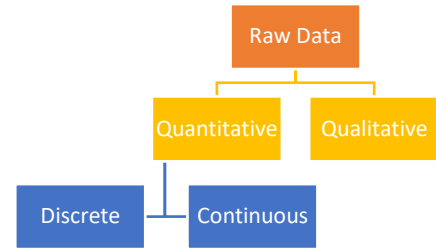


Edexcel GCSE Statistics (9-1) Revision Notes

Chapter 1: Collection of Data

Types of Data

- **Raw Data** – **Unprocessed**. Just been collected. Needs to be ordered, grouped, rounded, cleaned.
- **Qualitative** – **Non-numerical**, descriptive data such as eye/hair colour or gender. Often subjective so usually more difficult to analyse.
- **Quantitative** – **Numerical** data. Can be measured with numbers. Easier to analyse than qualitative data. Example, height, weights, marks in an exam etc.
- **Discrete** – Only takes **particular values** (not necessarily whole numbers) such as shoe size or number of people.
- **Continuous** - Can take **any value** e.g. height, weight.
- **Categorical** – data that can be **sorted** into non-overlapping categories such as gender. Used for qualitative data so that it can be more easily processed.
- **Ordinal (rank)** – quantitative data that can be given an **order** or ranked on a rating scale, e.g. marks in an exam.
- **Bivariate** – Involves measuring **2 variables**. Can be qualitative or quantitative, grouped or ungrouped. Usually used with scatter diagrams where the two axes represent the two different variables. One variable is often called the explanatory variable and the other the response variable.
- **Multivariate** – Made up of **more than 2 variables** e.g. comparing height, weight, age and shoe size together.



Grouping Data

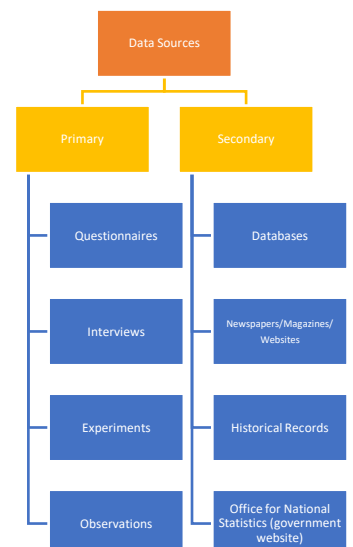
Grouping data using tables makes it easier to spot patterns in the data and quickly see how the data is distributed.

- Discrete data can be grouped into classes that do not overlap e.g. 0-10, 11-15... (they do not have to have equal class width). Uses smaller intervals when there is a lot of data close together in that range and wider classes for data that is more spread out.
- Continuous data can be grouped using inequalities. The class intervals must not have gaps between them or be overlapping so inequality symbols must be used with one of the symbols being $<$ and the other \leq .
- **Pros:**
 - Makes the data easy to read and understand.
 - Easy to spot patterns and compare data.
- **Cons:**
 - Loses accuracy of data as you no longer know exact data values.
 - Calculations made from these will only be an estimate e.g. mean.

Data Sources

- **Primary** – Data that you have **collected yourself**, or someone has collected on your behalf.
- **Secondary** – Data that has **already been collected**.

	Advantages	Disadvantages
Primary Data	<ul style="list-style-type: none"> • Accurate • Collection method known • Can find answers to specific questions 	<ul style="list-style-type: none"> • Time consuming • Expensive
Secondary Data	<ul style="list-style-type: none"> • Cheap • Easy • Quick • Data from some organisations can be more reliable than data collected yourself 	<ul style="list-style-type: none"> • Method of collection unknown • Data may be out of date • May contain mistakes • May come from unreliable source • May be difficult to find answers to specific questions

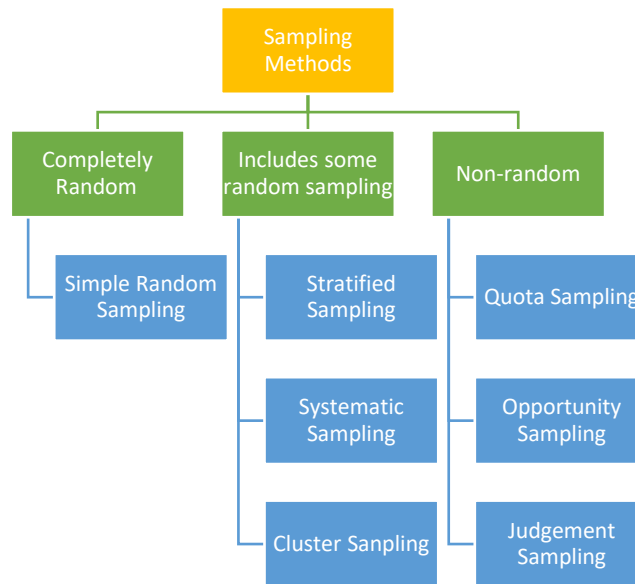


Populations and Sampling:

- **Population** – **Everyone** or everything that could be involved in the investigation e.g. when investigating opinions of students in a school the population would be all the students in the school.
- **Census** – A **survey** of the **entire population**.
- **Sample** – A **smaller number from the population** that you actually survey. The data obtained from the sample is then used to make conclusions about the whole population, so it is important that the sample represents the population fairly.
- **Sampling Frame** - A **list of all the members of the population**. This is where you will choose the sample from. E.g. electoral roll, school register.
- **Sampling Unit** – The **people that are to be sampled** e.g. students in a school.
- **Biased sample** – a sample that **does not represent the population fairly**. Example, if surveying students at a mixed school and the sample only contains girls. Avoid bias by using random sampling methods.

	Advantages	Disadvantages
Census	<ul style="list-style-type: none"> • Unbiased • Accurate • Takes into account entire population 	<ul style="list-style-type: none"> • Time consuming • Expensive • Lots of data to manage • Difficult to ensure whole population is used
Sample	<ul style="list-style-type: none"> • Cheaper • Quicker • Less data to consider 	<ul style="list-style-type: none"> • May be biased • Not completely representative

Sampling Methods:



- **Random Sample** – Every item/person in the population has an **equal chance** of being selected.
 - **Method:**
 - Assign a number to every member in the population.
 - Mention the random sampling technique you are going to use e.g. a random number table or a random number generator on a calculator.
 - Select the numbers chosen from your population.
 - Ignore any repeats and choose another number.
 - **Random Sampling Techniques:**
 - Pick numbers/names out of a hat (only works for small samples)
 - Using a random number table
 - Using the random number generator function on a calculator or computer.
 - **Advantages:**
 - Sample is representative as every member of the population has an equal chance of being selected.
 - Unbiased
 - **Disadvantages:**
 - Need a full list of population (not always easily obtainable)
 - Not always convenient as it can be expensive and time consuming.
 - Needs a large sample size

- **Stratified Sample** – the size of each strata (group) in the sample is in **proportion** to the sizes of strata in the population. E.g. if group A accounts for 10% of the population, in the sample group A will also be 10% of the sample size.
 - **Method:**
 - Split the population into groups (usually done for you in the exam)
 - Use the formula ***stratified sample*** = $\frac{\text{strata}}{\text{total}} \times \text{sample size}$ to calculate sample size for **each group**. (remember to check totals if you rounded numbers and adjust accordingly if your total sample size after stratification is bigger/smaller than sample size in the question).
 - Use random sampling to select members from each strata/group.
 - **Advantages:**
 - Sample is in proportion to population, so sample represents the population fairly.
 - Best used for populations with groups of unequal sizes.
 - **Disadvantages:**
 - Time consuming
- **Systematic Sampling** – choosing items in the population at regular **intervals**.
 - **Method:**
 - Divide your population size by sample size to calculate the intervals, e.g. $400/40 = 10$ so choosing every 10th item in the population.
 - Use random sampling to generate a number between 1 and 10 (or the answer to your calculation from above) to choose a starting point e.g. 7.
 - Select every 10th item after the 7th e.g. 7th, 17th, 27th, ..., until you obtain your sample size.
 - **Advantages:**
 - Population is evenly sampled.
 - Can be carried out by a machine.
 - Sample is easy to select.
 - **Disadvantages:**
 - Not strictly a random sample as some member of the population cannot be chosen.
- **Cluster Sampling** – The population is divided into natural groups (clusters), groups are chosen at random and every member of groups are sampled. Useful for large populations e.g. when surveying lots of different towns in a country.
 - **Advantages:**
 - Economically efficient – less resources required.
 - Can be representative if lots of small clusters are sampled.
 - **Disadvantages:**
 - Clusters may not be representative of the population and may lead to a biased sample.
 - High sampling error.
- **Quota Sampling** – Population is grouped by characteristics and a fixed amount is sampled from every group.
 - **Method:**
 - Group population by characteristics e.g. gender and age
 - Select quota (amount) for each group e.g. 30 men under 25, 40 women over 30 etc.
 - Obtain sample by finding members of each group until quota is reached.
 - **Advantages:**
 - Quick to use.
 - Cheap.
 - Do not need sample frame or full list of the population.
 - **Disadvantages:**
 - NOT RANDOM – biased as interviewer is choosing who will be in the sample so every member of the population does not have an equal chance of being selected.

- **Opportunity Sampling** – Using the people/items that are available at the time. E.g. interviewing the first 10 people you see on a Monday morning.
 - **Advantages:**
 - Quick
 - Cheap
 - Easy
 - **Disadvantages:**
 - NOT RANDOM. The sample has not been collected fairly so it may not represent the population and every member of the population has not been given an equal chance to be selected.
- **Judgement Sampling** – When the researcher uses their own judgement to select a sample, they think will represent the population. E.g. A teacher choosing students to interview about their opinion on a new after school club.
 - **Advantages:**
 - Easy
 - Quick
 - **Disadvantages:**
 - NOT RANDOM.
 - Quality of sample depends on the person selecting the sample. The researcher may be biased and unreliable in the sample they select.

Petersen Capture-Recapture - Used to estimate the size of large or moving populations where it would be impossible to count the entire population. Your answer is only an ESTIMATE.

$$\frac{M}{N} = \frac{m}{n} \quad N = \frac{Mn}{m}$$

$$\frac{\text{First Capture}}{\text{Total (N)}} = \frac{\text{Tagged}}{\text{Second Capture}}$$

N is the population size to be estimated.

