\ &= \frac{n\mu}{n}\\ &= \mu\end{aligned}$$

- As $\mathrm{E}\left(\overline{X}\right) = \mu$ this shows the formula will produce an **unbiased estimate** for the population mean

Why is there a divisor of *n*-1 in the unbiased estimate for the variance?

- You **do not need to learn this proof**
  - It is simply here to help with your understanding
- Suppose the population of X has mean μ and variance σ^2
- Take a sample of *n* observations
  - $X_1, X_2, \dots, X_n$
  - $\mathrm{E}(X_i) = \mu$
  - $\mathrm{Var}(X_i) = \sigma^2$
- Using the formula for a linear combination of *n* independent variables:

$$\begin{aligned}\mathrm{Var}\left(\overline{X}\right) &= \mathrm{Var}\left(\frac{X_1 + X_2 + \ \dots + X_n}{n}\right)\\ &= \frac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ \dots + \mathrm{Var}(X_n)}{n^2}\\ &= \frac{\sigma^2 + \sigma^2 + \ \dots + \sigma^2}{n^2}\\ &= \frac{n\sigma^2}{n^2}\\ &= \frac{\sigma^2}{n}\end{aligned}$$

- It can be shown that $\mathrm{E}\left(\overline{X}^2\right) = \mu^2 + \dfrac{\sigma^2}{n}$
    - This comes from rearranging $\mathrm{Var}\left(\overline{X}\right) = \mathrm{E}\left(\overline{X}^2\right) - \left[\mathrm{E}\left(\overline{X}\right)\right]^2$
- It can be shown that $\mathrm{E}\left(X^2\right) = \mathrm{E}\left(X_i{}^2\right) = \mu^2 + \sigma^2$
    - This comes from rearranging $\mathrm{Var}(X) = \mathrm{E}\left(X^2\right) - [\mathrm{E}(X)]^2$
- Using the formula for a linear combination of *n* independent variables:

$$\begin{aligned}
\mathrm{E}\left(S_n^2\right) &= \mathrm{E}\left(\frac{\sum X_i^2}{n} - \overline{X}^2\right) \\
&= \frac{\sum \mathrm{E}\left(X_i^2\right)}{n} - \mathrm{E}\left(\overline{X}^2\right) \\
&= \frac{\sum\left(\mu^2 + \sigma^2\right)}{n} - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\
&= \frac{n\left(\mu^2 + \sigma^2\right)}{n} - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\
&= \mu^2 + \sigma^2 - \left(\mu^2 + \frac{\sigma^2}{n}\right) \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n\sigma^2 - \sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}$$

- As $\mathrm{E}\left(S_n^2\right) \neq \sigma^2$ this shows that the sample variance is not unbiased
    - You need to multiply by $\dfrac{n}{n-1}$
    - $\mathrm{E}\left(S_{n-1}^2\right) = \sigma^2$

Exam Tip

- Check the wording of the exam question carefully to determine which of the following you are given:
    - The **population variance**: σ^2
    - The **sample variance**: s_n^2
    - An **unbiased estimate** for the **population variance**: s_{n-1}^2

Worked Example

The times, X minutes, spent on daily revision of a random sample of 50 IB students from the UK are summarised as follows.

$$n = 50 \qquad \sum x = 6174 \qquad s_n^2 = 1384.3$$

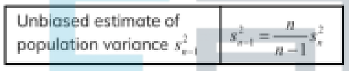
Calculate unbiased estimates of the population mean and variance of the times spent on daily revision by IB students in the UK.

Unbiased estimate of population mean $\bar{x} = \frac{\Sigma x}{n}$

$\bar{x} = \frac{6174}{50} = 123.48$

$\boxed{\bar{x} = 123 \text{ minutes } (3sf)}$

Formula booklet



Unbiased estimate of population variance s_{n-1}^2	$s_{n-1}^2 = \frac{n}{n-1} s_n^2$

$S_{n-1}^2 = \frac{50}{49} \times 1384.3 = 1412.55...$

$\boxed{S_{n-1}^2 = 1410 \text{ minutes}^2 \text{ } (3sf)}$

4.7 Binomial Distribution

4.7.1 The Binomial Distribution

Properties of Binomial Distribution

What is a binomial distribution?

- A binomial distribution is a **discrete probability distribution**
- A **discrete random variable** X follows a **binomial distribution** if it **counts the number of successes** when an experiment satisfies the following conditions:
  - There are a **fixed finite number of trials (*n*)**
  - The outcome of each trial is **independent** of the outcomes of the other trials
  - There are **exactly two outcomes** of each trial (**success or failure**)
  - The **probability of success is constant (*p*)**
- If X follows a binomial distribution then it is denoted $X \sim \mathrm{B}(n,\, p)$
  - *n* is the **number of trials**
  - *p* is the **probability of success**
- The **probability of failure is 1 – p** which is sometimes denoted as **q**
- The formula for the probability of *r* **successful trials** is given by:
  - $\mathrm{P}(X = r) = {}^{n}C_{r} \times p^{r}(1-p)^{n-r}$ for $r = 0, 1, 2, \ldots, n$
    - ${}^{n}C_{r} = \dfrac{n!}{r!(n-r)!}$ where $n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$
  - You will be expected to use the distribution function on your **GDC to calculate probabilities** with the binomial distribution

What are the important properties of a binomial distribution?

- The **expected number (mean)** of successful trials is

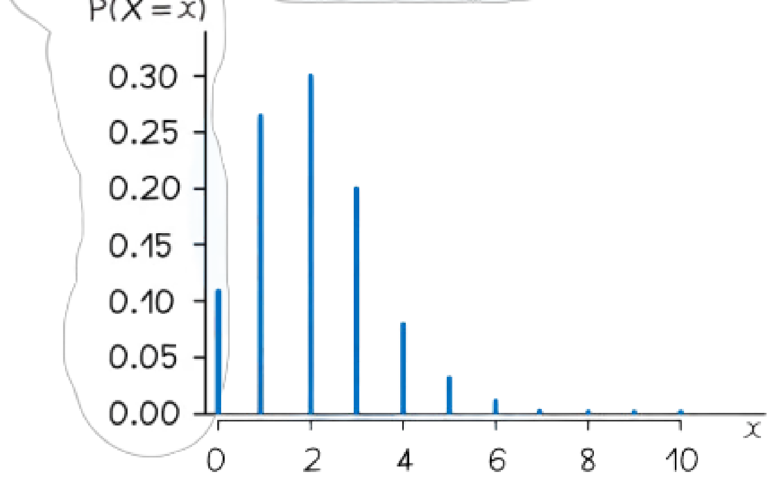
$$\mathrm{E}(X) = np$$

  - You are given this in the **formula booklet**
- The **variance** of the number of successful trials is

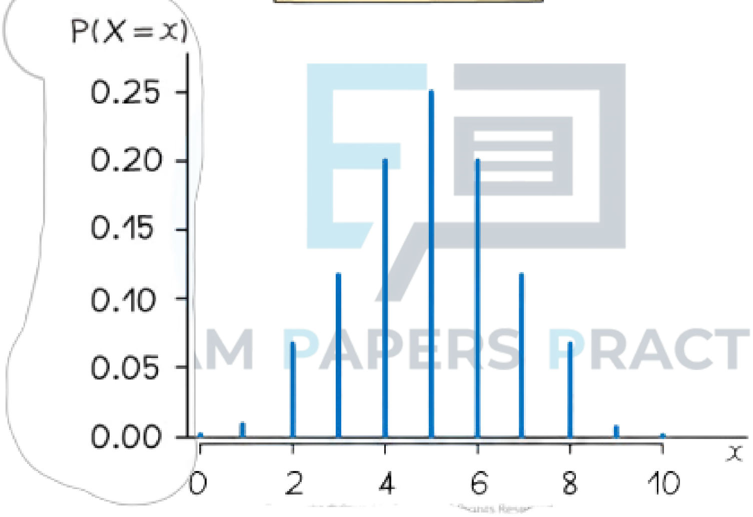
$$\mathrm{Var}(X) = np(1-p)$$

  - You are given this in the **formula booklet**
  - Square root to get the **standard deviation**
- The distribution can be represented visually using a vertical line graph
  - If *p* is **close to 0** then the graph has a **tail to the right**
  - If *p* is **close to 1** then the graph has a **tail to the left**
  - If *p* is **close to 0.5** then the graph is **roughly symmetrical**
  - **If p = 0.5** then the graph is **symmetrical**

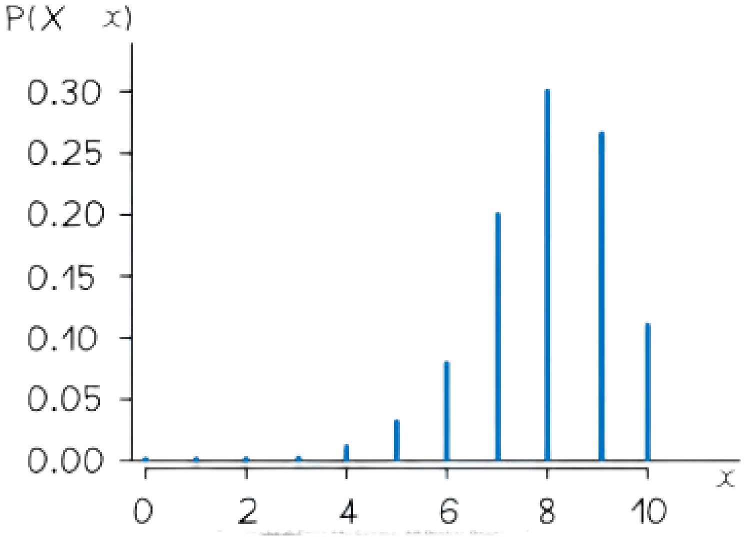
$X \sim B(10, 0.2)$



$X \sim B(10, 0.5)$



$X \sim B(10, 0.8)$



Modelling with Binomial Distribution

How do I set up a binomial model?

- **Identify** what a **trial** is in the scenario
  - For example: rolling a dice, flipping a coin, checking hair colour
- **Identify** what the **successful outcome** is in the scenario
  - For example: rolling a 6, landing on tails, having black hair
- **Identify** the **parameters**
  - *n* is the number of trials and *p* is the probability of success in each trial
- Make sure you **clearly state** what your **random variable** is
  - For example, let X be the number of students in a class of 30 with black hair

What can be modelled using a binomial distribution?

- Anything that satisfies the **four conditions**
- For example: let T be the number of times a fair coin lands on tails when flipped 20 times:
  - A trial is flipping a coin: There are 20 trials so ***n* = 20**
  - We can assume each coin flip does not affect subsequent coin flips: they are **independent**
  - A success is when the coin lands on tails: **Two outcomes** - tails or not tails (heads)
  - The coin is fair: The probability of tails is constant with ***p* = 0.5**
- Sometimes it might **seem like there are more than two outcomes**
  - For example: let Y be the number of yellow cars that are in a car park full of 100 cars
    - Although there are more than two possible colours of cars, here the trial is whether a car is yellow so there are two outcomes (yellow or not yellow)
    - Y would still need to fulfil the other conditions in order to follow a binomial distribution

- Sometimes a **sample may be taken from a population**
  - For example: 30% of people in a city have blue eyes, a sample of 30 people from the city is taken and X is the number of them with blue eyes
    - As long as the population is large and the sample is random then it can be assumed that each person has a 30% chance of having blue eyes

What can not be modelled using a binomial distribution?

- Anything where the number of trials is **not fixed** or is **infinite**
  - The number of emails received in an hour
  - The number of times a coin is flipped until it lands on heads
- Anything where the outcome of **one trial affects** the outcome of the **other trials**
  - The number of caramels that a person eats when they eat 5 sweets from a bag containing 6 caramels and 4 marshmallows
    - If you eat a caramel for your first sweet then there are less caramels left in the bag when you choose your second sweet
  - Anything where there are **more than two outcomes** of a trial
    - A person's shoe size
    - The number a dice lands on when rolled
  - Anything where the **probability of success changes**

- The number of times that a person can swim a length of a swimming pool in under a minute when swimming 50 lengths
  - The probability of swimming a lap in under a minute will decrease as the person gets tired
  - The probability is **not constant**

Exam Tip

- An exam question might involve different types of distributions so make it clear which distribution is being used for each variable

Worked Example

It is known that 8% of a large population are immune to a particular virus. Mark takes a sample of 50 people from this population. Mark uses a binomial model for the number of people in his sample that are immune to the virus.

a)

State the distribution that Mark uses.

A trial is checking if a person is immune to the virus

A success is if the person is immune.

Let X be the number of people in the sample immune to the virus

$\boxed{X \sim B(50, 0.08)}$

Number of people in sample

Probability of being immune to the virus

b)

State two assumptions that Mark must make in order to use a binomial model.

Mark needs to assume that:

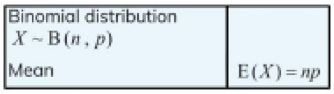
- each person in the population has an 8% chance of being immune
- the sample is random and the people are independent a person being immune does not affect the immunity of others

For example:
If all 50 came from the same family then they would not be independent

c)

Calculated the expected number of people in the sample that are immune to the virus.

Formula booklet



Binomial distribution $X \sim \mathrm{B}(n, p)$ Mean	$\mathrm{E}(X) = np$

$E(X) = 50 \times 0.08$

$\boxed{\text{4 people}}$

4.7.2 Calculating Binomial Probabilities

Calculating Binomial Probabilities

Throughout this section we will use the random variable $X \sim \mathrm{B}(n, p)$. For binomial, the probability of X taking a non-integer or negative value is always zero. Therefore any values of X mentioned in this section will be assumed to be non-negative integers.

How do I calculate P(*X* = *x*): the probability of a single value for a binomial distribution?

- You should have a **GDC** that can calculate **binomial probabilities**
- You want to use the "**Binomial Probability Distribution**" function
  - This is sometimes shortened to BPD, Binomial PD or Binomial Pdf
- You will need to enter:
  - The '*x*' value - the value of *x* for which you want to find $\mathrm{P}(X = x)$
  - The '*n*' value - the **number of trials**
  - The '*p*' value - the **probability of success**
- Some calculators will give you the option of **listing the probabilities** for **multiple values of *x* at once**
- There is a formula that you can use but you are expected to be able to use the distribution function on your GDC
  - $\mathrm{P}(X = x) = {}^{n}\mathrm{C}_{x} \times p^{x}(1 - p)^{n-x}$
    - ${}^{n}\mathrm{C}_{x} = \dfrac{n!}{r!(n - r)!}$

How do I calculate P(*a* ≤ *X* ≤ *b*): the cumulative probabilities for a binomial distribution?

- You should have a **GDC** that can calculate **cumulative binomial probabilities**
  - Most calculators will find $\mathrm{P}(a \le X \le b)$
  - Some calculators can only find $\mathrm{P}(X \le b)$
    - The identities below will help in this case
- You should use the "**Binomial Cumulative Distribution**" function
  - This is sometimes shortened to BCD, Binomial CD or Binomial Cdf
- You will need to enter:
  - The lower value - this is the **value *a***
    - This can be zero in the case $\mathrm{P}(X \le b)$
  - The upper value - this is the **value *b***
    - This can be *n* in the case $\mathrm{P}(X \ge a)$
  - The '*n*' value - the **number of trials**
  - The '*p*' value - the **probability of success**

How do I find probabilities if my GDC only calculates P(*X* ≤ *x*)?

- To calculate P(*X* ≤ *x*) just enter *x* into the cumulative distribution function

- To calculate P(*X* < *x*) use:
  - $\mathrm{P}(X < x) = \mathrm{P}(X \le x - 1)$ which works when *X* is a binomial random variable
    - P(*X* < 5) = P(*X* ≤ 4)

- To calculate P(*X* > *x*) use:
  - $\mathrm{P}(X > x) = 1 - \mathrm{P}(X \leq x)$ which works for any random variable *X*
    - P(*X* > 5) = 1 - P(*X* ≤ 5)

- To calculate P(*X* ≥ *x*) use:
  - $\mathrm{P}(X \geq x) = 1 - \mathrm{P}(X \leq x - 1)$ which works when *X* is a binomial random variable
    - P(*X* ≥ 5) = 1 - P(*X* ≤ 4)

- To calculate P(*a* ≤ *X* ≤ *b*) use:
  - $\mathrm{P}(a \leq X \leq b) = \mathrm{P}(X \leq b) - \mathrm{P}(X \leq a - 1)$ which works when *X* is a binomial random variable
    - P(5 ≤ *X* ≤ 9) = P(*X* ≤ 9) - P(*X* ≤ 4)

What if an inequality does not have the equals sign (strict inequality)?

