A SHORT COURSE ON BIOSTATISTICS

Prepared by Nikolai Bogduk
Newcastle Bone and Joint Institute
University of Newcastle

for the subject – Critical Reasoning
in the course for the degree of Master of Pain Medicine

INTRODUCTION

If you note the name of the subject, it reads – Critical Reasoning. It does not read – Statistics.

The purpose of the subject is to equip you with the tools and skills to evaluate information. It is designed to make you a good consumer of information, and
a responder to information. But it is not intended to have you be able, necessarily, to generate information.

Consequently, you are not required to be able to calculate complex statistics or to apply inferential statistical tests to raw data. However, in order to be a good consumer you must have a certain proficiency in biostatistics. This short course is designed to equip you with that proficiency.

The course does not derive or explain the mathematical basis of the more complicated, statistical tests. For that, either you can take for granted that the statisticians have worked it out, or you can consult a textbook of statistics.

The course summarises the more commonly used tests, and provides a set of operational rules. The rules dictate when certain tests should be used, and what they mean. The objective is to arm you, as a consumer, with an insight that allows you to be discerning and demanding. You will be able to recognize when a correct test has been applied, and when an inappropriate one has been used. This protects you, as a consumer, from being fooled by statistics. It also allows you to be a cynical participant at conferences, should you wish to do so, when speakers present wrongful data. In addition to protecting yourself, you can protect others in the audience who are less skilled in statistics, less erudite, less eloquent, and more gullible.

LESSON 1. DESCRIPTIVE STATISTICS

The objective of clinical studies is to determine some sort of rule that can be used in the future to render clinical practice more accurate or more efficient. That rule might apply for a diagnostic test or for a therapeutic procedure. The reason for determining the rule is to be able to apply it to anyone and everyone for whom the rule is relevant.

If a rule is to be useful, it has to apply to everyone and anyone (within reason). Pursuing rules is pointless if the rules are to be different to each and every patient. The whole point of defining rules is so that they can be applied in general.

In statistical terms, all individuals to whom a rule might apply are referred to as the **population**. To mathematicians, this means every possible person in the universe.

When an investigator seeks a rule that will apply to everyone, they do not have access to everyone; they cannot study the entire population. When conducting an experiment the investigator has access only to a **sample** of the population.

For example, an investigator, interested in showing how well a treatment works, is seeking to establish a rule that will apply to anyone and everyone with the condition in question. However, when conducting an experiment, the investigator can evaluate the treatment in only a sample of patients.

The extent to which this sample is representative of the population governs how readily rules derived from the sample can be generalized to apply to the population. Statistical tests can be applied to determine how representative the sample is of the population for which the rule is intended.

A preliminary step, however, is to describe the data harvested. That description yields variables that can used to test the data.

Data exhibit two features: nature and distribution. Both features predicate what sorts of tests are applicable to the data. Not all data can be treated in the same way.

Nature

By nature, data can continuous or categorical.

Continuous data are data in which any individual within a sample can express any value along a continuum. For mathematicians, this means that the difference between any two values can infinitely large, or infinitesimally small, i.e. anything.

Thus, if the variable is systolic blood pressure, one individual might have a value of 120, another might have a value of 130. But others might have values of 121 and 122. It is even possible that still others might have values of 120.5 and 120.6, or 120.55 and 120.56; and so on. For mathematicians, there is no limit to the possible differences. Any individual might assume any (exact) value.

For clinicians, tiny differences are not appropriate, and will ultimately be limited by the resolution of the device used to measure the variable. Nevertheless, in principle, the variable can effectively assume any value.

Categorical data differ in that it is not possible for an individual to assume any, infinitesimally different value, (even if you did have devices to resolve such differences). Someone, somehow, makes a decision to fit variables into distinct categories.

One can have two categories, such as good – bad, yes – no, or present – absent. Under such conditions, there are no grey zones; there are no other options. Either something is present, or it is not present; it either it is good or it is bad.

One can have three categories, such as good, bad, and indifferent; or yes, no, and maybe. One can have multiple categories, such as poor, fair, reasonable, good, very good; or nil, trivial, mild, moderate, severe, excruciating.

Whatever, their number, the categories define a spectrum of possible values, but not a continuum, in the sense that values cannot fall between categories.

Distribution

Regardless of the nature of the data, the values exhibited by individuals within a sample will have a distribution, defined as the number of individuals that exhibit particular values. The distribution differs according to the nature of the variable and the conditions under which is measured.

One type of distribution is the **normal** distribution, also known as the Gaussian distribution. When graphed, this distribution assumes a bell-shaped curve. In a mathematically idealized normal distribution, the data can extend from minus infinity to plus infinity (Fig. 1).

More practically, the distribution lies between two finite values (Fig. 2), but the essence of the distribution is that many and most individuals in the sample express values that are in the middle of the range, and progressively fewer individuals express values that are extremely low or extremely high.

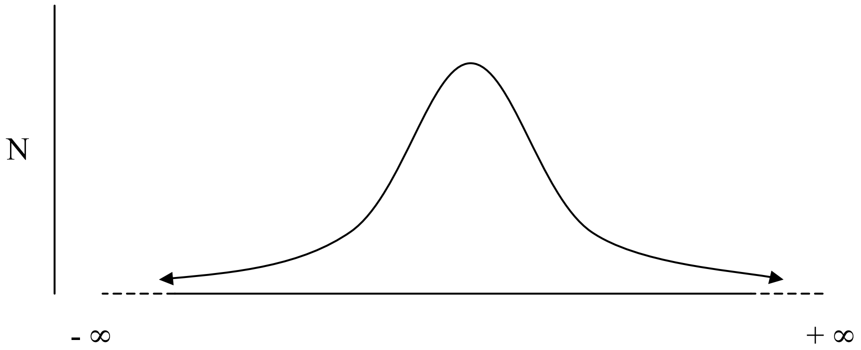


Figure 1. The ideal normal distribution, with possible values extending from minus infinity to plus infinity.

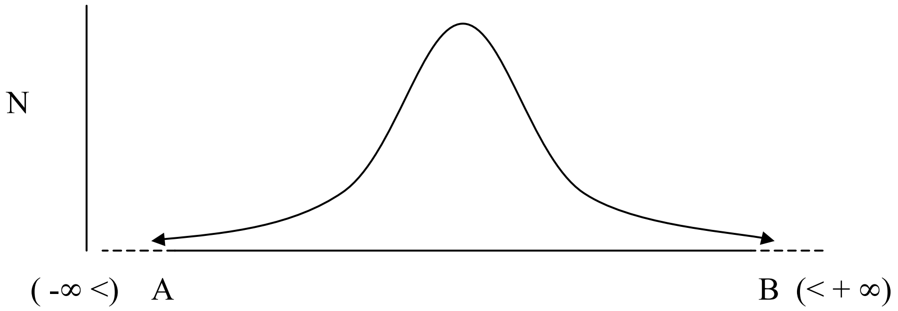


Figure 2. A practical normal distribution, with possible values extending between some value (A) greater than minus infinity, and some value (B) less than plus infinity.

Such distributions can be summarized and described by certain statistics (Fig. 3). The central value of the distribution is the **mean** value (μ), and the spread of the distribution is the **standard deviation** (s). The standard deviation describes how widely the values are spread. Indeed, the nature and magnitude of the standard deviation is such that 68% of the individuals in the sample express values within one standard deviation either side of the mean (Fig. 4), and 95.4% of individuals express values within two standard deviations of the mean (Fig. 5).

It is also true that 95% of the values exhibited by a sample fall within 1.96 standard deviations of the mean (Fig. 6). [I raise this observation because we will repeatedly encounter the magic numbers: 95% and 1.96.]

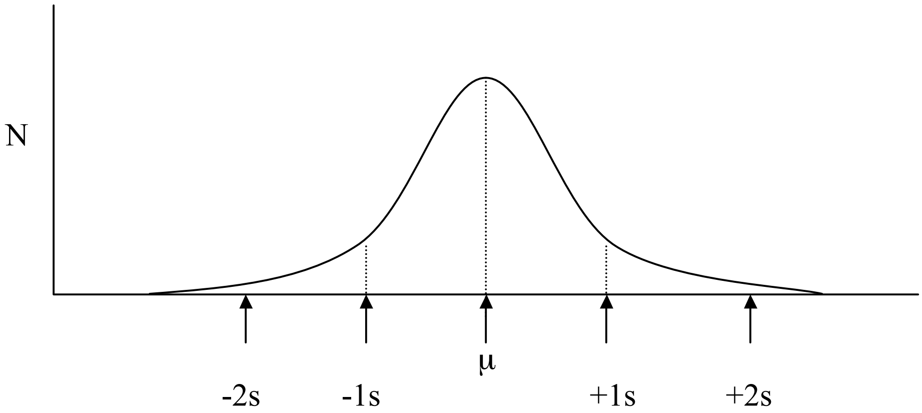


Figure 3. The parameters of a normal distribution. μ: mean. s: standard deviation.

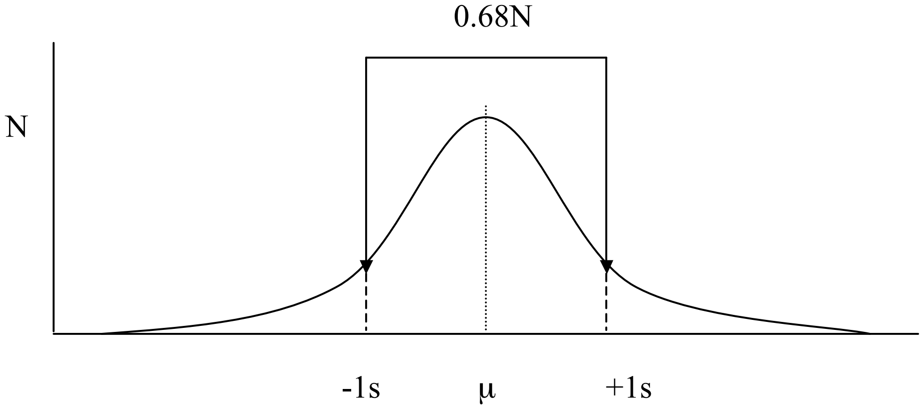


Figure 4. In a normal distribution, 68% of the individuals express scores that fall within one standard deviation of the mean.