M is the number of members of the population that are captured initially and tagged.

n is the number of members of the population that are captured subsequently.

m is the number of members of this subsequent captured population that are tagged.

Method:

1. Take a sample of the population
2. Mark each item
3. Put the items back into the population and ensure **they are thoroughly mixed**
4. Take a second sample and count how many of your sample are marked
5. The proportion of marked items in your new sample should be the same as the proportion of marked items from the population in your first sample.

Assumptions:

- Population has not changed – no births/deaths
- Probability of being caught is equally likely for all individuals.
- Marks/tags not lost
- Sample size is large enough and is representative of the population.

Experiments – used when a researcher in how changes in one variable affect another.

- **Variables:**
 - **Explanatory (Independent) Variable** – The variable that is changed.
 - **Response (dependent) variable** – The variable that is measured.
 - **Extraneous Variables** – Variables you are not interested in but that could affect the result of your experiment.

- **Laboratory Experiments** – Researcher has **full control** over variables. Conducted in a lab or similar environment.
 - **Example** - measuring reaction times of people of different ages.
 - Explanatory variable - age
 - Response variable - reaction time.
 - Extraneous variables - gender, health condition, fitness level etc.
 - **Advantages:**
 - Easy to replicate – makes results more reliable.
 - Extraneous variables can be controlled so results are more likely to be valid as you can be sure that other factors are not affecting your results.
 - **Disadvantages:**
 - People may behave differently under test conditions than they would under real-life conditions – could affect validity of results.

- **Field Experiments** – Carried out in the everyday environment. Researcher has **some control** over the variables. They set up the situation and controls the explanatory variable but has less control over extraneous variables.
 - **Example** – Testing new methods of revision.
 - Explanatory variable – method of revision
 - Response variable – results in exam
 - Extraneous variables – amount of revision pupils does, ability of pupils.
 - **Advantages:**
 - More accurate – reflects real life behaviour.
 - **Disadvantages:**
 - Cannot control extraneous variables.
 - Not as easy to replicate – less reliable than lab experiments.

- **Natural Experiments** - Carried out in the everyday environment. Researcher has **no/very little control** over the variables. Explanatory variables are not changed but instead researchers look at something that already exists in the world and how it affects other things.
 - **Example** – the effect of education on level of income
 - Explanatory variable – level of education
 - Response variable – income
 - Extraneous variables – IQ, other skills people may have, personal circumstances
 - **Advantages :**
 - Reflects real life behaviour
 - **Disadvantages:**
 - Low validity – extraneous variables are not controlled which may affect results instead of explanatory variable.
 - Difficult to replicate.
 - Cannot control extraneous variables.

Simulation – A way to model random events using random numbers and previously collected data. These could be used to help you predict what could actually happen in real life.

Easier and cheaper than actually collecting the data.

Steps:

1. Choose a suitable method for getting random numbers – dice, calculator, random number tables.
2. Assign numbers to the data.
3. Generate the random numbers.
4. Match the random numbers to your outcomes.

Example:

You sell milk, dark and white chocolates in a shop. $P(\text{milk}) = 3/6$, $P(\text{white}) = 1/6$, $P(\text{dark}) = 2/6$.

Simulate the choice of chocolates that the next 10 customers will buy.

We are not looking at theoretical probability for each chocolate otherwise we could just work out $3/6$ of 10 and so on. We are using these to assign numbers to generate random numbers from that will tell us which chocolate each customer will choose. So, a bit more like experimental probability/relative frequency without the real-life situation.

1. Use a dice as there are 6 numbers in this scenario.
2. $\frac{3}{6}$ of 6 is 3 so assign numbers 1, 2, 3 on the dice to milk chocolate. $\frac{1}{6}$ of 6 is 1 so assign the next number, 4, to white chocolate. Assign numbers 5 and 6 on the dice to dark chocolate.
3. Roll the dice 10 times to generate the random numbers and record the results. E.g. 3,3,4,5,1,5,1,3,5,2.
4. Match the numbers to the outcomes – M, M, W, D, M, D, M, D, M.
You now know for the next ten customers you need 6 milk chocolates, 1 white chocolate and 3 dark chocolates.

Note that these results do not match with the probabilities in the question and they won't always as this is mimicking real life situations. Also remember that since this is a simulation these results are not necessarily accurate. To get a more reliable simulation repeat the simulation lots of times.

Questionnaires/Interviews:

A source of primary data

Questionnaire – A set of questions used to obtain data from the population/sample. Can be carried out via post, email, phone or face to face. The person completing the questionnaire is called the respondent.

Questions can be open or closed.

Open questions: Allows any answer. However, the wide range of different answers makes it difficult to analyse the data.

Closed questions: Has a fixed number of non-overlapping option boxes that only allow for specific answers or opinion scales. This makes data easier to analyse.

Features of a good questionnaire:

- Easy to understand
- Uses simple language
- Avoid leading questions such as “do you agree...?” – makes the respondent want to agree.
- Questions are relevant to the investigation
- Includes a time frame/unit in the question.
- Includes non-overlapping, exhaustive option boxes.
- Questions should not be offensive/personal/embarrassing
- Questions which are easy to analyse the results.

Problems with Questionnaires:

- **Non – response:** when people in the sample do not respond to the questionnaire. Could be due to people not wanting to answer the questionnaire or not understanding the questions.
 - Follow up on people who have not responded.
 - Collect each questionnaire yourself.
 - Offer an incentive to complete the questionnaire such as the opportunity to win a prize.
 - Use a pilot survey to test response rate or understandability of questions.
- **Sensitive questions:** Includes questions about people's health, age, weight, salary etc.
May make people uncomfortable so they may not answer truthfully which could distort the results. You can make respondents more comfortable by making the questionnaire anonymous and allowing them to answer the questionnaire in private or by using the random response method.

Random Response Method:

Uses a random event to decide how to answer a question which ensures that people who answer the question remain anonymous. You can use the survey results to calculate an estimate for the proportion of people who answered yes to the sensitive question.

Steps:

1. Find total who answered questions.
2. Find prob. (heads) if it is a coin.
3. Estimate no. of heads – Prob x total
4. Estimate number of “yes” answers that were truthful;
Yes answer – estimated no of heads
5. Estimate proportion of people who did the crime = D/C

Pilot Study: A **small scale replica** of the study to be carried out. Used to test the design and methods of the questionnaire.

Advantages:

- Helps you spot any questions that are unclear or ambiguous.
- Gives you an idea of the response rate
- Allows you to check the time and costs of the study.
- You can check that closed questions include all the possible answers.
- Can use pilot study to check that the questionnaire collects all the information needed.

Interviews: where you question each person individually.

Involves lots of specific questions or a list of topics.

Can be carried out face to face or over the phone or internet.

	Advantages	Disadvantages
Interview	<ul style="list-style-type: none"> • Interviewer can explain questions • Interviewer can put people at ease when having to answer personal qs • Respondents can explain their answers • High response rate 	<ul style="list-style-type: none"> • Less likely to answer personal questions and may be less honest • Time consuming • Expensive • Smaller sample size than questionnaire • Interviewer bias - interviewer may interpret answers to suit their opinion • Respondent may try to impress/guess the answer the interviewer wants.
Anonymous Questionnaire	<ul style="list-style-type: none"> • Respondents more likely to answer personal questions • No interviewer bias • Easy to send questionnaires to large sample size • Quick • Cheap 	<ul style="list-style-type: none"> • Some questions may not be understood • Researchers may not understand some of the responses • Low response rate

Problems with Collected Data:

Outliers - values that do not fit in with the pattern or trend of the data.

Can be extreme values or incorrectly recorded. If incorrectly recorded, these can be ignored. If extreme values, you need to decide whether or not to include them in the data as they may distort/skew your results.

Cleaning Data – fixing problems with the data. This could be done by:

- Identifying and correcting/removing incorrect data values or outliers.
- Removing units or symbols from the data,
- Putting all the data in the same format e.g. m/cm, capital/lowercase, words/letters.
- Deciding what to do about missing data.

Controlling Extraneous Variables:

- **Control Groups** – The control group (sometimes called a comparison group) is used in an experiment as a way to ensure that your experiment actually works. It's a way to make sure that the treatment you are giving is causing the experimental results, and not something outside the experiment.
 - Use random selection to select 2 groups of people, control and experimental groups.
 - Give the test group the treatment, control group no treatment
 - Compare results from 2 groups to see how effective treatment is
- Conditions must be exactly the same for both groups, only treatment must be different.

- **Matched pairs** - 2 groups of equally matched (age/gender etc.) people used to test effect of a particular factor. Everything in common except factor being studied.
The “pairs” don’t have to be different people — they could be the same individuals at different time. For example:
The same study participants are measured before and after an intervention.
The same study participants are measured twice for two different interventions.
The purpose of matched samples is to get better statistics by controlling for the effects of other “unwanted” variables. For example, if you are investigating the health effects of alcohol, you can control for age-related health effects by matching age-similar participants.

Hypotheses and Investigations:

Hypothesis - A statement (not a question) that can be tested by collecting and analysing data.

Stages of an Investigation:

- **Planning** – choose hypothesis, what data to collect (variables), how you will record data (data collection tables)
- **Collecting Data** – choosing data sources (primary/secondary), collection methods (questionnaire/interviews), control factors.
- **Processing and Representing data** – choosing diagrams and calculations.
- **Interpreting Results** – drawing conclusions from the results of the diagrams and conclusions
- **Evaluating methods** – looking at the strengths and weaknesses of your data collection methods, planning and diagrams and how well they helped to test the hypothesis.

Chapter 2 – Processing and Representing Data

Tables

Databases – Tables with a collection of data. They are a form of secondary data is the data is available online and, in most cases, easily accessible.

These tables usually contain information from real-life statistics, and you will be asked in the exam to extract and interpret information from it. These questions have multiple parts and many 1 marker sub-questions. You need to be able to use these tables to identify values, calculate totals/differences/percentages, describe trends and explain inconsistencies. One of the main inconsistencies will be that the percentages do not add up to 100% and this is due to rounding errors because individual percentages for columns/rows in the tables have been rounded.

