

TOPIC 4: INFERENCE TESTING

Specification says you need to know:

- Probability and significance: use of statistical tables and critical values in interpretation of significance; Type I and Type II errors.
- Factors affecting the choice of statistical test, including level of measurement and experimental design. When to use the following tests: Spearman's rho, Pearson's r, Wilcoxon, Mann-Whitney, related t-test, unrelated t-test and Chi-Squared test.

LESSON 23: PROBABILITY AND SIGNIFICANCE

Learning objectives: You should be able to:

1. Explore what is meant by probability
2. Determine what is meant by significance and chance
3. Explore what is meant by a Type I and Type II error

Starter: Let's practice probability

For example with a pack of cards:

- 1) What is the probability that if I pick one card it will be from a red suit?

.....

- 2) What is the probability that if I pick one card it will be a picture card?

.....

- 3) What does it mean if the probability of something happening is 0?

.....

- 4) What does it mean if the probability of something happening is 1?

.....

- 5) If the probability of it raining is 0.05 and the probability is the same every day, how many days can I expect it to rain in

- a) 100 days
- b) 200 days
- c) 40 days

Activity 1: Definition

Define the term **Probability**:

Probability and Significance:

Probability, or p , is expressed as a number between 0 and 1. 0 means an event will not happen, 1 means that an event will definitely happen. The P value will always be found to be between 0 and 1 due to the way in which it is calculated. To calculate the probability that a particular outcome will occur, it has to be divided by the number of possible outcomes.

One way to work out the probability of something occurring is to use this formula:

$$\text{Probability} = \frac{\text{number of particular outcomes}}{\text{number of possible outcomes}}$$

Sometimes it is a little more complicated to work out the probability of something. For example, there may be a probability that one in nine of the entire population will develop cancer at some point in our lives. However, the probability of us doing so is greatly increased by other factors such as lifestyle choices (e.g. diet and smoking). In this case, the researcher would need to break down the sample into groups according to these other factors. This enables the psychologist to work out **conditional probability** – the probability of something happening **if something else occurs**.



Probability and Significance continued....

Researchers use statistical tests to work out how probable it is that something might occur. For example, research has found a link between aggression and playing video games in children but we cannot accurately say that there is a high probability of this occurring across all children in the country without running a statistical test.

Research studies should have an alternate hypothesis (research and experimental). This states the relationship between variables in correlation and difference between conditions in experiments. A null hypothesis is also used suggesting there is no relationship or difference. Carrying out a statistical test allows us to either accept or reject these hypotheses (depending on the result).

Activity 2: Types of hypotheses

Experimental hypotheses can either be:

- D _____ or
- N _____

A significant result is one where there is a low probability that chance factors were responsible for any observed difference, correlation or association in the variables tested.

The question is how large an effect (difference or relationship) is required for psychologists to conclude that a result is significant (i.e. not due to chance)?

It is useful to think about the words that commonly occur in different types of hypotheses. This can often help when building your own.

E.g. In experiments:

- Directional hypotheses – more/greater
- Non-directional hypotheses – a difference
- Null hypotheses – no difference

In correlations (this is harder to generalise):

- Directional hypotheses – positive/negative, as x increases y decreases.....
- Non-directional hypotheses – a correlation/ relationship
- Null hypotheses – no relationship/correlation

FINALLY – a top tip is to always write hypotheses in pencil first as there are so many steps (e.g. getting the direction right, operationalising) that an error first time through is quite likely. This also encourages careful checking.

Activity 3: Hypotheses and significance (<https://www.youtube.com/watch?v=0zZYBALbZgg>)

In each of the following cases, write a suitably operationalised:

a) Null hypothesis H_0

b) Directional alternative hypothesis H_1

- 1 An investigation to see if people who regularly watch football are better at remembering a set of football scores than people who never watch football.

H_0

H_1

- 2 An investigation to see if people farming in Wales or England suffer more symptoms of stress.

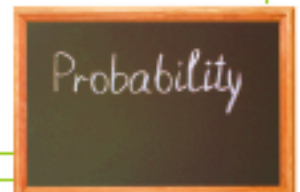
H_0

H_1

- 3 An investigation into the relationship between alcohol consumed per month and work absence.

H_0

H_1



- 4 What is another term for a one-tailed hypothesis?

- 5 When would you use a one-tailed hypothesis?

- 6 What is the usual level of significance employed in psychology? Write as decimal and percentage.

- 7 Justify why psychologists use the level stated above.

- 8 Under what circumstances might a more stringent level of significance be used and what would the level be?

- 9 What is a Type II error?

- 10 When is a Type II error most likely to occur?



Activity 4: Probability levels (http://onlinestatbook.com/2/logic_of_hypothesis_testing/errorsM.html)

Psychologists have concluded that for most purposes, the 5% level of significance will be used ($p=0.05$)

Why 0.05?

$p = 0.05$ represents a 5 per cent probability (or 1 in 20).

Why 0.01?

Sometimes 0.01 level is preferred to 0.05. A common reason is that the findings are likely to be controversial or raise ethical dilemmas. The researcher would want to be more stringent and only present the findings as significant if there was a very small probability that the null hypothesis was true. OR the study is theoretically important

Which errors would a researcher be making if their significance level was:

a) Too lenient?

b) Too strict?

		Truth	
		H_1 is correct. There is something going on	H_0 is correct. There is nothing going on
Test result	Reject H_0	True positive	False positive (likely when p too lenient, i.e. 10%) TYPE 1 ERROR
	Accept H_0	False negative (likely when p too stringent, i.e. 1%) TYPE 2 ERROR	True negative

Type 1 and Type 2 errors

If you use level of significance that is too high (too lenient), such as let's say 10%, then you may reject a null hypothesis that is true. This is called Type 1 error and its likelihood is increased if the significance level is too high. Type 2 error occurs when we accept a null hypothesis that is not true. Type 2 occurs when the significance level is too low (stringent) such as 1%. Therefore we can say that at 5% significance, 5% uncertainty is usually acceptable if it is not a life and death matter. If for example, research is on a new drug then psychologists will select 1% significance level because they have to be very careful about taking chances. So in the nutshell:

Type 1 error occurs when the null hypothesis is rejected but it should not have been.

Type 2 error occurs when the null hypothesis is accepted but it should not have been.

The likelihood of type 1 error is increased if the significance level is too high (eg. 10%). Remember – Type 1 error is made by a researcher who wants to be the best and first at everything, so he puts the significance level too high so that his hypothesis is correct and he wins the 1st prize.

The likelihood of type 2 error is increased if the significance level is too low (eg. 1%). Remember – Type 2 error is made by a caring research who doesn't mind coming 2nd. He only allows himself very low possibility of chance (1%) and he is not bothered if he needs to reject his alternative hypothesis.

Observed and critical value

The value that we achieve at the end of the calculations of statistical test is called the **observed value** (because it is based on the observation we have made). To decide if this observed value is significant, this figure is compared to another number found in **table of critical values**. This is the **critical value** and is the value that our result must be greater/lower (depending on the test) in order for the null hypothesis to be rejected. This is how to remember this rule: if there is a letter **R** in the name of the test, then the observed value must be **gReateR** than critical value, so for Spearman's chi-square observed value must be greater than critical value. For Mann-Whitney and Wilcoxon observed value should be less than critical value to be significant. You don't need to worry too much about remembering this, as it is given to you at the bottom of statistical table.

In order to find the appropriate critical value in the table you need to know:

- Number of participants (N)
- Whether to use one-tailed test (for directional hypothesis) of two-tailed test (for non-directional hypothesis)
- Significance level selected, usually $p \leq 0.05$

Activity 5: <https://www.youtube.com/watch?v=hZxznfnt5v8>

Before we have a lesson on choosing a statistical test, let's recap some of the basics

Activity 6: Revising qualitative and quantitative data

This requires you to practise questions on first qualitative and quantitative data and then types of quantitative data, i.e. nominal/ordinal/interval.

Revising qualitative and quantitative data

In each of the following cases, imagine you are writing a questionnaire to collect data and state one question that would enable you to collect **quantitative** data and one that would yield **qualitative** data.

1. A researcher wants to measure stress levels in teachers.
2. A researcher wants to measure differences in smoking behaviour and attitudes towards smoking between teenage boys and girls.
3. A researcher wants to investigate sleep problems in new mothers.
4. A researcher wants to investigate levels of anger in rail passengers whose train has been delayed.

Types of data

Quantitative data can be classified as **Nominal**, **Ordinal** or **Interval**.

In each of the cases below identify the data produced.

