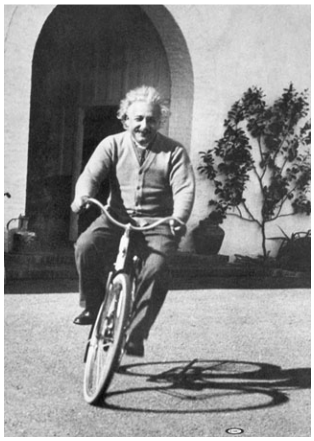
UNDERSTANDING STATISTICS



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

Dr Alan McLintic
Department of Anaesthesia
Middlemore Hospital
Auckland

"MBO QBTTFE BXBZ TVEEFOMZ CVU
QFBDFGVMMZ BU IJT IPNF JO .JTTJF
#BZ "VDLMBOE PO 4FQUFNCFS
5IJT JT B MBTUJOH HJGU UP USBJOF

INTRODUCTION

When the first plane flew into the twin towers, the initial response was that a terrible accident had occurred. When the second plane flew in, people immediately realised that this was not an accident but a deliberate act. No statistical analysis was required to reach this conclusion; it was an intuitive process which we experience day in day out. Goldfinger knew this too:

"*They have a saying in Chicago, Once is happenstance. Twice is coincidence. Three times is enemy action*".

What does formal statistical analysis tell us about an event in a trial? Usually it comes down to one thing: it formalises the above thought process and produces a Probability that the event could have happened by chance. It will *never* allow us to absolutely *prove* that an effect was more than a chance event. We should always leave the clichéd hyperbole of 'clinically proven' to alternative medicine and advertising agencies.

Statistics is never having to say you're certain

Statisticians have designed tests that at first appear abstract, but in fact are based on common sense and intuition. For example, if we were asked if there is any difference in skin colour between African Americans, Scotsmen and Chinese we subconsciously perform ANOVA and would say 'yes'. This is because the overall differences in skin colour (the between groups variance) far outweighs the confusion caused by the pallid Michael Jackson and the bronzed Rod Stewart and Bruce Lee (the within groups variance).
What if you see a strange person outside your house twice in one week? The question arises, is this a co-incidence or are they …a stalker! You will be subconsciously comparing the current proportion of two weirdo sightings per seven days against the normal proportion of zero sightings per seven days. How frequent does the new event need to be before you believe it to be more than coincidence? This is the thought process behind the Chi-square test.
Most other tests also have an intuitive basis – look for it and you'll understand statistical methods better.

AJM May 2008

CONTENTS

TYPES OF DATA

Two main groups of data:

Categorical (qualitative)

Each individual patient can only belong to one of a number of distinct categories.

Binary — Two categories: male /female; alive/dead

Nominal data: — The categories have names and there is no order. Eg blue, brown, green and grey eyes; blood types: O,A,B,AB.

Ordinal data: — There is an order to the categories eg cancer staging; severity of pain, ASA scores, APGAR scores,

Discrete numerical data: Where the variable can only take certain whole number values. eg pain score from 0-10. This data may be analysed as 'continuous' if the sample is large enough. Note: Age is a continuous numerical variable but is often treated as *discrete* and may also be counted into age range groups and handled as *categorical.*

Continuous numerical (quantitative)

The variable has a numerical value

Parametric and non parametric data:
Parametric data is continuous numerical data from a normal distributed 'population'.
Non-parametric data is all the rest but usually refers to continuous numerical data from a severely non-normal distribution or when the sample is too small to be sure what the parent distribution was like. (Large numbers of Ordinal data are often treated as continuous numerical data and are usually handled with non-parametric tests).

Interval data and ratio data: A largely irrelevant sub-classification of statistical data. Interval data increase at constant intervals but do not start at a true zero. eg gauge pressure or temperature on the Celsius scale (20 °C is not twice as hot as 10 °C). Ratio data is a type of interval data in which there is a true zero; eg absolute temperature or absolute pressure (Absolute pressure of 200 kPa is twice as great as 100 kPa)

DESCRIPTIVE STATISTICS

What is *descriptive statistics*?

Describing the data from a sample. This usually consists of a summary measure accompanied by a measure of the spread of data in the sample. A summary measure with a narrow spread of data is indicative of a real trend in the sample. If there is a wide spread of data one can't be sure that the summary measure represents a real trend. Graphical displays are also part of descriptive statistics.

Summarizing data

Arithmetic mean

Often referred to as simply the *mean*. The sum of observations divided by the number of observations. Has most relevance if the data are symmetrical.

Median

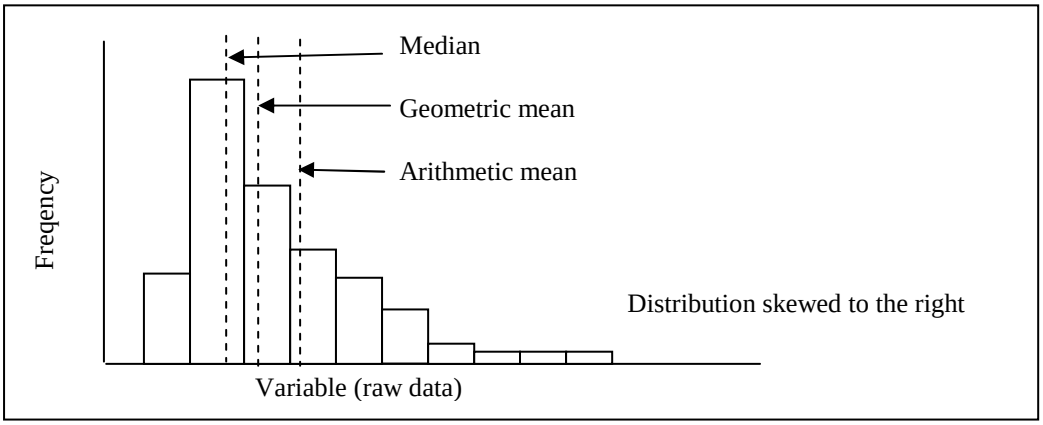
Middle of a series of observations. If there is an even number of observations, the median is the average of the two middle observations. The median is the same as the mean if the data are symmetrical but will be different if the data are skewed. The median is the summary of choice in non-parametric data

Mode

The value that occurs most frequently. The same as the median and mean if the data are symmetrical.

Geometric mean

If data are skewed to the right, plotting the log of x against the frequency produces a much more symmetrical distribution. The arithmetic mean of the log values can then be calculated . The anti-log of this mean is termed the geometric mean and will be closer to the median than the mean of the raw data.



Measures of variability

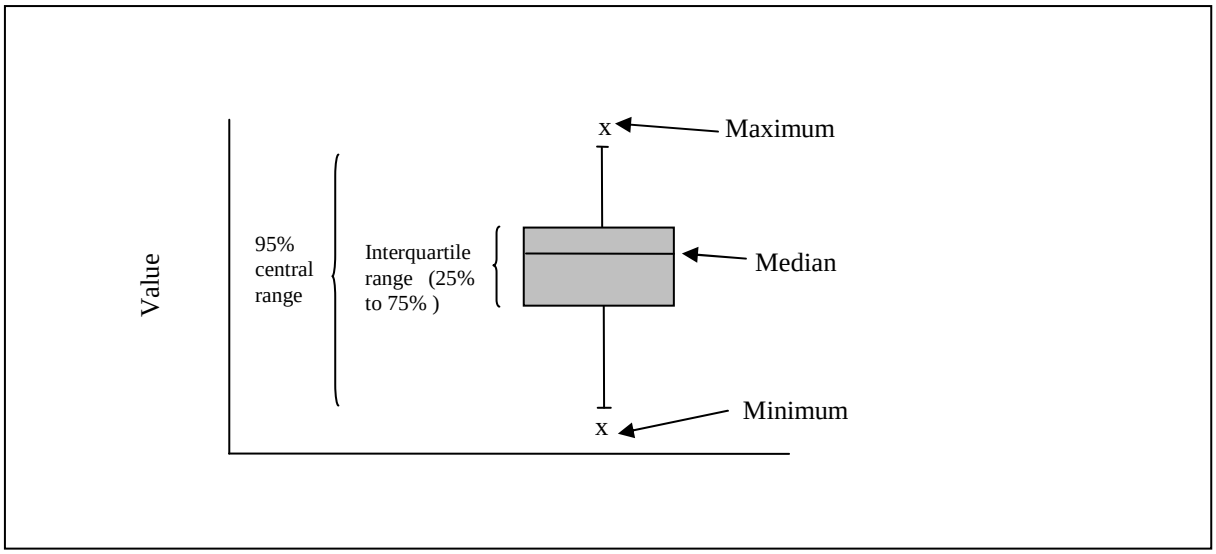
Measures of variability describe the average dispersion of data around a mean. The most commonly used measures of variability are the range, percentiles, standard deviation and the standard error of the mean.

Range

The smallest and largest values in a sample. May be used in reporting skewed and other non-parametric data along with the median Problem is that the range is too influenced by outliers so the inter-quartile range(see below) is preferred.

Percentiles

Definitions vary. The simplest is that they are a method of ranking by dividing the scores or results into 100 parts. Thus, if your score was on the 65th percentile, 65% of scores lie below you. The interquartile range is the range around the median within which 50% of the scores lie. It has the advantage of not being influenced by outliers. Another commonly used range which is sometimes referred to as the 'normal' range, is the 95% central range. This excludes the outer 2.5% of observations.



Box and whisker plot. Plots the median, 25th and 75th percentiles and the maximum and minimum values. The 95% central range may also appear in the B and W plots.

Standard deviation (SD, s or σ)

A measure of the average spread of *individual values* around the sample or population mean. The symbol *s* is use for the SD of the sample data and σ is used for population data.

Calculation of SD

To calculate the SD from a sample you square the differences between each value and the sample mean, sum them (*sum of squares*) and divide this by (n-1) to give the *variance*. The SD is the square root of the *variance*.

$$SD = \sqrt{\left(\frac{\sum (x_i - \bar{x})^2}{n-1} \right)}$$

Degrees of freedom

The term (n-1) is called the *degrees of freedom* (d.f). It is the number of totally independent observations that are possible in a sample where the mean is known. It is one less than the sample number because, if n-1 observations are totalled, the last one can be deduced.
An intuitive explanation for degrees of freedom is hard to find in the texts. It is, however, enough to know that using (n-1) gives a better estimate of the population variance from sample data.

When is the standard deviation used ?

i) When reporting sample data, the SD gives an indication at a glance as to whether the sample mean represents a real trend in the sample.

ii) The SD of a large, randomly selected sample can be assumed to be close to that of the population from which it was drawn.

iii) The SD is used to calculate the SE

iv) Any individual data point in a normal distribution can be described as a multiple of SD's from the population mean. This is called z transformation. This has less importance than z transformation of means.

v) Standardised difference. This is an effect size expressed in multiples of the standard deviation. It is used in sample size calculations. See later.

Standard error of the mean (SE)

An estimation of the spread of *sample means* around the population mean. If you were to take multiple samples from a normally distributed population and plot the sample means you would end up with a normally distributed plot of sample means. (See Figure page 9) The SE is an estimate of the spread of sample means in this *theoretical* distribution. But, you do not need to take multiple samples and plot their means to estimate the SE. It is 'guess-timated' from the data in a single sample.

Calculation of the SE

The SE is an estimate based on the number in the sample and the sample SD.

$$SE = \frac{SD}{\sqrt{n}}$$

Intuitively, this makes sense because the variability among sample means will be increased if there is a) a wide variability of individual data and b) small samples.

Why do we need the SE?

i) An indication of the *precision* of the sample mean as an estimate of the population mean (See Page 10 for further explanation)

ii) The SE is used in parametric tests to quantify the magnitude of an effect size. The effect size is expressed as multiples of the SE. This is called z-transformation. (See below)

iii) The SE is used to calculate confidence intervals (See below)

DISTRIBUTIONS

An empirical frequency distribution is one in which the observed data are plotted against their frequency. Theoretical distributions are those which are described by a mathematical model and are used to analyse data.

The Normal Distribution

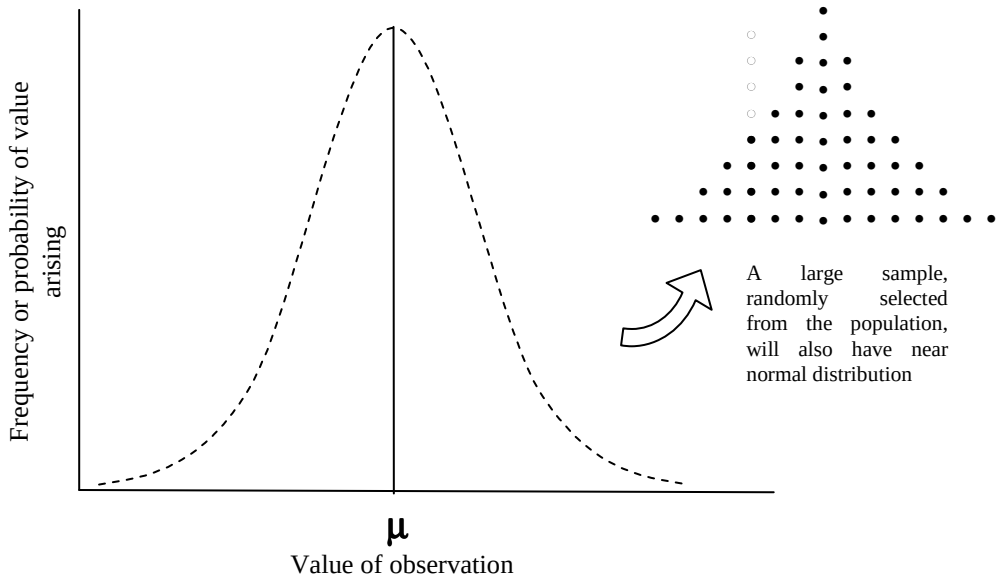


Figure 1. The normal distribution curve

Features of a normal distribution

1. An observation that is normally distributed within a population has a *norm* and random *independent* factors have caused variation on that norm. Most values cluster around the norm with fewer and fewer values towards the tails. Extreme values *do* exist though.

2. It can be completely described by its mean and SD.

3. Because the variation is random, there is equal spread of values above and below the norm. The average value (mean) is the same as the central value (median) and the most common value (mode).

4. A normal distribution (Gaussian) curve can be plotted to illustrate the *frequency* of observations within the population or the *probability* of an observation arising in the population. The curve is bell shaped, symmetrical and *theoretically* of infinite size with tails that never reach the x axis.

5. A large (n > 100) randomly selected sample from a normally distributed population would also have near normal distribution.

6. The mean and standard deviation of such a sample is likely to be close to the mean and standard deviation of the population from which it was sampled.

7. The smaller the sample the less likely it will have 'normal' geometry and the less likely that the mean and standard deviation will match those of the population.

8. If multiple large samples were to be randomly selected from a normally distributed population the plot of the sample means would also have normal distribution.

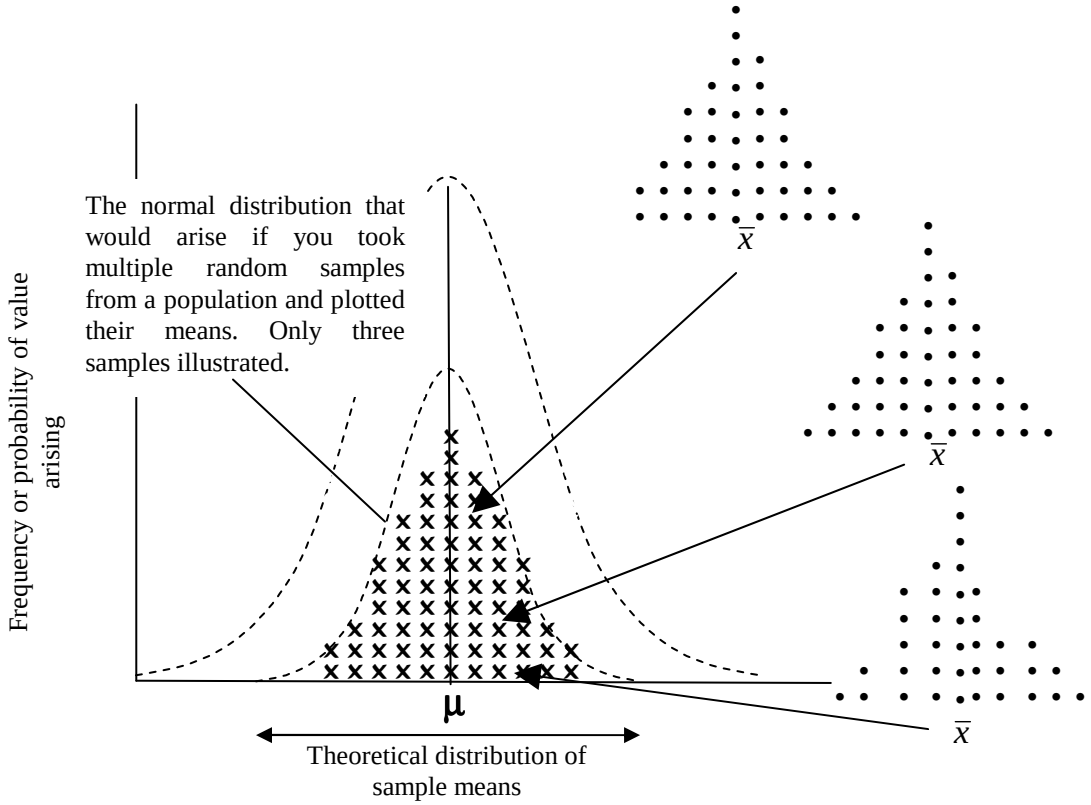


Figure 2. The theoretical distribution of sample means

Precision and the standard error

Suppose you repeated the above exercise with huge samples of, say, 1000 patients each. Clearly, in this situation, the sample means would be close estimations of the population mean and so a plot of the means of multiple samples would have a very tight distribution. The SE of this distribution would thus be small.
Now suppose you repeated the exercise with small samples of 10 patients each. In this situation the sample means may not be close estimations of the population mean and the plot of multiple sample means would be much more spread-out. The SE of this distribution would be large. Thus, the SE is an indication of the *precision* of a sample mean as an estimation of the population mean.

