**13 A chapter on probability? What are the chances of that happening? How likely it is that all the sections are named after Sherlock Holmes books?**

The red headed league

I have decided that I want to find out how many red headed people live in the UK. I could look at every person in the UK. (Known as the **Population**). This would be called a **census.** But there are 66.65 million people in the UK. So I thought I would just take a sample and count them. I have chosen the island of Ivorbowlocockileekie in Scotland. It has a population of 6665 so it makes the maths nice and easy.



And I discover 60% of the population have red hair.

I am on the train back to Birmingham, when I realise I have made a mistake.

Yes I should take a sample, but I need to be more careful.

**The sample must be selected at random from the population** and the number selected must be as the result of a **power calculation**. If I select a number that is too small, I will never be able to find a statistically significant result even if it exists. Whereas if the number is too great, I may find a statistically significant difference that is not clinically significant.

The sign of the four

The thing we are looking out for is known as a **variable** and recording it makes it a **datum** record, more than one is **data.** There are two types of data:

1. Non Numerical or **Qualitative** (such as red hair)
2. Numerical or **Quantitative**

Quantitate can be subdivided into **Discrete** (such as shoe size or money)

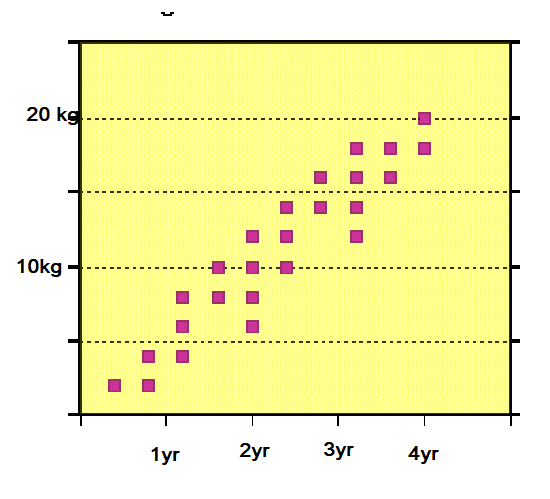
Or **Continuous** such as height or weight.

A word of warning.

Some things appear to be Quantitative data but are not really. For example PAR scores. A score of 4 is not necessary twice a score of 2 or half a score of 8 because the score is made of many factors and some of them are multiplied by a weighting factor. The same is true of periodontal charting it uses numbers but it is not really quantitative.

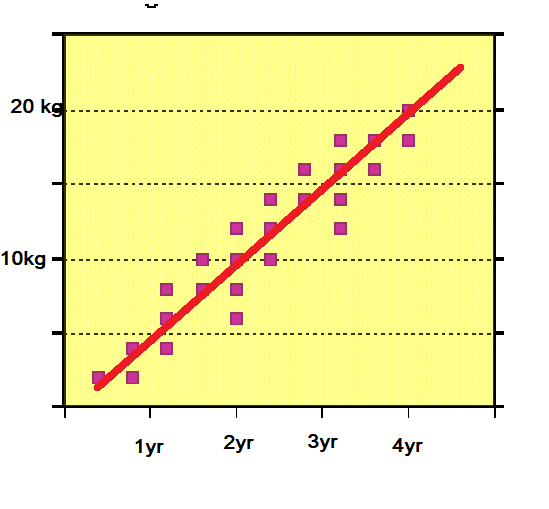
Correlation.

If we plot weight against age for gorillas we may get this:

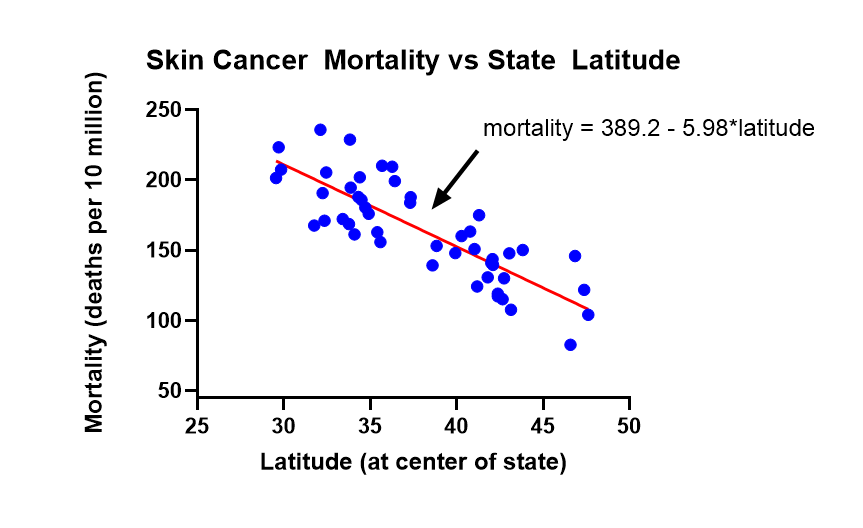


Not unexpectedly, as they get older they get heavier.

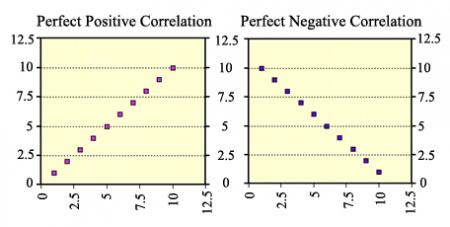
We can draw a line indicating the trend.



Because the line goes upwards this is known as a positive correlation and has a + sign

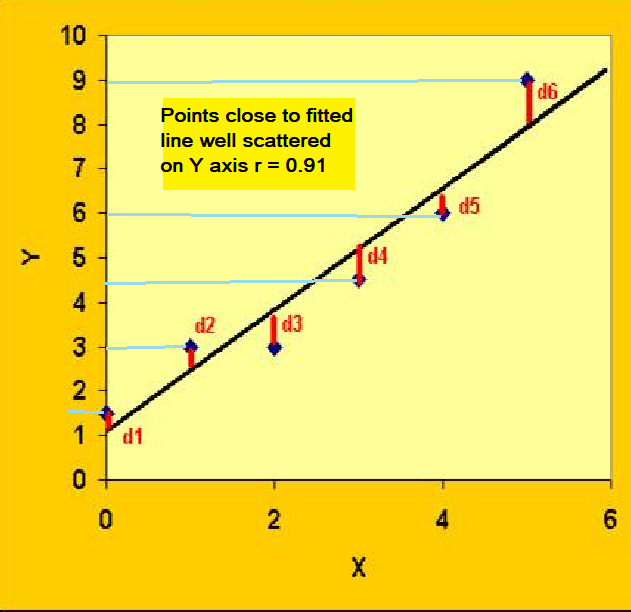


This is a negative correlation. The angle of the line is not important. You can change it by altering the units. A perfect correlation would be r =1 (either plus or minus)



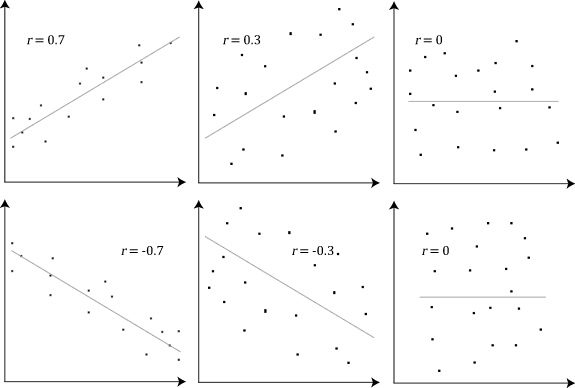
r = +1 r = -1

But in real life you will see **Scatter** and you can measure the scatter by measuring the distance of the points from the best fit line and relate it to the spread of the points measured from the y axis. If the points are close to the best fit line and spaced on the y axis. Then the correlation will be high. If they are away from the best fit line and are closer together on the y axis then the correlation will be low.



