	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from some organisations can be more reliable than data collected yourself	• Method of collection unknown • Data may be out of date • May contain mistakes • May come from unreliable source • May be difficult to find answers to specific questions

	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

Types of Data

- **Quantitative** – numerical observations or measurements.
- **Qualitative** – non-numerical observations
- **Continuous** – can take any value on a continuous numerical scale
- **Discrete** – can only take particular values
- **Categorical** – can be sorted into non-overlapping categories e.g. gender.
- **Ordinal** – can be written in order of be given a rating scale.
- **Bivariate** – involves a pair of related data.
- **Multivariate** – involves sets of 3 or more related data values.
- **Primary** data – collected by, or for, the person using it.
- **Secondary** data – has already been collected by someone else.

Sampling

- **Population** – everything or everybody that could possibly be involved in an investigation.
- **Census** – a survey of a whole population
- **Sample** – a smaller number of items from the population
- **Biased Sample** – not representative of everyone in the population
- **Sampling frame** – a list of people/items that are to be sampled.

Collecting Data

- **Questionnaire** – a set of questions designed to obtain data.
  - Features of a good questionnaire: short questions, simple language, no leading questions, non-overlapping boxes, time frame in question, option boxes are exhaustive, no personal questions
- **Open question** – has no suggested answers
- **Closed question** – has a set of given answers to choose from
- **Pilot survey** – a small scale version of the survey to test the design and methods of that survey.
- **Hypothesis** – a statement made as a starting point of an investigation
- **Cleaned data** – made by identifying and assessing extreme values, missing data and errors before it is used.
- **Extraneous variables** – variables you are not interested in but could affect the result of your experiment
- **Control group** – used to test the effectiveness of a treatment.
  - Use random selection to select 2 groups of people
  - Give the test group the treatment, control group no treatment
  - Compare results from 2 groups to see how effective treatment is
- **Matched Pairs Test** – 2 groups of equally matched (age/gender etc.) people used to test effect of a particular factor. Everything in common except factor being studied.
- **Simulations** – model random real life events, to help you predict what could actually happen. May be easier and cheaper than collecting real life data.

Petersen Capture-Recapture Formula

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

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.

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

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.

Random Response Method

Uses a random event to decide how to answer a question

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

Types of Data

Discrete Data = Raw value, grouped data with no inequalities, Cumulative Frequency Step polygons
Continuous data = grouped data with inequalities, CF curves, histograms

Systematic Sampling

To find nth interval = $\frac{Population\ size}{Sample\ size}$

Stratified Sampling

Members of each stratum (group) are in **proportion** to the size of the stratum. Sample from each strata is selected using random sampling.

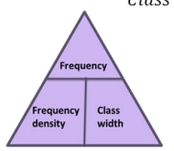
$$Number\ selected\ from\ strata = \left(\frac{strata\ size}{total\ population}\right) \times 100$$

Random Sampling

Every member of the population has an **equal chance** of being selected – representative of population.
- Number everyone in population
- Use random number generator to select enough for sample size.
- Match numbers to individuals.
- Ignore repeats/numbers outside range

Histograms

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



Cumulative Frequency

- Running total of the frequencies of each class interval
- **For discrete data**, use a **step polygon** – plot the points and join by going across and then up.
- CF Curves – used for continuous data
- CF Step polygons – for discrete data.
- Plot points at UPPER BOUND of class interval.

Comparative Pie Charts

Used to compare 2 sets of data with different frequencies
Area of 2 circles should be in the same ratio as the two total frequencies.
To compare **total frequencies**, compare the **areas**.
To compare **proportions**, compare individual **angles**.
Larger pie chart = larger frequency

