Second year research methods and statistics pack

Name-

Teacher-

Contents

Assessing and improving reliability and validity

You will have used validity and reliability as a way of evaluating research in the first year but in the second year you need to know how assess and improve on them.

Reliability-"think consistency" -the extent to which a test or measurement produces consistent results i.e. if a test is repeated using the same method, design etc. and gets the same results every time then it can be said to be reliable.

Internal reliability-extent to which something is consistent within itself e.g. a set of scales should measure the same weight between 50 and 100 grams as between 200 and 250 grams.

External reliability-extent to which a test measures consistently over time.

Assessing the reliability of observations

Inter-observer reliability-you should know this one! This is a way of assessing how reliable an observation is. You always need to have two or more observers carrying out an observation to ensure reliability. You compare their observation schedules or records to see if they are similar by correlating the observer's scores. If there is a correlation of 0.8 (or 80% similar findings) then you can say you have inter-observer reliability, i.e. that observers are observing and categorising behaviour consistently.

Improving observational reliability

- Always have more than one observer
- have clearly defined (operationalised) and separate observational criteria.
- train observers so they know exactly what to look for
- Do a pilot study so you know the observers are applying the observational categories properly.

Assessing the reliability of self-report techniques

Questionnaires and interviews are measuring an aspect of a person so we want to be sure that the measurement is reliable i.e. we would obtain the same set of answers or the same score on a personality test every time we took it. Your score may vary. For example if you take a mood test and score highly one week when you've just got an A on your mock but then score low the next week when you fall out with your best friend this is natural but we need to be sure that any change is down to the person and not that the test is unreliable.

Split-half method-assesses INTERNAL reliabiltiy by splitting a test in two and making the same participants do both halves. If the results are the same for both halves then it indicates the test has internal validity.

Test-retest method-assesses external reliability by giving participants the same test on two occasions with normally a week or two apart so they don't remember the answers. If the results are the same then external reliability is established. This can also be done with interviews to test the reliability of the interviewer.

Improving reliability

- Make questions clear and precise (closed are more reliable than open for example)
- Pilot a questionnaire beforehand to check if the questions are clear enough
- Use the same interviewer with each participant or fully train if using more than one

Assessing the reliability of experiments

Lab experiments are thought to be the most reliable as they can have the most carefully controlled procedure, instructions, conditions. So when looking at research you want to see if these things are all carefully controlled and that participants were all tested under the same conditions.
Also reliability could be to do with the measurement of the DV. For instance in Rutter's research they used IQ tests as one of the ways to measure Romanian orphans progress so we would be looking at the reliability of these tests.

Improving reliability

- Use exactly the same procedures for all participants
- Use the same conditions for all participants.
- If repeated by other researchers they need to replicate the research in exactly the same way as the original

For the exam remember……

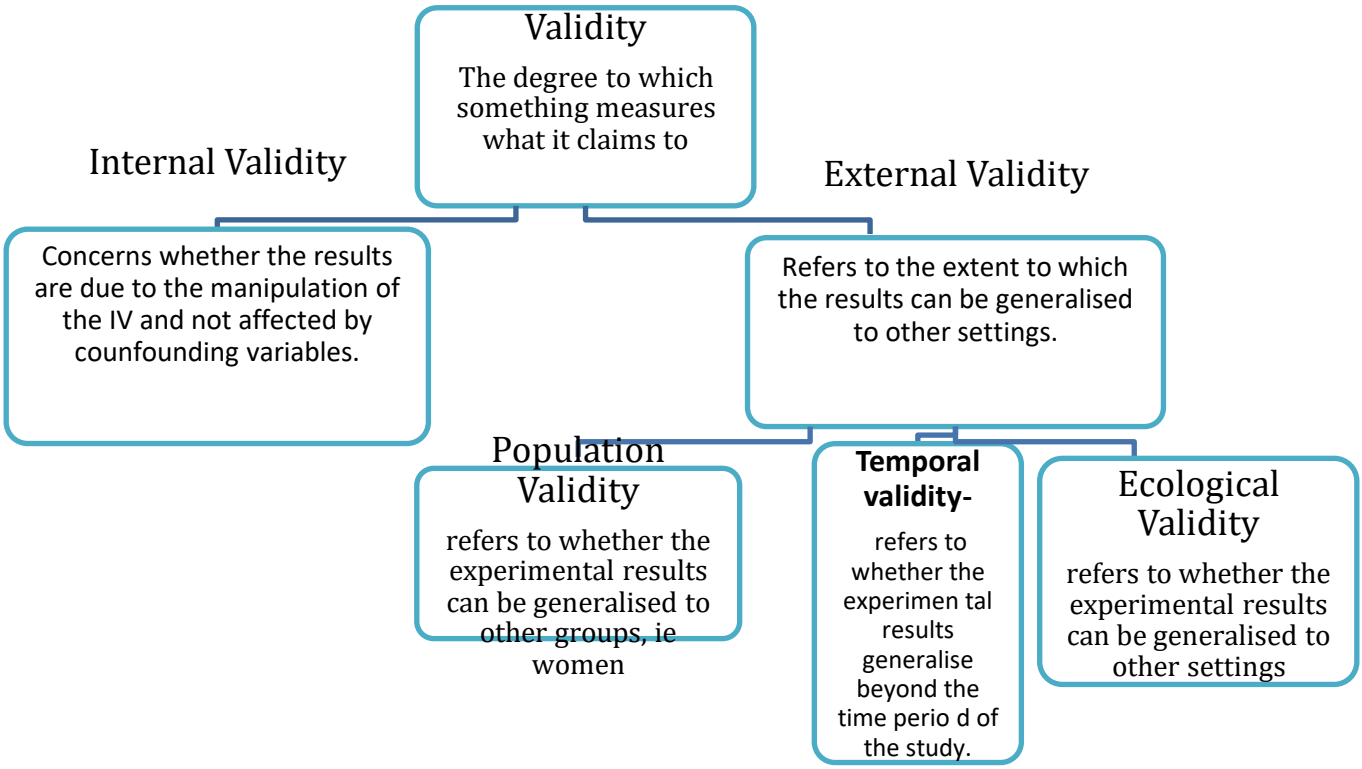
Unreliable results cannot be trusted (see replicability)

Results can be reliable but not valid. If you add up 1+1 100 times and get 3 each time then this is reliable but is still not valid as is added up incorrectly every time.

Stretch yourself- Can you?

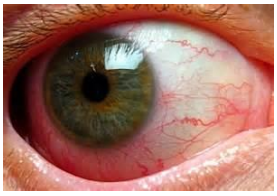
1. Define reliability (1 mark)
2. Explain how you can assess whether a set of data from a questionnaire is reliable (3 marks)
3. Explain how inter-observer reliability is assessed (3 marks)
4. Explain how inter-observer reliability can be improved (3 marks)
5. Explain how you can assess and improve on the reliability of an experiment (2+2 marks)

Validity-"think legitimacy or accuracy" Remember this diagram from the first year?



Assessing validity

Imagine we are testing whether men or women are more stressed at work, using a questionnaire to measure stress. We need to know if the questionnaire is actually measuring stress, maybe the questions are designed so badly that people don't' really understand what to put so are guessing for example.