When r = 1 to 0.8 this is considered to be a very high correlation. (The same would be true of minus 1 to minus 0.8)

Here are some examples:



Correlation measures the strength of the association between two numerical sets of data. HOWEVER it is not a measure of causation Mobile phone use has gone up and family sizes have gone down, the figures may have a correlation coefficient that is r = - 0.6 but this doesn’t mean mobile phones **caused** a reduction in family size.

**r²** gives you a measure of how much one variable is predicted in the other. So -0.6 squared is r² = 0.36 which is very low.

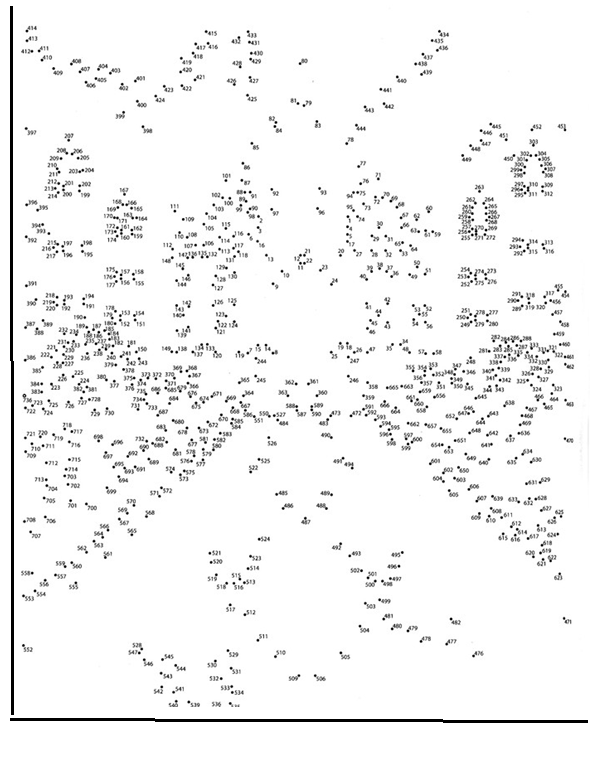
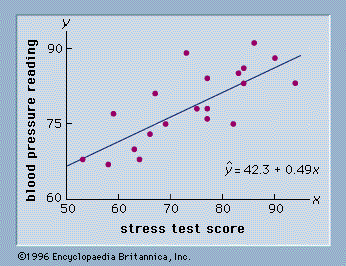


Fig r = 0 (I would not join the dots if I were you, I think it might be rude)

Simple linear **Regression.**

This is the one that looks like correlation, but here the slope of the graph is important. You have two variables one you know. The independent variable is controlled by the experimenter and the dependent variable changes as a result of changes in the independent variable.



The Hounds of the Baskerville’s

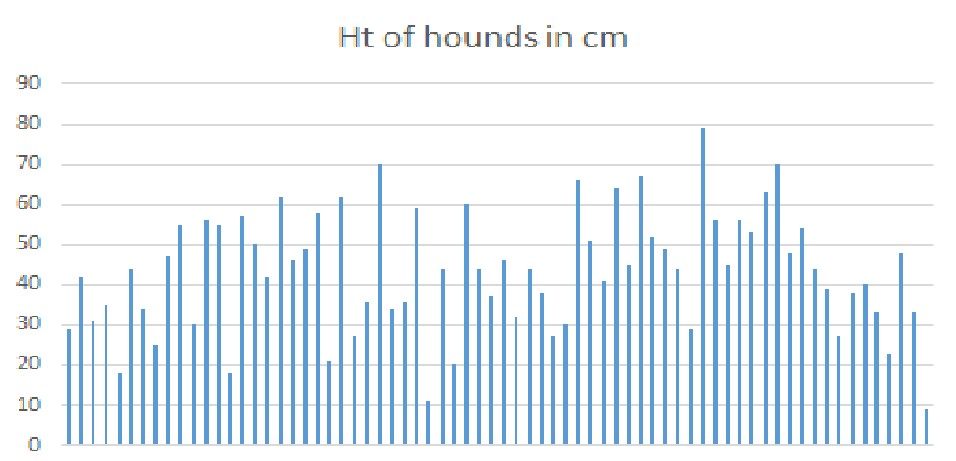
The Baskervilles love their dogs. They have hundreds of them (in this case that is the “population” under investigation). They believe their dogs are bigger and stronger than their neighbours. We have already learnt that we need to take a sample. And that must be taken at random using a random number generator. But we have a little problem. There are both male and female hounds and it is said that the males are bigger than the females. So we plan to take two separate random samples their size related to the number of males versus females in the population. We also plan to exclude all puppies from the group as they will obviously be smaller.

Our randomised sample contains 70 dogs varying in size from a massive Great Dane called Hamlet almost 80cm tall to a tiny Yorkshire terrier called “Richard the Turd” who is only 8cm tall.



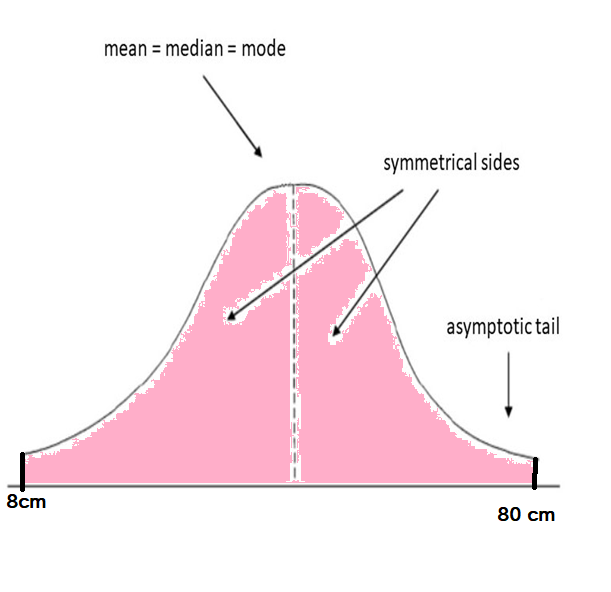
Fig Hamlet and Richard the Turd

We could make a simple bar chart of their heights based on their kennel number. But it would mean nothing at all.