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

Type of Experiment	Description	Variables	Advantages	Disadvantages
Laboratory Experiments	Experiments conducted in a **controlled environment**. Example: to investigate effect of colour on taste perception people are given apple juice coloured red, green and without colouring.	Explanatory Variable: Colour of juice. Researcher changes this and sees what effect his has on taste perception. Response Variable: The flavours people identify.	Easy to replicate because you can copy the experiment exactly. Extraneous variables can be controlled, such as putting all the juices in identical cups.	Test subjects may behave differently under test conditions than in real life.
Field Experiments	Experiments carried out in test subjects' **everyday environment**. Researcher sets up the situation and **varies one or more variables**. Example: testing a new method of teaching timetables by giving students a test, teaching new method and retesting	Explanatory Variable: New teaching method. Response Variable: Marks in test	More likely to reflect real life behaviour.	Cannot control extraneous variables e.g. some teachers may be better at motivating students than others. Difficult to replicate exactly.
Natural Experiments	Experiments carried out in test subjects' **everyday environment** but researcher has **no control over any variables**. Example: Looking at effect of level of education on income.	Explanatory Variable: Level of education Response Variable: Income	More likely to reflect real life behaviour.	Cannot control variables so difficult to replicate.

	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

GCSE STATISTICS

Method	Description	Example	Advantages	Disadvantages
Random Sampling	Every member of the population has an equal chance of being selected.	Pulling names from a hat, using random number tables or generators on a calculator.	Sample is more likely to be representative of the population, provided it is large. Choice of members of sample is unbiased.	Needs a full list of the whole population. Needs a large sample size. Can be expensive and time consuming.
Judgement Sampling	Use your judgement to select a sample that is representative of the population	In test marketing, a judgement is made as to which cities would constitute the best ones for testing the marketability of a new product.	Easy way of selecting a sample.	The quality of the sample depends on the person selecting the sample.
Opportunity Sampling	Use the people (or objects) that are available at the time.	Going in the street and interviewing the first 10 people you see	Quick. Convenient.	Samples are unrepresentative of the population.
Cluster Sampling	The population is divided into groups (clusters) and then a group (or groups) are chosen at random.	For a population of all secondary school students, clusters can be secondary schools. A random sample of schools are selected, and all their students are sampled.	Economically efficient – less resources required. Can be representative of large numbers of small clusters used.	Clusters may not be representative of the whole population. Higher sampling error
Systematic Sampling	Choose a starting point from the sampling frame at random, and then choose items at regular intervals.	Every 5th person on a list, or every 10th item from a production line.	Sample easy to select. Population is evenly sampled.	Not strictly a random sample as some member of the population cannot be chosen.
Quota Sampling	Group the population by characteristics and interview a quota (number) from each group.	100 males under 20, 100 females under 20, 300 males 20-50 etc.	Quick and cheap to carry out	People are chosen to take part in a non-random way, so the sample is unrepresentative.
Stratified Sampling	Contains member of each strata (group) in proportion to the size of the that strata. The sample from each stratum is selected randomly	For a class with twice as many boys as girls, a stratified sample would include twice as many boys as girls.	Representative of the population – less bias. Best used when population has unequal groups.	Time consuming.

Scatter Diagrams

- Use for **bivariate** data.
- X- axis = Explanatory (independent) variable.
- Y - axis = Response (dependent) variable.
- **Correlation** – an association between 2 variables. As one variable increases, the other variable increases or decreases.
- **Causal Relationship** – When a change in one variable **directly** causes a change in another variable.
- Correlation does not necessarily imply causation. In real life situations, multiple factors interact to cause variables to change.
- **Line of Best Fit** (LOBF): To get a good fit, draw your line through the **mean point** $(\bar{x}, \bar{y})$.
- **Interpolation** – using your LOBF to estimate **within** the range of values already plotted. Usually reliable.
- **Extrapolation** – Extending LOBF and reading values **outside** of the range of values plotted. Less reliable.
- **Equation of LOBF:**
  - $y = ax + b$ a = gradient, b = y-intercept
  - **LOBF** is called **regression line.**
  - $a = \frac{y_2 - y_1}{x_2 - x_1}$ $b = y_1 - ax_1$

SRCC/PMCC

SRCC – measures the strength of the correlation between two variables.
SRCC is ALWAYS between -1 and 1.
- If r = 1 there is strong +ve correlation
- If r = 0 there is no correlation
- If r = -1 there is strong -ve correlation
- $SRCC = 1 - \frac{6\sum d^2}{n(n^2-1)}$ d = diff between ranks
- **SRCC** is best used for data that can be ranked and data that is **non-linear**.

PMCC – tests for **linear correlation**.