Is my sample from a normal distribution ?

- Plot the data and superimpose a normal distribution with the same mean and SD. 'Eye-ball' the fit.
- Normal Plot. This is a plot of the ordered sample values against what you would expect from a Normal distribution of the same size. Eye-ball the fit – it should be a straight line.

- Goodness of fit calculation. Computer algorithm will give you the likelihood that your data is normally distributed. Not thought to be superior to the subjective methods above.

The Standard Normal Distribution

z transformation

z transformation is where the difference between an observation and its population mean, or a sample mean and its population mean is converted to a multiple of, respectively, SD's or SE's. The resulting multiples are called z points, z values or standard normal deviates.

$$z = \frac{x_i - \mu}{SD} \quad or \quad z = \frac{\bar{x} - \mu}{SE}$$

Standard normal distribution / z distribution

If the data of *any* normal distribution were to be converted to z values, a standard curve with fixed, known proportions arises. This is called the standard normal distribution curve (SNDC). By simple arithmetic, you will see that the SNDC always has a mean of zero and a SD of one.

The SNDC is a theoretical distribution of infinite size and the area under the curve (AUC) contains all possible chance variations of the population mean. The probability of any variable occurring within the total AUC is, therefore, 1. In fact, because the probability densities of all the proportions of the SND are known, it can be used to determine the probability of *any* z value occurring through a chance variation of the population mean . (Fig 5 and 6).

Example: if your analysis results in a z value of 1.96 and look this up in z tables, you will find that 0.025 will appear against this z value. This means that the proportion of the SNDC with z values greater than 1.96 is 2.5%. This is the fundamental principle behind parametric testing.

Percentage points

z values are also referred to as percentage points of the SND. The most important are the 5 % percentage points ($z_{0.05}$) which are ± 1.96. These values exactly encompass 95% of the SND.

SND and sample means

The plot of the means of an infinite number of samples is also normally distributed and has SND geometry when the means are expressed as *z* values. Here, however, z values are multiples of *standard errors* from the population mean. Thus the SND can be used to describe the probability of any sample mean arising as a random variation of the population mean. It is much more common for us to be using this sort of z value, as medical research is usually interested in comparing sample means rather than individual values.

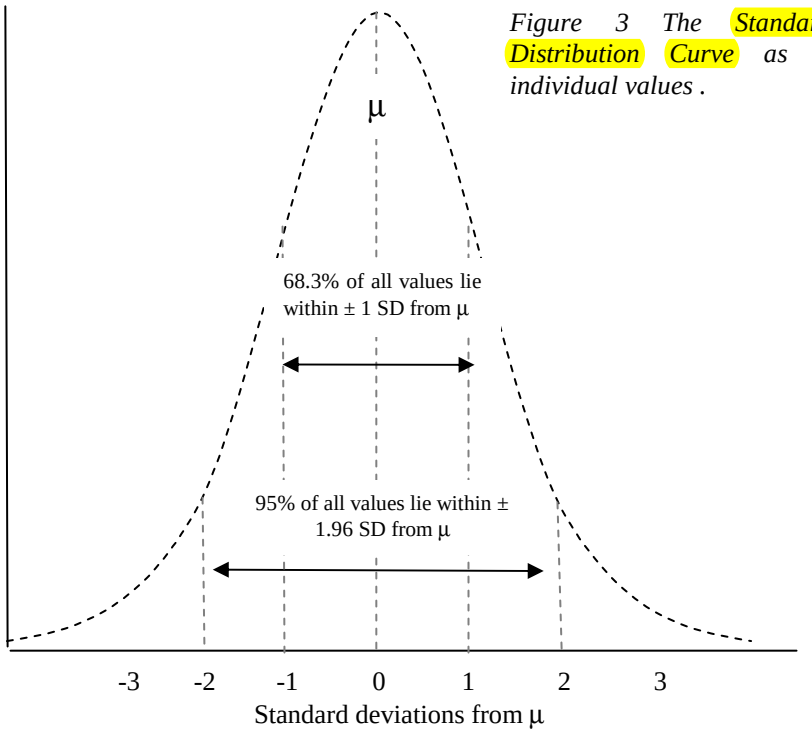


Figure 3 The Standard Normal Distribution Curve as applied to individual values .

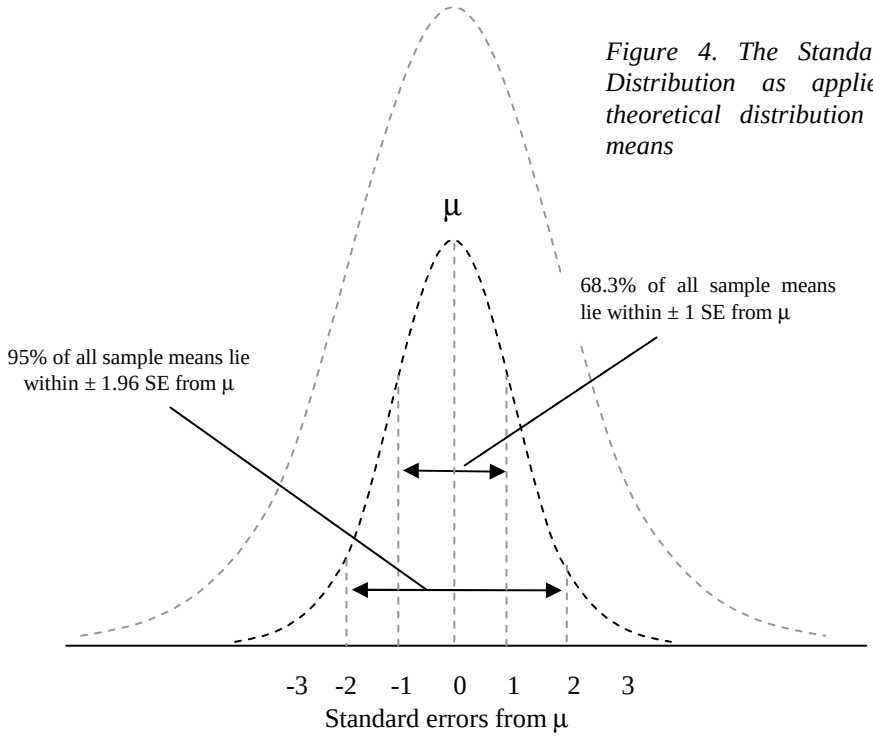


Figure 4. The Standard Normal Distribution as applied to the theoretical distribution of sample means

The t-distribution

Derived by W.S Gossett under the pseudonym *Student.* Probability density distribution for parametric data but for samples that are too small to use in the z distribution. For the latter you require large samples with near *normal* geometry. The t-distribution is consulted according to the degrees of freedom (n-1) of your sample. As the sample size gets bigger you will find that the parameters of the t-distribution become closer to those of the SND. (See later under t-test)

The Chi-squared distribution

The chi-squared distribution is derived from the normal distribution (and therefore continuous numerical data) and describes the distribution of the variance of samples taken from the ND. The shape of the distribution depends on the degrees of freedom. As the degrees of freedom become greater the distribution becomes more Normalised.

Skewed distributions and data transformations:

If a variable has a distribution that is skewed to the right (positively skewed), the data may be transformed so that the distribution becomes more 'Normal', thus allowing parametric tests to be used. The transformation of choice here is to plot frequency against the log of x. Parametric tests are made on the transformed data and conclusions made. (Eg the mean log of X is significantly different from the mean log Y). Summary measures are often back-transformed (antilog), for example to produce the geometric mean.

Binomial distribution

Describes the probability of different proportions of a binary outcome arising in a fixed number of observations.

Example

A binomial distribution might be used to display the probabilities of different proportions of heads arising during sets of coin tosses or the chances of turning up a disease of known incidence in a sample of specified size.

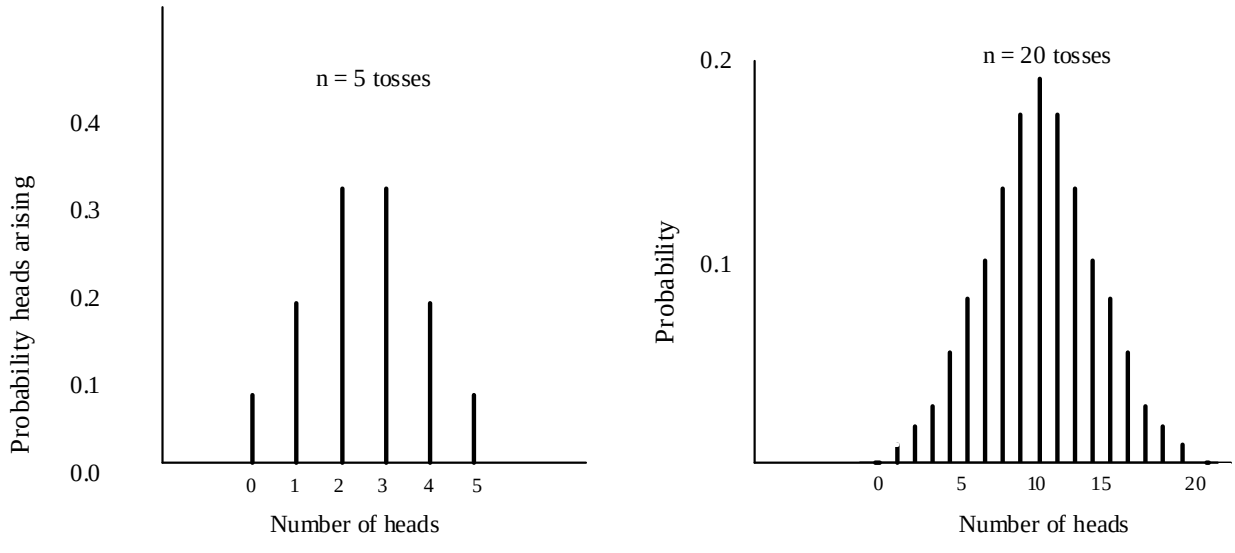


Figure5. Two binomial distributions. The first displays the probabilities of all possible proportions of heads arising in five tosses of a coin. The second is the probability of all possible proportions of heads if the coin is tossed 20 times

Population proportion

The most likely proportion (the norm) in the population is called the population proportion (π). In the above example it would be 0.5 (50 % heads). In another example one third of people have blue eyes, so $\pi = 0.33$.

Sample size

As the sample size gets bigger, it becomes more likely that the proportion of a particular observation within the sample will be the same or similar to that of the population proportion (π). Thus, if you were to toss a coin only four times you would have a good chance of turning up a proportion of heads far removed from 0.5. If you were to toss a coin 1,000 times it is likely that the resulting proportion of heads would be very close to 0.5.

Distribution shape

The larger the sample, the closer the binomial distribution is to a normal distribution. This is the case, even if π is not 0.5. See below. This fits in with the statements above in that, with large samples, finding results far from the norm is rare.

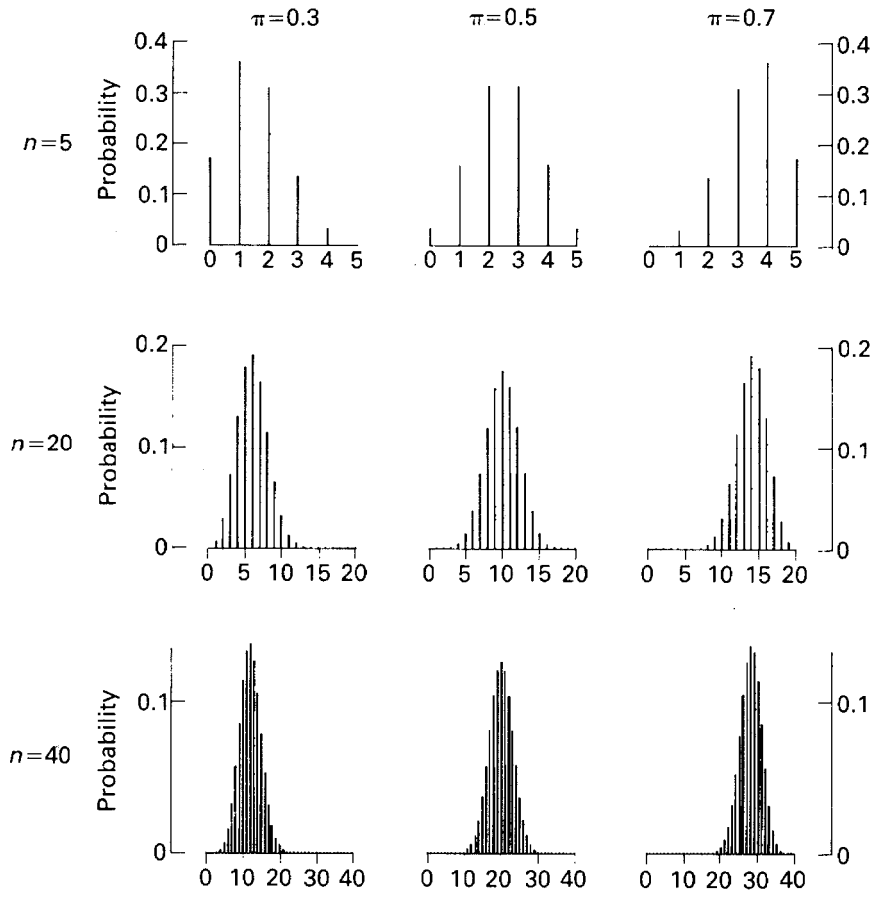


Figure 7. Examples of the way the binomial distribution changes with sample size. (Reproduced from Kirkwood BR, The essentials of medical statistics, Blackwood Scientific Publications 1988)

Total probability

Total of all outcomes must be 1.0

The binomial formula

The probability of a specific proportion arising in a sample is not eye-balled from the above graphs but calculated using the

	binomial formula. The input into this formula is the proportion you are seeking, the population proportion and the sample number. (See appendix)
Hypothesis tests	As the binomial distribution can be approximated to a normal distribution, hypothesis tests such as the *normal approximation test* can be carried out to determine, for instance, the probability of a particular proportion or greater arising in a binomial distribution.
Example of the application	If vomiting is known to occur after general anaesthesia in 0.3 cases (π = 0.3) and a study of TIVA in 100 patients results in vomiting in only 0.2 cases, is the difference in proportions significant or is 0.2 simply a random variation of 0.3?

Poisson distribution

	Describes probability of a number of events occurring in a fixed time period or in a region of space. The events occur randomly and independent of each other at some average rate (μ) The probability is calculated from an exponential formula and depends on prior knowledge of one parameter only, the mean number of occurrences per unit time period (or unit region of space). See appendix for formula.
Example	If the number of adverse incidents in theatre over a two year period is known, what is the probability of more than 5 incidents in one day?

HYPOTHESIS TESTING

What is an hypothesis test?	A process by which we test a specific hypothesis on a set of data. The result will be couched as a rejection or acceptance of the hypothesis. Hypothesis tests may be parametric or non-parametric.

Principles of hypothesis testing	The null and alternative hypotheses are defined Data is collected A test statistic is calculated to test the hypothesis The statistic is compared to values in a probability distribution A *P-value* is produced which is compared with a significance level The hypothesis is either accepted or rejected

Null hypothesis (H_0)	The standpoint that that an effect found experimentally is simply a chance event. For example, in a comparison of the effects of Drug A with Drug B, the Null Hypothesis would be that there is no real difference between the drugs and that any difference detected is simply due to chance. An hypothesis test is then carried out to determine the likelihood of A and B being simply random variations of each other. On the basis of this, the Null Hypothesis is either accepted or rejected.
Alternative hypothesis (H_1)	An alternative hypothesis that holds by default if the null hypothesis is not true. For example, if the Null hypothesis is that *Drug A does not alter blood pressure*, the Alternative Hypothesis will be that *Drug A does alter blood pressure*. Note that, by using the term *alter* rather than specifying *raise* or *lower*, the H_1 is a two-tailed Alternative Hypothesis. This is the most common situation because we cannot usually state beforehand that Drug A (if it had an effect) could only move the blood pressure in one specific direction.
P value:	The probability that an effect could have occurred by chance alone if the Null Hypothesis is true The *P* value is calculated from your study results and is the proportion of the SNDC which is more extreme than the z value. *P* is then compared with the pre-set *alpha* and the H_0 accepted or rejected. *NB*: It is not correct to say that the *P value is the probability that the H_0 is true.* The H_0 is either accepted or rejected.
Alpha value	The alpha value is the *significance level* and is the limit at which your *P* value will be deemed too large for a difference to be regarded as statistically significant. Alpha is set by the investigators at the study design stage. In medical research an alpha value of 0.05 is usually selected.
Comparing *P* with alpha	Supposing you are comparing two means, A and B. At the design stage you have set alpha at 0.05 and your calculated *P* value later turns out to be 0.045. The latter means that, if H_0 is true and differences between A and B are merely due to chance, the detected difference could occur 4.5 % of the time. Because you have set alpha at 0.05, you may claim the difference as statistically significant. and reject the null hypothesis. The downside is that you have a 4.5 % chance that your statement is a false positive one.