As the data represents real-world statistics you may be asked to explain reasons for trends. Think about the data in terms of real-life rather than just an exam question. What real-life situation may affect the data you have?

Make	September 2016		September 2017		% change in sales
	sales	market share (%)	sales	market share (%)	
Ford	49 078	10.45	39 696	9.31	-19.12
Volkswagen	33 722	7.18	36 332	8.53	7.74
BMW	32 595	6.94	31 465	7.38	-3.47
Mercedes-Benz	31 988	6.81	31 430	7.37	-1.74
Vauxhall	41 697	8.88	31 058	7.29	-25.52
Audi	31 113	6.62	29 619	6.95	-4.80
Nissan	27 807	5.92	28 810	6.76	3.61
Toyota	18 888	4.02	19 222	4.51	1.77
Hyundai	17 039	3.63	16 587	3.89	-2.65
Kia	15 340	3.27	15 706	3.69	2.39
Land Rover	14 629	3.11	14 504	3.40	-0.85
Peugeot	16 130	3.43	12 810	3.01	-20.58
Renault	17 275	3.68	12 378	2.90	-28.35
Mini	13 119	2.79	12 282	2.88	-6.38

(Source: www.smmr.co.uk)

Two-Way Tables – Has information in two categories and has two variables so the data is called bivariate data.

To find missing values, start with the row or column that has only one value missing. Make sure the grand totals for the rows and columns add up to the same number.

When comparing data from two-way tables, write about comparisons between rows/columns but also individual cells.

Age	male	female	Total
18 to 22	2	4	
23 to 29	15		
30 to 36			21
Total	30	30	








(Source: www.wtatennis.com and www.atpworldtour.com)

Pictograms

Uses pictures or symbols to represent a particular amount of data. Always has a key to show the amount each symbol represents.

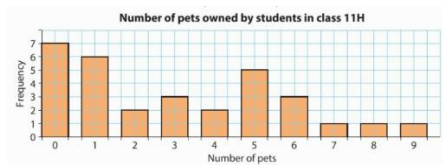
When drawing a pictogram, make sure that:

- Each symbol is the same size
- The symbols represent numbers that can be easily divided to show different frequencies, e.g. for a symbol that represents 4, you can draw a quarter of the symbol to show a frequency of 1.
- Spacings are the same in each row.
- There is a key to show the frequency that each symbol represents.

Hip-hop		Key:  represents 2 members
Indie rock		
Metal		
Pop		
R&B		
Other		

Bar Charts

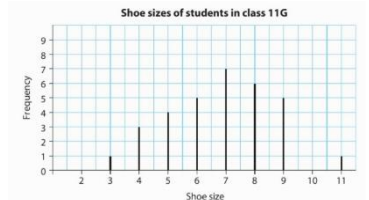
• Simple Bar Charts



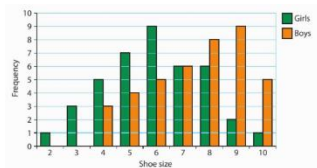
- Bars are equal width
- Equal gaps between bars
- Frequency on y-axis

• Vertical Line Graph

Similar to simple bar chart but with lines instead of bars.



• Multiple Bar Charts

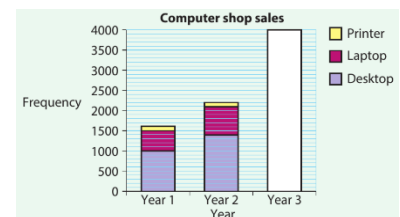


Can be used to compare two or more sets of data.

Has more than one bar for each class represented by different colours which is shown in the key.

• Composite Bar Charts

Has single bars split into different sections for each different category. Usually used to compare different times/days/years. The frequency of each component should be calculated by subtracting the upper frequency of that component with the lower frequency. Do not just read off the y-axis (unless looking at total frequencies or the bottom component).



Stem and Leaf Diagrams

A good way of organising data without losing any of the detail – All the original data is in the diagram but looks simple. It also shows the shape of the distribution – whether most of the data lies at the beginning, the end or is distributed in the middle.

Each value is split into a 'stem' and 'leaf' – Stems can be more than one digit, leaves are single digits only. No need for commas in between leaves. Leaves must be written in order from smallest to largest – this makes it easier to find mode and median.

How to draw one:

- 1) Put the **first digits** of each piece of data in **numerical order** down the left hand side.
- 2) Go through each piece of data in turn and put the remaining digits in the **correct row**.
- 3) Re-draw the diagram, putting the pieces of data in **numerical order**.
- 4) Add a **key**.

• Back-to-back Stem and Leaf Diagrams

- Shows two sets of data sharing the same stem so that you can easily compare them.
- Numbers closest to the stem are smallest.
- Use two different keys for each set of data.

Pie Charts

A way of displaying data to show how something is shared or divided into categories, Each sector shows what proportion that category represents of the total data,

Area of Pie Chart = Total Frequency

Angles add up to 360° .

How to draw a pie chart:

1. Total up the frequency
2. Calculate the angle for each frequency. $360/\text{frequency}$
3. Calculate the angle for each category will be by multiplying your previous answer by the frequency.
4. Make sure all the angles now add up to 360.
5. Draw the pie chart.
6. Label the sectors.

Interpreting Pie Charts – Remember pie charts show proportion and not numbers.

Comparative Pie Charts

Can be used to compare two sets of data of different sizes. The areas of the two circles should be in the same ratio as the two frequencies.

Why? Drawing two pie charts the same size can be misleading.

Area of Pie Chart = Total Frequency

So, the larger the pie chart, the greater the frequency.

To compare the total frequencies, compare the areas.

Working out radius of second pie chart:

1. Divide both areas (this gives you the area scale factor)
2. Square root answer (this gives you the scale factor for radius)
3. Multiply by radius of first pie chart.

$$r_2 = r_1 \frac{\sqrt{F_2}}{\sqrt{F_1}}$$

- If pie chart B is larger than pie chart A then pie chart B has a greater frequency.
- If both pie charts then have the same angle for a sector that means that sector has a greater frequency in pie chart B even though the proportions are the same because it has a larger area.

Population Pyramids

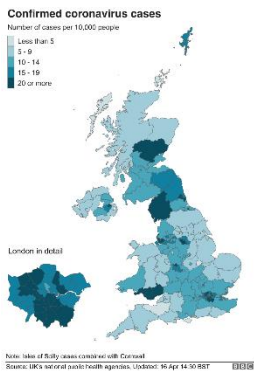
Shows distribution of ages in a population, in numbers or proportion/percentages.

They are used to compare two sets of data, usually genders or two geographical areas.

When comparing the data look at the shape of the distribution.

- If it looks like a pyramid with smaller bars at the top that means there is a higher proportion of younger people in the population and less older people. This could be because of short life expectancy (how long people live), high birth rates or high death rates.
- If the diagram looks more or less straight that means there is a similar proportion of older and younger people in the population which could be because of lower birth/death rates or that the life expectancy is increasing.
- An upside-down pyramid with larger bars at the top and smaller bars at the bottom shows that the population has a larger proportion of older people compared to younger people. This could be because of low birth/death rates, longer life expectancy or the location might be far from the city or a coastal area where older people are retiring to.

Choropleth Maps (not Chloropeth)



Think colour by numbers.

They split a geographical area into different regions which are then shaded.

The darker the shading the higher the frequency for that area.

Each map has a key to show what the shading represents.

Interpreting:

The area of the map which is shaded darkest has the highest proportion/percentage.

Look at the key for the shading to read off percentages/numbers.

Cumulative Frequency Diagrams

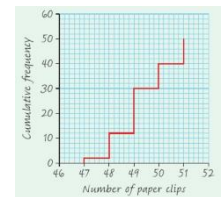
Cumulative frequency is a running total of the frequencies.

To work out CF for a class interval, add all the frequency for that class interval and the CF of the previous class interval.

Use upper bounds for x-axis when plotting points.

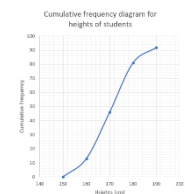
CF Step Polygons – Use for **discrete** data.

Plot the points using upper bound of class interval and join points using straight lines by going *across then up*.



CF Curves – Use for **grouped continuous** data.

Plot points using upper bound of class interval and connect with a *smooth curve*.



Estimating values from CF diagrams:

- **Median**
 - Work out median value by dividing total frequency by 2.
 - Find on Y-axis
 - Draw horizontal line from that value to curve/line
 - Read off value from x-axis
- **Interquartile Range (IQR)**
 - Work out 25% and 75% values
 - Find on y-axis
 - Draw horizontal line from that value to curve/line
 - Read off values from x-axis
 - Subtract them (Big one – small one)
- **Estimating more than/greater than values**
 - Draw a vertical line from the value in the question on the axis to the curve.
 - Read off corresponding y-axis value.
 - Subtract from total frequency.

Histograms

Represents continuous data from grouped frequency tables.

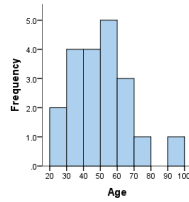
No gaps between bars.

Equal Class Widths

x-axis = data

y-axis = frequency

Looks like bar charts without gaps.

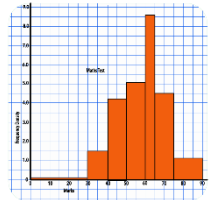


Unequal Class Widths

Area of bar = frequency

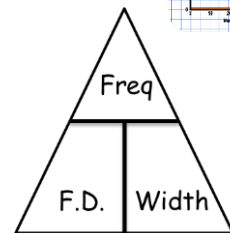
Y-axis = Frequency Density (not frequency)

The idea is that the frequency density reflects the 'concentration' of things within each range of values.



$$\text{Frequency Density} = \frac{\text{Frequency}}{\text{Class Width}}$$

$$\text{Frequency Density} \times \text{Class Width} = \text{Frequency}$$



Drawing Histograms:

1. Calculate class widths for each class interval
2. Calculate frequency density for each class interval using $FD = F/CW$ formula.
3. Draw a suitable scale on y-axis labelled frequency density.
4. Draw bars using frequency density data. (Remember the bars have no gaps in between)

Estimating frequencies from histograms:

With these questions you are using the class widths and frequency density from the histogram to work out frequencies. Be careful when calculating class width as some intervals may not include the entire bar.

