Study Guide 13: How to Make Sense of Numbers

Dr David Chinn,
Research & Development Office,
Queen Margaret Hospital, Dunfermline, Fife.
david.chinn@nhs.scot 01383 623623 (ext 20943)
Alternative contact: Dr Frances Quirk frances.quirk@nhs.scot 01383 623623 (ext 20941)

Contents Page

Disclaimer

I am an epidemiologist, not a statistician. These notes are written from my experience of working in the field of medical research for over 40 years. I have sought to give what I hope is a clear and simple explanation of some rather complex statistical principles. I do not profess to be an expert in statistics and a 'proper' statistician reading this guide may take issue with some of my explanations. Accordingly, I would encourage the reader to refer to one of the many excellent introductory books available on statistics for further guidance; some titles are given in the references and further reading.

(1) Overview and learning outcomes

This guide is for those who need to interpret numbers but have little or no knowledge of statistics. The content is appropriate for those who may need to critically appraise published (quantitative) articles. The focus is on interpreting rather than generating the results of a statistical analysis. However, some detail on the statistics is provided to facilitate the explanations, though this can be omitted without loss of the overall message. Examples are used throughout and at the end of reading this guide you should be able to:

- Distinguish between absolute and relative measures
- Describe and interpret a confidence interval
- Explain the distinction between confidence intervals and P-values
- Interpret the results of one-sample, unrelated (independent) and related (paired) t-tests
- Understand the limitations of the t-test
- Interpret the chi-square test for comparing proportions
- Interpret and make sense of the results from a drug trial.
- Make sense of data presented in graphs
- Understand the concept of linear correlation
- Interpret and make sense of the odds ratio
- Interpret and make sense of a Run Chart and Control Chart
- Interpret and make sense of a Funnel Plot
- Be aware of some of the common pitfalls in published statistics

The terminology can also be challenging so we have provided a glossary at the end.

Associated NHS Fife study guides:

7	How to plan your data collection and analysis
10	An introduction to medical statistics
11	How to calculate sample size and statistical power
12	How to choose a statistical test
14	An introduction to SPSS

(2) Introduction

Statistics is the science of assembling and interpreting numerical data. It is concerned with estimation and with describing uncertainty. We use descriptive statistics to estimate, for example, the prevalence of asthma in children within a community, the proportion of patients with hospital acquired infection, the length of stay (in hospital), the demographics of patients attending a particular clinic, the benefit of a drug on some physiological response etc. When presented with such figures it can be a challenge to make sense of them in terms of what they tell you and, just as importantly, what they do not. Consider the following typical statements:

- Mortality in Group A was 60% higher than that in Group B
- The mean length of stay was 4.3 days but the median was only 1 day
- Mean age of disease onset was 38.2 years in men and 43.3 years in women (mean difference 5.1 years, 95% confidence interval 3.5 to 6.7, P=0.009, unmatched t-test)

- The average reduction in diastolic blood pressure was 9 mm Hg (95% CI 4.5 to 13.1, P<0.01, paired t-test)
- Lung size correlated strongly with height in adult men (r=0.71, P=0.001)
- The number needed to treat with drug A was 38 whereas that for drug B was 12
- As an example of a real problem, consider the following:

In the 1990s a survey was undertaken of NHS Board members who held responsibility for commissioning services. A questionnaire was sent with details of 4 rehabilitation programmes. Respondents were told that each programme cost about the same. They were asked to review the information presented on the different outcomes and select the best programme suitable for funding. The 4 programmes with their associated outcomes were:

Prog 1 – with an absolute reduction in deaths of 3%
Prog 2 – with an increased survival from 84% to 87%
Prog 3 – with reduced death rates by 19%
Prog 4 – 33 patients needed to avoid 1 death

Which programme would you have chosen? *(we'll return to this question in a later section)*

Confused? This guide should help you identify the strengths, limitations and interpretation of these types of statistical results.

A single statistic will have only limited utility. For example, if you are told the average height of school children aged 13 years is 152 cm this ignores any difference between boys and girls and you cannot assume that all children are this height. Also, the single value of 152 cm tells you nothing about the minimum or maximum heights of the group of children. So, what figures do you need to describe fully these details for a population?

Any set of measurements that describes data from a sample has two important properties: the average, central or 'typical' value and the spread of values about that average. We use descriptive measures to describe 'typical' values (also called measures of location), such as the mean, median, mode and the spread of values such as the variance, standard deviation (SD), interquartile range (IQR) and confidence intervals of the mean and other estimates. You will hear these terms used widely in describing data and further details of their derivation are given in the NHS Fife study guide 'Introduction to Medical Statistics'.

Other ways of describing data include bar charts showing frequencies for different groups within a sample, histograms for a variable showing frequencies of data split into ranges, pie charts depicting proportions and scatterplots exploring the relationship between two quantitative variables.

(3) Awareness of Numbers: Absolute and Relative Measures

Numbers may be presented as absolute figures (e.g. the number of people who died from road accidents in a year) and in relative figures (e.g. the proportion, or percentage, of people who were aged 18-25 years, amongst those who died in road accidents in a year).