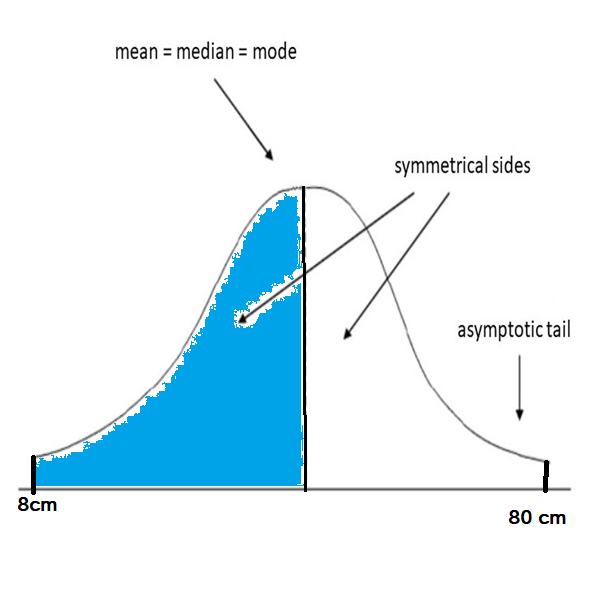
However if we make a bar chart based on their height, in this case with 8 cm boundaries something interesting happens.

Yes this is known as a **Normal Distribution** also referred to as **Bell shaped curve** or a **Gaussian curve** (all three terms mean the same thing). Just to confuse you in statistics normal distribution may also be referred to as **Parametric.**

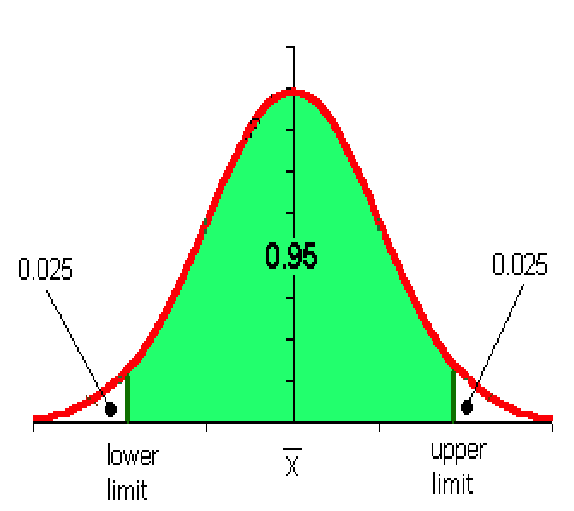


Expressed a simple graph it looks like this. The central line is the **mean** which is familiar to us as the arithmetic average (add together all the heights and divide by the number of dogs). It is also the **median** (Write down a list of all the heights in order and select the middle one). It may also be the **Mode** (the most commonly occurring value)

The area under the graph is the **probability**. For probability we use the term **p.** The highest possible value of p is p =1. This is the pink area in out graph and it tells us that 100% of our dogs are between 8 and 80cm. This is known as the **range.** The range can be quite a poor indication of the spread of data because it can be affected by outliers.



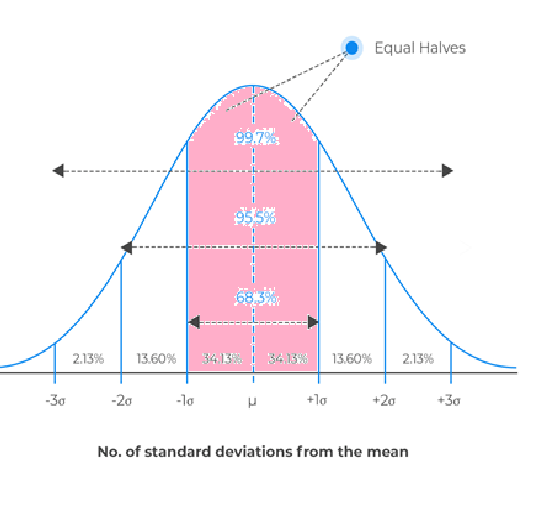
As we know the graph is symmetrical so that the blue area has a probability of p= 0.5 and this means half of the dogs have a height below the mean



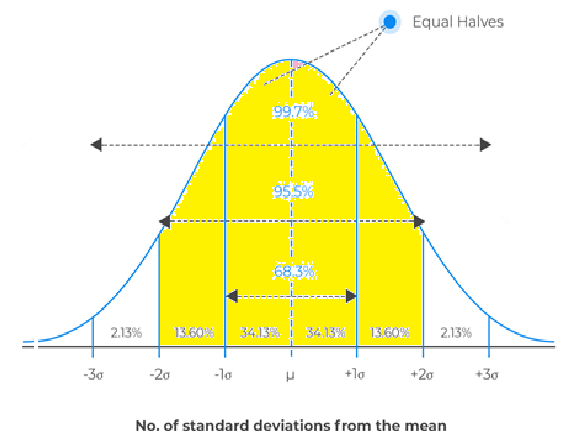
This graph shows you a normal distribution with the 95% confidence limits marked. Again it is the area under the graph that gives you the probability. There is a p=0.95 of a randomly selected value falling in the green area and a p=0.05 of it falling in the white area (Two areas each p = 0.025, these are the two tails. In some studies you may only be interested in one tail, for example individuals whose height is above the 95% confidence limit. This would be a single tailed study)

You will see 95 and 90% confidence levels in some papers but is more common to see **standard deviation. (Also called standard error of the mean)**

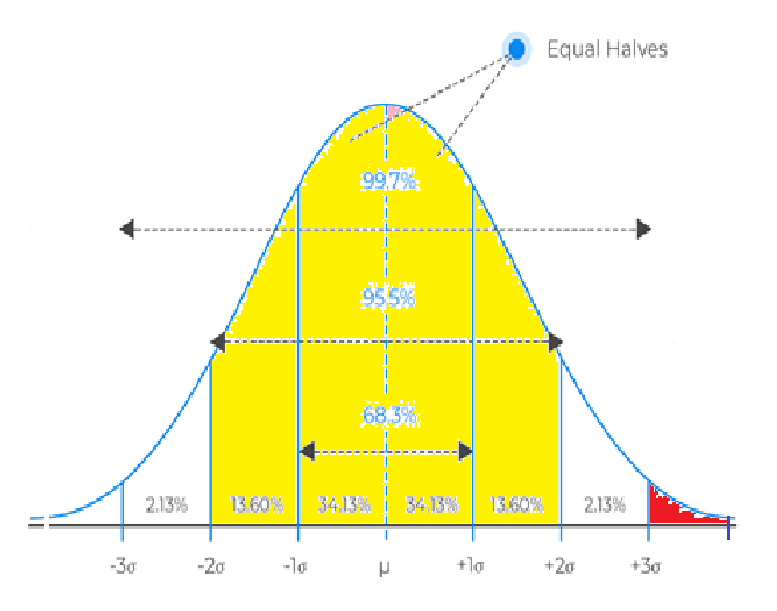
**Standard deviation** derives from the **variance,** a clever way of showing how widely the data is spread from the mean by squaring the distances from the line to the median. Because you square them you get rid of the problem that the positive figures on the right side of the graph will be cancelled out by the negative figures on the left. Because the square of a negative figure is always positive -2 x -2 =4 etc. So standard deviation is the square root of the variance. (If you have the 95% confidence limit there is a clever formula to calculate the Standard Deviation. Just ask your uncle Google)