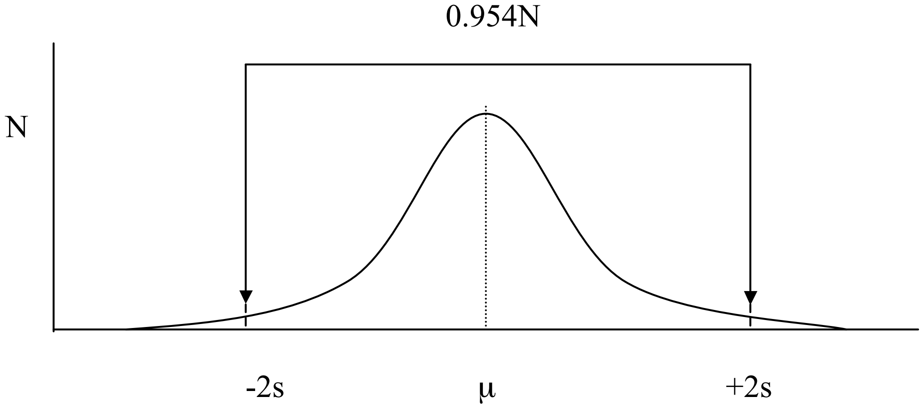


Figure 5. In a normal distribution, 95.4% of the individuals express values that fall within two standard deviations of the mean.

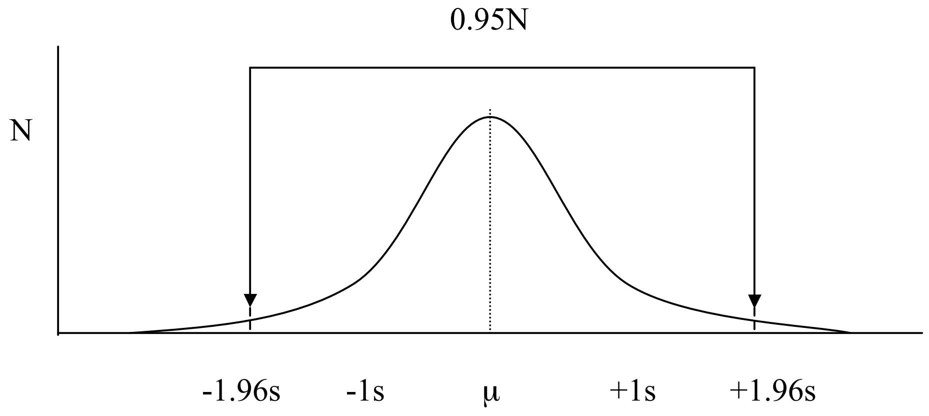


Figure 6. In a normal distribution, 95% of the individuals express scores that fall within 1.96 standard deviations of the mean.

These various numbers are derived from the mathematical equation for a Gaussian distribution, which is:

$$\text{p (x)} = \frac{1}{\sqrt{2\pi s}} e^{-(x-\mu)^2/(2s^2)}$$

By integrating this equation, one can calculate the area under the curve between any two points, which in turn indicates the probability of x falling within that range. Upon doing so, it transpires that 68% of the total area under the curve lies between x – s and x + s; 95.4% of the area lies between x – 2s and x + 2s; and 95% lies between x – 1.96s and x + 1.96s.

For those fascinated by mathematics, it can be shown that the second differential of p(x), i.e. d^2p/dx^2, is zero when $x = \mu \pm s$. This occurs at the point of inflection of the bell-curve, i.e. where is stops coming down and starts to curve outwards, which is at the junction of its external convexity and its external concavity (Fig. 3). Although this may sound esoteric, it comes in handy. If you know the mean and the standard deviation, you can readily sketch the distribution, for the mean marks the top of the bell-curve, and the standard deviation marks where it starts to flare.

Another statistic that is commonly used in statistical calculations is the **variance**. It describes not the spread of the values, but by how much they vary (from the mean). It turns out that mathematically, the variance is the square of the standard deviation (s^2).

Data that assume a normal distribution as said to be **parametric**, which means that the data are symmetrically distributed about a central, mean value. Other distributions do not exhibit this feature. They are said to be **non-parametric**.

Examples of non-parametric distributions are depicted in figures 7 and 8. The first is a flat distribution. The second is a skewed distribution.

The flat distribution (Fig. 7) has values spread across a range, just like a normal distribution, but there is not a predominance of individuals expressing central values. Just as many individuals express high and low values as express middle values. Such a distribution cannot be held to be normal or parametric unless and until more individuals express middle values, and the curve of the distribution takes on the central hump of the bell shape (Fig. 1).

The skewed distribution has most of the individuals in a sample expressing low values, at the bottom end of a range, and a few individuals stretching out into the high end of the range (Fig. 8).

The statistics – mean and standard deviation do not apply to such distributions because the data are not symmetrically distributed about a central core of values. Instead, the statistics – median and interquartile range, apply (Figs, 9 and 10).

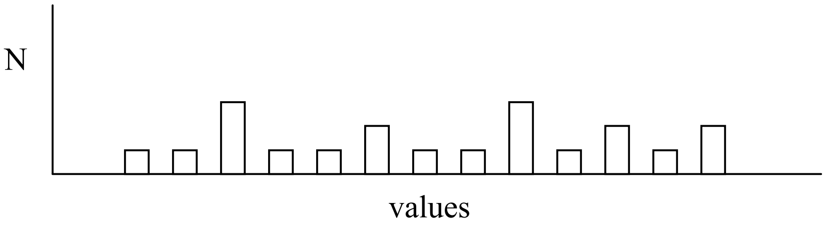


Figure 7. A non-parametric, flat distribution.

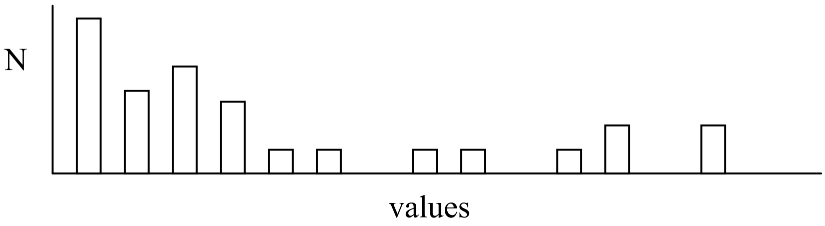


Figure 8. A non-parametric, skewed distribution.

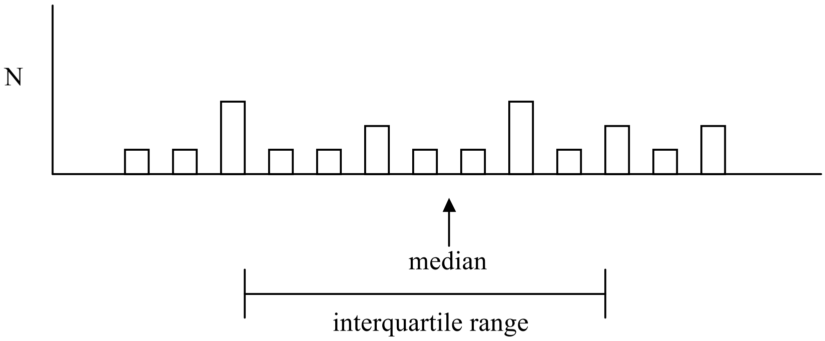


Figure 9. The median value and interquartile range of flat distribution..

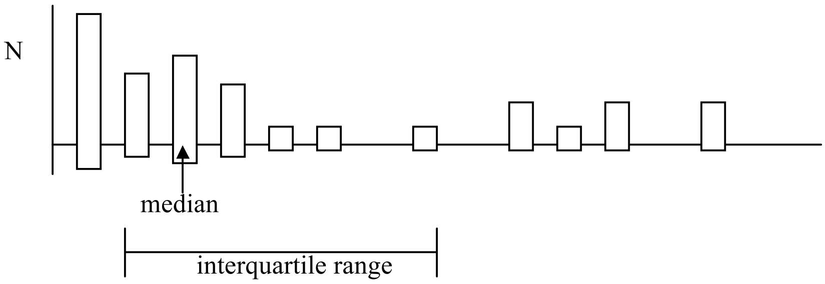


Figure 10. The median value and interquartile range of a skewed distribution.

The median is the "middle" value expressed by the sample, in that just as many individuals express values equal to or less than the median value as express values equal to or greater than the median value. It does not matter how far the values spread. All that is required that equal **numbers** of individuals (i.e. values) are represented above and below the median.

The interquartile range is the values between which 50% of the individuals are represented, below which 25% are represented, and above which the remaining 25% are represented. It provides an indication of where "most" of the values lie.

Take-Home Messages

Computers and calculators will do what you tell them. Consequently, if you enter a stack of numbers and ask for the mean and standard deviation, the computer will obediently provide them for you. That, however, does not guarantee that calculating the mean and standard deviation was the correct thing to do.

