DP IB Psychology: HL

Your notes

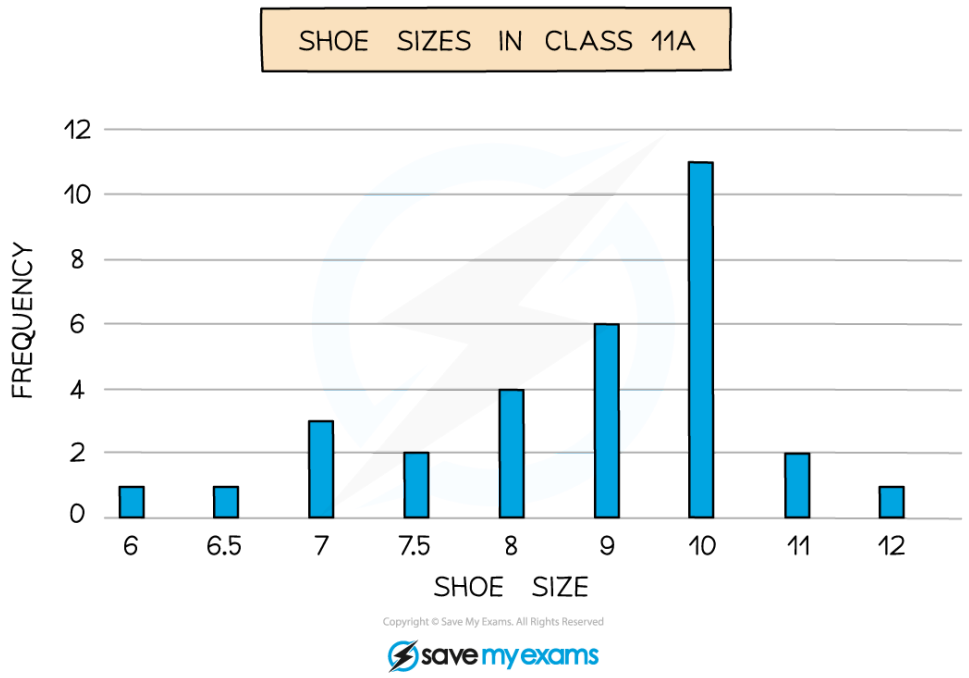
Data Analysis & Interpretation (HL Only)

Contents

Your notes

Bar graphs

- A type of **graphical** display can be achieved using a **bar graph**
- The data shown on the x-axis of a bar graph is **discrete** (not continuous)
  - E.g., scores on a memory test; number of 'yes' answers ticked on a **questionnaire**
- A bar graph uses **categorical** data which does not necessarily fall into any particular order
  - If a researcher had conducted an experiment with three conditions, they could use a bar graph to display the means of each condition
- Bar graphs **do have gaps** between each category on the x-axis (unlike histograms)
  - The x-axis shows the categories/conditions
  - The y-axis shows the score/percentage per category/condition

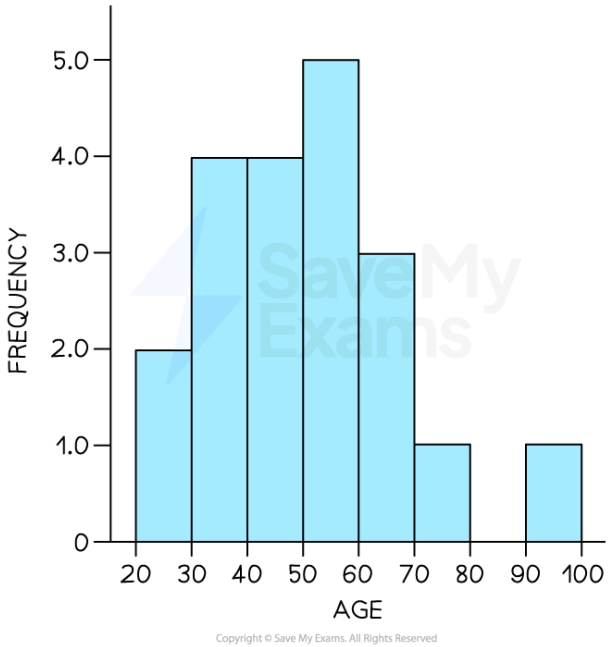


An example of a bar chart

Histograms

- On a histogram, the **x-axis** represents the **categories** that have been measured, e.g.,
  - the number of goals scored across one football season
  - the number of marks in a psychology mock exam across one year group

- On a histogram, the **y-axis** represents the **frequencies** of each category occurring, e.g.,
  - the frequency of the number of two goals scored in one match
  - the frequency of question 5 on the mock exam being awarded full marks
- A histogram thus, shows **continuous data**
  - Any category with zero frequency is represented by a space (a gap) in the chart
- Histograms do **not** have **gaps** between the bars; the bars touch each other