Type 1 Error (Alpha error):

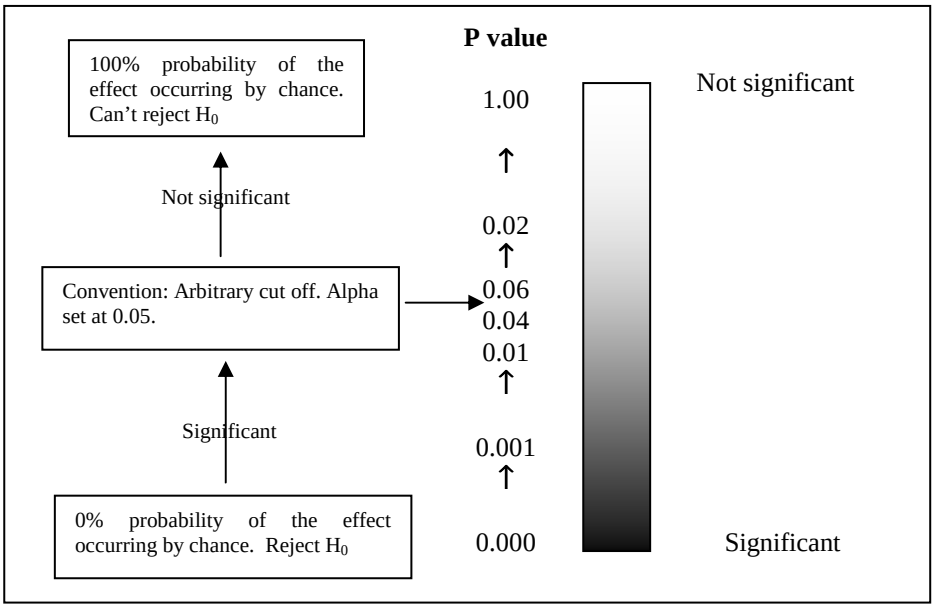
Frequency where we erroneously conclude there is a difference when there isn't one. False positive frequency. The magnitude of the potential alpha error is determined by the alpha value selected.

Type II Error (Beta error)

Frequency where we are unable to detect a difference when there is one. False negative frequency. One common cause of this is using a sample size which is too small. (see later)

Limitations of hypothesis testing

- The selection of 0.05 as the significance level is very common, but is totally arbitrary and does not usually have a clinical basis.
- A statistically significant difference does not necessarily imply a clinically significant difference.
- The greater the alpha value, the greater the likelihood of more false positive error (alpha or type 1 error)
- If a 'significant' result is presented as $P < 0.05$, rather than by giving the exact P value, the reader is prevented from drawing their own conclusions about the degree of significance. For example, you wouldn't know if P had been 0.0001 (highly significant) or 0.047 (barely significant).



Interpretation of P values: The continuum between significance and non-significance

PARAMETRIC TESTS

Basis of parametric tests:	Parametric tests are tests that are based on the parameters of the normal distribution. They determine the likelihood that a difference has occurred by chance variation rather than because of a real effect.
What assumptions are made?	The data is *continuous and numerical*. (Large numbers of discrete data may also be treated as parametric.) The samples have *similar variance* and are taken *randomly* from a *normally distributed population.*

Normal test (z test) : One sample/Unpaired

Definition	A parametric test for very large samples *(texts vary as to the minimum size but rarely used when n < 100)* or in the unusual situation where we know the population variance. The One sample z test determines the likelihood that the mean of a large sample ($\bar{x}$) is simply a random variation of a specified number (μ).
Basic principle	A standpoint is adopted that (μ) is the mean of a normal distribution and my sample is part of this distribution. Any numerical difference between $\bar{x}$ and μ has, thus, simply occurred by random variation. If this is true, the SD of the large sample will be the same as that of the proposed 'parent' distribution. The sample SD can thus be used to calculate the SE of this proposed parent distribution and to convert $\bar{x}$ to a z value on it. Z tables are then consulted to determine the probability of finding a value more extreme than $\bar{x}$ in this distribution.

<u>Step by step example</u>

Question	The average height of normal UK 4 yr old boys is 102 cm. (μ = 102 cm) Does the height of a sample of 100 immigrant 4 yr olds differ from this number?
Results	n = 100, sample mean ($\bar{x}$) = 99, SD of sample = 9.8
Null hypothesis	There is no real difference between the mean of 102 and the height of immigrant boys. In other words, 99 is just a random variation of 102.
Alternative hypothesis	The average height of UK and immigrant 4 yr olds is *different.* (This has to be a two tailed H_1 as we have no reason to suppose beforehand that the test results could *only* vary in one direction from 102 cm)
First assumption	Assume that 102 is the mean of a normal distribution of heights (Distribution A) and that our sample is part of that distribution.
Second assumption	(a) Assume that, as the sample is large, the SD of the sample is approximately the same as that of Distribution A. The SD turned out to be 9.8cm. (b) From the data in your sample, you can estimate the SE of the theoretical distribution that would occur if you took multiple sample means from Distribution A.

$$SE = \frac{9.8}{\sqrt{100}} = 0.98$$

Thus, the Null Hypothesis is proposing our sample mean (99) belongs to a theoretical distribution of sample means with mean 102 and SE 0.98.

Z transformation

The difference between 99 ($\bar{x}$) and 102 is then estimated in terms of SE's. (See figure 5)

$$z = \frac{99 - 102}{0.98} = -3.06$$

Consult z tables

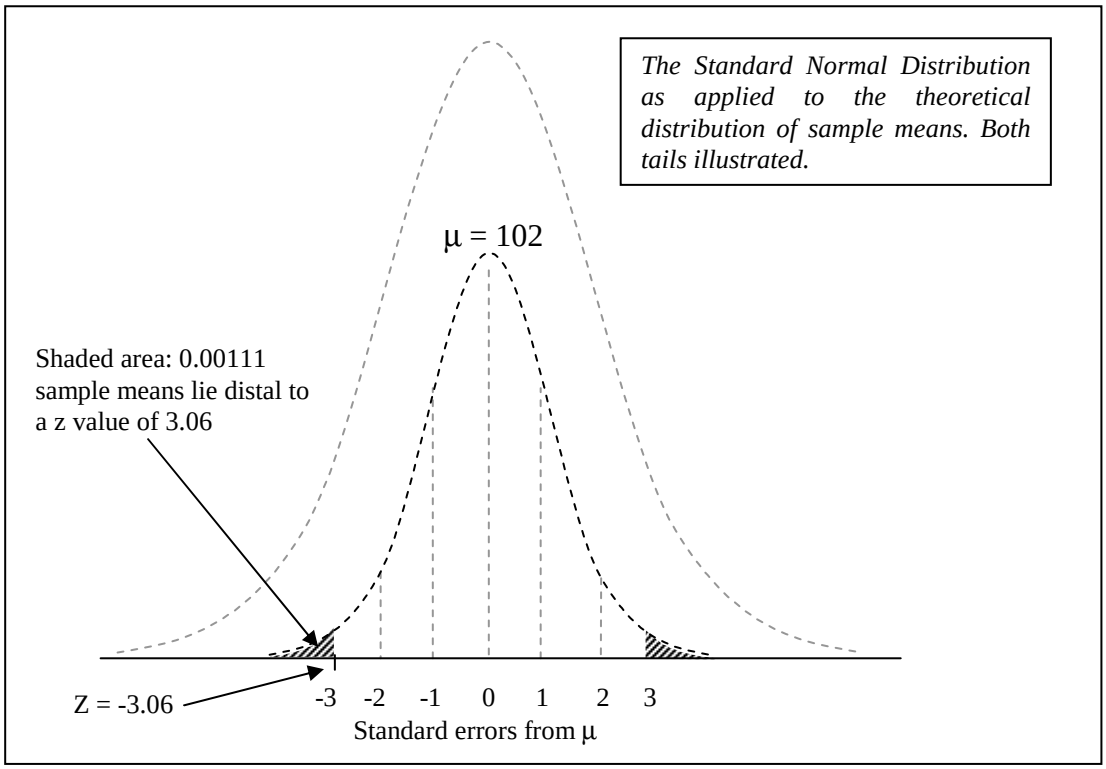
z tables are consulted to determine the proportion of the normal distribution that lies below a z value of -3.06. The answer is 0.00111.
As we didn't know at the start of the study which way, if any, the sample mean would vary from 102, we must double the 0.00111 to 0.00222. (See two tailed tests below)

Conclusion

If the H_o was true, there is a 0.2 % chance that the sample mean of 99 could be a random variation of 102. P = 0.002 In other words there is a strong likelihood that the immigrant boys have a different height from the UK norm of 102. There is a 0.2 % chance that this statement is false positive.
If the significance level was set at 0.05, the above result would be regarded as highly significant.

One tail vs two tailed tests

As already stated, an effect will be regarded as significant if the probability that that an effect of that *magnitude* could have occurred through chance alone (*P*) was less than 5%. Thus, if you have only been able to state a two-sided alternative hypothesis, this probability is made up of the proportion in SNDC distal to your test z value plus the proportion that is distal to the equivalent z value in the other tail. This is a *two–tailed test* and is the usual test, given that H_1 is rarely able to state beforehand that an effect could only occur in one particular direction.
If you know beforehand that A could *only* vary in *one* direction compared with B (eg A could only be greater than B), the *P* value is simply the proportion of the SNDC *greater* than the test z value. This approach is called a *one-tailed test*.



The Standard Normal Distribution as applied to the theoretical distribution of sample means. Both tails illustrated.

Normal test (z test): Two samples

Definition	A parametric test to compare the means of two very large samples or in the unusual situation where we know the population variance.
Basic principle	The null hypothesis that there is no difference between the sample means is tested by determining the likelihood that the difference between the means could be found in a normal distribution around a mean of zero.
<u>Step by step example</u>	Are the post-operative morphine requirements of technique A and B different?
Basic trial design	70 patients given technique A and 80 patients given technique B. Post operative morphine requirements noted for each
Null hypothesis	There is no difference in morphine requirements. Any numerical difference is due to random variation alone.
Alternative hypothesis	The morphine requirements of A and B are different
Results	Group A (n = 70); mean morphine requirement 10.6 mg, SD 1.4 Group B (n = 80); mean morphine requirement 11.5 mg, SD 1.3 Mean difference in Morphine requirement ($\Delta \bar{x}$) = 0.9 mg
First assumption	Imagine the situation in which the techniques were *exactly* the same. The *difference* in post-operative morphine requirement ($\Delta \bar{x}$) would most likely be 0 mg or close. If that trial was repeated over and over and $\Delta \bar{x}$ plotted each time, a normal distribution

would occur around a mean of zero. Call it Distribution Δ. The null hypothesis is that the mean difference between our two samples (0.9 mg) is part of this distribution

Second assumption

Because 0.9 mg is assumed to come from Distribution Δ, and the samples were large and randomly selected, the SD of our samples can be used to estimate the SE of Distribution Δ. The calculation of the SE for a two sample test is not as simple as before, requires a combining of the SD's, and is found in the appendix.

Z transformation

The difference between 10.6 and 11.5 is then estimated in terms of SE's by dividing the difference by the SE.

$$SE = 0.222 \qquad z = \frac{10.6 - 11.5}{0.222} = \frac{-0.9}{0.222} = -4.054$$

Consult z tables

z tables are consulted to determine the proportion of the normal distribution that lies below a z value of - 4.054. The most extreme z value listed is 3.29 which corresponds to a two-tailed *P* value of less than 0.001.

Conclusion

Statistically, this is a highly significant difference. If the H_o is true, there is less than 0.1% chance that it could occur by random variation alone. The H_o is rejected at the 0.05 level

Student's t-test

Definition

The *t*-test is a parametric test for the means of samples which are from a normally distributed population, but which are too small for the Normal test.

Basic principles

The calculation of the test statistic is very similar to those of the Normal Test although the statistic is called a *t*-statistic rather than a *z*-value. However, because the samples are small, we can no longer assume that the sample SD is the same as that of its population because, as noted before, the SD becomes larger as the sample size becomes smaller. We cannot, therefore, use the probability densities of the SND. Instead we take the t-statistic to t-distributions. These have been adjusted to take sample size into account, becoming flatter and flatter as sample size decreases. The *t* statistic must be used with the *t*-distribution appropriate to the sample's degrees of freedom. The *t*-distribution becomes nearly Normal when n > 60 and negligibly different when n > 100

One-sample t-test

Determines the likelihood of a sample mean being different from a specified number

Basic calculation of *t*-statistic

$$t = \frac{\bar{x} - \mu}{SE}$$

Two-sample or unpaired t-test

Determines the likelihood of the means of two independent samples being different.

Basic calculation of *t*-statistic

$$t = \frac{\bar{x}_1 - \bar{x}_2}{SE_{\text{difference between means}}} \quad \left(\text{see appendix for calculation of SE}\right)$$

Paired t-test

Determines the likelihood of two sample means being different where the samples are the same individuals in a before-and-after an intervention.

In a before-and-after intervention on the same subjects there is likely to be less intrinsic variability within the samples. This means that a small difference in means is more obvious and will have greater significance. (An electrical analogy is that there is a greater signal to noise ratio) The calculation of the paired *t*-statistic takes this into consideration and results in a more powerful test. Thus, a false negative result is less likely than if an unpaired test is used.

Another instance where a paired test is indicated is when there are two groups of different patients but they have been matched with each other

Basic calculation of *t*-statistic

$$t = \frac{\bar{x}_{\textit{differences before and after treatment}}}{SE_{\textit{differences before and after treatment}}}$$

CONFIDENCE INTERVALS

Definition (parametric tests)

The 95% confidence interval (CI) in a parametric test is the range around the sample mean within which you predict with 95% certainty, that the true value (the population mean) lies.

More general definition

CI = Estimate ± a multiple of SE where multiple depends on assumed distribution and level of confidence

How is it calculated?

In the SNDC, 95 % of sample means should lie between 1.96 SE above and 1.96 SE below the population mean. It follows, therefore, that there is a 95 % probability that the population mean lies within ± 1.96 SE of any large sample mean that has been randomly selected from that population. *Think about those two statements carefully and understand why, if the first is true, the second must also be true.* This is called the 95 % CI for the population mean.

$$CI = \bar{x} \pm (z_{0.05} \times SE)$$

The above refers to a SNDC which has a $z_{0.05}$ of 1.96 (*see percentage points)* If the t-test is being used, the 5 % percentage points vary, depending on the degrees of freedom. For example $t_{0.05}$ with 19 degrees of freedom is 2.09 and $t_{0.05}$ with 9 degrees of freedom is 2.26.

$$CI = \bar{x} \pm (t_{0.05} \times SE)$$

What are the usual causes of wide CI's?

Small samples
Large variance within samples

What information does the CI give you?

Descriptive

The CI gives an indication of the precision of the sample mean as an estimate of the population mean. The wider the confidence interval, the greater the imprecision and the greater the potential difference between the sample mean and its population mean.

Inferential

Generally speaking, hypothesis tests produce a 'reject' or 'accept' answer devoid of any indication of statistical significance. Examination of the *P* value itself is necessary to provide this information.

Confidence intervals allow more scope for reader judgement on significance. The alternative to the hypothesis test is to examine whether a population mean of interest falls within the 95% CI of your sample. If not, it is a 95 % probability that your sample is from a different population. In addition, a reader may look within a CI for a clinically significant value of their own choosing. Similarly, sample means with overlapping CI's cannot be regarded as different, and graphical presentation of several means with CI's allow instant visual comparisons to be made. See Forest plots and meta-analysis.

Specific applications

.

Odds ratios. OR's are frequently presented with confidence intervals. An OR of 1.0 suggests no risk associated with the exposure. If an OR is presented with a CI that has a range that includes the number 1.0, then that OR cannot be regarded as significant. For example an OR of 2.1 (0.6 – 3.6) could not be regarded as significantly different from 1.0.

Forest plots (see later) Graphical representation of the trials included in a metaanalysis. The results of each trial are plotted with their confidence intervals. Allows easy comparison of the significance of each trial.
The pooled OR in a meta-analysis is often presented as a diamond, the width of which is the confidence interval of the pooled OR. If the width encroaches on a line marked at 1.0 on the x-axis, the pooled OR cannot be regarded as significantly different for 1.0.