The rule is that ***mean and standard deviation apply only to normally distributed, parametric data,*** *(or data that at least resemble a normal distribution). Mean and standard deviation* ***do not apply to non-parametric data****. If for no other reason, the standard deviation cannot apply to skewed distribution because the distribution is not symmetrical; the data do not spread beyond the peak of the skew; yet the calculated standard deviation implies (wrongly) that it does.*

These rules are more than just statistical pedantry. It seems that some investigators in the past, who were studying the efficacy of analgesics, simply entered their raw data; asked for a mean and standard deviation; and got them. Subsequently, they used statistical tests to compare data in order to show that the analgesics were effective. When others reviewed the raw data they found that they were not normally distributed. Therefore, the use of means and standard deviations was inappropriate. Moreover, when the data were re-described using non-parametric statistics, and tested using non-parametric tests, the evidence of efficacy disappeared. So, ***wrong conclusions can be drawn if data are wrongly described****.*

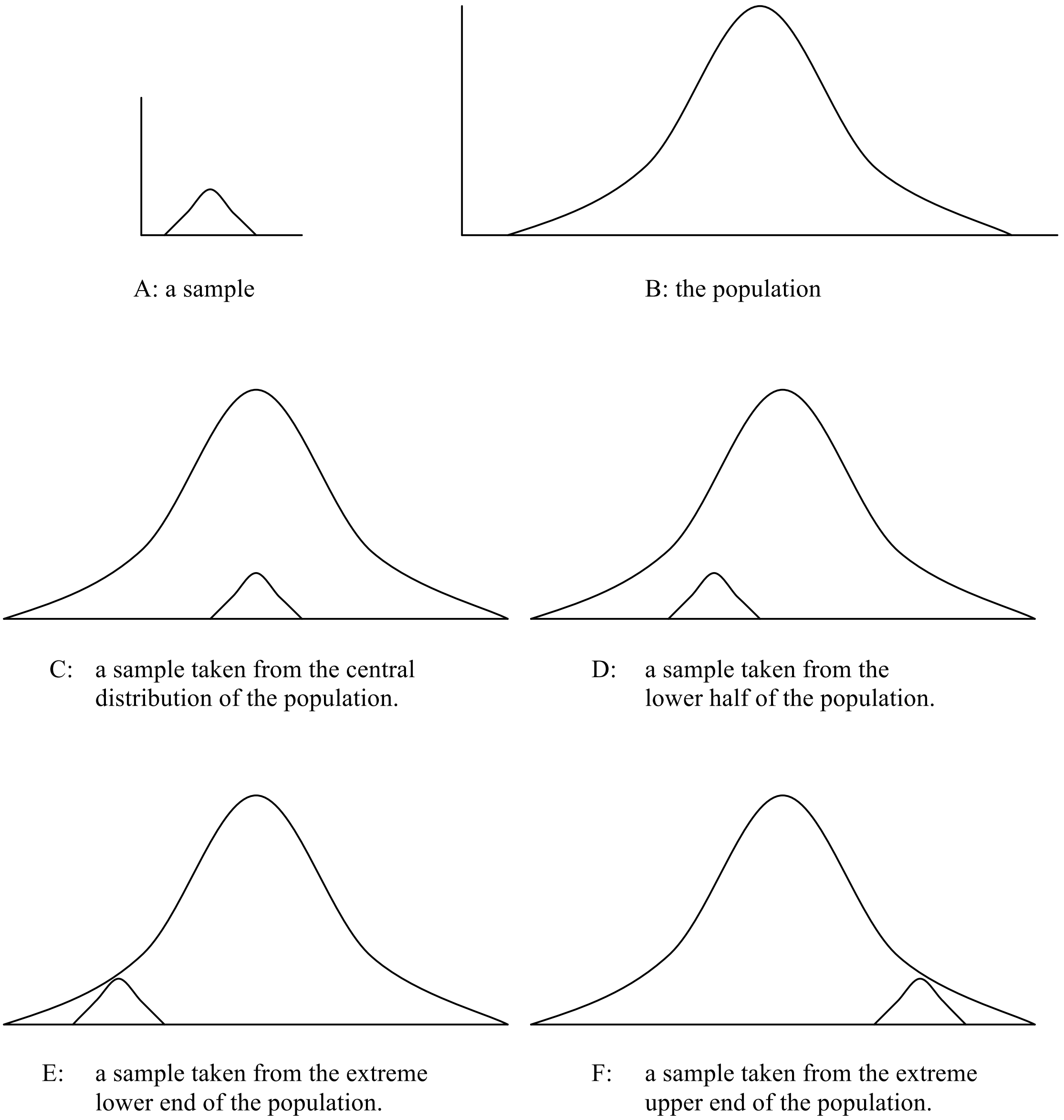
A simple way to check if the data are normally distributed is to ***plot them and read the graph****. Does it look like a bell curve? If so, the mean and standard deviation are appropriate. If not, the median and interquartile range apply. Be perceptive in this regard; for if you are not, you could get fooled. Don't fall for the three-card trick.*

LESSON 2. TESTING THE SAMPLE

Because an investigator does not have access to an entire population, they are obliged to use a sample when investigating a phenomenon. There is no guarantee, however, that the sample that they select is necessarily representative of the population. Therefore, any rule that they devise, based on the sample, may not be applicable to the population, in general.

Even though it is taken from the correct population, any sample may represent the more central values of the population or more extreme values (Fig. 11).

The laws of Statistics dictate that any set of multiple samples will themselves be normally distributed within the range expressed by the population. Thus, if you or another investigator were to repeat the same experiment using new and different samples, some of the samples would come from the lower end of the population, others would come from the upper end, and still others would come from the middle of the population (Fig 12).



A: a sample

B: the population

C: a sample taken from the central distribution of the population.

D: a sample taken from the lower half of the population.

E: a sample taken from the extreme lower end of the population.

F: a sample taken from the extreme upper end of the population.

Figure 11. Examples of how samples might relate to the population.

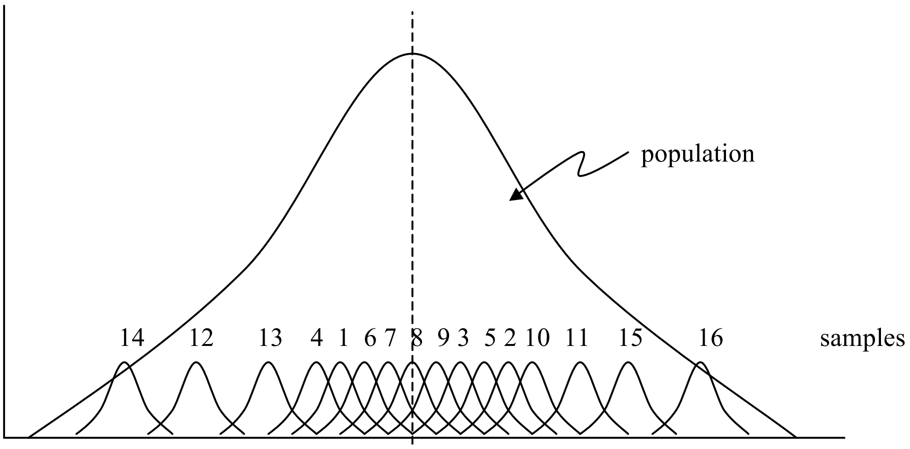


Figure 12. A representation of how 16 different samples might relate to the population.

It also transpires that most samples will be taken from the central regions of the population; few will be drawn from the extremes. Expressed in another way, when you take a single sample, you do not know if that sample is accurately representative of the population or if it is an extreme sample; but the chances are high that it comes from the central regions of the population, and the chances are low (but not nil) that it comes from an extreme region.

Consequently, no sample can be guaranteed to be truly representative of the population, but certain operations can be performed to determine just how representative it might be.

The mean of a sample might not coincide with the true mean of a population. However, if the mean (μ), variance (s^2), and size (n) of a sample are known, it can be estimated where, in relation to the observed mean, the true mean might be. The statistic that can be calculated is the **standard error of the mean (sem)**, where:

$$\text{sem} = \sqrt{\frac{s^2}{n}}$$

When the sample in question is the distribution of possible values of the true mean, the sem is homologous to the standard deviation. The range: $\mu \pm$ sem, indicates where the true mean could lie in relation to the sample mean, with 68% confidence. In other words, there is a 68% chance that the true mean lies within one standard error of the sample mean. The latter also means that there is a 32% chance that the true mean lies somewhere beyond one standard error of the sample mean.

One can also calculate the **95% confidence intervals** of the mean. Mathematically the 95% confidence intervals are:

$$\mu \pm 1.96 \text{ sem.}$$

This measure establishes the range in which one can expect, with 95% confidence, where the true mean lies. It also says that there is a 5% chance that the true mean lies somewhere outside this range, i.e. a 2.5% chance that it is less than μ - 1.96 sem, and a 2.5% chance that it is greater than μ + 1.96 sem.

Note how the sem is a function of the sample size. As n is made larger, the sem gets smaller. Therefore, the larger the sample size, the more confident one can become that the observed mean is representative of the true mean. Conversely, the smaller the sample size, the larger the sem, and the further the true mean may be from the observed mean; and the less confident you can be that the observed mean is representative of the true mean.

Take-Home Messages

The sem is a descriptive statistic that pertains expressly to how well an observed mean relates to the true mean. In operational terms, it reveals how well a mean derived from a sample reflects the true mean.

Textbooks or journal articles might publish data, such as normative data on range of movements. They might specify the mean range of movement. That mean was derived from a sample that the authors studied. You, however, will not see patients from that sample. You are dealing with the population. If you were to study a sample of your own patients, you might not encounter the same mean range of movement as the published mean. Nor should you be surprised if you don't. What you should note is the published sem. If the published data are correct, and if they do apply to your patients, the mean that you observe should be within ± 1.96 standard errors of the published mean. There is a 95% chance that this will be the case. There is, however, a 5% chance that your mean will not fall within that range. Should that be the case you have grounds to consider whether the published data are not applicable to your population, or you have collected an unusual sample.

In essence, when you practise, you are repeating observational experiments that others have already performed. Your patients might constitute a sample of the population that is slightly different from the one reported in the literature. The sem tells you how different you are allowed to find your sample to be.

In a totally different context, watch out. Undisciplined authors sometimes use the sem to indicate (falsely) the spread of their data. They do so because the sem is conspicuously smaller than the standard deviation. Consequently, by using sem for error bars instead of sd, they make their graphs look better. If what you are interested in is the spread of the data, and not how accurate the mean is, you are entitled to object; for the author, or speaker, is withholding information from you (perhaps to try to fool you). Error bars come in sd's not in sem's. Don't be fooled.

LESSON 3. DRAWING INFERENCES USING STATISTICAL TESTS

Statistics are used to determine relationships between samples, in order to establish rules that might be applied in general, to populations. In broad terms, the types of relationships most often sought are differences and correlations. The statistical tests that can be used differ according to what relationship is being sought, and the nature of the data.

DIFFERENCES

In Pain Medicine, differences are typically sought between two samples in order to test if a treatment is effective. One sample is treated with the treatment in question, while the other sample is treated with a control or comparison treatment. The variable measured is typically pain, but may be other variables such as physical disability or psychological distress. The question that arises is: are (the results in) the two samples different?

Parametric Data

If the variable is normally distributed in each of the two samples to be compared, their distributions can be plotted. This allows for a visual inspection of the data to determine, in the first instance, if there appears to any difference (Fig. 13).