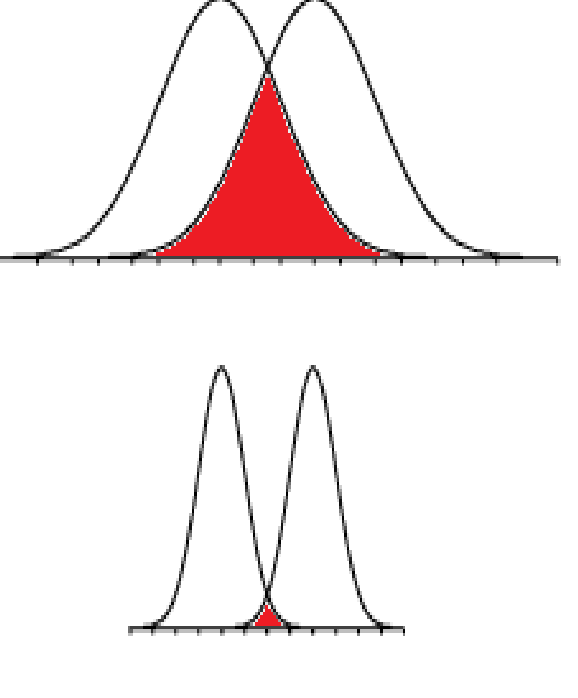
The pink area is one standard deviation from the mean there is a 68.3% chance that a random sample would fall in this area so p = 0.683



The yellow area is two standard deviations from the mean. There is a 95.5% chance that a random sample would fall within this area p=0.955 and there is a 99.7% chance that a random sample would fall within 3 standard deviations



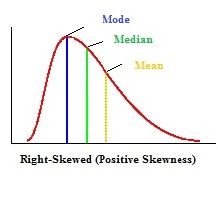
The probability of a random sample falling in the red area is p=0.0015 or 0.015%

If we have two normal distributions. Let us say the results of treatment A and the results of treatement B we can draw them together. The red area shows the probability that a random sample will show no difference in the treatments.

Clearly in the lower graph there is a greater separation of the two results with very little overlap. We would need very few tests to establish a difference between the two treatments. For the upper graph if we took enough random samples we would establish that there might be a difference. This is an indication of the **power** of the study. Power is the probability that it will detect a statistically significant difference.

In orthodontics the problem is that if you do a huge number of tests you may get a statistically significant difference but it would not be **clinically significant**.

Statistical tests using normally distributed data have a higher power. Sometimes the graph is **skewed,** an example would be weight of a human population

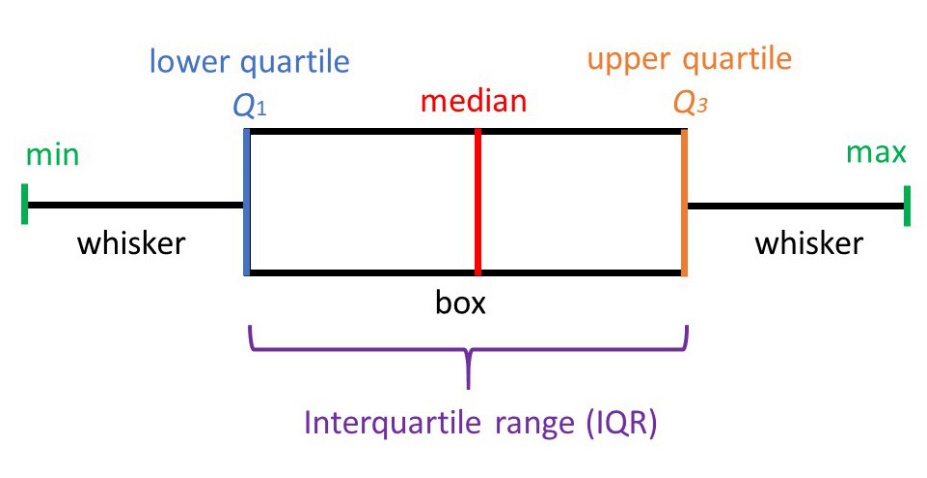


Which is skewed to the right, meaning there is a bigger scope to be overweight than underweight. Using a logorithmic scale on the x axis can convert it to a normal distribution. This is called **Logerithmic transformation**.

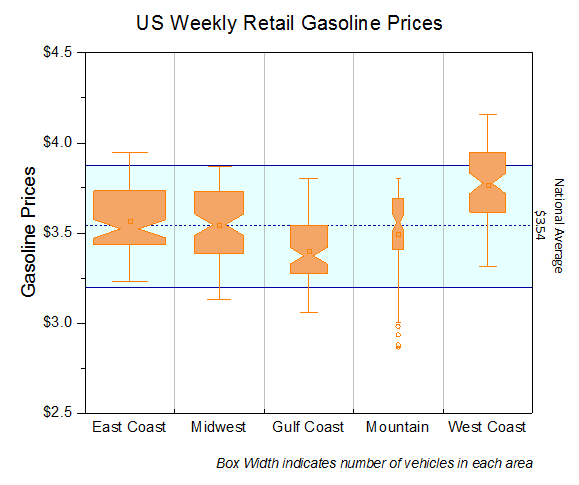


Fig Logerithmic Transformation.