Face validity-does the self-report measure, in this case the stress test, look like it is actually measuring stress? This is just a quick "eyeballing" or intuitive measure whereby you just look over the questionnaire (or pass to an expert) to see if on the face of it, it looks like it's measuring stress.

Concurrent validity-This is when you use a well-established, validated test to compare with your new test. So in this case you'd get participants to take your stress test and a well-established one at the same time. If participants got similar scores then this confirms the concurrent validity of your test.

Improving validity

Questionnaires

- Review questionnaires or tests if when assessed they have low face or concurrent validity
- Assure responses are anonymous

Experimental research

- Use a control group so that the researcher is better able to assess whether changes in the DV were due to the IV
- Standardise procedures to reduce investigator effects and participant reactivity.
- Reduce demand characteristics by using double blind (research doesn't know the aims) or single blind (participants don't know the aims) procedures.

Observations

- Observations tend to have high ecological validity especially if covert.
- Make sure that behaviour categories are not too broad, overlapping or ambiguous.

Qualitative methods

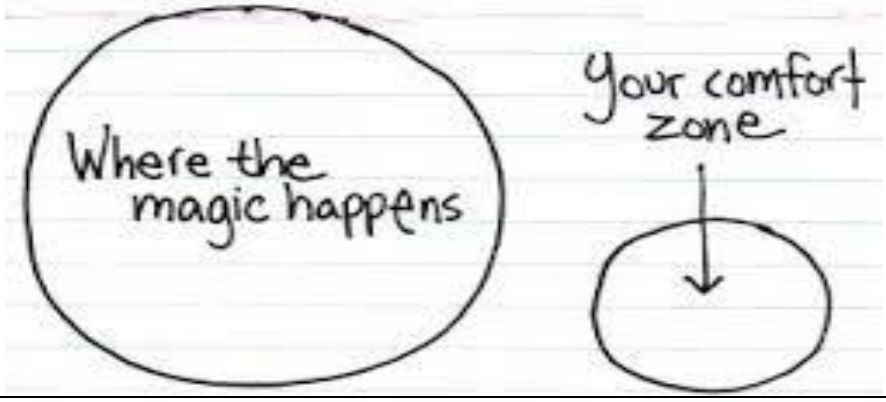
- Interviews and case studies that produce qualitative data are said to be higher in ecological validity because of the depth and detail involved being better able to reflect the participants reality.
- Interpretive validity (the extent to which a researchers interpretation of events matches the participant) can be demonstrated by using direct quotes and being coherent in reporting.
- Triangulation can be used to improve validity by using a number of different sources as evidence for example interviews with friends, family, personal diaries, observations etc.

For the exam remember……

The specification specifically mentions face and concurrent validity so you must be a prepared for a question about these alone

Stretch yourself- Can you?

1. Define validity (2 marks)
2. Explain what is meant by face validity and concurrent validity (2+2 marks)
3. Explain the term ecological validity using an example from a research study (3 marks)
4. Explain why temporal validity might be a problem in research (2 marks)
5. Explain two ways in which validity can be assessed (2+2 marks)
6. A research wants to improve the validity of their interview, explain two ways this can be achieved (2+2 marks)



Content analysis

A content analysis is a bit like doing an observational study but instead of observing actual people a researcher makes their observations indirectly through books, films, adverts, photos, songs, diaries etc. In fact content analysis is the analysis of the content of any artefact.

The researcher will make *three* decisions.

1. Sampling method-what to use, for example choosing which channels to watch, for how long, what length of time. If analysing book content then do you look at every page, or say every fifth page?

2. Coding the data. What behavioural categories need to be used? For example, if a researcher was performing a content analysis from the diaries of someone with depression, they need to develop specific categories and tally each time they are reported in the diary. Decisions about behavioural categories may involve a *thematic analysis (see below).*

3. Method of representing data-Should the data be quantitative, so you count the number of times a person's diary mentions feeling sad? Or should it be qualitative where you would describe themes so pull out descriptions of passages where the person says they have felt sad.

Thematic analysis

This is a qualitative analytical method for organising, describing and interpreting data. It is a very lengthy process as is painstaking and each item is gone through repeatedly and with careful consideration.

There are many ways to do it but one is detailed below

General principles	Applied to Finnish study (2005) on adolescents' peer and school experiences using interviews	Analysis of graffiti
1. Read and reread the data, become immersed in the content, don't make notes	Read and re-read the interview transcripts, in this case 234 pages of notes!	Study the photographic or written record of a wide range of graffiti
2. Break the data into meaningful units- small bits of text which are able to independently convey meaning e.g. sentences or phrases	All the answers to the questions e.g. How is your family involved in your school activities? Were put together and then each statement was compressed into a briefer statement.	Each item of graffiti would be a unit.
3. Assign a label or code to each unit. These codes are your initial behavioural categories. You will	Each compressed statement was given a label such as "parental help" "siblings help".	Each unit of graffiti is given a code to describe its meaning such as "humour", "advice", "love".

have developed some ideas whilst reviewing the data in step one.		
4. Combine simple codes into larger categories/themes and then instances can be counted examples given.	The categories were grouped into larger units producing eight main categories. For example; **Enablement**-"yeah, ever since my childhood we've always had lots of kids visiting" (girl, 15 years) **Negligence**-"My sister is not at all interested in my friends" (girl, 16 years).	Larger order categories are developed which combine units such as "interpersonal concerns".
5. A check can be made on the emergent categories by collecting a new set of data and applying the categories). They should fit the data well if they represent the topic area investigated.		

Evaluation

Strengths

- Tends to have high ecological validity-because it is largely based on what people actually do with real communications that tend to be current and relevant such as newspaper articles.

- Establishing reliability is easy and straightforward. Of all the research methods, content analysis scores highest with regard to ease of replication. Usually the materials can be made available for others to use

Limitations

- Purely descriptive- so does not reveal underlying reasons for behaviour or attitudes etc. Gives us the what but not why

- Lack of cause and effect-as not performed under controlled conditions with extraneous variables like observer bias a problem due to interpretation of the meaning of the behavioural categories then causality cannot be established.

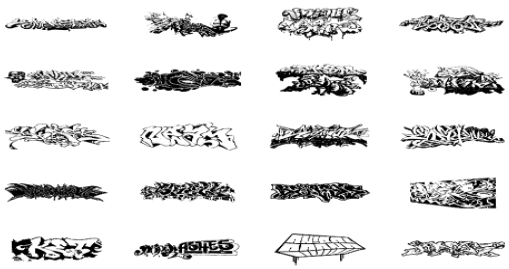
Stretch yourself- Can you?

1. Explain what is meant by content analysis (2marks)
2. Explain how observer bias might affect the findings of a content analysis (3 marks)
3. Briefly outline what is involved in thematic analysis (3 marks)
4. Give one strength and one limitation of content analysis (2+2 marks)

For the exam-remember!

Immerse yourself

Break the data up into meaningful units



Label or code the units

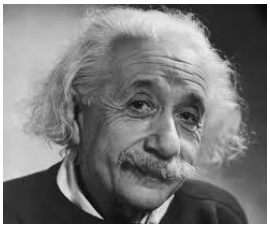
"advice"

Create larger categories from the units

"Interpersonal concerns"

Check your larger categories using new data

Features of science



"No amount of experimentation can ever prove me right; a single experiment can prove me wrong" Albert Einstein

So what makes science scientific?

