IB Maths DP

YOUR NOTES

4. Statistics & Probability

CONTENTS

4.1 Statistics Toolkit
- 4.1.1 Sampling & Data Collection
- 4.1.2 Statistical Measures
- 4.1.3 Frequency Tables
- 4.1.4 Linear Tranformations of Data
- 4.1.5 Outliers
- 4.1.6 Univariate Data
- 4.1.7 Interpreting Data

4.2 Correlation & Regression
- 4.2.1 Bivariate Data
- 4.2.2 Correlation & Regression

4.3 Probability
- 4.3.1 Probability & Types of Events
- 4.3.2 Conditional Probability
- 4.3.3 Sample Space Diagrams

4.4 Probability Distributions
- 4.4.1 Discrete Probability Distributions
- 4.4.2 Expected Values

4.5 Binomial Distribution
- 4.5.1 The Binomial Distribution
- 4.5.2 Calculating Binomial Probabilities

4.6 Normal Distribution
- 4.6.1 The Normal Distribution
- 4.6.2 Calculations with Normal Distribution
- 4.6.3 Standardisation of Normal Variables

YOUR NOTES
⬇

4.1 Statistics Toolkit

4.1.1 Sampling & Data Collection

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

YOUR NOTES
⬇

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 where the sample is **not random**

What sampling techniques do I need to know?

Simple random sampling

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

Systematic sampling

- **Systematic sampling** is where a sample is formed by choosing members of a population at regular intervals using a list
- To carry this out you would...
  - calculate the size of the interval $k = \frac{\text{size of population } (N)}{\text{size of sample } (n)}$
  - choose a random starting point between 1 and *k*
  - select every *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 (called strata) and then a random sample is taken from each group (stratum)
- The proportion of a stratum that is sampled is equal to the proportion of the population that belong to that stratum
- To carry this out you would...
  - Calculate the number of members sampled from each stratum

YOUR NOTES
⬇

- - $\frac{\text{size of sample } (n)}{\text{size of population } (N)} \times \text{number of members in the stratum}$
  - Take a random sample from each stratum
- **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

YOUR NOTES
⬇

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

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

Quota sampling

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

Mark could improve his sampling method by increasing his sample size

YOUR NOTES
⬇

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 Statistical Measures

YOUR NOTES
⬇

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

YOUR NOTES
↓

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

Mode = 43

Median is the middle when in order

29 43 43 51 64 70

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

Median = 47

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

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

Mean = 50

YOUR NOTES
⬇

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

YOUR NOTES
↓

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

Range = 41

Find upper and lower quartiles using GDC

$Q_1 = 43$ and $Q_3 = 64$

$IQR = Q_3 - Q_1$

64 − 43

IQR = 21

By hand

Range

29 43 43 51 64 70

IQR

YOUR NOTES
⬇

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 $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
  - This can be rewritten as

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

- The formula for **standard deviation** is $\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
  - This can be rewritten as

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2}$$

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

YOUR NOTES

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

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

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

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

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

YOUR NOTES
⬇

4.1.3 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**

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

  - **Standard deviation**

YOUR NOTES
⬇

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

- 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

YOUR NOTES
↓

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

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

Median = 1

c)
the mean.

Formula Booklet

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

$$\bar{x} = \frac{\Sigma fx}{n} = \frac{11\times 0 + 5\times 1 + 8\times 2 + 6\times 3}{11 + 5 + 8 + 6} \quad \frac{39}{30}$$

Mean = 1.3

d)
the standard deviation.

Use GDC $\sigma_x = 1.159...$

Standard deviation = 1.16 (3sf)

YOUR NOTES

YOUR NOTES
↓

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**

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

YOUR NOTES
⬇

- **Standard deviation**

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

- 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 see 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
  - $\bar{x} = 0.333$ (3sf) rather than $\bar{x} = \frac{1}{3}$

YOUR NOTES

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

Modal class : $160 \le h < 165$

b)

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

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

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

Mid-interval value = 162.5 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	$\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$ $\quad n = \sum_{i=1}^{k} f_i$

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

Estimated mean = 162.1 cm

4.1.4 Linear Tranformations of Data

YOUR NOTES

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 $\bar{x}$ 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 $k\bar{x}$
- 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 $\bar{x} \pm a$

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

YOUR NOTES
↓

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

$31 \times 2 + 10$

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

Standard deviation of standardised scores = 10

YOUR NOTES
⬇

4.1.5 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 $\boldsymbol{x < Q_1 - 1.5 \times \text{IQR}}$ or $\boldsymbol{x > Q_3 + 1.5 \times \text{IQR}}$
- 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

YOUR NOTES
↓

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.