- For a binomial distribution (as it is discrete) you could **rewrite all strict inequalities** (< and >) as **weak inequalities** (≤ and ≥) by using the identities for a binomial distribution
  - $\mathrm{P}(X < x) = \mathrm{P}(X \leq x - 1)$ and $\mathrm{P}(X > x) = \mathrm{P}(X \geq x + 1)$
  - For example: P(*X* < 5) = P(*X* ≤ 4) and P(*X* > 5) = P(*X* ≥ 6)
- It helps to think about the **range of integers** you want
  - Identify the smallest and biggest integers in the range
- If your range has no minimum or maximum then use 0 or *n*
  - $\mathrm{P}(X \leq b) = \mathrm{P}(0 \leq X \leq b)$
  - $\mathrm{P}(X \geq a) = \mathrm{P}(a \leq X \leq n)$

- $\mathrm{P}(a < X \leq b) = \mathrm{P}(a + 1 \leq X \leq b)$
  - P(5 < *X* ≤ 9) = P(6 ≤ *X* ≤ 9)

- $\mathrm{P}(a \leq X < b) = \mathrm{P}(a \leq X \leq b - 1)$
  - P(5 ≤ *X* < 9) = P(5 ≤ *X* ≤ 8)

- $\mathrm{P}(a < X < b) = \mathrm{P}(a + 1 \leq X \leq b - 1)$
  - P(5 < *X* < 9) = P(6 ≤ *X* ≤ 8)

Exam Tip

- If the question is in context then write down the inequality as well as the final answer
  - This means you still might gain a mark even if you accidentally type the wrong numbers into your GDC

Worked Example

The random variable $X \sim \mathrm{B}(40, 0.35)$. Find:

i)

$\mathrm{P}(X = 10)$.

Identify n and p $n = 40$ $p = 0.35$

Use binomial probability distribution on GDC

$P(X = 10) = 0.057056...$

$$\boxed{P(X = 10) = 0.057 \quad (3sf)}$$

ii)

$\mathrm{P}(X \leq 10)$.

Identify upper and lower values

$P(X \leqslant 10) = P(0 \leqslant X \leqslant 10)$

Use binomial cumulative distribution on GDC

$P(X \leqslant 10) = 0.121491...$

$$\boxed{P(X \leqslant 10) = 0.121 \quad (3sf)}$$

iii)

$\mathrm{P}(8 < X < 15)$.

Identify upper and lower values

$P(8 < X < 15) = P(9 \leqslant X \leqslant 14)$

Use binomial cumulative distribution on GDC

$P(9 \leqslant X \leqslant 14) = 0.541827...$

$$\boxed{P(8 < X < 15) = 0.542 \quad (3sf)}$$

4.8 Normal Distribution

4.8.1 The Normal Distribution

Properties of Normal Distribution

The binomial distribution is an example of a discrete probability distribution. The normal distribution is an example of a **continuous** probability distribution.

What is a continuous random variable?

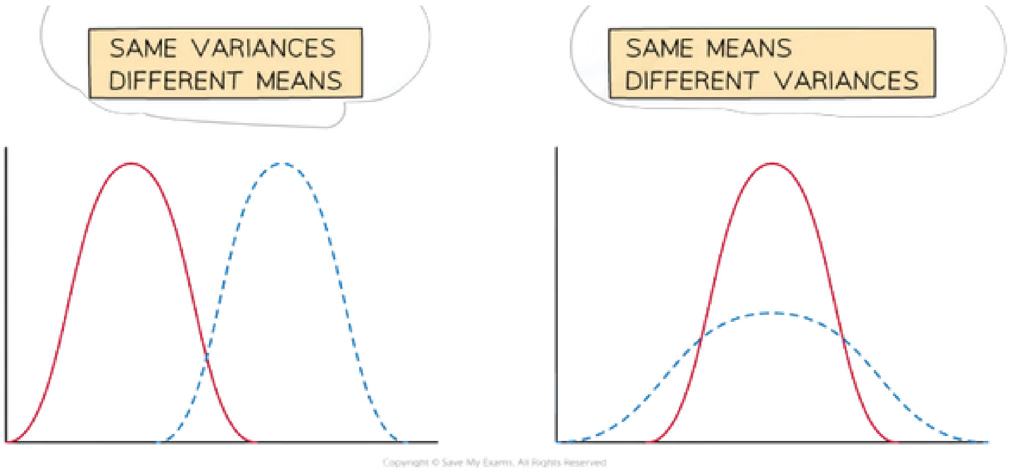
- A continuous random variable (often abbreviated to CRV) is a random variable that can take **any value** within a range of infinite values
  - Continuous random variables **usually measure** something
  - For example, height, weight, time, etc

What is a continuous probability distribution?

- A continuous probability distribution is a probability distribution in which the random variable X is continuous
- The probability of X being a **particular value is always zero**
  - $\mathrm{P}(X = k) = 0$ for any value *k*
  - Instead we define the **probability density function** $\mathrm{f}(x)$ for a specific value
    - This is a function that describes the **relative likelihood** that the random variable would be close to that value
  - We talk about the **probability** of X being within a **certain range**
- A continuous probability distribution can be represented by a continuous graph (the values for X along the horizontal axis and probability **density** on the vertical axis)
- The **area under the graph** between the points $x = a$ and $x = b$ is equal to $\mathrm{P}(a \le X \le b)$
  - The **total area under the graph equals 1**
- As $\mathrm{P}(X = k) = 0$ for any value *k*, it does not matter if we use strict or weak inequalities
  - $\mathrm{P}(X \le k) = \mathrm{P}(X < k)$ for any value *k* when *X* is a **continuous random variable**

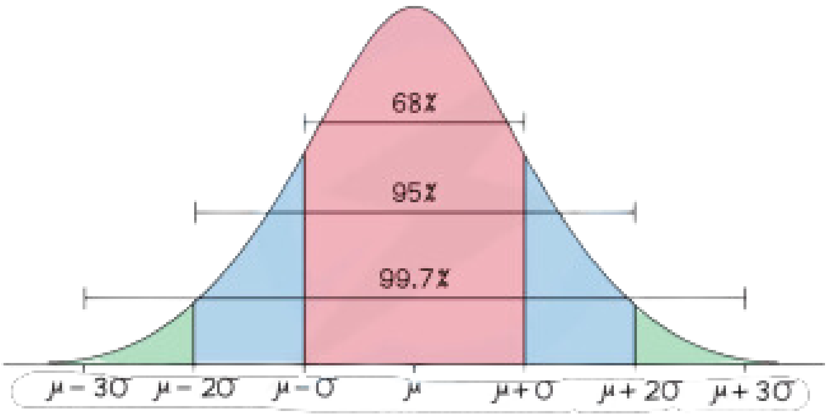
What is a normal distribution?

- A normal distribution is a **continuous probability distribution**
- The **continuous random variable** X can follow a normal distribution if:
  - The distribution is **symmetrical**
  - The distribution is **bell-shaped**
- If X follows a normal distribution then it is denoted $X \sim \mathrm{N}(\mu, \sigma^2)$
  - μ is the **mean**
  - σ^2 is the **variance**
  - σ is the **standard deviation**
- If the **mean** changes then the graph is **translated horizontally**
- If the **variance** increases then the graph is **widened horizontally** and **made taller vertically** to maintain the same area
  - A **small variance** leads to a **tall** curve with a **narrow** centre
  - A **large variance** leads to a **short** curve with a **wide** centre



What are the important properties of a normal distribution?

- The **mean** is μ
- The **variance** is σ^2
  - If you need the **standard deviation** remember to square root this
- The normal distribution is symmetrical about
  - Mean = Median = Mode = μ
- There are the results:
  - Approximately **two-thirds (68%)** of the data lies within **one standard deviation** of the mean ($\mu \pm \sigma$)
  - Approximately **95%** of the data lies within **two standard deviations** of the mean ($\mu \pm 2\sigma$)
  - Nearly **all of the data (99.7%)** lies within **three standard deviations** of the mean ($\mu \pm 3\sigma$)



Modelling with Normal Distribution

What can be modelled using a normal distribution?

- A lot of real-life continuous variables can be modelled by a normal distribution provided that the population is large enough and that the variable is **symmetrical** with **one mode**
- For a normal distribution X can take any real value, however values far from the mean (more than 4 standard deviations away from the mean) have a probability density of **practically zero**
  - This fact allows us to model variables that are not defined for all real values such as height and weight

What can not be modelled using a normal distribution?

- Variables which have **more than one mode** or **no mode**
  - For example: the number given by a random number generator
- Variables which are **not symmetrical**
  - For example: how long a human lives for

Exam Tip

- An exam question might involve different types of distributions so make it clear which distribution is being used for each variable

Worked Example

The random variable S represents the speeds (mph) of a certain species of cheetahs when they run. The variable is modelled using $\mathrm{N}(40,\ 100)$.

a)
Write down the mean and standard deviation of the running speeds of cheetahs.

$\mu = 40$ and $\sigma^2 = 100$

↑

Square root to get standard deviation

Mean $\mu = 40$
Standard deviation $\sigma = 10$

b)
State two assumptions that have been made in order to use this model.

We assume that the distribution of the speeds is

- symmetrical
- bell-shaped

4.8.2 Calculations with Norma lDistribution

Calculating Normal Probabilities

Throughout this section we will use the random variable $X \sim \mathrm{N}(\mu,\ \sigma^2)$. For *X* distributed normally, *X* can take any real number. Therefore any values mentioned in this section will be assumed to be real numbers.

How do I find probabilities using a normal distribution?

- The **area under a normal curve** between the points $x = a$ and $x = b$ is equal to the **probability** $\mathrm{P}(a < X < b)$
    - Remember for a normal distribution you do not need to worry about whether the inequality is strict (< or >) or weak (≤ or ≥)
        - $\mathrm{P}(a < X < b) = \mathrm{P}(a \le X \le b)$
- You will be **expected to use** distribution functions on your **GDC** to find the probabilities when working with a normal distribution

How do I calculate P(*X* = *x*): the probability of a single value for a normal distribution?

- The probability of a **single value** is **always zero** for a normal distribution
    - You can picture this as the area of a single line is zero
- $\mathrm{P}(X = x) = 0$
- Your GDC is likely to have a "**Normal Probability Density**" function
    - This is sometimes shortened to NPD, Normal PD or Normal Pdf
    - **IGNORE THIS FUNCTION** for this course!
    - This calculates the **probability density function** at a point **NOT the probability**

How do I calculate P(*a* < *X* < *b*): the probability of a range of values for a normal distribution?

- You need a **GDC** that can calculate **cumulative normal probabilities**
- You want to use the "**Normal Cumulative Distribution**" function
    - This is sometimes shortened to NCD, Normal CD or Normal Cdf
- You will need to enter:
    - The 'lower bound' - this is the value *a*
    - The 'upper bound' - this is the value *b*
    - The 'μ' value - this is the mean
    - The 'σ' value - this is the standard deviation
- **Check the order carefully** as some calculators ask for standard deviation before mean
    - Remember it is the standard deviation
        - so if you have the **variance** then **square root it**
- **Always sketch** a quick diagram to visualise which area you are looking for

How do I calculate P(*X* > *a*) or P(*X* < *b*) for a normal distribution?

- You will still use the "**Normal Cumulative Distribution**" function
- $\mathrm{P}(X > a)$ can be estimated using an **upper bound that is sufficiently bigger** than the **mean**
    - Using a value that is more than 4 standard deviations **bigger than the mean** is quite accurate
    - Or an easier option is just to input lots of 9's for the upper bound (**99999999…** or **10^{99}**)
- $\mathrm{P}(X < b)$ can be estimated using a **lower bound that is sufficiently smaller** than the **mean**

  - Using a value that is more than 4 standard deviations **smaller than the mean** is quite accurate
  - Or an easier option is just to input lots of 9's for the lower bound with a negative sign (**-99999999…** or **-10^{99}**)

Are there any useful identities?

- $\mathrm{P}(X < \mu) = \mathrm{P}(X > \mu) = 0.5$
- As $\mathrm{P}(X = a) = 0$ you can use:
  - $\mathrm{P}(X < a) + \mathrm{P}(X > a) = 1$
  - $\mathrm{P}(X > a) = 1 - \mathrm{P}(X < a)$
  - $\mathrm{P}(a < X < b) = \mathrm{P}(X < b) - \mathrm{P}(X < a)$
- These are useful when:
  - The mean and/or standard deviation are unknown
  - You only have a diagram
  - You are working with the **inverse distribution**

Exam Tip

- Check carefully whether you have entered the standard deviation or variance into your GDC

Worked Example

The random variable $Y \sim \mathrm{N}(20, 5^2)$. Calculate:

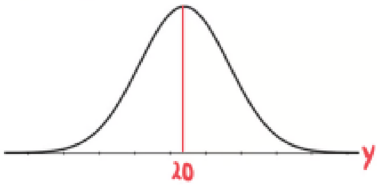
i)

$\mathrm{P}(Y = 20)$.

Identify μ and σ

$\mu = 20$ $\sigma^2 = 5^2$ so $\sigma = 5$

Sketch!

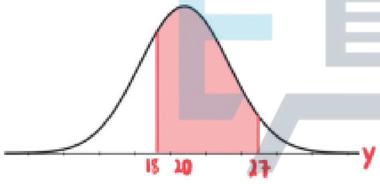


$\boxed{\mathrm{P}(Y=20) = 0}$

ii)

$\mathrm{P}(18 \le Y < 27)$.

Sketch!



Using GDC

Lower = 18

Upper = 27

We can use ≤ or <

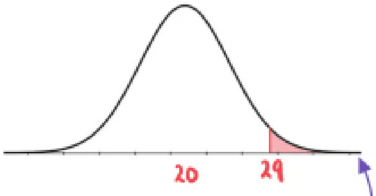
$\mathrm{P}(18 < Y < 27) = 0.574665...$

$\boxed{0.575 \text{ (3sf)}}$

iii)

$\mathrm{P}(Y > 29)$

Sketch!



No upper bound so choose a big number

Using GDC

Lower = 29

Upper = 99999

$\mathrm{P}(Y > 29) = 0.035930...$

$\boxed{0.0359 \text{ (3sf)}}$

Inverse Normal Distribution

Given the value of P(*X* < *a*) how do I find the value of *a*?

- Your **GDC** will have a function called "**Inverse Normal Distribution**"
  - Some calculators call this InvN
- Given that $\mathrm{P}(X < a) = p$ you will need to enter:
  - The 'area' - this is the value *p*
    - Some calculators might ask for the 'tail' - this is the left tail as you know the area to the left of *a*
  - The 'μ' value - this is the mean
  - The 'σ' value - this is the standard deviation

Given the value of P(*X* > *a*) how do I find the value of *a*?

- If your calculator **does** have the **tail option** (left, right or centre) then you can use the "Inverse Normal Distribution" function straightaway by:
  - Selecting 'right' for the tail
  - Entering the area as '*p*'
- If your calculator **does not** have the **tail option** (left, right or centre) then:
  - Given $\mathrm{P}(X > a) = p$
  - Use $\mathrm{P}(X < a) = 1 - \mathrm{P}(X > a)$ to rewrite this as
    - $\mathrm{P}(X < a) = 1 - p$
  - Then use the **method for P(*X* < *a*)** to find *a*

Exam Tip

- Always check your **answer makes sense**
  - If P(*X* < *a*) is **less than 0.5** then *a* should be **smaller than the mean**
  - If P(*X* < *a*) is **more than 0.5** then *a* should be **bigger than the mean**
  - A sketch will help you see this

Worked Example

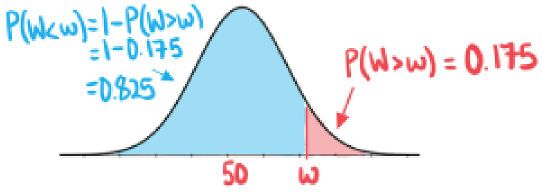
The random variable $W \sim \mathrm{N}(50,\ 36)$.

Find the value of w such that $\mathrm{P}(W > w) = 0.175$.

Identify μ and σ

$\mu = 50 \quad \sigma^2 = 36$ so $\sigma = 6$

Sketch!



P(W>w) is less than 0.5 so w is bigger than the mean

Area from left is 0.825

Use Inverse Normal Distribution function on GDC

$w = 55.6075\ldots$

$\boxed{w = 55.6 \ (3sf)}$

4.9 Further Normal Distribution (incCentral Limit Theorem)

4.9.1 Sample Mean Distribution

Combinations of Normal Variables

What is a linear combination of normal random variables?

- Suppose you have *n* **independent** normal random variables $X_i \sim \mathrm{N}\left(\mu_i,\ \sigma_i^2\right)$ for *i* = 1,2,3, ..., *n*
- A linear combination is of the form $X = a_1X_1 + a_2X_2 + \ ... + a_nX_n + b$ where *a*ᵢ and *b* are constants
- The mean and variance can be calculated using results from random variables
  - $\mathrm{E}(X) = a_1\mu_1 + a_2\mu_2 + \ ... + a_n\mu_n + b$
  - $\mathrm{Var}(X) = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ ... + a_n^2\sigma_n^2$
    - The variables **need to be independent** for this result to be true
- A **linear combination of *n* independent normal random variables** is also a **normal random variable** itself
  - $X \sim \mathrm{N}\left(a_1\mu_1 + a_2\mu_2 + \ ... + a_n\mu_n + b,\ a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \ ... + a_n^2\sigma_n^2\right)$
  - This can be used to find probabilities when combining normal random variables

What is meant by the sample mean distribution?

- Suppose you have a population with distribution *X* and you take a random sample with *n* observations $X_1, X_2, ..., X_n$
- The sample mean distribution is the distribution of the values of the sample mean
  - $\bar{X} = \dfrac{X_1 + X_2 + \ ... + X_n}{n}$
- For an individual sample the sample mean $\bar{x}$ can be calculated
  - This is also called a point estimate
  - $\bar{X}$ is the distribution of the point estimates

What does the sample mean distribution look like when *X* is normally distributed?