To see median, range and inter-quartiles in action you might go for **Box-plot and whisker graphs.**



They can give you useful information in an easy to understand way. Sometimes the median is marked not with a line but with a notch. (But I find notched box-plot graphs quite difficult to say)



A study in scarlet



Let’s look at some of the really wicked things that you can do when undertaking research:

1. A poor sample. We have already mentioned that we must have a proper sample. We need to state clearly if some groups from the population are excluded. We need to randomise and if the population has relevant sub groups (for example male and female).we need to randomise maintaining the sub groups.
2. We must account for missing data. For example supposing you are doing a study comparing a functional appliance with extractions and some functional cases are not going so well. **You must not say” they are obviously not wearing their appliances” and remove them from the study.** A study where patients who are withdrawn must still be included in the analysis. This is called an **intention to treat study**
3. You must have groups that record all the outcomes. For example a famous paper by Eriksson and Kurol on treatment of unerupted canines showed that 78% of permanent canines erupted following the extraction of deciduous canines but there was no control group and it is possible that 79% could have erupted without extractions. So for good research we need a **Control** group and this group must be matched with our randomised sample so it is called **A Randomised Controlled Trial.**
4. If I am measuring my own results I may feel disposed to give the benefit of the doubt to my favoured appliance. So I need to get an independent person to measure the results so that I cannot bias them. Now we have a **Blinded Randomised Controlled Trial.** In some circumstances, for example drug trials, it is possible that the person giving the drug can be given a specially prepared drug for the trial (say marked trial A and Trial B tablets) and that they too have no idea which drug is being given. This is a **Double Blinded Randomised Controlled Trial.** Let’s suppose my trial involves patients treated at a Birmingham Hospital and a control of patients from their waiting list. Then there is a question. “Are Birmingham patients typical of the rest of the country?”

Fig Are Birmingham patients typical of the rest of the country?

The answer is to carry out the trial in many places at the same time. This is a **Multi-centred Randomised Trial.** A secondary benefit is that it makes recruiting numbers for the trial easier.

1. The control groups must be relevant. Some of the studies of functional appliances use control groups of a different age (in some it is the same patients either before or after the functional treatment) ideally they should have a similar skeletal pattern. You could expect that a control from a growth study with many class III cases would have different growth patterns to patients who start skeletal II.
2. Retruded contact and Intercuspal position. I confess it seems just to be me that is worried about this issue. It comes up in two different types of research:
3. Functional appliances. Yes you know Kevin O’Brian says 70% of the change is dento-alveolar and this implies that 30% is skeletal change. BUT this is measured to the B point and the B point moves forward if the mandible is postured forwards AND you often see the mandible postured forwards in patients who have been wearing functional appliances, especially the twin block because it tends to cause a posterior open bite and this encourages forward posture to get a bite. Do I believe that in a trial where the OJ is reduced from 7mm to 2mm that 1.5mm could be posture? You bet I do. And this would explain why some trials seem to show a skeletal change occurring in the functional phase that disappears during the fixed phase where the Retruded contact bite is re-established.
4. Face-Mask therapy. I repeat my little rant from the Chapter on class III malocclusions.

**Mandall *et al* JO 2010 37 p149**  a real RCT but start radiographs not in Retruded Contact Position. ANB increased by 2.1 SNA increased by 1.4 which was not significantly significant.

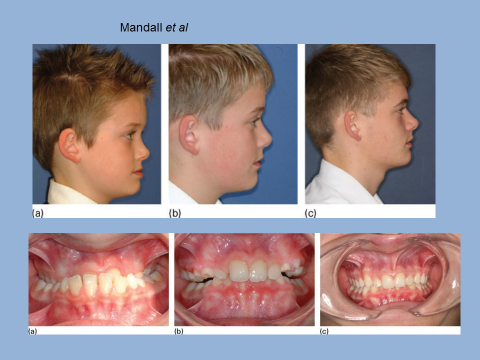


Fig this looks like a very severe class III malocclusion but I suspect he could bite edge to edge at the start.

This is an important paper.It has the EOF prize for the best scientific paper and the Bristol notes say it **proves** facemask therapy produces orthopaedic change.

**The case for:**

1. This is a genuine multi-centre prospective randomised clinical trial
2. Its authors are a group of well-respected British orthodontists.
3. It was published in a well-respected refereed journal.
4. Subsequent papers suggested more of the control cases needed surgery.

**The case against:**

1. The start radiographs were not taken in retruded contact position but in inter-cuspal position. It is well known that class III cases with positive overbites often have an initial contact edge to edge bite. Then they move into a different mandibular position with the posterior teeth in contact.

This was investigated by Gravely in 1984 BJO vol 11 pages 85-91. He found a difference in ANB of 2.7 degrees between RCP and inter-cuspal position in these cases. So an increase of 2.1 does not seem so impressive.

(*Just a word about Gravely. This paper used 50 pairs of radiographs one taken in RCP and one in ICP he concluded that as the mandible closed up into occlusion the chin came forwards as a result of the hinge closure of the mandible and with a reduction in the MM angle but not by a forward movement of the condyle which remained in the glenoid fossa)*

1. The appliance was designed to move the maxilla forwards but the increase in SNA was not significantly different from the controls
2. At any time they could have resolved the problem by taking radiographs of the controls in RCP and compared these radiographs with the treatment group, but they did not do so.
3. They concluded more of the controls needed surgery, but is this because they had their assessment done still on radiographs taken in inter-cuspal position and photographs again taken in inter-cuspal position. (They did not examine the actual patients.) Orthodontists have a 100 year history of attributing orthopaedic change that turns out to be an illusion.

Statistically what we see is two matched groups, both had class III malocclusions and like other groups of mild to moderate class IIIs many could achieve incisor contact but because the biting forces on the incisors was too high moved the mandible to a position where the molars are in contact. After Gravely we believe this is an upward and forward swing. Finishing in a position where the condyle is back in the fossa. The jaw has closed like a hinge.

At the end of active treatment the treatment group can bite in retruded contact position but the control group still have the upward and forward close. We are 95% certain that there is a difference between the two groups but is this orthopaedic change or because of the removal of the abnormal closing pattern?

Later on photographs and cephalograms of the treatment group were compared with photographs and cephalograms of the control group. Again all records were taken in centric occlusion so the controls were displaced from centric relationship. We still cannot be sure that the decision that the control group were more in need of surgery was a result of the treatment or the displacement,

1. Too small numbers. You need a power calculation. Far too many studies have concluded that there is no difference between two treatments. When in fact the study never had sufficient power to show a difference between the two groups.
2. Too large numbers. In some circumstances there is a danger if the numbers are too large because it may bring to light a difference which gets the magical accolade of “statistically significant” but the difference is too small to have any clinical significance.
3. Blunderbuss cephalometrics. I think it is a shame that some people are saying we have had enough cephalometric studies, but I do see the problem. Somebody carries out a treatment and takes cephs before and after and they do the same for a suitably matched control. The digitised results are put through a software measuring package which measures 10 different angles?

What is the probability of at least one of these angles being significant at the p=0.05 level? It would be (50%). Then when they write it up they forget to mention all the other angles and imply that this was the focus of their investigation from the start. No, you should select one or two measures which logically are related to the treatment. If the study seems to suggest some strange finding then you need a new study and a new power calculation.