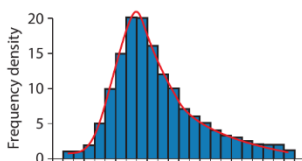
1. Find the bars that cover the range you need from the question.
2. Work out the frequency for each bar using the $FD \times CW = F$ formula.
3. Add the frequencies.

To compare histograms, they need to have the same class intervals and frequency density scales.

When comparing histograms, describe the shape of the distribution and what this shows.

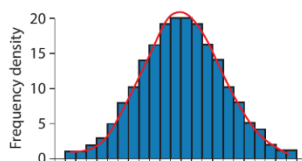
The Shape of a Distribution

This is the shape formed by the diagram. It can be positive, negative or symmetrical.

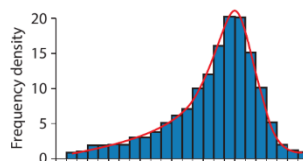


This distribution has positive skew. Most of the data values are at the lower end. Example: The age at which a person learns to write.

The distribution is stretched out in the positive direction →.



This distribution is symmetrical. It has no skew. Example: The lengths of leaves on a tree.



This distribution has negative skew. Most of the data values are at the upper end. Example: The age at which a person dies.

The distribution is stretched out in the negative direction ←.

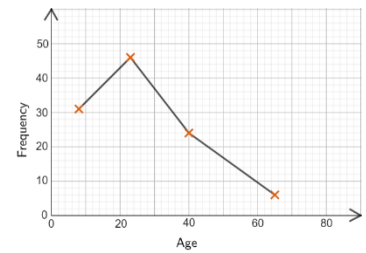
Frequency Polygons

Similar to histograms with equal class widths but without the bars.

Uses mid-points of class intervals and points are plotted and then joined together with straight lines.

Common errors:

- Midpoints not used
- Joined together at the bottom
- Points not joined together with straight lines but with a curve instead.



Misleading Diagrams

Diagrams can be misleading because of their shape or because of axes and scales.

Types of Misleading Diagrams:

- Pictograms – Same symbol and size needs to be used for all the diagrams and key needed.
- 3D charts – They distort parts of the diagram making it difficult to read off values.
- Colours – Some colours may make parts of the diagram stand out more thus making it seem more important when it may not be.
- Lines that are drawn too thick making it difficult to read information from the diagram.

Axes and Scales that can be misleading:

- Scales that do not start at zero.
- Missing values on the scales.
- Axes that are unevenly scaled.
- Axes that are not labelled.
- Not using a key.

Chapter 3 – Summarising Data

Averages

A measure of central tendency (represents the 'centre' of a set of data). Includes mode, median and mean.

Mode

The one that appears the **most** (remember the Mo in mode and Mo in most) – the most common value.

Modal Class – the class with the highest frequency (the frequency value is not the mode but the column/row next to it).

Median

The **middle** value.

Discrete Data:

1. Put the numbers in order from smallest to largest.
2. The **median is the $\frac{1}{2}(n + 1)$ th value** – this means find your total frequency, add 1 to it and then divide by 2. The answer is not your median but the position your median will be in the list of data.
Example: If total frequency is 23, the median position will be $\frac{1}{2}(23+1) = 12^{\text{th}}$ number in the list after being put in order.
3. Find the median position from the list of values. This is your median value.
If the median position is a decimal value such as 7.5 then you would find the 7th and 8th values in the list and then divide by 2.

If data is in a frequency table, add the frequency values (like you would do with cumulative frequency) until you reach a row that includes the median position in between it. The median is the category/class that contains the $\frac{1}{2}(n + 1)$ th value.

Grouped Data:

For grouped continuous data (has classes with the inequality symbols), the median is the **$\frac{1}{2}$ nth value**.

The median class is the class interval which contains the median position.

Sometimes you may be asked to work out an estimate for the median value rather than the median class.

For grouped data your median will always be an estimate as you do not know exact values.

Estimate Median using Linear Interpolation:

1. Use $\frac{1}{2}n$ to find the median position.
2. Find Cumulative Frequency (CF) of the frequency column until you reach the class interval that contains the $\frac{1}{2}$ nth value. This is the group that contains the median.
3. Find the median's position in the group and see how many more values you need in that class to get to the median.
Do this by subtracting the CF of the group above from your $\frac{1}{2}$ nth value.
4. Divide this number by the frequency for the median class.
5. Multiply your answer by the class width.
6. Add your answer to the lower bound for the class interval. This is your estimate for the median value.

Mean (also called Arithmetic Mean)

The **sum of all the values divided by the number of values.**

Discrete Data:

1. Add all the values.
2. Divide by the number of values.

$$\text{Formula for Mean: } \bar{x} = \frac{\sum x}{n}$$

Where \bar{x} = mean

Σ = Greek letter sigma which is the symbol for 'sum of'

x = data values

n = number of data values

So Σx means sum of all values

Frequency Table (not grouped):

1. Add an extra column to the table and label as $f \times x$.
2. Multiply the values in the first two columns for each row and write the answer in the new third column.
3. Add up the third column – this gives you the total (Σfx)
4. Add up the frequency column (Σf)
5. Divide answer to 3 by answer to 4.

$$\text{Formula: } \frac{\sum fx}{\sum f}, \quad f \text{ stands for frequency}$$

$$\sum fx = \text{total of 3}^{\text{rd}} \text{ column}$$

$$\sum f = \text{Total Frequency}$$

Frequency Table (grouped):

1. Add 2 extra columns to the table and label as midpoint and $f \times \text{midpoint}$.
2. Calculate the midpoint of the class intervals
3. Multiply the midpoint and frequency values for each row and write answers in last $f \times \text{midpoint}$ column.
4. Add up the $f \times \text{midpoint}$ column – this gives you the total (Σfx)
5. Add up the frequency column (Σf)
6. Divide answer to 4 by answer to 5

$$\text{Formula: } \frac{\sum(f \times \text{midpoint})}{\sum f}$$

Weighted Mean

For data that has **different number of values or weights in each group**. It is used to combine different sets of data where one set is more important (has more weighting) than another – example: Maths and English for progress 8 scores has twice the weighting as other subjects or subjects where controlled assessment is 25% and exam is 75% of final mark.

$$\text{Weighted Mean} = \frac{\sum(\text{weight} \times \text{value})}{\sum \text{weights}}$$

Geometric Mean

The n th root of the product of all the values. Useful to compare things with different properties that aren't immediately comparable if they are out of different frequencies or values are dependent on previous values. Useful for looking at average growth rates.

$$\text{Geometric Mean} = \sqrt[n]{\text{value}_1 \times \text{value}_2 \times \dots \times \text{value}_n}$$

Where n is the total number of values

Transforming Data

For large data values you may want to make the numbers smaller so that it saves you time working with large numbers even on a calculator (it is easy to make mistakes typing in large numbers).

You can find the mean by taking away the same large number from all the values (so that you're left with smaller values), find the mean of these numbers and then add that number back on.

This works best with whole numbers.

For decimal values use **Linear Transformation**:

1. Add/subtract a number from the data values.
2. Multiply/divide them by a number

Steps 1 and 2 can be done the other way around depending on the data set.

Example: For values 1.04, 1.09, 1.03, 1.12 you might want to subtract 1 from all the values first and then multiply by 100 to make them whole numbers.

For values, 0.00152, 0.00149, 0.00141, you might want to multiply by 100,000 first to turn them into whole numbers and then subtract 140 to make them into small numbers that are easier to perform calculations with.

3. Find the mean of the new numbers.
4. Reverse what you did to the original numbers.

How Changes to your Data affect the Averages

Mode – could change only if the new value changes which value appears the most. Could also make the data bimodal if there are now two values that appear the same amount.

Median –
 If you add a value that is greater than the median, the median might increase.
 If you add a value that is smaller than the median, the median might decrease.
 If you remove a value that is greater than the median, the median might decrease.
 If you remove a value that is smaller than the median, the median might increase.
 If you add/remove one value that is greater and one that is smaller than the median, the median stays the same.

Mean –
 If you add a value that is greater than the mean, the mean increases.
 If you take away a value that is less than the mean, the mean increases.
 If you add a value that is less than the mean, the mean decreases.
 If you take away a value that is greater than the mean, the mean decreases.
 If you replace a value in your data with another number that is greater/smaller than the original, the mean will also change.

Deciding which Average to Use

	Advantages	Disadvantages
Mode	<ul style="list-style-type: none"> • Easy to use • Always a value in the data • Unaffected by extreme values • Can be used with quantitative and qualitative data 	<ul style="list-style-type: none"> • There may not be a mode or may be more than one mode. • Cannot be used to calculate measures of spread. • Not always representative of the data – can include extreme values and can be a misleading value far from the mean.
Median	<ul style="list-style-type: none"> • Easy to find when data is in order • Unaffected by outliers/extreme values • Best to use with skewed data • Can be used to calculate quartiles, IQR and skew. 	<ul style="list-style-type: none"> • May not be data value • Not always representative of the data.
Mean	<ul style="list-style-type: none"> • Uses all the data • Can be used to calculate standard deviation and skew. 	<ul style="list-style-type: none"> • May not be a data value • Always affected by extreme values or outlier. • Can be distorted by open-ended classes.

Measures of Dispersion

Range

How **spread** out the data is.

The difference between the biggest and smallest values.

$$\text{Range} = \text{Largest Value} - \text{Smallest Value}$$

For data from tables the largest value is the biggest number from the first column and the smallest value is the first number from the first column.

Interquartile Range (IQR)

"Between Quartiles"

The middle 50% of the data when in order.

$$\text{Interquartile Range} = \text{Upper Quartile} - \text{Lower Quartile}$$

Lower Quartile (LQ) – The value $\frac{1}{4}$ of the way through the data. 25% of the data is less than the LQ.

Upper Quartile (UQ) – The value $\frac{3}{4}$ of the way through the data. 25% of the data is above than the UQ.