- | | |
|--|--|
| 1. Heart rate in newborn babies. | |
| 2. How many people in a class would vote for each of the main political parties. | |
| 3. Ratings of how happy each student feels about their psychology test score. | |
| 4. The time taken for participants to complete a questionnaire. | |
| 5. Patients' ratings of satisfaction with the service of their dentist. | |
| 6. The number of calories consumed by a participant on each day. | |
| 7. How many boys and girls each choose to do either an apprenticeship or a degree. | |
| 8. Favourite foods of a group of nurses. | |
| 9. A study of the main mode of transport to school of Year 7. | |
| 10. The weight of participants beginning a diet study. | |

Extension activity

Thinking of studies that you have covered so far, state **two** that have collected quantitative data and **two** that have collected qualitative data and state what data you are referring to.

Null and alternate hypotheses

Researchers begin their investigation by writing a **hypothesis**. This may be **directional** if there is previous consistent previous research, or **non-directional**. These hypotheses are often referred to as an **alternative hypothesis**, as it is alternative to the **null hypothesis**. This states there is no difference or relationship between conditions.

The **statistical test** determines which hypothesis is 'true' and thus whether we accept or reject the null hypothesis.

Chance

Chance refers to something with no cause. It just happens. WE decide on a probability that we will 'risk'. You can't be 100% certain that an observed effect was not due to chance, but you can state how certain you are.

Levels of significance and probability

Actually 'true' is the wrong word. Statistical tests work on the basis of probability rather than certainty. All statistical tests employ a **significance level** – the point at which the research can claim to have discovered a significant difference or correlation within the data. In other words, the point at which the researcher can reject the null hypothesis and accept the alternative hypothesis.

The usual significance level in psychology is 0.05 (or 5%).

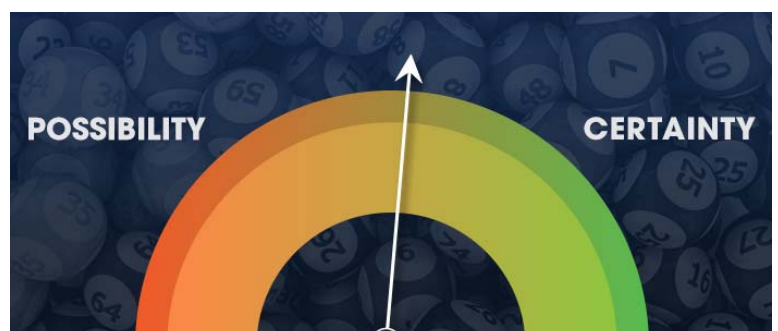
This is properly written as $p \leq 0.05$ (p means probability).

This means the probability that the observed effect (the result) occurred by chance is equal to or less than 5%. In effect, this means that even when a researcher claims to have found a significant difference/correlation, there is still up to 5% probability that the observed effect occurred by chance – that it was a 'fluke.'

Psychologists can never be 100% certain about a particular result as they have not tested all members of the population under all possible circumstances! For this reason, psychologists have settled upon a conventional level of probability where they are prepared to accept that results may have occurred by chance – this is the 5% level

Lower levels of significance

Occasionally, a more stringent level of significance may be used (such as 0.01) in studies where there may be a human cost – such as drug trials – or 'one off' studies that could not, for practical reasons, be repeated in future. In all research, if there is a large difference between the calculated value and critical values – in the preferred direction – the researcher will check more stringent levels, as the *lower* the p value is, the more statistically significant the results.



Activity 2: Pregnancy tests

Pregnancy tests are not 100% reliable so women who suspect they are pregnant are advised to take more than one test in order to make sure.

If the result says you are not pregnant – in what way could this be a Type 2 error.



Since type I and type II errors are hard to remember, the "VERTICAL LINES" help you with related concepts and formulas!

Type **I** has one vertical line.

P has one vertical line.

Therefore, associate with false **p**ositive rate.

Type **II** has two vertical lines.

n has two vertical lines.

Therefore, associate with false **n**egative rate.

Use statistical tables

The critical value

Once a statistical test has been calculated, the result is a *number* - the **calculated value** (or observed value). To check for statistical significance, the calculated value must be compared with a **critical value** – a number that tells us whether or not we can reject the null hypothesis and accept the alternative hypothesis.

Each statistical test has its own **table of critical values**, developed by statisticians. These tables look like very complicated bingo cards (you will see plenty of examples over the next few lessons). For some statistical tests, the calculated value must be equal to or greater than the critical value, for other tests the calculated value must be equal to or less than the critical value.

Using tables of critical values

How does the researcher know which critical value to use? There are three criteria:

One tailed or two tailed test? You use a one tailed test if the hypothesis was directional and a two tailed test for a non-directional hypothesis. Probability levels *double when* two tailed tests are being used as they are more *conservative* predictions.

The number of participants in the study. This usually appears as the *N* value on the table. For some tests **degrees of freedom (df)** are calculated instead.

The **level of significance** (or P value). As discussed, the 0.05 level of significance is the standard level in psychological research.

Key terms

Alternative hypothesis – A testable statement about the relationship (difference, association etc.) between two or more variables.

Null hypothesis – An assumption that there is no relationship (difference, association etc.) in the population from which a sample is taken with respect to the variables being studied.

Probability – A numerical measure of the likelihood or chance a certain event will occur, where 0 indicates statistical impossibility and 1 statistical certainty.

Statistical test – This gives the probability that a particular set of data did not occur by chance.

Type 1 error – The incorrect rejection of a true null hypothesis (a false positive)

Type 2 error – The failure to reject a false null hypothesis meaning the researcher accepts a null hypothesis that was not true (a false negative)

Critical value – when testing a hypothesis, the numerical boundary or cut-off point between acceptance and rejection of the null hypothesis.

LESSON 24: CHOOSING A STATISTICAL TEST

Learning objectives: You should be able to:

1. Explore factors that affect the choice of statistical tests
2. Determine whether focus is on difference or correlation, experimental design and levels of measurement
3. Understand when to use each of the following tests: Spearman's rho, Pearson's r, Wilcoxon, Mann-Whitney U, Related t-tests, Unrelated t-tests and Chi squared test

Starter:

<https://www.youtube.com/watch?v=oHGr0M3TlcA>

<https://www.youtube.com/watch?v=fv5SW5-u2M>

There are two types of statistics, descriptive statistics (such as averages, measures of dispersion and graphs covered in year 1) and inferential statistics which we started in year 1 when we researched the sign test. Inferential tests are designed by statisticians who work out the probabilities of certain results so we can decide whether to accept or reject a null hypothesis. They are called 'inferential' because the statisticians make an inference (deduction) about whole populations based on small samples.

Statistical testing

In year 1 you had a brief introduction to the concept of statistical testing using the example of a sign test. You should recall that a statistical test is used to determine whether a difference or an association found in a particular investigation is statistically **significant** – that is, more than could have occurred by **chance**. The outcome of this has implications for whether we accept or reject the null hypothesis. This lesson we will focus on which statistical test is used under what circumstances. There are three factors that help to decide this:

1. Whether the researcher is looking for a *difference* or **correlation**
2. In the case of a difference, what **experimental design is being used**.
3. The **level of measurement**.

1. Difference or correlation

The first thing to consider when deciding which statistical test to use relates to the aim or purpose of the investigation – namely, is the researcher looking for a difference or correlation. This should be obvious from the working of the hypothesis. In this context, 'correlation' can include correlational analyses as well as investigations that look for an **association**.

Activity 1: difference or correlation?

- Researching an association between time spent on homework (1/2 hour to 3 hours) and number of G.C.S.E. passes (1 to 6). _____
- Researching whether different verbs used in a question influences the accuracy of eyewitness testimony in the form of estimated speed given. _____
- Researching the relationship between watching violence on T.V. and violent behaviour in adolescence. _____
- To see if people work harder when they eat breakfast before coming to school compared to not eating breakfast when coming to school. _____

2. Experimental design

You will also remember from Year 1 studies that there are three types of experimental designs. These are:

Repeated measures and matched pairs are referred to as a **related design**. In repeated measure the same participants are used in all conditions of the experiment. In a matched pairs design participants in each condition are not the same but have been 'matched' on some variables that are important for the investigation which makes them 'related'. For this reason, both designs are classed as *related*.

As participants in each condition of independent groups design are different, this design is *unrelated*.

Thus the researcher chooses from two alternatives here: *related* or *unrelated*.

Remember this doesn't count when looking at a correlation, this only occurs when looking at the difference.

Activity 2: Experimental designs

Decide which are the following are Independent, repeated or matched. Also in brackets put 'R' for related or 'U' for unrelated

- At a day-care, the staff has had problems with the children behaving badly every day. They begin to test to see how the children react if the staff gives them candy when they are good and the same children no candy when they are bad. The staff hopes that the incentive for the children will improve their behaviour.