1. Not deciding beforehand. Great care is needed at the beginning not just in assessing sample size, randomisation and the make-up of control groups. If the treatment involves a new appliance you need a **pilot study** for the clinicians to learn how to use the appliance and the laboratory to learn how to make them. You need to write a **Protocol** which defines the **Primary Outcome** and the research needs to be registered.
2. The research should be published even if it does not show a statistically significant difference between the sample and control groups. There is strong evidence that a lot more papers with a positive result are published. But this can cause bias. Imagine 50 papers on functional appliances 25 show statistically significant change and they are all published. 25 show no difference but only one is submitted for publication. A subsequent **Meta-analysis** will be biased.
3. The write up needs transparency so that another researcher can try to revalidate the results.

This looks quite complicated but we have some help in the form of **The Consolidated Standard Of Reporting Trials.**

As you can see the idea is based on the proper publication of a trial but it can be used as a checklist to use when setting up a trial.

Here is the Consort 2010 statement

**Title and Abstract**

[1a Title](http://www.consort-statement.org/checklists/view/32-consort/66-title)– Identification as a randomised trial in the title.

[1b Abstract](http://www.consort-statement.org/checklists/view/32-consort/67-abstract) – Structured summary of trial design, methods, results, and conclusions

**Introduction**

[2a Background](http://www.consort-statement.org/checklists/view/32-consort/69-background) – Scientific background and explanation of rationale

[2b Objectives](http://www.consort-statement.org/checklists/view/32-consort/70-objectives) – Specific objectives or hypothesis

**Methods**

[3a Trial design](http://www.consort-statement.org/checklists/view/32-consort/72-trial-design) – Description of trial design (such as parallel, factorial) including allocation ratio

[3b Changes to trial design](http://www.consort-statement.org/checklists/view/32-consort/73-changes-to-trial-design) – Important changes to methods after trial commencement (such as eligibility criteria), with reasons

[4a Participants](http://www.consort-statement.org/checklists/view/32-consort/75-participants) – Eligibility criteria for participants

[4b Study settings](http://www.consort-statement.org/checklists/view/32-consort/76-study-settings) – Settings and locations where the data were collected

[5 Interventions](http://www.consort-statement.org/checklists/view/32-consort/78-interventions) – The interventions for each group with sufficient details to allow replication, including how and when they were actually administered

[6a Outcomes](http://www.consort-statement.org/checklists/view/32-consort/80-outcomes) – Completely defined pre-specified primary and secondary outcome measures, including how and when they were assessed

[6b Changes to outcomes](http://www.consort-statement.org/checklists/view/32-consort/81-changes-to-outcomes) – Any changes to trial outcomes after the trial commenced, with reasons

[7a Sample size](http://www.consort-statement.org/checklists/view/32-consort/83-sample-size) – How sample size was determined

[7b Interim analyses and stopping guidelines](http://www.consort-statement.org/checklists/view/32-consort/84-interim-analyses-and-stopping-guidelines) – When applicable, explanation of any interim analyses and stopping guidelines

[8a Randomisation: sequence generation](http://www.consort-statement.org/checklists/view/32-consort/86-randomisation-sequence-generation) – Method used to generate the random allocation sequence

[8b Randomisation: type](http://www.consort-statement.org/checklists/view/32-consort/87-randomisation-type) – Type of randomisation; details of any restriction (such as blocking and block size)

[9 Randomisation: allocation concealment mechanism](http://www.consort-statement.org/checklists/view/32-consort/89-randomisation-allocation-concealment-mechanism) – Mechanism used to implement the random allocation sequence (such as sequentially numbered containers), describing any steps taken to conceal the sequence until interventions were assigned)

[10 Randomisation: implementation](http://www.consort-statement.org/checklists/view/32-consort/91-randomisation-implementation) – Who generated the allocation sequence, who enrolled participants, and who assigned participants to interventions

[11a Blinding](http://www.consort-statement.org/checklists/view/32-consort/93-blinding) – If done, who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how

[11b Similarity of interventions](http://www.consort-statement.org/checklists/view/32-consort/94-similarity-of-interventions) – If relevant, description of the similarity of interventions

[12a Statistical methods](http://www.consort-statement.org/checklists/view/32-consort/96-statistical-methods) – Statistical methods used to compare groups for primary and secondary outcomes

[12b Additional analyses](http://www.consort-statement.org/checklists/view/32-consort/97-additional-analyses) – Methods for additional analyses, such as subgroup analyses and adjusted analyses

**Results**

[13a Participant flow](http://www.consort-statement.org/checklists/view/32-consort/99-participant-flow) – For each group, the numbers of participants who were randomly assigned, received intended treatment, and were analysed for the primary outcome

[13b Losses and exclusions](http://www.consort-statement.org/checklists/view/32-consort/100-losses-and-exclusions) – For each group, losses and exclusions after randomisation, together with reasons

[14a Recruitment](http://www.consort-statement.org/checklists/view/32-consort/102-recruitment) – Dates defining the periods of recruitment and follow-up

[14b Reason for stopped trial](http://www.consort-statement.org/checklists/view/32-consort/103-reason-for-stopped-trial) – Why the trial ended or was stopped

[15 Baseline data](http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data) – A table showing baseline demographic and clinical characteristics for each group

[16 Numbers analysed](http://www.consort-statement.org/checklists/view/32-consort/107-numbers-analysed) – For each group, number of participants (denominator) included in each analysis and whether the analysis was by original assigned groups

[17a Outcome and estimation](http://www.consort-statement.org/checklists/view/32-consort/111-outcomes-and-estimation) – For each primary and secondary outcome, results in each group, and the estimated effect size and its precision (such as 95% confidence interval)

[17b Binary outcomes](http://www.consort-statement.org/checklists/view/32-consort/112-binary-outcomes) – For binary outcomes, presentation of both absolute and relative effect sizes is recommended

[18 Ancillary analyses](http://www.consort-statement.org/checklists/view/32-consort/114-ancillary-analyses) – Results of any other analyses performed, including subgroup analyses and adjusted analyses, distinguishing pre-specified from exploratory

[19 Harms](http://www.consort-statement.org/checklists/view/32-consort/116-harms) – All important harms or unintended effects in each group

**Discussion**

[20 Limitations](http://www.consort-statement.org/checklists/view/32-consort/118-limitations) – Trial limitations, addressing sources of potential bias, imprecision, and, if relevant, multiplicity of analyses

[21 Generalisability](http://www.consort-statement.org/checklists/view/32-consort/120-generalisability) – Generalisability (external validity, applicability) of the trial findings

[22 Interpretation](http://www.consort-statement.org/checklists/view/32-consort/122-interpretation) – Interpretation consistent with results, balancing benefits and harms, and considering other relevant evidence

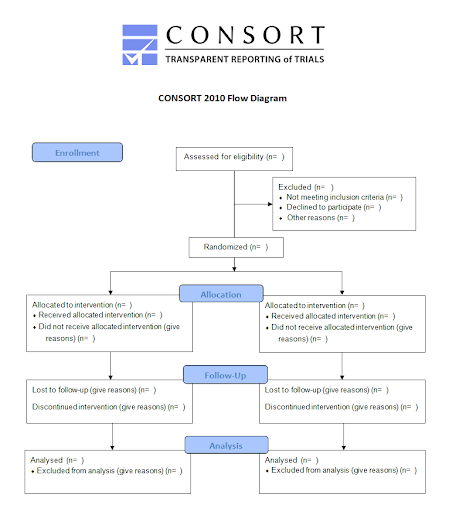
**Other Information**

[23 Registration](http://www.consort-statement.org/checklists/view/32-consort/124-registration) – Registration number and name of trial registry

[24 Protocol](http://www.consort-statement.org/checklists/view/32-consort/126-protocol) – Where the full trial protocol can be accessed, if available

[25 Funding](http://www.consort-statement.org/checklists/view/32-consort/128-funding) – Sources of funding and other support (such as supply of drugs), role of funders

You will also be familiar with the CONSORT flow diagram.