Be wary of the use of percentages in headlines as the way summary data are presented can be misleading. In the mid-1970s Barbara Castle, the Minister of Health, announced a 30% salary increase for student nurses. This *relative* amount was generous and looked impressive but student nurses were poorly paid and 30% of a small, *absolute* number is itself still a small number! A percentage salary increase of 10% looks impressive when inflation is running at about 2% but work out for yourself what is the absolute increase in salary for those currently employed (2020) on the National Minimum Wage which, for a person aged over 25 years is £8.72 per hour. A 10% increase amounts to 87p per hour.

Consider the headline “70% of deaths from Swine ‘flu are in women”. Think about what it is you are being presented with. Make the distinction between absolute and relative figures. When you are given a percentage to consider ask what base number it relates to. In this example is the 70% estimate based on 10 or 100 cases. For an extra woman (instead of a man) dying from Swine ‘flu the percentage for the group of 10 patients would increase from 70 to 80%; for the group of 100 the percentage would increase from 70 to 71%.

In the 1990s the Department of Health published routine summary statistics on health service activity and outcomes for different trusts in England. A journalist picked up on one aspect and the following newspaper headline appeared:

“Infant mortality in Gateshead 50% higher than national average.”

This caused great concern locally, particularly amongst parents of new born babies.

The infant mortality rate is the number of infants who die in their first year of life as a proportion of the total number of live births. The rate is considered a good reflection of the state of the health services in a country. In the UK infant mortality has dropped markedly since 1900 when the rate was about 140 / 1000. At the end of the century it was about 6 / 1000 (or about 4% of the rate at the start of the century). How does this relate to numbers? The population of the UK has grown over the century but the number of births has declined. In England & Wales in 1901 the commonest causes of death in infants under a year of age were atrophy, debility and premature birth. The total number of infants dying was huge at 140,648. In 1998 the commonest causes of death in infants under a year of age were ‘Neonatal’ and Sudden Infant Death Syndrome (SIDS). In total 3,625 infants died (or about 2½% of the 1901 total) (Table 1).

How else could these figures be reported? A comparison of rates differs from a comparison of the absolute number of deaths (Table 2). This is because the number of births in each year is itself different, about 1,005,000 in 1901 and about 600,000 in 1998.

Table 1. The number and principal causes of death in infants aged 0-1 years, England and Wales, 1901 and 1998.

Principal cause and number of deaths, England and Wales			
1901		1998	
Cause	Deaths	Cause	Deaths
Atrophy, debility	18,685	Neonatal	2,418
Premature birth	18,562	Sudden Infant Death (SID)	234
Convulsions	15,513	Anomalies of the heart	42
Diarrhoea	13,233	Ill-defined Intestinal infection	41
Enteritis	13,084	Asphyxia	39
Bronchitis	11,694	Other diseases of the lung	39
Bronchial pneumonia	6,228	Meningococcal infection	25
Whooping Cough	4,793	(Whooping Cough	2)
Others	*38,856*	*Others*	*785*
Total	140,648	Total	3,625

Table 2. Alternative ways of reporting data on change in infant mortality, England and Wales, 1901-1998

	1901	1998
Number of deaths age <1 yr	140,648	3,625
Rate / 1000 live births	140	6
Comparison of rates	6/140 (%) = 4.3%,	or 95.7% reduction,
	or 140/6 = 23 fold decrease	
Comparison of number of deaths	3,625 / 140,648 = 2.6%,	or 97.4% reduction,
	or 140,648 / 3,625 = 39 fold decrease	

Consider the statement "Infant mortality in Gateshead 50% higher than the national average" and what this means in terms of the actual number of infants dying. In absolute terms it was a difference between 9/1000 live births (in Gateshead) compared with a national average of 6/1000 live births (Table 3).

Table 3. A comparison of relative and absolute measures of infant mortality

	National average	Gateshead
Rate / 1000 live births	6	9
Comparison of rates	9/6 (%) = 150%,	**or 50% increase**
	or 9/6 = 1½ fold increase	
Absolute terms	**3 extra deaths per 1000 live births**	

The journalists were correct in that infant mortality in Gateshead was 50% higher than the national average (the *relative* comparison). But the difference in *absolute* terms worked out at just 3 extra deaths per 1000 live births. The Director of Public Health was interviewed on local television and tried hard to make this distinction stating that the extra deaths were small in number and related to impoverished living arrangements, poor social circumstances, poverty etc, common to city dwelling communities, but each time the journalist would emphasise the relative statistic. The message here is:

Be wary of rates based on small numbers.

Always consider the base number on which a percentage or rate is based!