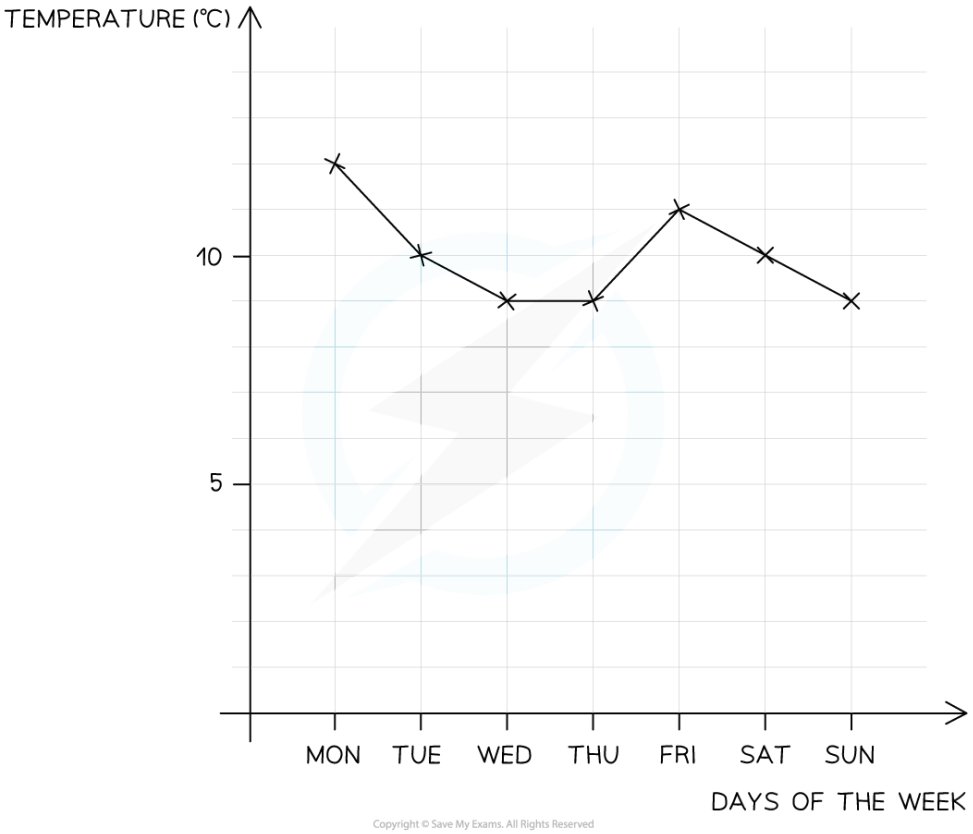
An example of a histogram

Line graphs

- A line **graph** shows how a quantity (**continuous** data) changes over time
  - E.g., How the outside temperature changes during a week (shown below)
    - This could be of interest to psychologists who wish to investigate the effect of temperature on behaviour
- **Measurements** of the quantity are taken at particular times
  - **Measurements** should be taken at **regular** time intervals
  - These are then **plotted** as points on a time series graph and **joined** together with **straight lines**
  - The straight lines help to **identify patterns** and **features** in the data
- Line graphs can show changes over **short** or **long periods** of **time**
  - E.g., Changes to the memory scores of participants 30 seconds after being shown a list of words

Your notes

- Or changes in memory scores of one group of people studied over several **years**
- Sometimes a line graph may have more than one data set
  - E.g., one line for temperature and one line for number of arrests made on that day

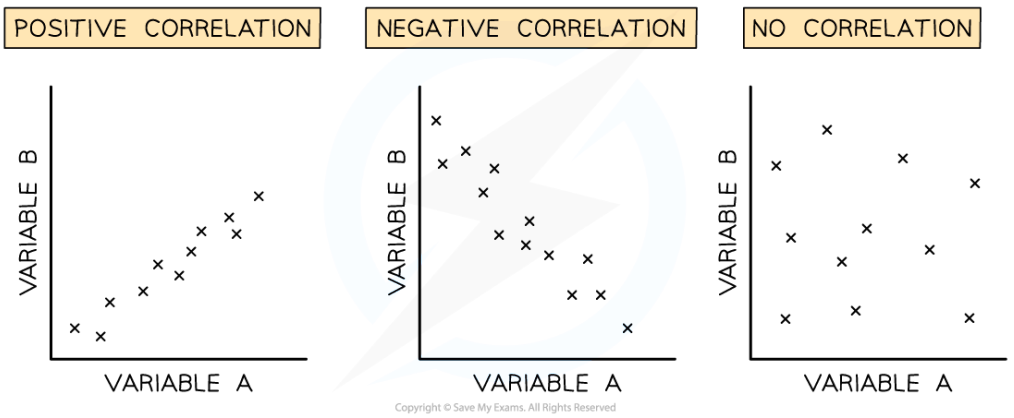


An example of a line graph

Your notes

Scatterplots

- Scatterplots are used to display the results of **correlations**
- A scatterplot shows the **point** at which **two separate pieces of data** meet
- Each **co-variable** can be presented along the **x-axis or the y-axis**
  - E.g., a strong **positive** correlation will be shown regardless of which axis is chosen per **co-variable**
- The arrangement of points on the scatterplot will indicate whether there is a positive correlation, a **negative** correlation or **no** correlation