4.1.6 Univariate Data

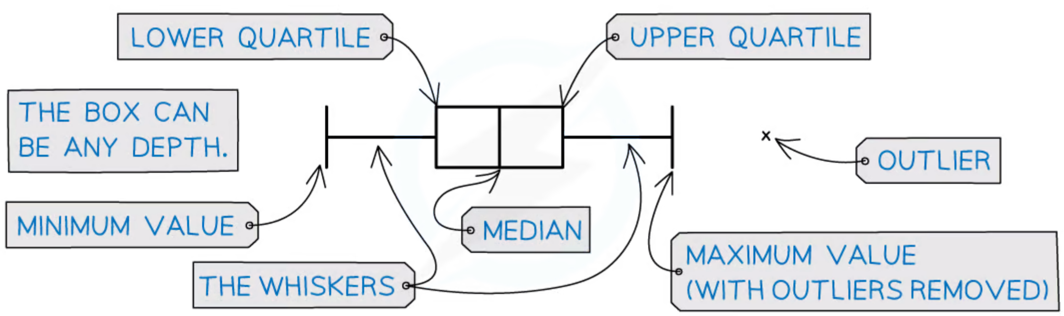
YOUR NOTES ↓

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

YOUR NOTES

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

$Q_1 = 1.1$ m $\quad Q_2 = 1.3$ m $\quad Q_3 = 1.4$ m

$IQR = Q_3 - Q_1 = 1.4 - 1.1$

$IQR = 0.3$ m

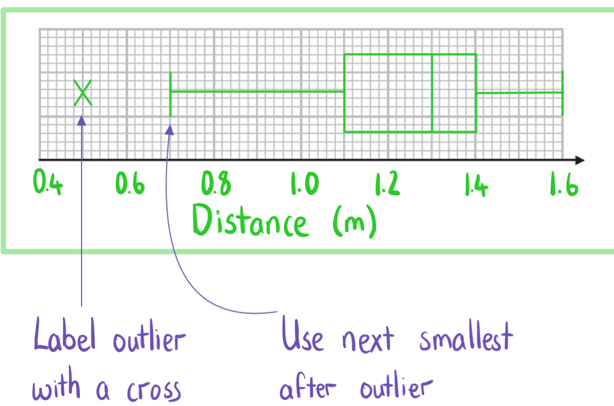
$Q_1 - 1.5 \times IQR = 1.1 - 1.5 \times 0.3 = 0.65$

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

0.5 m is an outlier

b)

Draw a box plot for the data.



YOUR NOTES
⬇

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**

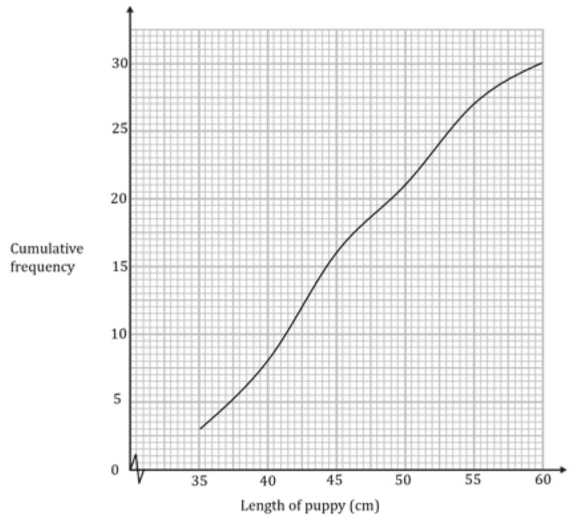
What are cumulative frequency graphs useful for?

- 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

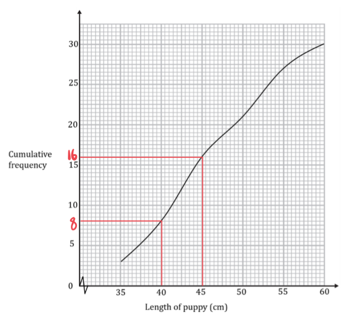
YOUR NOTES

Worked Example

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



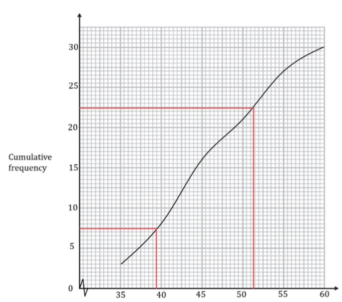
a)
Given that the interval $40 \leq l < 45$ was used when collecting data, find the frequency of this class.



$16 - 8$

Frequency = 8

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



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

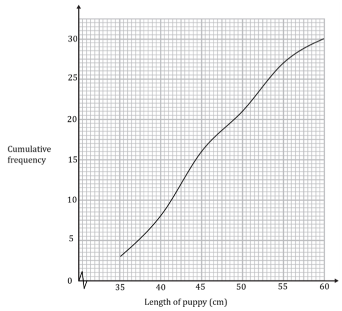
$\frac{3}{4} \times 30 = 22.5 \quad Q_3 = 51.4$

$IQR = Q_3 - Q_1 = 51.4 - 39.5$

$IQR = 11.9$ cm

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

YOUR NOTES
⬇



$30 - 22 = 8$ puppies longer than 51 cm

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

26.7 % (3sf)

YOUR NOTES
⬇

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

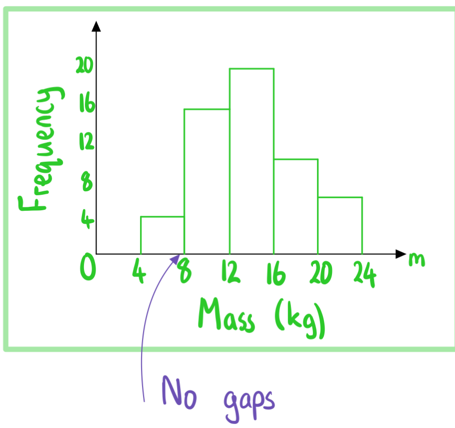
YOUR NOTES

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

Modal class : $12 \le m < 16$

YOUR NOTES
↓

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 r**ange 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?

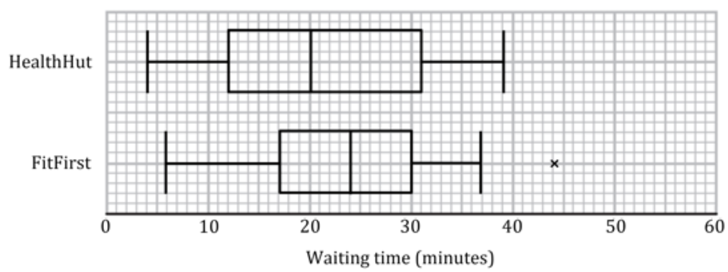
- Compare a **measure of central tendency**
  - If the data **contains outliers – use the median**

YOUR NOTES

- 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.