Discrete Data

LQ = $\frac{1}{4}(n+1)$ th value

UQ = $\frac{3}{4}(n+1)$ th value

1. Put the data in order, smallest to largest.
2. Work out the lower and upper quartiles using the above formulae.
If your LQ is 2.75th value, divide the interval between the 2nd and 3rd values into quarters and use this to work out what the LQ value will be – find the value $\frac{3}{4}$ of the way between the 2nd and 3rd values.
If 2nd value = 2 and 3rd value = 4, then 2.75th value = $((4-2)/4)*3 + 2^{\text{nd}} \text{ value} = 0.5*3 + 2 = 1.5+2=3.5$
3. IQR = UQ – LQ

Grouped Data

LQ = $\frac{1}{4}n$ th value

UQ = $\frac{3}{4}n$ th value

1. Draw your CF curve.
2. Use the above formulae to find the positions for LQ (25%) and UQ (75%).
3. Draw lines from the 25% and 75% marks on the y-axis. The corresponding x-axis values give you your LQ and UQ values.
4. IQR = UQ – LQ

Interpercentile Range (IPR)

The difference between 2 percentiles.

Percentiles – divide the data into 100 equal parts.

Gives a more flexible view of the spread of the data e.g. could be used to analyse the gap between highest and smallest earners.

Interpercentile Range (IPR) = Value of larger percentile – Value of smaller percentile

1. Divide the percentiles you need by 100, then multiply that decimal by the total frequency. E.g. for 70th percentile with frequency 80, $(70/100)*80 = 56^{\text{th}}$ position.
2. Find the position on the y-axis of your graph and read across to find the corresponding x-values.
3. To find the interpercentile range, subtract the two percentiles that you calculated.

If calculating IPR from a table, carry out linear interpolation as you would when finding median from a table.

Interdecile Range

The difference between 2 deciles – usually the difference between the first and ninth deciles.

Deciles – divides the data into 10 equal parts.

Interdecile Range = 9th decile – 1st decile

Standard Deviation (SD)

A measure of **how far all the values are from the mean** value, or how spread out they are.

The smaller the SD, the closer the data is to the mean,

The larger the SD, the more spread out the data is from the mean.

Discrete Data

Formulae: $\sigma = \sqrt{\frac{1}{n} \sum (x - \bar{x})^2}$ OR $\sigma = \sqrt{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2}$ \bar{x} = mean, σ = standard deviation

Using the first formula:

1. Calculate the mean.
2. Subtract the mean from each data value and square the answer – it might be useful to do this in a table.
3. Add up all the answers to step 2.
4. Divide by the number of values.
5. Square root.

Using the second formula:

1. Calculate the mean.
2. Square each value
3. Add up the answers to step 2.
4. Divide by number of values.
5. Subtract the square of the mean from your answer to step 4.
6. Square root.

Frequency Table (not grouped)

Formulae: $\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$ OR $\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2}$ $\sum f = n$ (total frequency)

$$\frac{\sum fx}{\sum f} = \text{mean}$$

Using the first formula:

1. Calculate the mean.
2. Create a new column for $x - \bar{x}$. Subtract mean from each value in the first column.
3. Square each answer to step 2 – create new column.
4. Multiply each answer in step 3 by corresponding frequency – create new column.
5. Add answers to step 4 – add the last column.
6. Divide answer to step 5 by total of frequency column.
7. Square root.

Using the second formula:

1. Add three columns: fx , x^2 and fx^2 and calculate these values. Remember to add these columns.
2. Calculate the mean.
3. Substitute your values into the formula and work out the answer.

Grouped

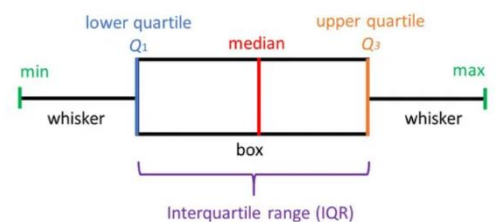
For grouped frequency tables, follow the same step as for frequency table but use the midpoint for x . You may need to create an extra column to your table for the midpoint before carrying out the above steps.

Box Plots

Divide the data into sections that each contain approximately 25% of the data in that set. Represents important features of the data and gives a summary of the spread/skew of the data.

Box Plots include 5 pieces of information about the data:

1. Minimum Value – the lowest score, shown at the far left of the diagram
2. Lower Quartile (LQ) – 25% of data is below this
3. Median – Mark the middle of the data – 50% of the data is above/below this value
4. Upper Quartile (UQ) – 25% of data is above this value/75% of data is below it.
5. Maximum Value – The highest score, shown at the far right of the diagram



The total length of the box plot represents the range.

The box represents the middle 50% and the IQR.

Drawing Box Plots:

1. Calculate your LQ, UQ, median and identify your minimum and maximum value.
2. Mark these 5 points on your diagram – the minimum and maximum values with small lines and the other three with bigger lines.
3. Draw a box around the big three lines.
4. Connect the box to the min/max points using horizontal lines.

Outliers

Values that are **far from the rest of your data** and don't fit the general pattern.

Can show errors in the data

Including outliers may misrepresent your data but not including them could falsify your data.

They **distort the data** so you need to identify them.

Outliers are more than 1.5 X IQR above UQ or below LQ.

$$\text{Outliers are values } > UQ + (1.5 \times IQR) \\ \text{or } < LQ - (1.5 \times IQR)$$

1. Work out IQR
2. Find 1.5 x IQR
3. Subtract this value from LQ and add to UQ.
4. These values are now your new min/max points for your box plot. Any values in your data outside of this range are outliers.
5. Mark outliers with an X on your box plot.

Outliers can also be found using the mean and standard deviation – they are values more than 3 SD away from the mean.

$$\text{Outliers} = \text{Values outside } \bar{x} \pm 3\sigma$$

Interpreting box plots – Compare median for measure of average and range or IQR for measure of spread.

Remember to compare in context of the question for full marks.

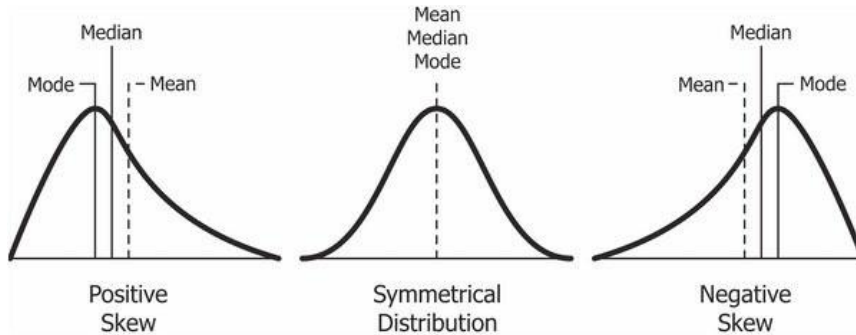
Compare skewness of both box plots.

Skewness

Describes the shape of the distribution and tells you how the data is spread out.

If the data is skewed, it means most of the values are more on one side of the median.

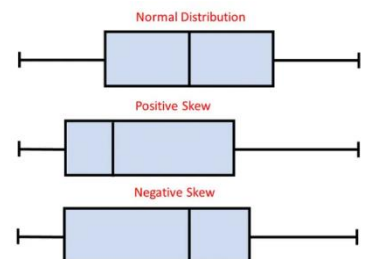
Types of Skew:



- **Positive Skew** – most values are at the beginning of the data set and values towards the end are more spread out. The majority of the data is low and there are very few higher values. The tail of the curve goes in the positive direction of the x-axis.
Mean > Median > Mode
- **Negative Skew** - most values are at the end of the data set and values towards the beginning are more spread out. The majority of the data is higher and there are very few low values. The tail of the curve points towards the negative direction of the x-axis.
Mean < Median < Mode
- **Symmetrical** – There is no skew. The data is evenly distributed on both sides of the median.
Mean = Median = Mode

Skewness on Box Plots:

- **Normal Distribution/Symmetrical** – When median is halfway between LQ and UQ.
- **Positive Skew** – Median closer to LQ.
- **Negative Skew** – Median closer to UQ.



Skewness using the Formula:

$$\text{Formula: } \text{Skewness} = \frac{3(\text{mean} - \text{median})}{\text{standard deviation}}$$

- **Positive Value** = positive skew. The larger the value, the larger the skew.
- **Negative Value** = Negative Skew. The smaller the value, the stronger the skew.
- **Value of 0** = No Skew/Symmetrical.

Comparing Data Sets

Compare using a **measure of average (mean/median/mode) and spread (range/IQR/SD)** or skewness. Always make reference to individual values and mention which data set is larger/smaller than the other clearly.

Always **interpret in context** – link back to the scenario in the question and labels on axes.

Example Comparisons and Interpretations of Data

Replace 'data sets' and 'results' with appropriate keyword from the question.

- Comparing Averages:

Mean/median/mode for data set A is larger than mean/median/mode data set B so on average data set A is more ... than data set B.

- Comparing spread:

Range/IQR/SD for data set A is larger than that of data set B so the 'results' of data set A are more spread out/less consistent than those of data set B.

Data A has a smaller range/IQR/SD than data set B which means the 'results' for 'data set A' are more consistent.

Remember lower SD means values are closer to the mean and therefore similar.

- Comparing Skew:

Box Plot for data set A is positively skewed so majority of 'results' were low with few higher 'results'.

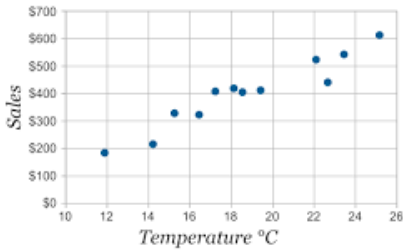
Box plot for data set A is negatively skewed so majority of 'results' were high with few lower 'results'.

When comparing data make sure to pair the appropriate values of average and spread.

Average	Measure of Spread
Mode	Range
Median	Range/IQR
Mean	Range/SD

Chapter 4 – Scatter Diagrams and Correlation

Scatter Diagrams



Used for **bivariate** data to show if there is a relationship between two variables.

Explanatory variable (independent – the one that you are changing) is plotted on the x-axis.

