

Variables, Correlations and Experiments

by
Jeff Standen

VARIABLES

What is a variable? Basically it's anything that can change in a measurable way. For example, heat is a variable. How do we measure it? We can use a thermometer. We can use a variety of instruments for measuring forces, distances, electricity, speed and a host of other qualities.

How might we measure some psychological variables? For example, how could we measure ability at maths? We have to convert the concept of ability to something we can measure that represents ability. So, we could ask somebody to carry out some mathematical calculations, and give them a correctness score. Children at school would call it a maths test! How could we measure stress? We are often saying how stressful an event is? How can we compare the stress of two events? Well, we know that stress cause some bodily changes, so if we measure the changes, we can compare two levels of stress. One way is to measure the galvanic skin response, or the micro electric charge we all have on our bodies. You know when you walk across the nylon carpet and you touch the brass door handle and you get an electric jolt? That's one effect of it.

When we convert a variable such as stress into something measurable, we **operationalise** that variable. So, how do we operationalise attitude? We would use a rating scale. How would we operationalise memory? We would ask our participant to remember something, and then recall it. The amount recalled would be an operationalisation of memory. (Sociologists would call the amount recalled an 'indicator' of memory – except they wouldn't be doing memory.)

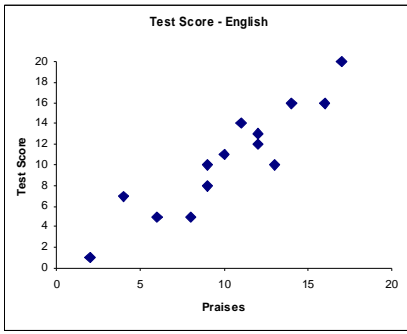
Correlation

Correlation is nothing more than the relationship between two variables. Love makes the heart go stronger – that is a correlation. The higher the 'love' measurement, the higher the 'heart is strong' measurement. Imagine a rainy day with lots of puddles. The longer it rains, the bigger the puddles. We could draw a graphical representation of this – it would be a scattergram. Can we say that the size of the puddles is due only to the time it's been raining? What other things might be affecting the size? Well, the porosity of the ground, perhaps. Puddles will be deeper on stone pavements than grass verges. But water may splash into the puddles from passing traffic. And surely any child in wellies would love to jump in a puddle and splash it everywhere! So we can't be sure that all of the puddle size is caused by the amount of rain falling. The correlation is between two variables: the size of the puddle and the length of time it has been raining. But water splashed in by traffic is a variable – we will call this an intervening variable.

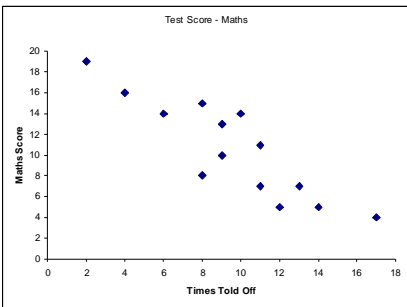
Here is an example closer to psychology.

- A teacher wants to see the effect of praising her pupils when they behave well during an English class.
- Every time a pupil does something "good" she publicly praises that pupil.
- She counts the number of praises per pupil.
- At the end of the lesson she gives them all a test based on that lesson.
- She collects the scores...

Pupil	Times Praised	Test Score	Results for English Class test
A	12	12	
B	4	7	
C	9	8	
D	2	1	
E	14	16	
F	10	11	
G	9	10	
H	17	20	
I	11	14	
J	9	5	
K	13	10	
L	16	16	
M	12	13	
N	6	5	



Scattergrams of praise lesson (above) and punishment lesson (Below)



She predicts that the pupils who are praised most will score highest in the English test. It is clear from the results – especially when they are presented on a scattergram – that there is a relationship between praises and success. The results show that there is a positive correlation between the number of praises and the English test score.

- Next, the teacher wants to see the effect of punishing her pupils when they behave badly during a Maths class.
- Every time a pupil does something “naughty” she tells that pupil off.
- She counts the number of tellings off per pupil...

This time, the results show a negative correlation between the number of times punished and the maths test score.

A positive correlation describes a relationship where, as one variable increases, so does the other.

A negative correlation describes a relationship where, as one variable increases, so the other variable decreases.

Generally speaking, psychologists look at the shape of the scattergram (the eyeball test) in the first instance. But it is possible to do a mathematical calculation to determine the coefficient of correlation. This is nothing more than a number that describes the correlation. If the correlation was a perfect relationship, the scattergram would be a straight line. For a perfect positive correlation, the coefficient is +1. For a perfect negative correlation, the coefficient is -1. (You might like to think what a coefficient of 0 actually means!)

But psychologists are usually dealing with people, who for some reason never seem to fit the results *exactly perfectly* and so the scattergram is really scattered! The two scattergrams here have coefficients of +0.92 and -0.87, in other words not far from perfect but certainly not a straight line. Psychologists might also want to know how far from perfection a correlation can be and still say there is a meaningful relationship between the two variables. This can be calculated, and it does depend on the number of points in the scattergram. When you come to study inferential statistics, this will be important, but for now, knowing what a coefficient of correlation is will be sufficient.

Let us return to the teacher in the classroom. Can she say with certainty that the improved English results were caused by praising pupils? No, she can't. There may be alternative explanations for the results. Some pupils may be more proficient at English. Some pupils may have been unwell, or tired. Any of these could affect the results. Similarly, punishment may not be the cause of the maths scores, for the same reasons. These alternative explanations are a result of intervening variables. It is because of the intervening variables that we cannot say that changes in one variable are caused by changes in the other. In technical terms, we say ‘there is no causal relationship between the variables’. This is the main weakness of correlation as a research method. However, it does have the strengths that it will show the existence of a relationship and that may be the starting point of more research.

Experiments

THE EXPERIMENT

When we say ‘an experiment’ we often think of any piece of research whether it’s in chemistry, physics, or psychology. But in fact, the experiment is a specific method of research which is probably the most studied of them all. I have found that a useful way to start studying the experiment is first, to compare it with correlation. In the correlation method, we look at the relationship between two variables, but we cannot say that changes in one variable cause changes in the other. We have already seen that intervening variables get in the way. Basically, the art of an experiment (and I believe it is an art!) is to get rid of the effects of all those intervening variables. Except, as you will see, we call them something else.

To begin our study of the experiment, we’ll make a perfect cup of coffee.

First problem: who says it’s perfect? How will we measure perfection? We can use a rating scale, say a score of 10 means absolute perfection, the coffee that legends are made of... while a score of zero would involve words like mud, vomit or similar disgusting terms.

So what goes into our perfect cup of coffee? Well, coffee, definitely. But what about the other ingredients? My preference is for medium strength black with no sugar. But yours may be strong white with loads of sugar and not too warm. In fact there are so many possibilities it is almost impossible to make a universally perfect cup of coffee. Let’s start with the brand of coffee. I have this theory that expensive well advertised household name coffee, such as Kenco (Whoops! No advertising allowed.) will make a better cup of coffee than Supermarket Own Brand coffee (I’ll call this “Cheapo” brand). I will test this theory and I will use the experimental method to test it.



The purpose of an experiment is to test a theory. The way I am going to test this theory is to compare the taste of Kenco with the taste of Cheapo. And the theory predicts that Kenco will taste better. So now we’ll unpack all the technical bits and design our experiment.

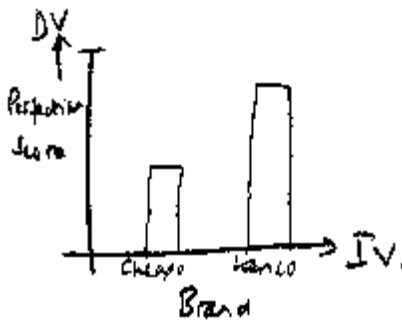
Basically, we can ask a sample of people, or participants, to taste Cheapo and Kenco and ask them to score the taste on our perfection scale.