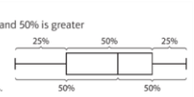
Measures of Dispersion/Spread

Range:
Largest Value - smallest value

Interquartile range (IQR):

IQR = UQ – LQ
- Lower Quartile (LQ) – $\frac{1}{4}$ = $\frac{1}{4}$ (n+1)th value (disc) or $\frac{1}{4}$nth value (cont.)
- Upper Quartile (UQ) – $\frac{3}{4}$ = $\frac{3}{4}$ (n+1)th value (disc) or $\frac{3}{4}$ nth value (cont.)

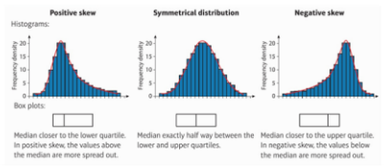
In a distribution:
- 50% of the data in a distribution is less than the median, and 50% is greater than the median.
- 25% of the data is less than the lower quartile
- 25% of the data is greater than the upper quartile
- 50% of the data is between the lower and upper quartiles.



- Outliers:
  < LQ - (1.5 x IQR) and > UQ + (1.5 x IQR)
- Interpercentile Range – difference between 2 percentiles
- Interdecile range – difference between two deciles
- Standard Deviation – A measure of how far all the values are from the mean value, or how spread out they are.

  Discrete: $\sigma = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$ OR $\sigma = \sqrt{\frac{\sum x^2}{n} - (\frac{\sum x}{n})^2}$

  Grouped Data: $\sigma = \sqrt{\frac{\sum f(x-\bar{x})^2}{\sum f}}$ OR $\sigma = \sqrt{\frac{\sum fx^2}{\sum f} - (\frac{\sum x}{\sum f})^2}$
- Skewness:
  - Skew = $\frac{3(mean-median)}{mode}$
  - Positive Skew = Mean > median > mode
  - Negative Skew = Mean < median < mode



Measures of Central Tendency/Averages

Mode – most common value
Modal Class – class with the highest frequency

Median:
Discrete data – $\frac{1}{2}$(n + 1)th value
Continuous data (with inequalities) – $\frac{1}{2}$nth value
To estimate median from grouped data:
- Find cumulative frequency of freq column until you get to $\frac{1}{2}$nth value and find the class interval with the median.
- See how many more values you need in that class to get the median.
- Divide this number by the freq for that class.
- Multiply your answer by the class width.
- Add your answer to the lower bound for the class interval.

Estimated median = $L + \frac{\frac{n}{2} - F}{f} \times w$, where:
- L is the lower boundary of the class containing the median
- n is the total number of values
- F is the cumulative frequency of the intervals before the one containing the median
- f is the frequency of the median class interval
- w is the width of the median class interval.

Mean ($\bar{x}$):
Discrete data - $\bar{x} = \frac{\sum x}{n}$ $\sum x$ = sum of all values
n = number of values

Continuous data - $\bar{x} = \frac{\sum fx}{\sum f}$ f = frequency
x = first column (use midpoints for grouped data)

Weighted Mean – for data that has different number of values or weights for each group.

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

Geometric Mean = $\sqrt[n]{value1 \times value\ 2 \times \cdots \times valuen}$

	Advantages	Disadvantages
Mode	• Easy to find • Can be used with quantitative and qualitative data • Unaffected by open ended or extreme values • Always a value in the data	• May be no mode or more than one mode • Cannot be used to calculate measures of spread
Median	• Easy to calculate • Unaffected by outliers • Best to use with skewed data • Can be used to calculate quartiles, IQR and skew.	• May not be a data value
Mean	• Uses all the data • Can be used to calculate standard deviation and skew.	• May not be a data value • Always affected by extreme values • Can be distorted by open-ended classes

Comparing Data Sets

- Compare using a measure of average (mean/median/mode), measure of spread (range, IQR, standard deviation) or skewness.
- Always make reference to individual values and mention which data set is larger/smaller than which one clearly.
- Always interpret in context (link back to scenario in question).