PARAMETRIC TESTS FOR MULTIPLE SAMPLES

Analysis of variance (ANOVA)

Definition	Determines whether there is a difference among three or more samples by comparing the variability between the groups (which should be large if there is a difference) with the variability within the groups (which should be as small as possible). ANOVA does not, however, tell you *which* of the samples is different.
Types of ANOVA	One way ANOVA: Comparing one observation in three or more groups Multiple ANOVA: Comparing more than one observation in three or more groups Repeated measures ANOVA: Comparing one variable in the same group at different times

Intuitive explanation of ANOVA

Post operative morphine requirements are compared after three different anaesthetic techniques (A, B and C). The morphine requirements of the three samples are plotted below with their respective means. It can be seen that, in each of the samples, the values cluster tightly around the mean. This implies that there is a real trend within each sample and, therefore, that the mean is likely to be a close representation of the true effect of each technique. Therefore, as the mean morphine requirement of Technique C is larger than the other two it is likely that Technique C has a real difference from the other two. In ANOVA terminology, the *between* groups variability is large but the *within* groups variability is small.
How certain would you be though, if the results looked like those in the second figure? The means are the same as before, with Technique C seemingly resulting in a larger morphine requirement. However, there is no trend within that sample. The results are spread out and we can no longer say with any confidence that the sample mean represents a 'true effect' (ie the effect that would occur if you used technique C on the whole population)
In ANOVA terminology, although the *between* groups variability is still large, the *within* groups variability is also large and a difference is not so certain.

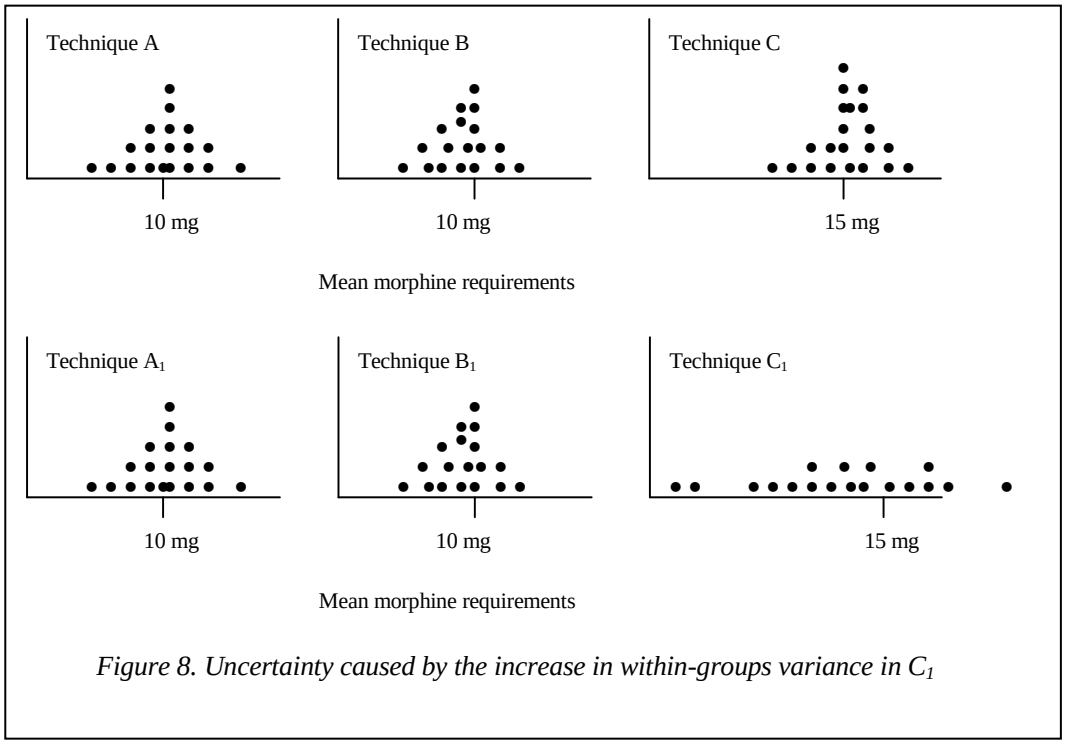


Figure 8. Uncertainty caused by the increase in within-groups variance in C_1

Calculation

The test statistic is a ratio called *F*. This is the ratio of the variances:

$$F = \frac{\text{Between groups MS}}{\text{Within groups MS}}$$

The larger the *F* statistic is above 1, the more likely there is a difference between the groups. The *F* statistic is then located in the *F* distribution tables at the appropriate degrees of freedom. This produces a *P* value which is the probability that *F* could have occurred by chance if the H_0 (no difference between the groups) was true. As before, the smaller this probability, the more likely it is that a real difference exists.

Small print: Simplistically, the between groups mean square is the mean square of the difference between the individual sample means and the grand mean. The within groups mean square is obtained by summing the individual SS for each sample and averaging the result.

Do you know *which* group is different?

No. ANOVA tells you that one is different but not which one is different. *Post hoc* tests must be carried out to determine this. Examples include t-testing with Bonferroni's correction, Scheffé, Neuman-Keuls, Tukey's Honestly Significantly Difference (HSD), and Dunnett's test.

T-tests with Bonferroni's correction

Definition

Correction factor which allows a *t*-test to be used to make comparisons between three or more samples.

Problem

Normally a *t*-test should only be used for one comparison eg the means of two samples.
If there are *three* samples and you wish to determine if any one sample is different from the other two, you must make *three* separate comparisons. The problem is that, if alpha is 0.05, every time you make a comparison, you are risking up to 5 % Type I error. Therefore, by the time you have come to your conclusion about the three samples, there is potentially a 15 % risk of Type I error.

Bonferroni's correction factor

To compensate for the above, instead of looking up the critical value for *t* at the $\alpha = 0.05$ level, you must look up the critical value for *t* at $\alpha = 0.05$ / number of planned comparisons.
Therefore, for three comparisons, the critical value for *t* would be looked up at $\alpha = 0.0167$ and for four patients (six separate comparisons) $\alpha = 0.0083$. In other words, your test value would have to lie in the outer 0.00415 of the t-distribution to be considered 'significantly' different.
As the number of comparisons increases it will get harder and harder to demonstrate a difference between the samples. This is, therefore, a less powerful method than ANOVA.

NON-PARAMETRIC TESTS

When are non-parametric tests appropriate?	Distribution of data is severely non-normal Ordinal or discrete quantitative data Small samples
Characteristics of non-parametric tests	Based on ranking Results are reported with the *median* and *range* rather than mean and SD Less powerful than parametric tests. Type II error more common
Assumptions	Samples are randomly selected Observations are independent
Wilcoxon rank sum test	Non-parametric equivalent to the unpaired *t*-test
Basic principle	The two samples are combined, ordered and ranked from lowest to highest. The samples are then separated again and the ranks summed in each. The next step is to determine whether there is a significant difference between the sums of the two groups. In the Wilcoxon rank sum test, tables list different sample sizes against rank sum ranges. If the smaller of your rank sums lies outside the relevant range, a difference is significant.
Mann - Whitney U Test	Non-parametric equivalent to the unpaired *t*-test
Basic principle	Rank all patients from smallest value to largest value and sum the rankings in each sample. The *U* statistic is then calculated to assess the likelihood of a difference between the rank sums. The equation is complicated and involves the sample size and rank sum. The U statistic is then located in *U* probability tables.
Wilcoxon paired-sample test	Non-parametric equivalent to the paired *t*-test
Kruskal-Wallis	Non-parametric equivalent of one-way ANOVA. Gives likelihood of a difference among the groups but not *which* one is different. This can be determined later using a Mann-Whitney *U* test
Friedman's test	Equivalent of repeated-measures ANOVA. Again based on ranking.
Spearman's rank order	Non-parametric equivalent of Pearson correlation coefficient

LINEAR REGRESSION AND CORRELATION

Purpose	Used to compare the relationship between two continuous variables where the relationship appears to be linear. eg blood pressure and blood loss.
Linear regression:	The drawing of a line that best describes / predicts the relationship between two variables
Correlation:	The assessment of the closeness of association between two continuous variables.

Linear regression

Assumptions:	-The relationship is *linear* -Observations are independent of each other. Multiple observations from the same patient or repeated measures over time are not permitted. -One variable must be an explanatory or independent variable and the other is the response or dependent variable. Not to be used for comparing two dependent values such as two measuring techniques. -For each value of x there is, potentially, a Normal Distribution of observed values of y
Process	The data are fed into a computer. Explanatory variables (from which observations are to be made) are plotted on the x axis and the dependent (outcome of interest) variables are placed on the y axis. The computer draws the best-fit line through the points by choosing a course which minimises the sum of the squared vertical distances between the individual points (y_i) and their imaginary equivalents ($\hat{y}$) on the line. This is called *least squares fit* and the plot of $\hat{y}$ at x is called the *regression of y on x.*

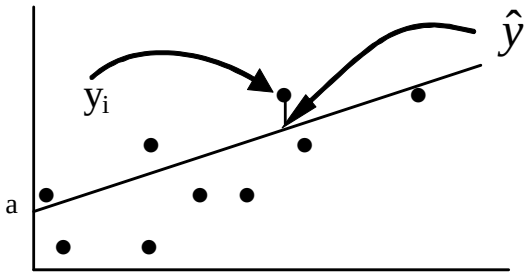


Figure 9. The regression of y on x

The computer then calculates the equation which describes the line and the proposed relationship:

$$\hat{y} = a + b.x$$

Where:

$\hat{y}$	predicted points on regression line
b	slope of line; defines the proposed relationship; *regression coefficient*
a	Intercept of y axis when x = 0

Values for *b*	b > 0	Positive relationship
	b < 0	Negative relationship
	b = 0	A line of no slope, therefore, no relationship
How precise is *b*?	The larger the sample the closer *b* will be to the true effect in the population. The precision can be gauged by reporting *b* with SE and CI.	

Could *b* be a random variation of zero?

The likelihood of this can be determined from the CI, which should not include zero, or by comparing *b* with 0 in a one sample *t*- test.

Example

Heights (cm) of children (independent variable) are plotted against their anatomical dead-space (ml) (dependent variable) to determine whether there is a relationship. (ref.eBMJ) Resulting regression equation:

$y = -82.4 + 1.033\ x$

Thus, if height was 110 cm, the anatomical deadspace would be:

$y = -82.4 + (1.033 \times 110) = 31.2$ ml

What influences the variation of y?

1.If y did not vary at all with x, y would be a horizontal line at the mean of y.

2.If y varies linearly with x, there is a slope to the line. If a perfect fit, the variation is said to be entirely due to the *regression.*

3. Random effects mean that the measured value of y may not be exactly on the predicted line ie there is *Residual* scatter

4.Non-random effects may also influence the scatter of y about the line. For example, as x increases the scatter of y might increase or decrease. If this is the case we need to transform data or use a different test.

Coefficient of determination (R^2)

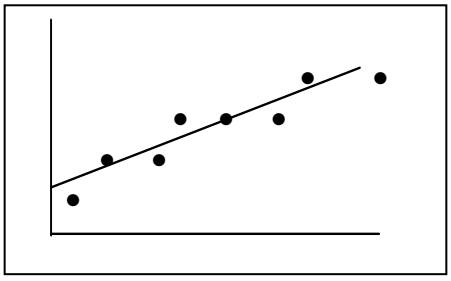
Allows us to subjectively assess the goodness of fit of the line to the data points by calculating the proportion of the total variation that is explained by the regression. For a regression line to have a good fit most of the variation of y will be due to 2 and little due to 3.

R^2 is the ratio of the variation explained by the regression (regression sum of squares) to the total variability (regression SS + residual SS). Thus, for example, if there was a perfect fit of the line to the points, the residual SS would be 0 giving an R^2 equal to 1.0

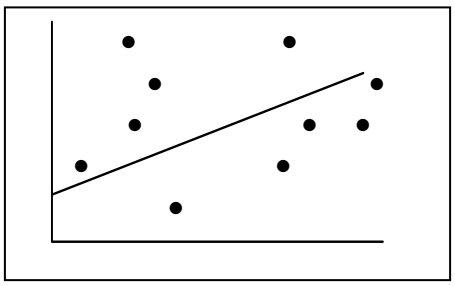
.

$$R^2 = \frac{\text{Regression SS}}{\text{Total SS}}$$

In the above example R^2 was 0.716. This means that 72 % of the variation between children, in the size of the anatomical dead-space, is accounted for by the height of the child.



Most of the variation of y is caused by the regression line. Good fit of line to data points. R^2 high



Lots of variation of y caused by residual scatter. Fit of line less good. R^2 smaller.

Pearson correlation coefficient (r)

Correlation is the assessment of how *likely* is the proposed linear relationship.

Values of *r*

1.0 or -1.0	Perfect correlation
0	No association at all
0.2 – 0.4	Mild association
0.4 – 0.7	Moderate association
0.7 – 1.0	Strong association

Calculation

First, if you haven't already confirmed linearity by regression, do a scatter plot to check that the relationship is linear. Computer software will calculate r but the equation is.

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

Note: This is mathematically equivalent to the square root of the coefficient of determination (R^2). In other words $r^2 = R^2$

$$r = \sqrt{\frac{\text{RegressionSS}}{\text{TotalSS}}}$$

Significance test for *r*

A *t*-test is used to test whether *r* is significantly different from zero.

In the above example the correlation co-efficient turned out to be 0.846 which suggests a strong association.

Spearman's rank correlation (r_s)

A non-parametric equivalent for Pearson's correlation coefficient. Used when the sample size is small (< 10 patients), where the variables are not Normally distirbuted or where one variable tends to increase in some fashion with respect to another but not necessarily linearly.

Basic method

The variables are ranked separately. The differences between the pairs of ranks for each patient is calculated, squared and summed. The sum is used in Spearman's rank correlation equation (see appendix) to give r_s which is interpreted in the same way as r.

Key Points

- Correlation is the assessment of the closeness of the relationship between two continuous variables.
- If the relation is linear, the test used is the Pearson correlation coefficient (r)
- The closer r is to 1, the more likely there is a relationship
- Linear regression is the drawing of the line that best describes the relationship between two linearly related variables.
- The equation of the line takes the form of y = a + bx where b, the regression coefficient, is the slope of the line and describes the proposed relationship
- The goodness of fit of the line to the data points is given by R^2

MULTIVARIATE ANALYSIS

Multivariate analysis

Definition

While univariate methods such as t-test or relative risk assess the relationship between an outcome and a single predictor variable, multivariate methods must be used to assess the relationship between multiple variables and an outcome. Some examples of multivariate methods are listed in Table.

Multivariate methods	Characteristic	Risk score
Multiple linear regression	Outcome on a numerical scale but binary confounders such as smoking and sex can factored in	Regression coefficient
Logistic regression	Outcome on a binary scale	Fitted regression coefficient Odds ratio Probability of an outcome
Mantel-Haenszel X^2test	Assesses relative influence of several groups of categorical data on an outcome. Similar in use to logistic regression	OR
Proportional hazards	Outcome time to event eg death	Hazard score
Discriminant analysis	More than two outcome categories	

Multiple linear regression

Definition

A method used to assess the impact of several variables on an outcome which has a numerical scale. In other words, after one variable x_1 has been shown to influence y, would another variable x_2 further influence y?

Example of a problem

Is there a relationship between birth weight and the dependent variables of maternal height and period of gestation? Birth weight can be shown in separate analysis to be linearly related to maternal height and gestation, but is gestation still important when the height of the mother is taken into account (and vice versa)?

Basic principle

The basic principle is to carry out separate regressions on the variables and add them one by one (largest first, then next largest etc) in the multiple regression equation below. At the same time, ANOVA and correlation analysis determine whether the addition of each successive variable improves the prediction of an outcome or whether it increases the residual scatter to a point of no significance.

$$y = a + b_1 x_1 + b_2 x_2 ....$$

Binary confounders

Binary confounders such as smoking and sex can be factored into the equation

Logistic regression

Definition

Regression analysis where the outcome is a binary categorical variable such as death. Often used in attempting to identify important factors in the production of an adverse outcome. If the predictive variables are binary as well, their relationship with the outcome can be expressed as an odds ratio (OR).

AGREEMENT

The Bland-Altman plot

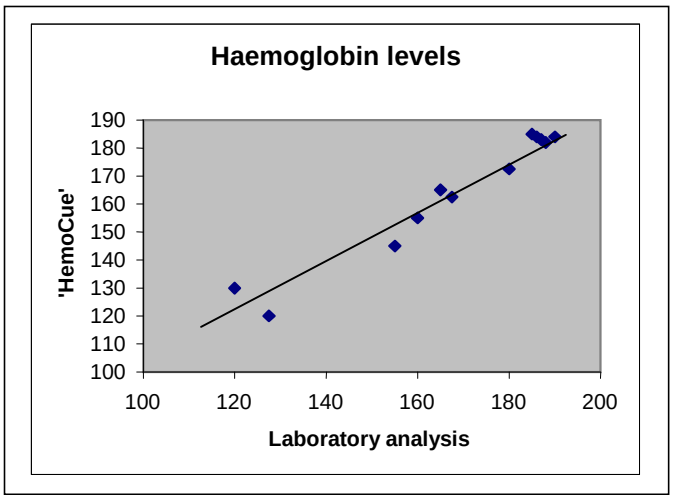
Why not use regression and correlation to assess agreement between measurement techniques?