Science is a system of acquiring knowledge through a process known as the scientific method or process which is defined as the observation, identification, description, experimental investigation and theoretical explanation of phenomena. The prime feature of science is one that is dependent on **empirical** methods of observations i.e. is based upon sensory information rather than simply on thought and beliefs. Basing a theory on an idea you've suddenly had whilst dreaming last night obviously isn't very scientific!

Science involves making predictions (**theory construction**) tested by scientific, **empirical** observations. Such observations are made without bias or expectation by the researcher (**objectivity**) and are performed under controlled conditions. In this way theories and hypothesis are validated (found to be true) or **falsified** (found to be untrue) using **hypothesis testing**. This is repeated and more evidence collected for a theory to be accepted **(replicability).**

The key features of science are:

- Objectivity (and the empirical method)
- Replicability and falsifiability
- Theory construction
- Hypothesis testing
- Paradigms and paradigm shifts

Objectivity

A good scientist should be objective – the findings of their research should not be influenced by any biases that they hold or exert. In other words, they must keep a "critical distance" during research. To lessen the possibility of this unconscious bias researchers aim to-

-use standardised instructions
-fully operationalise variables
-use the double-blind technique
-use carefully controlled conditions

If research is objective it increases our confidence in the results of the research and it also increases the replicability of the research since bias did not contribute towards it. Laboratory experiments tend to be the most objective as they are associated with the greatest level of control. ***Objectivity is the basis of the empirical method*** and whilst most bias is unconscious there have been incidences of fraud. Peer review and replication are important to prevent the publication of such unscientific and flawed research as there can be disastrous consequences when flawed research impacts on real life. An example of such fraud was the false claim that the MMR (measles/mumps/rubella) vaccine was linked to autism. The researcher however had falsified results, in fact there was no link, and it was discovered that he was

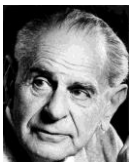
being paid by lawyers who wished to sue vaccine companies. Scan the QR code to find how this lack of objectivity and scientific fraud has had a terrible impact on public health.

Replicability

This relates to the reliability of the findings: so, if it is possible to carry out the research again and find the same or similar results, the research is replicable. If it is replicable we can have confidence in the findings and it increases our "trust" in the findings. To enable others to replicate a study, psychologists should publish full and precise details of their research. If research is replicable it guards against scientific fraud (for instance, researchers may have simply made their findings up or made a mistake-see real life example below) and allows us to rule out that the finding was a one off caused by something about the original study, such as an atypical sample being tested.

Fleischmann (1989) claimed to have created cold fusion (a way of creating abundant cheap energy) however replications did not get the same result and they realised they had made an error in there procedure that only by replication they were able to realise.

Falsifiability



Karl Popper (1934) argued that the key criterion of a scientific theory was its falsifiability. A genuine good scientific theory should have the possibility of it being proven false and should stand up to hypothesis testing (i.e. falsifiability really means can a hypothesis be proven wrong). Popper believed unfalsifiable theories were unscientific and called them pseudoscience, for example the study of paranormal activity such as ghosts or the belief in the existence of god. How can you ever prove these to be false?
Theories that survive attempts to falsify them become the strongest theories but not because they are necessarily true but because they have not been proved false. This is why in your Psychology essays you never write "this proves that the theory is correct" but instead "this adds supports for the theory".

A theory is a general set of laws or principles that have the ability to explain particular events or behaviours.

Theory construction

Science tests theories. Initial observations (that are carefully arranged and unbiased) yield up information about the world which is then made into theories that try and account for this information. Predictions in the form of testable hypothesis are formulated and tested, producing data that can be statistically analysed to see if the theory can be refuted or falsified.
There are two schools of thought as to how you should construct your theory. Should it be straight after you observations **(deductive model)** or at the end once you have tested your hypothesis and drawn conclusions **(inductive)?**

Observations
↓
Testable hypothesis
↓
Conduct a study to test hypothesis
↓
Draw conclusions
↓
Propose a theory

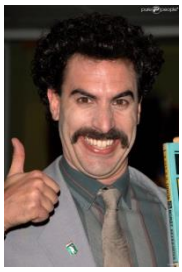
Inductive

Observations
↓
Propose a theory
↓
Testable hypothesis
↓
Conduct a study to test hypothesis
↓
Draw conclusions

Deductive

Real life example of the *inductive method* of theory construction

Simon Baron-Cohen (yes he is related to Sacha Baron-Cohen, he's his first cousin) is a British psychologist who has carried out much research into autism, which is a severe developmental disorder that causes difficulties in social interaction. He has developed a theory about autism and the male brain, the development of this theory is shown below.



Unbiased observations

Autism is more common in males and Baron Cohen studied theory of mind, which is the ability to realize that other people think and feel different things to you. He noted that autistic children did not have a theory of mind causing them to lack empathy. Many autistic children have unusually strong obsessional interests, for example in railway timetables. Baron Cohen also noted the different ways that boys and girls who are not autistic play: many girls seem person oriented in their play - they play with dolls and teddy bears as if they are real people whereas many boys are more interested in toy cars and collecting things such as football stickers.
This led Baron Cohen to initial ideas that there is something different about the ways that males' and females' brains work, and that this difference is in evidence very early on.

Testable hypothesis

One piece of work that he supervised (it was actually carried out by Jennifer Connellan and Anna Batkti) had these hypotheses:

One day old baby girls will spend more time looking at a human face than a mechanical object.
One day old baby boys will spend more time looking at a mechanical object than a human face.

Conduct a study to test hypothesis

To test these hypotheses babies saw Connellan's face and a mobile (hanging toy) over their crib. Connelan was not told the gender of the babies; the babies were videoed so it was possible to tell where they were looking. The tapes were then analysed to see how long the babies looked at the face and the mobile and only then was the gender of the babies revealed. Both hypotheses were supported by the results of the study.

Theory construction

This study, combined with much other work, led Baron Cohen to develop his empathizing-systemizing theory. It states that the female brain is predominantly hard-wired for empathy, which is the cognitive skill of identifying another person's emotions and thoughts, and the affective aspect of responding to these with an appropriate emotion. The male brain is predominantly hard-wired for systemizing (understanding and building systems) which refers to skills such as finding out how systems work, predicting them or inventing new ones. Baron-Cohen describes autism as the extreme male brain because autism involves minimal empathy and maximum systemizing. The theory hypothesizes that systemizing gave an evolutionary advantage to male hunter gatherers and empathizing gave an evolutionary advantage to female carers.

Example of how to reference a book

Reference
Baron-Cohen, S. (2003) *The Essential Difference, London,* Penguin

Want to know more? Thinking of doing Psychology at University? Want to stretch yourself? This qr code links to a lecture by Baron-Cohen on his work into autism and the male brain. It is really fascinating, I promise!

Scan below for clip that explains paradigm shifts using pom poms

A **paradigm** is a shared assumption about the subject matter of a discipline and the methods of study (Kuhn 1962)

Paradigms and paradigm shifts

Kuhn (1962) argued that induction and deduction was not how science worked. Instead he argued that scientific advancement occurs not through steady progress but instead by revolutionary paradigm shifts.