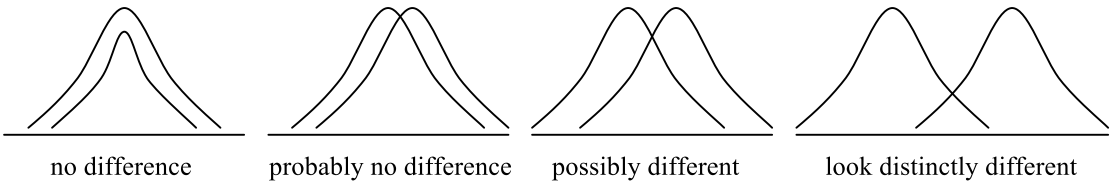


Figure 13. Comparing two samples visually, to assess if two samples look different.

The laws of Statistics reveal that even if two samples are actually drawn from the same population, there is a certain probability that the two samples will nevertheless look different. However, the greater the apparent difference, the less the probability that samples are from the same population.

This probability has a distribution that is not exactly a normal distribution. It resembles a normal distribution but is slightly skewed. It is called the t-distribution.

In formal terms, the t-distribution describes the probability that a particular difference between two parametric samples has arisen by chance and, therefore, that the apparent difference is not due to some external effect (such as the effect of a treatment).

A test based on the t-distribution is the **t-test**. This test takes into account the variance in each of the two samples and their sizes, and determines the probability that the difference between the two means has arisen simply by chance alone (Fig. 14).

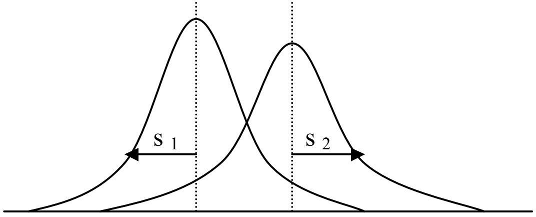


Figure 14. The parameters pertinent to comparing two parametric samples.

The t-test does not prove that the two samples are different. It establishes only the probability that the difference arose by chance alone. By convention, a difference is considered unlikely to have occurred by chance alone if the t-test reveals that the probability (P-value) of the difference is less than 0.05.

Note that, conversely, this P-value means that there is a 5% chance that there is no real difference: that the two samples happen to be freak subsets of the same population, i.e. they belong to the same family and, therefore, are not really different, despite appearances.

Another version of this converse interpretation is that there is a 5% chance that an investigator will conduct an experiment and will find a difference, but this difference is a fluke. In other words, 1 in 20 repetitions of the same experiment will, by chance alone, yield a result that falsely looks as if there is a difference between the two samples. For this reason, skeptical consumers require that experiments with dramatic new results should be repeated. It cannot be guaranteed that the first experiment in a field did not happen to be that 1 in 20 fluke result.

Continuous Non-Parametric Data

The t-test cannot be applied to non-parametric data, because the data are not normally distributed. Therefore, the means and standard deviations required for the t-test are not available; or if artificially calculated, they do not faithfully reflect the true distribution of the data.

For non-parametric data, rank tests, such as the Wilcoxon test or the Mann-Whitney test, can be used. In essence, these tests line up the data from two samples, opposite one another, and ask if the values in one sample are consistently larger than those in the other sample.

Figure 15 illustrates two sets of data between which there is no difference. Each value in sample A has a corresponding value in sample B to which it is equal.

Figure 16 illustrates two samples in which there are some differences between the values, but there are just as many values in sample A that are larger than values in sample B, as there are values in sample B that are larger than values in sample A. Overall there is no nett difference between the samples.

In figure 17, the samples are distinctly different. The two samples share some values that are equal, but for every value in sample B there is a value in sample A that is larger. The values in sample A are consistently larger, and the sample A is distinctly different from sample B.

sample A	10	11	11	12	13	13	13	14	14	15	16	17
sample B	10	11	11	12	13	13	13	14	14	15	16	17
differences	0	0	0	0	0	0	0	0	0	0	0	0

Figure 15. Two samples between which there are no differences.

sample A	11	11	12	12	13	13	13	14	14	15	16	17
sample B	10	11	11	12	13	13	13	14	15	16	16	17
differences	+1	+1	0	0	0	0	0	0	-1	-1	0	0

Figure 16. Two samples between which, overall, there are no differences.

sample A	11	12	12	12	13	13	14	14	15	15	17	17
sample B	10	10	10	11	11	12	12	13	13	14	14	14
differences	+1	+2	+2	+1	+2	+1	+2	+1	+2	+1	+3	+3

Figure 17. Two samples between which there are consistent differences.

The Mann-Whitney test calculates the probability of encountering a certain magnitude of differences between two samples when the samples actually belong to the same population. As with parametric tests, the convention is that the differences are considered unlikely to have occurred by chance alone if the probability (P-value) is less than 0.05. Under those conditions, one is entitled to draw the inference that the two samples do not belong to the same population and, therefore, are different.

Categorical Data

When data are categorical, statistics such as means and medians do not apply, for there is no range over which data are distributed; there is no standard deviation or variance. The variables occur in absolute categories. There may be multiple categories, but the archetypical form involves two categories per sample. The principles of dealing with two categories can apply to multiple categories but the mathematics gets a little bit more complicated.

When tabulated, categorical data form a contingency table with four values: the positive and negative variables for each of two samples (Fig. 18).

Any differences between the samples will be expressed in the form of differences between the proportion of positive variables that each sample expresses.

If

$P_A = a/(a+b)$

and

$P_B = c/(c+d)$

the ratio: P_A / P_B, will indicate how different sample A is from sample B.

	Variable	
	Positive	**Negative**
Sample A	**a**	**b**
Sample B	**c**	**d**

Figure 18. A table for the display of categorical data.

However, from a mathematical perspective, the numbers in a contingency table could have arisen by chance alone. By looking at the various ways that numbers inside such a table might be shuffled so that the rows add up to (a+b) and (c+d) respectively, and the columns add up to (a+c) and (b+d) respectively, mathematicians have determined the probability that the numbers inside the table produce certain ratios. Some ratios are common; some are rare.

It transpires that the distribution of possible ratios is in a particular form, called the chi-squared (χ^2) distribution. Upon adjusting for the sample sizes, one can determine what the probability is that a particular value of the ratio: P_A/P_B, arose by chance alone. By convention, if this probability is less than 0.05, one can infer that the observed ratio between the samples is unlikely to have arisen by chance alone. This is the basis for the **chi-squared (χ^2) test**.

A different test applies when the numbers in a table are all small: each less than 10. Under these circumstances, the chi-squared (χ^2) lacks specificity. It is generous, and produces probabilities that are lower than justified. This idiosyncrasy arises because of the shape of the chi-squared (χ^2) distribution.

For small numbers, the appropriate test is the **Fisher's exact test**. Like the chi-squared (χ^2) the Fisher's exact test determines all the possible ways that four small numbers might be arranged in a table, and calculates the probabilities of particular combinations of numbers having arisen by chance alone. From these calculations, the probability can be determined that the observed table of numbers has arisen by chance. If that probability is less than 0.05, one can infer that the table is unlikely to have arisen by chance.

Another approach to testing differences between proportions is to compare their confidence intervals. For any proportion (p), its 95% confidence intervals are given by the formula:

$$95\%\ \text{CI} = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

where n is the sample size.

Two proportions can be inferred to be distinctly different (with 95% confidence) if their confidence intervals do not overlap. For example,

if

$P_A = 0.8$, and $n_A = 10$, and

$P_B = 0.4$, and $n_B = 10$

95% CI of P_A $= 0.8 \pm 0.25$
$= 0.55$ to 1.05

95% CI of P_B $= 0.4 \pm 0.3$
$= 0.1$ to 0.7

Since $P_B + 0.3$ overlaps $P_A - 0.25$, they are not necessarily different. However,

if

$P_A = 0.8$, and $n_A = 50$, and

$P_B = 0.4$, and $n_B = 45$

95% CI of P_A $= 0.8 \pm 0.11$
$= 0.69$ to 0.91

95% CI of P_B $= 0.4 \pm 0.14$
$= 0.26$ to 0.54

Since $P_B + 0.14$ does not overlap $P_A - 0.11$, the proportions are likely to be different, with 95% confidence.

In these examples, although the original proportions were respectively the same, the second set was significantly different because the sample sizes on which they were based were larger.

Take-Home Messages

When looking for differences between samples,

> *for parametric data, the t-test is appropriate for testing the difference between two means;*
>
> *for non-parametric data, a rank test such as the Mann-Whitney test should be used to test the difference between two medians;*
>
> *for categorical data, the difference between proportions is tested, for which the chi-squared test can be used; but if the numbers are small, the Fisher's exact test should be used.*

By convention, these tests are significant when P < 0.05,

but that still means that there is 5% chance that the observed difference might be accidental.

Proportions can also be tested for difference by comparing their 95% confidence intervals.

LESSON 4. POWER

When an experiment is conducted, two samples are compared. Those samples respectively represent two populations. The objective of the experiment is define a rule that will eventually apply to the populations (i.e. generalisation). However, the rule is inferred from the behaviour of the samples.

Errors may occur that preclude the rule being generalized to the populations. These errors arise if the samples chosen are not representative of the average population. Rather, for reasons beyond the control of the investigator, the samples happen to be freak samples of one population or the other.

The errors that can occur are ones in which, between two samples,

a difference is found that is not a real difference; this is known as a false-positive, or **type I error**;

a difference that really should be there is not detected by the experiment; this is known as a false-negative, or **type II error**.

The possibility of such errors arises when the two populations in question are not completely different. Although substantially different, their distributions nevertheless overlap to some extent (Fig. 19).

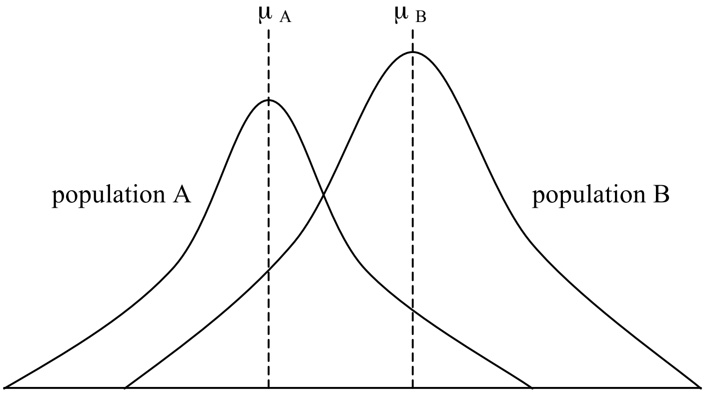


Figure 19. Two populations with significantly different means, but whose distributions overlap to some extent.