- To see whether boys are more superstitious than girls

- In order to compare the effectiveness of two different types of therapy for depression, depressed patients were assigned to receive either cognitive therapy or behaviour therapy for a 12-week period. The researchers attempted to ensure that the patients in the two groups had a similar severity of depressed symptoms by administering a standardised test of depression to each participant, then pairing them according to the severity of their symptoms.

- In order to assess the effect of organisation on recall, a researcher randomly assigned student volunteers to two conditions. Condition one attempted to recall a list of words that were organised into meaningful categories; condition two attempted to recall the same words, randomly grouped on the page.

- To assess the effectiveness of two different ways of teaching reading, a group of 5-year-olds were recruited from a primary school. Their level of reading ability was assessed, and then they were taught using scheme 1 for 20 weeks. At the end of this period, their reading was reassessed, and a reading improvement score was calculated. They were then taught using scheme 2 for a further 20 weeks and another reading improvement score for this period was calculated. The reading improvement scores for each child were then compared.
-

- In order to assess the effects of fatigue on reaction times, a researcher gave participants a target detection test in which they pressed a button every time a dot appeared on a screen. The time between the dot appearing and the button being pressed was recorded. The participants did the test twice, once first thing in the morning, and once last thing at night.
-

3. Levels of measurement

Quantitative data can be divided into different levels of measurement and this is the third factor influencing the choice of statistical test. There are three levels of measurement: nominal, ordinal and interval.

Nominal

Nominal data: Data is presented in the form of categories – hence nominal is sometimes referred to as **categorical data**. E.g. no. of boys and girls in your year group, you count the number in each male or female category.

Nominal data is **discrete** in that one item can only appear in one of the categories. E.g. choosing your favourite football team, the vote only appears in one category.

Ordinal

Ordinal data: is ordered in some way. E.g. asking people in the class to rate how much you like psychology on a scale of 1 to 10, where 1 is you don't like it and 10 is absolutely love it. As this may affect my ego, we won't be doing this task 😊.

Ordinal data does not have equal intervals between each unit unlike in interval. E.g. someone who rates psychology as an 8 does not enjoy it twice as much as someone who has put 4.

Ordinal data also lacks precision because it is based on subjective opinions rather than objective measures. For these reasons, ordinal data is sometimes referred to as 'unsafe' data as it lacks precision. Due to this ordinal data is not used as part of a statistical test, instead raw scores are converted into ranks (e.g. 1st, 2nd, 3rd etc.) and it is the ranks – not the scores – that are used in calculations. We will go over this again when we look at each statistical test.

Interval

Interval data in contrast to ordinal is based on numerical scales that include units of equal, precisely defined size. In this sense it is 'better' than ordinal as it provides more detail and is preserved. To think of ordinal think of methods you would use in maths and science such as a stopwatch (time), thermometer

(temperature) or weighing scales (weight). For example if we recorded how long it took each participant to complete a written recall test in psychology, we would have collected interval data.

Interval data is the most precise and sophisticated form of data in psychology and is a necessary criterion for the use of parametric tests.

Activity 3: What's the level of measurement?

Identify whether the following would produce nominal, ordinal or interval data.



- Time taken to sort cards into categories

- People's choice of the Sun, The Times or the Guardian



- Participants' sense of self-worth, estimated on a scale of 1-10



- Judges in a dancing competition giving marks for style and presentation.



- Participants' reaction to aversion stimuli measured using a heart rate monitor



- A set of medical records classifying patients as either chronic, acute or 'not yet classified'



Some of the data produced in psychology is quite difficult to classify. For example, show the 'no. of words recalled' in a memory test be treated as interval or ordinal data?

Strictly speaking, this would only be interval if the words are all of equal difficulty (so units of measurement are all equivalent). This would be difficult to achieve as some words will always be more memorable than others! For this reason, it is probably 'safe' to treat no. of words recalled as ordinal data and rank the set of scores accordingly.

You must always provide reasoning when deciding which level of measurement is appropriate.

The table below shows levels of measurement and their relation to the appropriate measures of central tendency and measures of dispersion.

Level of measurement	Measure of central tendency	Measure of dispersion
Nominal	Mode	n/a
Ordinal	Median	Range
Interval	Mean	Standard deviation

Note that the range and standard deviation cannot be calculated on nominal data as such data is in the form of frequencies. It is not appropriate to use the mean or the standard deviation for ordinal data as the intervals between the units of measurement are not of equal size.

Choosing a statistical test

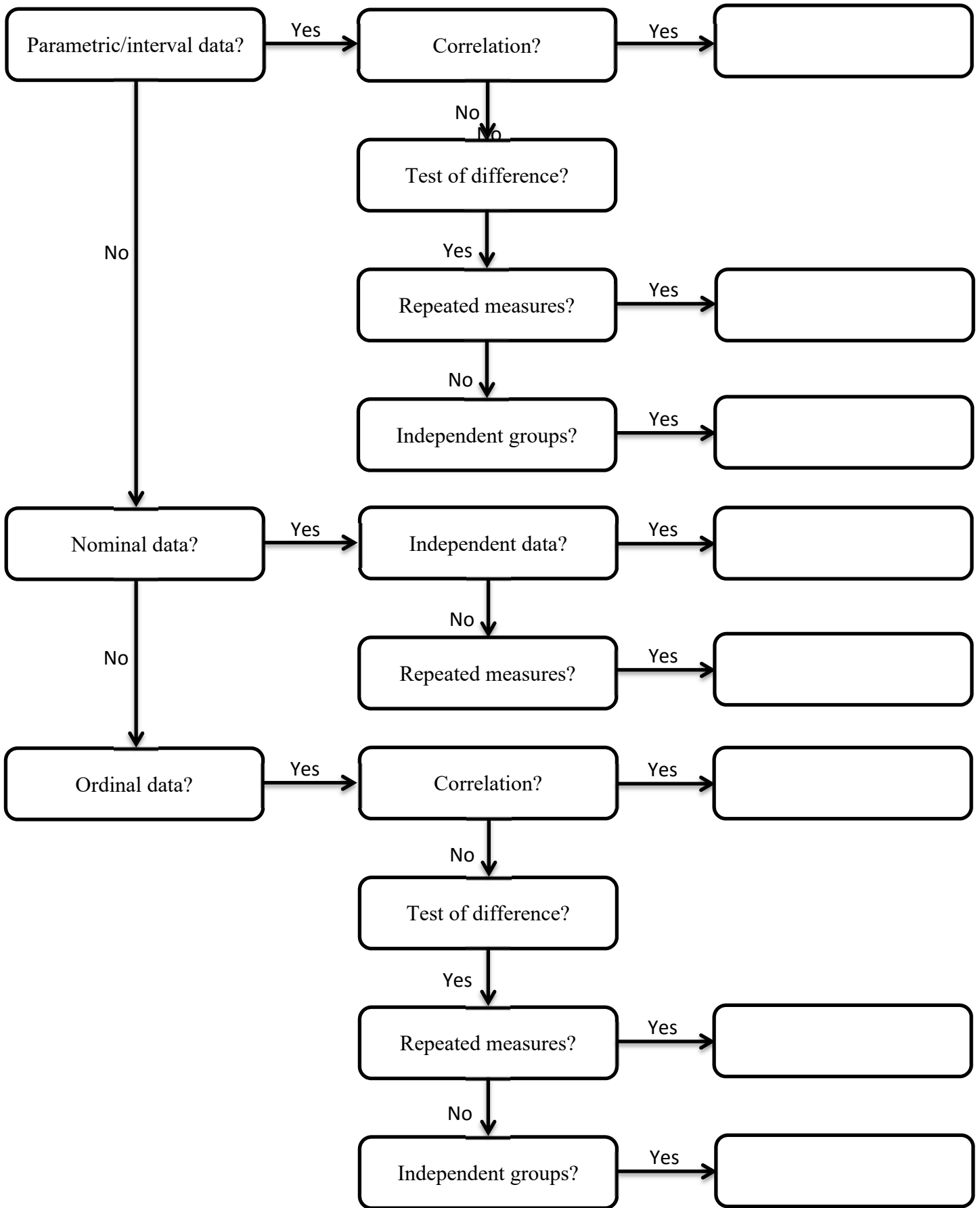
In a nutshell	Test of difference		Test of association or correlation
	Unrelated design	Related design	
Nominal data	Chi-square	Sign test	Chi-square
Ordinal data	Mann-Whitney	Wilcoxon	Spearman's Rho
Interval data	Unrelated t-test	Related t-test	Pearson's r

Note that Chi-square is a test of both difference and correlation. Data items must be unrelated.