YOUR NOTES
⬇

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

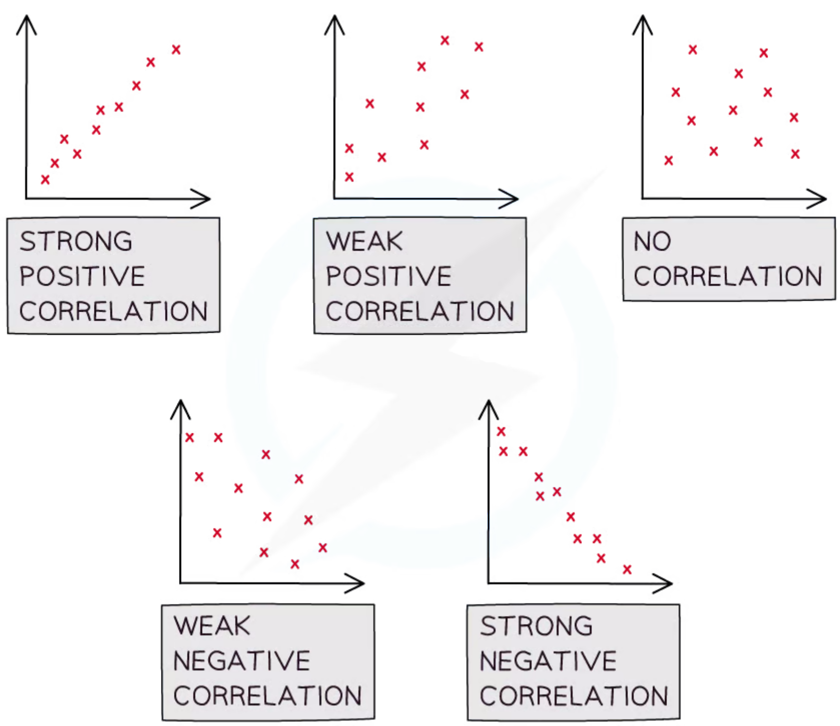
- If you use scatter diagrams in your Internal Assessment then be aware that finding outliers for bivariate data is different to finding outliers for univariate data
  - (x, y) could be an outlier for the bivariate data even if x and y are not outliers for their separate univariate data

YOUR NOTES
⬇

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

YOUR NOTES
⬇

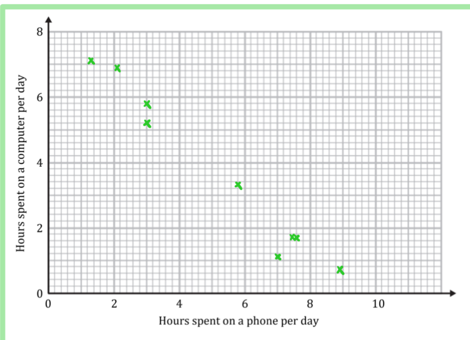
YOUR NOTES

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.



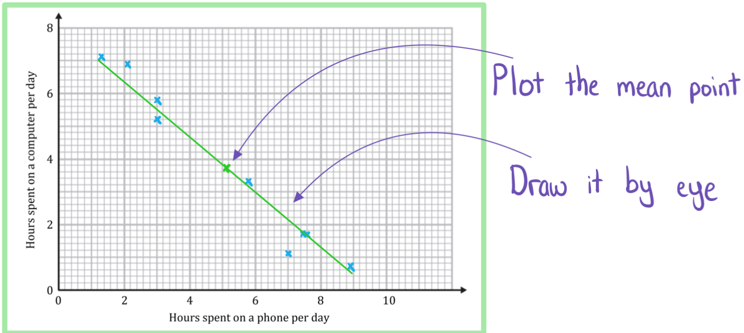
b)
Describe the correlation.

Strong negative linear correlation

c)
Draw a line of best fit.

YOUR NOTES
⬇

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



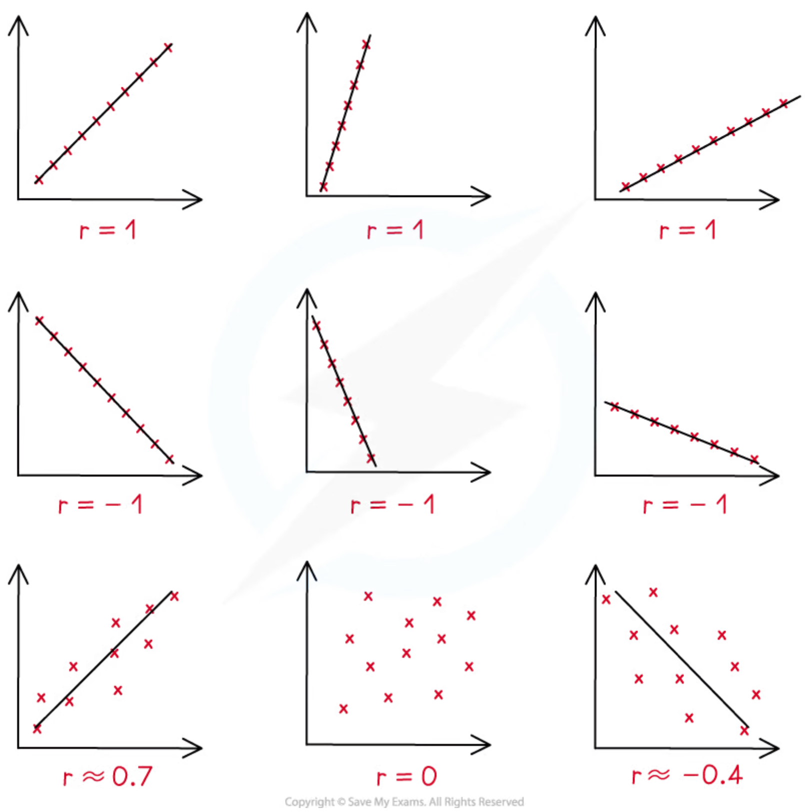
4.2.2 Correlation & Regression

YOUR NOTES

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 $-1 \leq r \leq 1$
  - 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

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

YOUR NOTES
⬇

- $S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$ 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

YOUR NOTES
⬇

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
- It can be calculated by either looking at:
  - **vertical distances** between the line and the data values
    - This is the **regression line of *y* on *x***
  - **horizontal distances** between the line and the data values
    - This is the **regression line of *x* on *y***

How do I find the regression line of *y* on *x*?