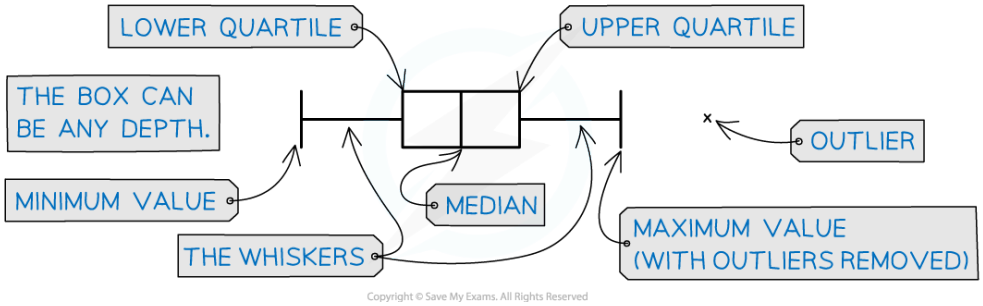
Scatterplots showing the different types of correlation

Your notes

Your notes

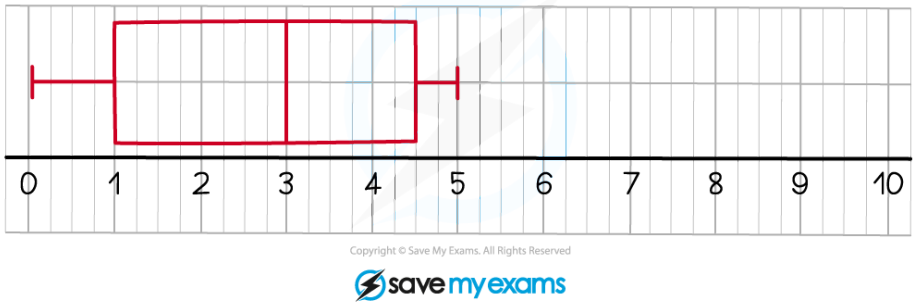
Box & whisker plots

- **Box** and **whisker plots** are used when researchers are interested in splitting data up into **quartiles**
- Often, data will contain **extreme values**
  - E.g., there are more people with an **IQ** of around 100 than there are people with an IQ of 130
    - If a researcher collects 50 IQ scores, 49 being from 'normal' people but one from a 'genius' (with an IQ of 130), then the 'genius' score would not fit in with the other data
- Using quartiles and drawing a box and whisker plot allows researchers to split the data so that they can see what is happening at the **low, middle and high points**
  - This allows them to also consider any possible extreme values
- Box and whisker plots include the following five values:
  - Lowest data value
  - Lower quartile
  - **Median**
  - Upper quartile
  - Highest data value
- On graph paper, box and whisker plots are drawn with the five values marked by **short vertical lines**
  - The middle three values then form a box with the **median line inside**
    - the median will not necessarily be in the middle of the box!
  - The box represents the **interquartile range** (middle 50% of the data)
  - The lowest data value and highest data value are joined to the box by horizontal lines:
    - these are the 'whiskers'
    - they represent the lowest 25% of the data and the highest 25% of the data



The key features of a box plot

- The box and whisker plot below shows the number of items recalled by a class of four-year-old children when tested once a week over a month:
  - The median is 3 items
  - The upper quartile is 4.5 items
  - The lower quartile is 1 item
  - The whiskers show the maximum value of 5 items and the minimum value of 0 items



A box and whisker plot

Your notes

Your notes

Frequency tables

- A **frequency table** measures the **number of times** a behaviour/action/phenomenon occurs
  - E.g., the number of times litter is dropped
  - the number of times red is chosen for a T-shirt design
  - the number of goals scored by players in one season
- To **organise** and **make sense** of **frequency data,** a researcher will arrange it into a **frequency table**:

Score: Number of goals scored in one match by school team in one season	Tally	Frequency
1	I	1
2	III	3
3	𝍸 𝍸 I	11
4	𝍸 𝍸 𝍸 𝍸	20
5	𝍸 III	8

- The frequency table above reveals that the **mode** for goals scored is 4, as this happened 20 times in one season
- The **median** score is 5, as this sits in the middle of the ordered data set, i.e., a frequency of 8
- The **mean** score is calculated as follows:
  - Multiply each score by its frequency e.g. 1 x 1, 2 x 3, etc.
  - Add the total of these scores; in this case, it is 1 + 6 + 33 + 80 + 40 = 160
  - Divide this score by the total number of goals in the first column (10) to find the mean; in this case, it is 10.6
- The **range** is calculated by subtracting the lowest score from the highest score; in this case, it is **5 – 1 = 4**