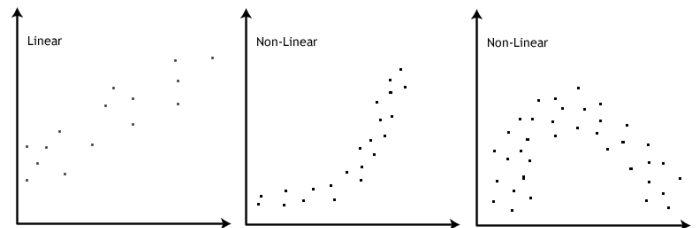
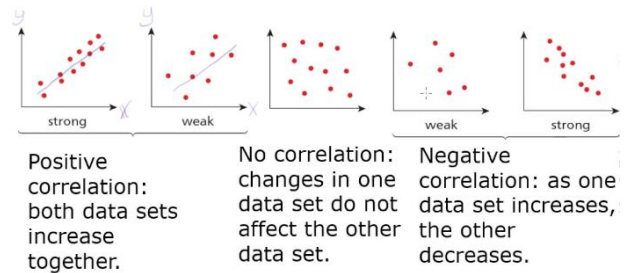
Response Variable (dependent – the one you are measuring) is plotted on the y-axis.

Plot the points with crosses. Do not join them up.

Correlation

The relationship between two variables.

- **Positive Correlation** – As one variable increases, so does the other.
- **Negative Correlation** – As one variable increases, the other decreases.
- **Zero Correlation** – The points are randomly scattered.
- **Linear Correlation** – When the points lie close together near a straight line.
- **Non-Linear Correlation** – When the points lie close together but the pattern formed by them is a curve.



Causal Relationships

Causation – When one variable causes a change in another.

Correlation shows that there may be a link between two variables. Correlation does not imply causation.

Example:

Causal Relationship – increase in temperature = Increase in ice cream sales

Correlation only – Sales of chocolate and sales of clothes having a positive correlation.

Multiple Factors – In real life situations there are usually multiple factors interacting to cause variables to change.

Example: A positive correlation between fat in liver and reaction time does not mean one causes the other. There could be a third variable, such as amount of alcohol consumed, which both variables depend on.

Line of Best Fit (LOBF)

A straight line drawn through the middle of the points so the points are evenly scattered on either side of the line.

Needs to be a straight line.

Needs to be close to as many points as possible.

Has to go through the mean point.

The closer the points are to the LOBF, the stronger the correlation.

$$\text{Mean Point } (\bar{x}, \bar{y}) = (\text{Mean of } x \text{ values}, \text{Mean of } y \text{ values})$$

Interpolation and Extrapolation

Using the LOBF to make predictions of unknown values.

Interpolation – When the LOBF is used to make predictions **within the range of data** given (you don't need to extend your LOBF more).

Tends to be **reliable** provided the LOBF is correct.

Extrapolation – When the LOBF is used to predict values **outside of the range of values** given (you may need to extend your LOBF for this).

Not always reliable as trends may change.

Values estimated from extrapolation are less reliable the further they are from the range of data.

Equation of LOBF

LOBF is also known as Regression Line.

$$\text{Eqn of LOBF: } y = ax + b$$

Where a is the gradient and b is the y-intercept.

Gradient, a – The rate of increase of the response variable in relation to explanatory variable.

Y-intercept, b – The value of the response variable when explanatory variable is 0.

Drawing Regression Line:

1. Use equation of LOBF to calculate coordinates of two points.
2. Plot these points on the scatter graph.
3. Join them up with a straight line (extend your line a little if you need to).

Finding Equation of LOBF/Regression Line:

1. Select 2 points on the graph – pick 2 points on the corners of the square for more accurate readings. Call these 2 points (x_1, y_1) and (x_2, y_2) .
2. Calculate the gradient, a .

$$a = \frac{y_2 - y_1}{x_2 - x_1}$$

3. Find the y-intercept, b .
You could do this by:
 - a. Read the point of the graph by extending the LOBF.
 - b. Use the formula $b = y_1 - ax_1$ using one of the 2 points you used for the gradient.
4. Write the equation in the form $y = ax + b$ with the correct values for a and b .

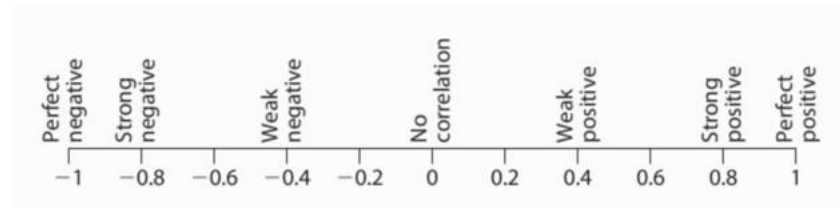
Spearman's Rank Correlation Coefficient (SRCC), r_s

Measures the strength of the correlation between 2 variables.

SRCC is always between -1 and 1.

The closer the value is to 0, the weaker the correlation.

The further the value is from 0, the stronger the correlation.



- If r_s near 1, there is a strong positive correlation
- If $r_s = 0$, there is zero correlation
- If r_s near -1, there is a strong negative correlation.

$$SRCC, r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

Where d = difference between ranks
 n = number of values

Calculating SRCC:

1. Rank both sets of data (largest to smallest)
2. Find the difference between each pair of ranks
3. Square the differences
4. Add the square of differences
5. Find the value of n – count the number of pairs of data.
6. Substitute into the formula – remember the 1 at the beginning.
7. Interpret your SRCC value in terms of correlation and strength of correlation – make this in context of the question.

Pearson's Product Moment Correlation Coefficient, PMCC

Measures the strength of linear correlation between two variables.

PMCC also between -1 and 1.

It is calculated using actual data values and not ranks so can be used for data that can't be ranked – don't worry you won't have to calculate PMCC.

- If r near 1, there is a strong positive correlation
- If $r = 0$, there is zero correlation
- If r near -1, there is a strong negative correlation.

SRCC vs PMCC

SRCC	PMCC
Measures the strength of correlation between 2 variables	
Have correlation between -1 and 1	
Tests for linear and non-linear correlation	Tests for linear correlation only
Best used for data that can be ranked	Can be used for data that can't be ranked as well

If there is a non-linear positive relationship between 2 variables then the SRCC and PMCC will both be positive but the SRCC will be closer to 1, or -1 for negative relationship.

Chapter 5 - Time Series

Time Series Graphs

A line graph with **time plotted on the x-axis**.

Used to spot trends – usually going up or down or fluctuating (going up and down).

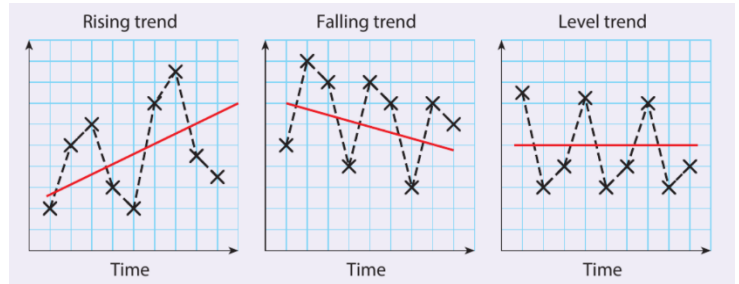
Time series is a set of data collected over a period of time at equal intervals – could be years, quarter, months or days.

Trend Lines

Shows the **general trend** of the data.

When drawing trend lines ignore fluctuations and follow the general pattern (is the data going upwards or downwards)

Trend line may show a rising (upwards) trend, falling (downwards) trend or level trend.



Moving Averages

An average worked out for a given number of successive observations

A good way to see trends in data with large variations – they smooth out fluctuations and make the trend line more accurate.

Plot moving averages at the midpoint of the time interval. Do not join the up – use LOBF.

Example: For 3 point moving average,

Term	Autumn 2000	Spring 2001	Summer 2001	Autumn 2001	Spring 2002	Summer 2002
Number of people	520	300	380	640	540	500

1st Value = $(520+300+380)/3=400$, Plot at middle of these 3 values so at Spring 2001.

2nd Value = $(300+380+640)/3=430$, Plot at Summer 2001

3rd Value = $(380+640+540)/3=520$, Plot at Autumn 2001

4th Value = $(640+540+500)/3=560$, Plot at Spring 2002

Connect them with a LOBF

For 4 point moving averages the midpoint would be halfway between the 2nd and 3rd values.

Seasonal Variations

Variations may be:

- A general trend – shown by the trend line
- Seasonal Variations – a pattern that repeats at a specific point every cycle.

Seasonal Variations in a time series follow a regular time period, like days of the week or seasons. Consider real life scenarios that may cause these variations – you need to interpret these in context of the question.

The Seasonal Variation at a point is how much the value varies from the trend.

Calculating the Seasonal Variation at a point:

$$\text{Seasonal Variation} = \text{Actual Value} - \text{Trend Value}$$

Estimated Mean Seasonal Variation (EMSV) –

Also called Average Seasonal Effect.

The average of all the seasonal variations for the same point in each cycle.

Example: the average of all the quarter 4 Seasonal Variations between 20012 and 2016.

$$\begin{aligned} \text{Estimated Mean Seasonal Variation} \\ = \text{Mean of all the seasonal variations for that season} \end{aligned}$$

Predicting Values

The trend line and estimated mean seasonal variations can be used to predict future values.

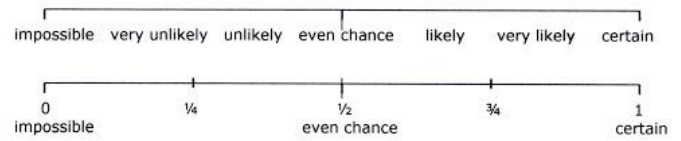
$$\text{Predicted Value} = \text{Trend Line Value (from graph)} + \text{EMSV}$$

Chapter 6 – Probability

Simple Probability

Probability is a measure of **how likely** an event is to happen.

Probabilities can be written as fractions, decimals or percentages.



An **outcome** is a possible result of an experiment or trial.

Example: when rolling a dice there are 6 different possible outcomes; 1, 2, 3, 4, 5 and 6.

An **event** is a specific thing that has a probability of happening. Example: rolling an even number.

$$P(\text{event}) = \frac{\text{Number of successful outcomes}}{\text{Total number of outcomes}}$$

The probabilities of all outcomes add up to 1.

The **expected frequency** of an event is the number of times you expect the event to happen. This does not mean it will actually happen this many times.