- 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** $(\bar{x}, \bar{y})$ will lie on the regression line

How do I find the regression line of *x* on *y*?

- The **regression line of x on y** is written in the form $x = cy + d$
- c is the **gradient** of the line
  - It represents the change in *x* for each individual unit change in *y*
    - If c is **positive** this means *x* **increases** by c for a unit increase in *y*
    - If c is **negative** this means *x* **decreases** by |c| for a unit increase in *y*
- *d* is the ***x* – intercept**
  - It shows the value of *x* when *y* is zero
- You are expected to use your **GDC** to find the equation of the regression line
  - It is found the same way as the regression line of *y* on *x* but with the two data sets **switched around**
  - Remember the **mean point** $(\bar{x}, \bar{y})$ will lie on the regression line

How do I use a regression line?

- The regression line can be used to decide what type of correlation there is if there is no scatter diagram
  - If the gradient is **positive** then the data set has **positive correlation**
  - If the gradient is **negative** then the data set has **negative correlation**
- The regression line can also be used to **predict** the value of a **dependent variable** from an **independent variable**

YOUR NOTES
⬇

  - The equation for the y on x line should only be used to make predictions for y
    - Using a y on x line to predict x is not always reliable
  - The equation for the x on y line should only be used to make predictions for x
    - Using an x on y line to predict y 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
- The y on x and x on y regression lines intersect at the mean point $(\overline{x}, \overline{y})$

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

YOUR NOTES

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.

$r = 0.79433...$

$r = 0.794$ (3sf)

b)
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.

a is the coefficient of x $a = 0.943579...$

b is the constant term $b = -18.05398...$

$y = 0.944x - 18.1$

c)
Write down the equation of the regression line of x on y, giving your answer in the form $x = cy + d$ where c and d are constants to be found.

Swap the two sets of data

c is the coefficient of y $c = 0.668700...$

d is the constant term $d = 30.52410...$

$x = 0.669y + 30.5$

d)
Use the appropriate regression line to predict the score on the maths test of a student who got a score of 63 on the English test.

YOUR NOTES

$y = 63$ so use x on y line

$x = (0.668700...) \times 63 + (30.52410...) = 72.652...$

Maths score 72.7

YOUR NOTES
⬇

4.3 Probability

4.3.1 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**

YOUR NOTES
⬇

- 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'

$$P(A) + P(A') = 1$$

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

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

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

- 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

$$P(A \cap B) = P(A)P(B|A)$$

- 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

YOUR NOTES
↓

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 $\mathrm{P}(T)$.

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

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

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

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

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

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

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

YOUR NOTES
⬇

Independent & Mutually Exclusive Events

What are mutually exclusive events?

- Two events are **mutually exclusive** if they **can not both happen at once**
    - 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:
    - $\mathrm{P}(A \cap B) = 0$
- **Complementary events** are mutually exclusive

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*

YOUR NOTES
↓

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"

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

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

$P(D \cap C) = 0.5$

b)
Two events, Q and R, are such that $P(Q) = 0.8$ and $P(Q \cap R) = 0.1$.
Given that Q and R are independent, find $P(R)$.

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

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

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

c)
Two events, S and T, are such that $P(S) = 2P(T)$.
Given that S and T are mutually exclusive and that $P(S \cup T) = 0.6$ find $P(S)$ and $P(T)$.

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

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

$0.6 = 2P(T) + P(T)$ ← $P(S) = 2P(T)$

$0.6 = 3P(T)$

$P(T) = 0.2$ and $P(S) = 0.4$

4.3.2 Conditional Probability

YOUR NOTES
↓

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 should use
  - $\mathrm{P}(A|B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$
  - This is given in the **formula booklet**
  - This can be rearranged to give $\mathrm{P}(A \cap B) = \mathrm{P}(B)\mathrm{P}(A|B)$
  - By symmetry you can also write $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B|A)$

How do I tell if two events are independent using conditional probabilities?

- If A and B are two events then they are independent if:
  - $\mathrm{P}(A|B) = \mathrm{P}(A) = \mathrm{P}(A|B')$
- Equally you can still use $\mathrm{P}(A \cap B) = \mathrm{P}(A)\mathrm{P}(B)$ to test for independence
  - This is given in the **formula booklet**

YOUR NOTES

Worked Example

Let R be the event that it is raining in Weatherville and T be the event that there is a thunderstorm in Weatherville.

It is known that $\mathrm{P}(T) = 0.035$, $\mathrm{P}(T \cap R) = 0.03$ and $\mathrm{P}(T|R) = 0.15$.

a)
Find the probability that it is raining in Weatherville.

Formula booklet

Conditional probability	$\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$

$$P(T|R) = \frac{P(T \cap R)}{P(R)}$$

Substitute the values in

$$0.15 = \frac{0.03}{P(R)}$$

$$P(R) = \frac{0.03}{0.15}$$

$$P(R) = 0.2$$

b)
State whether the events R and T are independent. Give a reason for your answer.

If R and T are independent then $P(T|R) = P(T)$

$P(T|R) = 0.15$ and $P(T) = 0.035$

R and T are not independent as $P(T|R) \neq P(T)$

4.3.3 Sample Space Diagrams

YOUR NOTES

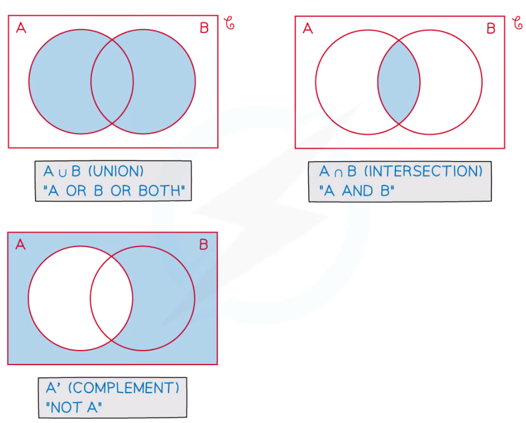
Venn Diagrams

What is a Venn diagram?