*Also the three tests at the bottom in blue are **parametric tests***

There are a large number of statistical tests that are used by psychologists. These are divided into parametric and non-parametric tests. Parametric tests are preferred to non-parametric because they are more powerful. However they can only be used if certain criteria are met.

Activity 4: Selecting a statistical test – use the table to help fill in the blanks



Remember matched pairs counts as repeated measures (related) design because the two groups of participants are related

Parametric criteria

The related t-test, unrelated t-test and Pearson's r are collectively known as **parametric tests**. Parametric tests are more powerful and robust than other tests. If a researcher is able to use a parametric test they will do so, as these tests may be able to detect significance within some data sets that non-parametric tests cannot.

There are three criteria that must be met in order to use a parametric test:

1. Data must be **interval level** – parametric tests use the actual scores rather than ranked data
2. The data should be drawn from a population which would be expected to show a **normal distribution** for the variables being measured. Variables that would produce a **skewed distribution** are not appropriate for parametric tests.

It is not the sample that must be normally distributed but the population (normal distribution means most items cluster around the mean with an equal number of items above and below the mean).

3. There should be *homogeneity of variance* (Some common statistical procedures assume that **variances** of the populations from which different samples are drawn are equal.) the set of scores in each condition should have similar dispersion or spread. One way of determining variance is by comparing the standard deviations in each condition; if they are similar, a parametric test may be used. In a related design it is generally assumed that the two groups of scores have a similar spread.

Activity 5

If a researcher compared two related sets of data and was looking to see if they were different, why would it be preferable to use a related t-test instead of a Wilcoxon?

Activity 5: Which test am I?

1.

- I am used to test for differences between sets of scores.
- When the data is ordinal.
- When the design is unrelated.

2.

- I am used as a test of correlation.
- Where data is nominal.

3.

- I am a test for a difference between two sets of scores.
- I am used where data is interval.
- I am used where the data is unrelated.

4.

- I am a test for correlation.
- When data is ordinal.

5.

- I am a test for the difference between two sets of scores.
- I am used where data is interval.
- I am used where the data is related.

6.

- I am a test of correlation.
- I am used where data is interval.

7.

- I am used to test for differences between sets of scores.
- When the data is at least ordinal.
- When the design is related.

8.

- I am used to test for differences between sets of scores.
- When the data is nominal.
- When the design is related.

9.

- The alternative hypothesis suggests that there is a difference in scores on a GCSE-style maths test between 18-year-olds that have continued to study Maths post-GCSE and those that have not.

10.

- Students want to investigate the correlation between happiness scores and how many friends people have on Facebook.

11.

- An investigation where a correlation between participants' ages and scores on a memory test is being tested.

12.

- A researcher wants to investigate whether reaction times increase after a person is given a dose of caffeine equivalent to three cups of coffee. The participant's base reaction times are compared with those 30 minutes after the caffeine intake.

13.

- In an investigation into gender and conformity, female participants were tested twice in an Asch-style line task test. In one condition the confederates were 6 females and in the second condition the confederates were 6 males. The number of times that the female participants conformed in each case was compared.

14.

- An observation is designed to test whether men or women are more likely to go through traffic signals when the signal is showing red.

15.

- Sixth form students were asked to rate themselves as either sociable or not sociable at the start and the end of their two-year courses.

16.

- An investigation comparing attractiveness ratings of a sports car between a group of drivers and a group of non-drivers.

Key terms

Statistical test determines whether a significant difference or correlation exists (consequently whether the null hypothesis should be rejected or retained). Procedures draw a logical conclusion (inferences) about the population which samples are drawn.

Levels of measurement quantitative data can be measured differently, the lower the level the less precise

Chi square a test for association (difference or correlation) between two variables or conditions. Data should be nominal level using an unrelated (independent) design

Mann-Whitney a test for a significant difference between two sets of scores. Data should be at least ordinal level using an unrelated design

Wilcoxon a test for significant difference between two sets of scores. Data should be at least ordinal level using a related design

Spearman's rho a test for correlation when data is at least ordinal level

Pearson's r a parametric test for correlation when data is at least ordinal level

Related t-test a parametric test for difference between two sets of scores. Data must be interval with a related design

Unrelated t-test a parametric test for difference between two sets of scores. Data must be interval with an unrelated design

Calculated value In a statistical test the value of the test statistic that must be reached to show significance.

Critical value in a statistical test the value of the test statistic that must be reached to show significance

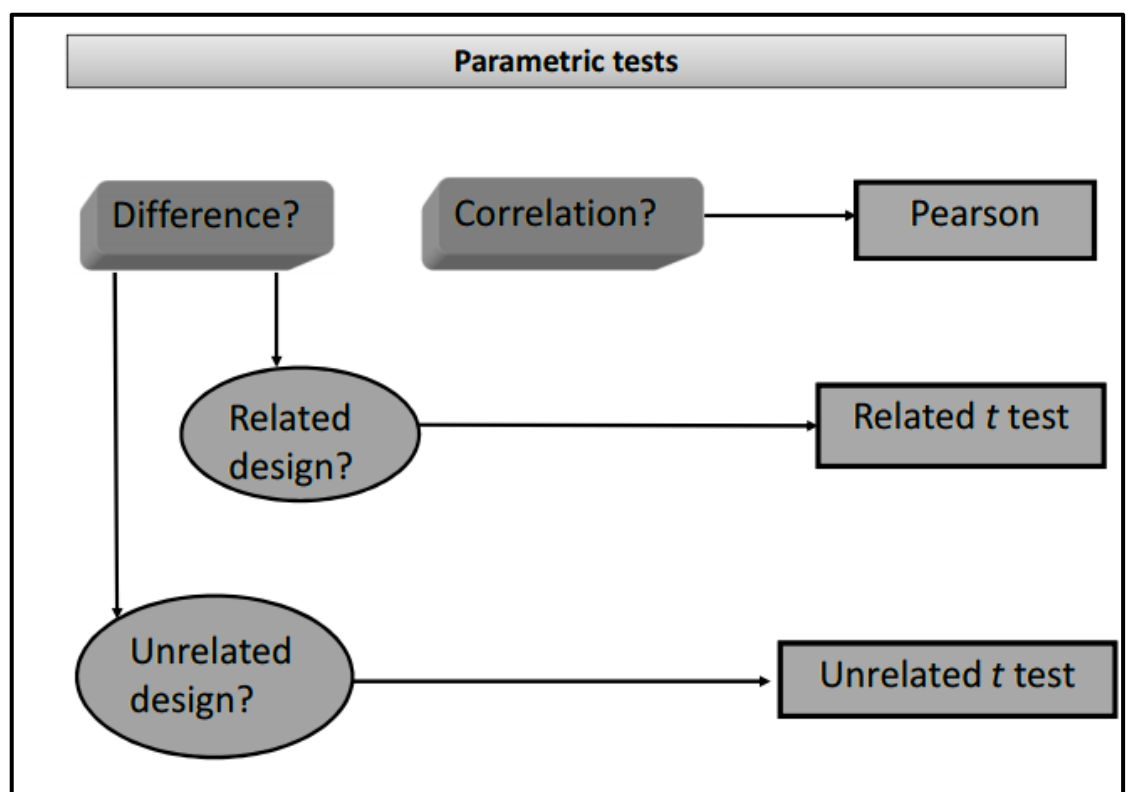
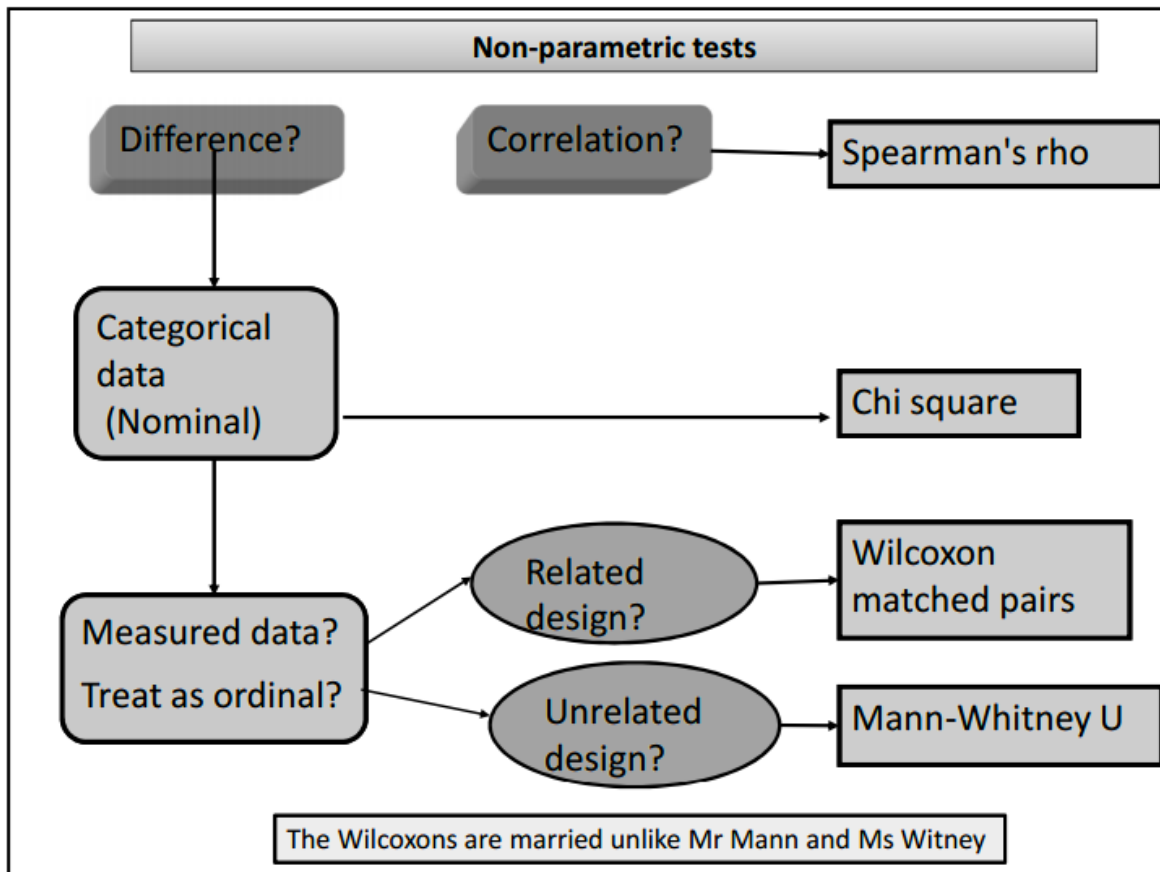
Degrees of freedom the number of values that are free to vary given that the overall total values are known