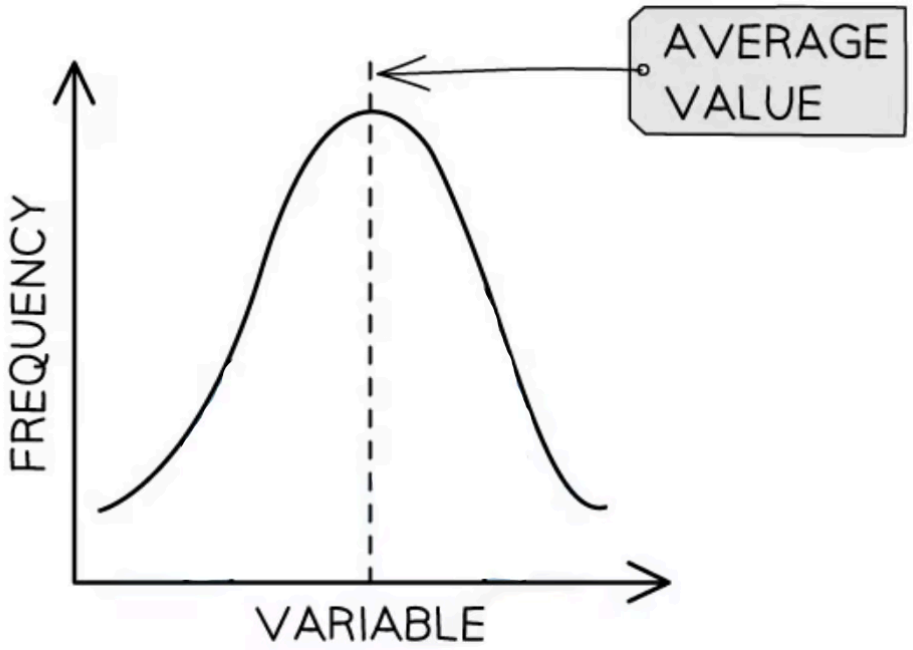
Your notes

Distributions

- Distribution in psychology refers to the **spread of data around the mean** for a specific **sample** or **population**
- Researchers in psychology are interested in the extent to which one data set **varies** from the mean
  - Do most scores **cluster** around the **mean**?
  - Are the scores spread **symmetrically?**
  - Are they **skewed**?

Normal distributions

- A **normal distribution** is **symmetrical** around the mean, with most scores being close to it, showing a **peak** in the middle where the **mean value** is located
- The **shape** of a normal distribution is known as the '**bell curve**', as the measurement outline looks like a bell
- Most scores (when the data is **normally distributed**) will fall within the central part of the bell curve
- Extreme **outliers** will fall within the '**tail ends**' of the curve (small scores on the extreme left and high scores on the extreme right of the curve)
- The tail ends **never actually touch the x-axis,** as there is **no assumption** as to there being one, final extreme high or low score
- Examples of data that is normally distributed are height, weight, shoe size
- The normal distribution can be used to test for signs of **deviance** from the **norm,** e.g.,
  - people who score beyond two **standard deviations** of the mean may rank as having an extremely high or low score, such as
    - **IQ**
    - scoring high on a scale to indicate **postpartum depression**
    - scoring low on an **empathy scale** which tests for **psychopathy**



The normal distribution (note the bell shape of the curve)

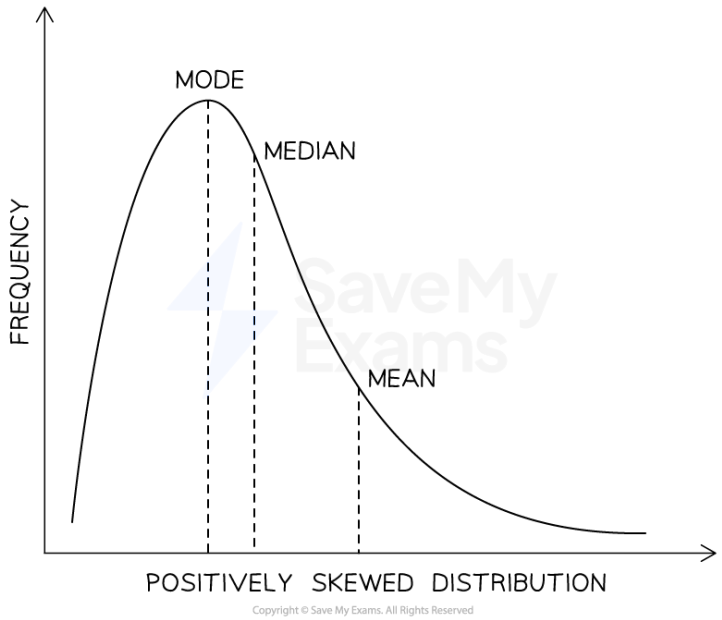
Your notes

Skewed distributions

- In a perfect normal distribution the mean, mode and median all appear at the **peak** of the curve, i.e., they have **similar values**
  - Scores to the **left** of the peak represent people who have scored **less than the mean**
  - Scores to the **right** of the peak represent people who have scored **more than the mean**
- There are some behaviours/conditions/test scores which do **not fit neatly** into a normal distribution; this represents **skewed distributions**
- A skewed distribution describes a graph curve where one tail is **longer** than the other
  - There is **asymmetry** in the graph curve; it is not bell-shaped
  - The two halves of the distribution **do not mirror each other** because the data is **not distributed equally** on both sides of the distribution's peak
  - The **mean** is the measure of **central tendency** which is most affected by skewed distributions, as it takes all scores in the data set into account

Positive skew

- A **positive skew** is one in which most of the values are found towards the **left side** of the graph, giving a long tail on the right