- Averages:
  Mean/median/mode for data A is larger than mean/median/mode for data B so on average data A is more….than data B.
- Spread:
  Range/IQR/SD for data A is larger than that of data B so the 'results' of data A are more spread out/less consistent than those of data B.
  Data A has a smaller range/IQR/SD than data B which means the 'results' for data A are more consistent.
  Lower SD means values are closer to mean.
- Skewness:
  Box plot for data A is positively skewed o majority of 'results' were low with fewer higher 'results'.
  Box plot for data A is negatively skewed so majority of 'results' were high with few lower results.

Time Series

- A line graph with time plotted on the x-axis.
- **Trend line** shows the general trend of the data – ignore fluctuations and just follow the general pattern. Place the line roughly halfway between highest and lowest point for each year.
- Trend line may show rising (upwards), falling (downward) or level trend.
- **Seasonal variations** – variations in a time series following a regular time period, like days of the week or seasons. Think about real life scenarios that may cause these.

- **Moving Averages** – An average worked out for a given number of successive observations. They smooth out fluctuations in the data and make the trend line more accurate.
- Plot them at the midpoint of the time interval and do not join them up – use a LOBF.

- **Seasonal Variation = Actual Value – Trend Value**
- **Estimated mean SV (EMSV) = Mean of all the SV for that season**
  (e.g. average of all quarter 4 SVs.), Also called average seasonal effect.
- **Predicted Value = Trend line value + EMSV**
- Reliability of prediction depends on how far into the future prediction is made (further = less reliable) and how good the EMSV as trends and variations can unexpectedly change.

Probability

- $\boldsymbol{Prob.of\ an\ Event, P(event) = \frac{Number\ of\ successful\ outcomes}{Total\ number\ of\ possible\ outcomes}}$
- Outcome – something that can happen as a result of a trial e.g. heads or tails when flipping a coin.
- **Expected Frequency** – The number of times you expect the event to happen, not necessarily what will happen.
- $\boldsymbol{Expected\ Frequency = P(outcome) \times number\ of\ trials}$
- $\boldsymbol{Estimated\ Probability = \frac{no.of\ trials\ with\ successful\ outcomes}{total\ number\ of\ trials}}$
- As the number of trials increase the estimate for the probability gets closer to the true value.
- Estimated prob is also called relative frequency.

- Risk – probability of an event occurring for negative events.
- $\boldsymbol{Risk\ of\ event = \frac{No.of\ trials\ in\ which\ event\ happens}{total\ no\ of\ trials}}$
- **Absolute Risk** – how likely an event is to happen.
- **Relative risk** – how much more likely an event is to happen for one group compared to another group.
- $\boldsymbol{Relative\ Risk = \frac{Risk\ for\ those\ in\ the\ group}{Risk\ for\ those\ not\ in\ group}}$

- **Mutually Exclusive** – CANNOT happen at same time
- **Exhaustive** – The set contains ALL the possible outcomes
- For a set of mutually exclusive and exhaustive events the sum of probabilities is equal to 1.
- Addition Law for M.E. events:
  - P(A or B) = P(A) + P(B)
  - P(A) + P(not A) = 1
  - P(not A) = 1 – P(A)
- General Addition Law (for non-M.E. events – can occur together)
  - P(A or B) = P(A) + P(B) – P(A and B)
- $P(A \cap B)$ = P(A and B). On a Venn diagram this is the intersection or middle/overlapping part.
- $P(A \cup B)$ = P(A or B). On a Venn diagram this is the union of A and B and includes everything in both circles.

- Independent Events – unconnected events. The outcome of one event does not affect the outcome of the other event.
- For two independent events, A and B: P(A and B) = P(A) x P(B).
- For 3 independent events, P(A and B and C) = P)A) X P(B) X P(C)

- P(at least 1) = 1 – P(none)

- **Conditional Probability** – opposite of independent events, when an event affects another.
- P(B|A) = P(B given that A happens). The event that happened first comes last in the bracket.
- How to know its conditional probability? Phrases like 'given that', 'if' or tells you about one group and asks you to work out the prob of second event from 'that'/'this' group.
- $\boldsymbol{P(B|A) = \frac{P(A\ and\ B)}{P(B)}}$ $\qquad$ $\boldsymbol{P(A\ and\ B) = P(B|A) \times P(A)}$
- For two independent events A and , P(A) = P(A|B). You can use this formula to test if two events are independent. If P(A) and P(A|B) are not equal, the events are not independent and are instead conditional.