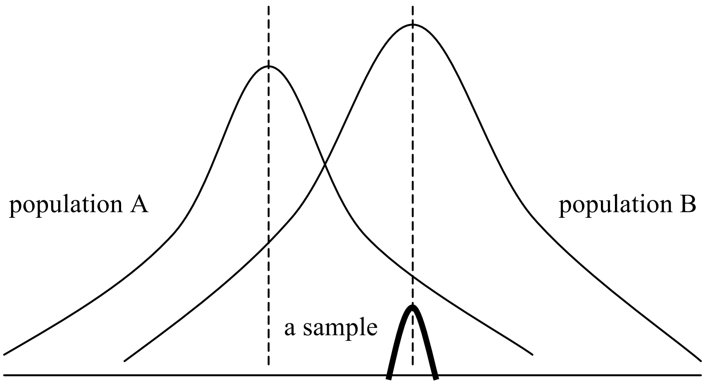


Figure 20. The basis for a type I error. A sample may be selected, that looks like a subset of population B and, therefore, looks different from population A. However, it is possible that the sample is actually a subset of the extreme distribution of population A.

Let's say that an investigator recruits a small sample whose mean coincides exactly with the mean of population B (Fig. 20). This sample looks different from population A, ostensibly in the same way that population B looks different from population A. The investigator wishes to infer that his sample is evidence of the difference between populations A and B.

However, because population A overlaps population B, there is a chance that the sample recruited was not a subset of population B, but a subset of population A. In other words, instead of being a sample recruited from the central regions of population B, it is a sample that belongs to the right-hand tail-region of the distribution of population A. If that is the case, the sample belongs to population A and, by definition, cannot be different. This is the statistical basis of the type I error.

In essence, a type I error occurs when a freak sample is recruited that, unfortunately, creates the illusion of a significant difference, but the difference is not real. The sample does not come from a significantly different population; it is just an extreme version of the control population.

No matter what the investigator wishes, he or she cannot tell that the sample was not a freak sample of population A, instead of the preferred population B. All that an investigator can do is be aware of this possibility, and take measures to reduce its probability. The possibility of an error cannot be eliminated; but its probability can be reduced.

A similar, but converse possibility arises. Let's say that in recruiting a true sample of population B, an investigator accidentally recruits a sample from the extreme left-hand end of the distribution of population B (Fig. 21). What the observer finds is that the mean of the sample coincides exactly with the mean of population A. Alas, the investigator infers that his sample is not different from population A and, therefore, that there must be no difference between population A and population B.

The truth is, however, that population B really is different from population A. The lack of observed difference arose because of the freak sample recruited. This is the basis for the type II error.

In essence, the type II error arises when a freak sample is recruited that, unfortunately, fails to demonstrate a difference when a difference is due. The sample comes from the extreme end of the test population, which overlaps into the central region of the control population.

By convention, the probability of a type I error is called α; the probability of a type II error is called β.

Figuratively, α may be conceived as the extent to which the control population overlaps under the test population, and β is the extent to which the test population overlaps under the control population (Fig. 22). Samples taken from the α zone will result in type I errors, and samples taken from the β zone will result in type II errors.

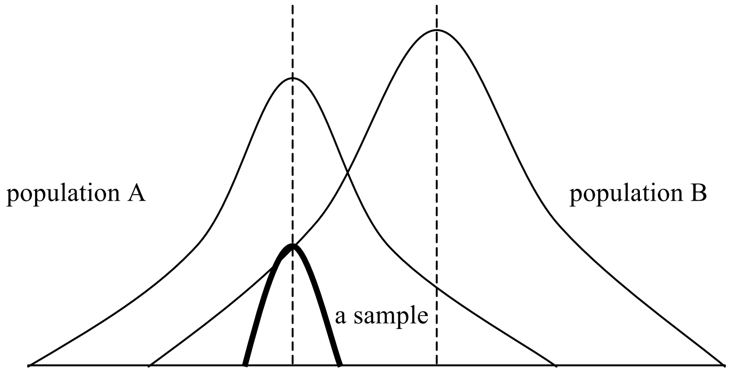


Figure 21. The basis for a type II error. A sample may be selected, that really is a subset of population B, but comes from the extreme end of that population, and fails to look different from population A.

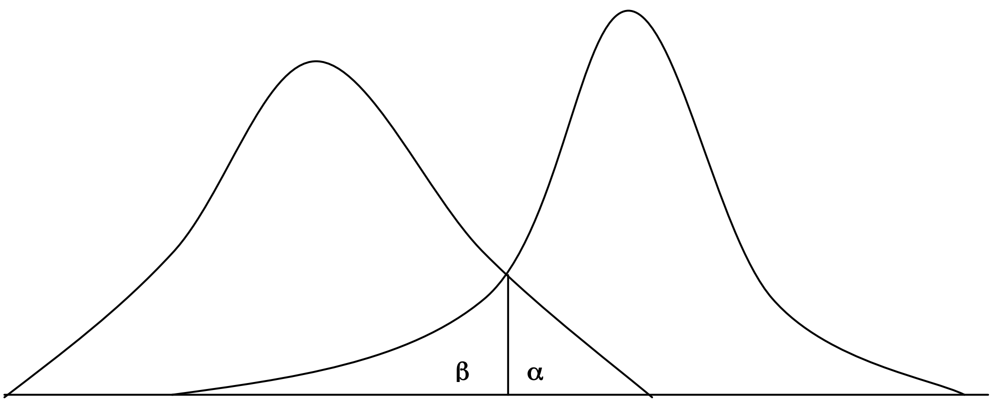


Figure 22. A graphic representation of the basis for type I and type II errors.

The measures that an investigator can take to reduce errors are

1. not to look for small differences as significant, but look only for large differences, or
2. use large samples.

By looking for large differences only, the investigator would need to rely on recruiting samples from the extreme left-hand end of the control population and from the extreme right-hand end of the test population (Fig. 23). Doing so avoids samples from the overlap zones.

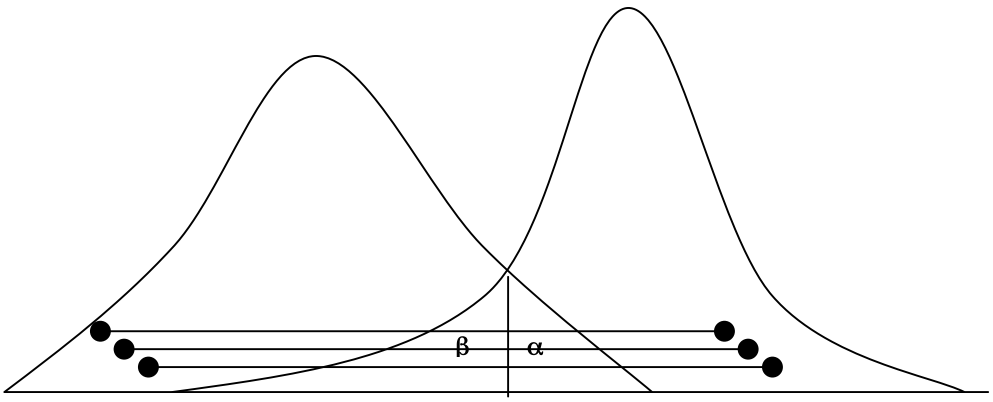


Figure 23. A graphic representation of how to avoid type I and type II errors, by seeking only large differences.

This strategy, however, prejudices the success of the experiment. If the two populations overlap, a difference will be found only if the appropriately extreme samples are recruited, which relies on luck. Also, if the two samples overlap greatly, it becomes unlikely that the appropriate samples will accidentally be recruited, and a difference will not be found. In that event, however, what thwarts the investigator is not his or her experiment but the fact that the two populations are so like one another, that the differences that might be detected are in general smaller than what the investigator can pursue without risk of a type I or type II error. Under such conditions, the differences may be so small that it is not worth conducting the experiment to "prove" them.

Increasing sample size captures an idiosyncrasy of statistics. While so long as sample sizes are small, there is a real chance that they might come from anywhere under the distribution of the population (including the α zone or β zone). Increasing sample size reduces the probability of this happening. The larger the sample size, the more it will tend to represent the true distribution of the population; and the less likely that the sample is a freak. At the risk of indulging too much figurative licence, understand that if the sample size is large, it cannot fit under the extreme ends of the distribution (Fig. 24).

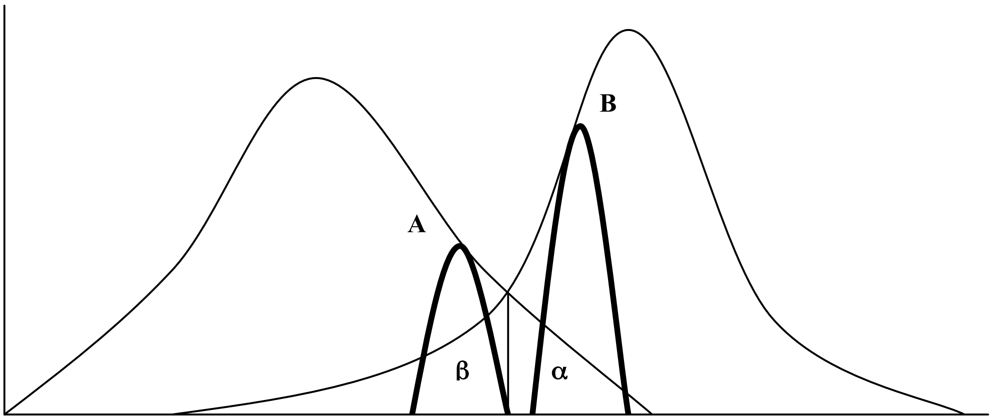


Figure 24. A graphic representation of how to reduce type I and type II errors by increasing sample size. Sample A will not "fit" under the α zone. Nor will sample B "fit" under the β zone of the overlap.