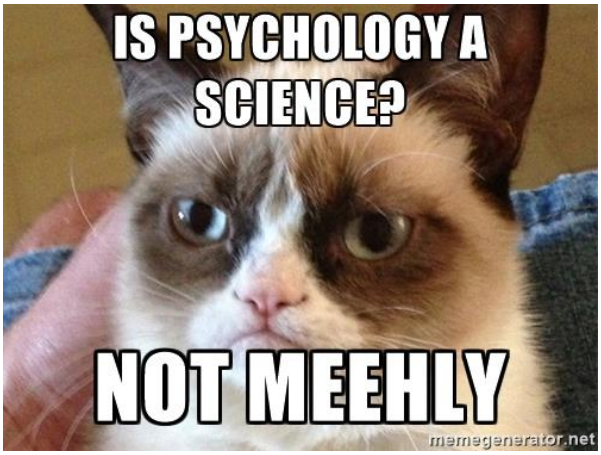
Kuhn argued that a particular science had one paradigm at a time, and that at any time one particular paradigm is believed by all scientists working in the field. He also stated that scientists will go to great lengths to defend their paradigm against falsification, by adding extra bits on to the theory to explain any anomalous data that may have been found. Alternatively, the failure of a result of research to conform to the paradigm is seen not as a problem for the paradigm, but as the mistake of the researcher in carrying out flawed research. So, changing paradigms is difficult because it is changing everyone's belief system and it requires an individual scientist to break with his or her peers.

A change of paradigm, or scientific revolution, can occur when the evidence is mounting up against the existing paradigm; this is known as a paradigm shift. The new paradigm can explain not only all the findings that fitted the previous paradigm but also all the findings that did not fit in with it. Paradigm shifts can be swift but they do require scientists to significantly change their thinking about their subject.

An example of a paradigm shift would be Darwin's theory of evolution. Previously it was believed god created all of the universe and all the animals within it but Darwin showed that humans are part of nature, not above it, and that all animal life, including human, is related by descent from a common ancestor.

So is Psychology a science then?

Kuhn says no! He states that there are fields of enquiry called pre-science, where no one agreed paradigm has yet emerged; he argued that these fields of enquiry would become a science when a central paradigm emerged. He believed that psychology was in fact a pre-science because it has yet to establish its paradigm as there are a number of theoretical perspectives in Psychology that have different ideas and ways of investigating the human subject. Some people argue that this isn't the case and that Psychology has a number of paradigms such as behaviourism, cognitive psychology etc. Also some believe Psychology has already gone through several paradigm shifts from Wundt's early structuralism to the dominant cognitive neuroscience model of today.
Regardless of which stance you take it clear that by adopting a scientific model of enquiry Psychology gives itself greater credibility and on an equal footing with other more established sciences.



For the exam remember……

Science is dependent on empirical methods so empiricism is at the heart of science

The key features of science are (taken straight from the specification)

- **Objectivity (and the empirical method)**
- **Replicability and falsifiability**
- **Theory construction**
- **Hypothesis testing**
- **Paradigms and paradigm shifts**

Stretch yourself- Can you?

1. Briefly explain **one** reason why it is important for research to be replicated (2 marks).
2. Explain how theory construction is a feature of science (3 marks)
3. Describe what a paradigm shift is (2 marks)
4. Explain how paradigm shifts contribute to scientific understanding (3 marks)
5. Briefly explain one reason why it is important for research to be objective (2 marks)
6. Explain two ways in which the objectivity of a piece of research can be increased (2+2 marks)
7. Explain what is meant by the empirical method. Refer to an example of psychological research in your answer (3 marks)
8. Explain what falsifiability means (2 marks)
9. Briefly explain one reason why it is important for research to be falsifiable (2 marks)

Inferential testing

Introduction to statistical testing

There are two kinds of statistics: descriptive statistics and inferential statistics. Descriptive stats give us convenient and easily understood summaries of the data e.g. graphs, averages but we can't draw any firm conclusions from them, they are just an overview. In order to draw firmer conclusions and to accept or reject hypothesis inferential statistics are needed.

So what are inferential statistics then?

Inferential statistics are tests that allow judgements (inferences) to be made about whole populations based on just the sample used in a study. We can't obviously test everybody in the world so inferential statistics is a way of dealing with this issue.

Imagine you are testing the effect an energy drink has on participants, data is collected for each condition (one group has the drink, one doesn't, or the same participants are tested with and without the drink). From casually looking at the results for the two conditions the psychologist may suspect that they have found a genuine effect, people in the energy drink condition spoke more; however, the differences between the results of the two conditions may simply have been caused by chance (or sampling error) rather than a genuine effect i.e. the energy drink. Using inferential statistics allows us to see if there **actually is** a genuine effect and if there is then we can generalise to other people. If the results are judged to be caused by a genuine effect, they are called **significant.**

It's all about the Null hypothesis

Inferential tests allow us to reject or retain (keep) the null hypothesis. Null hypotheses state that there is no difference between the two conditions and that any difference is due to chance factors. If the null hypothesis is retained the research hypothesis (also known as the alternative hypothesis) must be rejected. If the null hypothesis is rejected, the research hypothesis is retained.

Null- There will be no difference in the amount of words recorded during a 3 minute group logic problem task if participants have just finished drinking a litre of water or a litre of energy drink.
Research-There will be a difference in the amount of words recorded during a 3 minute logic problem task if participants have just finished drinking a litre of water or a litre of energy drink.

N.B For a correlation, the relationship between the two sets of data is looked at. The null hypotheses would state that there is no correlation and that any relationship is due to chance factors.

Breaking down the complex language-it's not complicated just looks it!

So you have counted the amount of words recalled in each of your conditions and although it looks like the energy drink has had an effect (mean average words for without drink 123 and with drink 262) you can't say it has until you do an inferential test. You do the test (don't worry about how just yet) and get an answer of 12. This answer is referred to as the **observed value** or **calculated value.**

So what do you do with this answer?

If the results have "worked" i.e. energy drink has made people speak more then we call it **significant,** if it doesn't "work" then it is **not significant.** If the result is **not** significant (i.e. the energy drink caused no difference) the null hypothesis is retained (kept), If the null hypothesis is rejected the result **is** significant i.e. participants who drank the energy drink were recorded to use more words.

How do you know if it significant though?

The **observed or calculated value** (i.e. the answer, for this example 12) is compared to something called the **critical value** to judge the probability that the result occurred by chance i.e. to see if it worked or in Psychological terms to see if it is **significant.** For some tests the observed value must be equal to or more than ≥ the critical value (Chi-squared, Spearmans rho, related t-test, unrelated t-test, Pearson's r) for some it must be equal to or less than ≤ the critical value (sign test, Wilcoxon, Mann Whitney).

What on earth is a critical value then? Where does it come from?

Critical values are found in critical value tables that exist for each statistical test (see critical value tables at the end of the pack). Psychologists use them to compare against their answer **(the observed value)** to be able to see if their findings are significant or not. Don't worry yourself too much about where they come from, just think that some people who really love statistics have sat down and worked out what the observed value (answer) must be for each test to judge the answer as significant!

How do you read a critical value table? How do we compare it with the observed value to decide if it is significant?

Table of critical values of the sign test (S)
Calculated value of S must be ≤
than the critical value to be significant.