Example: P(Heads on a coin) = $\frac{1}{2}$ so if you flip a coin 10 times you would expect the coin to land on heads 5 times. This may not always happen in real life if you try this but is what should happen in theory.

$$\text{Expected Frequency of Event A} = P(A) \times \text{number of trials}$$

Experimental Probability

In real-life situations the outcomes of all event aren't equally likely do you have to use results of previous trials to predict future probabilities.

Trial – Each experiment that happens.

$$\text{Estimated Probability} = \frac{\text{Number of trials with successful outcomes}}{\text{Total number of trials}}$$

Estimated Probability is also called Relative Frequency.

The more trials there are, the more accurate the probability should be.

As the number of trials increases, the relative frequency will get closer to the theoretical probability.

Risk

Probability of an event occurring for **negative events**.

Relative frequency can be used to predict bias and assess risk.

For Bias:

A fair coin should land on heads and tails approximately $\frac{1}{2}$ the time each. If the coin is biased it will land on one side more than the other. You can check this by increasing the number of trials and seeing if the P(heads) is getting closer to the theoretical probability of $\frac{1}{2}$.

Risk is when collected data is used to predict how likely a negative event is to happen e.g. a house being flooded or the chance of an 18 year old having a car accident – mostly used by insurance companies to decide how much to charge you.

$$\text{Risk} = \frac{\text{Number of trials in which event happens}}{\text{Total number of trials}}$$

2 types of risk:

1. **Absolute Risk** – how likely an event is to happen. This is just relative frequency.
2. **Relative Risk** – How much more likely an event is to happen for one group compared to another group (e.g. comparing the probability of developing lung cancer for smokers and non-smokers).

$$\text{Relative Risk} = \frac{\text{Risk for those in the group}}{\text{Risk for those not in the group}}$$

Sample Space Diagrams

Sample Space – A list of all the possible outcomes.

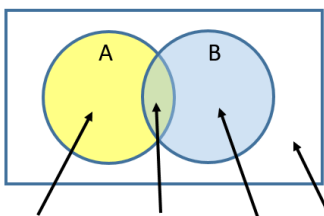
Example: When rolling a fair six-sided dice the sample space is 1, 2, 3, 4, 5, and 6.

Sample Space Diagram – A table used to represent the outcomes of two events.

	1	2	3	4	5	6
1	1,1	2,1	3,1	4,1	5,1	6,1
2	1,2	2,2	3,2	4,2	5,2	6,2
3	1,3	2,3	3,3	4,3	5,3	6,3
4	1,4	2,4	3,4	4,4	5,4	6,4
5	1,5	2,5	3,5	4,5	5,5	6,5
6	1,6	2,6	3,6	4,6	5,6	6,6

Example: The table on the right shows all the possible outcomes if you roll 2 fair six-sided dice. You can see that there are a total of 36 probabilities.

Venn Diagrams



Uses overlapping circles to represent all the outcomes of two or three events happening.

Each region of a Venn diagram represents a different set of data.

The whole rectangle represents all the possible outcomes.

Venn diagrams can be used to work out probabilities.

Objects here are in set A but not set B	Objects here are in both sets A and B	Objects here are in set B but not set A	Objects here are not in set A or set B.
---	---------------------------------------	---	---

Completing Venn Diagrams:

1. Draw and label the Venn diagram
2. Fill in any known values.
3. Use letters to label any area where you don't know the formulae.
4. Work out missing values.
5. The sum of all probabilities in a Venn diagram must equal to 1.

Mutually Exclusive and Exhaustive Events

Mutually Exclusive Events – Events that **CANNOT happen at the same time**.

Example: Getting heads and tails on a coin on the same flip.

For 2 mutually exclusive events, A and B:

$$P(A \text{ or } B) = P(A) + P(B)$$

Exhaustive Events – If the set **contains ALL the possible outcomes**.

Example: When rolling a fair dice, the events P(even) or P(odd) are a pair of exhaustive events as they cover all the possible outcomes you can have when rolling a dice.

The sum of mutually exclusive, exhaustive events is equal to 1.

$$P(A) + P(\text{not } A) = 1$$

$$P(\text{not } A) = 1 - P(A)$$

Addition Law

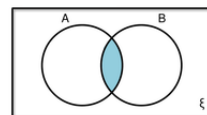
Also known as the **General Addition Law**.

Used for events that are **not mutually exclusive** – events that can happen together.

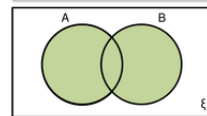
When two events can happen together and you want to find the probability of both of them happening you don't want to include the overlap – this is the intersection part of the Venn diagram, $P(A \cap B)$.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

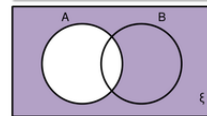
Also written as: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$



The **i**ntersection is where two sets overlap.
 $A \cap B$
This means **A and B**.



If you put two sets together, you get the **U**nion.
 $A \cup B$
This means **A or B**. Think marriage, they become 1!



The **complement of A** is the region that is not A.
 A'
This means **not A**. Or outside of A

$P(A \cap B) = P(A \text{ and } B)$. The intersection/overlap part of the Venn diagram.

$P(A \cup B) = P(A \text{ or } B)$. On a Venn diagram this is the union of A and B and includes everything in both circles, including the intersection.

Independent Events

Unconnected Events. The outcome of one event does not affect the outcome of the other event.

Example: Flipping a coin and then rolling a dice. The coin landing on tails will not affect what number the dice lands on.

Multiplication Law for Independent Events:

For 2 independent events, A and B:

$$P(A \text{ and } B) = p(A) \times P(B)$$

For 3 independent events, A, B and C:

$$P(A \text{ and } B \text{ and } C) = P(A) \times P(B) \times P(C)$$

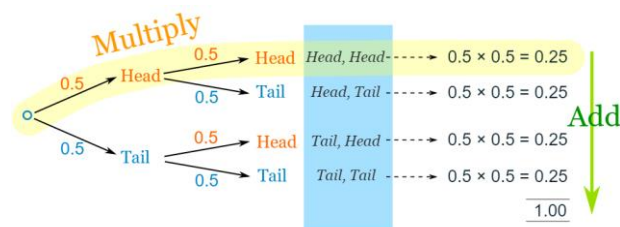
$$P(\text{at least 1}) = 1 - P(\text{none})$$

Tree Diagrams

Each branch of a tree diagram represents an outcome.

The probabilities on a set of branches add up to 1.

Always **multiply along the branches** to get the end results.



Add probabilities down columns (after multiplying).

Replacement: The denominator stays the same on the

second set of branches. The question will indicate that the item has been replaced or put back

Without replacement: This is not clearly stated in the question. Usually uses words such as 'takes another', 'takes two' – doesn't mention anything about replacing the item. The denominator on the second set of branches will be 1 less than that of the first set.

Conditional Probability

Opposite of independent events.

When one event affects the chances of another event happening.

Example: If there are 2 green and 4 white balls in a bag and you take a white ball the first time and don't put it back, this changes the probability of taking a green or white ball the second time. $P(\text{white first time}) = 4/6$, $P(\text{white second time}) = 1/5$, $P(\text{green second time}) = 4/5$. So the chances of selecting a specific colour ball the second time depends on which colour was chosen the first time as choosing white first time increases the chances of selecting green the second time.

Notation:

$P(B|A) = P(B \text{ given that } A \text{ happens})$. The event that happens first comes last in the bracket.

How to know it is conditional probability?

Phrases like '**given that**', '**if**' or the questions starts by telling you about one group and asks you to work out the probability of a second event **from 'that'/'this'** group.

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

$$P(A \text{ and } B) = P(B|A) \times P(A)$$

For two independent events, A and B $P(A) = P(A|B)$.

This formula can be used to test if 2 events are independent. If $P(A)$ and $P(A|B)$ are not equal, the events are not independent but are conditional.

Chapter 7 - Index Numbers

Calculating Index Numbers

Used to compare price changes over time.

Simple Index numbers compare the price change of an item with its base year price.

The index number for each year is that year's price as a percentage of its base year price.

The base year price has index number of 100 (meaning 100%) – this is the original amount.

$$\text{Index Number} = \frac{\text{Price}}{\text{Base Year Price}} \times 100$$

Index numbers show rates of change:

- If the index number is >100, the value has increased.
- If the index number is <100, the value has decreased.

RPI, CPI and GDP

Retail Price Index (RPI) – Shows rate of change (inflation/deflation) of prices of everyday goods, such as mortgage, food and heating.

RPI is calculated monthly by comparing prices to the same month of the previous year – this is because there can be seasonal variations.

Consumer Price Index (CPI) – Official measure of inflation used by the UK Government.

It is similar to RPI but does not include mortgage payments.

Pensions and benefits in the UK are updated each year in line with the CPI.

CPI is weighted to reflect importance of different items the average shopping basket. The weightings change each year to reflect consumer spending.

Gross Domestic Product (GDP) – Value of goods and services produced in a country in a given amount of time.

If the GDP falls in two (or more) successive quarters the economy is in recession.

Weighted Index Numbers – Takes into account proportions (similar to weighted mean).

Weightings reflect the importance of different items.

$$\text{Weighted Index Number} = \frac{\sum(\text{index number} \times \text{weight})}{\sum \text{weights}}$$

Chain Base Index Numbers

Compares prices from each year with that of the previous year.

Show how values change from year to year.

The previous year is used as the base year – this means the base year will change every time.

$$\text{Chain Base Index Numbers} = \frac{\text{price}}{\text{last year's price}} \times 100$$

RPI and CPI are chain base index numbers and show how values change monthly or annually.

CPI is published monthly by ONS.

Rates of Change

Crude Rates tell you how things change in every 1000 – usually births, deaths, marriages or unemployment.

They need to be recorded to make plans for the future e.g. high birth rate means more schools will be required.

Crude Rate – how many times a particular event occurs per 1000 of the population in a given time.

Crude Birth Rate – Number of births per 1000 of the population.

Crude Death Rate – Number of deaths per 1000 of the population.

$$\text{Crude Rate} = \frac{\text{number of births/deaths}}{\text{total population}} \times 1000$$

Crude rates can be misleading when used for comparing against another area which has different distribution of ages.