Second problem: Participant A likes his coffee black, like me, but with sugar. Participant B likes coffee white, with sugar. When I come to compare their scores, how can I be sure that the difference in scores is due to the brand of coffee in each cup, or whether there was milk, or sugar, or whatever? The answer is, we must make all the cups of coffee in exactly the same way, using the same amount of milk, sugar, water, coffee etc., and the only difference between the two cups will be in the brand used. Kenco or Cheapo. We can use a variety of methods to estimate the most popular combination of milk, sugar and coffee etc., and we will probably make our cups

of coffee using 250ml of boiling water from the same urn, 25ml of milk, 25g of sugar and 15g of coffee granules.

We will ask our participants to drink a cup of Cheapo, then rate it for perfection, followed by a cup of Kenco, and rate that for perfection. We can predict, according to the theory, that participants will give a higher score to Kenco than to Cheapo. Providing we have made our cups of coffee the same, we should be able to say that the difference in scores is only due to the difference in brand.

Let's look at some technical terms for what we've done so far. There are quite a few variables here. The amount of each ingredient is a separate variable. The volume of milk, the weight of coffee etc., all of these are variables. The brand of coffee is also a variable, and in this experiment it has two values. If we were testing a third brand, say Decaff, then it would have three values. And so on. Another variable is the perfection score. This has a whole spectrum of values from 0 to 10.



Try and draw a rough bar chart in your mind - even on a scrap of paper. If you represent the Variable you're measuring as the vertical or y axis, and the Variable you're manipulating as the x-axis, then IV and DV will ALWAYS be as shown here.

Experiments are designed so that change in one of the variables is shown to be caused by, and only by, changes in another variable. We are investigating changes in the variable called "Brand of Coffee", to which we assign the values "Cheapo" and "Kenco". In other words, we are manipulating the values of this variable to suit our own purpose. The scores on the perfection scale (another variable) are dependent on the brand of coffee being tasted.

In this experiment, the brand of coffee is called the Independent Variable. The Perfection Score is called the Dependent Variable. All the other variables – amount of milk, temperature of water, sugar etc. – which might interfere with the result, are called Extraneous Variables.

EXTRANEIOUS VARIABLES are variables which, if not controlled, can interfere with the results

In an experiment, the independent variable is manipulated to produce changes in the dependent variable, and provided all the extraneous variables are held constant, or "controlled", then we can say confidently that changes in the dependent variable are caused by changes in the independent variable.

When you come to carry out the experiment, however, some niggling little doubts and complications show up. We don't seem to have allowed for other variables which might interfere with the result.

We asked our participants to drink a cup of Cheapo (Yeuchh!), then rate it. Then we asked them to drink a cup of Kenco... "well, it tastes better, but I can still taste the Cheapo. In fact, the Kenco is nothing like I imagined, I can only taste Cheapo."

An **ORDER EFFECT** is one of the commonest extraneous variables

You couldn't be sure that the difference in scores was completely due to the brand, because of the order of tasting. This is an example of an "order effect" and is another extraneous variable. You could offer Kenco first, then Cheapo, but then the Cheapo might not score as badly as it should. Here are three suggestions to solve this problem:

COUNTERBALANCING is a method of carrying out experimental conditions to avoid order effects

1. Offer half your participants Kenco followed by Cheapo; then the other half Cheapo followed by Kenco. This way, the effect will be reduced, although not completely eliminated. This technique is known as ‘counterbalancing’.

2. Take a week to do the experiment. Offer your participants Kenco this week, then Cheapo next week. Or, vice-versa. Or, half Kenco this week and half Cheapo this week, followed by the reverse next week. That way, the taste should have dissipated. But over a week, participants’ tastes may have changed.

3. Give half of your participants Kenco only, and the other half Cheapo only. Ah, you say, but different people have different likings for coffee. Yes, there may be some individual differences but perhaps these may not interfere as much as the taste of the coffee.

EXPERIMENTAL DESIGN

Repeated Measures

Independent Subjects

Matched Pairs

This brings us to the question of **experimental design**, which has a specific meaning. A design where participants take part in all experimental conditions, in this case where they drink both brands of coffee, is known as a “**repeated measures**” design – think of it in terms of the experimenter repeating the measurement for each participant. A design where the conditions are divided between participants, in this case where participants only drink one brand, is known as an “**independent subjects**” design. There is a third type of design – “**matched pairs**” – where individual differences are controlled as far as possible, perhaps in only one characteristic, say aggression, for example, but more likely individual differences would be controlled by using identical twins. This is a very expensive and difficult technique, but it is not rare.

So, counterbalancing, taking a week, or using a different design – which is it to be? Each method has its strengths and weaknesses. The fact is there is a trade off between each design, and you must make your own design decision. I would be inclined to use an independent measures design.

3. So, we make a design decision. We hand our participants a cup of coffee. But we are after all psychology students, and we do know about body language and stuff. How can we be sure that our participant is ignorant as to the brand? If they guess from the experimenter’s attitude or body language that they’ve got Cheapo, how can we be sure that the score reflects the taste or the expectation of the taste? In other words if they guess it’s Cheapo they’ll score it low.

EXPERIMENTER BIAS

This is an example of Experimenter Bias. It can be a serious extraneous variable, and can be controlled by using a “double blind” procedure. The experimenter will know which cup is which, but another person will give out the cups without knowing which is which.

4. What about someone who doesn’t like sugar in their coffee? They should always have the option of withdrawing from the experiment. This would be an ethical issue. But if they didn’t like the

**SITUATIONAL vs.
PARTICIPANT
VARIABLES**

coffee, they would give it a lower score whichever the brand. But we can assume that they will be consistently low scoring, and as we are looking for a difference, then it is not a problem.

5. What about someone who smokes, or is not well, or has a hangover, or has some other individual characteristic that could interfere with the results. Well, these are individual differences, and sometimes they can be reflected in a sample which, if it is truly random, may have a higher than expected representation of that characteristic. If you are unfortunate enough to have an extreme sample, then you would probably repeat with a new sample.

It is clear then that there are extraneous variables which can arise because of the way the experiment is carried out; and those which arise because of individual participant differences. Extraneous variables resulting from the method and the environment are known as 'situational variables' Those arising from the participants are known as 'Participant Variables'

A third type of extraneous variable is known as a 'Confounding Variable' These are those interfering little variables that you just can't easily control. Examples are where the coffee gets cold because participants chatted over their coffee, or where the milk was 'off' and it wasn't noticed in time. You learn from your confounding variables!!

HYPOTHESES AND PREDICTIONS

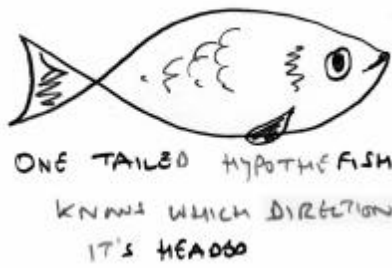
It is important when planning our research that we have a particular objective or aim. Why are we actually doing the experiment? How do we know if we have achieved our objective? We have already said that the purpose of an experiment is to test a theory. The theory for our coffee experiment is that Kenco will make a better cup of coffee than Cheapo. If the theory is correct, then the results of the experiment should be predictable: Participants will give a higher perfection score to Kenco than to Cheapo. But, as you will have guessed, the prediction needs to be dressed up to include exactly the information that we want – no more, no less. There are thus two experimental conditions, we'll call them Condition A (Participants drinking Cheapo coffee) and Condition B (Participants drinking Kenco coffee). The predicted description of the results is therefore:

**ALTERNATIVE
OR
EXPERIMENTAL
HYPOTHESIS**

Participants will give a higher perfection score in Condition A (Cheapo coffee) than in Condition B (Kenco coffee).

This prediction, which is no more than a specific statement of the expected results, is known as the Alternative Hypothesis. (Or sometimes the Experimental Hypothesis.)