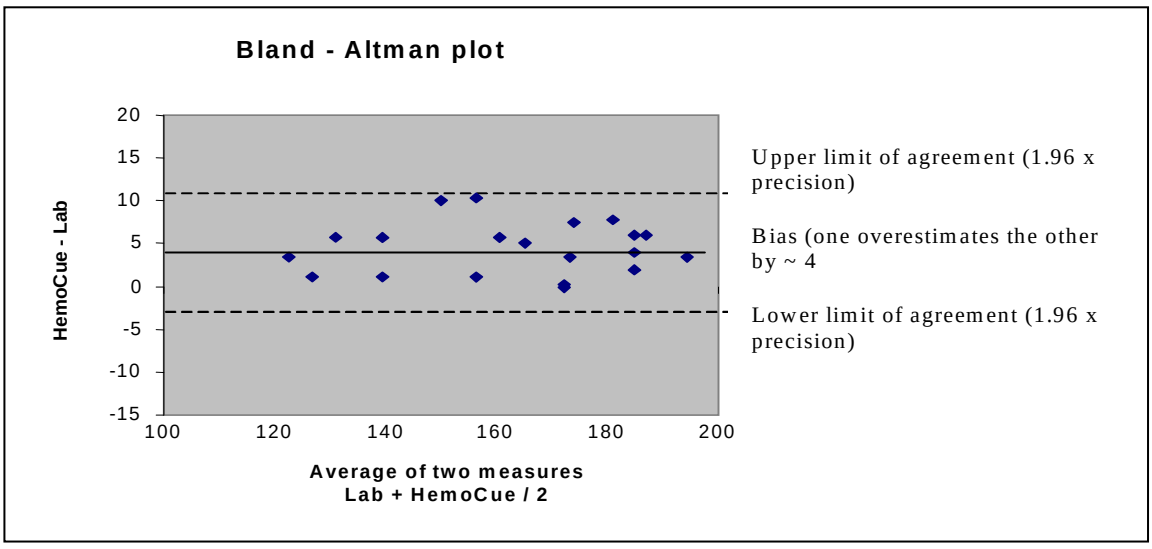
If the results of one measuring technique differ consistently from another by a constant amount, the correlation will appear strong but, in fact, the agreement is *poor*.

What is the Bland - Altman plot?

The Bland - Altman method is to plot the differences between each set of measurements against the mean of each set of measurements. If there is close agreement, a line around zero will be formed. If there is a consistent difference between the measurements, the plot will form a line above or below zero.

The graph on the right illustrates the plot of haemoglobin measurements by a laboratory technique against those made using a theatre 'HemoCue' device. The plot is a straight forward linear regression and shows that there is close correlation between the techniques...but is there agreement?





Figures 12 a and b. Comparison of a linear regression model and the Bland - Altman plot.

The Bland - Altman plot shows that there is a fairly consistent difference between the two techniques, with HemoCue values being about four units greater than lab results. This graphically illustrates that although correlation maybe strong, *agreement* is not. If agreement were 'strong' the mean values would follow a line of zero difference.

What is Bias and Precision ?

The mean difference between the measures is called the bias and the standard deviation of the difference is called the precision. The bias tells us how well the two measures agree in general. For example, a bias of +1.5 suggests that one produces readings which are, on average, 1.5 units greater than the other.
The precision gives us an indication of the spread of measures in the individual's study. If there is close agreement between the measures, there will be very little spread. If there is a large spread there will be uncertainty in the prediction of agreement or otherwise. As with the 5 percent percentage points in parametric testing, the standard deviation can be multiplied by 1.96 to give the limits of agreement, within which 95% of all values should lie.

The Kappa statistic

Definition

The Kappa statistic is used to assess the agreement or reliability between two observers who are performing a test which has a *categorical* variable.

Example

Two clinicians auscultating a group of elderly patients to determine which has aortic stenosis and which don't.

Basic principle

A 2 x 2 contingency table is constructed with the two observers' predictions. The actual agreement between the clinicians would seem, at first, to be easy to determine from this but the Kappa statistic adjusts the value by removing the proportion of agreement that is expected by chance alone.

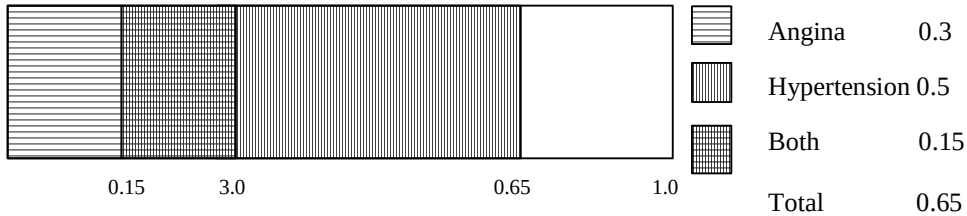
The kappa statistic

$$\kappa = \frac{A - E}{1 - E}$$

Where A is the proportion of times the observers agree and E is the proportion of agreement expected by chance.

PROPORTIONS

See also	Binomial distribution Chi square Calculation of power and sample size
Arithmetical handling of proportions	What is the probability of *two specific events* occurring when we know the *probability of each* occurring separately ?
Multiplicative rule	Used to calculate the probability of both of two, *unrelated, independent*, events occurring.
Example	What is the probability of a couple who are planning to have two children, having two girls?
Calculation	The probability of the first child being a girl is 1/2. The next time the probability of having a girl is 1/2 again but the *overall* probability that the second child is also a girl is half of this ie 1/2 x 1/2 = 1/4.
Additive rule	This is used when two events can occur together and allows you to calculate the probability of the occurrence of one or other of the two events or both events together.
Example	What is the probability that a randomly selected individual has either angina or hypertension (HBP) or both, when the prevalence of angina in the population is 0.3 and the prevalence of HBP is 0.5.
Calculation	The answer is 0.65 (rather than 0.8) because some of the 0.5 with HBP will also have angina and vice versa. See figure below



Probability (angina or HBP or both) = $P_{angina} + P_{hypertension} - (P_{angina} \times P_{hypertension})$

Figure. 13. Example of the multiplicative rule in proportions

CHI-SQUARE (X^2)

Definition	A test to assess whether there is likely to be a real difference in the *frequency* of a *categorical* event between two or more groups. The principle is to construct a *contingency* table and to compare the *observed* frequencies with those which would be *expected* if there is no difference between the groups.
Assumptions	The samples are randomly selected from a population, observations are independent and expected frequencies are not small.
Step by step example	Is there a difference in the frequency of vomiting between exposed and non-exposed patients. (Exposure might be antiemetic)

1. Construct a contingency table of the *observed* frequencies (***O***) :-

	Vomiting	*No vomiting*	*Total*
Exposure	8	22	30
No exposure	16	9	25
	24	31	55

Observed frequencies

2. Construct a contingency table for the frequencies that would be *expected* if the exposure made no difference to outcome. The calculation is straightforward:

Expected frequencies (***E***) = (column total x row total) / overall total

	Vomiting	*No vomiting*	*Total*
Exposure	13.1	16.9	30
No exposure	10.9	14.1	25
	24	31	55

Expected frequencies

3. For each cell calculate: $\frac{(O-E)^2}{E}$

4. Sum all four cells to get X^2 $X^2 = \sum \frac{(O-E)^2}{E} = 6.2842$

5. Consult the X^2 distribution. The calculated value for X^2 is located in the X^2 distribution tables at the appropriate degrees of freedom. (Degrees of freedom for X^2 = product of (number rows -1) and (number columns -1).

In the above example there is one df and the probability of there being no difference is p = 0.0122.

Further points about the X^2 test

The $\boldsymbol{X^2}$ test is carried out on the *actual numbers*, not the percentages or proportions.

The $\boldsymbol{X^2}$ test mentioned above is strictly called the Pearson $\boldsymbol{X^2}$ test.

The $\boldsymbol{X^2}$ distribution is a continuous frequency distribution of probabilities which is consulted in numerous tests other than the above.

Sample size: The probabilities obtained from $\boldsymbol{X^2}$ assumes the distribution of probabilities is continuous. This is not the case. The data is discrete and there are only a limited number of probabilities possible for each scenario. This leads to an increase in the possibility of Type 1 error particularly if numbers are small

Fisher's exact test

This is the preferred alternative to the Pearson $\boldsymbol{X^2}$ test in 2 x 2 contingency tables where the sample size is insufficient. Sufficiency is defined as:
All expected values must exceed 1
80% of expected values must exceed 5.

Fisher's exact test is a complex calculation and may engage a computer for several hours in a large contingency table.

Yates continuity correction

An alternative to Fisher's but little used now. The *continuity* correction is applied to reduce the overall value of $\boldsymbol{X^2}$.

$$X^2 = \sum \frac{[(O-E)-0.5]^2}{E}$$

Other forms of Chi-square	
Chi-square for larger tables	Estimates if there is a difference in the frequency of an observation among several groups
Chi-square test for trend	If there is a natural order to the groups, this test looks for an increasing or decreasing trend Usually used for a table where there are two rows (eg fatty diet and non-fatty diet) and several columns of a variable that increases in value (eg skin fold thickness). Does the proportion of those in the fatty diet group increase with increasing skinfold thickness? The above test is able to determine whether such a relationship exists and is a much more powerful test than the usual chi-square.
McNemar's chi-square test	Used with paired data, such as the frequency of an observation in a single group of patients before and after an intervention.
Mantel-Haenszel	A multivariate test which can be used to assess the impact of confounders on a group outcome.

RISK

Chi-square vs risk analysis

While Chi-square assesses whether there is likely to be a real *numerica*l difference in the frequency of an event between groups, risk analysis gives an indication of the *strength of association* between the groups. There are several ways to score risk. Three of the most common are the relative risk, odds ratio and number needed to treat.

Incidence (rate) of disease

Quantifies the number of *new* cases of disease that develop in a population at risk during a specified time period

Prevalence of disease

Quantifies number of *existing* (new and old) cases of disease in a population at a given point or period in time

Cohort studies

A study where a sample of patients. some of whom are exposed to a risk factor and some not, are followed over time to determine which develop the disease. Almost always *prospective* although it is possible to follow a cohort retrospectively. The most commonly used risk score is the *relative risk* although *odds ratios* may also be used.

	Cancer	**No Cancer**
Smoking	a	b
Non-smoking	c	d

Case-control studies

A study in which *cases* are identified retrospectively as having a disease (eg DVT) and compared with *controls* without disease. The number of *cases* and *controls* which had the exposure of interest (eg OC pill) is compared. Risk analysis is with the *odds ratio.*

	DVT	**Controls**
OC pill	a	b
No OC pill	c	d

Relative risk (risk ratio)

Definition

Relative risk is the ratio of the *incidence* of disease among exposed to the incidence among non-exposed. Also called the *incidence risk*.

$$RR = \frac{\text{incidence among exposed}}{\text{incidence among non - exposed}} = \frac{\left(a \big/ a+b\right)}{c \big/ c+d}$$

Key points

-RR is a true risk in that a RR of 3.0 means there is three times the risk and a RR of 0.5 implies the risk has been halved. A RR of 1.0 implies no association.

-The RR is reported with a CI. If the CI includes 1.0 the RR is not significant.
-The RR is a common risk score in cohort studies.

Odds ratio (OR)

Definition

The odds of disease is the number of cases who have disease divided by the number who do not have the disease. The odds ratio is the odds of the disease in exposed over the odds of the non-exposed.

$$OR = \frac{\text{odds of disease in exposed}}{\text{odds of disease in non - exposed}} = \frac{\left(a/b\right)}{c/d}$$

Key points

-Unlike RR, the OR does not give an *exact* value for risk. In general they tend to overstate the risk, being smaller than the RR for values over 1.0 and less than RR for values under 1.0. However, the OR is approximately the same as the RR when the outcome is rare.
- As with RR, an OR of 1.0 implies no association
- OR is reported with a CI. If the CI includes 1.0 the OR is not significant.
- In retrospective case-control studies, the OR must be used rather than RR because there is no information on the numbers of all exposed and non-exposed.

A useful acronym?

Backward	co**H**ort
OR	**A**head (ie prospective)
co**N**trol	**R**R
Exposure	**D**isease

Number needed to treat (NNT)

The number of patients who need to be treated in order to avoid one adverse event. The NNT is the reciprocal of the absolute risk reduction.

What advantage does this have over RR?

The NNT gives the RR some relevance in terms of the magnitude of clinical effect. For example, if the incidence of an adverse event is only 0.6:1000 (0.06 %), a 33% reduction in risk (RR = 0.33) will produce an absolute risk reduction of only 0.02 %. The NNT to prevent one adverse event would be 5000. However, if the adverse event had an incidence of 6:100 (6 %), a 33 % risk reduction would produce an absolute risk reduction of 2 % and a NNT of only 50.

Re-cap on calculation of NNT

An intervention reduces mortality from 45 % to 25 %
Absolute risk reduction = 20%
NNT = 100 / 20 = 5

For other risk scores

See appendix

Confounding variables in case-control studies

Definition	A form of bias that occurs when the demographics of the groups studied are different and those demographics influence the outcome.
Example	In a study that aims to compare the incidence of urinary retention between PCEA and PCA, the mean age should be the same in both groups as the elderly have a higher incidence of retention than the young. As well as age, other common confounders include gender, BMI, coexisting medical conditions.
Prevention of confounding problems:	
Design stage	Large samples Randomization Stratum matching eg several studies with different age groups Matched design (see appendix)
Analysis stage	Subdivide into different age groups and analyse each separately. Mantel-Haenszel test. A multivariate test which can be used to assess the impact of confounders on a group outcome. Logistic regression

PREDICTIVE ABILITY OF TESTS

How is the quality of a test assessed?

Compare test with a gold standard or assess its ability to predict a clinical outcome

Example

The value of intra-operative intra-operative ST segment depression might be assessed by comparing with echo detection of regional wall motion abnormalities or by its ability to predict an adverse cardiac outcome such as myocardial infarction

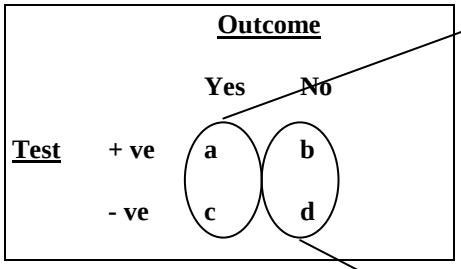
What methods are there to assess a test?

Sensitivity and specificity
Positive and negative predictive values
Receiver operating characteristic curve

Sensitivity

The ability of a test to detect the *disease*; the proportion of disease that was correctly identified; the true positive rate.

ie true positive rate = a / a + c = 1 - false negative rate



Specificity

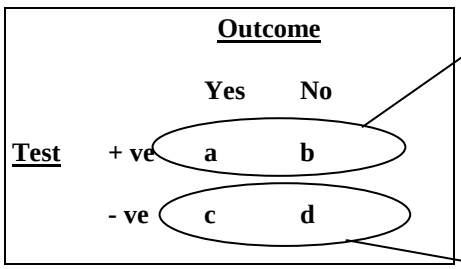
The proportion of *no-disease* that was correctly identified
The true negative rate. = d / b + d = 1 - false positive rate
(*I usually think of a highly specific test as one with few false positives*)

Key points

Particularly important in screening tests
Not affected by the prevalence of disease
Increased sensitivity is usually at the expense of specificity

***Positive predictive value*:**

The proportion of a test's positive *results* which are true positives.
a / a + b



Negative predictive value

The proportion of a tests negative results which are truly negative.
d / c + d

Key points	Positive and negative predictive values take into consideration the prevalence (*prior probability*) of the disease. If a disease is common, such as ischaemic heart disease in vascular patients, intra-operative ST segment depression is truly likely to represent ischaemia. If ST segment depression occurred in obstetric patients it is unlikely to be a true positive. This would be reflected in a higher PPV in the vascular group than the obstetric group. The sensitivity and specificity of the test would, however, remain te same. An example of this is found in the appendix. When recalling the differences between sensitivity/specificity and predictive values remember that sensitivity/specificity looks at *disease* and absence of *disease* and determines how much was picked up correctly. On the other hand, positive predictive values (PPV) and negative predictive values (NPV) look at the subject from the point of view of the tests positive and negative *results* and determines which were correct.

Receiver operating characteristic curve

What is it?	A plot of sensitivity against false positive rate for several values of a diagnostic test.
What is it used for?	Used to illustrate the trade-off between sensitivity and specificity in tests that produce results on a numerical scale, rather than as an absolute positive or negative result. The ROC curve can be used to compare different tests or to help choose the cut-off points.
How is it formed?	Take Troponin I levels in the diagnosis of myocardial infarction, for example. Several different Troponin plasma concentrations would be chosen and compared against a gold standard in diagnosing MI, such as echocardiographic evidence of new and permanent wall motion abnormality. The sensitivity and specificity of each chosen Troponin level would be determined and plotted.
What would a good test look like?	The ideal cut-off point is one which picks up a lot of disease (high sensitivity) but has very few positives (high specificity). One is usually a trade off for the other. A test that produced one false positive for every true positive is very poor and the plot would follow the diagonal. The ideal cut-off point would, in most cases, be high on the left hand side of the graph and would lead to a large AUC. If the consequences of a false positive result were worse than those of a false negative result the chosen cut-off point would be lower and further to the left.

Likelihood ratio

What is it?	The LR is a statistical tool which enables you to assess the actual chances of a patient having a target disorder (the post test odds) if a test result has reached a particular level. To calculate the post test odds, the pre test odds (usually the prevalence of disease in the population) are multiplied by the LR.