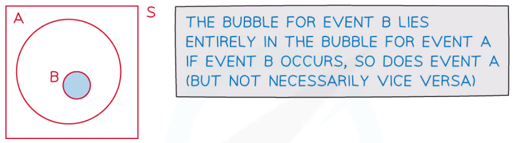
- A Venn diagram is a way to illustrate **events** from an **experiment** and are particularly useful when there is an overlap between possible **outcomes**
- A Venn diagram consists of
  - a **rectangle** representing the **sample space (*U*)**
    - The rectangle is labelled *U*
    - Some mathematicians 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**

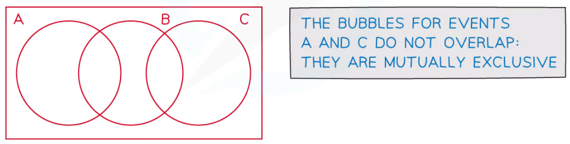
What do the different regions mean on a Venn diagram?

- 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



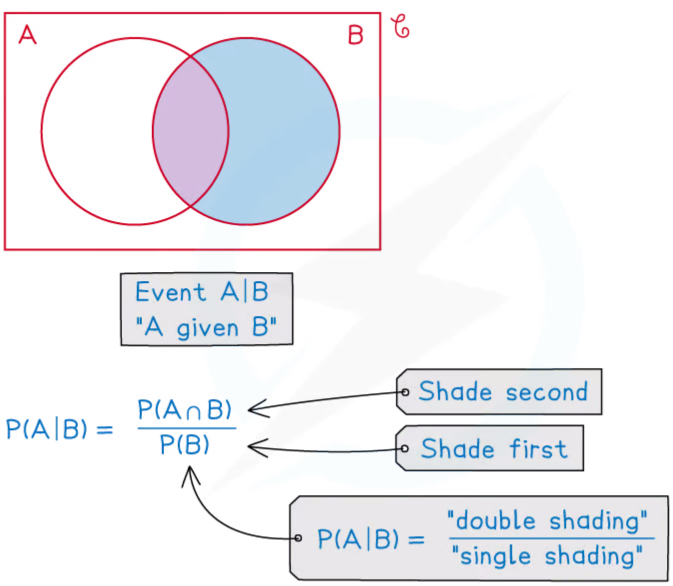
YOUR NOTES
⬇





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
  - $P(A \cap B) = P(A)P(B)$
- To find the conditional probability $P(A | B)$
  - 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



YOUR NOTES
⬇

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

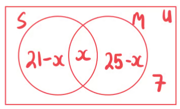
YOUR NOTES
↓

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.

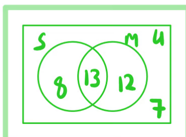
Find the centre first



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

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

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

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

Tree Diagrams

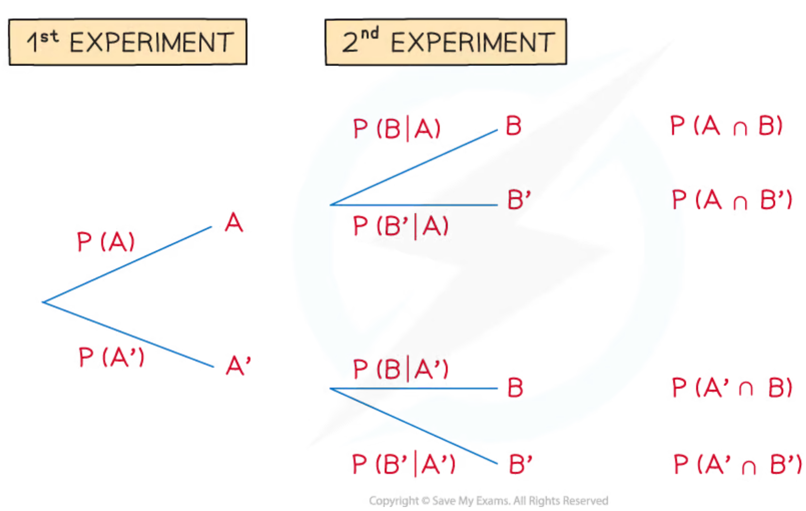
YOUR NOTES
↓

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: $P(A \cup B) = P(A \cap B) + P(A \cap B') + P(A' \cap B)$
  - **subtract** the probabilities of the combined outcomes that are not part of that event from 1
    - For example: $P(A \cup B) = 1 - P(A' \cap B')$



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

YOUR NOTES
↓

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

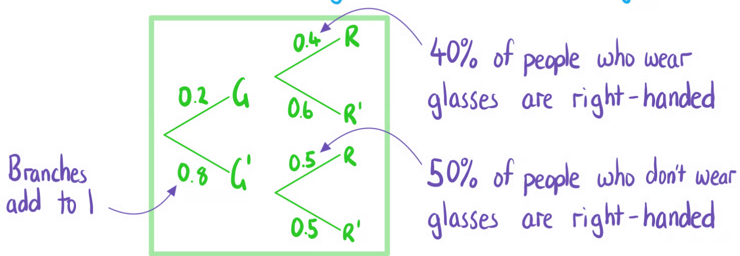
YOUR NOTES
↓

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.

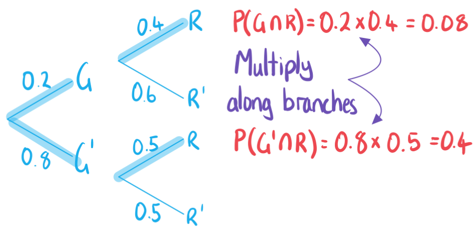
a)
Draw a tree diagram to represent the information.