Level of significance for a one tailed test				
	.05	.025	.01	.005
Level of significance for a two-tailed test				
	.10	.05	.02	.01
N				
5	0			
6	0	0		
7	0	0	0	
8	1	0	0	0
9	1	1	0	0
10	1	1	0	0
11	2	1	1	0
12	2	2	1	1
13	3	2	1	1
14	3	2	2	1
15	3	3	2	2
16	4	3	2	2
17	4	4	3	2
18	5	4	3	3
19	5	4	4	3
20	5	5	4	3
25	7	7	6	5
30	10	9	8	7
35	12	11	10	9

To read a critical value table you need to know-

1. Your level of significance -across the top of the table (we'll explain what this is later).

2. How many participants you used, this is known as N (the column down the left hand side)

*this is slightly different for unrelated data see** below

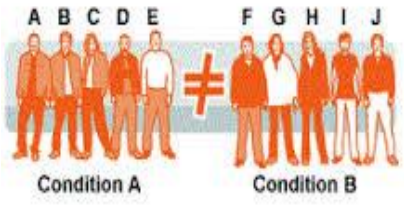
3. Whether your test was one tailed (directional) or two tailed (non-directional).

You simply find your value of N and your level of significance and where the two meet on your table is your critical value

Example one- You have 20 participants, a one tailed test at 5% significance (0.05) then your critical value is **5**

Example two- You have 9 participants with a level of significance of 1% (as is a socially sensitive study) and a non-directional hypothesis. Your critical value is? **0**

**Critical value tables for independent data (i.e. independent groups design) are read in a slightly different way (T-test, Chi squared, Mann-Whitney). When the data is related (repeated measures) then the same participants are used in each condition so you will always have the same value of N for each condition. However if unrelated you might have different amount of participants in each condition so need to do it differently just in case.



For the **Mann-Whitney** test you just call the number of participants in the first condition N1 and second condition N2 and you find the value of N1 along the top and N2 down the side and where they meet that's your critical value.

The unrelated T-test and **Chi squared** have degrees of freedom (df) instead of an N value. For the unrelated T-test this is worked out using the equation Df=(N1+N2)-2.

Chi Squared has degrees of freedom calculated slightly differently (see statistics example three).

Critical value table for the T-test with a two tailed hypothesis. T ≥critical value of T to be significant		
df	0.1	.05
1	6.314	12.706
2	2.92	4.303
3	2.353	3.182
4	2.132	2.776
5	2.015	2.571
6	1.943	2.447

Working out a critical value for the unrelated T-test

1. Find your level of significance

2. Your degrees of freedom worked out using

Df=(N1+N2)-2

3. Whether your test was one tailed (directional) or two tailed (non-directional).

You simply find your value of df and your level of significance and where the two meet on your table is your critical value

Example one- you have 11 participants in condition1 and 13 in condition 2, with a two tailed test at 5% level of significance. Your Df is (4+3)-2=5
Your critical value is **2.571**

Probability

Even if your results do "work" we are not saying that we are 100% sure that the findings are significant i.e. that drinking one litre of energy drink makes you talk more. We are saying that we are pretty sure. This is because there may be chance factors i.e. maybe you have a person in the energy drink condition who just never shuts up talking anyway and the energy drink has not caused this. We work this out using probability.

Psychologists have to decide how certain they want to be that the results are not just down to chance. Normally psychologist say they want to be 95% sure which means there is a 5% chance that the findings weren't because of the energy drink but that they just occurred by chance. We call this level of chance our **level of significance**.

However, when a piece of research is socially sensitive or considering something like the effects of a new drug on health for instance, it is particularly important not to accept the study as significant if it's not and was caused by chance alone. In this instance a 1% level of significance is likely to be used. This means that

we are only prepared to accept a 1% chance or less of wrongly rejecting the null hypothesis. Imagine the severe consequences of incorrectly accepting a new drug to be safe for humans to use when actually it is not! That's why in these cases we need to be almost 100% sure.

Remember
< means less than, **≤** means equal or less than, **<<** means a lot less than.
> means more than, **≥** means equal or more than, **>>** means a lot more than.

Level of significance

So the level of significance is the point at which the researcher can claim to have discovered a significant difference (or correlation) within the data and can reject the null hypothesis and accept the alternative hypothesis.
In psychology our level of significance is usually 5% but is written as a decimal and referred to it as **p.**

p can be any value between 0 and 1 but is normally **p** ≤ 0.05

p=0.05 means that there is a 5%, or a 1 in 20 chance, or less, of the results having occurred by chance alone. This means that statistically speaking, if this result is obtained 20 times, in 19 of those studies, the difference will be caused by the IV, (or there is a relationship between two or more variables), whereas in one of the studies, the difference/correlation will have occurred through chance alone.

How is p expressed?

If p < 0.05 this means that the probability of the null hypothesis being true is less than 5 in 100, or is less than 5%.

If p ≤ 0.025 this means that the probability of the null hypothesis being true is equal to or less than 2.5 in 100, or equal to or less than 2.5%.

If p << 0.01 this means that the probability of the null hypothesis being true is a lot less than 1 in 100, or a lot less than 1%.

Type 1 and Type 2 errors

By using significance levels, we are always at risk of either rejecting the null hypothesis when it is true or retaining the null hypothesis when it is false. This is why scientific evidence can never be taken as *fact*, and theories are *never proved*.

A **type I** error occurs when a null hypothesis is rejected when it should not have been. In other words, they have accepted their results as significant when, in fact, they are down to chance alone. The likelihood of a type 1 error mirrors the level of significance employed. A type 1 error is sometimes known as a **false positive**.

A **type II** error occurs when a null hypothesis is retained when it should not have been. In other words, we have accepted that our results are down to chance, when in fact they are not. A type 2 error is often referred to as a **false negative**. It is more likely to occur with a higher level of significance (e.g. p=0.01).

We are more likely to a type I error is we are too **lenient** with the significance level e.g. 10% rather than 5%. A type II error is more likely to occur if we are too stringent e.g. 1% or 0.01. A way of trying to prevent these errors is to not be too lenient or stringent and so that is why we use significance level of 5% or 0.05.

The pregnancy test says you are pregnant but you are not-an example of a type I error!

For the exam remember……

Inferential statistics allows us to make judgments about whole populations based on the sample used in a study

If a study "works" we say it is **significant**, if it doesn't "work" we say it is **not significant**

If research is **significant** we **reject the null hypothesis** and accept the alternative hypothesis

If the research is **not significant** we **reject the alternative hypothesis** and accept the null hypothesis

The answer to a statistical test is called the **observed value** or **calculated value**

We decide if the result (observed value) is significant by comparing it to a **critical value** (you find this using a **critical value table,** all of them are at the back of your pack**)**

You need to know **the level of significance**, number of participants **(N)** and the **direction of the hypothesis** to work out the critical value.

For independent groups design there is a different way of working out N which is called **degrees of freedom**

We are never 100% sure that our results are significant when we say they are significant.

The **level of significance** we usually use in psychology is 5% or $p \leq 0.05$

A 5% level of significance means that we are **95%** sure that the results are **not down to chance**

A **Type l error** is a **false positive**- incorrectly accepted as significant when it's not

A **Type ll error** is a **false negative** –incorrectly accepting as not significant when it is.

Stretch yourself- Can you?