Positive skew

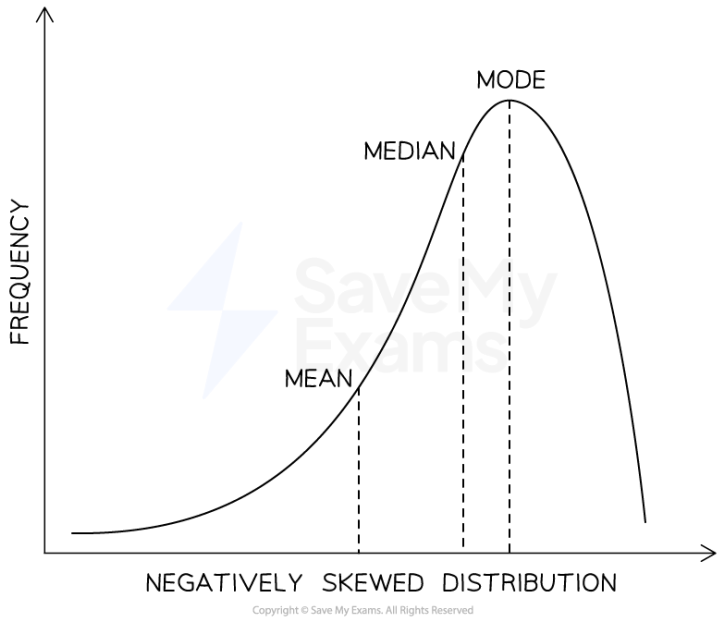
Examples of positively skewed data

- The age at which people take on their first job
  - In a **population** aged 16–80, most scores will likely be at the lower end of the measure
- A very difficult maths test might see most students score at the lower end of the mark scale
  - As the test is so difficult, there are very few in the class who scored towards the right end of the tail (where the high scores reside)

Negative skew

- A **negative skew** is one in which most of the values are found towards the **right side** of the graph, giving a long tail on the left

Your notes



Negative skew

Examples of negatively skewed data

- The age at which people retire
  - In a **population** aged 16–80 most scores will likely be at the higher end of the scale
- A very easy maths test might see most students score at the higher end of the mark scale
  - As the test is so easy, there are very few in the class who scored towards the left end of the tail (where the low scores reside)

Your notes

Your notes

Descriptive statistics & measures of central tendency

- **Descriptive statistics** include **measures of central tendency**, as they describe the central or **typical value of a data set**
- Measures of central tendency are used to **summarise** large amounts of data into typical mid-point scores

The mean

- This calculates the **average score** of a **data set**
- The mean indicates what a researcher would **expect to find** (as the average score) if they were to **replicate** the **procedure** of a given study
- The mean is **calculated** using the **total** score of all the **values** in the data set **divided by the number of values** in that set
- E.g., 4 + 6 + 7 + 9 = 26
  - 26 ÷ 4 = 6.5
    - **mean = 6.5**

The median

- This calculates the **middle value** of a data set (the **positional average**)
- The data has to be arranged into **numerical order** first (with the lowest score at the beginning of the list)
  - E.g., 20, 43, 56, 78, 92, 67, 48 is ordered into 20, 43, 48, **56**, 67, 78, 92
  - **Median = 56** as this is the value at the halfway point in the set
- Sometimes there may be two middle numbers in a set of data
  - E.g., 15, 16, **18**, **19**, 22, 24
  - The halfway point is between 18 and 19
  - In this case, **add the two middle values** (18 + 19 = 37)
  - **Divide the total by 2** (37 divided by 2 = 18.5)
  - Thus, the **median = 18.5**

The mode

- This calculates the **most frequently** occurring score in a data set
- Some data sets may have:

- **no mode**
- **two** modes (known as **bi-modal**)
- **more than two** modes (known as **multi-modal**)

- The mode is used when the researcher cannot use the mean or the median
  - E.g., a researcher wishes to measure **how many times** litter is dropped in a **naturalistic observation**
    - E.g., with a data set of 3, 3, 3, 4, 4, 5, **6, 6, 6, 6**, 7, 8, count the most frequently occurring number
    - Thus, the **mode = 6**

Your notes

Descriptive statistics & measures of dispersion

- **Measures of dispersion** calculate the **spread of scores** and how much they **vary** in terms of how **distant** they are from the **mean or median**
  - A data set with **low dispersion** will have scores that **cluster** around the measure of central tendency (the mean or median)
  - A data set with **high dispersion** will have scores that are **spread apart** from the central measure with **much variation** among them
  - If a data set contained exactly the **same score** per participant (e.g., everyone scored 15 out of 20 on a memory test), then the dispersion score would be **zero,** as there would be **no variation** at all in the scores (plus the mean, mode and median would be identical = 15)

The range

- This describes the **difference** between the **lowest and the highest scores** in a data set
- The range provides information as to the **gap** between the highest and lowest scores
- To calculate the range subtract the **lowest value from the highest value** in the data set, e.g.,
  - to calculate the range of 4, 4, 6, 7, 9, 9, subtract the lowest number (4) from the highest number (9)
  - The **range is 9**
  - When dealing with data that has been rounded, +1 is added to the data set to account for any **rounding up or down** which has been applied to the original scores
    - 9 - 4 = 5 + 1
    - Thus, **the range = 6**

Standard deviation

- This calculates how a set of scores **deviates** from the **mean**

- Standard deviation provides **insight** into how clustered or spread out the scores are from the mean
  - A **low standard deviation** indicates that the scores are clustered **tightly** around the mean, which indicates the **reliability** of the data set
  - A **high standard deviation** indicates that the scores are more **spread out** from the mean, which indicates **lower** reliability
- **Normal distributions** have a low standard deviation, as they reflect the fact that the scores are clustered close to the mean
- There are **six steps** to calculating the standard deviation
  1. Calculate the mean
  2. **Subtract** the mean from **each score** in the data set
  3. **Square** the scores which have just been calculated at step 2
  4. **Add** all of the **squared scores** together
  5. **Divide** the total squared score by the **number of scores minus 1**
  6. Work out the **square root** of the **variance** (using a calculator)