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



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

Find options that contain R



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

$P(R) = 0.48$

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

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

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

YOUR NOTES
⬇

4.4 Probability Distributions

4.4.1 Discrete Probability Distributions

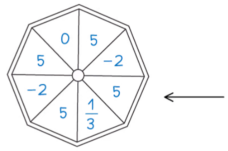
Discrete Probability Distributions

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)
- $\text{P}(X = x)$ 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?

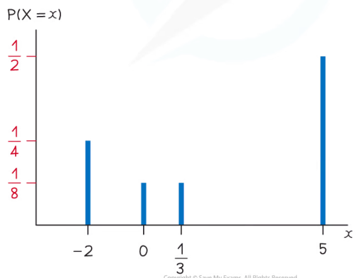
- A **discrete probability distribution** fully describes **all the values** that a discrete random variable can take along with their **associated probabilities**
  - This can be given in a **table**
  - Or it can be given as a **function** (called a discrete probability distribution function or "pdf")
  - 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 $\sum \text{P}(X = x) = 1$
- 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
- $P(X < k)$ would be used for phrases such as:

YOUR NOTES

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

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

a)

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

Construct a table

x	-3	-1	2	4
$P(X=x)$	$9k$	k	$4k$	$16k$

Substitute in the values of x
e.g. $P(X=-3) = k(-3)^2 = 9k$

The probabilities add up to 1

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

$30k = 1$

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

b)

Calculate $P(X \leq 3)$.

Substitute k into the probabilities

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

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

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

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

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

YOUR NOTES
⬇

4.4.2 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

YOUR NOTES
⬇

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

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

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

Expected value = $4.35

b)
Determine whether the game is fair.

A game is fair is expected gain/loss is 0

Prize − cost

$4.35 - 5 = -0.65$

Expected loss is $0.65 so game is not fair

YOUR NOTES
⬇

4.5 Binomial Distribution

4.5.1 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:
  - $\mathrm{P}(X=r) = {}^nC_r \times p^r(1-p)^{n-r}$ for $r = 0, 1, 2, \ldots, n$
    - ${}^nC_r = \dfrac{n!}{r!(n-r)!}$ 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

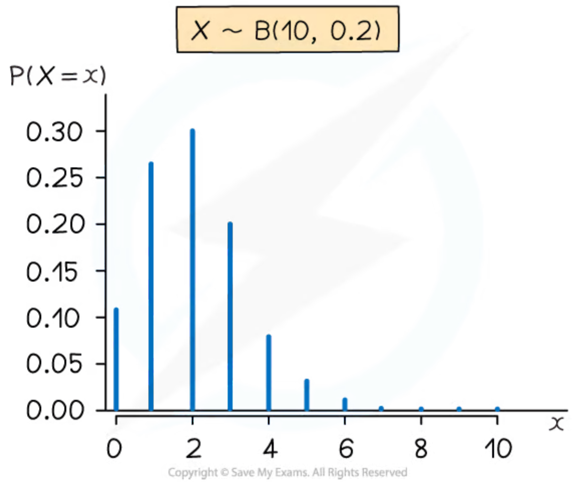
$$\mathrm{E}(X) = np$$

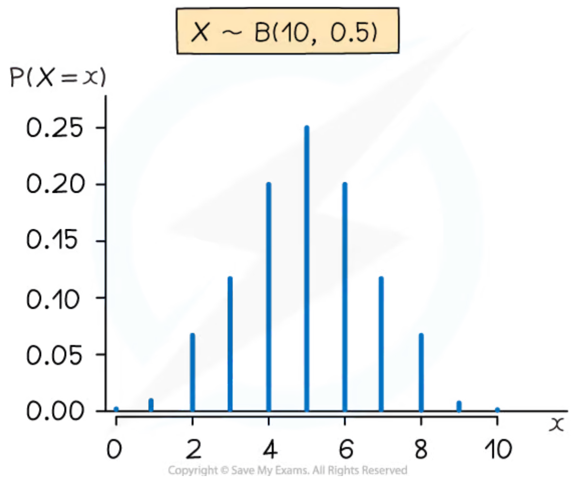
  - You are given this in the **formula booklet**
- The **variance** of the number of successful trials is

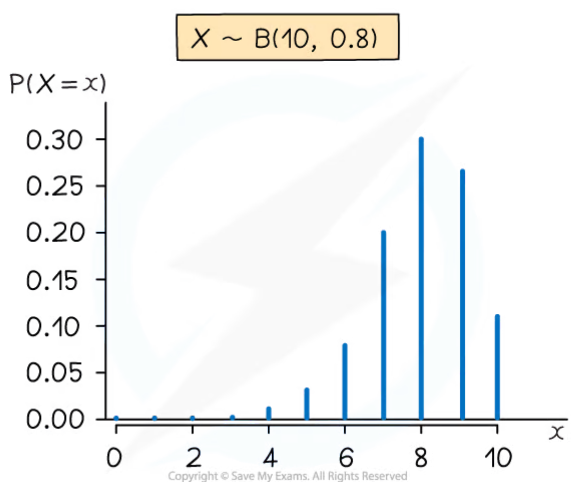
$$\mathrm{Var}(X) = np(1-p)$$

  - 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**

YOUR NOTES
↓


X ~ B(10, 0.2)
P(X = x)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
x
Copyright © Save My Exams. All Rights Reserved


X ~ B(10, 0.5)
P(X = x)
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
x
Copyright © Save My Exams. All Rights Reserved


X ~ B(10, 0.8)
P(X = x)
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
2
4
6
8
10
x
Copyright © Save My Exams. All Rights Reserved

YOUR NOTES
⬇

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 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**

YOUR NOTES

Exam Tip

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

YOUR NOTES
↓

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

$X \sim B(50, 0.08)$

Number of people in sample (50); Probability of being immune to the virus (0.08)

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 $X \sim B(n, p)$	
Mean	$E(X) = np$

$E(X) = 50 \times 0.08$

4 people

4.5.2 Calculating Binomial Probabilities

YOUR NOTES
↓

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 $\mathrm{P}(X \le x)$ just enter *x* into the cumulative distribution function