Standard Populations – represent the whole population. It is a hypothetical population of 1000 people used to represent the whole population, taking into account the number of people with different age/gender/income.

$$\text{Standard Population} = \frac{\text{number in age group}}{\text{total population}} \times 1000$$

Standardised Rate – Allows you to compare the same age group in different populations by using the standard population – allows for more realistic comparisons.

$$\text{Standardised Rate} = \frac{\text{Crude Rate}}{1000} \times \text{Standard Population}$$

To find the standardised rate for the entire population, add up the rates for each group.

Chapter 8 - Probability Distributions

A probability distribution is a **list of all the possible outcomes together with their expected probabilities.**

Example:

When flipping a fair coin, the probability distribution, for x outcomes, would be:

X	Heads	Tails
P(x)	$\frac{1}{2}$	$\frac{1}{2}$

Binomial Distributions

A type of probability distribution where there are **only two possible outcomes.**

Examples:

Event = flipping a coin, Outcomes = heads or tails

Event = Rolling a six on a dice, Outcomes = Success (if it lands on 6) or Failure (if it does not and on 6).

Notation: It is written as **B (n, p)** where **n=number of trials** and **p=probability of success.**

Conditions for Binomial Distribution:

1. **Fixed** number of trials (n)
2. Each trial has **2 outcomes**, success (p) or failure (q) e.g. rolling a six on a dice or not 6.
3. All the trials are **independent** of each other – the outcome of one doesn't affect the others.
4. **Probability** of success is **constant** – it stays the same for every trial.

If these conditions are met, an event can be modelled using the binomial distribution.

These conditions can also be used to explain if the binomial distribution is a suitable model – In an exam question you would need to show if the event described in the question meets each of these conditions. If all 4 are met then the binomial distribution can be used otherwise not.

Finding Probabilities using the Binomial Distribution: Use $(p + q)^n$ to find the probabilities.

1. **Identify the 2 outcomes** and their probabilities
2. **Expand $(p + q)^n$** where n is the number of trials. Leave p and q as letters for now.
3. To find the probability of x successes, **find the term that has p to the power of x successes.** E.g. for 3 successes find the term that has p^3 .
4. **Substitute** the values of p and q (the probabilities of success and failure) into that term and **calculate e.g.** for 5 trials and 3 successes you would use the $10p^3q^2$ term. For the event rolling a six on a dice $p=1/6$ and $q=5/6$. So your probability of landing on 6 3 times would be $10 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^2$.

Finding the Probabilities/Coefficients:

1. **Pascal's Triangle** – The coefficients of a binomial distribution follow the pattern of Pascal's triangle.

It starts with 1 in row 0 and has 1s down both sides. The other numbers are found by adding the 2 numbers directly above.

Memorise this triangle or how to work out the values so you don't have to expand $(p + q)^n$

Example: For $(p + q)^4$, the expansion would be:

$$1p^4 + 4p^3q + 6p^2q^2 + 4p^1q^3 + 1q^4$$

For each term the power of p decreases by 1 and the power of q increases by 1, starting with p=n and q=0 until you reach p=0 and q=n.

1							
1	1						
1	2	1					
1	3	3	1				
1	4	6	4	1			
1	5	10	10	5	1		
1	6	15	20	15	6	1	
1	7	21	35	35	21	7	1

2. **nCr button on calculator** – This is on top of the ÷ button.

N=number of trials and r=number of successes.

Example: For 5 trials with 3 successes, type '5', 'nCr', '3', '=' and you will get 10 which is the coefficient for the probability you would use – $10p^3q^2$ (3 success and 2 failures) and then you just need to substitute your probabilities of p and q into $10p^3q^2$ and calculate.

To find a range of probabilities, work out their individual probabilities and then add them up.

Example: P (3 or more successes for 5 events), work out the probabilities for 3, 4 and 5 successes using the above methods and add up the answers.

For questions that ask for the probability of 'at least 1 success' work out the probability of 0 successes and subtract the answer from 1 rather than working out all the individual probabilities.

The mean (or expected value) of the binomial distribution, **B (n, p) is np.**

Example: for B (6, ½) the mean is $6 \times \frac{1}{2} = 3$. This is the expected number of times the event would happen for n trials. In the above example if you flipped a fair coin 5 times you would expect it to land on tails an average of 3 times.

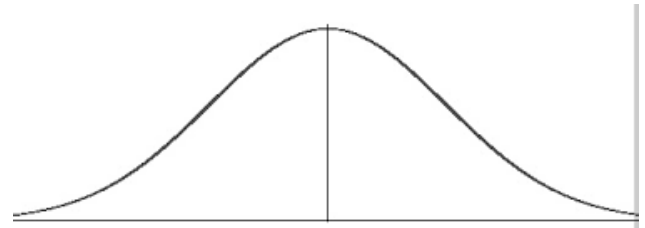
Normal Distributions

Drawn as a **smooth, bell-shaped curve**.

It is a common model for real-life situations such as weights of apple or marks in an exam.

Most of the data is in the middle with similar values and a fewer on either end.

A **larger standard deviation (SD)** will result in a lower curve and **smaller SD** will give a curve with a higher maximum height.



Notation: $N(\mu, \sigma^2)$ where μ = mean and σ^2 = variance (the square of standard deviation - σ =standard deviation, SD).

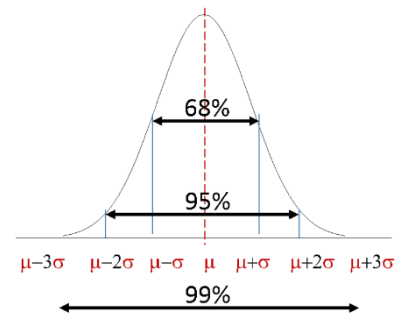
Conditions for Normal Distribution:

1. The data is **continuous** (heights, weights, time)
2. The distribution is symmetrical – there is a peak in the middle at the mean.
3. **Mode, median and mean are all approximately equal.**

Normal distribution is not suitable for data that is skewed (positively skewed or negatively skewed).

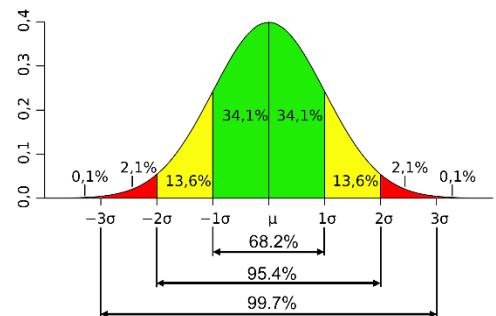
Important properties of a Normal Distribution:

- Approximately **68%** of data values lie within **1 standard deviation of the mean ($\mu \pm \sigma$)**
- Approximately **95%** of data values lie within **2 standard deviations of the mean ($\mu \pm 2\sigma$)**
- Approximately **99.8%** (you may see this as **99.7%** in some places – both are correct) of data values lie within **3 standard deviations of the mean ($\mu \pm 3\sigma$)**



For each property half the area lies either side of the mean.

- For 1 SD, 34% lies between μ and $\mu + \sigma$ and 34% between μ and $\mu - \sigma$
- For 2 SD, 47.5% lies between μ and $\mu + 2\sigma$ and 34% between μ and $\mu - 2\sigma$
- For 3 SD, 49.9% lies between μ and $\mu + 3\sigma$ and 34% between μ and $\mu - 3\sigma$



Example: For a data set with mean=30 and SD=3,

68% of the sample lie within 1SD so 30 ± 3 which is in the range 27-33

95% of the sample lie within 2SD so $30 \pm 2(3)$ which is in the range 24-36

99.8% of the sample lie within 3SD so $30 \pm 3(3)$ which is in the range 21-39

Sketching a Normal Distribution:

1. **Work out 3 SDs** either side of the mean
2. **Draw your x axis** stretching 3 SD either side of the mean.
3. **Sketch a bell-shaped curve** centred on the mean and ending at 3 SD from the mean.
4. If you are sketching more than one normal distribution curve on the graph the curves may not all be the same height. The larger the SD the lower the peak of the curve.

Calculating number of SDs: *Number of SD from mean* = $\frac{\text{value} - \text{mean}}{\text{standard deviation}}$

This is useful when you need to use standard deviations and means to work out the probability of an event.

Example:

For data set with mean=1000 and SD=15, you can calculate the probability of the data being between 960 and 1030.

$$\frac{960-1000}{15} = -2 \quad \frac{1030-1000}{15} = 2$$

So the data lies within 2 SDs of the mean which means the probability is 95%.

Standardised Scores

Used to compare 2 samples of data to see how far above or below the average individual values are.

Standardised scores tells you how many standard deviations away from the mean the data values are.

$$\text{Standardised Score} = \frac{\text{Score} - \text{Mean}}{\text{Standard Deviation}}$$

- **Positive score** means the value is above the mean.
- **Negative score** means the value is below the mean.
- **Score of zero** means the value is equal to the mean.

It is useful to compare individuals' performance in exams against the whole class for 2 different subjects.

Example:

English: Mean=60 SD=5 Mark=54

Maths: Mean=70 SD=8 Mark=65

$$SS (\text{English}) = (54-60)/5 = -1.2$$

$$SS (\text{Maths}) = (65-70)/8 = -0.675$$

This person did better in Maths in terms of actual mark but also compared to the rest of the class because they have a better standardised score.

Quality Assurance

Involves checking samples to make sure products are all of the same quality and standard.

It is about ensuring samples selected for checking quality are as close as possible to the target value so that products are all of a similar quality.

How it works:

1. Regular samples taken (the sampling technique used by manufacturer's will vary)
2. Sample mean, median and range calculated.
3. These are plotted on control charts to see how far they are from the value you'd expect them to be if the manufacturing process was working correctly.

Control Chart – A time series chart used for **quality assurance**. It has 5 lines:

- **Target Value** – this is the **middle line** – you want your sample values that you plot to be close to this line.
- **Upper and Lower Warning Lines (Inner 2 lines)** – These are **2 SD above and below the target value**. 2SD=95% so only 5% of sample averages or range should fall outside these lines. If a sample average/range plotted is above/below warning line another sample is taken and checked to see if there is a problem and production stopped if there is.
- **Upper and Lower Action Limits (Outer 2 lines)** – These are **3SD above and below the target value**. Almost all of the sample average/range should fall within these lines. If a sample average/range is outside of these lines production is stopped immediately and machinery is reset.