Significance level a statistical term indicating that the research indicating that the research findings are sufficiently strong to enable a researcher to reject the null hypothesis under test and accept the research hypothesis.

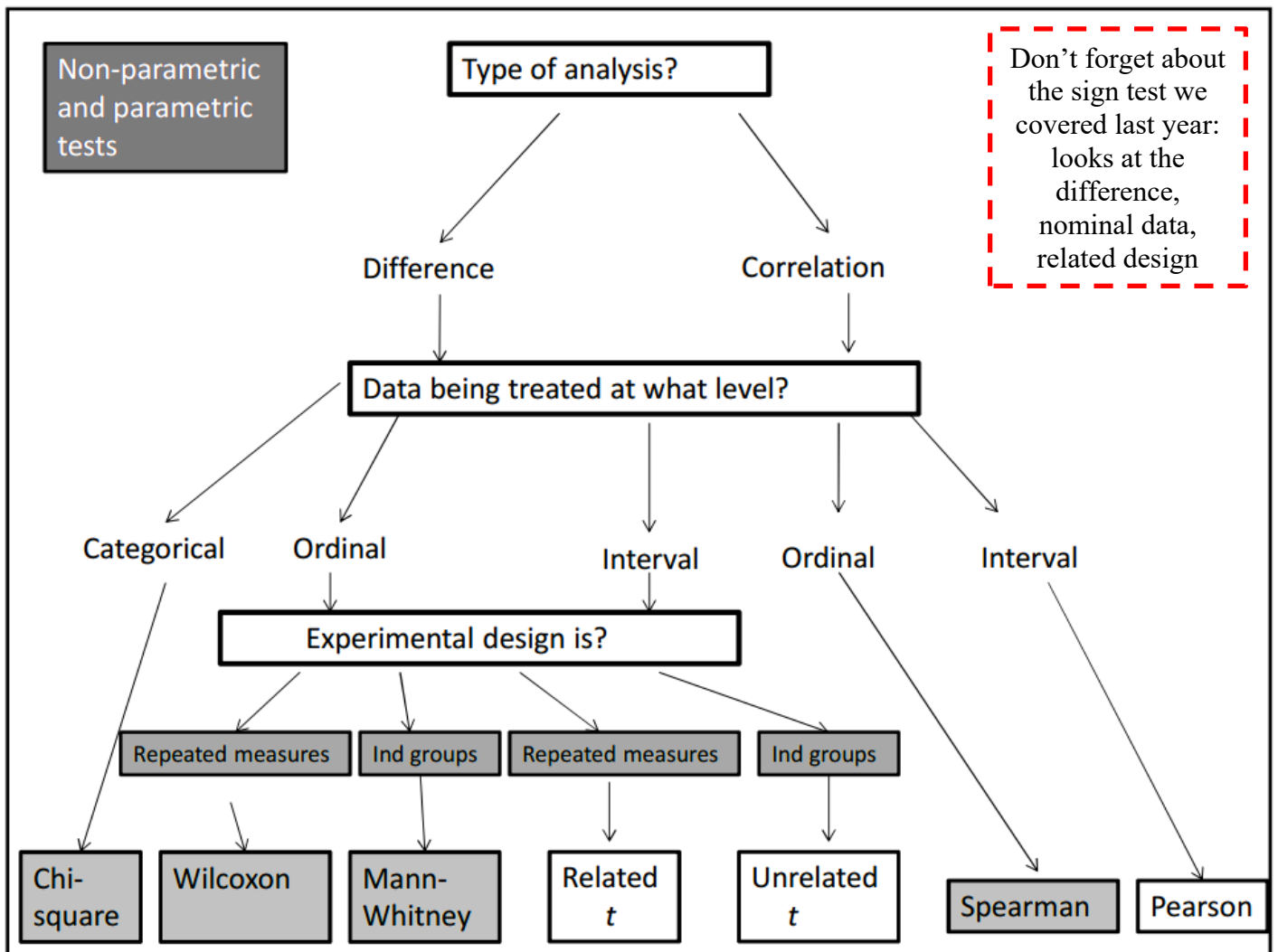
LESSON 25: CHOOSING A STATISTICAL TEST

Learning objectives: You should be able to:

1. Explore factors that affect the choice of statistical tests
2. Determine whether focus is on difference or correlation, experimental design and levels of measurement
3. Understand when to use each of the following tests: Spearman's rho, Pearson's r, Wilcoxon, Mann-Whitney U, Related t-tests, Unrelated t-tests and Chi squared test



Don't forget about the sign test we covered last year: looks at the difference, nominal data, related design



Starter: Guess the test. For each scenario name the test and bullet point the reasons why you have gone for this test

1. Bandura's Bobo study involved two groups of children, one group saw a violent model and the other group saw a non-violent model. Both groups were later rated for their aggressiveness in a play situation. Was the violent model group more aggressive?

.....

.....

.....

.....

2. Same question. In a similar study, children were not rated but were simply classed as aggressive or not.

.....

.....

.....

.....

3. A researcher selected a simple random sample from among the psychology student population in her University department. She wished to show that the same participants would perform a simple task faster in front of an audience than when alone.

4. A study on intelligence looked at the similarity of IQ scores between identical twins. Was there a relationship?

5. A study looked at the effects of exercise on time spent sleeping. For a week before the study each participant recorded how many hours they spent asleep in order to work out their average time spent sleeping. They then spent a day doing extreme exercise and reported how long they slept that night. Did exercise increase sleep?

6. Another sleep study looked at the same issue but this time they asked one participant to keep a diary for a month. On each day she had to write down how many hours of sleep she had and also list all the activities she had done during the day. At the end of the month the researchers calculated a score for the amount of activity each day and compared this with the number of hours sleep for that day. Was there a relationship between daytime activity and hours of sleep the same night?

7. A researcher selected a simple random sample from among the psychology student population in her University department. One group of participants was asked to perform a task with brief instructions. The other group received no instructions. Time to complete the task was measured. The researcher predicted that the instruction group would perform faster.

8. A researcher recorded whether people were male or female and whether they struck a match towards their body or away from their body. What analysis would determine a sex difference in match striking behaviour?

9. In order to support the theory that people's mood is better on sunnier days participants were asked to rate their mood on a standardised scale every lunchtime for several weeks. Number of hours of sunshine for each day was also recorded and the researchers looked at the relationship between averages for sunshine and mood level.

10. In a research study 20 participants were identified as high extroverts and 23 were identified as high introverts. It was predicted that the high extroverts would score higher on a test of risk taking.