How do you calculate the LR?	The likelihood ratio of a positive test result (LR+) is sensitivity divided by 1- specificity. The likelihood ratio of a negative test result (LR-) is 1- sensitivity divided by specificity.
Example from Greenhalgh	Prevalence of iron deficiency anaemia 5% Pre-test probability = 0.05 ∴ Odds of having IDA = 0.05 / 0.95 = 0.053 LR of IDA when Ferritin level between 18 and 45 ug = 3 Post test odds of IDA = 0.53 x 3 = 0.159 Which is equivalent to post test probability = 14%

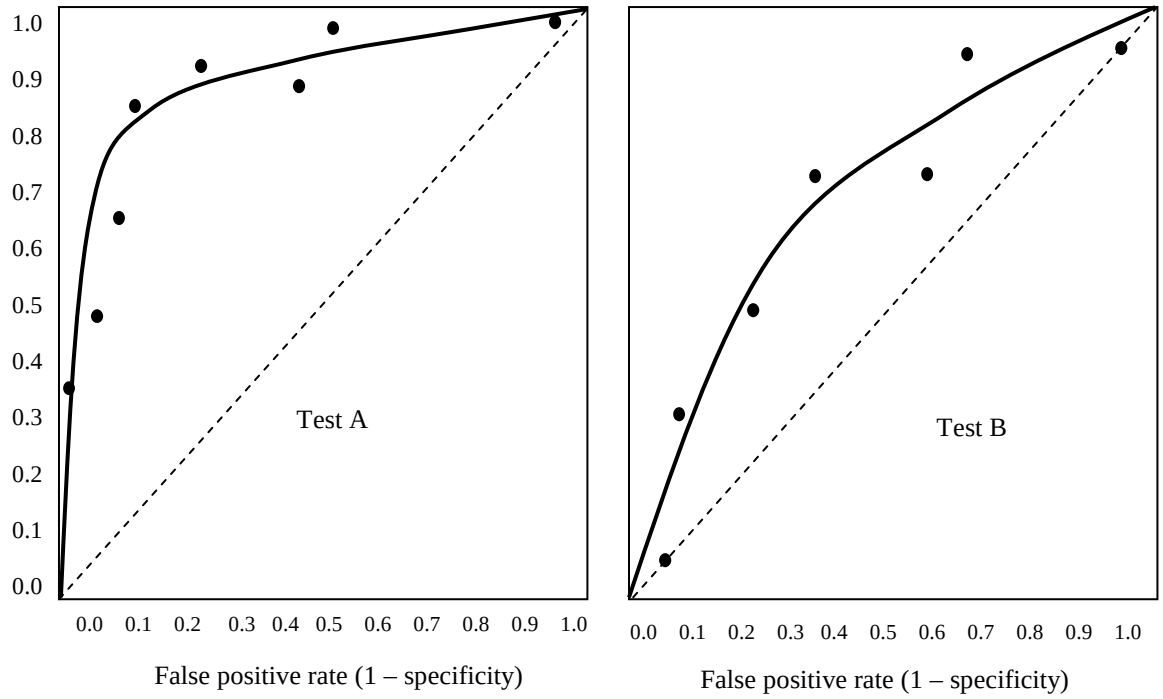


Figure 14. ROC curves for two tests. Test A can be seen to be a better test than B because the ROC curve extends much higher into the top-left part of the graph. This means that there are results in that portion that have high sensitivity (high pick-up rate) and high specificity (few false positives). the quality of the test can be quantified by measuring the area under the ROC curve..

POWER AND THE CALCULATION OF SAMPLE SIZE

The problem

It is unethical and a waste of time and money to embark on a study to see if a drug is effective, if there is a significant chance of a false negative result. The commonest cause of a false negative result is that the sample size is too small. The larger the sample, the more likely it is that the true effect of the intervention will be demonstrated.

An example

We all know that coins come down, on average, 50% heads. This is the 'true' effect. But how many times would you have to toss a coin to convince a sceptic that, on average, it *does* produce 50% heads? If you only tossed it six times, the *most likely* proportion would be 50% heads but it could quite easily be markedly different from that, with even 0% or 100% being eminently possible! If, on the other hand, you tossed it one million times you would almost certainly produce the 'true' value of 50% heads (± a very small decimal point) with proportions markedly different from that being almost impossible. In research we cannot have samples of one million, so we have to compromise and say *'What is the smallest sample I need to be almost certain of producing the true result?'*.

Power

The Power of a study is the chance of it successfully demonstrating the 'true' result. Power can also be expressed as one minus the false negative rate or (1-β error).

Required input to sample calculation

1) First decide what will be regarded as the desired **effect size**. This might, for example, be a desired fall in blood pressure when an anti-hypertensive is to be compared with a placebo. The actual input into the calculation differs depending on the sort of study you are carrying out. You may perhaps be required to input both a desired **successful effect** (π) and what will be regarded as no effect (the **null hypothesis value**, π_0). The smaller the effect size the larger the required sample.

2) Next you decide on how certain you want to be of picking up the true effect. In other words, you decide on the **Power**. Conventionally we usually want a power of 80 to 90 %. Remember that the higher the power the larger the sample that is required.

3) Choose your significance level, the **alpha** value. Usually we will choose 0.05. The smaller the alpha value, the larger the required sample.

4) Studies that are assessing the difference between sample means require a prediction of **variance** within samples. This can be '*guesstimated*' from pilot studies or literature searches. The larger the variance, the larger the required sample

Basis of calculation

Remember that *even if a drug has the desired/predicted effect*, a study will not necessarily reproduce that *exact effect.* Because of random variation, there is a range of possible results centred around the true value. There will also be a spread of possible outcomes around what you regard as the null hypothesis value. The two distributions of positive and negative possibilities are likely to

overlap. The only portion of the distribution of all possible positive results that can be regarded as significantly different from the distribution of negative results, is that which are beyond the 5% percentage point of the negative distribution. If you want a power of 80%, this proportion must be at least 80% of the total possible positive results.

As we know, as the sample sizes get bigger, their distributions become narrower and narrower as the majority of results cluster around the true values (ie the mean). There will, thus, be less and less overlap between the chosen negative and positive value distributions as sample sizes get bigger.

In the practical setting this means that if you choose a large sample, and the drug has the effect you desire, the study result is likely to be near to the true result and less likely to fall within the 'non significant' range

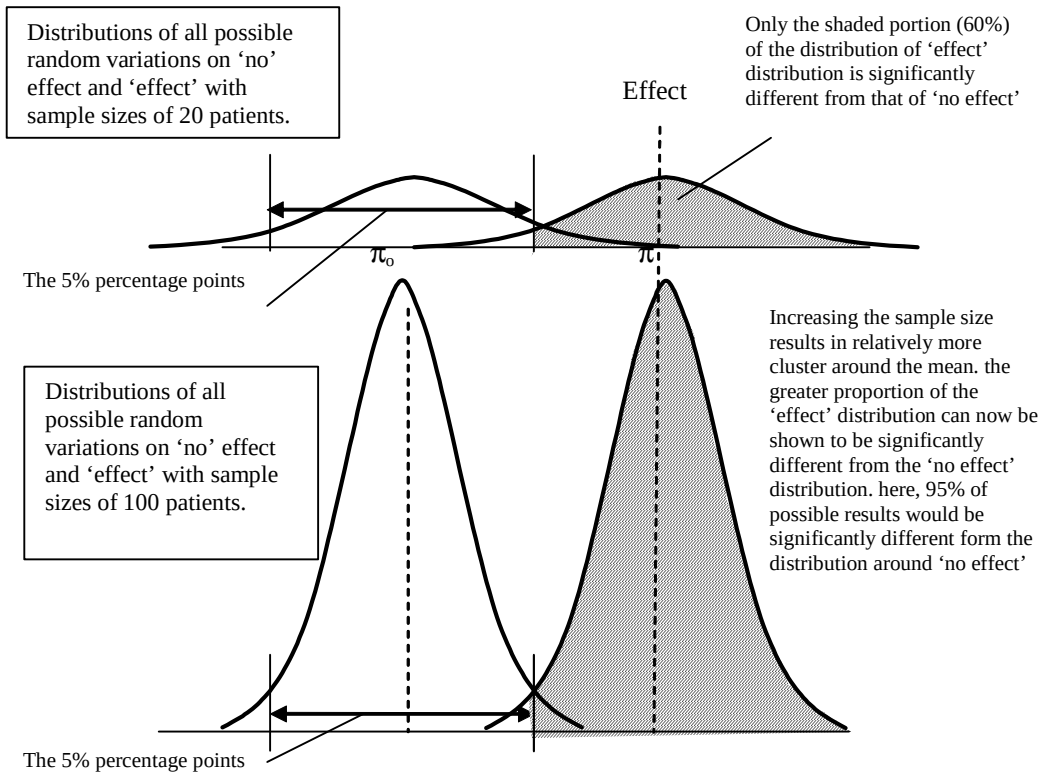


Figure 15. Illustration of how 'no effect' and 'effect' distributions separate as sample size becomes larger. This increases the probability of showing that an effect is significantly different from ' no effect'.

Sample size calculation

Approaches:

1.) Formulae

There are complex formulae for estimating the required sample size. Different formulae are required for different study designs. These can be obtained from text books but mostly we simply use computer software. The basic input to each formula is similar in concept:

Comparing Proportions:

Significance level (α)
Desired power (90 %) (actual input is 1-power or the β error)
π = proportion of interest (0.7 here)
π_0 = null hypothesis proportion (0.5)

Comparing the means of two samples:

Significance level (α)
Desired power (actual input is 1-power or the β error)
Proposed difference between the means (Δ)
Standard deviations of the samples

2) Lehr's quick formulae

For unpaired t-test or Chi-squared test there is a quick formula for calculating the sample size for a power of 80% and an alpha value of 0.05. It requires the calculation of the *Standardized difference* which is the effect size divided by the standard deviation.

$$\text{Sample size} = \frac{16}{(\text{standardized difference})^2}$$

(For a power of 90% the numerator is 20)

3) Altman's normogram

This normogram can be obtained from texts. A line is drawn between a column of Standardized differences and a column of Powers. This line will intersect a third column of sample sizes (N). The sample size can be read off at two different alpha values. For an Unpaired t-test you use N/2 for each sample.

Further points regarding sample size calculation

Adjustment for losses/non-compliance

In most studies there will be losses for various reasons such as loss to follow-up and non-compliance.

Losses
Non compliance

Inflate sample size by: 1/ (1-predicted total loss rate)
Inflate by $1/(1\text{-predicted total non-compliance rate})^2$

Precision based calculation

The above sample size calculations are aimed at power based hypothesis tests. Precision based sample size calculations are used when you want to estimating a variable to within a certain level of precision. Thus, instead of factoring in Power you stipulate a certain CI for the estimate. The narrower the CI, the greater the sample size required

Sequential trial design

This allows a clinical trial to be carried out so that, as soon as a significant result is obtained, the study can be stopped, thus minimising the sample size, cost and morbidity.
As each patient is tested, their results are plotted on a graph. The upper and lower borders of the graph are drawn depending on the number of patients tested, the power and the desired significance level. If A is better than B,

the line moves one up and one to the right and, if B is better than A, the line moves one down and one to the right. When the line crosses an upper or lower border, a significant difference has been attained and the study stopped. If the line crosses the right hand border there is no difference between the groups and the study is also stopped. (Figure 16)

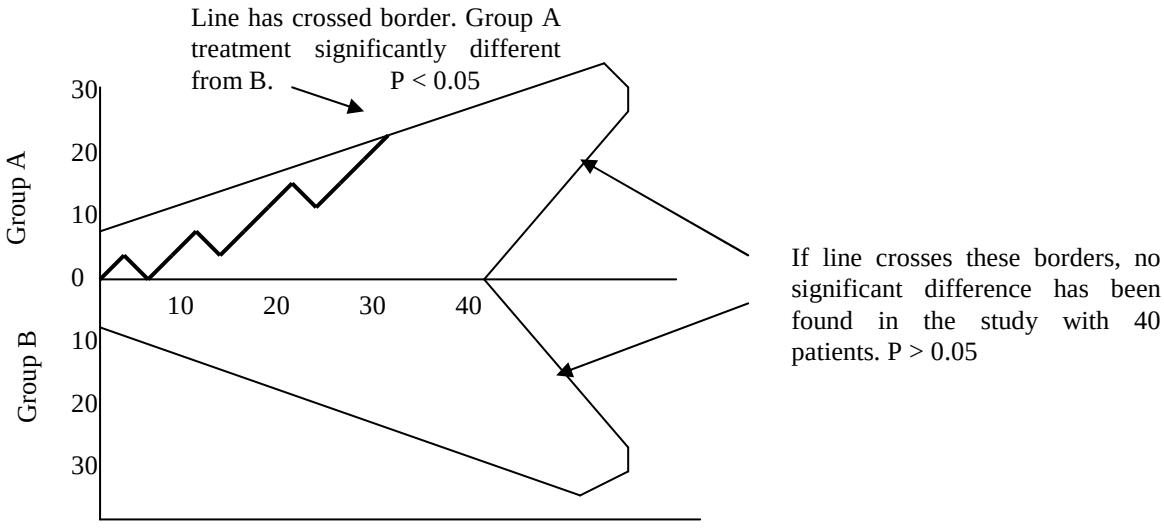


Figure 16. Sequential trial design

ERRORS IN RESEARCH

Random error

Definition	Error introduced by a lack of precision in conducting the study. Reduced by meticulous technique and by studying large numbers.

Bias

Definition	The introduction of a systematic error. Not reduced by increasing sample size. *Experimental bias* specifically is a bias towards a result expected by the human experimenter. The list of potential areas for experimental bias is huge. Sackett in Journal of Chronic Disease 1979 produced a comprehensive list. The basic classification and some examples are listed below.
1. Reading up on the field	Only reading articles that agree with experimenters view Only reading articles with positive results Only reading hot, topical articles
2. Specifying and selecting sample	Inadequate sample bias Volunteer bias – volunteers respond differently to patients Confounders not accounted for Randomisation inadequate
3. Executing the experiment	Withdrawal bias – withdrawals not accounted for Compliance bias Contamination bias Influencing experiment through personality Inadequate or bogus control
4. Measuring outcomes	Insensitive measure bias Expectation bias – pushing result towards expectation Recall bias Instrument bias End digit preference – rounding measurements up needlessly
5. Analysing the data	Post hoc significance- assuming cause and effect Data dredging / torture Over simplifying Repeated peeks bias – should not make judgements until pre-determined sample numbers achieved Using wrong test – (see below)
6. Interpreting analysis	Correlation / cause and effect bias Magnitude bias – suggesting the effect is bigger than it is Significance bias - suggesting significance greater than it is

See chapter on Study Design for compressed alternative version

COMMON BIAS		
PROBLEM	EXPLANATION	PREVENTION
Selection bias	One group has different risk than the other	Randomization, Cross over
Detection bias	Observations in one group not sought as diligently as in the other.	Blinding
Observer bias	The observer is able to be subjective about the outcome	Observer blinding, outcome

		measure design
Recall bias	The patients' 'treatment group' allocation influences the way they report past history and symptoms. Eg if the patient knows they are in the placebo group they may exaggerate their 'untreated' symptoms	Patient blinding
Response bias	Patients who enrol in a trial may not represent those of the population as a whole. e.g the obese patients who enrol in a weight loss medication trial may be more motivated than those in the general population.	Random selection from population
Regression to the mean	Random effects may cause a rare, extreme variation on a measurement. If the measurement is repeated, the likelihood is that the measurement will be less extreme. Thus, if a treatment had been given after the first measurement, it would erroneously appear, on the basis of the second measurement, that it had had an effect.	Control group
Hawthorne effect	The actual process of studying and following up patients influences the outcome. eg chronic headache may improve in patients who are being studied and regularly followed up.	Control group; mask intention of study from patient

Sample size too small

This introduces a form of bias in that a false negative result is more likely. Type II error increased.

Confounding

Definition

A form of bias that occurs when the demographics of the groups studied are different and those demographics influence the outcome.

Example

In a study that aims to compare the incidence of urinary retention between PCEA and PCA, the mean age should be the same in both groups as the elderly have a higher incidence of retention than the young. As well as age, other common confounders include gender, BMI, coexisting medical conditions.

Prevention of confounding problems:

Design stage

Large samples
Randomization
Stratum matching eg several studies with different age groups
Matched design (see appendix)

Analysis stage

Subdivide into different age groups and analyse each separately.
Mantel-Haenszel test

Errors at analysis stage	
ERROR	COMMENT
Parametric tests used instead of non-parametric	This error may occur if: - the population is not normally distributed - the sample is too small to be sure of its' population distribution - ordinal data are treated as interval data
Non-parametric test used instead of parametric	Parametric tests are more 'powerful' and should be used when appropriate
Multiple inter-group t-test comparisons instead of ANOVA	Increases the chance of type I error
Paired data treated as unpaired	Increases the chance of type II error
One tailed test used instead of two tailed test	Increases chance of type I error
Chi- square used when numbers too small	- Yates correction should be used in 2 X 2 tables - Fisher's exact test should be used if expected value for two or more cells is < 5

Errors at presentation and publication stage	
ERROR	COMMENT
Failure to report data points or SD or SEM	Unprocessed raw data is helpful in interpretation
Reporting mean with SEM rather than SD	SEM is a smaller, more processed number. Gives false impression of a trend in a sample
Failure to give explicit details of study design and statistical analysis	
Publication bias	Negative studies less likely to be submitted and / or published than positive ones All well conducted studies should be submitted and (ideally) published. In meta-analysis, absent negative study should be sought for by way of funnel plot analysis.