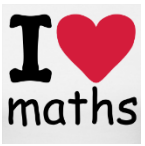
1. The psychologist found the results were significant at $p<0.05$. What is meant by 'the results were significant at $p<0.05$'? (2 marks)
2. Define what is meant by the critical value in statistical testing (2 marks)
3. What is meant by the term type l error (1 mark?)
4. What is the probability of making a type l error at $p \leq 0.1$ (1 mark)
5. How can you reduce the risk of having a type l or type all error? (2 marks)
6. A researcher is testing the effectiveness of a new drug to reduce depression. What level of significance should be used and why? (3 marks)
7. Distinguish between a Type I error and a Type II error (4 marks)

Carrying out statistical testing

You can only be asked to carry out one statistical test (the sign test), the others (Spearman's rho, Pearson's r, Wilcoxon, Mann Whitney, related t-test, Chi-Squared) you will simply be given the observed value of and will have to interpret the data.

Before starting to look at how to carry out tests you need to know all about levels of measurement, which are the different types of data we use in psychology.

Level of Measurement (type of data)

Nominal data (Sometimes referred to as Category data) 	The simplest form of data, data represented in the form of categories for example the number of boys or girls in your class or colour of cars you count in BHASVIC car park.  Nominal data is discreet in that only one item can appear in each category i.e. a red car can't be in the green car column as well.
Ordinal data 	This type of data is put in an order in some way for example if you were asked to rate how much you like Psychology on a scale of 1-10 or to put a list of subjects in order of how much you like them. Ordinal data does not have equal intervals as you might have put maths first, psychology second and French third but you might like psychology almost as much as maths but may not like French that much at all.  We convert raw scores to ranks for statistical testing (1st ,2nd,3rd) and it is the ranks and not the scores that are used in the calculations.
Interval/ratio data 	This data has units with equal intervals between points unlike ordinal but it can still be ordered like ordinal data. Examples include time, weight, temperature, distance. The difference between ratio and interval is that ratio data has an absolute zero point for example if you have zero money in your bank you have no money at all. Zero degrees however doesn't mean there's no temperature as interval data only has an arbitrary zero point.
Scores on a memory test-ordinal or interval? Some data in psychology is harder to decide on for example scores on a memory test. It seems like it has equal intervals between scores but the words would all have to be of equal difficulty but as this is very difficult to achieve it is safer to treat it as ordinal data and to rank the scores. As long as you give a reason for this in the exam you'll be fine.	

The sign test

This is a simple statistical test. In order to complete it you need to know-

1. When it is appropriate to use it?

-If a difference is predicted between two sets of data
-The data is related i.e. is a repeated measures design? (You can also use a matched pairs design because the participants are paired and therefore count as one person tested twice).
-If it is nominal data (but we often turn data into nominal data as in the example below).

2. How to do a sign test

We will be testing the null hypothesis:
Null- There will be no difference in the amount of words recorded during a 3 minute group logic problem task if participants have just finished drinking a litre of water or a litre of energy drink.
Experimental-There will be a difference in the amount of words recorded during a 3 minute logic problem task if participants have just finished drinking a litre of water or a litre of energy drink.

Step one- Convert the data into nominal data by working out which participants produced a higher word count after the energy drink and which produced a lower word count.

We do this by subtracting the score for water from the score for energy drink. If the answer is negative we simply record this sign, if the answer is positive we record the plus sign. If there is **no difference** then we just miss out that score **and that participant**.

Participant	Energy drink	Water	Sign of difference
1	110	122	-
2	59	45	+
3	206	135	+
4	89	90	-
5	76	42	+
6	141	87	+
7	152	131	+
8	98	113	-
9	198	129	+
10	57	62	-
11	267	176	+
12	282	240	+
13	134	157	-
14	103	103	**No diff**
15	88	108	-
16	201	121	+
17	267	231	+
18	322	200	+
19	249	207	+
20	90	104	-

It doesn't matter whether you subtract energy drink form water or the other way around as long as you always subtract in the same direction.

Step two

Next add up all the pluses and minuses. So for our data we have **12 +** and **7 –**
So we can say
pluses (13) = participants who spoke more words after drinking the energy drink than the water
minuses (7) = amount of participants who spoke more words after the water.

Step three

The calculated or observed value is called S and it is the one with the **fewest** scores so here it is 7 so S=7

Step four

Remember we have to compare our answer (calculated value) with the critical value to see if our results are significance.
You know how to find a critical value and the critical value table for the sign test is on page 15 so cover up the answer below and find it for this example.

The value of n is 19 (because one value is ignored as there was no difference)
It is a non-directional or two tailed test
Level of significance is p ≤0.05
So the critical value is **4**
Now look at the top of the sign test critical value to table to see if your observed value needs to be bigger or smaller that the critical value.

It states that the value of S must be equal to or less (≤) than the critical value to be significant.

S=7
critical value= 4

As 7 >4 (more than the critical value) then result is NOT significant.

3. How to report the conclusion that can be drawn

So our results are not significant as S>4. This means that we have to accept our null hypothesis and reject our research hypothesis so even though there was a difference in the mean number of words spoken between the two conditions, it was not statistically significant i.e.

There was no difference in the amount of words recorded during a 3 minute group logic problem task if participants have just finished drinking a litre of water or a litre of energy drink.

For the exam……..

Now turn to the exam practice section and try some of the examples. You need to practice these questions to check that you fully understand the sign test.

Remember you will only be asked to calculate the sign test.

The other statistical tests (that you don't have to calculate just interpret)

You'll notice that there is reference made to statistical **testS, t**hat is because there a number of different tests to choose from. So how do you pick the correct statistical test? This depends on a number of factors specified by the nature of the research and the type of data collected.

The seven tests that you can be asked about (but won't have to calculate) in the exam:

1) **Chi-squared (χ^2)**
2) **Wilcoxon T**
3) **Mann-Whitney U**
4) **Spearman's Rho**
5) **Unrelated t-test**
6) **Related t-test**
7) **Pearson's r**

How do you know which test to use? **You need to learn the following criteria, but there is no need for you to understand why this is the case**

a) **Difference or correlation?**-Whether the researcher is testing for differences between groups (i.e. an experiment) or a correlation between two co-variables

b) **Level of measurement** (nominal, ordinal, interval and ratio)

c) In a test of difference, whether the **experimental design** is an independent groups, repeated measures or matched pairs

a) A test of difference or a correlation?

Laboratory experiments, field experiments, natural and quasi experiments are all testing for differences between groups. The researcher is trying to establish the probability that changes in the DV are caused by the experimental manipulation or naturally occurring IV. Correlational research is attempting to show how two co-variables are linked

Which it is should be obvious from the wording of the hypothesis. In this context correlation can include correlational analysis as well as investigations that are looking for an association. So look for the word **correlation** or **association** in the hypothesis for a correlation.

b) Level of Measurement (type of data)- is the data nominal, ordinal or interval/ratio? (see page 19)

c) The Experimental Design

An independent groups design uses different participants in each group, whereas the repeated measures design uses the same participants in each group. For mathematical reasons, these tests treat the repeated measures and match-pairs design as the same category.

NB: correlations do not have an experimental design.