11. Students observe whether males or females do or don't walk under a ladder. They want to see whether one sex is more 'superstitious' than the other. What test do they need to use?

12. A psychologist claims to have a very well-standardised measurement scale. What statistical test would be used to check its test/re-test reliability?

13. Two groups of people are selected. Scores have been used to place people in one of two groups: high 'initiative-taking' and low 'initiativetaking'. They are asked to select just one of three possible activities they would prefer to do. The choices are rock-climbing, dancing or reading a book. What test would demonstrate a significant difference between the choices of the two groups?

14. The time is recorded for the same group of participants to read out loud a list of rhyming words and a list of non-rhyming words. What test is appropriate for showing that rhyming words take less time to read? What test would show that people tend to read at consistent rates?

Use of statistical tables

The critical value

Once a statistical test has been calculated, the result is a *number* – the **calculated value** (or observed value). To check for statistical significance, the calculated value must be compared with a **critical value** – a number that tells us whether or not we can reject the null hypothesis and accept the alternative hypothesis.

Each statistical test has its own **table of critical values**, developed by statisticians. These tables look like very complicated bingo cards (you will see plenty of examples over the next few lessons). For some statistical tests, the calculated value must be equal to or greater than the critical value, for other tests the calculated value must be equal to or less than the critical value.

Using tables of critical values

How does the researcher know which critical value to use? There are three criteria:

One tailed or two tailed test? You use a one tailed test if the hypothesis was directional and a two tailed test for a non-directional hypothesis. Probability levels *double when* two tailed tests are being used as they are more *conservative* predictions.

The number of participants in the study. This usually appears as the *N* value on the table. For some tests **degrees of freedom (df)** are calculated instead.

The **level of significance** (or P value). As discussed, the 0.05 level of significance is the standard level in psychological research.

Lower levels of significance

Occasionally, a more stringent level of significance may be used such as (0.01) in studies where they may be a human cost – such as drug trials – or ‘one-off’ studies that could not, for practical reasons, be repeated in future. In all research, if there is a large difference between the calculated and critical values – in the preferred direction – the researcher will check more stringent levels, as the *lower* the p value is, the more statistically significant the result.

Activity 1: Use of non-parametric tests

A researcher is interested in the relationship between stress and number of days away from work. She develops a questionnaire to assess the number of Life Change Unit's a worker has encountered over the last year and obtains details of absence rates for the same period of time.

a suitable, fully operationalised NULL hypothesis

a suitable fully operationalised two tailed (non-directional) hypothesis

a suitable experimental design (where appropriate)

a suitable inferential test

explain your choice of test

A researcher decides to investigate whether there is a difference in the perceptual abilities of males and females. The researcher sets up two slide shows, both involve a scene in a bookshop, in one slide show familiar items only are presented, in the other show there are a number of 'unusual' items (e.g. a microwave oven; a knife and fork; a calculator etc.) items dotted around the shop.

The research notes down how many of the 'unusual' items are noted by each participant and then compares the number of items recalled by men and women

a suitable, fully operationalised NULL hypothesis

a suitable fully operationalised one tailed (directional) hypothesis

a suitable experimental design (where appropriate)

a suitable inferential test

explain your choice of test

A psychology student believes that males are more likely to cross at a pelican crossing when the pedestrian light is red, than are women.

They set up an observation point at a local pelican crossing and note the number of times men or women cross on the 'red' or 'green' light.

a suitable, fully operationalised NULL hypothesis

a suitable fully operationalised two tailed (non-directional) hypothesis

a suitable experimental design (where appropriate)

a suitable inferential test

explain your choice of test

A psychologist is interested in finding out whether memory is influenced by background music. They set up 2 conditions, in one the participants are given headphones playing soft music and are told to memorise a list of 15 words; in the other condition participants are asked to wear headphones which are silent and are then told to memorise another list of 15 words.

At the end of the experiment the researcher compares the two sets of scores.

a suitable, fully operationalised NULL hypothesis

a suitable fully operationalised one tailed (directional) hypothesis

a suitable experimental design (where appropriate)

a suitable inferential test

explain your choice of test

Activity 2: Accepting or rejecting the null hypothesis

A researcher is interested in the relationship between stress and number of days away from work. She develops a questionnaire to assess the number of Life Change Unit's a worker has encountered over the last year and obtains details of absence rates for the same period of time.

She uses 16 participants to test the hypothesis:- ***'The higher the number of LCU encountered over a year, the higher will be the number of days off due to illness'***

As a result of the above study the researcher carries out a Spearman's ρ correlation test to check whether they can be accepted at the 5% level.

The result of the Spearman's test is 0.41

Can she accept the hypothesis at the 5% level?

Reason:-

A researcher decides to investigate whether there is a difference in the perceptual abilities of males and females. The researcher sets up two slide shows, both involve a scene in a bookshop, in one slide show familiar items only are presented, in the other show there are a number of 'unusual' items (e.g. a microwave oven; a knife and fork; a calculator etc.) items dotted around the shop.

The researcher notes down how many of the 'unusual' items are noted by each participant and then compares the number of items recalled by men and women

The researcher uses an independent groups design of 12 males and 15 females to test the hypothesis:-

'Males will identify more unusual items in a slideshow than will women'

As a result of the above study the researcher carries out a Mann-Whitney U test to check whether the hypothesis can be accepted at the 5% level.

The result of the Mann-Whitney U test is 47

Can the researcher accept the hypothesis at the 5% level?

Reason:-

A psychology student believes that males are more likely to cross at a pelican crossing when the pedestrian light is red, than are women.

They set up an observation point at a local pelican crossing and note the number of times men or women cross on the 'red' or 'green' light.

The student develops the hypothesis:- ***'There will be a significant difference between the number of males and females who cross a pelican crossing on the red light'***

As a result of the above study the researcher carries out a Chi-Square Test to check whether they can be accepted at the 5% level.

The result of the Chi-Squared test is 4.65

Can the student accept the hypothesis at the 5% level?

Reason:-

A psychologist is interested in finding out whether memory is influenced by background music. They set up 2 conditions, in one the participants are given headphones playing soft music and are told to memorise a list of 15 words; in the other condition the same participants are asked to wear headphones which are silent and are then told to memorise another list of 15 words.

The researcher uses a Wilcoxon T test to see if the results are significant at $p < 0.05$; they use 20 participants to test the hypothesis:- ***'People will remember more words if they are listening to music during the learning process than if they are not'***

As a result of the above study the researcher carries out a Wilcoxon T test to check whether they can be accepted at $p < 0.05$

The result of the Wilcoxon T test is 53

Can the researcher accept the hypothesis at the 5% level?

Reason:-

Activity 3: Practice interpreting inferential statistical tests

Are these results significant?

Don't forget, if they are significant, you reject your null hypothesis and accept the alternative one; if they aren't, you retain the null and reject the alternative.

1. Using Spearman's rho

- the observed value of rho is 0.33, $N = 20$, with a one tailed hypothesis
- the observed value of rho is 0.13, $N = 12$, with a two tailed hypothesis
- the observed value of rho is - 0.81, $N = 30$, with a one tailed hypothesis

2. Using Chi square

- the observed value of chi is 1.65, $df = 3$, with a two tailed test
- the observed value of chi is 6.35, $df = 1$, with a one tailed test
- the observed value of chi is 9.49, $df = 2$, with a two tailed test

3. Using Mann Whitney

- the observed value of U is 18.3, $N_1 = 8$ and $N_2 = 10$, with a one tailed test
- the observed value of U is 38.73, $N_1 = 12$ and $N_2 = 12$, with a two tailed test
- the observed value of U is 3.43, $N_1 = 5$ and $N_2 = 8$, with a one tailed test

4. Using Wilcoxon T

- the observed value of T is 14.5, $N = 20$, with a two tailed test
- the observed value of T is 23.5, $N = 12$, with a one tailed test
- the observed value of T is 4.39, $N = 10$, with a two tailed test

Activity 4: Critical Values and how to write a response

Spearman's rho example:-

H_1 = There is a significant positive correlation between test scores and the amount of time spent studying for the test (1-tailed)

H_0 = There is no significant correlation between test scores and the amount of time spent studying for the test

$N = 10$

$r_s = 0.88$

Critical value = 0.564

$p = 0.05$ (significance level)

As the calculated value of rho (0.88) is higher than the critical value (0.564), the null hypothesis can be rejected and the experimental hypothesis can be accepted. It can be reported that there is a significant positive correlation between test scores and the amount of time spent studying for the test ($r_s = 0.88$, $N = 10$, $p < 0.05$, one-tailed).