SYSTEMATIC REVIEWS AND META-ANALYSIS

Systematic review	A systematic review is a highly structured process in which an attempt is made to answer a specific clinical question by collating and analysing the data from all relevant trials. The key elements of a systematic review are outlined in the table below.
Meta-analysis	The *mathematical* process by which the data from several trials are combined to give a single pooled estimate of effect. Usually part of a systematic review.

Systematic reviews: Key steps	
Focussed clinical question (FCQ)	A systematic review addresses a specific clinical question
Inclusion and exclusion criteria	The studies are selected using clearly defined predetermined inclusion and exclusion criteria. Importantly, studies must only be included which address the specific FCQ. Other considerations will include the type of trial, language, outcome measures, methodology etc
Sources for the search	The sources for the search are decided beforehand and are clearly documented. Sources are likely to include all online databases, a hand search of anaesthesia journals, reference lists from journals, citations, and personal consultations with experts.
Outcome measures	Trials addressing the same FCQ may have slight variations in outcome measures. The protocol for the systematic review must define the specific outcome measures to be used.
Validation	The fact that the studies have been published does not necessarily mean that they have been adequately validated. Once studies have been identified they must be properly validated before inclusion in the meta-analysis. Validation will often be carried independently out by two individuals, one preferably without expertise in the topic under review. They assess adequacy of treatment allocation concealment, blinding, consistency of trial management, patient withdrawals during the trial etc
Assessment of heterogeneity	See below
Meta-analysis	See below
Reliability of pooled result	The pooled result becomes more credible when there is a big difference in treatment effect, a statistically significant difference in treatment effect, consistency across the studies, indirect evidence to support the difference, biological plausibility.
Sensitivity Analysis	Once a pooled effect measure has been reached it is worth re working the analysis by, for example, using an alternative mathematical model or by excluding outliers or by excluding trials of arguable quality. If you find that fiddling with the criteria in this way makes very little difference to the conclusions, the findings are relatively *robust.* If the findings disappear, the conclusions should be expressed more cautiously
Conclusions and discussion	

Heterogeneity

What is heterogeneity?	Heterogeneity is diversity among study results greater than you would expect by chance alone.
Types of heterogeneity	*Clinical* heterogeneity is where there are significant differences in patient demographics between the studies *Methodological* heterogeneity is where there are significant differences in the conduction and methods between trials.

	Statistical heterogeneity is usually a consequence of the above two and is a significant difference between the results of the studies. When the term 'heterogeneity' is used alone it usually refers to statistical heterogeneity.
Assessing heterogeneity	A utopian view of meta-analysis is that it can combine results from several identically conducted small studies and reach a more reliable conclusion on the large pool of data. In this ideal situation, the selected study results will not differ much from each other because they should all be conducted on the same types of patients and with the same methodology. Any difference in results will be small and due to chance. What, then, if the results of the selected studies differ rather more from each other? The reviewer must decide whether the differences between study results are so big that chance can't account for it all. The latter would imply that there may be differences in trial methodology which have resulted in truly different effects, in which case trying to combine them to get a single pooled effect is inappropriate. There are two tests to help make an objective decision:
i) Chi-square test for heterogeneity	The more significant the test result, the less likely it is that the differences between trials are due to chance alone. Unlike much of medical statistics, an alpha value of 0.1 is employed here. Thus, a *P* value less than 0.1 is an indication that heterogeneity is significant and that perhaps the trials are not combinable. As a rule of thumb, the Chi-square should not be more than the degrees of freedom (number of trials – 1). If it is, heterogeneity is probably significant.
ii) I^2	I^2 is the percentage of variation across the trials that is due to heterogeneity rather than chance alone. I^2 less than 25% is low heterogeneity, > 50% significant and > 75% is high.
Strategies for heterogeneity	Ignore Check data Do not undertake a meta-analysis Explore and report the cause Random effects meta-analysis Change effects measure Exclude outlier studies

Meta-analysis

Fixed effects model	A meta-analysis technique that takes the standpoint that there is a single treatment effect or one true answer. Any variation between studies is solely due to random variation on that one true answer. The final estimate is, therefore, the best estimate of the proposed single treatment effect.
Random effects model	Takes the standpoint that there are a variety of similar treatment effects. The final result is therefore the average of several treatment effects.
Pooled treatment effect	The pooled treatment effect is calculated as a weighted average, with larger studies' results carrying more weight in the calculation. The final pooled effect is commonly presented as an OR or RR.

The Forest plot

A graphical display of the results of a meta-analysis.

The following example is taken from a systematic review that examined hypotension during spinal surgery for caesarean section

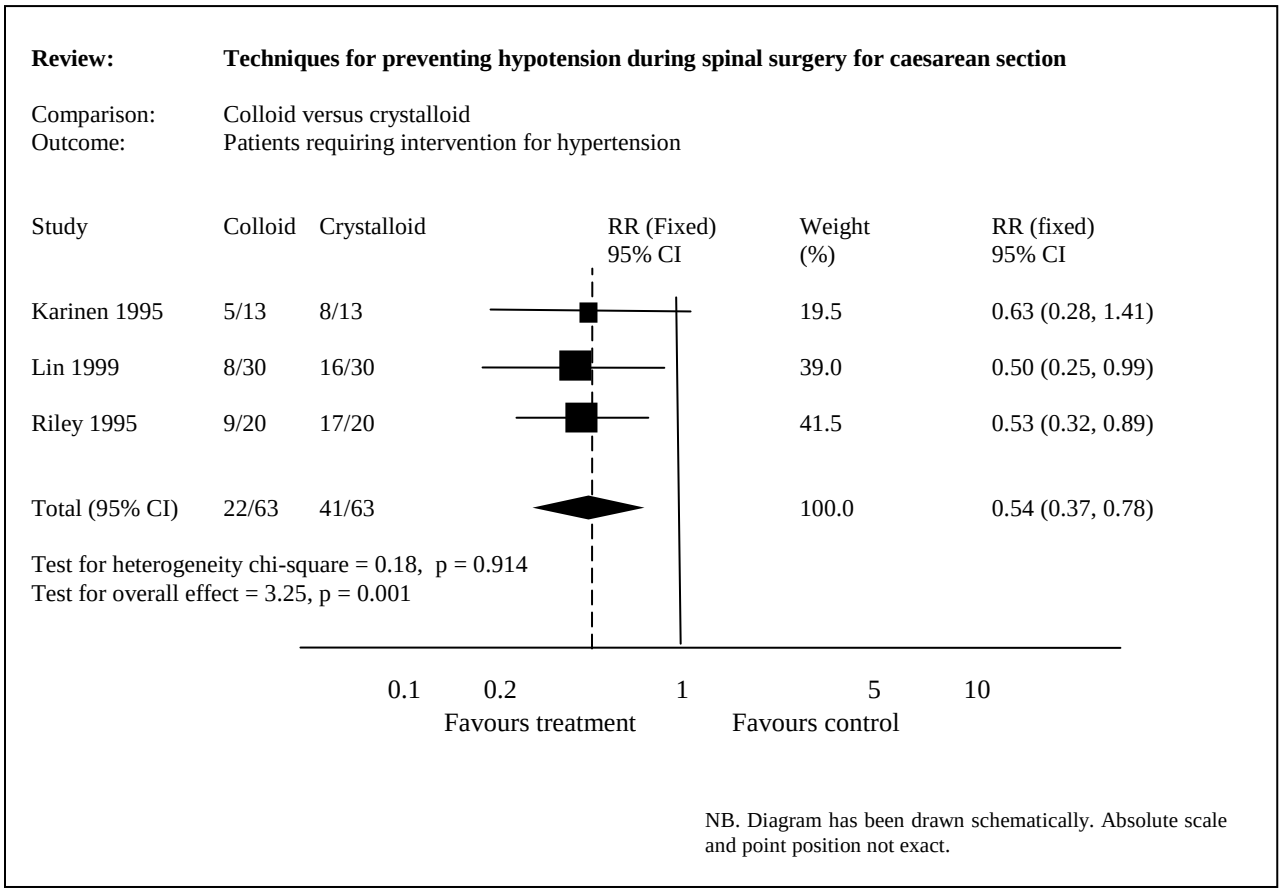


Figure 17: Forest plot

Note the following points:

i) The comparison and outcome measure are stated at the top left corner. There may be several meta-analyses looking at different comparisons in each systematic review.
ii) Individual studies are plotted as boxes on the vertical axis. Size of box represents weighting of each study.
iii) Weighting predominantly by sample size. Confidence interval of study is plotted.
iv) Effect measure (commonly the OR or RR) is plotted on a log scale on the horizontal axis. This means that increases and decreases in risk of the same magnitude have the same visual separation on the horizontal scale.
v) Vertical reference line drawn at position of no treatment effect. (OR or RR = 1)
vi) Pooled OR or RR displayed as a diamond and, sometimes, a vertical dashed line. Width of the diamond is the confidence interval of the pooled effect.
vii) A *P* value is given for the strength of the overall effect.
viii) The result of the test of heterogeneity is displayed with a *P* value. If $P < 0.1$, there is significant heterogeneity.
ix) I^2 statistic may also be displayed.
x) The mathematical model (fixed or random)is displayed

Funnel plot

An adjunct to meta-analysis to aid the detection of bias. The funnel plot is a scatter plot of treatment effect against a measure of study size, the latter generally being represented by sample size. Normally one would expect larger studies to cluster near the true effect and smaller studies to have more scatter. In other words, the precision in estimating the true effect increases as the study size increases. The scatter of smaller studies should, therefore, be symmetrical about the 'true effect' because the scatter should be solely due to random independent factors.

If there is asymmetry (a 'hole') there may be bias, particularly selection bias. The commonest cause of this is publication bias where small 'negative' studies have failed to be published. Other causes for an asymmetrical funnel plot include poor methodology of small studies, true heterogeneity and fraud.

Note: Increasingly, the SE is plotted on the y axis instead of the sample size. This is because imprecision may arise even in large studies, if the effect size is small.

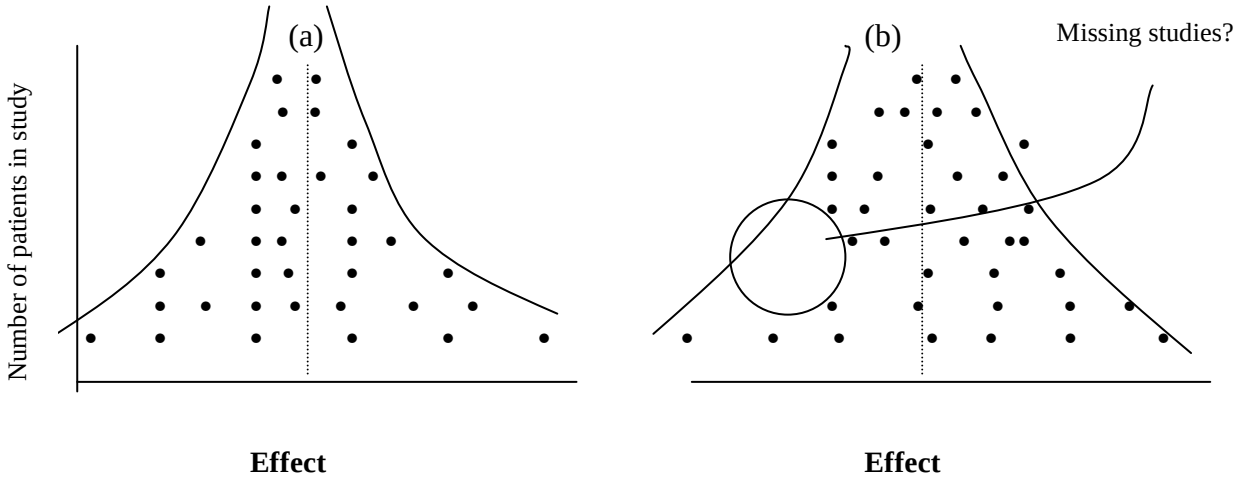


Figure 18: The Funnel Plot

Random variation will cause a spread of study results around the 'true' result. The larger the study, the closer it's result will be to the 'true' result. A plot of results of all the studies against their size should, therefore, give a *funnel shape*.(a) Large holes in the funnel suggest there has been publication bias. (b)

EVIDENCE BASED MEDICINE

Basic points

Definition	The conscientious use of mathematical estimates derived from high quality research to make decisions about the clinical management of individual patients.
Key components of EBM	Ask an answerable question Track down best evidence Appraise / validate evidence critically Implement results in clinical practice Evaluate performance

Thus, EBM requires you not only to read papers but to read the right papers at the right time and then alter your behaviour.

Advantages	Poor quality evidence may lead to morbidity / mortality Traditional reviews prone to bias - systematic process more reliable RCT's reduce bias and weighted heavily in EBM Mass of literature too much to read - EBM provide summaries

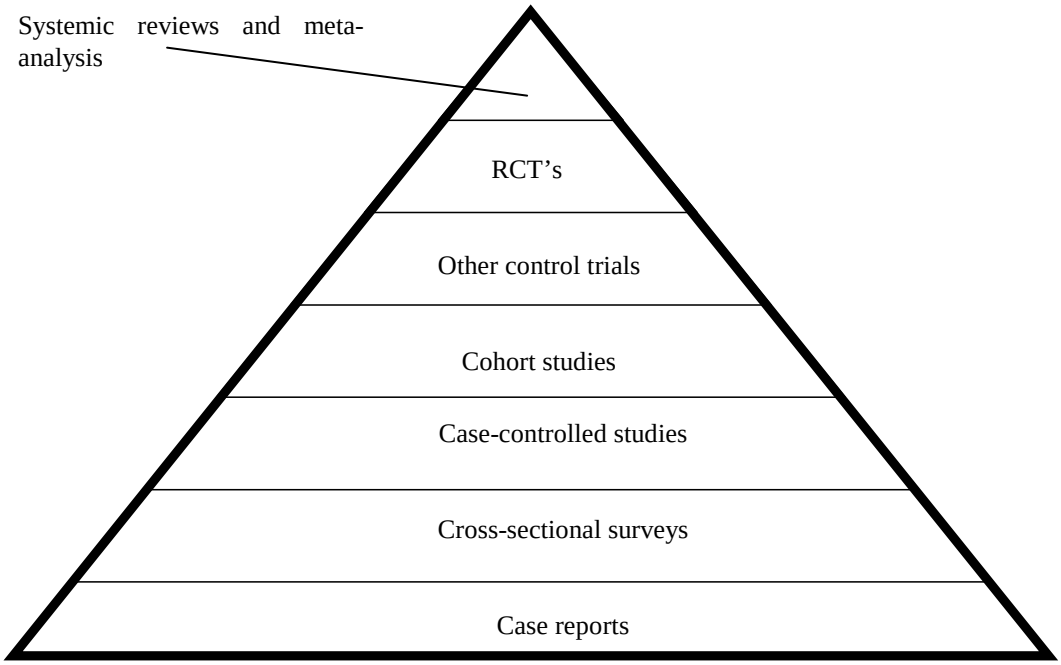
Criticisms of EBM

Criticism	Defence
Generalisations about populations not necessarily appropriate for individual patient	No one is advocating cookbook medicine
Co-morbidity in RCT's may be less than those of clinicians' patients	As above. Best evidence always useful but clinician must realise that it is not necessarily applicable to each individual patient.
Variability in RCT smaller than population so effect stands out. Over-estimation of intervention effect?	
Surrogate end points often used	
Expertise and clinical experience being devalued	Decisions on the application of EBM to the individual patient requires expertise and clinical experience
EBM has always been carried out	Reading papers has , but EBM has not

Hierarchies of evidence

The hierarchy must not be regarded as all-powerful. The results of a well conducted trial in an inferior hierarchical category, will be more valid than those of a poorly conducted one in a superior category.
There are several versions of the hierarchies including those of US Preventive Services Task Force, the NHMRC (Australia) and The Oxford Centre for EBM. The latter is the most comprehensive. Two are given below.

NHMRC (Aus) Levels of evidence (abridged)	
I	Systematic review
II	Properly designed RCT
III – 1	Pseudo – RCT
III – 2	Non-randomised cohort with concurrent control; case-controlled study
III – 3	Comparative study with historical control;
IV	Case series



Traditional hierarchy of evidence: Based on Trish Greenhalgh: How to read a paper. Blackwell Publishing.

The following hierarchy has a different function being a categorization of Risk vs benefit. This one produced by the US preventative service task force.

