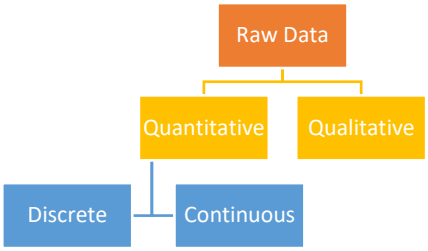
Chapter 1: Collection of Data

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 quantitative 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

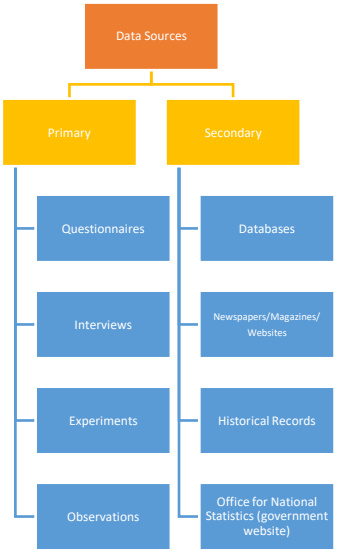
Grouping data using tables makes it easier to spot patterns in the data and quickly see how the data is distributed.

- 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 ≤.
- 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	• Accurate • Collection method known • Can find answers to specific questions	• Time consuming • Expensive
Secondary Data	• Cheap • Easy • Quick • Data from some organisations can be more reliable than data collected yourself	• 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



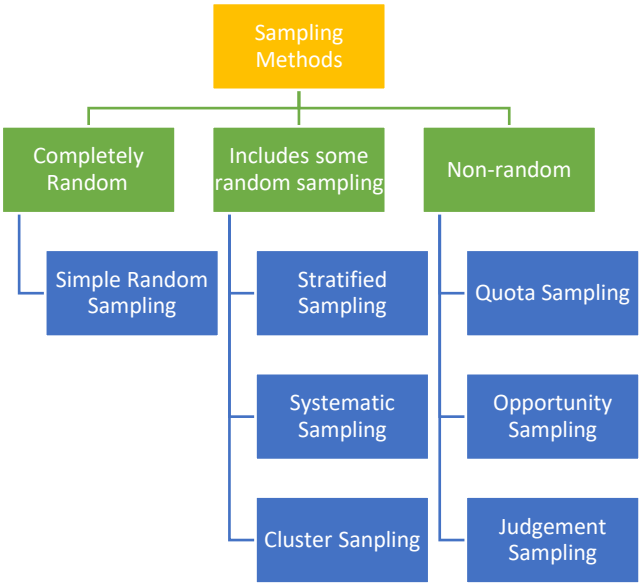
Populations and Sampling:

- **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.

	Advantages	Disadvantages
Census	• Unbiased • Accurate • Takes into account entire population	• Time consuming • Expensive • Lots of data to manage • Difficult to ensure whole population is used
Sample	• Cheaper • Quicker • Less data to consider	• May be biased • Not completely representative

- **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.

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 $\boldsymbol{stratified\ sample} = \frac{strata}{total} \times sample\ size$ 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 10^{th} 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 10^{th} item after the 7^{th} e.g. 7^{th}, 17^{th}, 27^{th}, …, 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.

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

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

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.

<u>Simulation</u> – 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(milk) = 3/6, P(white) = 1/6, P(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. 3/6 of 6 is 3 so assign numbers 1, 2, 3 on the dice to milk chocolate. 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	• Interviewer can explain questions • Interviewer can put people at ease when having to answer personal qs • Respondents can explain their answers • High response rate	• 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	• Respondents more likely to answer personal questions • No interviewer bias • Easy to send questionnaires to large sample size • Quick • Cheap	• 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.

  <u>Hypotheses and Investigations:</u>
  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.