Mann-Whitney U example:-

H_1 = Children using the maths scheme attain significantly higher scores than children not using the maths scheme (1-tailed)

H_0 = There is no significant difference in scores between children using the maths scheme and children not using the maths scheme

N_1 = 9 (number of people in the smaller group)

N_2 =10

U = 8

Critical value = 24

p = 0.05 (significance level)

Since the observed value of U (8) is less than the critical value (24), the null hypothesis can be rejected and the experimental hypothesis can be accepted. It can be reported that children using the maths scheme attain significantly higher scores than children not using the maths scheme ($U=8$, $N_1= 9$, $N_2=10$, $p<0.05$, one-tailed).

Wilcoxon example:-

H_1 = Participants recall significantly fewer emotionally threatening words than neutral words (1-tailed)

H_0 = There is no significant difference in the number of emotionally threatening words and neutral words recalled

N =9

T = 7

Critical value = 8

p = 0.05 (significance level)

Since the observed value of T (7) is less than the critical value (8), the null hypothesis can be rejected and the experimental hypothesis can be accepted. It can be reported that participants recall significantly fewer emotionally threatening words than neutral words ($T=7$, $N= 9$, $p<0.05$, one-tailed).

Chi-squared example:-

H_1 = There is a significant association between subject studied and personality (2-tailed)

H_0 = There is no significant association between subject studied and personality

$\chi^2=3.22$

df = 1 (degrees of freedom)

Critical value = 3.84

p = 0.05 (significance level)

As the observed value of χ^2 (3.22) is less than the critical value (3.84), the experimental hypothesis must be rejected and the null hypothesis must be accepted. It can be reported that there is no significant association between subject studied and personality ($\chi^2=3.22$, $df= 1$, $p>0.05$, two-tailed).

Now you try....

Look at the information given, use the critical value tables in your book and write up the results as shown above.

Spearman's rho

$N = 16$

$r_s = 0.12$

Critical value =

$p = 0.05$ (significance level for a 1-tailed test)

Accept

Reject

Mann-Whitney U

$N_1 = 5$ (number of people in the smaller group)

$N_2 = 14$

$U = 56$

Critical value =

$p = 0.05$ (significance level for a 1-tailed test)

Accept

Reject

Wilcoxon

$N = 14$

$T = 20$

Critical value =

$p = 0.05$ (significance level for a 1-tailed test)

Accept

Reject

Chi-squared:-

$H_1 =$ There is a significant association between subject studied and personality (2-tailed)

$H_0 =$ There is no significant association between subject studied and personality

$\chi^2 = 1.72$

$df = 6$ (degrees of freedom)

Critical value =

$p = 0.05$ (significance level for a 1-tailed test)

Accept

Reject

Spearman's rho:-

H₁= There is a significant correlation between shoe size and attractiveness rating (2-tailed)

H₀= There is no significant correlation between shoe size and attractiveness rating

N= 7

r_s= 0.52

Critical value =

p = 0.05 (significance level)

.....
.....
.....
.....
.....
.....

Mann-Whitney U:-

H₁= People with pink hair score significantly higher in IQ tests than people without pink hair (1-tailed)

H₀= There is no significant difference in IQ test scores between those with and without pink hair

N₁= 18 (number of people in the smaller group)

N₂=20

U= 128

Critical value =

p = 0.05 (significance level)

.....
.....
.....
.....
.....
.....

Wilcoxon:-

H₁= There is a significant difference in the number of comprehensible words spoken by participants before and after alcohol has been consumed (2-tailed)

H₀= There is no significant difference in the number of comprehensible words spoken by participants before and after alcohol has been consumed

N=6

T= 3

Critical value =

p = 0.05 (significance level)

.....
.....
.....
.....
.....
.....

Chi-squared:-

H_1 = There is a significant association gender and dress wearing (2-tailed)

H_0 = There is no significant association between gender and dress wearing

$\chi^2=9.34$

$df= 9$ (degrees of freedom)

Critical value =

$p = 0.05$ (significance level)

.

.

.

.

.

.

Statistical Tests of Significance

These calculate whether the difference between two sets of results is large enough for us to be confident (at the $p < 0.05$ level) that any changes seen in the dependent variable are due to the manipulation of the independent variable.

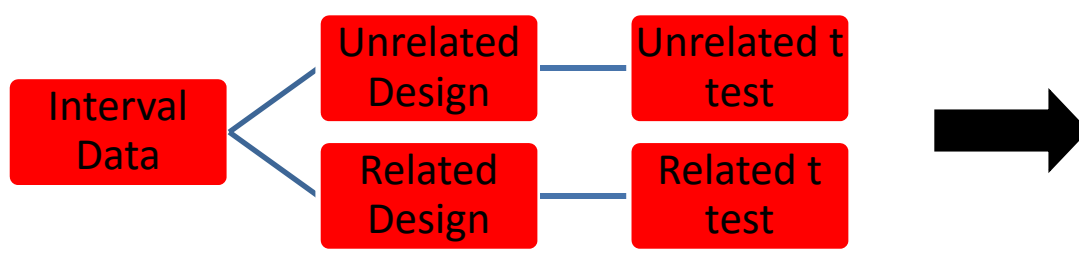
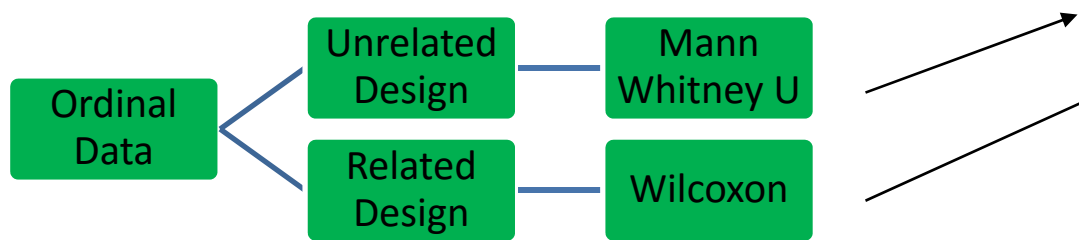
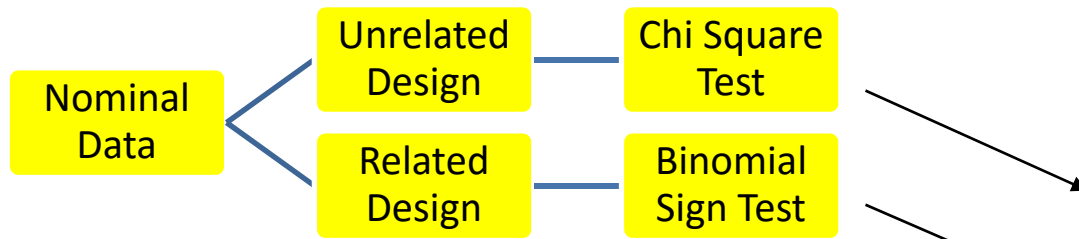
These are the statistical tests that you need to know for the exam in June:

- Chi Square
- Sign Test
- Mann Whitney U
- Wilcoxon
- Related t test
- Unrelated t test
- Spearman's Rank
- Pearson's Product Moment

How do I choose which test to use?

You can only use one of these tests on your data so it's important that you make the right choice. The choice you make depends upon three things:

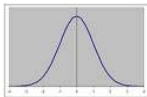
1. Are you testing for a difference or a correlation?
2. Is your data nominal, ordinal or interval?
3. Was your experimental design related (the same participants in each condition) or unrelated (different participants in each condition)



These tests are called non-parametric tests. They are used when you are testing for a **DIFFERENCE** between two conditions.

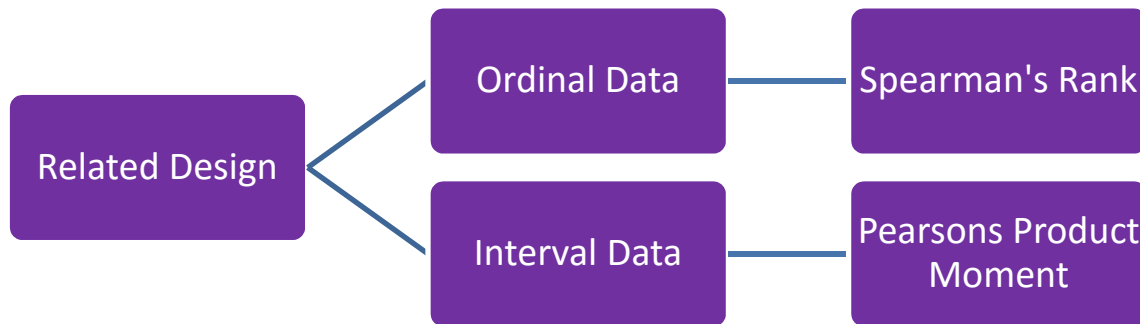
These tests are called Parametric Tests. They are very powerful. You can **ONLY** use them with interval data **AND IF**:

1. The data has homogeneity of variance (the data in both conditions does not vary too much between the participants)
2. The data is normally distributed



They are used when you are testing for a **DIFFERENCE** between two conditions





These tests test for a **relationship**. You would use them if you are looking for a relationship between two co-variables.