I think this is getting a bit complex. Let’s go back to the very beginning

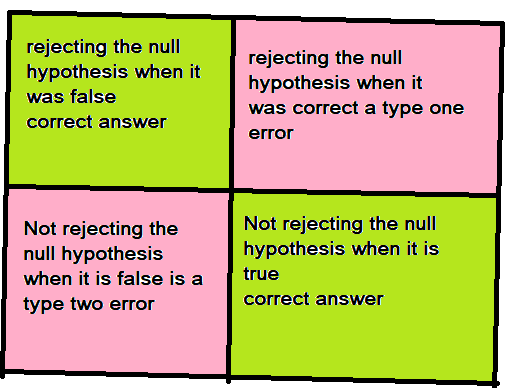
The Greek Interpreter

Regrettably some research is done to further the career of doctors and dentists but the proper reason for doing it is because you have a **question you want to know the answer to.**

I would like to know if I can use TADs to reduce the face height in long face patients.

So the first thing I have to do is express that as a **Null Hypothesis (Ho)** In this case the Null hypothesis will be that the treatment with fixed appliances and TADs placed in the palate does not reduce the face height as measured by the MM angle. After my trial I will find out if I can **reject the null hypothesis** with a probability of p = 0.05 or less and accept the alternative which is called HA. (Notice in statistics I can never be absolutely certain). In this case that use of TADs does allow me to reduce the MM angle.

I have decided that I will do so if p=0.05. This means I am 95% certain I am right. Or to put it another way there is a 5% chance of me making a **type one error** also called **alpha α.** A Type one error in this case would be rejecting the null hypothesis that TADs did not reduce the MM angle when in fact the null hypothesis was true, The probability is p=0.05



A type one error is when we say there is an effect and there isn’t one

A type two error is where we miss an effect which does exist

There is also a **type two error** also known as **Beta β.** In our study this is the chance that we would accept the null hypothesis even if there was a benefit from TADs reducing MM angle. So that 1 minus β is the probability that you have correctly rejected the null hypothesis and is therefore the **Power** of the study. (Power = 1-β)

In a lot of studies the power is set at 0.8. or 80% (Sometimes it is set at 90%)

Power increases as sample size increases. A larger power lets you see smaller differences, but there is a danger that the difference will be so small it is of no clinical significance. A larger sample size increases the complexity and the cost of the study. So we need a Power calculation to see how many subjects we need in our study.

To do a power calculation we need:

1. Delta Д size of the difference you want to detect.

2. Power β (which is usually set at 80 or 90%)

3. Significance level α which is set at 5% or 1%

4. Standard deviation sigma Ϭ

So α and β are simple, for delta we can think what we believe to be important. For example in our study to see if TADs can reduce the mm angle I would suggest that a reduction of 4° or more would be a significant reduction. How are we going to get the standard deviation when we haven’t done the trial yet? Well we could look at previous studies or in this case just use SD of MM angle in cephalometric studies which is 5°.

Once it would have seemed important to know the actual formula (Which you could Google) but now you would use a statistic package and just enter the four figures.

Now we have sorted out our sample size we need to know which test to use. Unfortunately there are a bewildering number of them. Some software packages guide you but the best answer is to get professional statistical advice.

As a guide

For studies where there is a normal distribution of data

**For categorical data:**

* One Sample proportion test
* Two sample proportion test
* Chi square test

**For numerical data**

* t test
* Paired t test (paired data in orthodontics is typically the same group of patients after some treatment or time interval although it could be a twin study)
* Analysis of variance (ANOVA).

**For non-parametric data**

These tests are less powerful than parametric tests. What you do is rank the data and apply the statistical test to the rankings. It is like giving them a position in class. It is the rank you use for the analysis rather than the actual value and you use the median rather than the mean.

* Wilcoxon’s sign rank test for paired data.
* Wilcoxon’s two sample rank test
* Mann-Whitney U test

Footnote Tea or beer



**William Sealy Gosset** (13 June 1876 – 16 October 1937) was an English statistician, chemist and brewer who served as [Head Brewer](https://en.wikipedia.org/wiki/Head_brewer) of [Guinness](https://en.wikipedia.org/wiki/Guinness)  and was a pioneer of modern statistics.

Gosset published under the [pen name](https://en.wikipedia.org/wiki/Pen_name) **Student** and developed most famously [Student's t-test distribution](https://en.wikipedia.org/wiki/Student%27s_t-distribution).

In his job as Head Experimental Brewer at [Guinness](https://en.wikipedia.org/wiki/Guinness), the self-trained Gosset developed new statistical methods– now central to the design of experiments, to proper use of significance testing on repeated trials, and to analysis of [economic significance](https://en.wikipedia.org/w/index.php?title=Economic_significance&action=edit&redlink=1) and more, such as his small-sample, stratified, and repeated balanced experiments on [barley](https://en.wikipedia.org/wiki/Barley) for proving the best [yielding](https://en.wikipedia.org/wiki/Crop_yield) varieties. Gosset acquired that knowledge by study, by trial and error, by cooperating with others, and by spending two terms in 1906–1907 in the Biometrics laboratory of [Karl Pearson](https://en.wikipedia.org/wiki/Karl_Pearson). Gosset and Pearson had a good relationship. Pearson helped Gosset with the mathematics of his papers, including the 1908 papers, but had little appreciation of their importance. The papers addressed the brewer's concern with small samples. Gosset's first publication came in 1907, "On the Error of Counting with a Haemacytometer," in which – unbeknownst to Gosset aka "Student" – he rediscovered the [Poisson distribution](https://en.wikipedia.org/wiki/Poisson_distribution). Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery so the Guinness Board of Directors allowed its scientists to publish research on condition that they do not mention "1) beer, 2) Guinness, or 3) their own surname"., Gosset seems to have taken his pen name "Student" from his 1906–1907 notebook on counting yeast cells with a haemacytometer, "The Student's Science Notebook" Thus his most noteworthy achievement is now called Student's, rather than Gosset's, [t-distribution](https://en.wikipedia.org/wiki/Student%27s_t-distribution) and test of [statistical significance](https://en.wikipedia.org/wiki/Statistical_significance).