Your notes

The effect of outliers

- An **outlier** is a score or value that falls far beyond the other values in a data set
- These extreme values can be caused by:
  - **variability** within the data
    - E.g., two people in a sample of 50 have abnormally good memory
  - **novel** data
    - E.g., people **self-report** the number of times they look at their fitness score on their smart watch
  - **errors** in how the data has been collected
  - E.g., some participants' memory scores were mistakenly not added to the statistical analysis
- Outliers can significantly affect calculation and **interpretation** of the mean
  - In a data set with outliers, the median is preferred over the mean, as it is not affected by extreme values
    - E.g., a data set comprising scores of 4, 6, 3, 7, 16, 2, 9, 4 would not be calculated using the mean due to the presence of the value 16 as this is significantly higher than the other values

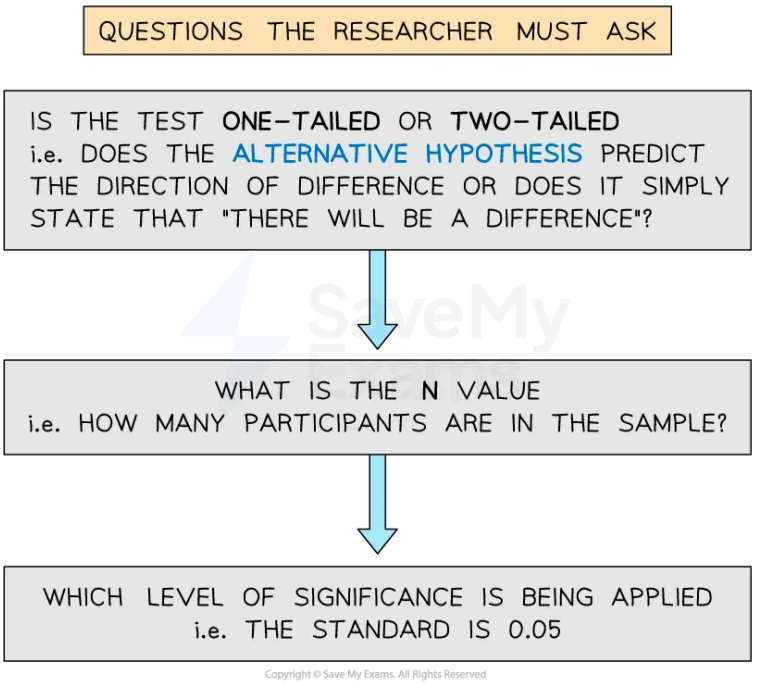
Your notes

Probability & significance

- To assess this, they use a **level of significance**, which reflects how likely it is that chance factors are responsible for the results
- The level of significance is expressed as **p (probability)**:
  - **p < 0.05** means that the probability of the result occurring by chance is less than or equal to **5%**
    - This is the **standard threshold** in psychology
  - **p < 0.01** means that the probability of chance is less than or equal to **1%**.
    - This more stringent level is used when:
      - There may be a **human cost** (e.g., drug trials)
      - Previous research findings are **contradictory**
- Researchers consult **statistical tables** to identify the **critical value** for their test
- If the calculated value meets or exceeds the critical value, they can **reject the null hypothesis** and conclude the result is significant

Using statistical tables

- Once the researcher has conducted their research and carried out a statistical test, the test produces an **observed** (or calculated) **value,** which is used to determine whether the results of their study are significant
- The **observed/calculated value** needs to be compared to the **critical value** in the **critical values table** to determine significance
- To find the **critical value** from the table, the researcher must ask the following questions, which will help them to use the critical values table properly:



Determining significance in statistical testing - A Level psychology diagram

Type I & type II errors

- A **Type I error** occurs when the null hypothesis is **rejected** when it should have been **accepted**
  - The researcher claims that the results **are significant** when in fact they are **not** (also known as a '**false positive**')
- A Type I error is more likely to happen when the researcher uses a probability value that is **too high, e.g.,**
  - 0.1 rather than 0.05
  - 0.06 rather than 0.05
- A **Type II error** occurs when the null hypothesis is **accepted** when it should have been **rejected**
  - The researcher claims that the results are **not significant** when in fact they **are** (also known as a '**false negative**')
- A Type II error is more likely to happen when the researcher uses a probability value that is **too low, e.g.,**
  - 0.01 instead of 0.05
  - 0.03 instead of 0.05
- Using a 0.05 significance level **guards against** making either a Type I or a Type II error

Your notes

Your notes

Hypotheses

- A **hypothesis** is a **testable statement** written as a **prediction** of what the researcher **expects** to find as a result of their experiment
- Where the **aim** of a study is expressed in **general terms** and outlines the focus of the study; **hypotheses must be precise** and **unambiguous**
- There are two types of hypothesis:
  - The **null** hypothesis (H_0)
  - The **alternative** hypothesis (H_1)

Alternative hypothesis (H_1)