In order to predict the direction of the difference in scores, we should be informed by the theory. Here we really do know the direction from the theory- "Kenco is better than Cheapo", so expect a higher score for Kenco. But what if the theory doesn't give us an



NULL HYPOTHESIS

Enough of the coffee!
What about a real life
Experiment...



indication of direction? For example, suppose we were doing the experiment, but instead of Kenco and Cheapo, we were comparing Decaffeinated coffee with regular coffee? Any theory of mine would say that there is a difference, but I don't know which coffee is better. The only prediction I could make based on the information I have is that there will be a difference – I don't know the direction. The experimental hypothesis in that case would be

Participants will give a different perfection score in Condition A (Regular coffee) than in Condition B (Decaffeinated coffee).

When we know the direction, we describe a one-tailed hypothesis. When we don't know the direction, we are describing a two-tailed hypothesis. See the fish pictures for an easy way to remember this.

There is another prediction we need to make. There are basically two possible outcomes from our experiment. The first is that the theory we are testing is correct, we have controlled all the variables and our results match the prediction. But what if the theory is incorrect. What would be the predicted results if Kenco and Cheapo were the same? There would be no difference in perfection scores. This prediction of the outcome would be:

There will be no difference in Participants' perfection scores for either Cheapo coffee (Condition A) or Kenco coffee (Condition B).

This is known as the Null Hypothesis, and should be stated alongside the Experimental Hypothesis.

(When we carry out an inferential test, strangely enough we are actually testing the null hypothesis. When we observe a difference, we say that there is an effect and we measure the probability of getting the same results again if the null hypothesis is true. We would expect to get a small probability – think about it. More later.)

A MEMORY EXPERIMENT

As a student of psychology, one of the easiest experiments to carry out will almost certainly involve memory. We'll consider one here. There is a theory that people are more likely to remember something that has meaning than something that has no meaning. You would find it easier to remember a group of three letter words, e.g. hat, dog, top etc., than a group of three letter nonsense items, e.g. gdx, csz, ufw etc. (An item consisting of three characters, such as gdx, or perhaps dog, or even s5k, is called a trigram.) So we can test that theory by asking participants to learn a list of meaningful trigrams (words!) and a list of nonsense trigrams, and comparing the number of items recalled. We would allow participants say, one minute in each condition to commit as many of the items as possible to memory. They should have the same time each because if the timings were different we couldn't be sure that the difference in results was due to the experimental effect or to extra time to learn some items.

Here are some trigram lists, Condition A And Condition B, of the sort that were used in class to demonstrate a memory Experiment:

JIG	GIN	EOP	GJI
MOB	BAR	VFI	BOH
FIN	FEW	BLR	MYF
GET	DAY	CXN	ZES
FUR	BAD	MUV	XHY
YET	SON	LQI	CDO
MAY	LAG	OJM	JRS
DAM	JOY	VXP	PBN
NOW	RIP	KEB	SIG
LOT	DOG	NUZ	ODR
Condition A Meaningful Trigrams		Condition B Nonsense Trigrams	

Participants would be asked to memorise the meaningful condition, recall as many as they can then do the same with the nonsense condition.

The mean number of trigrams recalled from each condition would be plotted as a bar chart.

Here are some results from carrying out this experiment with some students. The figures are the number of items recalled by each participant.

The bottom two results are the mean of each condition.

In the next section we will show how to present these results.

mean- ing	no mean-
15	6
14	5
12	4
16	6
15	7
19	9
18	9
15	7
17	7
16	6
13	4
15	6
20	10
18	8
17	7
16	6
14	4
11	3
15	5
12	3
9	1
11	2
18	8
mean 15.04	5.78

(An extraneous variable). We would predict that participants would recall more words than nonsense trigrams. Compare the memory experiment with the coffee experiment:

Coffee

Theory: Kenco makes a better cup of coffee than Cheapo

Operationalisation: “Better cup of coffee” is measured on a rating scale of “perfection”

Experimental Hypothesis: Participants will give a higher perfection score in Condition A (Cheapo coffee) than in Condition B (Kenco coffee).

Null Hypothesis: There will be no difference in Participants’ perfection scores for either Cheapo coffee (Condition A) or Kenco coffee (Condition B).

Independent Variable: The brand of coffee. It has two values, Cheapo and Kenco.

Dependent variable: The perfection score

Extraneous variables: These include – but there are probably others! –

- ◆ Differences in the amount of ingredients, e.g. sugar, milk etc.
- ◆ Order effects
- ◆ Individual differences
- ◆ Experimenter bias

Memory

Theory: Meaningful items are easier to remember than nonsense items

Operationalisation: “Easier to remember” is operationalised by counting the number of items recalled.

Experimental Hypothesis: Participants will recall more items from the Condition A (meaningful items) than from Condition B (nonsense items)

Null Hypothesis: There will be no difference in participants’ recall between Condition A (meaningful items) and Condition B (nonsense items)

Independent Variable: The meaningfulness of the trigrams. There are two values, Meaningful (i.e. words) and meaningless (nonsense trigrams).

Dependent variable: The number of items recalled

Extraneous variables: These include – but there are probably others! –

- ◆ Order effects
- ◆ The time allowed for learning the items
- ◆ The environment – i.e. all in the same room at the same time, where noise and distraction is at a minimum. If it is noisy, learning will be impaired.

Reporting your Results

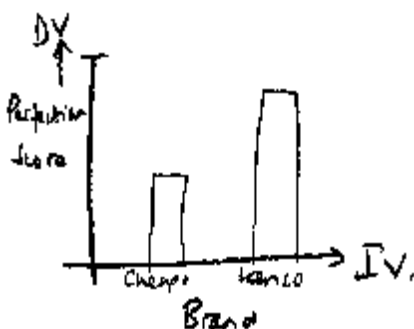
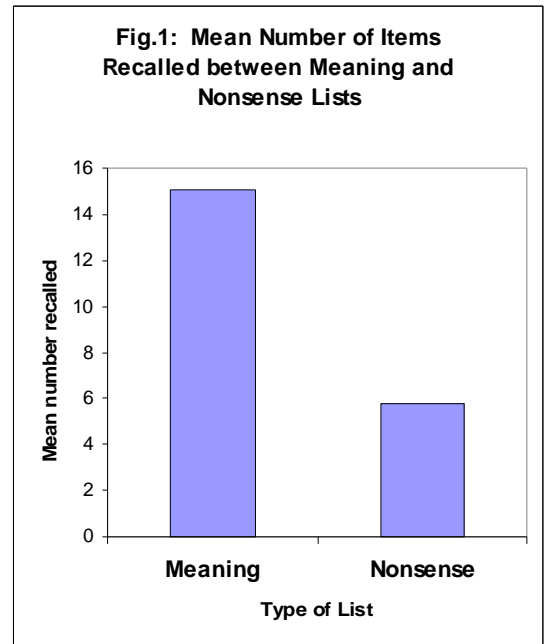
Some notes on describing data

So, you've finished your experiment and you've got the results. All these lovely numbers or times or questionnaire answers. What are you going to do with them? Some sort of analysis I hope. My guess is that you will want to organise the data in some way so that you can understand it. Perhaps you may have designed your research so that you already have a ready made analysis and you know what to do with it. Great, in an ideal world, yes. But the real world, well, sit down, look at the figures. Go and make a cup of coffee. Look at the figures again. Look at a really good research methods book. Make another cup of coffee. Repeat process for several days. Realise deadline is approaching. Weep. Swear. Give up...



It needn't be like that. What do we actually want to do in our analysis? Suppose we have a series of results from participants, say, the number of items recalled in a memory experiment. We want to see what most people got, and we want to see if what most people got was representative of the group of participants. In other words, we want to see what was the average number of items recalled and we want to know that most participants got somewhere near that average.

We are using the data obtained from the memory experiment in the previous section.