In a correlation, your design will only **ever** be a related design.

Your data may be ordinal or interval (but **NEVER** nominal).

The Pearson's Product Moment is classed as a parametric test because it uses **only** interval data.

Activity 5: For each of the following scenarios, decide which statistical test you would use to analyse the data. Justify your answer.

1. Individuals who are depressed often have low arousal and say that they feel time passes slowly. A researcher wanted to compare time estimation by people who are depressed with time estimation by people who are not depressed.

Sixty college students volunteered to participate in the study. Half were male and half were female. They were asked to complete a standardised depression scale, where a high score on the scale indicated a high level of depression. Those above the median score for the total sample were placed in the 'depressed' group and those on or below the median score were placed in the 'non-depressed' group.

All participants then carried out a task. When the task was over, they had to estimate the duration of the task. Participants were tested individually. There were no clocks in the room and the participants removed their watches before entering the room. The researcher wanted the task to be neither pleasant nor unpleasant, as he believed that a neutral experience was an important element of the design. The task involved scanning rows of four letters as quickly as possible, crossing out any letter 'A's. An example is given below:

QPUD DFRK CSAX MJLA
SFWH BUIF AQLP GTTE
XAYO YRAC TGRW AUIF

After five minutes, the participants were asked to stop the task. They were then asked to estimate how much time, in minutes and seconds, had passed whilst they were doing the task.

1. Test: _____

2. Justification: _____

2. Janet is 43 years old. She has received therapy for an obsessive-compulsive disorder. Her symptoms include excessive hand-washing, taking numerous showers and repeatedly cleaning the toilet.

A psychologist conducted a case study of Janet before and after she received therapy. He used a questionnaire to measure the levels of anxiety associated with Janet's disorder. He made observations of Janet's behaviour at her home for two hours per day for one week before and one week after therapy. For each observation, he recorded the number of times Janet washed her hands, took a shower and cleaned the toilet. He calculated the average frequency of these behaviours, before and after therapy.

The results of the psychologist's observations are shown in the table below (**Table 1**).

Table 1: The average daily frequency of Janet's behaviours before and after therapy

Behaviours frequency after therapy	Average frequency before therapy	Average
Hand-washing	40	10
Showering	7	3
Cleaning the toilet	15	2

1. Test: _____

2. Justification: _____

3. A researcher wanted to compare the effectiveness of two therapies for people who had a phobia of flying. A newspaper advertisement was used to recruit a sample of fifty volunteers who were afraid of flying. The participants were randomly allocated to Therapy A (Group 1) or Therapy B (Group 2).

Each participant's anxiety about flying was tested before and after therapy. The assessment involved the use of an Anxiety Scale on which participants were asked to rate how they felt at the time. A high score indicated extreme anxiety and a low score indicated mild anxiety.

For the **pre-therapy assessment** of anxiety about flying, all participants experienced realistic pre-flight conditions in an airport terminal and on board an aircraft. Although participants expected the plane to take off, it remained on the ground. Once the engines were turned off, participants were asked to fill in the Anxiety Scale on board the aircraft.

For the next four weeks, participants attended weekly sessions for either Therapy A or Therapy B. At the end of this period, participants experienced the same procedure as for the pre-therapy assessment and filled in the Anxiety Scale as before. This was the **post-therapy assessment**.

1. Test: _____

2. Justification: _____

4. A major bank's call centre introduced the playing of music to its customers whilst they were waiting for their telephone call to be answered. A psychologist wanted to find out whether or not the playing of music affected the length of time that customers were prepared to wait for their telephone call to be answered. A questionnaire was sent to 150 customers. They had all used the service before music was introduced to the telephone line. In the questionnaire, they were asked whether they were now prepared to wait on the telephone for a shorter time, the same length of time or a longer time.

Sixty customers returned the questionnaires. The data in **Table 1** summarise their replies.

Table 1: The number of customers who replied that they were now prepared to wait for a shorter time, the same length of time and a longer time

Waiting times	Number of customers
Shorter time	12
The same length of time	25
Longer time	23
Total number of customers	60

1. Test: _____

2. Justification: _____

5. A health psychologist wanted to investigate whether there was a relationship between workplace stress and the number of days absent from work.

The psychologist obtained a random sample of thirty nurses from a nearby hospital. Their ages ranged from 21 to 60 years.

The psychologist obtained the number of days each nurse was absent from work in the month of April. This information was obtained from their personnel files.

The psychologist interviewed each nurse. As part of the interview, each nurse completed a psychological test to measure his/her stress level. A high score on the test indicated a high level of stress and a low score, a low level of stress.

1. Test: _____

2. Justification: _____

Overview of inferential statistics

Inferential statistics are used by psychologists (and other researchers) to determine the likelihood that an observed effect is due to chance. By 'observed effect' we mean a difference between two sets of scores, an association between variables or a correlation. The statistics allow the researchers to *make inferences* about the populations from which the samples are drawn.

You need to know which test is appropriate in different circumstances. One of the decisions to make is whether to use a **parametric** or **non-parametric test**.

Parametric tests are powerful statistical tests meaning that they are better able to detect a significant effect. This is because they are calculated using the actual scores rather than the ranked scores. However, this sensitivity can also be a problem if the data is inconsistent or erratic.

Criteria for parametric tests

- the tests should only be used on data of interval status
- the data will come from a sample drawn from a **normally distributed** population
- there is **homogeneity of variance** between conditions



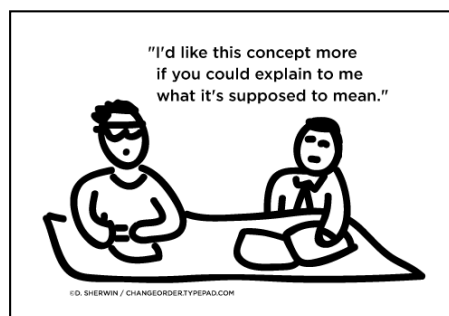
Normally distributed: a set of data distributed so that the middle scores are most frequent and extreme scores are least frequent

Homogeneity of variance: the deviation of scores (measured by the range or standard deviation for example) is similar between populations

Parametric tests are also robust, meaning they are able to cope with data which do not fully meet the three criteria. The only essential criterion is that the data must be *interval level*.

If this criterion is not met then a non-parametric test needs to be used instead. Non-parametric tests are calculated using ranks, which means they are less sensitive but better able to cope with any inconsistency in the data.

You do not need to know **how** to carry out these statistical tests, but you do need to know the criteria for choosing each test and how to read the statistical tables. Quite simply you need to **learn this information**.



Statistical test table

	Non-parametric Test	Non-parametric Test	Parametric Test
Experimental Design	Nominal Data	Ordinal Data	Interval Data
Repeated Measures and matched pairs	Sign Test	Wilcoxon Test	Related T Test
Independent Groups	Chi-Squared Test	Mann-Whitney U Test	Un-related T Test
Correlations	Chi-Squared Test	Spearman Rho Test	Pearson's Product Moment

- The first step in working out which statistical test should be used is finding out **what experimental design** has been used.
- Experimental Design refers to the way the participants are used in research:
 - If a participant's before and after scores that are being used and the participant is used twice, this is a **repeated measures design**.
 - If the participants are only used once in that one group's response is compared with another group, this is an **independent group design**.
 - If an independent groups design is used but the participants in one group are matched with participants in the other group, in terms of specific characteristics, e.g. intelligence, this is a **matched pairs design**.
 - If no participants are being used as such, but their two sets of scores are, this is a **correlation design**.
- Once you have found the experimental design, you have to work out **what type of data has been generated**.
 - **Nominal data is data that can be put into specific categories**, e.g. aggressive, non-aggressive. The information can be presented in bar charts where the bars do not touch because the data is separate.
 - **Ordinal data is data that is continuous**. It can be used to represent test scores, e.g. in a memory test, the scores may be 0-10, 11-20, 21-30. The numbers continue. The information can be presented in a histogram where the bars touch. This is because the data is continuous.
 - **Interval data is also continuous, but it represents things that are set in stone that the researcher cannot manipulate**. E.g. time, weight, length etc. E.g. a minute is always a minute and there is nothing the psychologist can do to change this.