- The H_1 should include the **independent variable (IV)** and the **dependent variable (DV)**
- Both the IV and the DV in the H_1 should be **operationalised**, which involves specifics on how each variable is to be **manipulated** (IV) and **measured** (DV)
- There are two different types of H_1:
  - **Directional (one-tailed)**
  - **Non-directional (two-tailed)**
- A **directional** hypothesis predicts the **direction** of the difference in conditions, i.e., it state that one condition will **outperform** the other
  - E.g., Participants who drink 200ml of caffeine before taking a memory test will correctly **recall more items** out of 15 than participants who drink 200ml of water before taking the same memory test
- A **non-directional** hypothesis does not predict the direction of the difference in conditions, i.e., it simply predicts that a difference will be shown
  - E.g., There **will be a difference** in the number of correctly recalled items out of 15 depending on whether participants have drunk 200ml of caffeine or 200ml of water before taking a memory test

Null hypothesis (H_0)

- All published psychological research must include the **null hypothesis (H_0)**; this is what **all** research starts with
- The H_0 begins with the idea that the IV will **not** affect the DV
  - It is the default assumption unless empirical evidence proves otherwise

Testing hypotheses

- The researcher must then write the H_0, which assumes '**no difference**'

- E.g., There **will be no difference** in the number of correctly recalled items out of 15 depending on whether participants have drunk 200ml of caffeine or 200ml of water before taking a memory test
- The researcher runs the experiment, uses statistical testing and then must form one of two conclusions:
  - If the result shows no difference between conditions (i.e., it is not **statistically significant**), then the H_0 must be **accepted**
  - If the result shows a difference between conditions (i.e., it is statistically significant), then the H_0 can be **rejected** (and the H_1 is then accepted)

Your notes

Hypotheses in correlational research

- Hypotheses for **correlational investigations** are written in the same way as experimental hypotheses, apart from **one crucial difference**
  - Instead of using the term 'difference', you have to use the term 'relationship or correlation', e.g.,
    - There will be a **relationship** between the number of cups of caffeine drunk and the number of hours slept per night across one week
      - This is a non-directional hypothesis
    - There will be a **negative correlation** between the number of cups of caffeine drunk and the number of hours slept per night across one week
      - This is a non-directional hypothesis
    - There will be **no relationship** between the number of cups of caffeine drunk and the number of hours slept per night across one week
      - This is a null hypothesis

Factors affecting the choice of a statistical test

- A statistical test determines if a **difference/correlation** is **statistically significant** according the level of significance applied
- Applying a statistical test to a data set determines:
  - if the outcome is **due to chance** or to the **effect of the IV** on the DV
  - whether the null hypothesis can be accepted or rejected
- There are 3 distinct criteria that a researcher must consider before deciding which statistical test to use:
  - Have they conducted a test of **difference** or a test of **correlation?**
  - If they have conducted a test of difference, did they use an **independent measures design, repeated measures** design, or a **matched pairs** design?
    - an **unrelated** design refers to independent measures/groups

- a **related** design refers to repeated measures and matched pairs
- Have they collected nominal, ordinal or interval data?
- The table below illustrates which test should be used and when:

	Tests of Difference		Tests of association or correlation
	Unrelated design	**Related design**	
Nominal data	Chi-Squared	Sign test	Chi-Squared
Ordinal data	Mann Whitney U	Wilcoxon T	Spearman's rho
Interval data (Parametric tests)	Unrelated t-test	Related t-test	Pearson's r

- Chi-Squared is a test of both difference and association
- Spearman's rho and Pearson's r are tests of correlation

Parametric & non-parametric tests

- Parametric tests assume the following:
  - **A normal distribution**
    - Occurs when data is **symmetrical around the mean**
    - Most scores cluster near the mean; fewer are at the extremes
    - Produces the familiar **bell curve** shape
    - E.g., height is a measurement that has a normal distribution
  - **The use of interval data or ratio data**
    - Requires the most **sensitive and precise** level of measurement
  - **Homogeneity of variance**
    - If the set of scores per data set/condition are **similar** in terms of their **dispersion**
    - If both conditions have similar standard deviations, this suggests the data is equally spread and clustered around the mean
- Non-parametric tests do **not** follow the same criteria as parametric tests
  - There is **no assumption of a normal distribution**
    - Useful when data is skewed or not continuous
    - E.g., scores on a **memory test**
  - Non-parametric tests use **nominal** or **ordinal** data

Your notes

- Non-parametric tests do **not** depend on homogeneity of variance
- Parametric tests are more **powerful** and **precise** than non-parametric tests
  - More likely to **detect a significant difference or correlation** if one truly exists

Your notes

Your notes

The correlation coefficient

- A **correlation** is not a research method but an **analysis of the relationship** between two co-variables.

  In correlational research:

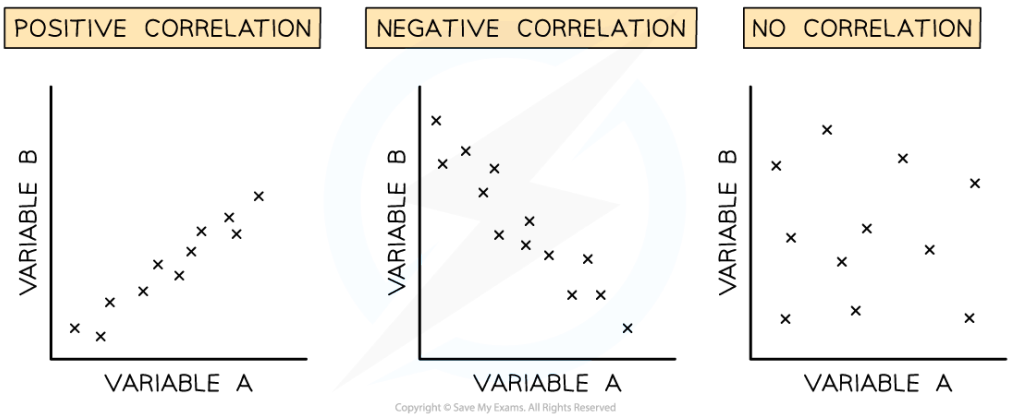
  - Variables are **not manipulated** (no IV).
  - Instead, two co-variables are **measured and compared** to identify relationships