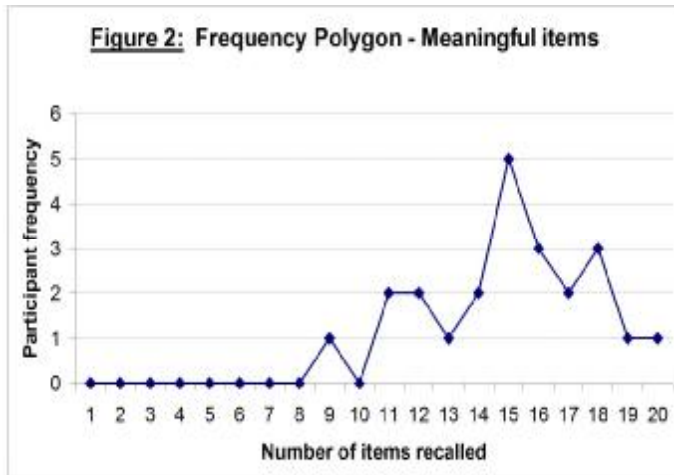
Hmm... Looks familiar



Figure 1 is a bar-chart illustrating the results. This is the typical graph you should be using when you write up an experiment. It is quick to draw, it shows clearly that there are differences between the conditions, and it conforms to the KISS principle. Also, the dependent variable will always be the Y-axis, and the independent variable will always be the X-axis.

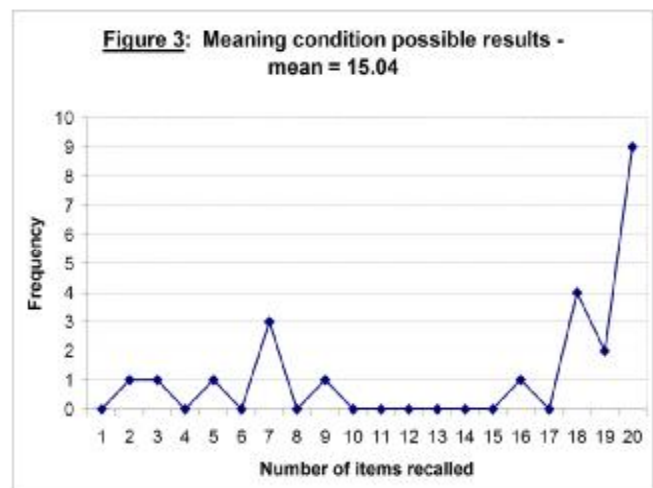
Let's take a closer look at the meaning list. The mean has been calculated for you. The results seem to be more or less around the mean, with a couple of high ones and a low one.

The range of recall scores is 11. That's not too bad, considering the largest possible range can only be 20. But it would be useful to see how the individual results are dispersed. It is possible for the results

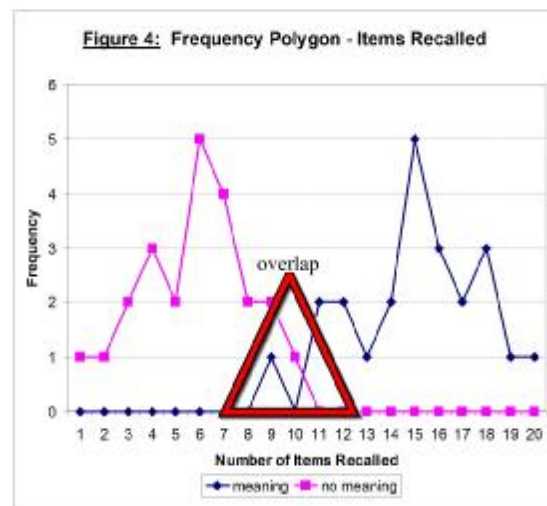


to be unbalanced and the mean and range on their own may not give a true representation of how the group of participants performed. It is possible to show this graphically, using a frequency diagram. First, you need to group your data into piles of recall scores. In other words, the numbers of participants recalling particular scores. One participant recalled 9, two participants recalled 12 words, and five recalled 15 words. The frequency diagram (technical name – frequency polygon) looks like figure 2:

It can be seen that there is a fairly even spread of results about the mean. But it might not be like that, and the mean and range on their own won't tell you. Figure 3 is a frequency polygon of some results which also have a mean of 15.04. They are so spread out that you can't tell much from them about anybody's memory.



You can do the same with the non-meaning trigram results. The third frequency polygon (Figure 4) shows both sets of data on the same chart. So, in this experiment, our results are fairly evenly spread about the mean for each condition. There is some overlap but this does not appear to affect the overall results to any great degree. (This overlap can actually be a part of a larger problem – see later when you learn about statistical significance.) The overlap means that where this happens, you couldn't be sure how many results were anomalous, in other words, weren't what you expected. Here, we expected that participants would recall more in the meaningful condition than in the non-meaning condition. But you can bet that some people might recall more of the non-meaning than meaning. No particular reason, it might just happen! It's because we are dealing with real people, not atoms and molecules.



Perhaps by now you are expecting me to tell you about a statistic that will tell you how your results are dispersed. But I bet you don't want to mess around with frequency diagrams and all that hassle. Quite right too. I'm leading up to standard deviation - which really is quite helpful. But first, some background.

What does the standard deviation actually do for us?

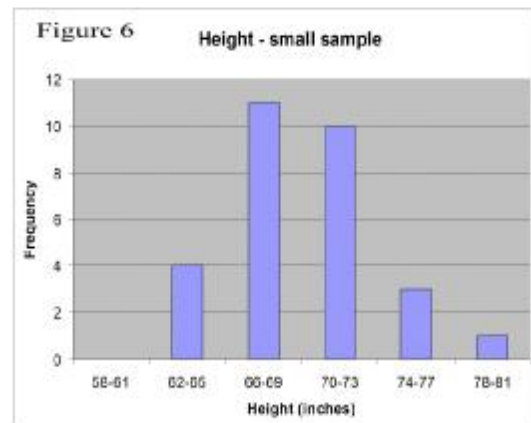
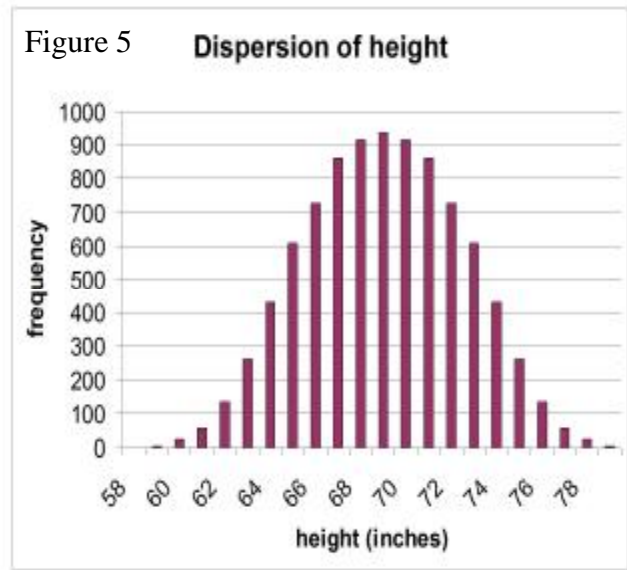
It is a descriptive statistic which gives us a measure of the dispersion of our data. In other words, it tells us how spread out our results are. But there is more. It is mathematically designed to be very helpful. To understand how helpful, you need to know about "the normal distribution" and the "normal curve".

It is a fact that most measurements of real life human attributes are "normally distributed", in other words, they fit a normal curve. You know what a frequency polygon is. Fig 5 is a frequency diagram of height of a few thousand people.

Intelligence is another attribute that is normally distributed. So also is heartbeat speed, foot size and many psychological attributes. For example, skill at remembering things is normally distributed. Instead of measuring thousands of people, you could perhaps only measure about 20. The spread of heights would still fit the normal curve, although it would be a bit jagged. And if you were measuring millions, the curve would be very smooth. But it would still be the same shape.

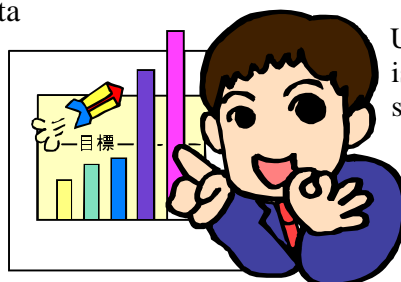
Provided the data are normally distributed, the standard deviation will be the same for the small sample or the mega sample – in fact the way it's calculated, you can say that what happens in your sample will be the same for the whole target population.