U.S. Preventive service task force Categories of recommendation	
Level A	Good scientific evidence that benefits substantially outweigh risks
Level B	At least fair scientific evidence that benefits substantially outweigh risks
Level C	At least fair scientific evidence that benefits exist but balance benefit / risk balance too close to make general recommendations
Level D	At least fair scientific evidence that risks outweigh benefit
Level I	Scientific evidence is lacking, poor quality or conflicting

STUDY DESIGN

1. Specify research objective	Occurrence of disease Hypothesis testing Understand disease causation Evaluate intervention
2.Specify target population	Inclusion / Exclusion criteria
3. Specify outcomes	Primary – should directly answer the research objective. Preferably *not* be a surrogate outcome Secondary – few as possible to avoid experiment-wide Type 1 error and temptation for 'data torture'.
4. Requirement for control group if intervention study	
5. Sample size estimation	Influences power and precision May require a pilot study. Pilot study should be descriptive with confidence intervals but not a hypothesis test.
6. Confounding	Confounding variable is one which is associated with both outcome and main risk factor independently. Control in two ways: Design – Observational study by matching, Randomisation in intervention trials Analysis – Stratification
7. Prevent bias (abbreviated)	Selection bias: Randomisation. Observer bias : Allocation concealment and blinding Response bias: If patient or observer or analyst knows allocation they can affect outcome. Avoid with blinding and allocation concealment Recall bias: Likelihood of side effects/ disease recall by patient will be influenced by knowledge of their treatment group. Avoid with blinding and allocation concealment. Withdrawals: Tends to underestimate treatment effect / side-effects. Use Intention to treat analysis and analyse all randomised patients
8. Data handling	Responsibility, confidentiality, double data entry, database
9. Statistical analysis plan	Outline the statistical analysis strategy in protocol. Avoid data dredging

CLINICAL DRUG TRIALS

Phase I	Administration to (usually) healthy human volunteers to determine the pharmacokinetics and toxicology of the drug.
Phase II	Specific clinical trials to determine pharmacodynamics, efficacy and safety.
Phase III	Large clinical studies to determine cost-benefits, risks etc
Phase IV	Continued surveillance once drug is in marketed

APPENDIX

KEY POINTS IN STATISTICS

TOPIC	KEY POINTS
The normal distribution	1) A population in which there is a trend, a 'normal' value, and in which *random* (chance) variation has caused a spread around that trend. 2) The spread is such that most values still cluster round the norm. Extreme variations exist but are rare. The random effect works equally above and below the norm. 3) The shape of the plot is, therefore, bell shaped, symmetrical and theoretically of infinite size 4) Because the spread has occurred through chance, the distribution is symmetrical and the mean = mode = median. 5) A large sample taken from a normally distributed population also has a normal distribution. 6) The larger the sample, the more likely it's mean will be close to the population mean 7) If multiple samples are taken from a normally distributed population, the plot of their means will also be normally distributed.
Standard deviation	A measure of the spread of individual values around the mean of a population or a sample.
Use of standard deviation	1. The SD gives an indication of the spread of values within a sample and, therefore, the reliability of the sample mean as an indication of a trend in a sample 2. The SD of a large sample is similar to that of its population. This fact is used in parametric tests. 3. Used to calculate the SEM
Standard error of the mean	An estimate of how the means of multiple samples would be spread around the population mean.
Use of standard error	1. Derived from the SD, the SEM also gives an indication of the spread of values within a sample but it is more commonly regarded as an indication of the proximity of the sample mean to its population mean. 2. Used in parametric tests comparing sample means.
The standard normal distribution	The basic template of the normal distribution where data are described in multiples of SD's or SEM's from the population mean.
Parametric testing	Tests based on the *parameters* of the normal distribution. The tests determine the probability of an effect being due to chance alone.
z-value	An expression in multiples of SD's or SEM's, of the distance between a point and the population mean. Used in the normal (z) test.
Null hypothesis	An hypothesis to be tested that states that any difference found has occurred through chance alone.

P value	Calculated from the results. The *P* value is the probability of an effect having occurred through chance alone if the Null Hypothesis is true.
Alpha	The significance level. Decided at the design stage. It is the limit at which *P* will be regarded as being too large for statistical significance.
Alpha (Type 1) error	Chance of there being no difference when you say there is one. False positive rate.
Beta (Type II) error	Chance of there being a difference when you say there is none. False negative rate.
95% Confidence interval (parametric test)	The range above and below the sample mean within which you predict with 95% confidence that the true value (population mean) lies.
Confidence interval (general equation)	CI = estimation ± a multiple of SE where multiple depends on assumed distribution and level of confidence. CI's used when reporting OR's, RR's and many other statistics
T-test	A parametric test of means where the samples are too small to use the normal test.
One sample t-test	T-test where the mean of a sample is compared with a number
Two-sample t-test	T-test where two means are compared.
Paired t-test	T-test comparing the results of a single sample *before and after* treatment.
One–tailed	When there is only one direction that one group can vary from another, you only have to look in one tail for a significant result.
Two-tailed	If you don't know *for certain* which way the test result will vary compared with another value, you must look in both tails
Bonferroni's correction factor	Correction factor used to reduce alpha error when multiple t-test comparisons are used.
Analysis of variance	A method used to compare three or more parametric samples. The between groups variance must outweigh the within groups variance.
Non-parametric testing	Any test which is not parametric! Based on ranking when data is continuous
Regression	The drawing of the line that best describes the relationship between two continuous variables. Equation: $y = a + bx$
Correlation	The determination of the likelihood that the above relationship does exist.
The power of a study	The probability of a study being able to demonstrate a

	difference when a difference does exist. 1-false negative rate.
Information required when calculating sample size a) Comparing the numbers in two samples	Alpha 1-Power (ß) The predicted effect-size. Predicted standard deviations of samples
Information required when calculating sample size b) Comparing a proportions	Alpha 1-Power (ß) The proportion you're looking for Null hypothesis proportion
Chi square	Compares the frequency of a binary event within two or more groups Uses a contingency table Compares observed with expected values
Relative risk	The ratio of the incidences of an event with and without exposure
Odds ratio	The ratio of the odds of an event with and without exposure
Number needed to treat	The number of patients needed to be treated to produce one success or survivor
Sensitivity	The proportion of *disease* which is correctly identified. Highly sensitive test has very few false negatives
Specificity	The proportion of *'no-disease'* which is correctly identified. Few false positives.
Positive predictive value	The proportion of a test's positive *result*s which true positives.
Negative predictive value	The proportion of a test's negative *results* which are true negatives.
Bland-Altman plot	A plot used to assess agreement between two measuring techniques
Receiver operating characteristic (ROC) curve	A graph used to illustrate the trade off between sensitivity and specificity in tests that produce results on a numerical scale rather than as an absolute positive or negative.
Systematic review	A highly structured process in which an attempt is made to answer a specific clinical question by collating and analysing the data from <u>all</u> relevant trials.
Meta-analysis	The mathematical process of combining data from studies using similar treatments in a systematic manner.

GLOSSARY

a	intercept of y axis
AR	attributable risk
α	significance level
ANOVA	analysis of variance
b	regression coefficient
c.i.	confidence interval
d.f.	degrees of freedom
df	degrees of freedom
E	expected frequency
F	statistic of analysis of variance
H	statistic of Kruskal-Wallis test
H_o	null hypothesis
MS	mean square
μ	population mean
NNT	Need to treat
N	population size
NPV	negative predictive value
n	sample size
O	observed frequency
OR	odds ratio
P	probability
p	proportion of outcomes in a sample
PPV	positive predictive value
π	population proportion
π_o	null hypothesis proportion
q	statistic of the Student-Newman-Keuls test
q'	statistic of Dunnett's test
r	Pearson's correlation coefficient
r_s	Spearman's rank correlation
r^2	coefficient of determination
r x c	rows x columns table
RR	relative risk
σ	population standard deviation
σ^2	population variance
s	sample standard deviation
s	standard error of the regression
s.d. SD	standard deviation
s^2	sample variance
SS	sum of squares
s.e., SE, SEM	standard error of the mean
SND	standard normal deviate
t	statistic of the *t* test
U	statistic of the Mann-Whitney *U* test
x	individual value; explanatory value in linear regression
x_i	individual value
$\bar{x}$	sample mean
y	dependant value in linear regression
y_i	individual / actual value in linear regression
$\hat{y}$	predicted value in linear regression
$\bar{y}$	mean value of y in linear regression
z	standard normal deviate
!	factorial of a number (all integers from number down to 1 multiplied together)

BIBLIOGRAPHY

1. Asbury AJ. Statistics and clinical trials, in Nimmo WS, Rowbotham DJ, Smith G. Anaesthesia. Blackwell Scientific Publications: Oxford 1994. 606 - 22.

2. Burley DM. Drug trials and statistical validation. Scientific Foundations of Anaesthesia. 584 - 599

3. Fisher DM. Statistics in anesthesia, in Miller RD ed. Anesthesia. Churchill Livingstone. New York. 3rd edition. 1990. 685 - 712.

4. Gore SM, Jones IG, Rytter EC. Misuse of statistical methods: critical assessment of articles in BMJ from January to March 1976. BMJ.1977;1: 85-7.

5. Ingelfinger JA, Mosteller F, Thibodeau LA, Ware JH. Biostatistics in clinical medicine. 3rd ed. McGraw-Hill. New York. 1994.

6. Katz MA. A probability graph describing the predictive value of a highly sensitive diagnostic test. NEJM 1974; 291: 1115-7.

7. Kirkwood BR. Essentials of medical statistics. Blackwell Scientific Publications. Oxford. 1988.

8. Myles PS, Williams NJ, Powell J. Predicting outcome in anaesthesia: understanding statistical methods. Anesthesia and Intensive Care. 1994; 22:447-53.

9. Altman DG. Confidence intervals for the number needed to treat. BMJ 1998;317:1309-12.

10. Oakley Davis HT, Crombie IK. When can odds ratios mislead? BMJ 1998; 316: 989-91.

11. Herbison P. Problems with meta-analysis. NZ Med J. 1999; 112: 38-41.

12. Bland and Altman series. BMJ

13. Myles PS and Gin T. Statistical methods for anaesthesia and intensive care. Butterworth Heineman Oxford 2000

14. Users' guides to the medical literature: VI. How to use an overview. Oxman, AndrewF; Cook, Deborah J, Gordon H. JAMA; Nov 2,1994;272,17

15. Emmett, RS. Cyna, AM. Andrew, M. Simmons, SW. Techniques for preventing hypotension during spinal anaesthesia for caesarean section. Cochrane Pregnancy and Childbirth Group Cochrane Database of Systematic Reviews. 2, 2005.

16. Cochrane handbook for systematic reviews of interventions. 4.2.4. March 2005. The Cochrane Collaboration

17. Petrie A, Sabin C. Medical statistics at a glance. Blackwell Publishing. 2005

INDEX

APPENDIX

Formulae

The binomial equation

$$\text{Prob}(rA's) = \frac{n!}{r!(n-r)!}\pi^r(1-\pi)^{n-r}$$

n = sample size
π = population proportion
r = number of the outcome of interest
A = the outcome

Poisson formula

$$P(r) = \frac{\mu^r}{r!}e^{-\mu}$$

P(r) = probability of *r* occurrences
μ = mean number of occurrences per unit time

Calculation of SE in a two sample normal test

$$SE\ (\text{large samples}) = \sqrt{\frac{s_1^{\,2}}{n_1}} + \sqrt{\frac{s_2^{\,2}}{n_2}}$$

Calculation of SE in a two sample normal test

$$SE\ (\text{population } \sigma \text{ known}) = \sqrt{\frac{\sigma_1^{\,2}}{n_1}} + \sqrt{\frac{\sigma_2^{\,2}}{n_2}}$$

Calculation of the SE in a two sample *t*-test

$$SE = \sqrt{\frac{s_1^{\,2}(n_1-1) + s_2^{\,2}(n_2-1))}{n_1+n_2-2}}.\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

Spearman's rank correlation equation

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Risk scores

Incidence rates

Relative risk given per year exposed:
= *incidence* per year exposed / *incidence* per year non-exposed

Attributable risk

Indicator of <u>magnitude</u> of excess risk in absolute terms. *eg how many <u>extra</u> cases of cancer per year were due to smoking?*
= incidence among exposed - incidence among non-exposed
= a / a + b - c / c + d per unit population

Proportional attributable risk:

The proportion of disease among the exposed which is caused by the exposure (after the proportion which occurs without exposure has been taken into account) *eg of all the cases of lung cancer among smokers, what proportion can be attributed to smoking ?*

= AR / incidence among exposed = RR - 1/ RR

Population attributable risk

If the prevalence of the disease in the population is low, an exposure with a high relative risk may not actually cause many deaths. The PAR takes this into consideration by relating the overall incidence with the incidence among non-exposed.
= overall incidence - incidence among non-exposed

Matched study design

An example of a matched study

Is DVT associated with oral contraceptive use? A retrospective case-control study is carried out. Patients who have had DVT are found and matched with patients of with the same confounders but without DVT. Their exposure to the oral contraceptive pill is then determined.

Cases DVT		Controls No DVT
1	1 has same age, ASA, BMI as 11	11
2	2 has same age, ASA, BMI as 12	12
3	3 has same age, ASA, BMI as 13	13
4	4 has same age, ASA, BMI as 14	14
5	Etc	15
6	Etc	16
7	Etc	17
8	Etc	18
9	Etc	19
10	Etc	20

10 cases of DVT are found retrospectively and each is matched with a patient with the same confounders but with no DVT.

Cases DVT		Controls No DVT
1 and 11 were both on OC pill		
2		12
3		13
4 and 14 were both not on OC pill		
5		15
6		16
7		17
8 and 18 were both on OC pill		
9		19
10		20

Those pairs who both have the exposure or who both do not have the exposure are removed. Eg both on OC or both not on OC

Cases DVT		Controls No DVT
1 and 11 were both on OC pill		
(2)		12
3		(13)
4 and 14 were both not on OC pill		
(5)		15
6		(16)
(7)		17
8 and 18 were both on OC pill		
(9)		19
(10)		20

Pick out the cases and the controls who had used the OC pill. ( )

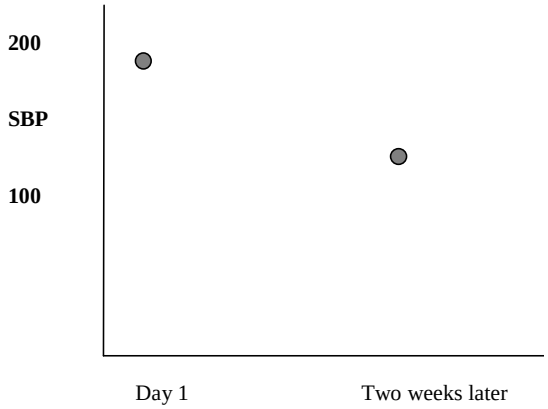
An odds ratio is then calculated to assess likelihood of risk associated with OC and DVT.

OR = ratio of discordant pairs =

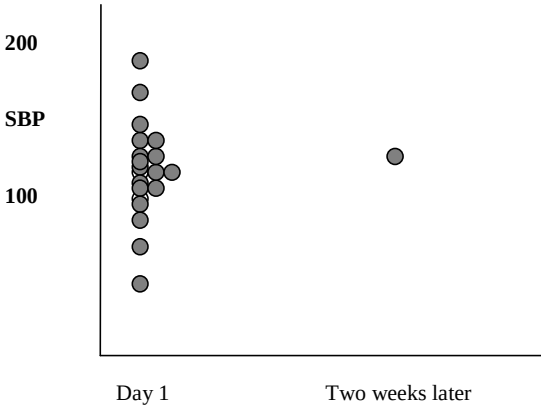
$$\frac{\text{number pairs where there is DVT + OC}}{\text{number pairs where there is no DVT + OC}}$$

= 5 / 2

Regression to the mean



An isolated recording on day 1 reveals an extremely high SBP. After two weeks of treatment there appears to be a dramatic improvement. But is this secondary to the treatment?



If 50 recordings had been taken on day 1, there would be a normal distribution of results based on random variation. His SBP has, in fact, not changed. It appears to have fallen because extreme variations are likely to *regress towards the mean*

Arithmetical demonstration of the effect of prevalence on predictive ability of a test:

A test of known sensitivity and specificity is used to predict an outcome in Population A. If it is then used in population B where the prevalence of the outcome is lower, prediction of a positive outcome becomes more difficult but prediction of a negative outcome is easier ie the PPV of the test becomes weaker and the NPV improves.

eg

Example 1. Prevalence of disease = **11 %** (100 / 991)

	yes	no		
+ve	29	80	109	Sensitivity = 29 /100 = 29 % (False negative rate = 71 %)
-ve	71	811	882	Specificity = 811 / 891 = 91 % (False positives = 9 %)
Totals	100	891	991	PPV test = 29 / 109 = **28 %**, NPV test = 811 / 882 = **92 %**

Example 2. Prevalence of disease = **4.8 %** (48 / 991)

	yes	no		
+ve	14↓	87↑	101	Sensitivity = 14 / 48 = 29 % (False negative rate = 71 %)
-ve	34↓	856↑	890	Specificity = 856 / 943 = 91 % (False positives = 9 %)
Totals	48	943	991	PPV test = 14 / 101 = **14 %**, NPV test = 856 / 890 = **96%**

(Based on an article by Myles PS, Williams NJ, Powell J. Predicting outcome in anaesthesia: understanding statistical methods. Anesthesia and Intensive Care. 1994; 22:447-53.)

Corrections
Introduction
Distribution of Sample means plot
Figure numbering