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

YOUR NOTES
⬇

- 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 binomial 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 binomial 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 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
  - $P(X < x) = P(X \le x - 1)$ 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 or maximum then use 0 or n
  - $P(X \le b) = P(0 \le X \le b)$
  - $P(X \ge a) = P(a \le X \le n)$

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

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

YOUR NOTES
↓

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

$P(X = 10) = 0.05705 6...$

$P(X = 10) = 0.057$ (3sf)

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

Identify upper and lower values

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

Use binomial cumulative distribution on GDC

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

$P(X \leq 10) = 0.121$ (3sf)

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

Identify upper and lower values

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

Use binomial cumulative distribution on GDC

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

$P(8 < X < 15) = 0.542$ (3sf)

4.6 Normal Distribution

4.6.1 The Normal Distribution

YOUR NOTES
↓

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

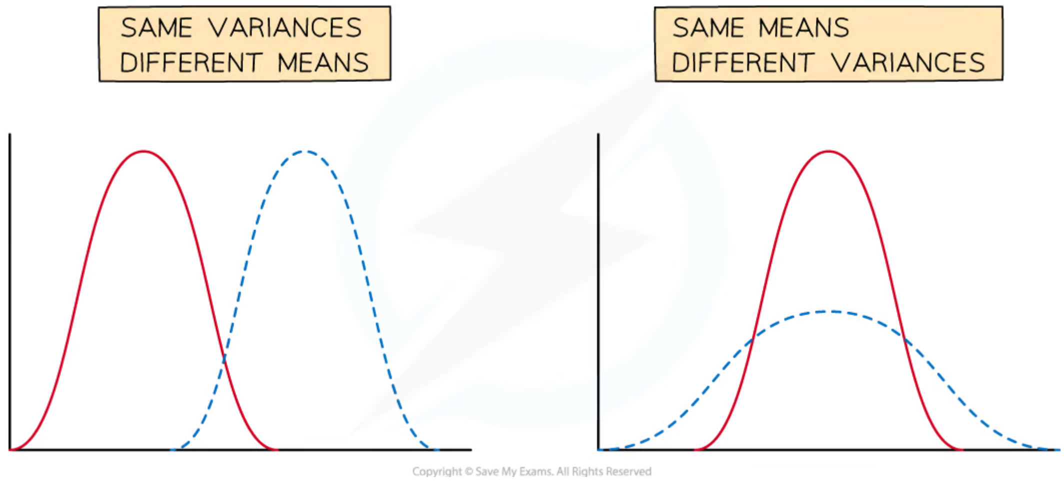
What is a continuous probability distribution?

- 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?

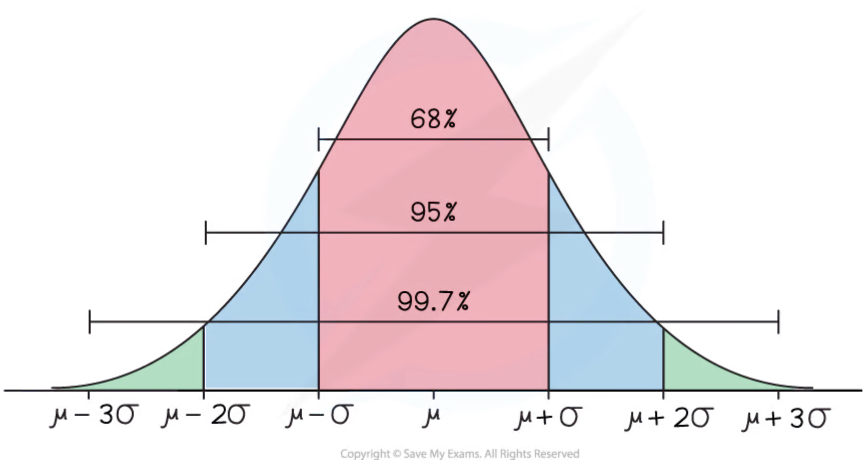
- A normal distribution is a **continuous probability distribution**
- The **continuous random variable** X can follow a normal distribution if:
  - The distribution is **symmetrical**
  - The distribution is **bell-shaped**
- If X follows a normal distribution then it is denoted $X \sim \mathrm{N}(\mu, \sigma^2)$
  - 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** changes then the graph is **stretched horizontally**
  - A **small variance** leads to a **tall** curve with a **narrow** centre
  - A **large variance** leads to a **short** curve with a **wide** centre

YOUR NOTES
⬇



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 ($\mu \pm 3\sigma$)



YOUR NOTES
⬇

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

4.6.2 Calculations with Normal Distribution

YOUR NOTES
⬇

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}**)

YOUR NOTES
⬇

- $P(X < b)$ 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?

- $P(X < \mu) = P(X > \mu) = 0.5$
- As $P(X = a) = 0$ you can use:
  - $P(X < a) + P(X > a) = 1$
  - $P(X > a) = 1 - P(X < a)$
  - $P(a < X < b) = P(X < b) - P(X < a)$
- 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

YOUR NOTES ↓

Worked Example

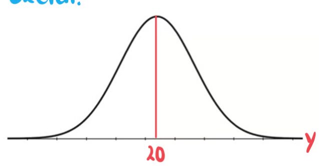
The random variable $Y \sim \mathrm{N}(20, 5^2)$. Calculate:

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

Identify μ and σ

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

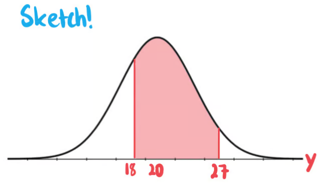
Sketch!



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

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

Sketch!



Using GDC
Lower = 18
Upper = 27

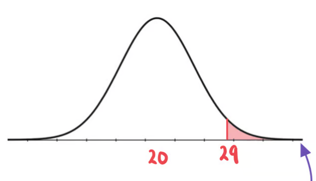
We can use $\le$ or $<$

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

0.575 (3sf)

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

Sketch!