Mathematical formulae have been derived that relate α, β, sample size, and difference to be detected. Using such formulae, an investigator can adjust the required difference to be detected and/or the sample size required, in order to reduce α and β to some desirable or tolerable minimum.

By convention, α is usually set as 0.05, i.e. let there be no more than a 5% chance of a type I error. [Do not confuse this with a P-value of 0.05. They are different entities, and have in common only the convention of 5% being a tolerable limit]. Meanwhile β is usually set as 0.20, which allows a 20% chance of a type II error.

The convention is tolerant and more generous about type II errors because, in general, they matter less than type I errors. If an investigator incurs a type II error he or she can do little harm, for they simply fail to detect (and thereupon proclaim) a difference. Bad luck; but the status quo does not change. In contrast, a type I error can lead to a false claim of difference when none is due. So, the standards are set tighter.

However, one is not obliged to follow the convention. The standards for either α or β can be tightened (or relaxed). In some situations it may be important for an investigator not to miss a difference. They might choose, therefore, to reduce β to 0.10, for example. Upon doing so, they need only to change their sample size in order to achieve that value of β.

Another convention is that the value: 1-β, is known as the **power** of the experiment. Literally this means the probability of not having a type II error. Practically, it indicates how strong the experiment is avoiding type II errors, and implies the extent to which the observed differences are likely to be real ones. Expressed in terms used above, the power indicates how likely the test sample is of having not come from the β zone.

Take-Home Messages

The unfortunate facts of life are that there is always some chance that a sample recruited for an experiment will be a freak sample. It might be a sample that accidentally fits the investigator's desired result, or it might be a sample that accidentally thwarts the desired (and correct) result. Respectively, these freak occurrences are known as type I and type II errors.

The probabilities of type I and type II errors are known respectively as α and β. They are a function of sample size, and the difference to be detected.

The power of an investigation is the complement of the probability of a type II error, and is equal to 1-β.

Good investigators would ensure that their experiment has adequate power, to be able to detect real differences, with little risk of error. This can be done by calculating the sample size required to ensure prescribed or desired values of α and β.

Increasingly this is becoming a requisite by journals, but unfortunately, in the past this has rarely been the case. Investigators have arbitrarily selected sample sizes. Those who found differences and claimed a positive result, did not calculate the probability of having a type I error. Those who failed to detect a difference did not account for the probability of a type II error. The sample sizes that have been used in the past have often been so small that the probabilities of type I or type II errors are considerable. For all the good intentions, for all the good (clinical) work of the investigators, the data they reported may be no better than noise (irrespective of the result reported). In such data the truth does not lie.

The values of α and β cannot be calculated expeditiously by hand, but they can be determined rapidly by a power program in a computer. Armed with such a device, a consumer can readily address any data presented to them, and determine for themselves if those data are confounded by probabilities of type I or type II errors that are too high; which renders the message of the study unacceptable.

LESSON 5. EFFECT-SIZE

Let's say an investigator reports that a treatment is more effective than a control treatment ($P<0.05$). Let us assume that the probability of type I and type II errors are properly low. We can accept that there is a statistically significant difference between the outcomes of the index group and the control group.

The question that should arise is: just how good is that difference?

One approach is to use a clinical and intuitive evaluation. You can look at the size of the difference and judge for yourself if it is any good. It is your prerogative to do so, and it is your prerogative to defend your own value judgments.

Some of you, however, may be reluctant to offer a judgment based on personal, subjective responses. For you, there are certain statistical devices available. These devices do not provide an absolute judgment, but they do offer quantitative assistance to guide your subjective response. One device pertains to parametric data. The other pertains to categorical data.

PARAMETRIC DATA

Let there be a sample of individuals who express a distribution of values of a certain variable. (This variable could, for example, be pain, disability, or any other continuous variable.) This sample exhibits a certain mean (μ_1) with a standard deviation (s_1). Let this sample undergo some sort of intervention (such as treatment of pain or better teaching). After the intervention the sample expressed a new mean (μ_2) and a new standard deviation (s_2). Let there be an obvious, and statistically significant, difference between the means. A statistic is available that allows one to comment on how good that difference is.

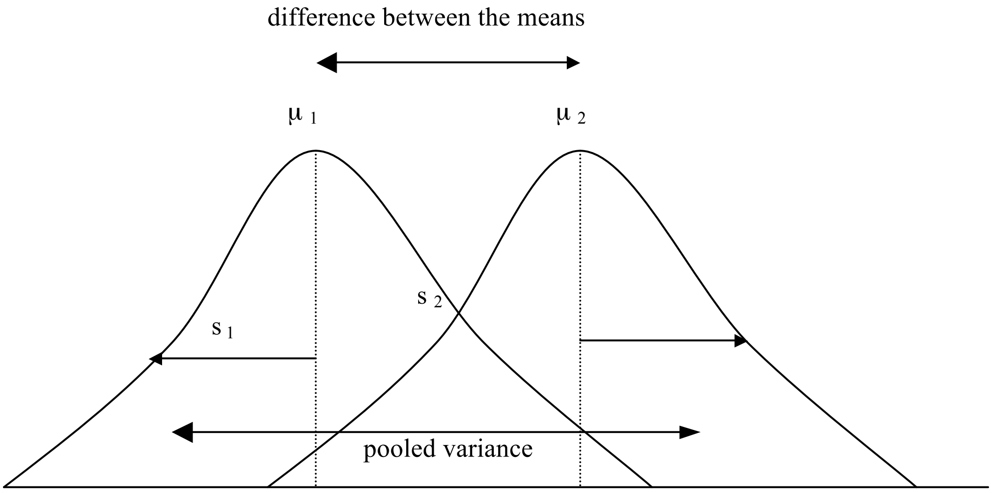


Figure 25. A graphic representation of effect-size for parametric data.

The statistic is Cohen's d[1]. It is defined algebraically as

$$\text{Effect-size} = \frac{\mu_1 - \mu_2}{s}$$

Expressed verbally, this statistic is the difference between the means divided by the standard deviation of the control sample.

Conceptually, the statistic expresses the change that occurs in the sample as a proportion of the natural difference within the sample. When the difference between the means is large in relation to the standard deviation, the effect of the intervention has been relatively large; and the statistic has a high value. Conversely, if the difference between the means is small, there has been little change, and the statistic will be small. If the difference is large, but if the variance is also large, the change in the mean is relatively insignificant, for it amounts to little more than "noise in the system". Under those conditions, the statistic will not be large.

Using this statistic, the effect-size can range in value from 0 to 1. A value of 0 means that there has been no effect. A value of 1 means that there has been a very large effect.

Cohen offered the following verbal correlates for intermediate values of effect-size:

0.20 = small
0.50 = medium
0.80 = large.

Take-Home Message

There is a statistic that can be used to quantify effect-size for parametric data. It compares the difference between the means in relation to the pooled variance of the samples.

Cynicism

There is an idiosyncrasy to Cohen's statistic. Because it is based on parametric data it operates only in the middle range of values. It cannot operate at extreme values, i.e. towards 0% or towards 100%. If the data are to be parametric they must be symmetrically distributed about a mean value. If the mean value is close to 0 or close to 100, the distribution cannot be symmetrical about these extremes (Fig 26).

Consequently, investigators who use Cohen's statistic cannot be reporting results in which large numbers of patients recover. They must be reporting clinically modest effects in the middle range of variables such pain scores or disability. If they had very good clinical results, they could not use Cohen's statistic.

Reference

Cohen J. Statistical Power Analysis for the Behvaioural Sciences. New York: Academic Press, 1977, pp 20-23, 40.

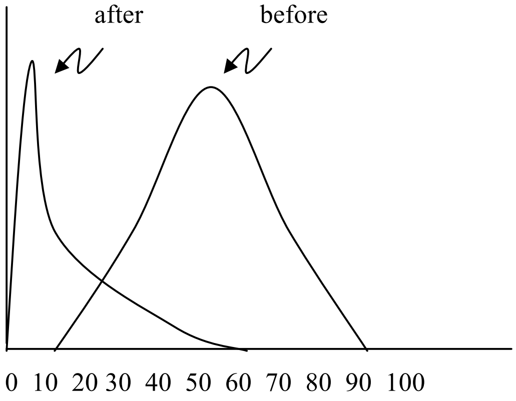


Figure 26. An illustration of how data centred around extremely low (or high) values cannot assume a normal distribution.

CATEGORICAL DATA

Categorical data are conveniently absolute. Either the treatment works or it doesn't. This makes the determination of effect size relatively straight-forward. Consider the following outcomes of a study of a treatment (Fig. 27).

For the purposes of the initial explanation, it does not matter what the definitions are of success or failure; that comes later. All that is important is that (somehow) a decision is made as to whether the treatment has been successful or not.

TREATMENT	**RESULT**	
	SUCCESS	**FAILURE**
INDEX	**a**	**b**
CONTROL	**c**	**d**

Figure 27. The categorical results of a clinical trial of an index treatment.

The proportion of patients who succeeded with the index treatment is a / (a+b).

Let this proportion be P_{index}, which is expressed as a decimal.

The proportion of patients who succeeded with the control treatment is c / (c+d).

Let this proportion be $P_{control}$, which is expressed as a decimal.

The attributable effect (AE) of the index treatment is the success rate that it achieves greater than the success rate achieved by the control treatment. The argument is that the control treatment provides non-specific effects, but these are also a component of the index treatment. The attributable effect of the index treatment is what remains when the success rate of the index treatment is discounted for these non-specific effects.

Mathematically,

$$AE = P_{index} - P_{control}$$

Since P_{index} and $P_{control}$ are both proportions, AE is also a proportion. It stipulates the proportion of patients treated whose successful outcome can be legitimately attributed to the effects of the index treatment above and beyond any non-specific effects.

Thus, if N patients are subjected to the index treatment one can expect N x AE patients to respond specifically to the unique effects of the index treatment. In other words, when N patients are treated, N x AE patients respond to the attributable effect of the treatment.

Consider then, what is the smallest number of patients that need to be treated for N x AE to be 1.