- If the population is normally distributed then the sample mean distribution is also normally distributed
- $\mathrm{E}(\bar{X}) = \mathrm{E}\left(\dfrac{X_1 + X_2 + \ ... + X_n}{n}\right) = \dfrac{\mathrm{E}(X_1) + \mathrm{E}(X_2) + \ ... + \mathrm{E}(X_n)}{n} = \dfrac{\mu + \mu + \ ... + \mu}{n} = \dfrac{n\mu}{n} = \mu$
- $\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\dfrac{X_1 + X_2 + \ ... + X_n}{n}\right) = \dfrac{\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ ... + \mathrm{Var}(X_n)}{n^2} = \dfrac{\sigma^2 + \sigma^2 + \ ... + \sigma^2}{n^2} = \dfrac{n\sigma^2}{n^2} = \dfrac{\sigma^2}{n}$
- Therefore you divide the variance of the population by the size of the sample to get the variance of the sample mean distribution
  - $X \sim \mathrm{N}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \mathrm{N}\left(\mu, \dfrac{\sigma^2}{n}\right)$

Worked Example

Amber makes a cup of tea using a hot drink vending machine. When the hot water button is pressed the machine dispenses W ml of hot water and when the milk button is pressed the machine dispenses M ml of milk. It is known that $W \sim \mathrm{N}(100,\ 15^2)$ and $M \sim \mathrm{N}(10,\ 2^2)$.

To make a cup of tea Amber presses the hot water button three times and the milk button twice. The total amount of liquid in Amber's cup is modelled by C ml.

a)

Write down the distribution of C.

$C = W_1 + W_2 + W_3 + M_1 + M_2$

$E(C) = E(W_1) + E(W_2) + E(W_3) + E(M_1) + E(M_2)$

$\mu = 100 + 100 + 100 + 10 + 10 = 320$

$Var(C) = Var(W_1) + Var(W_2) + Var(W_3) + Var(M_1) + Var(M_2)$

$\sigma^2 = 15^2 + 15^2 + 15^2 + 2^2 + 2^2 = 683$

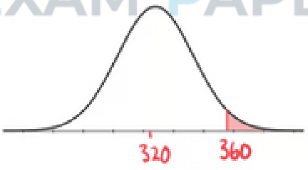
A linear combination of normal variables is also a normal variable

$\boxed{C \sim N(320, 683)}$

b)

Find the probability that the total amount of liquid in Amber's cup exceeds 360 ml.

Use normal distribution on GDC



$\mu = 320 \quad \sigma = \sqrt{683}$

Lower $= 360$ Upper $= 9999...$

$P(C > 360) = 0.062939...$

$\boxed{P(C > 360) = 0.0629 \ (3sf)}$

c)

Amber makes 15 cups of tea and calculates the mean $\overline{C}$. Write down the distribution of $\overline{C}$.

$\bar{C} \sim N(\mu, \frac{\sigma^2}{n})$

$\boxed{\bar{C} \sim N\left(320, \frac{683}{15}\right)}$

Central Limit Theorem

What is the Central Limit Theorem?

- The Central Limit Theorem says that if a sufficiently large random sample is taken from any distribution X then the sample mean distribution $\overline{X}$ can be approximated by a normal distribution
  - In your exam $n > 30$ will be considered sufficiently large for the sample size
- Therefore $\overline{X}$ can be modelled by $\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$
  - μ is the mean of X
  - σ^2 is the variance of X
  - n is the size of the sample

Do I always need to use the Central Limit Theorem when working with the sample mean distribution?

- No – the Central Limit Theorem is not needed when the population is normally distributed
- If X is **normally distributed** then $\overline{X}$ is normally distributed
  - This is true regardless of the size of the sample
  - The **Central Limit Theorem is not needed**
- If X is **not normally distributed** then $\overline{X}$ is approximately normally distributed
  - Provided the sample size is large enough
  - The **Central Limit Theorem is needed**

Worked Example

The integers 1 to 29 are written on counters and placed in a bag. The expected value when one is picked at random is 15 and the variance is 70. Susie randomly picks 40 integers, returning the counter after each selection.

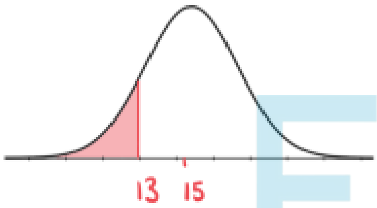
a)

Find the probability that the mean of Susie's 40 numbers is less than 13.

Let $\bar{X}$ be the mean of 40 numbers

n large ⇒ $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ $\quad$ $\bar{X} \sim N(15, \frac{70}{40})$

Use normal distribution on GDC



$\mu = 15$ $\quad$ $\sigma = \sqrt{\frac{70}{40}}$

Lower = -999... $\quad$ Upper = 13

$P(\bar{X} < 13) = 0.065285...$

$\boxed{P(\bar{X} < 13) = 0.0653 \text{ (3sf)}}$

b)

Explain whether it was necessary to use the Central Limit Theorem in your calculation.

The Central Limit Theorem was necessary as the random variable for the number picked out of the bag is not normally distributed.

4.9.2 Confidence Interval for the Mean

Confidence Interval for μ

What is a confidence interval?

- It is **impossible** to find the **exact value** of the **population mean** when taking a sample
  - The mean of a sample is called a **point estimate**
  - The best we can do is find an **interval** for which the **exact value is likely to lie**
  - This is called the **confidence interval for the mean**
- The **confidence level** of a confidence interval is the **probability** that the **interval contains the population mean**
  - Be careful with the wording – the population mean is a fixed value so it does not make sense to talk about the probability that it lies within an interval
    - Instead we talk about the probability of an interval containing the mean
  - Suppose samples were collected and a **95%** confidence interval for the population mean was constructed for each sample then for every 100 intervals we would **expect on average 95 of them to contain the mean**
    - 95 out of 100 is **not guaranteed** – it is possible that all of them could contain the mean
    - It is also possible (though **very unlikely**) that none of them contains the mean

How do I find a confidence interval for the population mean (μ)?

- You will be given data using a **sample from a population**
  - The population will be **normally distributed**
    - If not then the sample size should be large enough so you can use the **Central Limit Theorem**
- You will use the **interval functions** on your calculator
- Use a **z-interval** if the **population variance** is **known** σ^2
  - On your GDC enter:
    - the standard deviation σ and the confidence level α%
    - EITHER the raw data
    - OR the sample mean $\overline{x}$ and the sample size n
- Use a **t-interval** if the **population variance** is **unknown**
  - In this case the test uses the **unbiased estimate** for the variance s^2_{n-1}
  - On your GDC enter:
    - the confidence level α%
    - EITHER the raw data
    - OR the sample mean $\overline{x}$, the value of s_{n-1} and the sample size n
- Your GDC will give you the lower and upper bounds of the interval
  - It can be written as $a < \mu < b$

What affects the width of a confidence interval?

- The **width** of a confidence interval is the **range of the values** in the interval
- The **confidence level** affects the width

  - **Increasing** the confidence level will **increase the width**
  - **Decreasing** the confidence level with **decrease the width**
- The **size of the sample** affects the width
  - **Increasing** the sample size will **decrease the width**
  - **Decreasing** the sample size will **increase the width**

How can I interpret a confidence interval?

- After you have found a confidence interval for μ you might be expected to **comment on the claim** for a value of μ
- If the claimed value is **within** the confidence interval then there is **not enough evidence to reject the claim**
  - Therefore the **claim is supported**
- If the claimed value is **outside** the interval then there is **sufficient evidence to reject the claim**
  - The value is **unlikely to be correct**

Worked Example

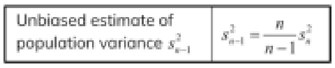
Cara wants to check the mean weight of burgers sold by a butcher. The weights of the burgers are assumed to be normally distributed. Cara takes a random sample of 12 burgers and finds that the mean weight is 293 grams and the standard deviation of the sample is 5.5 grams.

a)

Find a 95% confidence interval for the population mean, giving your answer to 4 significant figures.

The population variance is unknown so use a t- interval

Formula booklet



Unbiased estimate of population variance s_{n-1}^2	$s_{n-1}^2 = \frac{n}{n-1}s_n^2$

$s_{n-1}^2 = \frac{12}{11} \times 5.5^2 = 33$

Enter data into GDC

Confidence level $= 0.95$ $\quad \bar{x} = 293 \quad s_{n-1} = \sqrt{33} \quad n = 12$

Lower: 289.35...

Upper: 296.64...

$\boxed{289.4 < \mu < 296.6 \quad (4sf)}$

b)

The butcher claims the burgers weigh 300 grams. Comment on this claim with reference to the confidence interval.

300 is above the confidence interval which suggests that the butcher's claim is not true.

4.10 Poisson Distribution

4.10.1 Poisson Distribution

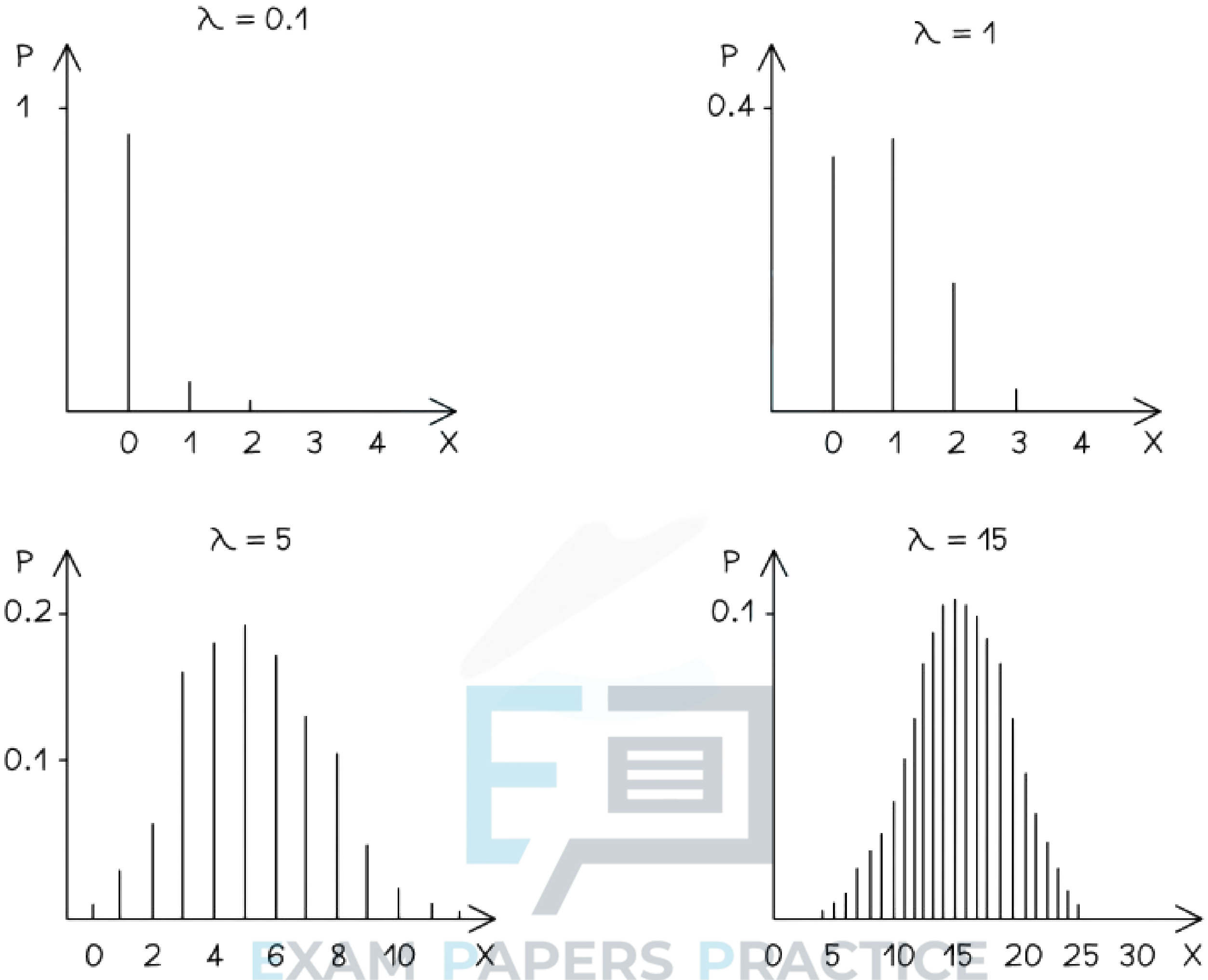
Properties of Poisson Distribution

What is a Poisson distribution?

- A Poisson distribution is a **discrete probability distribution**
- A **discrete random variable** X follows a **Poisson distribution** if it **counts the number of occurrences** in a fixed time period given the following conditions:
  - Occurrences are **independent**
  - Occurrences occur at a **uniform average rate** for the time period **(*m*)**
- If X follows a Poisson distribution then it is denoted $X \sim \mathrm{Po}(m)$
  - *m* is the average rate of occurrences for the time period
- The formula for the probability of *r* occurrences is given by:
  - $\mathrm{P}(X=r)=\dfrac{\mathrm{e}^{-m}m^{r}}{r!}$ for $r = 0, 1, 2, \ldots$
    - e is Euler's constant 2.718…
    - $r! = r\times(r-1)\times\ldots\times2\times1$ and $0!=1$
    - There is no upper bound for the number of occurrences
  - You will be expected to use the distribution function on your GDC to calculate probabilities with the Poisson distribution

What are the important properties of a Poisson distribution?

- The **expected number (mean)** of occurrences is *m*
  - You are given this in the **formula booklet**
- The **variance** of the number of occurrences is *m*
  - You are given this in the **formula booklet**
  - Square root to get the **standard deviation**
- The **mean** and **variance** for a Poisson distribution are **equal**
- The distribution can be represented visually using a vertical line graph
  - The graphs have **tails to the right** for all values of *m*
  - As ***m* gets larger** the graph gets **more symmetrical**
- If $X\sim\mathrm{Po}(m)$ and $Y\sim\mathrm{Po}(\lambda)$ are **independent** then $X+Y\sim\mathrm{Po}(m+\lambda)$
  - This extends to *n* independent Poisson distributions $X_i \sim \mathrm{Po}(m_i)$
    - $X_1+X_2+\ \ldots+X_n \sim \mathrm{Po}(m_1+m_2+\ \ldots+m_n)$


λ = 0.1
P
1
0
1
2
3
4
X
λ = 1
P
0.4
0
1
2
3
4
X
λ = 5
P
0.2
0.1
0
2
4
6
8
10
X
λ = 15
P
0.1
0
5
10
15
20
25
30
X

Modelling with Poisson Distribution

How do I set up a Poisson model?

- **Identify** what an **occurrence** is in the scenario
    - For example: a car passing a camera, a machine producing a faulty item
- Use **proportion** to find the **mean number of occurrences** for the given time period
    - For example: 10 cars in 5 minutes would be 120 cars in an hour if the Poisson model works for both time periods
- Make sure you **clearly state** what your **random variable** is
    - For example: let *X* be the number of cars passing a camera in 10 minutes

What can be modelled using a Poisson distribution?

- Anything that satisfies the **two conditions**
- For example, Let C be the number of calls that a helpline receives within a 15-minute period: $C \sim \mathrm{Po}(m)$
    - An occurrence is the helpline receiving a call and can be considered independent
    - The helpline receives calls at an average rate of *m* calls during a 15-minute period
- Sometimes a **measure of space** will be used instead of a time period
    - For example, the number of daisies that exist on a square metre of grass
- If the **mean** and **variance** of a discrete variable are **equal** then it might be possible to use a Poisson model

Exam Tip

- An exam question might involve different types of distributions so make it clear which distribution is being used for each variable

? Worked Example

Jack uses $\mathrm{Po}(6.25)$ to model the number of emails he receives during his hour lunch break.

a)

Write down two assumptions that Jack has made.

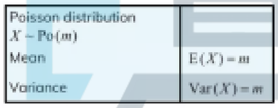
Jack has assumed that :

- the emails that he receives are independent
- he receives emails at a uniform average rate of 6.25 emails per hour during his lunch breaks

b)

Calculate the standard deviation for the number of emails that Jack receives during his hour lunch breaks.

Formula booklet



Poisson distribution $X \sim \mathrm{Po}(m)$	
Mean	$\mathrm{E}(X) = m$
Variance	$\mathrm{Var}(X) = m$

$\sigma^2 = 6.25$

$\sigma = \sqrt{6.25}$

Standard deviation = 2.5 emails

4.10.2 Calculating Poisson Probabilities

Calculating Poisson Probabilities

Throughout this section we will use the random variable $X \sim \text{Po}(m)$. For a Poisson distribution *X*, the probability of *X* taking a non-integer or negative value is always zero. Therefore, any values mentioned in this section for *X* will be assumed to be non-negative integers. The value of *m* can be any real positive value.

How do I calculate P(*X* = *x*): the probability of a single value for a Poisson distribution?