The normal curve has a special relationship with the standard deviation. If we say the area under the curve is the number of data points (and if you look carefully and think about it, because it is a frequency polygon, it must be the case) then the standard deviation can tell you how much of your data is in a particular section.



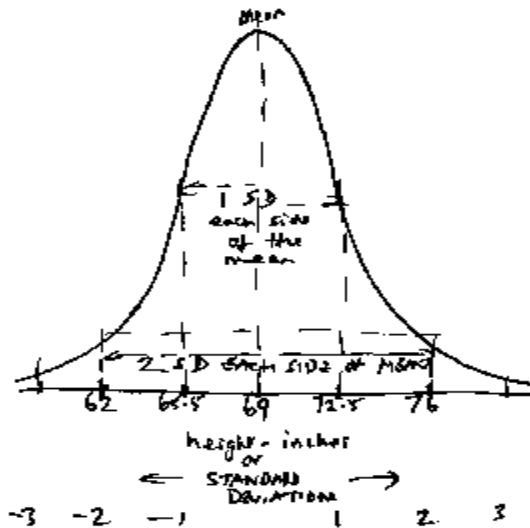
How do we calculate it?

I use a computer. The formula is easy if you know how, but it is sophisticated. Basically, you calculate the average of the differences between the mean of the data and your individual data points. Then, in order to get rid of annoying minus signs, you do a very magic trick – you square and then take the square root. That way, the standard deviation is always positive.



Using Microsoft Excel, which I guess is one of the most common spreadsheet programmes, you use the spreadsheet function STDEV

Figure 7.



The mean of the height data is 69 inches, and the standard deviation has been calculated as approximately 3.5 inches. In Fig 7, the shaded part is the area under the curve between 65.5 inches and 72.5 inches. That is, mean - 1 SD, or 69 - 3.5 inches, and the mean + 1 SD, or 69 + 3.5 inches.

2 standard deviations go to 62 inches and to 76 inches. The area under the curve lying between -2 SD and +2 SD accounts for 95% of the data points.

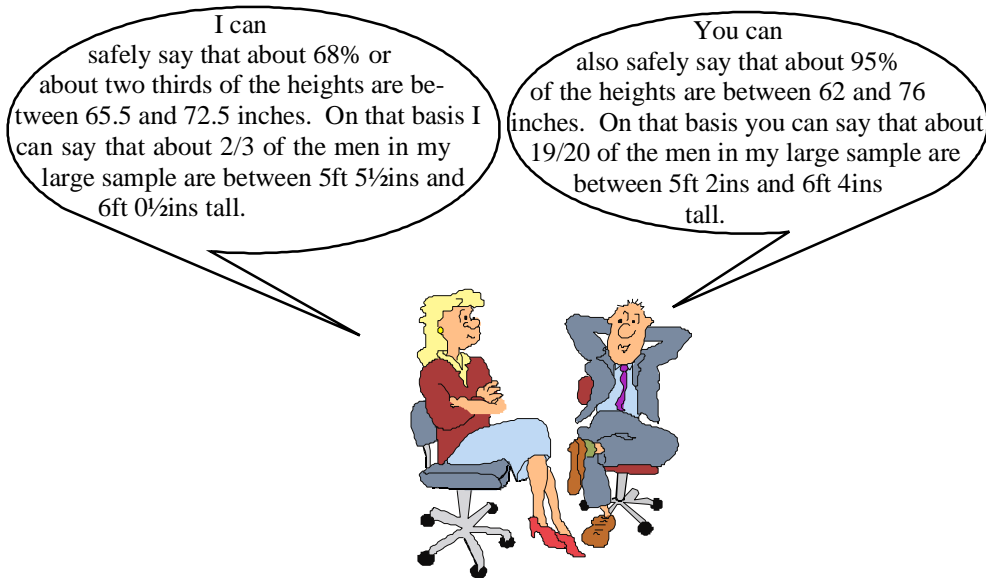
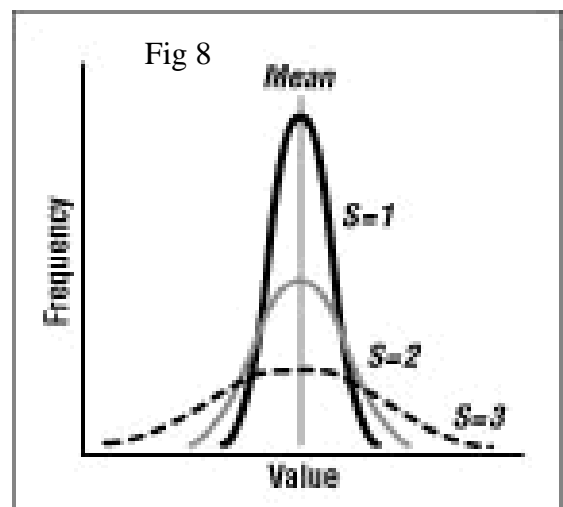


Figure 8 is a graphic representation of different standard deviations and how they are affected by the spread of data. A small SD, ($S=1$ in fig 8) shows a nicely peaked distribution, while a larger SD ($S=3$) shows data which is very wide.

Ideally, your data needs to be “tight”, as in $S=1$. So, look at the relationship between mean and SD. A standard deviation of about $\frac{1}{4}$ the mean indicates that $\frac{2}{3}$ of your data will be quite close to the average.

That is a very good reason to know the standard deviation. So, when you report your results, quote both mean and standard deviation!



Selection Of Participants Or 'Sampling'

What is Sampling?

When we sample, we choose people from a group who will take part in our research. It might be a good idea at this stage to learn some of the technical terms we use in sampling.

Participant

This is the term that we use for the poor unsuspecting person who is going to help us in our research – ethics permitting!! Previously, we used 'subjects', but that term implied a machine like being with no autonomy that was just carrying out instructions. A participant is someone who takes part and who interacts with the experimenter.

Population

'All of the cases within a given definition, from whom a sample is taken'. In simpler terms, this refers to everyone that we are sampling from. It might be 'all the people in the world' (unlikely!) or perhaps 'all the students aged 16-19 in the college', or even 'all my friends and family in Lincolnshire'

Target Population

This is a more specific way of saying 'the population you have selected from'.

Sample

The sample is the group of participants you have selected from the target population.

Generalise

To generalise is to say that what is true for the sample is also true of the population from which the sample has been drawn. This is in fact what we aim to do.

Sampling Bias

If there is a disproportionate representation of a particular subgroup in the sample, then sampling bias may be a problem. For example, if your sample contained many more females than males. Beware of this, but it isn't really important unless gender (or whatever is disproportionately represented in the sample) is an issue in the research. The same can be said for age, or socioeconomic status, or whatever.

Sampling Method

For basic, simplistic purposes, the methods can be divided into

- Opportunity Sampling
- Random Sampling
- The others



Opportunity Sample

This is the sampling method which will most likely be used by students doing coursework. It is cheap, convenient and quick.

Basically, it's asking the first people who come along.

The main disadvantage is that you can't be sure it is representative, so therefore you cannot generalise from an opportunity sample to a target population.

Random Sample

This is also often used by students doing coursework.

The definition of a random sampling method is that it is one where every participant has an equal chance of being selected from the target population.

There are two main ways of obtaining a random sample. The first involves obtaining all the names of the target population and drawing the participants' names from a hat. Another way uses a random number generator – a piece of computer software usually found in a spreadsheet – and matching the names to the random numbers.

The main strength of a random sample is that the results can be generalised to the target population. There are however some practical problems which may occur when, as a student, you are dealing with fairly small samples. For example, a random sample of 10 participants may contain one female, i.e. it may not be representative of the original target population. Having taken great pains to select a truly random sample, if some participants refuse to take part it would be difficult to replace them, especially as the sampling method is time consuming and probably expensive.

A common mistake is the belief that by choosing at random from people walking past in a street, or a college corridor, would constitute a random sample. This is not a random sample. The reason is that we would be biased in our choice of participant, particularly if the person gave out non-verbal signals that they did not want to be chosen. So, another advantage of the truly random sample is that it will be unaffected by experimenter bias.