If

$$N \times AE = 1$$

$$N = 1 / AE$$

Therefore, N needs to be (1/AE) before one patient is encountered who can legitimately be claimed to have responded to the unique effects of the index treatment. This number is known as the **number needed to treat (NNT)** [1,2]; and

$$NNT = 1 / AE$$

Example1:

if $P_{index} = 0.78$
and $P_{control} = 0.45$

$P_{index} - P_{control} = 0.33$

AE = 0.33

1 / AE = 3

NNT = 3

If 100 patients are treated, 33 respond to the attributable effects of the treatment.

For one patient to respond, at least 100 / 33 must be treated, i.e. 3

Example 2:

if $P_{index} = 0.56$

and $P_{control} = 0.36$

$P_{index} - P_{control} = 0.20$

$AE = 0.20$

$1 / AE = 5$

$NNT = 5$

If 100 patients are treated, 20 respond to the attributable effects of the treatment.

For one patient to respond, at least 100 / 20 must be treated, i.e. 5

Notice how the NNT increases as the difference in outcome is less.

One of the virtues of NNT is that it provides an index of the strength of a treatment in terms of a single number. The smaller the digit, the stronger the treatment. Conversely, the larger the number the weaker the treatment.

Another virtue is that NNT provides an index of resources required - both medical and economic. If the NNT is high, it means that doctors will need to treat large numbers of patients before they get an attributable effect. This consumes time and effort. The doctors might consider if this large effort is worthwhile; and whether their efforts might not be better spent using another treatment. A large NNT also means that funds are being expended on large numbers of patients in order to get gains in a minority. Doctors might reflect as to whether these funds might be better spent otherwise; or if as good a result might be achieved, on the average, by using less expensive treatments.

For example, the NNT for epidural steroids is about 11 [3]. Effectively, this means that 11 patients must be treated before one can be claimed to have responded to the specific effects of the injections. The cost of that success is not just the time and expense required for that one success, but the costs incurred for the other 10 patients.

Subscripts

The NNT is not a single measure of all of the effects of a treatment. It measures the power of a treatment only with respect the outcomes specified in the original table of data from which the NNT is derived. Therefore, the pedantic but accurate use of NNT notionally requires that it be specified. This might be done as a subscript, but is not done so in practice because of the typographic absurdities required; but the concept is conveyed by this notation.

If the success in question is "ability to walk 1km in 10 minutes", the NNT for that outcome would be recorded as

$NNT_{\text{ability to walk 1 km in 10 minutes}}$.

If the success in question is "achieving a reduction of at least 50% in VAS score", the NNT for that outcome would be

$NNT_{\text{reduction in VAS by 50\%}}$.

No-one uses this notation, but it is taken as understood. Readers should understand that authors leave this implicit. They expect readers to have noticed what outcome they are addressing. Therefore, readers should consult the methods and results sections to find out what the subscript would have been had the authors used the absurd notation.

This is not an example of academic pedantry or an idiosyncrasy. It is an important realisation lest NNT be abused. An NNT might look good, and be used to extol a treatment as successful and useful. But the treatment might not be as good as it sounds if the reader realises that NNT pertains to an unconvincing or uncompelling outcome.

For example, the NNT for many drug therapies in Pain Medicine is about 3, which is considered a good score. But readers might care to ask - exactly what was the outcome measure. The risk obtains of readers being lulled into believing that with an NNT of 3, they could expect that for every three patients that they treat, one will be totalled cured. This is not be the case, for the NNT in question actually referred to "patients lowering their VAS by 50%". It says nothing about patients being completely relieved. Similarly, even if the NNT did refer to "complete relief of pain", it says nothing about "return to work". For that one would need $NNT_{\text{return to work}}$.

Always ask for what is the implied subscript.

References

1. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995; 310:452-454.

2. Laupacis A, Sackett DL, Roberts RS. An assessment of clinically useful measures of the consequences of treatment. New Engl J Med 1988; 318:1728-1733.

3. McQuay HS, Moore A. Epidural steroids for sciatica. Anaesth Intens Care 1996;24:284-286.

Take Home Message

If categorical data are provided, the effect-size can be determined by calculating the attributable effect and the NNT.

The NNT provides a clear indication of how well the treatment works.

NNT's of 3 or less are good.

Beware, however, of unqualified NNT's. If subscripts are not provided, check what the author actually meant by "success". Although the NNT might look good, it might not be for an impressive variable, or for an impressive mount of improvement.

LESSON 6. ODDS RATIOS

The odds ratio is a statistic that is sometimes used by investigators to describe data in a contingency table that relates two variables (Table App.5). In such a table, if there is a relationship between the two variables, the values in the "a" and "d" cells should be greater than the values in the "b" and "c" cells. The odds ratio (OR) seeks to reveal this difference.

Variable One	**Variable Two**	
	Present	**Absent**
Present	**a**	**b**
Absent	**c**	**d**

Figure 28. A contingency table from the odds ratio can be calculated.

Explicitly,

$OR = ad / bc = (a/b) / (c/d) = (a/c) / (d/b)$.

In one sense, the odds ratio describes the "balance" between the two rows. It describes the extent to which the ratio of "a" as to "b" is greater than the ratio of "c" as to "d". Similarly it describes the "balance" between the columns. It describes the extent to which the ratio of "a" as to "c" is greater than the ratio of "d" as "b". In another sense, it describes the balance between the two diagonals. It describes how much the product of "a" and "d" is greater than the product of "b" and "c". Each of these interpretations reflects by how much the "a" and "d" values are greater than the "b" and "c" values. The greater the "imbalance" between these values the stronger is the putative relationship between the two variables.

Historically, odds ratios have been used in epidemiological studies to determine the strength of association between exposure to a risk factor for a disease and the subsequent development of that disease. However, it is quite legitimate to use the odds ratio to compare other variables, such as two clinical signs, or the outcomes of two different treatments.

An odds ratio greater than 1.0 indicates that there is some relationship between the two variables; and the greater the odds ratio the stronger the relationship. Alone, however, the odds ratio does not indicate how clinically significant that relationship might be. In general, relationships start to become clinically significant when the odds ratio exceeds 3.0. Values great than 1.0 but less than 3.0 reflect a definite but only slight relationship.

Mathematically, it can be shown that the odds ratio (OR) is related to the likelihood ratio (LR). Thus,

$$
\begin{aligned}
OR &= ad/bc \\
LR &= [a/(a+c)]/[b/(b+d)] \\
&= (a/b)[(b+d)/(a+c)]
\end{aligned}
$$

If we let "d" be some multiple of "b", and "a" be some multiple of "c", i.e.

if $b = m.d$

and $a = k.c$

then,

$$
\begin{aligned}
LR &= (a/b)[(md+d)/(kc+c)] \\
&= (a/b)[d(m+1)/c(k+1)] \\
&= (a/b)(d/c)[(m+1)/(k+1)] \\
&= (ad/bc)(m+1)/(k+1) \\
&= OR\ (m+1)/(k+1)
\end{aligned}
$$

Thus, whenever, "a" is greater than "c", and "d" is greater than "b", the odds ratio will be greater than the likelihood ratio.

The advantage of the likelihood ratio, however, is that it is directly related to prevalence and diagnostic confidence and so can be used to compute diagnostic confidence directly (as described above). The odds ratio could be used to the same effect, but would need to be reduced by the coefficient (m+1)/(k+1) in order to obtain the correct arithmetic result.

Take-Home Message

The odds ratio can be used to determine the strength of relationship between two categorical variables. It measures the "imbalance" between the diagonals of a contingency table. Values greater than 3.0 suggest a clinically significant relationship.

LESSON 7. CORRELATIONS

Sometimes investigators are interested in establishing a relationship between two variables. The objective is to establish a rule whereby knowing the value of one variable will predict the value of the other variable. For example, disability might be predictable by severity of pain. When the variables are categorical, the odds ratio can be used. When the variables are continuous, other devices apply.

The correlation between two continuous variables can be assessed in the first instance by plotting the data. In such a plot, each datum point represents a single patient. Where each point lies on the graph is determined by the values expressed by that patient for the respective variables under consideration (Fig. 29) are distributed either side of that line. However, it is possible mathematically to estimate how well the scatter of points approximates a line that reflects the ”average” distribution of points.

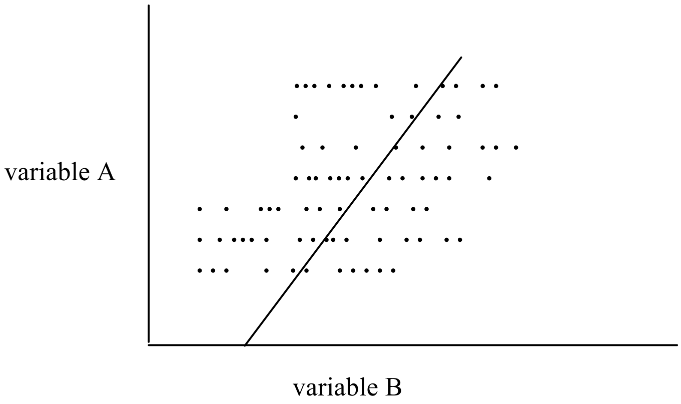


Figure 29. A plot of two sets of data between which there appears to be some correlation

The equation of the line can be determined by a process known as **linear regression**. The equation relates variables A and B to one another in terms of the scatter of A and B from their respective mean values.

How strongly the data are correlated is reflected graphically by how tightly they are clustered around the derived line. It is reflected numerically by a **correlation coefficient (r)**: a number whose value ranges from 0 to 1 (or from 0 to –1 if the relationship is inverse). When the magnitude of r is 1, the data lie perfectly on a straight line. When r is zero, there is no relationship between the variables. For values of r between 0 and 1, the relationship between the variables is proportionately stronger.

As with all statistics, it is possible that the derived correlation coefficient arose by chance. It is possible to calculate the probability that this is the case. Formulae are available by which to calculate the P-value of a correlation coefficient. A P-value of less than 0.05 is considered statistically significant. Such a P-value, however, does not alone indicate that there is necessarily a strong correlation. It shows only that the observed correlation is unlikely to have arisen by chance alone. In general, strong correlations obtain when the correlation coefficient is greater than 0.7.

Another worthwhile statistic is the square of the correlation coefficient (r^2). This statistic estimates how much of the variance between the two variables is explained by the derived equation relating them. The greater the scatter, the less one variable is directly related to the other. Some other factor contributes to the variance. Calculating r^2 is, therefore, important lest the investigator overstates the clinical significance of the reported calculation. When r^2 is small, something else is going on between the variables.