GCSE STATISTICS
Revision Notes
Ms Patel

Probability Distributions

- A list of all the possible outcomes together with their probabilities.

- Binomial Distribution: B(*n*, *p*)
- n = no. of trials, p = probability of success.
- Conditions: (use these to explain if binomial dist. is a suitable model)
  1. Fixed number of trials (n)
  2. Each trial has two outcomes (success (p)/failure(q)). E.g. on a dice, 6 or not 6.
  3. All the trials are independent of each other,
  4. Probability of success is constant (stays the same for each trial).
- Use $(p+q)^n$ to find probabilities:
  1. Identify the 2 outcomes and find their probabilities.
  2. Expand $(p+q)^n$ where n is number of trials. Leave p and q as letters for now.
  3. To find prob of x successes, find the term that has p to the power of x successes.
  4. Substitute values of p and q into that term and calculate.
- Mean of Binomial Distribution B(n,p) is *np*.
- You can find coefficients of binomial expansion on calculator using nCr button (on top of ÷) for n trials and r successes.

- Normal Distribution: N(μ, σ^2)
- μ = mean, σ^2 = variance
- Smooth bell shaped curve.
- Conditions:
  1. The data is continuous (heights, weights, time)
  2. The distribution is symmetrical
  3. Mode, median and mean are approximately equal.
- Not suitable for skewed data.
- 68% of data lie within 1 SD of the mean.
- 95% of data lies with 2 SD of the mean.
- 99.8% of data lies within 3 SD of the mean.

- Standardised Scores – used to compare how far above or below average individual values are.
- $\boldsymbol{Standardised\ Score = \frac{Score - Mean}{Standard\ Deviation}}$
- + = score > mean
- - = score < mean

- Quality Assurance – Involves checking sample to make sure products are of same quality.
- Control Chart – time series chart used for quality assurance. They have 5 lines:
  - Target value (middle line)
  - Upper and lower Warning Limits (inner two lines) – 2 SD above/below target value. If sample is above warning line, another sample is taken to see if there is a problem and if so production stopped.
  - Upper and lower Action Limits (outer two lines) – 3 SD above/below target value. If sample outside action limit production stopped immediately.

Index Numbers

- Compares the price change of an item with its base year price.
- It is a %.
- $\boldsymbol{Index\ Number = \frac{price}{base\ year\ price} \times 100}$
- Index number > 100 = increase in value
- Index number < 100 = decrease in value.

- **Retail Price Index (RPI)** – Shows rate of change (inflation/deflation) of prices of everyday goods such as mortgage, food and heating.
- **Consumer Price Index (CPI)** – official measure of inflation used by UK government. Similar to RPI but does not include mortgage payments.
- **Gross Domestic Product (GDP)** – Value of goods and services produced in a county in a given amount of time.
- If GDP falls for 2 or more quarters, economy is in recession.

- **Weighted Index Number** – take into account proportions (similar to weighted mean). Weightings reflect importance of different items.
- $\boldsymbol{Weighted\ index\ number = \frac{\sum(index\ number\ x\ weight)}{\sum weights}}$
- **Chain Base index Numbers** – compare prices from each year with previous year.
- $\boldsymbol{Chain\ Base\ I.N. = \frac{price}{last\ year's\ price} \times 100}$
- RPI/CPI are chain base that show annual or monthly % changes.

- **Crude Birth Rate** – Number of births per thousand of the population.
- **Crude Death Rate** – number of deaths per thousand of the population.
- $\boldsymbol{Crude\ Rate = \frac{number\ of\ births/deaths}{total\ population} \times 100}$
- **Standard Population** – hypothetical pop. Of 1000 people used to represent whole population.
- $\boldsymbol{Standard\ Population = \frac{number\ in\ age\ group}{total\ population} \times 1000}$
- **Standardised Rates** – allows you to compare same age group in different populations.
- $\boldsymbol{Standardised\ Rate = \frac{Crude\ Rate}{1000} \times standard\ population}$