Other sampling techniques

Quota sampling is where the sample selected contains specified groups in proportion to the mix in the target population. For example, a target population may contain 80 men and 120 women. A quota sample selected from this population might be contrived to contain 24 women and 16 men. This has the advantage that, if the

proportionate different groups were a methodological issue, then the sample would reflect the whole group in that respect. But it is not a truly random sample and there would be problems in generalisation to the target population.

Stratified sampling is a combination of random and quota sampling. The researcher identifies the different groups in the target population, say males and females. As in the quota sample example, let's say that the proportionate mix is 120:80 men: women. The stratified sample would consist of 24 men and 16 women, but the participants would have been selected at random from within the groups.

Similarly, **Cluster sampling** is selected at random from geographical clusters, rather than groups. Cluster and stratified sampling are efficient ways of selecting people into a sample, but they are not always representative and may not be generalisable to the target population.

Systematic sampling is the method whereby every n th person is selected from a list. There will be no experimenter bias in this method, but it may be over representative in, say one or other sex if we are choosing every 10th name and the list is paired by sex, such as an electoral roll!

A **self selecting sample** is one to beware of. Participants who answer an advert for volunteers to take part – which would be a self selected sample – might have particular reasons for coming forward. And the sample would only be representative of the people who saw the advert in the first place. Similarly, a sample of Sun Readers is definitely *not* representative of the British, I don't care what they say!!

Probability and Testing. Do my Results mean anything?

HOW MANY BEANS IN A TIN?

Have you noticed that sometimes there seems to be less beans and more sauce than usual? This is most probably because the bean hopper is nearly empty. The tins are packed by weight and the extra beans would be replaced by bean juice. I don't know how many beans there are to a tin, but let's say that a 500gram tin has nominally 1000 beans. Sometimes there may be 900 beans to a tin, and sometimes as many as 1100 beans to a tin (but nobody complains about that, do they?) But generally, we accept some variation in tins of beans. Apart from process deficiencies, there are also individual differences in beans – some may be bigger than others, for example. The question is, how much variation do we accept before we complain that our tin of beans is unacceptable?

Would we notice 5% of difference, that is down to 950 beans? Would we say that “on average, there's about 1000 beans to a tin, but I know there may be some variation. If there's only 950 beans, I'm going to refuse to accept that tin of beans.”

What you are actually doing is setting a level of significance that will help you make a statistical decision. The possibility of getting a tin of beans with as few as 950 beans is so small that there is something going on and I will not accept it. You have chosen a significance level of 5%. Under normal conditions, there is only a 5% chance of getting a tin of beans like that.

We'll come back to this later...

LIFE IN LAS VEGAS

Imagine the Grand Casino in Las Vegas. In America, they play a game called craps, which involves throwing a pair of dice and betting on the result. Imagine instead a simple game of dice. A dice is tossed and players bet on the number they hope is going to come up. Let's say that Fred Punter is betting on fours. If everything is above board, no cheating (by players or the casino!) and the dice is perfectly made, then we would expect four to come up once every six times.

The dice is rolled. It comes up four. Fred wins. He's feeling lucky, so he bets again. On four.

The dice is rolled. Four again! Lucky old Fred! He pockets some of his money, then bets again. On four. The dice is rolled. Four again!! Fred wins again. People are gathered around the game table. Will he play again? The pit boss is getting nervous.

Fred bets again, and big this time. The dice is rolled. FOUR!! A loud cheer goes up. The floor manager comes to look at the game. Fred is starting to feel really lucky. He bets again, on four.

The dice is rolled... FOUR!! AGAIN!!!. Fred wins. The pit boss closes the game, pending an enquiry. A nearby spectator is heard to say "What's the chance of that happening, I wonder?" Someone else says "There's something going on here!"

We'll look at this sequence of events in detail. First of all, we have said that, if everything is above board the chance of getting a four is one in six. The chance of getting a four is always one in six. But the chance of getting two fours in succession is $1/6 \times 1/6$, or $1/36$. The chance of getting five fours in succession is

$1/6 \times 1/6 \times 1/6 \times 1/6 \times 1/6$ or $1/7776$

In decimals, that's 0.000129.

How many fours in succession would you accept, if you were a pit-boss? Three? Four?

The pit boss was not prepared to go beyond five. It might have been less if he'd been there earlier! Why? Because he believed that the possibility of getting throws like this if everything was above board is so small that he would not accept that there was not something going on - an alternative possibility, cheating perhaps, or a defective dice. And if the alternative possibility were true, then a run of several fours would be quite normal.

SO WHAT DOES THIS HAVE TO DO WITH PSYCHOLOGY?

Although it is not immediately apparent, we have covered issues that include null and alternative hypothesis, statistical significance, probability, to name a few. But we'll start with one of the simple (imaginary) memory experiments: participants took part in two conditions of a memory recall experiment. Group 1 learnt and recalled a series of words in silence, while Group 2 learnt and recalled the words with soft classical music playing. The results showed that Group 2 with music were able to recall more items (mean = 24) than Group 1 in silence (mean = 20).

On the face of it, we could conclude that a theory that the presence of soft, classical music aids learning is correct. But the spread of results was such that some of the non music group recalled more than some of the music group. In other words, there was some overlap. It was only 2 out of 50 participants, but if we were testing a theory that soft classical music aids learning, we couldn't say on the face of it that the theory is correct. But we can use statistics to calculate the probability of getting these results again – in other words we can relate

to the Las Vegas pit boss. How many participants must be helped by the music before we can accept that there is something going on, that there is an effect there? And then we can think of the tins of beans. What percentage are we prepared to allow before we say that we are “within tolerance”? In psychology, particularly at student level, the accepted tolerance is 5%. In other words, if the statistic we can calculate from our results says that there is a less than 5% chance of repeating these results if nothing is going on, then our theory is correct.

Now we can get a bit more technical.

The state of affairs that would exist if there really was “nothing going on” is related to the null hypothesis. The null hypothesis is really a function of the means of populations. By mean, we are talking about the “average”. By population, in this case we are talking about all the possible means of all the possible results we could get if we repeated the experiment an infinite number of times. A mathematical statement of the null hypothesis would be:

$$H_0: \mu_1 = \mu_2$$

In non-technical language that would say:

“Under the null hypothesis, the population means of condition 1 and 2 are the same.”

The symbol “ μ ” is the Greek letter mu, and here it stands for “the population mean”. It is the mean value of all the possible scores that could be obtained from this experimental condition, if it were generalised to that part of the target population who carried out that condition. (I have to say that population here is one of the most confusing and ambiguous words going. It was probably designed by lawyers, not mathematicians. Why am I telling you this? So that you can read the text books!)

But the null hypothesis is not the true state of affairs. It’s like saying that the null hypothesis exists on a separate planet, where the effect we’re looking for doesn’t exist. So if the null hypothesis isn’t true, then we must find another explanation, to bring us back to our planet. We need an alternative hypothesis. Under the alternative hypothesis, the two means would not be the same. You can probably guess now what the next line means:

$$H_A: \mu_1 \neq \mu_2$$

“The alternative hypothesis is that the means of condition 1 and 2 are different”

And, if we are going to accept our theory as correct, then we need to specify how difficult it would be to get the same results again. We say that the possibility of these results being

replicated under the null hypothesis is less than 5%. Mathematicians say that the level of significance is 0.05. This is the probability expressed as a number between 0 (Impossible) and 1 (certainty). And so, the correct way to describe the results would include, in the technical way:

The results show that participants recalled more from condition 1 (with music) (mean = 20) than from condition 2 (without music) (mean = 24), and that the results were significant $p < 0.05$. Therefore the null hypothesis can be rejected and the alternative hypothesis accepted.

(Actually, there are two more pieces of information required in the result – the statistical test that was used and the critical value of the results. More later.)