Exam skills

Picking the right test

	Difference		Correlation
	Related data (Repeated measures, matched pairs)	**Independent data (independent groups design)**	**Related data**
Nominal	***Sign test***	Chi squared	Chi-squared
Ordinal	***Wilcoxon***	***Mann-Whitney***	Spearman's rho
Interval	related t-test	Unrelated t-test	Pearson's r

The tests in bold italics are the ones in which the observed value has to be ≤ the critical value and they form an L for less than to help you to remember. This will all become clear later one.

You need to learn this table so come up with your own mnemonic to remember the order, something like

Scoffing **C**heesy **C**hips Will **M**ake **S**omeone **R**ather **U**nderstandably **P**orky

*Extension material- Parametric or non-parametric?

Statistical tests are either classed as parametric or non-parametric. You don't need to know the differences between these for the exam and they are quite complex so we'll just tell you a little bit about them here in case you were wondering. Parametric tests are more powerful and robust that the other tests and so if psychologists can use them they will as these tests may be able to detect significance within some data sets that non-parametric tests cannot. The related t-test, unrelated t-test and

1. Data must be interval as parametric tests use the raw data rather than ranking them.
2. Data should be from a normal distribution (skewed distributions don't work)
3. The set of score should have similar distributions or spreads. Often we compare the standard deviations of the conditions and if they are similar than a parametric test can be used.

When and how to use the statistical test

Below we are going to show you examples of how you work through a statistical test, from deciding which one to use, to concluding what the results show. Try to work out the answer for each section for yourself by covering up the answers to see of you really understanding this. There are questions for you to try in the practical section also.

Example one-Familiarity leads to liking

Research has found that people like things that are familiar. In one study Zajonc (1968) told participants that he was conducting a study on visual memory and showed them photographs of 12 different men (face only), each for two seconds only. At the end, participants were asked to rate how much they liked the 12 different men on a scale form 0-6. Some photos were shown more often than other; for example one photo appeared 25 times, whereas others appeared only once.

Alternative hypothesis- People rate the more frequently seen face as more likeable than the less frequently seen face.
Null hypothesis-There is no difference in the likeability score for more and less familiar faces.

So what statistical test would you use?

Remember to first work out

1. Difference or correlation?
2. Which levels of measurement?
3. Which experimental design?

Justifying your statistical test

1. The hypothesis is looking for a test of difference, because it's looking at the difference in ratings of frequently seen photos and less seen photos.

2. The data is ordinal because there are not equal intervals between ratings as they are simply rating on a scale of 0-6.

3. As the two conditions are scores for the less frequently and more frequently seen photo as rated by the same person the experimental design is repeated measures and so the data is related.

So looking at the table, what test should we use?

	Difference		Correlation
	Related data (Repeated measures, matched pairs)	**Independent data (independent groups design)**	**Related data**
Nominal	Sign test	Chi squared	Chi-squared
ordinal	**Wilcoxon**	Mann-Whitney	Spearman's rho
interval	related t-test	Unrelated t-test	Pearson's r

Writing up the justification for a statistical test

The statistical test we would use for the research is Wilcoxon because the hypothesis is looking for a test of difference, with ordinal data that is related (repeated measures).

Writing up the results –You will be given the following info and asked to interpret it.

The observed/calculated value of T came out at T=18.5
N=11 (one score was omitted)
Significance level: 5% (0.05)
The hypothesis was directional.

Were the results significant? I.e. interpret the data

Remember you have to compare the observed value to the critical value to decide significance. The critical value table for Wilcoxon is in the critical value tables section, use it to find the critical value for this research and whether the observed value has to be less than or more than the critical value.

Found it? Good, the critical value for T =13 and T must be ≤ the critical value to be significant.

T=18.5 critical value =13

Write it up like this- T must be ≤ the critical value to be significant. 18.5>13 therefore the calculated value is not significant (at p≤0.05), we must accept the null hypothesis and reject the alternative hypothesis and conclude that there is no difference in the likeability score for more or less familiar faces.

Example two-helping behaviour and appearance

Piliavin (1969) investigated helping behaviour and looked specifically at whether people were quicker at offering help in an emergency situation to a man with a cane or a man who appeared drunk. The emergency situation was when a person appeared to collapse on an underground train. Participants were randomly assigned to condition A (man with cane collapses) or condition B (man appears drunk collapses).

Alternative hypothesis*:* The speed of offering help in an emergency situation is faster when the victim is carrying a cane than when appearing drunk
Null hypothesis*:* There is no difference in the speed of offering help to victims with a cane or appearing drunk.

So what statistical test would you use?

Remember to first work out

1. Difference or correlation?
2. Which levels of measurement?
3. Which experimental design?

Justifying your statistical test

1. The hypothesis is looking for a test of difference, because it's looking at the difference in speed in which someone helps either drunk or carrying a cane.

2. The data is interval as measured in seconds which has equal intervals.

3. It is independent groups design or unrelated as people either see a drunk man collapse or man with can but not both

NB the data is normal distribution, with a similar spread of scores.

So looking at the table, what test should we use?

	Difference		Correlation
	Related data (Repeated measures, matched pairs)	**Independent data (independent groups design)**	**Related data**
Nominal	Sign test	Chi squared	Chi-squared
ordinal	Wilcoxon	Mann-Whitney	Spearman's rho
interval	related t-test	**Unrelated t-test**	Pearson's r

Writing up the justification for a statistical test

The statistical test we would use for the research is the independent t-test because the hypothesis is looking at a test of difference, with interval data that is independent or unrelated (and the data is normally distributed with a similar spread of scores).

Writing up the results

The observed/calculated value of T came out at T=1.882
N1=10, N2=12
Significance level: 5% (0.05)
The hypothesis was directional

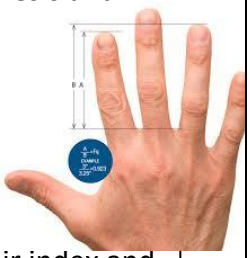
Were the results significant?

The critical value table for the T-tests are in the back of the pack, have a look at find the critical value for this research and find out whether the observed value has to be less than or more than the critical value. Remember the data is independent groups so you are looking for degrees of freedom not the value of N.

Found it? Good, Df=(10+12)-2 So Df=2 so the critical value is 1.724
T must be ≥ critical value to be significant

T=1.882 critical value =1.724

Write up like- T must be ≥ critical value to be significant. 1.882>1.724 therefore the calculated value is significant (at p≤0.05), we must accept the alternative hypothesis and reject the null hypothesis and conclude how a victim looks does effect the speed a person helps in an emergency situatic



Example three-mathematicians have long fingers!

A study by Brosnan (2008) found that boys with smaller finger length ratio between their index and ring fingers were more likely to have a talent in maths. The explanation is related to the effects of testosterone which is associated with reduced finger length ratio and increased numeracy skills.

Alternative hypothesis- The finger length ration between index finger and ring finger is negatively correlated to numeracy skills in boys.
Null hypothesis-There is no correlation between finger length ratio and numeracy in boys

Justifying your statistical test

1. The hypothesis is looking for a test of correlation, because it's looking at the link between finger length and numeracy and the word correlation is used which gives you a big clue!

2. The data is ordinal as numeracy skills are measured using a test a so may not have equal intervals between scores.

3. It is related data as the numeracy in boys and their finger length is being studied

Writing up your justification for your statistical test

The statistical test that would be used for this research is spearman's rho as the hypothesis is looking for a test of correlation for ordinal data which is related.