- You should have a **GDC** that can calculate **Poisson probabilities**
- You want to use the "**Poisson Probability Distribution**" function
  - This is sometimes shortened to PPD, Poisson PD or Poisson Pdf
- You will need to enter:
  - The '*x*' value - the value of *x* for which you want to find $\text{P}(X=x)$
  - The 'λ' value - the **mean number of occurrences** (*m*)
- Some calculators will give you the option of **listing the probabilities** for **multiple values of *x* at once**
- There is a formula that you can use but you are expected to be able to use the distribution function on your GDC
  - $\text{P}(X=x)=\dfrac{\text{e}^{-m}m^x}{x!}$
    - where e is Euler's constant
    - $x!=x\times(x-1)\times\ldots\times2\times1$ and $0!=1$

How do I calculate P(*a* ≤ *X* ≤ *b*): the cumulative probabilities for a Poisson distribution?

- You should have a **GDC** that can calculate **cumulative Poisson probabilities**
  - Most calculators will find $\text{P}(a\le X\le b)$
  - Some calculators can only find $\text{P}(X\le b)$
    - The identities below will help in this case
- You should use the "**Poisson Cumulative Distribution**" function
  - This is sometimes shortened to PCD, Poisson CD or Poisson Cdf
- You will need to enter:
  - The lower value - this is the **value *a***
    - This can be zero in the case $\text{P}(X\le b)$
  - The upper value - this is the **value *b***
    - This can be a very large number (9999...) in the case $\text{P}(X\ge a)$
  - The 'λ' value - the **mean number of occurrences** (*m*)

How do I find probabilities if my GDC only calculates P(*X* ≤ *x*)?

- To calculate P(*X* ≤ *x*) just enter *x* into the cumulative distribution function
- To calculate P(*X* < *x*) use:
  - $\text{P}(X<x)=\text{P}(X\le x-1)$ which works when *X* is a Poisson random variable

- $P(X < 5) = P(X \le 4)$

- To calculate $P(X > x)$ use:
  - $P(X > x) = 1 - P(X \le x)$ which works for any random variable X
    - $P(X > 5) = 1 - P(X \le 5)$

- To calculate $P(X \ge x)$ use:
  - $P(X \ge x) = 1 - P(X \le x - 1)$ which works when X is a Poisson random variable
    - $P(X \ge 5) = 1 - P(X \le 4)$

- To calculate $P(a \le X \le b)$ use:
  - $P(a \le X \le b) = P(X \le b) - P(X \le a - 1)$ which works when X is a Poisson random variable
    - $P(5 \le X \le 9) = P(X \le 9) - P(X \le 4)$

What if an inequality does not have the equals sign (strict inequality)?

- For a Poisson distribution (as it is discrete) you could **rewrite all strict inequalities** (< and >) as **weak inequalities** (≤ and ≥) by using the identities for a Poisson distribution
  - $P(X < x) = P(X \le x - 1)$ and $P(X > x) = P(X \ge x + 1)$
  - For example: $P(X < 5) = P(X \le 4)$ and $P(X > 5) = P(X \ge 6)$
- It helps to think about the **range of integers** you want
  - Identify the smallest and biggest integers in the range
- If your range has no minimum then use 0
  - $P(X \le b) = P(0 \le X \le b)$

- $P(a < X \le b) = P(a + 1 \le X \le b)$
  - $P(5 < X \le 9) = P(6 \le X \le 9)$

- $P(a \le X < b) = P(a \le X \le b - 1)$
  - $P(5 \le X < 9) = P(5 \le X \le 8)$

- $P(a < X < b) = P(a + 1 \le X \le b - 1)$
  - $P(5 < X < 9) = P(6 \le X \le 8)$

Worked Example

The random variables $X \sim \text{Po}(6.25)$ and $Y \sim \text{Po}(4)$ are independent. Find:

i)
$\text{P}(X = 5)$,

Use Poisson probability distribution on GDC

$m = 6.25 \ (\lambda) \quad x = 5$

$P(X = 5) = 0.15341...$

$\boxed{P(X = 5) = 0.153 \ (3sf)}$

ii)
$\text{P}(Y \leq 5)$,

Identify upper and lower bounds

$P(Y \leq 5) = P(0 \leq Y \leq 5)$

Use Poisson cumulative distribution on GDC $\quad m = 4 \ (\lambda)$

$P(Y \leq 5) = 0.78513...$

$\boxed{P(Y \leq 5) = 0.785 \ (3sf)}$

iii)
$\text{P}(X + Y > 7)$.

Form the distribution $\quad m = 6.25 + 4 = 10.25$

$X + Y \sim Po(10.25)$

Identify lower bound - no upper bound so use a large number 999...

$P(X + Y > 7) = P(8 \leq X + Y)$

Use Poisson cumulative distribution on GDC

$P(8 \leq X + Y) = 0.80146...$

$\boxed{P(X + Y > 7) = 0.801 \ (3sf)}$

4.11 Hypothesis Testing

4.11.1 Hypothesis Testing

Language of Hypothesis Testing

What is a hypothesis test?

- A hypothesis test uses a **sample of data** in an experiment to test a **statement** made about the **population**
    - The statement is either about a **population parameter** or the distribution of the **population**
- The hypothesis test will look at the probability of observed outcomes happening under set conditions
- The probability found will be compared against a given **significance level** to determine whether there is **evidence to support** the statement being made

What are the key terms used in statistical hypothesis testing?

- Every hypothesis test **must** begin with a clear **null hypothesis** (what we believe to already be true) and **alternative hypothesis** (how we believe the data pattern or probability distribution might have changed)
- A **hypothesis** is an assumption that is made about a particular **population parameter** or the **distribution of the population**
    - A **population parameter** is a numerical characteristic which helps define a population
        - Such as the mean value of the population
    - The **null hypothesis** is denoted H_0 and sets out the assumed population parameter or distribution given that no change has happened
    - The **alternative hypothesis** is denoted H_1 and sets out how we think the population parameter or distribution could have changed
    - When a hypothesis test is carried out, the null hypothesis is **assumed to be true** and this assumption will either be **accepted** or **rejected**
        - When a null hypothesis is accepted or rejected a **statistical inference** is made
- A hypothesis test will always be carried out at an appropriate **significance level**
    - The significance level sets the **smallest probability** that an event could have occurred by chance
        - Any probability smaller than the significance level would suggest that the event is unlikely to have happened by chance
    - The **significance level** must be set **before** the hypothesis test is carried out
    - The **significance level** will usually be 1%, 5% or 10%, however it may vary

One-tailed Tests

What are one-tailed tests?

- A **one-tailed test** is used for testing:
  - Whether a distribution can be used to model the population
  - Whether the population parameter has **increased**
  - Whether the population parameter has **decreased**
- **One-tailed tests** can be used with:
  - Chi-squared test for independence
  - Chi-squared goodness of fit test
  - Test for proportion of a binomial distribution
  - Test for population mean of a Poisson distribution
  - Test for population mean of a normal distribution
  - Test to compare population means of two distributions

Two-tailed Tests

What are two-tailed tests?

- A **two-tailed test** is used for testing:
  - Whether the population parameter has **changed**
- **Two-tailed tests** can be used with:
  - Test for population mean of a normal distribution
  - Test to compare population means of two distributions

Conclusions of Hypothesis Testing

How do I decide whether to reject or accept the null hypothesis?

- A sample of the population is taken and the **test statistic** is calculated **using the observations** from the sample
  - Your GDC can calculate the test statistic for you (if required)
- To decide whether or not to reject the null hypothesis you first need either the ***p*-value** or the **critical region**
- The ***p* - value** is the probability of a value being **at least as extreme** as the test statistic, assuming that the null hypothesis is true
  - Your GDC will give you the *p*-value (if required)
  - If the ***p*-value is less than the significance level** then the **null hypothesis** would be **rejected**
- The **critical region** is the range of values of the test statistic which will lead to the **null hypothesis** being **rejected**
  - If the **test statistic** falls within the **critical region** then the **null hypothesis** would be **rejected**
- The **critical value** is the boundary of the critical region
  - It is the least extreme value that would lead to the rejection of the null hypothesis
  - The **critical value** is determined by the **significance level**

How should a conclusion be written for a hypothesis test?

- Your conclusion **must** be written in the **context** of the question
- Use the **wording in the question** to help you write your conclusion
  - If **rejecting the null** hypothesis your conclusion should state that there is **sufficient evidence** to suggest that the null hypothesis is unlikely to be true
  - If **accepting the null** hypothesis your conclusion should state that there is **not enough evidence** to suggest that the null hypothesis is unlikely to be true
- Your conclusion **must not** be definitive
  - There is a chance that the test has led to an **incorrect conclusion**
  - The outcome is **dependent on the sample**
    - a **different sample** might lead to a **different outcome**
- The conclusion of a **two-tailed test** can state if there is evidence of a change
  - You should not state whether this change is an increase or decrease
  - If you are testing the difference between the means of two populations then you can **only conclude that the means are not equal**
    - You can not say which population mean is bigger
    - You'd need to use a **one-tailed** test for this

Exam Tip

- Accepting the null hypothesis does **not** mean that you are saying it is true
  - You are simply saying there is not enough evidence to reject it

4.11.2 Chi-squared Test for Independence

Chi-Squared Test for Independence

What is a chi-squared test for independence?

- A chi-squared (χ^2) **test for independence** is a hypothesis test used to test whether **two variables are independent** of each other
  - This is sometimes called a χ^2 **two-way test**
- This is an example of a **goodness of fit** test
  - We are testing whether the data fits the model that the variables are independent
- The chi-squared (χ^2) distribution is used for this test
- You will use a **contingency table**
  - This is a two-way table that shows the **observed frequencies** for the different combinations of the two variables
    - For example: if the two variables are hair colour and eye colour then the contingency table will show the frequencies of the different combinations

Why might I have to combine rows or columns?

- The **observed** values are used to calculate **expected** values
  - These are the expected frequencies for each combination **assuming that the variables are independent**
    - Your GDC can calculate these for you after you input the observed frequencies
- The **expected values** must all be **bigger than 5**
- If one of the expected values is less than 5 then you will have to **combine the corresponding row or column** in the matrix of **observed values** with the **adjacent** row or column
  - The decision between row or column will be based on which seems the **most appropriate**
    - For example: if the two variables are age and favourite TV genre then it is more appropriate to combine age groups than types of genre

What are the degrees of freedom?

- There will be a **minimum number of expected values** you would need to know in order to be able to calculate all the expected values
- This minimum number is called the **degrees of freedom** and is often denoted by ν
- For a **test for independence** with an $m \times n$ contingency table
  - $\nu = (m-1)\times(n-1)$
  - For example: If there are 5 rows and 3 columns then you only need to know **2 of the values** in **4 of the rows** as the rest can be calculated using the totals

What are the steps for a chi-squared test for independence?

- **STEP 1**: Write the **hypotheses**
  - H_0: Variable *X* is independent of variable *Y*
  - H_1: Variable *X* is not independent of variable *Y*
    - Make sure you clearly write what the variables are and don't just call them *X* and *Y*
- **STEP 2**: Calculate the **degrees of freedom** for the test
  - For an $m \times n$ contingency table
  - Degrees of freedom is $\nu = (m-1)\times(n-1)$

- **STEP 3**: Enter your **observed frequencies** into your GDC using the option for a 2-way test
  - Enter these as a matrix
  - Your GDC will give you a matrix of the **expected values** (assuming the variables are independent)
    - If any values are 5 or less then combine rows/columns and **repeat step 2**
  - Your GDC will also give you the χ^2 statistic and its p-value
  - The χ^2 statistic is denoted as χ^2_{calc}
- **STEP 4**: Decide whether there is evidence to reject the null hypothesis
  - EITHER compare the **χ^2 statistic** with the given **critical value**
    - If χ^2 statistic **>** critical value then **reject H_0**
    - If χ^2 statistic **<** critical value then **accept H_0**
  - OR compare the ***p*-value** with the given **significance level**
    - If p-value **<** significance level then **reject H_0**
    - If p-value **>** significance level then **accept H_0**
- **STEP 5**: Write your **conclusion**
  - If you **reject H_0**
    - There is sufficient evidence to suggest that variable X is not independent of variable Y
    - Therefore this suggests they are **associated**
  - If you **accept H_0**
    - There is insufficient evidence to suggest that variable X is not independent of variable Y
    - Therefore this suggests they are **independent**

How do I calculate the chi-squared statistic?

- You are **expected** to be able to use your **GDC** to calculate the χ^2 statistic by inputting the matrix of the observed frequencies
- Seeing how it is done by hand might deepen your understanding but you are **not expected** to use this method
- **STEP 1**: For each **observed frequency** O_i calculate its **expected frequency** E_i
  - Assuming the variables are independent
    - $E_i = \text{P}(X = x) \times \text{P}(Y = y) \times \text{Total}$
    - Which simplifies to $E_i = \dfrac{\text{Row Total} \times \text{Column Total}}{\text{Overall Total}}$
- **STEP 2**: Calculate the χ^2 statistic using the formula
  - $\chi^2_{calc} = \sum \dfrac{(O_i - E_i)^2}{E_i}$
  - You do not need to learn this formula as your GDC calculates it for you
- To calculate the p-value you would find the probability of a value being bigger than your χ^2 statistic using a χ^2 distribution with ν degrees of freedom

Exam Tip

Note for Internal Assessments (IA)

- If you use a χ^2 test in your IA then beware that the outcome may not be accurate if there is only 1 degree of freedom
  - This means it is a 2 × 2 contingency table

Worked Example

At a school in Paris, it is believed that favourite film genre is related to favourite subject. 500 students were asked to indicate their favourite film genre and favourite subject from a selection and the results are indicated in the table below.

	Comedy	Action	Romance	Thriller
Maths	51	52	37	55
Sports	59	63	41	33
Geography	35	31	28	15

It is decided to test this hypothesis by using a χ^2 test for independence at the 1% significance level.

The critical value is 16.812.

a)
State the null and alternative hypotheses for this test.

H_0 : Favourite subject is independent of favourite film genre

H_1 : Favourite subject is not independent of favourite film genre

b)
Write down the number of degrees of freedom for this table.

$\nu = (\text{rows} - 1) \times (\text{columns} - 1) = (3-1) \times (4-1)$

$\boxed{\nu = 6}$

c)
Calculate the χ^2 test statistic for this data.

Type matrix into GDC

χ^2 statistic $= 12.817...$

$\boxed{\chi^2_{calc} = 12.8 \ (3\text{ sf})}$

d)
Write down the conclusion to the test. Give a reason for your answer.

$12.8 < 16.812$

Accept H_0 as X^2 statistic < critical value. There is insufficient evidence to suggest that favourite subject is not independent of favourite film genre. Therefore this suggests they are independent.

4.11.3 Goodness of Fit Test

Chi-Squared GOF: Uniform

What is a chi-squared goodness of fit test for a given distribution?

- A chi-squared (χ^2) **goodness of fit test** is used to test data from a sample which suggests that the population has a given distribution
- This could be that:
  - the proportions of the population for different categories follows a **given ratio**
  - the population follows a **uniform distribution**
    - This means all outcomes are **equally likely**

What are the steps for a chi-squared goodness of fit test for a given distribution?

- **STEP 1**: Write the **hypotheses**
  - H_0: Variable *X* can be modelled by the given distribution
  - H_1: Variable *X* cannot be modelled by the given distribution
    - Make sure you clearly write what the variable is and don't just call it *X*
- **STEP 2**: Calculate the degrees of freedom for the test
  - For *k* outcomes
  - Degrees of freedom is $\nu = k - 1$
- **STEP 3**: Calculate the **expected frequencies**
  - Split the total frequency using the given ratio
  - For a uniform distribution: divide the total frequency *N* by the number of outcomes *k*
- **STEP 4**: Enter the **frequencies** and the **degrees of freedom** into your GDC
  - Enter the observed and expected frequencies as two separate lists
  - Your GDC will then give you the χ^2 statistic and its *p*-value
  - The χ^2 statistic is denoted as χ^2_{calc}
- **STEP 5**: Decide **whether** there is evidence to **reject the null** hypothesis
  - EITHER compare the **χ^2 statistic** with the given **critical value**
    - If χ^2 statistic > critical value then **reject H_0**
    - If χ^2 statistic < critical value then **accept H_0**
  - OR compare the **p-value** with the given **significance level**
    - If *p*-value < significance level then **reject H_0**
    - If *p*-value > significance level then **accept H_0**
- **STEP 6**: Write your **conclusion**
  - If you **reject H_0**
    - There is sufficient evidence to suggest that variable *X* does not follow the given distribution
    - Therefore this suggests that the data is **not distributed as claimed**
  - If you **accept H_0**
    - There is insufficient evidence to suggest that variable *X* does not follow the given distribution
    - Therefore this suggests that the data is **distributed as claimed**

Worked Example

A car salesman is interested in how his sales are distributed and records his sales results over a period of six weeks. The data is shown in the table.

Week	1	2	3	4	5	6
Number of sales	15	17	11	21	14	12

A χ^2 goodness of fit test is to be performed on the data at the 5% significance level to find out whether the data fits a uniform distribution.

a)

Find the expected frequency of sales for each week if the data were uniformly distributed.

If uniformly distributed all expected frequencies are equal

$$\text{Expected frequency} = \frac{15+17+11+21+14+12}{6}$$

$$\boxed{\text{Expected frequency} = 15}$$

b)

Write down the null and alternative hypotheses.