And now, here is the real low-down on inferential testing. There are three stages:

- Assume that the null hypothesis is true
- Calculate the probability of getting these results if the null hypothesis is true
- If this probability is small enough, reject the null hypothesis

Now, we don't actually want the null hypothesis to be true, and in reality it is unlikely to be completely true. But it is a starting point – we assume that it is true and then try and show that it isn't. (This is a bit like saying that the defendant in a trial is innocent until proven guilty. The aim is to break down the starting assumption.)

The calculation of the probability of getting the results if the null hypothesis is true is the purpose of the inferential test. We want this probability to be as small as possible.

If the probability is small enough – well, how small is small? Back to the tins of beans. We would probably accept about 0.05 to 0.1. (Remember, probability is expressed as a number between 0 and 1). The Las Vegas Casino Manager stepped in at $p = 0.000129$. The level usually accepted by psychologists is 0.05. Why that figure? Well, it is partly arbitrary, but is also based on the possibility of making a rather unfortunate series of errors.

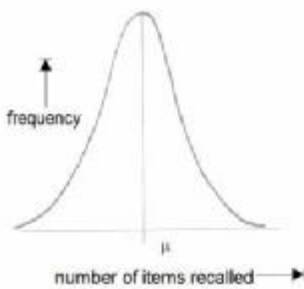
The type of error we can make depends on how strict is the level of significance. If it is not strict enough, then we might reject the null hypothesis when something really is going on. It's a bit like hanging an innocent man on weak evidence. Statisticians call this a Type 1 Error. On the other hand, if the level of significance is too strict, then we might keep the null hypothesis when we should be rejecting it. This is equivalent to letting the guilty person go free because the evidence isn't strong enough to convict. This is a Type 2 Error. The 0.05

level seems a reasonable trade off between these two types of error.

ABOUT CURVES AND DISTRIBUTIONS

If we were to carry out the memory experiment loads and loads of times, and count the frequency of recall for each condition, we could represent the results as a frequency distribution. (Sometimes called a distribution curve, frequency curve or bell-shaped curve, Figure 1) There would be one curve for each population, or experimental condition

Figure 1: A normal Distribution Curve



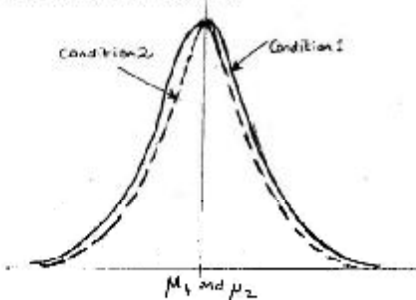
Under the null hypothesis, the curves should be virtually the same, because

$$\mu_1 = \mu_2$$

Remember?

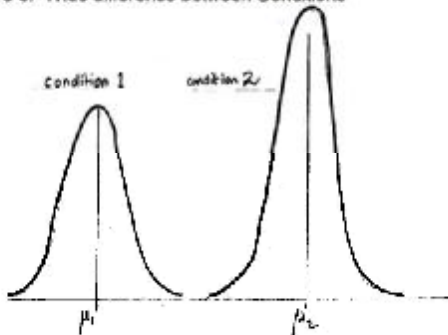
This is illustrated in Figure 2.

Figure 2: Distribution curves for H_0



If the effect we are measuring (the something that is going on) is very strong, then all the recall should be close to the mean, especially if there are no other effects interfering. If there is a wide difference between the two means μ_1 and μ_2 there are going to be two distinct spikes with little or no overlap. (Figure 3)

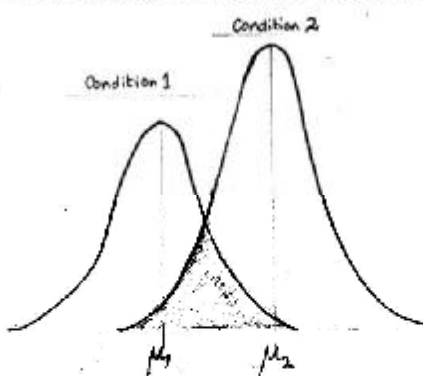
Figure 3: Wide difference between Conditions



But if the effect isn't very strong, or there are other effects at play, (extraneous variables, perhaps?) or the population consists of participants with a wide range of memory ability, then the frequency curves might be a bit more diffuse. (Figure 4)

We couldn't be sure if individual results are associated with population mean μ_1 or μ_2 .

Figure 4: Closer Means. We can't be sure that recall in the shaded area is from Condition 1 or Condition 2

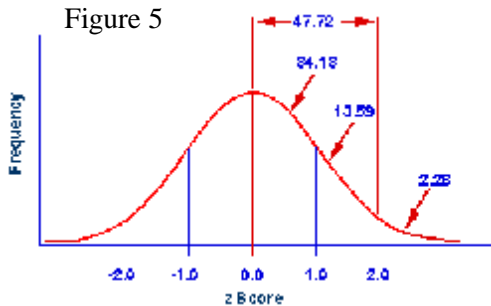


But all is not lost, because inferential tests which deal with differences in means are used to calculate the probability of getting the results under the null hypothesis, while at the same time taking the overlap into consideration. Basically, the more the overlap, the more likely we are to get "null hypothesis" results. On the other hand, the less the degree of overlap, the less the probability of getting H_0 results.

Remember the standard deviation? Well, now you can see the importance of that descriptive statistic. The smaller the standard deviation in relation to the mean, the tighter the frequency curve, hence the less likelihood of support for the null hypothesis.

A TIGER BY THE TAIL – MORE ABOUT HYPOTHESES

When we carry out an inferential test, we end up with a statistic (i.e. a number) which is related to our results. Imagine all the possible sets of results we could get from our population. Some of them will be extreme values, that is very unusual values – so unusual that the probability of getting these results under the null hypothesis is high. Now, I advise you to read that statement again. What I am saying is that there will be some values that are so extreme we would have to retain the null hypothesis.



Now look at figure 5. This again is a normal distribution curve.

The horizontal axis shows the number of standard deviations from the mean. (This number is also known as a Z-score, where $Z = \text{number of SDs from the mean}$). The diagram shows that 47.72 % of the scores are between 0 and 2 standard deviations from the mean. As a rough guide, we can say that the area under the tails outside the 2 standard deviations line is about 5% (I know, it's 2.28% under each tail, but let's keep it simple for the sake of explanation!) of the total area under the curve – this can be a useful piece of information if you have the occasional extreme result. But for now, think of it this way: Of all the possible results you can have, all the possible combinations of scores, if they are normally distributed, roughly 5% of them will be shared between those two tail ends. Or 2.5% at each end. If we are only considering positive deviations from the mean, we say we have a direction. Beyond that 2SD line are 2.28% of all the results. They are extreme results, and the probability of getting those results is small.



Army cadets in a normal curve.
The probability of being on a tail is very small

We are saying, if we are accepting a 5%, or 0.05 level of significance, that 95% of our results are not in the extreme area, while 5% are. In terms of memory recall, we are saying that people will recall more, or less, between two conditions, in other words we can state a direction of difference in recall. So, we can say that all 5% of our extreme values will lie at one end of the curve, or tail. If we can't say the direction, but only that there will be a difference in recall, we must accept that there may be some extreme values at each end of the tail. This will increase the probability of getting the results under the null hypothesis, and reduce the credibility of our results, and hence our theory. But that makes sense, after all. If we don't (or can't) specify a direction of difference, only that there is a difference, then we are not being so precise. We have to take into account the extreme results at both ends of the curve, or both "tails". One- and two-tailed- hypotheses should by now be connecting with this. The importance is, if we are counting extreme results under two tails, then there

are likely to be twice as many, or 10%. This is why tables show critical values for two-tailed tests at twice the corresponding level for one-tailed tests.

INFERENCEAL TESTS

We have already seen there are three steps to testing a hypothesis:

- Assume that the null hypothesis is true
- Calculate the probability of getting these results if the null hypothesis is true
- If this probability is small enough, reject the null hypothesis

Inferential tests are basically a mathematical procedure for calculating the probability of getting the results if the null hypothesis is true. The calculation we (or the computer!) make produces a statistic, which is nothing more than a number representing that probability. We compare our observed statistic with a previously determined number known as the critical value. The critical value is the value of the statistic corresponding to our accepted level of significance, and depending upon the type of test we are using, we can make a statement as to whether the results are so extreme that we must reject the null hypothesis in favour of an alternative hypothesis.