No upper bound so choose a big number

Using GDC
Lower = 29
Upper = 99999

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

0.0359 (3sf)

Inverse Normal Distribution

YOUR NOTES
↓

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*?

- 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*
- 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*'

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

YOUR NOTES
↓

Worked Example

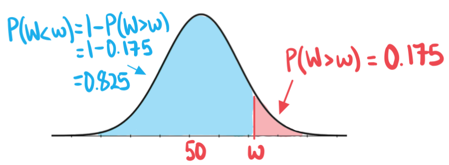
The random variable $W \sim N(50, 36)$.

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

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

$w = 55.6075...$

$w = 55.6$ (3sf)

YOUR NOTES
⬇

4.6.3 Standardisation of Normal Variables

Standard Normal Distribution

What is the standard normal distribution?

- The **standard normal distribution** is a normal distribution where the **mean is 0** and the **standard deviation is 1**
  - It is denoted by Z
  - $Z \sim \mathrm{N}(0,\ 1^2)$

Why is the standard normal distribution important?

- Any **normal distribution curve** can be transformed to the standard normal distribution curve by a **horizontal translation** and a **horizontal stretch**
- Therefore we have the relationship:
  - $Z = \dfrac{X - \mu}{\sigma}$
  - Where $X \sim \mathrm{N}(\mu,\ \sigma^2)$ and $Z \sim \mathrm{N}(0,\ 1^2)$
- Probabilities are related by:
  - $\mathrm{P}(a < X < b) = \mathrm{P}\left(\dfrac{a-\mu}{\sigma} < Z < \dfrac{b-\mu}{\sigma}\right)$
  - This will be useful when the mean or variance is unknown
- Some mathematicians use the function $\Phi(z)$ to represent $\mathrm{P}(Z < z)$

z-values

What are *z*-values (standardised values)?

- For a normal distribution $X \sim \mathrm{N}(\mu,\ \sigma^2)$ the *z*-value (standardised value) of an *x*-value tells you how many standard deviations it is away from the mean
  - If $z = 1$ then that means the *x*-value is 1 standard deviation bigger than the mean
  - If $z = -1$ then that means the *x*-value is 1 standard deviation smaller than the mean
- If the *x*-value is **more than the mean** then its corresponding *z*-value will be **positive**
- If the *x*-value is **less than the mean** then its corresponding *z*-value will be **negative**
- The *z*-value can be calculated using the formula:
  - $z = \dfrac{x - \mu}{\sigma}$
  - This is given in the **formula booklet**
- *z*-values can be used to compare values from different distributions

Finding Sigma and Mu

YOUR NOTES
⬇

How do I find the mean (μ) or the standard deviation (σ) if one of them is unknown?

- If the **mean** or **standard deviation** of $X \sim \mathrm{N}(\mu, \sigma^2)$ is **unknown** then you will need to use the **standard normal distribution**
- You will need to use the formula
  - $z = \dfrac{x-\mu}{\sigma}$ or its rearranged form $x = \mu + \sigma z$
- You will be given a **probability for a specific value** of
  - $\mathrm{P}(X < x) = p$ or $\mathrm{P}(X > x) = p$
- To find the unknown parameter:
- **STEP 1**: **Sketch** the normal curve
  - Label the known value and the mean
- **STEP 2**: **Find** the ***z*-value** for the given value of ***x***
  - Use the **Inverse Normal Distribution** to find the value of z such that $\mathrm{P}(Z < z) = p$ or $\mathrm{P}(Z > z) = p$
  - Make sure the direction of the inequality for Z is consistent with the inequality for X
  - Try to **use lots of decimal places** for the z-value or **store your answer** to **avoid rounding errors**
    - You should use at least one extra decimal place within your working than your intended degree of accuracy for your answer
- **STEP 3**: **Substitute** the known values into $z = \dfrac{x-\mu}{\sigma}$ or $x = \mu + \sigma z$
  - You will be given and one of the parameters (μ or σ) in the question
  - You will have calculated z in STEP 2
- **STEP 4**: **Solve** the equation

How do I find the mean (μ) and the standard deviation (σ) if both of them are unknown?

- If **both** of them are **unknown** then you will be given two probabilities for two specific values of ***x***
- The process is the same as above
  - You will now be able to **calculate two *z* -values**
  - You can form **two equations** (rearranging to the form $x = \mu + \sigma z$ is helpful)
  - You now have to **solve the two equations simultaneously** (you can use your calculator to do this)
  - Be careful not to mix up which z-value goes with which value of x

YOUR NOTES

Worked Example

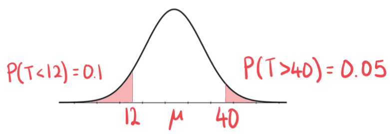
It is known that the times, in minutes, taken by students at a school to eat their lunch can be modelled using a normal distribution with mean μ minutes and standard deviation σ minutes.

Given that 10% of students at the school take less than 12 minutes to eat their lunch and 5% of the students take more than 40 minutes to eat their lunch, find the mean and standard deviation of the time taken by the students at the school.

Let $T \sim N(\mu, \sigma^2)$ be the time taken to eat lunch

STEP 1

Sketch the information



STEP 2

Find the corresponding z-values using inverse normal on GDC

$Z \sim N(0, 1^2)$

$P(Z < z_1) = 0.1 \Rightarrow z_1 = -1.2815...$

$P(Z > z_2) = 0.05 \Rightarrow P(Z < z_2) = 0.95 \Rightarrow z_2 = 1.6448...$

STEP 3

Form equations using $z = \frac{x - \mu}{\sigma}$ or $x = \mu + \sigma z$

$12 = \mu - (1.2815...)\sigma$

$40 = \mu + (1.6448...)\sigma$

STEP 4

Solve equations using GDC

$\mu = 24.26... \quad \sigma = 9.568...$

Mean = 24.3 mins (3sf)
Standard deviation = 9.57 mins (3sf)