H_0 : Number of sales can be modelled by a uniform distribution

H_1 : Number of sales can not be modelled by a uniform distribution

c)

Write down the number of degrees of freedom for this test.

$$\nu = 6-1$$

$$\boxed{\nu = 5}$$

d)

Calculate the p-value.

Type two lists into GDC

Observed	15	17	11	21	14	12
Expected	15	15	15	15	15	15

$p = 0.4933...$

$\boxed{p = 0.493 \quad (3sf)}$

e)

State the conclusion of the test. Give a reason for your answer.

$0.493 > 0.05$

Accept H_0 as p-value > significance level There is insufficient evidence to suggest that number of sales can not be modelled by a uniform distribution. Therefore this suggests it is uniformly distributed.

Chi-Squared GOF: Binomial

What is a chi-squared goodness of fit test for a binomial distribution?

- A chi-squared (χ^2) **goodness of fit test** is used to test data from a sample suggesting that the population has a **binomial distribution**
  - You will either be **given a precise binomial distribution** to test $\mathrm{B}(n, p)$ with an assumed value for *p*
  - Or you will be asked to test whether a binomial distribution is **suitable without being given an assumed value** for *p*
    - In this case you will have to calculate an **estimate** for the value of *p* for the binomial distribution
    - To calculate it divide the mean by the value of *n*
    - $p = \frac{\bar{x}}{n} = \frac{1}{n} \times \frac{\sum fx}{\sum f}$

What are the steps for a chi-squared goodness of fit test for a binomial distribution?

- **STEP 1**: Write the **hypotheses**
  - H_0: Variable *X* can be modelled by a binomial distribution
  - H_1: Variable *X* cannot be modelled by a binomial distribution
    - Make sure you clearly write what the variable is and don't just call it *X*
    - If you are given the assumed value of *p* then state the precise distribution $\mathrm{B}(n, p)$
- **STEP 2**: Calculate the **expected frequencies**
  - If you were not given the assumed value of *p* then you will first have to **estimate it** using the **observed data**
  - Find the probability of the outcome using the binomial distribution $\mathrm{P}(X = x)$
  - Multiply the probability by the total frequency $\mathrm{P}(X = x) \times N$
  - You will have to combine rows/columns if any expected values are 5 or less
- **STEP 3**: Calculate the **degrees of freedom** for the test
  - For *k* outcomes (after combining expected values if needed)
  - Degrees of freedom is
    - $v = k - 1$ if you were **given** the assumed value of *p*
    - $v = k - 2$ if you had to **estimate** the value of *p*
- **STEP 4**: Enter the **frequencies** and the **degrees of freedom** into your GDC
  - Enter the observed and expected frequencies as two separate lists
  - Your GDC will then give you the χ^2 statistic and its *p*-value
  - The χ^2 statistic is denoted as χ^2_{calc}
- **STEP 5**: Decide whether there is **evidence** to **reject the null** hypothesis
  - EITHER compare the **χ^2 statistic** with the given **critical value**
    - If χ^2 statistic **>** critical value then **reject** H_0
    - If χ^2 statistic **<** critical value then **accept** H_0
  - OR compare the *p*-value with the given significance level
    - If *p*-value **<** significance level then **reject** H_0
    - If *p*-value **>** significance level then **accept** H_0
- **STEP 6**: Write your **conclusion**

- If you **reject** $\mathbf{H_0}$
    - There is sufficient evidence to suggest that variable *X* does not follow the binomial distribution $\mathrm{B}(n, p)$
    - Therefore this suggests that the data **does not follow** $\mathrm{B}(n, p)$
- If you **accept** $\mathbf{H_0}$
    - There is insufficient evidence to suggest that variable *X* does not follow the binomial distribution $\mathrm{B}(n, p)$
    - Therefore this suggests that the data **follows** $\mathrm{B}(n, p)$

Worked Example

A stage in a video game has three boss battles. 1000 people try this stage of the video game and the number of bosses defeated by each player is recorded.

Number of bosses defeated	0	1	2	3
Frequency	490	384	111	15

A χ^2 goodness of fit test at the 5% significance level is used to decide whether the number of bosses defeated can be modelled by a binomial distribution with a 20% probability of success.

a)

State the null and alternative hypotheses.

H_0 : Number of bosses defeated can be modelled by the binomial distribution $B(3, 0.2)$

H_1 : Number of bosses defeated can not be modelled by the binomial distribution $B(3, 0.2)$

b)

Assuming the binomial distribution holds, find the expected number of people that would defeat exactly one boss.

Let $X \sim B(3, 0.2)$

Using GDC $P(X=1) = 0.384$

Expected $1000 \times 0.384 = 384$

$\boxed{\text{Expected frequency of } 1 = 384}$

c)

Calculate the p-value for the test.

Find the other expected frequencies

For 0: $1000 \times P(X=0) = 1000 \times 0.512 = 512$

For 2: $1000 \times P(X=2) = 1000 \times 0.096 = 96$

For 3: $1000 \times P(X=3) = 1000 \times 0.008 = 8$

Type two lists into GDC

Observed	490	384	111	15
Expected	512	384	96	8

$\nu = 4 - 1 = 3$

$p = 0.02426\ldots$

$\boxed{p = 0.0243 \ (3sf)}$

d)

State the conclusion of the test. Give a reason for your answer.

$0.0243 < 0.05$

Reject H_0 as p-value < significance level

There is sufficient evidence to suggest that the number of bosses defeated can not be modelled by the binomial distribution $B(3, 0.2)$.

Chi-Squared GOF: Normal

What is a chi-squared goodness of fit test for a normal distribution?

- A chi-squared (χ^2) **goodness of fit test** is used to test data from a sample suggesting that the population has a **normal distribution**
  - You will either be **given a precise normal distribution** to test $\mathrm{N}(\mu, \sigma^2)$ with assumed values for μ and σ
  - Or you will be asked to test whether a normal distribution is **suitable without being given assumed values** for μ and/or σ
    - In this case you will have to calculate an **estimate** for the value of μ and/or σ for the normal distribution
    - Either use your GDC or use the formulae
    - $\bar{x} = \dfrac{\sum fx}{\sum f}$ and $s_{n-1}^2 = \dfrac{n}{n-1} s_n^2$

What are the steps for a chi-squared goodness of fit test for a normal distribution?

- **STEP 1**: Write the **hypotheses**
  - H_0: Variable *X* can be modelled by a normal distribution
  - H_1: Variable *X* cannot be modelled by a normal distribution
    - Make sure you clearly write what the variable is and don't just call it *X*
    - If you are given the assumed values of μ and σ then state the precise distribution $\mathrm{N}(\mu, \sigma^2)$
- **STEP 2:** Calculate the **expected frequencies**
  - If you were not given the assumed values of μ or σ then you will first have to estimate them
  - Find the probability of the outcome using the normal distribution $\mathrm{P}(a < X < b)$
  - Multiply the probability by the total frequency $\mathrm{P}(a < X < b) \times N$
  - You will have to combine rows/columns if any expected values are 5 or less
- **STEP 3**: Calculate the **degrees of freedom** for the test
  - For *k* class intervals (after combining expected values if needed)
  - Degrees of freedom is
    - $v = k - 1$ if you were **given** the assumed values for **both** μ and σ
    - $v = k - 2$ if you had to **estimate either** μ or σ but **not both**
    - $v = k - 3$ if you had to **estimate both** μ and σ
- **STEP 4**: Enter the **frequencies** and the **degrees of freedom** into your GDC
  - Enter the observed and expected frequencies as two separate lists
  - Your GDC will then give you the χ^2 statistic and its *p*-value
  - The χ^2 statistic is denoted as χ^2_{calc}
- **STEP 5**: Decide whether there is **evidence** to **reject the null** hypothesis
  - EITHER compare the **χ^2 statistic** with the given **critical value**
    - If χ^2 statistic > critical value then **reject H_0**
    - If χ^2 statistic < critical value then **accept H_0**
  - OR compare the ***p*-value** with the given **significance level**

    - If *p*-value **<** significance level then **reject** H_0
    - If *p*-value **>** significance level then **accept** H_0
- **STEP 6**: Write your **conclusion**
  - If you **reject** H_0
    - There is sufficient evidence to suggest that variable *X* does not follow the normal distribution $\mathrm{N}(\mu,\ \sigma^2)$
    - Therefore this suggests that the data **does not follow** $\mathrm{N}(\mu,\ \sigma^2)$
  - If you **accept** H_0
    - There is insufficient evidence to suggest that variable *X* does not follow the normal distribution $\mathrm{N}(\mu,\ \sigma^2)$
    - Therefore this suggests that the data **follows** $\mathrm{N}(\mu,\ \sigma^2)$

Worked Example

300 marbled ducks in Quacktown are weighed and the results are shown in the table below.

Mass (g)	Frequency
$m < 450$	1
$450 \le m < 470$	9
$470 \le m < 520$	158
$520 \le m < 570$	123
$m \ge 570$	9

A χ^2 goodness of fit test at the 10% significance level is used to decide whether the mass of a marbled duck can be modelled by a normal distribution with mean 520 g and standard deviation 30 g.

a)
Explain why it is necessary to combine the groups $m < 450$ and $450 \le m < 470$ to create the group $m < 470$ with frequency 10.

Combine categories if expected frequencies are 5 or less

$300 \times P(X < 450 \mid X \sim N(520, 30^2)) = 300 \times 0.00981... = 2.944...$

The expected frequency is less than 5 so combine with the next category.

b)
Calculate the expected frequencies, giving your answers correct to 2 decimal places.

Let $X \sim N(520, 30^2)$

300 × probability

Mass (g)	Probability	Expected frequency
$m < 470$	0.047790...	14.34
$470 \le m < 520$	0.452209...	135.66
$520 \le m < 570$	0.452209...	135.66
$m \ge 570$	0.047790...	14.34

c)

Write down the null and alternative hypotheses.

H_0 : Mass of the marbled ducks can be modelled by the normal distribution $N(520, 30^2)$

H_1 : Mass of the marbled ducks can not be modelled by the normal distibution $N(520, 30^2)$

d)

Calculate the χ^2 statistic.

Enter the observed and expected frequencies into GDC $\nu = 4 - 1 = 3$

χ^2 statistic $= 8.162\ldots$

$\chi^2_{calc} = 8.16$ (3sf)

e)

Given that the critical value is 6.251, state the conclusion of the test. Give a reason for your answer.

$8.16 > 6.251$

Reject H_0 as χ^2 statistic > critical value.

There is sufficient evidence to suggest that the mass of the marbled ducks can not be modelled by the normal distribution $N(520, 30^2)$.

Chi-squared GOF: Poisson

What is a chi-squared goodness of fit test for a Poisson distribution?

- A chi-squared (χ^2) goodness of fit test is used to test data from a sample suggesting that the population has a Poisson distribution
  - You will either be **given a precise Poisson distribution** to test $\mathrm{Po}(m)$ with an assumed value for *m*
  - Or you will be asked to test whether a Poisson distribution is **suitable without being given an assumed value** for *m*
    - In this case you will have to calculate an **estimate** for the value of *m* for the Poisson distribution
    - To calculate it just calculate the mean
    - $m = \dfrac{\sum fx}{\sum f}$

What are the steps for a chi-squared goodness of fit test for a Poisson distribution?

- **STEP 1**: Write the **hypotheses**
  - H_0: Variable *X* can be modelled by a Poisson distribution
  - H_1: Variable *X* cannot be modelled by a Poisson distribution
    - Make sure you clearly write what the variable is and don't just call it *X*
    - If you are given the assumed value of *m* then state the precise distribution $\mathrm{Po}(m)$
- **STEP 2**: Calculate the **expected frequencies**
  - If you were not given the assumed value of *m* then you will first have to **estimate it** using the **observed data**
  - Find the probability of the outcome using the Poisson distribution $\mathrm{P}(X = x)$
  - Multiply the probability by the total frequency $\mathrm{P}(X = x) \times N$
    - If *a* is the smallest observed value then calculate $\mathrm{P}(X \le a)$
    - If *b* is the largest observed value then calculate $\mathrm{P}(X \ge b)$
  - You will have to combine rows/columns if any expected values are 5 or less
- **STEP 3**: Calculate the **degrees of freedom** for the test
  - For *k* outcomes (after combining expected values if needed)
  - Degree of freedom is
    - $\nu = k - 1$ if you were **given** the assumed value of *m*
    - $\nu = k - 2$ if you had to **estimate** the value of *m*
- **STEP 4**: Enter the **frequencies** and the **degree of freedom** into your GDC
  - Enter the observed and expected frequencies as two separate lists
  - Your GDC will then give you the χ^2 statistic and its *p*-value
  - The χ^2 statistic is denoted as χ^2_{calc}
- **STEP 5**: Decide whether there is **evidence** to **reject the null** hypothesis
  - EITHER compare the **χ^2 statistic** with the given **critical value**
    - If χ^2 statistic **>** critical value then **reject** H_0
    - If χ^2 statistic **<** critical value then **accept** H_0
  - OR compare the *p*-value with the given significance level
    - If *p*-value **<** significance level then **reject** H_0

        - If p-value > significance level then **accept** H_0
- **STEP 6**: Write your **conclusion**
    - If you **reject** H_0
        - There is sufficient evidence to suggest that variable X does not follow the Poisson distribution $\text{Po}(m)$
        - Therefore this suggests that the data **does not follow** $\text{Po}(m)$
    - If you **accept** H_0
        - There is insufficient evidence to suggest that variable X does not follow the Poisson distribution $\text{Po}(m)$
        - Therefore this suggests that the data **follows** $\text{Po}(m)$

Worked Example

A parent claims the number of messages they receive from their teenage child within an hour can be modelled by a Poisson distribution. The parent collects data from 100 one hour periods and records the observed frequencies of the messages received from the child. The parent calculates the mean number of messages received from the sample and uses this to calculate the expected frequencies if a Poisson model is used.

Number of messages	Observed frequency	Expected frequency
0	9	7.28
1	16	a
2	23	24.99
3	22	21.82
4	16	14.29
5	14	7.49
6 or more	0	b

A χ^2 goodness of fit test at the 10% significance level is used to test the parent's claim.

a)

Write down null and alternative hypotheses to test the parent's claim.

We are not given a specific Poisson distribution

H_0: Number of messages received can be modelled by a Poisson distribution

H_1: Number of messages received can not be modelled by a Poisson distribution

b)

Show that the mean number of messages received per hour for the sample is 2.62.

$$m = \frac{\Sigma fx}{\Sigma f} = \frac{0\times 9 + 1\times 16 + 2\times 23 + 3\times 22 + 4\times 16 + 5\times 14}{9 + 16 + 23 + 22 + 16 + 14} = 2.62$$

c)

Calculate the values of a and b, giving your answers to 2 decimal places.

Let $X \sim Po(2.62)$

$a = 100 \times P(X=1) = 100 \times 0.19074... = 19.074$ $\boxed{a = 19.07 \ (2dp)}$

$b = 100 \times P(X \geq 6) = 100 \times 0.05052... = 5.05$ $\boxed{b = 5.05 \ (2dp)}$

d)

Perform the hypothesis test.

Calculate degree of freedom $\nu = k - 2$ m was estimated

$\nu = 7 - 2 = 5$

Enter observed and expected frequencies in GDC

$p = 0.03515... < 0.1$

Reject H_0 as p-value < significance level.
There is sufficient evidence to suggest that a Poisson distribution can not model the number of messages received.

4.12 Further Hypothesis Testing

4.12.1 Hypothesis Testing for Mean (One Sample)

One-Sample z-tests

What is a one-sample *z*-test?

- A **one-sample *z*-test** is used to **test the mean (μ)** of a **normally distributed** population
  - You use a *z*-test when the **population variance (σ^2) is known**
- The **mean of a sample** of size *n* is calculated $\overline{x}$ and a normal distribution is used to test the **test statistic**
- $\overline{x}$ can be used as the test statistic
  - In this case you would use the distribution $\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)$
    - Remember when using this distribution that the standard deviation is $\frac{\sigma}{\sqrt{n}}$
- $z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ can be used as the test statistic
  - In this case you would use the distribution $Z \sim \mathrm{N}(0, 1^2)$
    - This is a more old-fashioned approach but your GDC still might tell you the *z*-value when you do the test
    - You will not need to use this method in the exam as your GDC should be capable of doing the other method

What are the steps for performing a one-sample *z*-test on my GDC?

- **STEP 1**: **Write the hypotheses**
  - $H_0: \mu = \mu_0$
    - Clearly state that μ represents the **population mean**
    - μ_0 is the **assumed population mean**
  - For a **one-tailed** test $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$
  - For a **two-tailed** test: $H_1: \mu \neq \mu_0$
    - The alternative hypothesis will depend on what is being tested
- **STEP 2**: **Enter the data** into your GDC and choose the **one-sample *z*-test**
  - If you have the raw data
    - Enter the data as a list
    - Enter the value of σ
  - If you have summary statistics
    - Enter the values of $\overline{x}$, σ and *n*
  - Your GDC will give you the *p*-value
- **STEP 3**: **Decide** whether there is **evidence to reject the null hypothesis**
  - If the *p*-value < significance level then reject H_0
- **STEP 4**: **Write your conclusion**

- If you **reject H_0** then there is **evidence** to suggest that...
  - The mean has decreased (for $H_1: \mu < \mu_0$)
  - The mean has increased (for $H_1: \mu > \mu_0$)
  - The mean has changed (for $H_1: \mu \neq \mu_0$)
- If you **accept H_0** then there is **insufficient evidence** to reject the null hypothesis which suggests that...
  - The mean has not decreased (for $H_1: \mu < \mu_0$)
  - The mean has not increased (for $H_1: \mu > \mu_0$)
  - The mean has not changed (for $H_1: \mu \neq \mu_0$)

How do I find the *p*-value for a one-sample *z*-test using a normal distribution?