Correlation coefficients are calculated differently according to whether the data are parametric or not. For parametric data, the correct test is the Pearson correlation. For non-parametric data the correct test is the Spearman Rank Correlation.

These tests, however, assess the overall correlation between the two samples of data. Such global measures are not sensitive to inhomogeneities in the correlation, i.e. although the correlation coefficient may be large, the correlation does not apply uniformly across the full spectrum of the data; a tight clustering of data at one end of the relationship may compensate for poor clustering at the other. For example, two observers measuring the same phenomenon might score identical values when the variable has small magnitudes but might disagree when the variable is large. They might nevertheless achieve a reasonable correlation; but that correlation applies only when the variable in question is small. They cannot claim credit for agreeing across all possible values of the variable.

Another statistic is available to cater for such anomalies. It is the Intra-Class Correlation Coefficient (ICCC). It looks for the consistency of correlation across the full range of possible values of the variable. It is an unmerciful test. In order to achieve a high ICCC the data have to be consistently correlated across the full range of possible values. Any deviation from such consistency is penalized. Values of the ICCC range from 0 to 1, with 1 indicating perfect correlation. High values, such as 0.9, are very difficult to achieve, and occur only when there is a consistently strong correlation. Conversely, regardless and despite what correlation coefficients other tests might provide for the same data, a modest or low ICCC reveals that the supposed correlation is not really that good.

Take-Home Message

Various tests are available to determine the strength of correlation between two sets of continuous data. For parametric data the test is the Pearson Correlation. For non-parametric data the test is the Spearman Rank Correlation. Both tests yield a correlation coefficient (r), whose value ranges from 0 to 1. The greater the value of r, the stronger is the correlation. The statistic r^2 estimates the extent to which the variance between the data is explained by any calculated relationship.

Correlation coefficients, however, do not measure consistency of correlation across all possible values of the variables in question. A more demanding test is the Intra Class Correlation.

For establishing general relationships, perhaps for descriptive purposes only, the Pearson or Spearman correlations may be quite adequate; but if an investigator is seeking to report a precise or accurate relationship between two variables, the ICCC should be used. Beware of authors or speakers who try to get away with having used only a Pearson correlation when an ICCC should have been used.

LESSON 8. SENSITIVITY ANALYSIS

An investigator (or a consumer) may encounter data that seem to support a particular conclusion, but upon inspection, the data are not entirely supportive. Consider the data in figure 30.

Upon inspection, the results of the index group seem to be somewhat better; fewer patients have poor results, and somewhat more get moderate to good results.

Indeed, a chi-squared test shows that P = 0.018. So there is a significant difference in the proportions of patients with various results.

However, that is not the same as saying that the index treatment is consistently or overall better than the control. For example, the number of patients with really good results is not particularly greater in the index group.

The question that arises is where in this table is the source of the good P-value?

ΔVAS%	CONTROL	INDEX
100	4	6
90	5	7
80	8	9
70	10	12
60	11	14
50	9	15
40	13	10
30	14	5
20	12	4
10	15	2
0	0	0

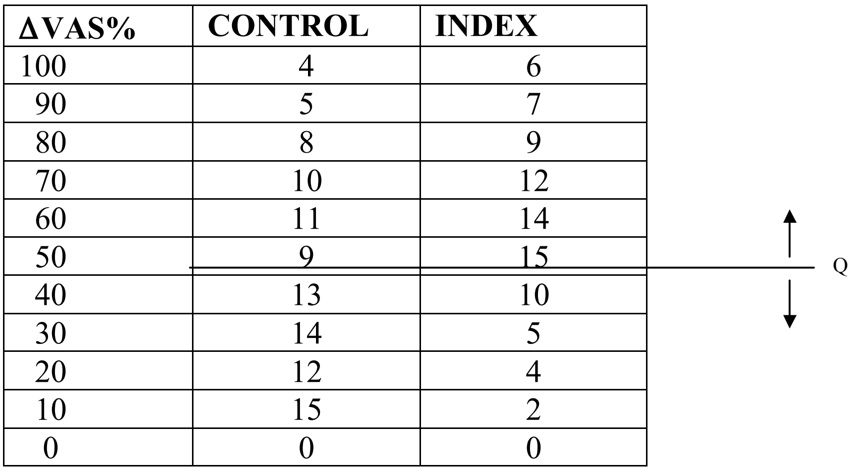
Figure 30. The results of a hypothetical controlled trial, listing the numbers of patients in each group who achieved the levels of pain-reduction indicated.

One approach to answering this question is a **sensitivity analysis**. Effectively this is a systematic trial and error approach, in which the data are tested and re-tested. The notion of "sensitivity" arises because the data are analysed, with the objective of detecting P-values, until the first significant P-value is found.

Any table of categorical data can be reduced to smaller tables based on the same data. A line can be drawn between any two rows in the table to yield values above the line and values below the line (Fig. 31). At each level, the larger table can be collapsed into a smaller, 2x2 table. A chi-squared test can then be applied to the smaller table to determine if there is a significant difference. If this process is systematically applied, from above downwards at each row, one can determine at which row the first significant P-value arises (Fig 32).

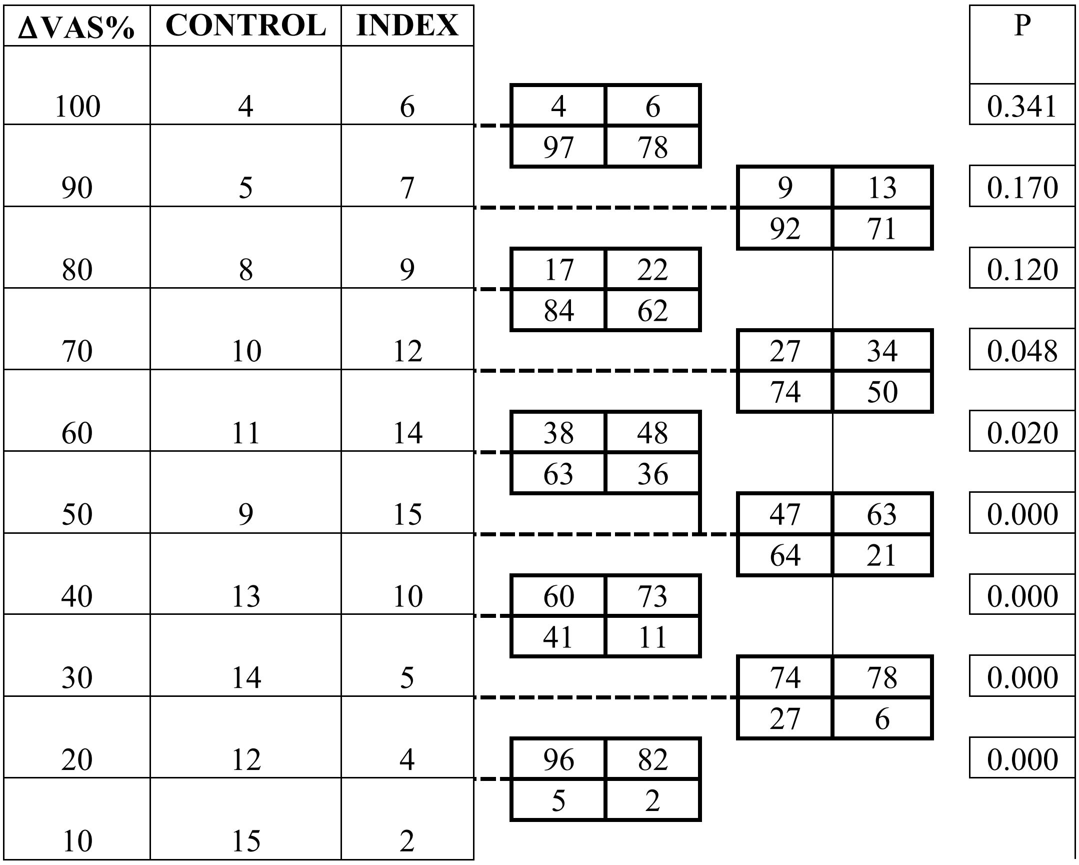
In the example shown in figure 32, it transpires that the first significant P-value occurs at row 70%. The tables above this level are not significant. For achieving 100% relief of pain, the index treatment is not significantly better than control. Nor is it better at securing 90%, or 80% relief. However, it is significantly better at securing 70% or better relief of pain (P = 0.048).

Subsequently, the index treatment is consistently better at securing 60% or greater relief (P = 0.020), greater than 50% relief (P = 0.000), and so on.



ΔVAS%	CONTROL	INDEX
100	4	6
90	5	7
80	8	9
70	10	12
60	11	14
50	9	15
40	13	10
30	14	5
20	12	4
10	15	2
0	0	0

Figure 31. A table of categorical results through which a line, Q, has been drawn to divide the table into outcomes of 50% or greater and less than 50%. Such a line could just as well be drawn between any two rows in the table to divide the outcomes into those above the line and those below the line.



ΔVAS%	CONTROL	INDEX			P
100	4	6	4 6 97 78		0.341
90	5	7		9 13 92 71	0.170
80	8	9	17 22 84 62		0.120
70	10	12		27 34 74 50	0.048
60	11	14	38 48 63 36		0.020
50	9	15		47 63 64 21	0.000
40	13	10	60 73 41 11		0.000
30	14	5		74 78 27 6	0.000
20	12	4	96 82 5 2		0.000
10	15	2			

Figure 32. Sensitivity analysis of a table of categorical data. Opposite each row of the table, the data above and below the row are collapsed to form a 2x2 table, which is subjected to a chi-squared test, and whose P-value is listed in the far right column. It transpires that the first three tables are not significant. The first significant P-value occurs at row 70%.

Thus, one can show, or one can detect, that

- the index treatment is significantly better than control, overall;
- but it is not significantly better than control for all levels of outcome;
- it is **not** significantly better at securing 100%. 90%, or 80% relief; but
- it is superior at securing 70% relief **or less**.

Any claims about the efficacy of the index treatment need to be couched in these terms.

Take-Home Message

Sensitivity analysis can be used to detect where, in a spectrum of results, significant differences start to appear; and reciprocally, over what range significant differences do not apply.