Writing up the results section **(sometimes you'll only get the observed value and will have to work out the rest from given data)**

The observed value of rho is -0.58

Is it significant?

-What is N? (see table to right)
-Is the hypothesis one or two tailed?
-What level of significance do we normally use in Psychology?
Find the critical value table in the critical value section of the pack and find out what it is for this example.
-Does the observed value need to be greater than less than?

Participant	Finger length ratio	Numeracy score
1	10	8
2	5.5	16
3	9	10
4	4	9
5	3	15
6	1	14
7	7	12
8	8	8
9	2	17
10	5.5	5

So were the results significant?

N=10
The hypothesis is one tailed
The level of significance is 5%
The critical value of rho=.564
The observed value must be ≥the critical value to be significant.

As the observed value of -0.58 is ≤ the critical value of 0.564 then the results are significant (at p≤0.05). We can therefore reject the null hypothesis and accept the alternative hypothesis and conclude that finger length ration between index finger and ring finger is negatively correlated with numeracy skills in boys.

Example four- Boys and their toys

A researcher wanted to investigate whether male or female children prefer to play with different sorts of toys and collected data in a 2x2 contingency table (see below).

***Alternative hypothesis*-** There will be a difference between gender and choice of toy (soft toy or car)

Null hypothesis- There will be no difference between gender and toy choice (car or soft toy).

A contingency table is simply a posh name for a table that displays **nominal** and **independent** data. The example here is called a 2x2 contingency table because there are two rows and two columns but you can have any number of rows and columns

2x2 contingency table to show toy preferences and gender

	Preferred toy		
Gender	Car	Soft toy	Total
Male	18	12	30
Female	10	20	30
Total	28	32	60

TOP TIP- if you see a contingency table you know the statistical test is chi-squared but remember you still have to justify the reason fully

Statistical test choice

This is looking for a test of difference
The data is nominal (category data)
The data is independent
So the correct statistical test is CHI SQUARED

Is it significant?

The calculated value of $x^2=4.29$
The hypothesis is two tailed
$p \leq 0.05$
It's independent so degrees of freedom not N need to be calculated. For chi squared it's simply

Df=(number of rows-1)x(number of rows-1)
As it's a 2x2 contingency table then Df=(2-1)x(2-1) so **Df=1**

Looking at the critical value table you can see that the calculated value≥ x^2 to be significant.
Critical value =3.84

Writing it up- calculated value≥ x^2 to be significant . 4.29≥3.84 so the results are significant (at 5%) and we accept our alternative hypothesis and reject our null hypothesis so we can say that there is a difference in toy choice between genders and looking at the contingency table it looks like males are more likely to choose the car and females the soft toy.

For the exam remember…..

Is the test significant?

Find the critical value-

-One or two tailed hypothesis?

-Level of significance? Usually 5% or 0.05

-Number of participants for related (N)

-or degrees of freedom for independent data

Does the observed value need to be ≤ or ≥ the critical value?

Choosing a statistical test

1. Difference of correlation?

2. Nominal, ordinal or interval data? (levels of measurement)

3. Experimental design? Independent groups or repeated measures/matched pairs.

Stretch yourself

1. Give an example of nominal data (2 marks)
2. Explain the difference between ordinal and interval data (3 marks)
3. Suggest why a researcher may choose to use p≤0.01 in preference to p≤0.05 (2marks)
4. Explain the relationship between the calculated value and critical value (3 marks)
5. Identify three pieces of information used to find a statistical test (3 marks)
6. When using the Wilcoxon test, is the calculated value greater than or less than the critical value? (1 mark).
7. What is a 4x4 contingency table? (2 marks)

For the exam……..

Now turn to the exam practice section and try some of the examples. You need to practice these questions to check that you fully understand when to use the tests and how to write them up.

Reporting Psychological explanations

Progress in science depends on communication between researchers. It is therefore essential to describe the results as accurately and effectively as possible. Also if all studies are written in the same standardised way then it is easier to replicate findings, check results and peer review.
The basic requirements are to communicate

- What was done
- Why was it done
- What was found
- What it means

The report takes the following format-

Abstract
Introduction
Method
Results
Discussion
References

Abstract

150-200 words. This allows the reader to get a quick picture of the study and its results.
It must very briefly mention the aims, hypothesis, method, results and conclusions.

Example of abstract for Mursteins matching hypothesis written for Psychology coursework (2006)

This study is designed to test the theory that individuals in heterosexual couples are of a similar level of attractiveness to each other. It was found by Murstein that when each half of a couple was rated by independent judges there was a tendency for the couple to be similar in terms of their physical attractiveness. The hypothesis for this study is there will be a significant positive correlation between the ratings of physical attractiveness of those in gay and lesbian couples. This was rated by 10 independent judges using opportunity sampling. 20 participants were used, 10 male and 10 female within the age range of 20-60. The participants were those in the photographs of couples found on the internet. The results were analysed using spearman's rank order correlation coefficient test. The calculated value of RS was 0.24242. The critical value of RS for a one tailed test with 0.05 significance level was 0.564. Therefore the null hypothesis was accepted. This shows that there was no significant positive correlation between the ratings of each half of the couple.

Introduction

The introduction begins with a review of previous research so the reader knows what other research has been done and understands the reasons for the current study. The focus of this research review should lead logical so the reader is convinced of the reasons for this particular research. The introduction should be like a funnel- beginning broadly and narrowing down to a particular research hypothesis. The researcher states their aims, research prediction and hypothesis.

Method

This section contains a detailed description of what the researcher did. It needs to be in enough detail that it could be easily and precisely replicated. It must include:

Design-The design is clearly stated e.g. laboratory experiment, repeated measures, IV/DV but all of the design decisions **must** be justified.
Participants- Information about sampling methods and the people who took part in the study e.g. how many, ages, jobs etc.
Apparatus-detail of any materials used that would be needed for a full replication.
Standardised procedure- This is a step by step description from beginning to end of everything that was done and everything that was said to participants to allow a full replication. This must include standardised instructions, debriefing, briefing.
Ethics-An explanation of how these were addressed within the study.

Results

The results section should summarise the key findings from the investigation and should include-
Descriptive statistics- this is the key findings outlined in a straight forward way so that readers can "eyeball" the data and should include tables, graphs, measures of central tendency and dispersion. See design section for how this is layed out.
Inferential statistics should include choice of statistical test with justification, calculated and critical values, the level of significance and whether the test was one or two tailed. The outcome is explained in terms of acceptance or rejection of the experimental and null hypothesis.

Discussion

After the results section in the investigation report, the researcher will include a discussion of the results. This will include a discussion of:

- **Summary of results**-The results and what they tell us
- **Relationship to background research**-Whether the study is in line with what the research studies quoted in the introduction suggest
- **Limitations and modifications**-Strengths and weaknesses of the research and how it could be improved
- **Implications and suggestions for further research.**

Referencing

The full details of any journal articles or books that are mentioned in the research report are given.
The format is:

For journal articles-Authors name, date, title of article, journal title, volume (issue number), page numbers

Gupta, S. (1991). Effects of time and day and personality on intelligence test scores. Personality and individual differences, 12 (11). 1227-1231

Book- Authors name, date, title of book, place of publication, publisher
Flanagan, C and Berry, D (2016). A-level Psychology. Cheltenham, illuminate publishing

Critical values tables