- The *p*-value is determined by the **test statistic** $\bar{x}$
- For $H_1: \mu < \mu_0$ the *p*-value is $P\left(\bar{X} < \bar{x} \mid \mu = \mu_0\right)$
- For $H_1: \mu > \mu_0$ the *p*-value is $P\left(\bar{X} > \bar{x} \mid \mu = \mu_0\right)$
- For $H_1: \mu \neq \mu_0$ the *p*-value is $P\left(\left|\bar{X} - \mu_0\right| > \left|x - \mu_0\right| \mid \mu = \mu_0\right)$
  - If $\bar{x} < \mu_0$ then this can be calculated easier by $2 \times P\left(\bar{X} < \bar{x} \mid \mu = \mu_0\right)$
  - If $\bar{x} > \mu_0$ then this can be calculated easier by $2 \times P\left(\bar{X} > \bar{x} \mid \mu = \mu_0\right)$

How do I find the critical value and critical region for a one-sample *z*-test?

- The critical region is determined by the **significance level** $\alpha\%$
  - For $H_1: \mu < \mu_0$ the critical region is $\bar{X} < c$ where $P\left(\bar{X} < c \mid \mu = \mu_0\right) = \alpha\%$
  - For $H_1: \mu > \mu_0$ the critical region is $\bar{X} > c$ where $P\left(\bar{X} > c \mid \mu = \mu_0\right) = \alpha\%$
  - For $H_1: \mu \neq \mu_0$ the critical regions are $\bar{X} < c_1$ and $\bar{X} > c_2$ where $P\left(\bar{X} < c_1 \mid \mu = \mu_0\right) = P\left(\bar{X} > c_2 \mid \mu = \mu_0\right) = \frac{1}{2}\alpha\%$
- The critical value(s) can be found using the inverse normal distribution function
  - When rounding the critical value(s) you should choose:
    - The **lower bound** for the inequalities $\bar{X} < c$
    - The **upper bound** for the inequalities $\bar{X} > c$
  - This is so that the probability **does not exceed the significance level**

Exam Tip

- Exam questions might specify a method for you to use so practise all methods (using GDC, *p*-values, critical regions)
- If the exam question does not specify a method then use whichever method you want
  - Make it clear which method you are using
  - You can always use a second method as a way of checking your answer

Worked Example

The mass of a Burmese cat, C, follows a normal distribution with mean 4.2 kg and a standard deviation 1.3 kg. Kamala, a cat breeder, claims that Burmese cats weigh more than the average if they live in a household which contains young children. To test her claim, Kamala takes a random sample of 25 cats that live in households containing young children.

a)

State the null and alternative hypotheses to test Kamala's claim.

Let μ be the population mean for the mass of Burmese cats

$$\boxed{\begin{array}{l} H_0: \mu = 4.2 \\ H_1: \mu > 4.2 \end{array}}$$

Testing for an increase

b)

Using a 5% level of significance, find the critical region for this test.

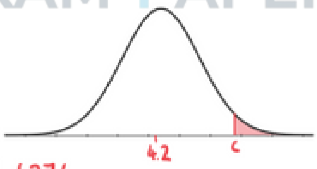
The population variance is known so use a z-test

Find the distribution of the sample means $N\left(\mu, \frac{\sigma^2}{n}\right)$

$\bar{C} \sim N\left(\mu, \frac{1.3^2}{25}\right)$ Square the standard deviation

The sample size

Critical region is $\bar{C} > c$ where $P(\bar{C} > c \mid \mu = 4.2) = 0.05$



Use inverse normal:

$\mu = 4.2$

$\sigma = \sqrt{\frac{1.3^2}{25}} = 0.26$

$P(\bar{C} < c) = 0.95$

$c = 4.6276...$

$$\boxed{\text{Critical region } \bar{C} > 4.63}$$

c)

Kamala calculates the mean of the 25 cats included in her sample to be 4.65 kg. Determine the conclusion of the test.

$4.65 > 4.6276...$ so 4.65 is in critical region

> Reject H_0 as test statistic is in critical region. There is sufficient evidence to suggest that Burmese cats weigh more if they live in a household which contains young children.

One-Sample t-tests

What is a one-sample *t*-test?

- A **one-sample *t*-test** is used to **test the mean (μ)** of a **normally distributed** population
    - You use a *t*-test when the **population variance (σ^2) is unknown**
    - You need to use the unbiased estimate for the population variance (s^2_{n-1})
- The **mean of a sample** of size *n* is calculated $\bar{x}$ and a *t*-distribution is used to test it
    - *t*-distributions are similar to normal distributions
        - As the sample size gets larger the *t*-distribution tends towards the standard normal distribution
- You won't be expected to find the critical value
    - The *p*-value can be found using the test function on your GDC

What are the steps for performing a one-sample *t*-test on my GDC?

- **STEP 1**: **Write the hypotheses**
    - $H_0: \mu = \mu_0$
        - Clearly state that μ represents the **population mean**
        - μ_0 is the **assumed population mean**
    - For a **one-tailed** test $H_1: \mu < \mu_0$ or $H_1: \mu > \mu_0$
    - For a **two-tailed** test: $H_1: \mu \neq \mu_0$
        - The alternative hypothesis will depend on what is being tested
- **STEP 2**: **Enter the data** into your GDC and choose the **one-sample *t*-test**
    - If you have the raw data
        - Enter the data as a list
    - If you have summary statistics
        - Enter the values of $\bar{x}$, s_{n-1} (sometimes written as s_x on a GDC) and *n*
    - Your GDC will give you the *p*-value
- **STEP 3**: **Decide** whether there is **evidence to reject the null hypothesis**
    - If the *p*-value < significance level then reject H_0
- **STEP 4**: **Write your conclusion**
    - If you **reject H_0** then there is **evidence** to suggest that...
        - The mean has decreased (for $H_1: \mu < \mu_0$)
        - The mean has increased (for $H_1: \mu > \mu_0$)
        - The mean has changed (for $H_1: \mu \neq \mu_0$)
    - If you **accept H_0** then there is **insufficient evidence** to reject the null hypothesis which suggests that...
        - The mean has not decreased (for $H_1: \mu < \mu_0$)
        - The mean has not increased (for $H_1: \mu > \mu_0$)
        - The mean has not changed (for $H_1: \mu \neq \mu_0$)

Worked Example

The IQ of a student at Calculus High can be modelled as a normal distribution with mean 126. The headteacher decides to play classical music during lunchtimes and suspects that this has caused a change in the average IQ of the students.

a)

State the null and alternative hypotheses to test the headteacher's suspicion.

Let μ be the population mean for the IQ of a student at Calculus High

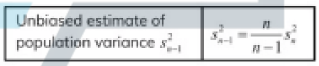
$$\boxed{\begin{aligned} H_0 &: \mu = 126 \\ H_1 &: \mu \neq 126 \end{aligned}}$$

Testing for a change

b)

The headteacher selects 15 students and asks them to complete an IQ test. The mean score for the sample is 127.1 and the sample variance is 14.7. Calculate the unbiased estimate for the population variance s^2_{n-1}.

Formula booklet



Unbiased estimate of population variance s^2_{n-1}	$s^2_{n-1} = \frac{n}{n-1} s^2_n$

$s^2_{n-1} = \frac{15}{14} \times 14.7$

$$\boxed{s^2_{n-1} = 15.75}$$

c)

Calculate the p-value for the test.

The population variance is unknown so use a t-test

Enter summary statistics into GDC using one-sample t-test

$\bar{x} = 127.1 \quad s_{n-1} = \sqrt{15.75} \quad n = 15$

$p = 0.3012\ldots$

$$\boxed{p = 0.301 \ (3sf)}$$

d)

State whether the headteacher's suspicion is supported by the test.

$0.3012... > 0.1$

Accept H_0 as p-value > significance level.
There is insufficient evidence to support the headteacher's suspicion.

4.12.2 Hypothesis Testing for Mean (Two Sample)

Two-Sample Tests

What is a two-sample test?

- A **two-sample test** is used to **compare the means (μ_1 & μ_2)** of **two normally distributed** populations
  - You use a ***z*-test** when the **population variances (σ_1^2 & σ_2^2) are known**
  - You use a ***t*-test** when the **population variances are unknown**
    - In this course you will assume the **variances are equal** and use a **pooled sample** for a *t*-test
    - In a pooled sample the data from both samples are used to estimate the population variance

What are the steps for performing a two-sample test on my GDC?

- **STEP 1: Write the hypotheses**
  - $H_0: \mu_1 = \mu_2$
    - Clearly state that μ_1 & μ_2 represent the **population means**
    - Make sure you make it clear which mean corresponds to which population
    - In words this means that the two population means are equal
  - For a **one-tailed** test $H_1: \mu_1 < \mu_2$ or $H_1: \mu_1 > \mu_2$
  - For a **two-tailed** test: $H_1: \mu_1 \neq \mu_2$
    - The alternative hypothesis will depend on what is being tested
- **STEP 2**: Decide if it is a ***z*-test or a *t*-test**
  - If the populations variances are **known** then use a ***z*-test**
  - If the populations variances are **unknown** then use a ***t*-test**
    - Assume the variances are equal and use a pooled sample
- **STEP 3**: **Enter the data** into your GDC and choose **two-sample *z*-test or two-sample *t*-test**
  - If you have the raw data
    - Enter the data as a list
    - Enter the values of σ_1 & σ_2 if a *z*-test
    - Choose the pooled option if a *t*-test
  - If you have summary statistics (only for a *z*-test)
    - Enter the values of $\bar{x}_1$, $\bar{x}_2$, σ_1, σ_2, n_1 & n_2
  - Your GDC will give you the *p*-value
- **STEP 4**: **Decide** whether there is **evidence to reject the null hypothesis**
  - If the *p*-value < significance level then reject H_0
- **STEP 5**: **Write your conclusion**
  - If you **reject H_0** then there is evidence to suggest that...
    - The mean of the 1st population is smaller (for $H_1: \mu_1 < \mu_2$)
    - The mean of the 1st population is bigger (for $H_1: \mu_1 > \mu_2$)
    - The means of the two populations are different (for $H_1: \mu_1 \neq \mu_2$)

- If you **accept H_0** then there is **insufficient evidence** to reject the null hypothesis which suggests that...
  - The mean of the 1st population is not smaller (for $H_1: \mu_1 < \mu_2$)
  - The mean of the 1st population is not bigger (for $H_1: \mu_1 > \mu_2$)
  - The means of the two populations are not different (for $H_1: \mu_1 \neq \mu_2$)

Worked Example

The times (in minutes) for children and adults to complete a puzzle are recorded below.

Children	3.1	2.7	3.5	3.1	2.9	3.2	3.0	2.9		
Adults	3.1	3.6	3.5	3.6	2.9	3.6	3.4	3.6	3.7	3.0

The creator of the puzzle claims children are generally faster at solving the puzzle than adults. A *t*-test is to be performed at a 1% significance level.

a)

Write down the null and alternative hypotheses.

Let μ_c be the population mean for children's times

and μ_A be the population mean for adults' times

$H_0: \mu_c = \mu_A$

$H_1: \mu_c < \mu_A$

It is claimed that children are quicker

b)

Find the *p*-value for this test.

Enter the data as two lists in GDC

Use 2-sample pooled t-test

$p = 0.007259...$

$p = 0.00726$ (3sf)

c)

State whether the creator's claim is supported by the test. Give a reason for your answer.

$0.00726 < 0.01$

Reject H_0 as p-value < significance level.
There is sufficient evidence to suggest that children are generally faster at solving the puzzle than adults. This supports the creator's claim.

Paired t-tests

What is a paired *t*-test?

- A paired test is where you take **two samples** but **each data point from one sample can be paired with a data point from the other sample**
  - These are used when one group of members are used twice and the two results for each member are paired
    - It could be to compare the sample before and after introducing a new factor
    - It could be to compare the sample under two different conditions
- For this test you use the differences between the pairs and treat them as one sample
  - As the variance of the differences is unlikely to be known you will use a *t*-test
  - For a paired test you need to assume the differences are normally distributed
    - You don't need to assume the populations are normally distributed

What are the steps for performing a paired *t*-test on my GDC?

- **STEP 1**: **Write the hypotheses**
  - $H_0: \mu_D = 0$
    - Clearly state that μ_D represents the **population mean of the differences**
    - In words this means the population mean has not changed
  - For a **one-tailed** test $H_1: \mu_D < 0$ or $H_1: \mu_D > 0$
  - For a **two-tailed** test: $H_1: \mu_D \neq 0$
    - The alternative hypothesis will depend on what is being tested
- **STEP 2**: **Enter the data** into your GDC and choose the **one-sample *t*-test**
  - Enter the differences as a list
    - Be consistent with the order in which you subtract paired values
  - Your GDC will give you the *p*-value
- **STEP 3**: **Decide** whether there is **evidence to reject the null hypothesis**
  - If the *p*-value < significance level then reject H_0
- **STEP 4**: **Write your conclusion**
  - If you **reject H_0** then there is evidence to suggest that...
    - The mean has decreased (for $H_1: \mu_D < 0$)
    - The mean has increased (for $H_1: \mu_D > 0$)
    - The mean has changed (for $H_1: \mu_D \neq 0$)
  - If you **accept H_0** then there is **insufficient evidence** to reject the null which suggests that...
    - The mean has not decreased (for $H_1: \mu_D < 0$)
    - The mean has not increased (for $H_1: \mu_D > 0$)
    - The mean has not changed (for $H_1: \mu_D \neq 0$)

Exam Tip

- If an exam question has two samples with the same number of members then consider carefully whether it makes sense to do a paired test or a two sample test
- The examiner might make it look like it is a paired test to trick you!

Worked Example

In a school all students must study French and Spanish. 9 students are selected and complete a test in both subjects, the standardised scores are shown below

Student	1	2	3	4	5	6	7	8	9
French score	61	82	77	80	99	69	75	71	81
Spanish score	74	79	83	66	95	79	82	81	85

The headteacher wants to investigate whether there is a difference in the students' scores between the two subjects. A paired *t*-test is to be performed at a 10% significance level.

a)

Write down the null and alternative hypotheses.

Let D be the French score minus the Spanish score for the students. Let μ_D be the mean difference for the whole population of students.

$$\boxed{\begin{array}{l} H_0: \mu_D = 0 \\ H_1: \mu \neq 0 \end{array}}$$

Testing for a difference in scores

b)

Find the *p*-value for this test.

Calculate the difference for each student d = French - Spanish

Student	1	2	3	4	5	6	7	8	9
d	-13	3	-6	14	4	-10	-7	-10	-4

Enter the differences into the GDC and use a t-test

$p = 0.2958...$

$$\boxed{p = 0.296 \text{ (3sf)}}$$

c)

Write down the conclusion to the test. Give a reason for your answer.

$0.2958... > 0.1$

Accept H_0 as p-value > significance level.
There is insufficient evidence to suggest that there is a difference in scores.

4.12.3 Binomial Hypothesis Testing

Binomial Hypothesis Testing

What is a hypothesis test using a binomial distribution?

- You can use **a binomial distribution** to test whether the **proportion** of a population with a specified characteristic has **increased** or **decreased**
  - These tests will always be **one-tailed**
  - You will not be expected to perform a two-tailed hypothesis test with the binomial distribution
- A sample will be taken and the **test statistic** x will be the **number of members with the characteristic** which will be tested using the distribution $X \sim \mathrm{B}(n, p)$
  - This can be thought of as the number of successes

What are the steps for a hypothesis test of a binomial proportion?

- **STEP 1: Write the hypotheses**
  - $\mathrm{H}_0 : p = p_0$
    - Clearly state that p represents the **population proportion**
    - p_0 is the **assumed population proportion**
  - $\mathrm{H}_1 : p < p_0$ or $\mathrm{H}_1 : p > p_0$
- **STEP 2**: Calculate the ***p*-value** or find the **critical region**
  - See below
- **STEP 3: Decide** whether there is **evidence to reject the null hypothesis**
  - If the p-value < significance level then reject H_0
  - If the test statistic is in the critical region then reject H_0
- **STEP 4: Write your conclusion**
  - If you **reject H_0** then there is evidence to suggest that...
    - The population proportion has decreased (for $\mathrm{H}_1 : p < p_0$)
    - The population proportion has increased (for $\mathrm{H}_1 : p > p_0$)
  - If you **accept H_0** then there is **insufficient evidence** to reject the null hypothesis which suggests that...
    - The population proportion has not decreased (for $\mathrm{H}_1 : p < p_0$)
    - The population proportion has not increased (for $\mathrm{H}_1 : p > p_0$)

How do I calculate the *p*-value?