DOING THE TEST

We need to generate a statistic from our data, or results, to calculate an associated probability.

The problem now is, *how?*

Mathematicians have solved the problem for us. There are a number of different statistics available and all you have to do is choose the correct test. They all depend on the following:-

- Whether you are analyzing an experiment or a correlation
- The type, or level, of data that has been collected
- If it is an experiment, the design
- Is the data from a normally distributed population?

Let's unpack these one at a time.

(a) Are you analysing an experiment or a correlation?

Remember an experiment is usually a comparison of data from two experimental conditions. A correlation is looking for a relationship or an association between two variables. Experiments have bar charts and correlations have scattergrams.

(b) The type, or level of data

For our purposes there are three levels of data: interval/ratio, ordinal and nominal.

(i) Interval/ratio data uses the sort of measurements we are used to. Basically, the distance between 1 and 2 is the same as the distance between 3 and 4 and so on. The distinction between interval and ratio data is important for some disciplines, but not for A-level social scientists. (It revolves around a thing called an 'arbitrary zero'. Temperature scales mainly have a zero which is near the freezing point of water. But there is an absolute zero, which could be utilised. However this is irrelevant for our purposes.)

Interval data has an arbitrary zero, and ratio data doesn't. Since you were learning to count you were likely to have been ratio data. Hence 'interval/ratio' data. There is a move towards calling it scale data, particularly in some statistical software applications.

(ii) Ordinal data

Sometimes called 'ranked' data. You may have come across it in two main designs:

- from a rating scale in a questionnaire
- from e.g. 'put in order the 5 most enjoyable...'

Rating scales

We often use two main types of rating scales (there are others!) for A-level practical work. We use either a Likert scale or one derived from it. A statement is presented and respondents indicate their agreement:

Very strongly agree	Strongly Agree	Agree	Neither agree nor disagree	Disagree	Strongly Disagree	Very Strongly Disagree
1	2	3	4	5	6	7

When the levels of agreement are quantified, i.e. represented as a scale of 1 to 7, you can see the difference between 1 (very strongly agree) and 2 (strongly agree) is not necessarily the same as the difference between 2 and 3 (agree). Another way to think of ordinal data is in terms of a horse race. The race was won by two lengths, with the third horse half a

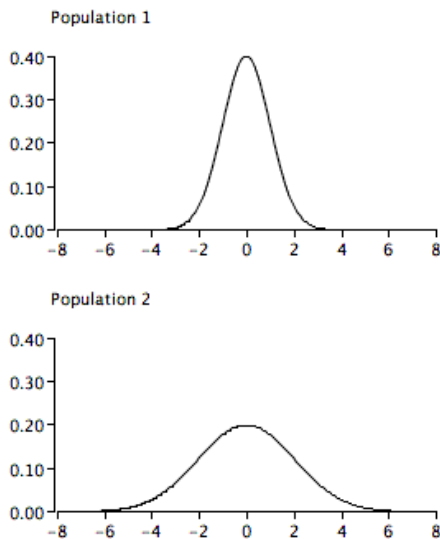


Figure 6: Results data from two conditions with dissimilar variance.

length behind. First, second and third are not equally spaced. This is ordinal data.

(iii) Nominal data

This is data representing categories. What we actually measure is the number of items or responses or whatever in these categories. Examples are male-female, young – old, or questionnaire responses in categories. Such data is often represented as ‘contingency tables’ or sometimes as ‘crosstabs’. For example, a psychologist was investigating types of aggressive behaviour displayed by old and young people. The data could be represented thus:

	Hitting	Swearing	Tutting
Young	27	30	3
Old	4	18	32

This is an example of nominal data in a contingency table.

(b) The Design of the experiment

The design in statistical tests is usually stated as either “related” which is equivalent to repeated measures, or unrelated, equivalent to independent subjects.

The concept of design has been dealt with earlier.

(d) Is the data drawn from a normally distributed population?

Generally speaking, when we test natural attributes, the data can be considered as normally distributed. A person’s height is part of a natural distributed set of data. So also is memory. height is something we are born with and different people have different heights – some are very tall, some are quite short. Similarly, memory ability varies between people and is also normally distributed.

Important: Some data is from e.g. IQ tests and personality questionnaires. Often, these tests state they are standardised, a mathematical procedure which adjusts the data so that they are normally distributed.

(e) are the variances similar?

This is known as ‘homogeneity of variance’. Is the spread of the data similar under each condition? The variance is the square of the standard deviation, so it follows that similarities can be estimated by looking at the standard deviation and the

mean. Figure 6 shows two sets of data with dissimilar variance.

There is a statistical test (Levene's test) which is available on advanced statistics packages, e.g. SPSS, but for A-level purposes, a guesstimate via standard deviation would be adequate.

PARAMETRIC TESTS

The most robust statistical procedures for testing our data are known as parametric tests. In order to "qualify" for a parametric test, the following conditions must be met by your data:

- (i) The level must be interval/ratio
- (ii) The sample must be taken from a normally distributed population
- (iii) There must be homogeneity of variance

Internal Validity

Were my results due to the effect I thought I was measuring?

External Validity

Can my results be generalised if conducted in different environments or using different participants?

If your data do not meet these criteria, all is not lost! Most of your data will in any case be analysed using non-parametric tests because of uncertainty over (ii) and (iii) above or because the level of data is not interval/ratio.

Most text books have some sort of flow diagram or decision tree to help you choose the correct test.

How to carry out the tests can be gathered from most text books if you wish to use the pencil and paper method, or there is no shortage of statistical software, some of which is free online.

RELIABILITY AND VALIDITY

These two concepts are extremely important when both designing and later evaluating research. But what are they? Basically, Validity means are you measuring what you think you are, and reliability means – erm – are my results reliable?

Let's take reliability first. **Internal reliability** refers to how consistently a method of measurement measures itself. Think of a ruler which is not graduated accurately or evenly. You would not be certain if the measured length was exactly 2 centimetres or 3 centimetres. It did not have internal reliability.

External reliability is lacking when I weigh myself on the bathroom scales and get one reading, then measure myself again on the same scales and find I am a kilogram lighter. The scales do not measure consistently and are therefore unreliable. External reliability refers to how consistently a method measures over time when repeated.

Examples in psychology often involve questionnaires or inventories, such as an IQ test. To test for internal reliability, some of the questions will be similar. If these questions give similar responses, then the test will have internal reliability. The split half method for testing reliability is similar. The test is carried out twice, but each half is correlated with the other. Again, the higher the correlation the better the internal reliability.

The test-retest method for investigating external reliability is as its name suggests. The test is administered and scored, then the same or very similar test is administered after, say three weeks and the two sets correlated. A high positive correlation indicates good external reliability.

Validity is really a way of asking 'am I measuring what I think I'm measuring, and if so, how well?'

We should look at two types of validity – internal and external. There are several 'sub-types' of validity, and they can all be classified as external or internal.

Construct Validity:

Does what I am trying to measure actually exist? For example, does IQ really exist? IQ is a construct, and we know that it is demonstrated by differences in IQ scores. So IQ has construct validity.

Concurrent Validity and Predictive Validity

If the method being used produces the same or similar results to an existing method, then it has concurrent validity. If it will predict future performance, e.g. in an IQ test, then it has predictive validity.

Content Validity and Face Validity:

If we are measuring aggression, then we would expect our test to have some reference to, e.g. violent behaviour. If so, it would have content validity.

Does our questionnaire look like a survey into eating habits – if so, it has face validity

Ecological Validity:

Does the method used appear to be rooted in the real world? Most memory experiments involve learning and recalling strings of items in a laboratory setting, which is not really how we remember experiences in real life! A criticism of laboratory experiments generally is that they do not have ecological validity.