Types of co-variables

- One or both of the co-variables could be **pre-existing**, e.g.,
  - School attendance (days present in Year 11) and number of GCSEs achieved.
  - Average August temperature and number of arrests for violent behaviour in a town
- One or both of the co-variables could be **collected data**, e.g.,
  - Number of arguments with a partner in a month and self-reported stress levels
  - Average hours of sleep in a week and number of caffeinated drinks consumed

How correlations are measured

- Each participant contributes **two scores** (one for each co-variable)
- Data are plotted on a **scattergraph**, with each point representing paired scores
- Scattergraphs typically show one of three outcomes:
  - **Positive correlation:** as one increases, the other increases (e.g., calories consumed and weight gained)
  - **Negative correlation:** as one increases, the other decreases (e.g., hours sitting and fitness level)
  - **Zero correlation:** no relationship (e.g., hair colour and IQ)



Types of correlation

Your notes

- Analysing the relationship between co-variables can be done by
  - visually 'eyeballing' the scattergraph to see the direction of the relationship (positive, negative or none at all)
  - calculating the **correlation coefficient,** which is expressed as a **numerical value**

The correlation coefficient

- A **numerical value** between **–1 and +1** showing both the **strength** and **direction** of a relationship
  - **+1** is a **perfect positive** correlation
  - **–1** is a **perfect negative** correlation
  - **0** is a **no correlation**
- Strength can be described as **weak, moderate, or strong** (applies to both positive and negative)
  - +0.03 is a **weak positive** correlation
  - –0.05 is a **moderate negative** correlation
  - –0.09 is a **strong negative** correlation
- The **correlation coefficient** represents both the **direction** and the **strength** of the r

Evaluation of the correlation coefficient

Strengths

- The correlation coefficient is a **quick and easy** way to analyse data
  - This is a strength, as it enables the researcher to access **large amounts** of data that would otherwise be impossible to gather if they tried to amass this from scratch
  - Large amounts of quantitative data mean that the research is high in **reliability**

- Correlation coefficients allow researchers to make **predictions** as to the relationship between co-variables
    - E.g., knowing that there is a relationship between school absence and GCSE results could be used to identify students **at risk** and to **implement interventions** to help them achieve their potential

Your notes

Limitations

- **Extraneous factors** connected to one or both co-variables may affect the result and lead to **invalid** conclusions being made
    - E.g., number of days of absence from school may be due to **illness** rather than to choice
    - a low GCSE score may be due to a **high turnover of teachers** in one school rather than to student absence
- Correlations cannot establish **cause and effect** — only association
- Correlation coefficients are useful for analysing **linear relationships** (height and shoe size)
    - They are **less successful** when dealing with **non-linear relationships** (number of hours worked and level of happiness)

Your notes

Thematic analysis

- A qualitative method for analysing data such as **books, diaries, interview transcripts, conversations, text messages, or film scripts**
  - Data is **organised into categories** (e.g., *early life*, *school experience*, *relationships*) and further divided into **sub-themes** (e.g., *conflict with sibling*, *bullying*, *abusive partner*)
- Thematic analysis can be used to analyse **primary** (e.g., interviews, conversations) **or secondary data** (e.g., published texts, films)
- The aim is to **summarise and interpret the main ideas** within the material in order to identify **patterns and conclusions**
- It is an **inductive method**: themes emerge from the data rather than being imposed beforehand
- Examples of approaches used in thematic analysis might include:
  - Analysing a transcript of a couple's argument, noting frequent references to "*blame*" or "*upset*"
  - Reviewing text messages to identify patterns of **coercive control** in a relationship
  - Examining a film script to explore how female characters are described by appearance while male characters are described by personality traits

Evaluation of thematic analysis

Strengths

- Qualitative data is rich in **meaning and detail,** which brings with it **external validity**
  - Sorting the data into themes means that **trends and patterns** emerge, which can provide **insight** into the topic
  - These patterns can be used for further investigation into the topic
- Thematic analysis enables researchers to investigate topics which might otherwise be off-limits due to **ethical concerns**
  - If secondary data is used which is in the public domain, then it should not compromise anyone's **privacy**
  - There is no need to gain **informed consent** to access or report on secondary data

Limitations

- The interpretation of themes can be **subjective**
  - The researcher's preexisting **biases** may influence the analysis, which would compromise the **validity** of the findings

- Researchers may not all practise **reflexivity** fully when conducting a thematic analysis
- The processes used in thematic analysis are **time-consuming** and repetitive
  - This means that researchers may lose focus and overlook important details
  - Consequently, it is less commonly used than quicker methods, despite its ability to provide deep insights into subjective experiences

Your notes