- The p-value is determined by the **test statistic** x
- The p-value is the probability that 'a value being **at least as extreme** as the test statistic' would occur if **null hypothesis were true**
  - For $\mathrm{H}_1 : p < p_0$ the p-value is $\mathrm{P}(X \le x \mid p = p_0)$
  - For $\mathrm{H}_1 : p > p_0$ the p-value is $\mathrm{P}(X \ge x \mid p = p_0)$

How do I find the critical value and critical region?

- The critical value and critical region are determined by the **significance level** $\alpha\%$
- Your calculator might have an **inverse binomial function** that works just like the inverse normal function

  - You need to use this value to find the critical value
  - The value given by the inverse binomial function is normally one away from the actual critical value
- For $H_1 : p < p_0$ the critical region is $X \le c$ where c is the critical value
  - c is the **largest integer** such that $\mathrm{P}(X \le c | p = p_0) \le \alpha\%$
    - Check that $\mathrm{P}(X \le c + 1 | p = p_0) > \alpha\%$
- For $H_1 : p > p_0$ the critical region is $X \ge c$ where c is the critical value
  - c is the **smallest integer** such that $\mathrm{P}(X \ge c | p = p_0) \le \alpha\%$
    - Check that $\mathrm{P}(X \ge c - 1 | p = p_0) > \alpha\%$

Worked Example

The existing treatment for a disease is known to be effective in 85% of cases. Dr Sabir develops a new treatment which she claims is more effective than the existing one. To test her claim she uses the new treatment on a random sample of 60 patients with the disease and finds that the treatment was effective for 57 of them.

a)

State null and alternative hypotheses to test Dr Sabir's claim.

Let p be the proportion of the population for which the new treatment is effective.

$$\boxed{\begin{aligned} H_0 &: p = 0.85 \\ H_1 &: p > 0.85 \end{aligned}}$$

Testing for an increase

b)

Perform the test using a 1% significance level. Clearly state the conclusion in context.

Let $X \sim B(60, p)$ be the number of people in the sample for which the new treatment is effective

Find the p-value and compare to the significance level

$p = P(X \geq 57 \mid p = 0.85) = 0.01483... > 0.01$

Accept H_0 as p-value > significance level.
There is insufficient evidence to suggest that the new treatment is more effective than the existing treatment.

4.12.4 Poisson Hypothesis Testing

Poisson Hypothesis Testing

What is a hypothesis test using a Poisson distribution?

- You can use **a Poisson distribution** to test whether the **mean number of occurrences** for a **given time period** within a population has **increased** or **decreased**
  - These tests will always be **one-tailed**
  - You will not be expected to perform a two-tailed hypothesis test with the Poisson distribution
- A sample will be taken and the **test statistic** x will be the **number of occurrences** which will be tested using the distribution $X \sim \text{Po}(m)$

What are the steps for a hypothesis test of a Poisson proportion?

- **STEP 1: Write the hypotheses**
  - $\text{H}_0 : m = m_0$
    - Clearly state that m represents the **mean number of occurrences** for the **given time period**
    - m_0 is the **assumed mean number of occurrences**
    - You might have to use **proportion** to find m_0
  - $\text{H}_1 : m < m_0$ or $\text{H}_1 : m > m_0$
- **STEP 2**: Calculate the **p-value** or find the **critical region**
  - See below
- **STEP 3: Decide** whether there is **evidence to reject the null hypothesis**
  - If the p-value < significance level then reject H_0
  - If the test statistic is in the critical region then reject H_0
- **STEP 4**: **Write your conclusion**
  - If you **reject H_0** then there is evidence to suggest that...
    - The mean number of occurrences has decreased (for $\text{H}_1 : m < m_0$)
    - The mean number of occurrences has increased (for $\text{H}_1 : m > m_0$)
  - If you **accept H_0** then there is **insufficient evidence** to reject the null hypothesis which suggests that...
    - The mean number of occurrences has not decreased (for $\text{H}_1 : m < m_0$)
    - The mean number of occurrences has not increased (for $\text{H}_1 : m > m_0$)

How do I calculate the p-value?

- The p-value is determined by the **test statistic** x
- The p-value is the probability that 'a value being **at least as extreme** as the test statistic' would occur if **null hypothesis were true**
  - For $\text{H}_1 : m < m_0$ the p-value is $\text{P}(X \le x \mid m = m_0)$
  - For $\text{H}_1 : m > m_0$ the p-value is $\text{P}(X \ge x \mid m = m_0)$

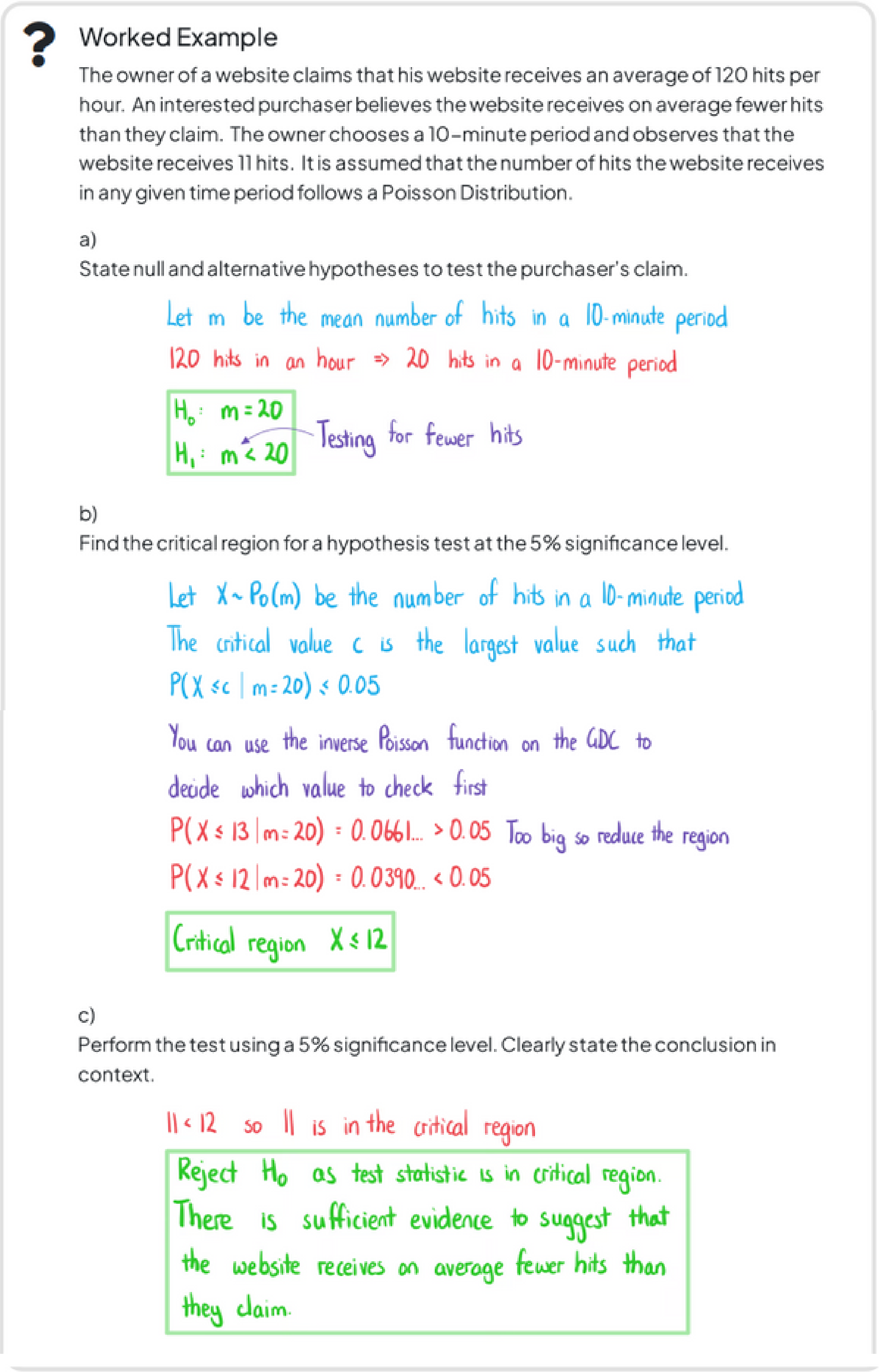
How do I find the critical value and critical region?

- The critical value and critical region are determined by the **significance level** $\alpha\%$

- Your calculator might have an **inverse Poisson function** that works just like the inverse normal function
  - You need to use this value to find the critical value
  - The value given by the inverse Poisson function is normally one away from the actual critical value
- For $H_1 : m < m_0$ the critical region is $X \le c$ where c is the critical value
  - c is the **largest integer** such that $\mathrm{P}(X \le c | m = m_0) \le \alpha\%$
    - Check that $\mathrm{P}(X \le c + 1 | m = m_0) > \alpha\%$
- For $H_1 : m > m_0$ the critical region is $X \ge c$ where c is the critical value
  - c is the **smallest integer** such that $\mathrm{P}(X \ge c | m = m_0) \le \alpha\%$
    - Check that $\mathrm{P}(X \ge c - 1 | m = m_0) > \alpha\%$

Exam Tip

- In an exam it is very important to state the time period for your variable
- Make sure the mean used in the null hypothesis is for the stated time period



Worked Example

The owner of a website claims that his website receives an average of 120 hits per hour. An interested purchaser believes the website receives on average fewer hits than they claim. The owner chooses a 10-minute period and observes that the website receives 11 hits. It is assumed that the number of hits the website receives in any given time period follows a Poisson Distribution.

a)

State null and alternative hypotheses to test the purchaser's claim.

Let m be the mean number of hits in a 10-minute period

120 hits in an hour ⇒ 20 hits in a 10-minute period

$$\boxed{\begin{aligned} H_0 &: m = 20 \\ H_1 &: m < 20 \end{aligned}}$$

Testing for fewer hits

b)

Find the critical region for a hypothesis test at the 5% significance level.

Let $X \sim Po(m)$ be the number of hits in a 10-minute period

The critical value c is the largest value such that

$P(X \leq c \mid m = 20) \leq 0.05$

You can use the inverse Poisson function on the GDC to decide which value to check first

$P(X \leq 13 \mid m = 20) = 0.0661... > 0.05$ Too big so reduce the region

$P(X \leq 12 \mid m = 20) = 0.0390... < 0.05$

$\boxed{\text{Critical region } X \leq 12}$

c)

Perform the test using a 5% significance level. Clearly state the conclusion in context.

$11 < 12$ so 11 is in the critical region

Reject H_0 as test statistic is in critical region. There is sufficient evidence to suggest that the website receives on average fewer hits than they claim.

4.12.5 Hypothesis Testing for Correlation

Hypothesis Testing for Correlation

What is a hypothesis test for correlation?

- You can use a **t-test** to test whether there is **linear correlation** between two normally distributed variables
  - If specifically testing for positive (or negative) linear correlation then a **one-tailed test** is used
  - If testing for any linear correlation then a **two-tailed test** is used
- A sample will be taken and the **raw data** will be given
  - You might be asked to calculate the **PMCC (Pearson's product-moment correlation coefficient)**

What are the steps for a hypothesis test for correlation?

- **STEP 1**: **Write the hypotheses**
  - $H_0: \rho = 0$
    - Clearly state that ρ represents **population correlation coefficient** between the two variables
    - In words this means there is no correlation
  - $H_1: \rho < 0$, $H_1: \rho > 0$ or $H_1: \rho \neq 0$
- **STEP 2**: Calculate the **p-value** or the **PMCC**
  - Choose a t-test for linear regression
  - Enter the data as two lists into GDC
- **STEP 3**: **Decide** whether there is evidence to **reject the null hypothesis**
  - If the p-value < significance level then reject H_0
  - If the absolute value of the PMCC is bigger than the absolute value of the critical value then reject H_0
    - If you are expected to use the PMCC you will be **given the critical value** in the exam
- **STEP 4**: **Write your conclusion**
  - If you **reject H_0** then there is evidence to suggest that...
    - There is a negative linear correlation between the two variables (for $H_1: \rho < 0$)
    - There is a positive linear correlation between the two variables (for $H_1: \rho > 0$)
    - There is a linear correlation between the two variables (for $H_1: \rho \neq 0$)
  - If you **accept H_0** then there is **insufficient evidence** to reject the null hypothesis which suggests that...
    - There is not a negative linear correlation between the two variables (for $H_1: \rho < 0$)
    - There is not a positive linear correlation between the two variables (for $H_1: \rho > 0$)
    - There is not a linear correlation between the two variables (for $H_1: \rho \neq 0$)

Worked Example

Jessica wants to test whether there is any linear correlation between the distance she runs in a day, d km, and the amount of sleep she has the night after her run, t hours. Over the period of a month she takes a random sample of 9 days, the results are recorded in the table.

Distance (d km)	1.2	2.3	1.5	1.3	2.5	1.8	1.9	2.0	1.1
Sleep (t hours)	7.9	8.1	7.6	7.3	8.1	8.4	7.8	7.9	6.8

a)

Write down null and alternative hypotheses that Jessica can use for her test.

Let ρ be the correlation coefficient between Jessica's distances and the hours of sleep she gets.

$$\boxed{\begin{array}{l} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}}$$

Testing for any linear correlation

b)

Perform the hypothesis test for linear correlation using a 5% significance level. Clearly state your conclusion.

Type the data in GDC and select a t-test for linear regression

$p = 0.03833... < 0.05$

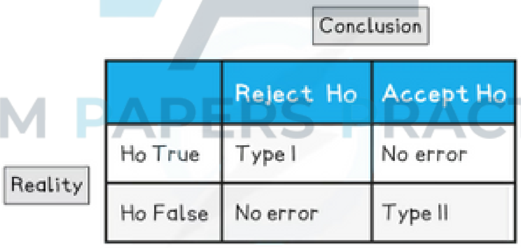
Reject H_0 as p-value < significance level. There is sufficient evidence to suggest that there is a linear correlation between the distance that Jessica runs and the hours she sleeps.

4.12.6 Type I & Type II Errors

Type I & Type II Errors

What are Type I & Type II errors?

- There are **four possible outcomes** of a hypothesis test:
    - H_0 is **false** and H_0 is **rejected**
    - H_0 is **true** and H_0 is **not rejected**
        - The test is **accurate** for these two outcomes
    - H_0 is **true** and H_0 is **rejected**
    - H_0 is **false** and H_0 is **not rejected**
        - The test has led to an **error** for these two outcomes
- A **Type I error** occurs when a hypothesis test gives **sufficient evidence to reject H_0** despite it **being true**
    - This is sometimes called a "**false positive**"
    - In a court case this would be when the defendant is found **guilty despite being innocent**
- A **Type II error** is when a hypothesis test gives **insufficient evidence to reject H_0** despite it **being false**
    - This is sometimes called a "**false negative**"
    - In a court case this would be when the defendant is found **innocent despite being guilty**



Conclusion

		Reject Ho	Accept Ho
Reality	Ho True	Type I	No error
	Ho False	No error	Type II

How do I find the probabilities of a Type I or Type II error occurring?

- You should calculate the probability of errors occurring **before a sample is taken**
- The probabilities are **determined by the critical region**
    - Equally it is **determined by the significance level** $\alpha\%$
    - Critical regions are determined such that:
        - They keep the **probability of a Type I error less than or equal to** the **significance level**
        - They **maximise** the **probability of a Type I error**
- The probability of a **Type I error** occurring is equal to the probability of **being in the critical region** if H_0 were true
    - P(Type I error) = P(being in the critical region | H_0 is true)
    - For a continuous distribution (normal, t, χ^2)
        - P(Type I error) = $\alpha\%$
    - For a discrete distribution (binomial, Poisson)
        - P(Type I error) $\leq \alpha\%$

- The probability of a **Type II** error occurring is equal to the probability of **not being in the critical region** given the actual value of the population parameter
  - P(Type II error) = P(not being in critical region | actual population parameter)
  - You need to know the actual population parameter in order to find the probability of a Type II error
- Once a sample has been taken you can determine which error could have occurred
  - If you **rejected H_0** then you could have made a **Type I error**
  - If you **accepted H_0** then you could have made a **Type II error**

Can I reduce the probabilities of making a Type I or Type II error?

- You can **reduce** the probability of a **Type I** error by **reducing the significance level**
  - However this will **increase** the probability of a **Type II error**
- You can **reduce** the probability of a **Type II** error by **increasing the significance level**
  - However this will **increase** the probability of a **Type I error**
- The only way to **reduce both** probabilities is by **increasing the size of the sample**

Exam Tip

- If an exam question asks you to find the probability of a Type I or II error then double check that the test has not been carried out yet
- The examiner could test your understanding of errors by asking you to state which error could not have occurred once the test has been carried out

Worked Example

Lucy can hit the target 70% of the time when she throws an axe with her right hand. She claims that the proportion, *p*, of her throws that hit the target is higher than 70% when she uses her left hand. Lucy uses the hypotheses $\mathrm{H}_0 : p = 0.7$ and $\mathrm{H}_0 : p > 0.7$ to test her claim. Lucy makes 100 throws and will reject the null hypothesis if the axe hits the target more than 77 times.

a)

find the probability of a Type I error.

Let $X \sim B(100, p)$ be the number of times Lucy hits the target when using her left hand.