Fig Mr Student unmasked. British statistician [William Sealy Gosset](https://en.wikipedia.org/wiki/William_Sealy_Gosset), known as "Student",

The final problem

So you have done your research and the p value is just less than p = 0.05.

Hurray that is “statistically significant”. But wait a minute what does that mean?



If there is a 5% chance (p = 0.05) that the soup is poisonous would you consume it? What if it was if the chance was 1% (p = 0.01)? You see statistics can never prove the soup is not poisonous, it can only give you the probability. However we can be more sure. We can look at other research of the same subject. This is called a **Systematic review.**

Often, systematic reviews include a **meta-analysis** component which involves using statistical techniques to synthesize the data from several studies into a single quantitative estimate or summary effect size. In contrast to traditional hypothesis testing which can give us information about statistical significance (i.e., did the intervention group differ from the control group) but not necessarily clinical significance (i.e., was this difference clinically meaningful or large).

**Effect sizes** measure the strength of the relationship between two variables, thereby providing information about the magnitude of the intervention effect (i.e., small, medium, or large). The type of effect size calculated generally depends on the type of outcome and intervention being examined as well as the data available from the published trials; however, some common examples include odds ratios (OR), weighted/standardized mean differences (WMD, SMD), and relative risk or risk ratios (RR).

Although systematic reviews are published in academic forums, there are also organizations and databases specifically developed to promote and disseminate them. For example, the **Cochrane Collaboration** ([www.cochrane.org](http://www.cochrane.org/)) is a widely recognized and respected international and not-for-profit organization that promotes, supports, and disseminates systematic reviews and meta-analyses on the efficacy of interventions in the health care field.

[Go to:](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3024725/)

8 Stages of a Systematic Review and Meta-Analysis

1. Formulate the review question

The first stage involves defining the review question, forming hypotheses, and developing a review title. It is often best to keep titles as short and descriptive as possible, by using the following formula: *Intervention* for *population* with *condition* (e.g., Dialectical behaviour therapy for adolescent females with borderline personality disorder). Reviews published with the Cochrane Collaboration do not need to be identified as such, but reviews published in other sources should also indicate in the title that they represent a systematic review and/or meta-analysis. If authors chose to conduct their review through the Cochrane Collaboration, they will also be required to register their title to the appropriate review group, which in essence “saves their spot” for this topic and provides access to further Cochrane support (e.g., assistance running search strategies).

2. Define inclusion and exclusion criteria

The Cochrane acronym PICO (or PICOC) which stands for population, intervention, comparison, outcomes (and context) can be useful to ensure that one decides on all key components prior to starting the review. For example, authors need to decide a priori on their population age range, conditions, outcomes, and type(s) of interventions and control groups. It is also critical to operationally define what types of studies to include and exclude (e.g., randomized controlled trials-RCTs only, RCTs and quasi-experimental designs, qualitative research), the minimum number of participants in each group, published versus unpublished studies, and language restrictions. For Cochrane Reviews, this information gets prepared, peer-reviewed, and published in a Protocol format first, which is then replaced with the full Review once it is completed.

3. Develop search strategy and locate studies

4. Select studies

Once a comprehensive list of abstracts has been retrieved and reviewed, any studies appearing to meet inclusion criteria would then be obtained and reviewed in full.

5. Extract data

6. Assess study quality

There has been a movement in recent years to better assess the quality of each RCT included in systematic reviews.

7. Analyse and interpret results

There are various statistical programs available to calculate effects sizes for meta-analyses, such as the Review Manager (RevMan) program endorsed by the Cochrane Collaboration. Effect sizes are stated along with a 95 % confidence interval (CI) range, and presented in both quantitative format and graphical representation (e.g., forest plots). Forest plots visually depict each trial as a horizontal diamond shape with the middle representing the effect size (e.g., SMD) and the end points representing both ends of the CI. These diamonds are presented on a graph with a centre line representing the zero mark. Often the left side of the graph (< zero) represents the side favouring treatment, while the right side (> zero) represents the side favouring the control condition.

8. Disseminate findings

Although reviews conducted through the Cochrane Collaboration get published in the online **Cochrane Database of Systematic Reviews**, they are often quite lengthy and detailed. Thus, it is also possible and encouraged to publish abbreviated versions of the review in other relevant academic journals, as long as they are clearly indicated as such.

Plain language summaries for families and patients are also commonly provided, and there is an expectation that reviews should be regularly updated to ensure they are always up-to-date and relevant. Indeed, participating in a review update or joining a well-established review team, can be a helpful way of getting involved in the systematic review process.

At last we have got to the questions.

Cover over the right side and see if you can answer these questions:

|  |  |
| --- | --- |
| 1. what is qualitative data | Non Numerical data like colour or sex |
| 2. What are the two types of quantitative data | Numerical data can be: A. discrete like hat sizes or the value of coins in your pocket or B. Continuous like the amount of liquid in a pond |
| 3. What would a correlation coefficient of R = 0.3 indicate | This shows a lot of scatter we can say that there is only a slight positive correlation between the data on the x and y axes |
| 4. What are the other names for a Gaussian curve | Normal distribution, Bell shaped cure |
| 5. What percentage of the population in a study lie between 1. 2. & 3 standard deviations | 68.3% lie within one standard deviation from the norm. 95.5% lie within two standard deviations and 99.7% lie within 3 standard deviations. |
| 6. In a box plot whisker diagram what is indicated by the ends of the whiskers and the ends of the box | The ends of the whiskers are the range and the ends of the box are the interquartile range so half the results would be covered by the box |
| 7. What is wrong with just forgetting about members of a trial whose treatment is not going well | It will bias the results making the treatment look better than it really is |
| 8. what are the advantages of a multi-centred trial | Increase the number of participants and lessens the effects of regional differences in the population (like red hair) |
| 9. Why publish research if it could not find a statistical benefit of the treatment | If negative papers are not published a future meta-analysis would falsely favour the treatment |
| 10 How could you use the CONSORT statement  (the Consolidated standard of reporting Trials) | You can use it to help you plan your study |
| 11 What is a type one error α | It is the chance of you rejecting the Null Hypothesis when you were wrong to do so. In effect you are saying you treatment works when it doesn’t. |
| 12 What factors increase the power of a study | Sample size and accuracy of measurement |
| 13 What four things are needed to do a power calculation | The power (usually set at 80% or 90%  Alpha (usually set at 5% or 1%)  Standard deviation. (Ideally you hope to get that from a previous related study)  Delta the size of a difference that you hope to detect and would consider meaningful |
| 14. what is a systematic review | An attempt to increase the power by amalgamating a large number of studies |
| 15. what is a meta-analysis | When studies give clear data it is possible to use statistical techniques to synthesize the data from several studies into a single quantitative estimate |