P(Type I error) = P(being in critical region | H_0 is true)

$$\begin{aligned} \text{P(Type I error)} &= P(X > 77 \mid p = 0.7) \\ &= P(78 \le X \le 100 \mid p = 0.7) \\ &= 0.04786... \end{aligned}$$

$$\boxed{\text{P(Type I error)} = 0.0479 \ (3sf)}$$

b)

Given that Lucy actually hits the target 80% of the time with her left hand, find the probability of a Type II error.

P(Type II error) = P(not being in critical region | true population parameter)

$$\begin{aligned} \text{P(Type II error)} &= P(X \le 77 \mid p = 0.8) \\ &= P(0 \le X \le 77 \mid p = 0.8) \\ &= 0.2610... \end{aligned}$$

$$\boxed{\text{P(Type II error)} = 0.261 \ (3sf)}$$

4.13 Transition Matrices & Markov Chains

4.13.1 Markov Chains

Markov Chains

What is meant by a "state"?

- States refer to **mutually exclusive events** with the current event **able to change over time**
- Examples of states include:
  - Daily weather conditions
    - The states could be: "sunny" and "not sunny"
  - Countries visited by an inspector each day
    - The states could be: "France", "Spain" and "Germany"
  - Store chosen for weekly grocery shop:
    - The states could be: "Foods-U-Like", "Smiley Shoppers" and "Better Buys"

What is a Markov chain?

- A **Markov chain** is a model that describes a **sequence of states** over a period of time
  - Time is measured in discrete steps
    - Such as days, months, years, etc
- The **conditions** for a Markov chain are:
  - The **probability** of a state being the **next state** in the sequence **only depends** on the **current state**
    - For example
      The 11th state **only depends** on the 10th state
      The first 9 states **do not affect** the 11th state
    - This probability is called a **transition probability**
  - The **transition probabilities do not change** over time
    - For example
      The probability that the 11th state is A given that the 10th state is B is equal to the probability that the 12th state is A given that the 11th state is B
- A Markov chain is said to be **regular** if it possible to reach any state after a finite period of time regardless of the initial state

What is a transition state diagram?

- A **transition diagram** is a **directed graph**
  - The **vertices** are the **states**
  - The **edges** represent the **transition probabilities** between the states
- The graph can contain
  - **Loops**
    - These will be the transition probabilities of the next state being the same as the current state
  - **Two edges between each pair** of vertices
    - The edges will be in opposite directions

  - Each edge will show the transition probability of the state changing in the given direction
- The **probabilities** on the **edges coming out** of a vertex **add up to 1**

Exam Tip

- Drawing a transition state diagram (even when the question does not ask for one) can help you visualise the problem

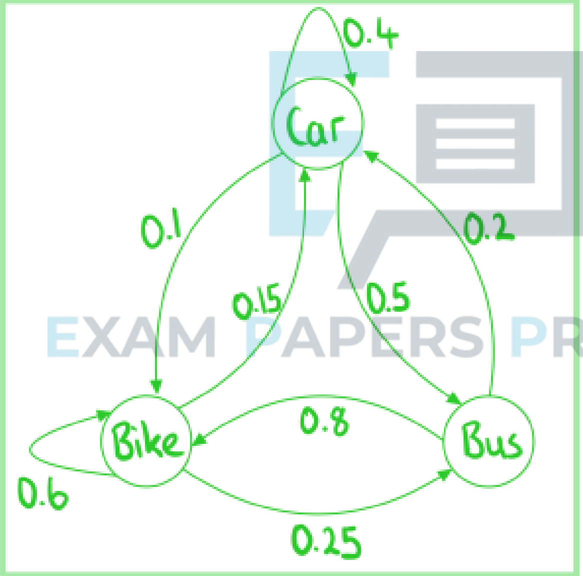
Worked Example

Fleur travels to work by car, bike or bus. Each day she chooses her mode of transport based on the transport she chose the previous day.

- If Fleur travels by car then there is a 40% chance that she will travel by car the following day and a 10% chance that she will travel by bike.
- If Fleur travels by bike then there is a 60% chance that she will travel by bike the following day and a 25% chance that she will travel by bus.
- If Fleur travels by bus then there is an 80% chance that she will travel by bike the following day and a 20% chance that she will travel by car.

Represent this information as a transition state diagram.

The probabilities on the arrows coming out of a state add to 1



4.13.2 Transition Matrices

Transition Matrices

What is a transition matrix?

- A **transition matrix *T*** shows the **transition probabilities** between the current state and the next state
  - The **columns** represent the **current states**
  - The **rows** represent the **next states**
- The element of ***T*** in the i^{th} row and j^{th} column gives the transition probability t_{ij} of :
  - the **next state** being the state corresponding to **row *i***
  - **given that the current state** is the state corresponding to **column *j***
- The probabilities in each **column** must **add up to 1**
- The transition matrix depends on how you assign the states to the columns
  - Each transition matrix for a Markov chain will contain the same elements
    - The rows and columns may be in different orders though
    - E.g. Sunny (S) & Cloudy (C) could be in the order **S then C** or **C then S**

What is an initial state probability matrix?

- An **initial state probability matrix $\mathbf{s}_0$** is a column vector which contains the **probabilities** of each state being chosen as the **initial state**
  - If you know which state was chosen as the initial state then that entry will be 1 and the others will all be zero
- You can find the **state probability matrix $\mathbf{s}_1$** which contains the probabilities of each state being chosen after **one interval of time**
  - $\mathbf{s}_1 = T\mathbf{s}_0$

How do I find expected values after one interval of time?

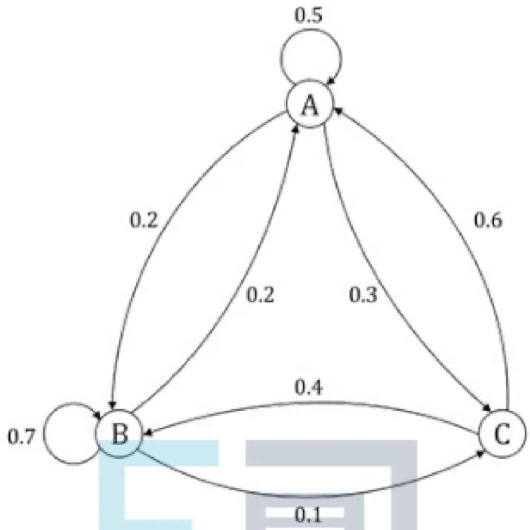
- Suppose the Markov change represents a **population moving between states**
  - Examples include:
    - People in a town switching gyms each year
    - Children choosing a type of sandwich for their lunch each day
- Suppose the **total population is fixed** and equals *N*
- You can **multiply the state probability matrix $\mathbf{s}_1$** by *N* to find the expected number of members of the population at each state

Exam Tip

- If you are asked to find a transition matrix, check that all the probabilities within a column add up to 1
- Drawing a transition state diagram can help you to visualise the problem

Worked Example

Each year Jamie donates to one of three charities: A, B or C. At the start of each year, the probabilities of Jamie continuing donate to the same charity or changing charities are represented by the following transition state diagram:



a)
Write down a transition matrix $\boldsymbol{T}$ for this system of probabilities.

Current state

$$\text{Next state} \quad \begin{array}{c|ccc} & A & B & C \\ \hline A & 0.5 & 0.2 & 0.6 \\ B & 0.2 & 0.7 & 0.4 \\ C & 0.3 & 0.1 & 0 \end{array}$$

$$\boxed{T = \begin{pmatrix} 0.5 & 0.2 & 0.6 \\ 0.2 & 0.7 & 0.4 \\ 0.3 & 0.1 & 0 \end{pmatrix}}$$

b)
There is a 10% chance that charity A is the first charity that Jamie chooses, a 10% chance for charity B and an 80% chance for charity C. Find the charity which has the highest probability of being picked as the second charity after the first year.

Write down the initial state vector $\quad s_0 = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.8 \end{pmatrix}$

$s_1 = Ts_0 \qquad s_1 = \begin{pmatrix} 0.5 & 0.2 & 0.6 \\ 0.2 & 0.7 & 0.4 \\ 0.3 & 0.1 & 0 \end{pmatrix} \begin{pmatrix} 0.1 \\ 0.1 \\ 0.8 \end{pmatrix} = \begin{pmatrix} 0.55 \\ 0.41 \\ 0.04 \end{pmatrix}$

Charity A has the highest probability of being the second charity picked.

Powers of Transition Matrices

How do I find powers of a transition matrix?

- You can simply use your **GDC** to find given powers of a matrix
- The power could be left in terms of an **unknown** *n*
    - In this case it would be more helpful to write the transition matrix in diagonalised form (see section **1.8.2 Applications of Matrices**) $T = PDP^{-1}$ where
        - D is a **diagonal matrix** of the **eigenvalues**
        - P is a matrix of **corresponding eigenvectors**
    - Then $T^n = PD^nP^{-1}$
        - This is given in the **formula booklet**
    - Every transition matrix always has an **eigenvalue equal to 1**

What is represented by the powers of a transition matrix?

- The powers of a transition matrix also **represent probabilities**
- The element of T^n in the i^{th} row and j^{th} column gives the **probability** t^n_{ij} of :
    - the **future state** after ***n* intervals of time** being the state corresponding to **row *i***
    - **given that** the **current state** is the state corresponding to **column *j***
- For example: Let T be a transition matrix with the element $t_{2,3}$ representing the probability that tomorrow is sunny given that it is raining today
    - The element $t^5_{2,3}$ of the matrix T^5 represents the probability that it is sunny in 5 days' time given that it is raining today
- The probabilities in **each column** must still **add up to 1**

How do I find the column state matrices?

- The column state matrix $\mathbf{s_n}$ is a column vector which contains the **probabilities** of each state being chosen after *n* intervals of time given the current state
    - $\mathbf{s_n}$ depends on $\mathbf{s_0}$
- To calculate the column state matrix you raise the transition matrix to the power *n* and multiply by the initial state matrix
    - $T^n s_0 = s_n$
        - You are given this in the **formula booklet**
- You can multiply $\mathbf{s_n}$ by the fixed population size to find the expected number of members of the population at each state after *n* intervals of time

Worked Example

At a cat sanctuary there are 1000 cats. If a cat is brushed on a given day, then the probability it is brushed the following day is 0.2. If a cat is not brushed on a given day, then the probability that is will be brushed the following day is 0.9.

The transition matrix $\boldsymbol{T}$ is used to model this information with $\boldsymbol{T} = \begin{pmatrix} 0.2 & 0.9 \\ 0.8 & 0.1 \end{pmatrix}$.

a)

On Monday Hippo the cat is brushed. Find the probability that Hippo will be brushed on Friday.

Identify the states with the rows/columns

$$\begin{array}{ccc} & \multicolumn{2}{c}{\text{Current}} \\ & B & B' \\ \text{Next} \quad B & 0.2 & 0.9 \\ B' & 0.8 & 0.1 \end{array}$$

Friday is 4 days after Monday

$$T^4 = \begin{pmatrix} 0.2 & 0.9 \\ 0.8 & 0.1 \end{pmatrix}^4 = \underbrace{\begin{array}{cc} B & B' \\ \end{array}}_{\text{Current}} \quad \begin{array}{c} \\ \\ \end{array}$$

$$T^4 = \begin{pmatrix} 0.2 & 0.9 \\ 0.8 & 0.1 \end{pmatrix}^4 = \begin{pmatrix} \boxed{0.6424} & 0.4023 \\ 0.3576 & 0.5977 \end{pmatrix} \begin{matrix} B \\ B' \end{matrix} \Big\} \text{Future}$$

Current = B

Future = B

$\boxed{0.6424}$

b)

On Monday 700 cats were brushed. Find the expected number of cats that will be brushed on the following Monday.

On Monday 700 brushed $\quad s_0 = \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix}$

Expected numbers after 7 days

$\text{Total} \times s_7 = \text{Total} \times T^7 s_0$

$$1000 \times \begin{pmatrix} 0.2 & 0.9 \\ 0.8 & 0.1 \end{pmatrix}^7 \begin{pmatrix} 0.7 \\ 0.3 \end{pmatrix} = \begin{pmatrix} 0.2 & 0.9 \\ 0.8 & 0.1 \end{pmatrix}^7 \begin{pmatrix} 700 \\ 300 \end{pmatrix} = \begin{pmatrix} 515.36309 \\ 484.63691 \end{pmatrix} \begin{matrix} B \\ B' \end{matrix}$$

$\boxed{515 \text{ cats}}$

Steady State & Long-term Probabilities

What is the steady state of a regular Markov chain?

- The vector **s** is said to be a **steady state** vector if it does not change when multiplied by the transition matrix
  - $\boldsymbol{T}\mathbf{s} = \mathbf{s}$
- **Regular Markov chains** have steady states
  - A Markov chain is said to be regular if there exists a **positive integer *k*** such that **none of the entries** are **equal to 0** in the matrix $\boldsymbol{T}^k$
    - For this course all Markov chains will be regular
- The transition matrix for a regular Markov chain will have **exactly one** eigenvalue equal to 1 and the **rest will all be less than 1**
- As n gets bigger $\boldsymbol{T}^n$ tends to a matrix where **each column is identical**
  - The column matrix formed by using **one of these columns** is called the steady state column matrix **s**
  - This means that the **long-term probabilities** tend to fixed probabilities
    - $\mathbf{s}_n$ tends to **s**

How do I use long-term probabilities to find the steady state?

- As $\boldsymbol{T}^n$ tends to a matrix whose columns equal the steady state vector
  - Calculate $\boldsymbol{T}^n$ for a large value of n using your GDC
  - If the columns are identical when rounded to a required degree of accuracy then the column is the steady state vector
  - If the columns are not identical then choose a higher power and repeat

How do I find the exact steady state probabilities?

- As $\boldsymbol{T}\mathbf{s} = \mathbf{s}$ the steady state vector **s** is the eigenvector of $\boldsymbol{T}$ corresponding to the eigenvalue equal to 1 whose elements sum to 1:
  - Let **s** have entries $x_1, x_2, \ldots, x_n$
  - Use $\boldsymbol{T}\mathbf{s} = \mathbf{s}$ to form a system of linear equations
  - There will be an infinite number of solutions so choose a value for one of the unknowns
    - For example: let $x_n = 1$
  - Ignoring the last equation solve the system of linear equations to find $x_1, x_2, \ldots, x_{n-1}$
  - Divide each value x_i by the sum of the values
    - This makes the values add up to 1
- You might be asked to **show this result using diagonalisation**
  - Write $\boldsymbol{T} = \boldsymbol{PDP}^{-1}$ where $\boldsymbol{D}$ is the diagonal matrix of eigenvalues and $\boldsymbol{P}$ is the matrix of eigenvectors
  - Use $\boldsymbol{T}^n = \boldsymbol{PD}^n\boldsymbol{P}^{-1}$
  - As n gets large $\boldsymbol{D}^n$ tends to a matrix where all entries are 0 apart from one entry of 1 due to the eigenvalue of 1
  - Calculate the limit of $\boldsymbol{T}^n$ which will have **identical columns**
    - You can calculate this by multiplying the three matrices ($\boldsymbol{P}, \boldsymbol{D}^{\infty}, \boldsymbol{P}^{-1}$) together

Exam Tip

- If you calculate T^{∞} by hand then a quick check is to see if the columns are identical
  - It should look like $\begin{pmatrix} a & a & a \\ b & b & b \\ c & c & c \end{pmatrix}$

Worked Example

If a cat is brushed on a given day, then the probability it is brushed the following day is 0.2. If a cat is not brushed on a given day, then the probability that is will be brushed the following day is 0.9.

The transition matrix $\boldsymbol{T}$ is used to model this information with $\boldsymbol{T}=\begin{pmatrix}0.2 & 0.9\\ 0.8 & 0.1\end{pmatrix}$.

a)
Find an eigenvector of $\boldsymbol{T}$ corresponding to the eigenvalue 1.

$\underline{v}$ is an eigenvector of T with eigenvalue 1 if $T\underline{v} = \underline{v}$

Let $\underline{v} = \begin{pmatrix}x_1\\ x_2\end{pmatrix}$

$T\underline{v} = \begin{pmatrix}0.2 & 0.9\\ 0.8 & 0.1\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix} = \begin{pmatrix}0.2x_1 + 0.9x_2\\ 0.8x_1 + 0.1x_2\end{pmatrix}$

$T\underline{v} = \underline{v}$ $\quad 0.2x_1 + 0.9x_2 = x_1 \Rightarrow 0.9x_2 = 0.8x_1 \Rightarrow 9x_2 = 8x_1$

$0.8x_1 + 0.1x_2 = x_2 \Rightarrow 0.8x_1 = 0.9x_2 \Rightarrow 8x_1 = 9x_2$

Find a solution $\quad x_1 = 9$ and $x_2 = 8$

$\boxed{\begin{pmatrix}9\\ 8\end{pmatrix} \text{ or any scalar multiple}}$

b)
Hence find the steady state vector.

Scale the elements so that they add to 1 $\quad \begin{pmatrix}\frac{9}{17}\\ \frac{8}{17}\end{pmatrix}$

The eigenvector corresponding to the eigenvalue 1, whose elements add to 1, is the steady state vector.

$\boxed{\begin{pmatrix}\frac{9}{17}\\ \frac{8}{17}\end{pmatrix}}$