(4) Awareness of Numbers: The Illusion of Accuracy

An audit of emergency admissions to a surgical unit was undertaken to estimate the proportion of GP referrals that were considered inappropriate. Ninety-three consecutive admissions were reviewed over four weeks and 18 were considered inappropriate by the surgeons. The proportion is 18/93 but this was reported as 19.3548 %. The reporting of figures to 4 decimal places implies a level of accuracy that, in this case, is simply not justified. Even reporting it to 2 decimal places is inappropriate because the sample is relatively small and you cannot estimate the accuracy of the *true* proportion with such a small sample. In this example, 19 % is perfectly satisfactory. The message here is:

Be wary of numbers reported to many decimal places which can give a spurious illusion of accuracy.

(5) Awareness of Numbers: The Level of Uncertainty

Statistics is concerned with estimation and with describing uncertainty. Each estimate from a sample is a snapshot of what you think is the *true* value in the population from which the sample was drawn. In the example above this is the proportion of inappropriate referrals being made by GPs to a local hospital surgery unit. The *level of uncertainty* is described in a confidence interval which is a *range of values* in which we believe the true value is likely to lie. It is common practice to report the 95% confidence interval (95% CI) which is interpreted as: ‘we are 95% confident that the true value lies between x and y’ (*see* Glossary). In the example above where the proportion was 19% based on 93 observations the 95% CI is 11% to 27%. A different sample of similar size would yield a different proportion but we would expect the estimated proportion to lie within this range of values (confidence interval). Now, it may be that the surgeons consider this range to be too wide. They decide to repeat the audit to obtain a better estimate, and narrower confidence interval. This would involve a larger study but the numbers needed would depend on the width of the confidence interval the surgeons decided was optimal. The message here is:

A single estimate must be interpreted in relation to its confidence interval and to increase precision you must increase the sample size.

Some information on the principles behind the derivation of confidence intervals is given below. However, more specific detail on how to calculate a confidence interval and how to determine the sample size needed for a study are given in the NHS Fife Study Guides: ‘An Introduction to Medical Statistics’ and ‘How to Calculate Sample Size and Statistical Power’.

(6) Confidence Intervals for a Continuous Variable

A continuous variable is one which can theoretically take any value within a given range (for example, height). When estimating some characteristic in a population (e.g. average height) we take a sample that we hope is representative of the population and calculate summary statistics from the sample. We can calculate the average value (the mean) for the characteristic and the standard deviation (SD) which is a measure of the variability in the data. The SD is the average of the deviation of individual values from the mean and describes the spread of data around the mean. The SD is calculated from the data itself and, in data that are distributed in a bell-shape (also called the 'Normal' or Gaussian distribution) about 68% of the observations will have a value between one SD below and one SD above the mean. Similarly, about 95% of the observations will have a value between 1.96 SD below the mean and 1.96 SD above the mean (Figure 1).

The mean ($\bar{x}$) and SD from the sample should be a good estimate of the *true* mean (μ) and SD (σ) of the population from which the sample is drawn, provided the sample is representative. However, the sample mean is unlikely to be exactly the same as the population mean. A different sample would give a different estimate of the population mean and the difference would be due to *sampling variation*. The relationship between the population true mean and SD and that estimated from the sample is shown in Table 4.

Table 4. Relationship between a sample and the population from which it is drawn

	Reality, 'Truth'	Estimate
Population:	Size = N	Sample: size 'n'
Mean:	μ	$\bar{x}$
SD:	σ	SD
		↓
		Standard Error of the mean, SE ($\bar{x}$) = SD /$\sqrt{n}$
		↓
		Confidence Interval of the estimated mean

The estimated SD is used to derive the *standard error* (SE) which is a measure of the accuracy of the estimated mean ($\bar{x}$).

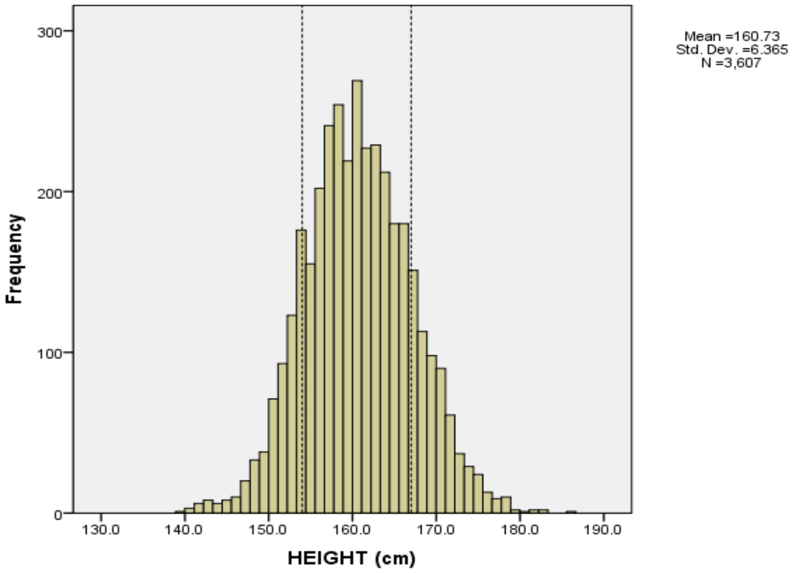
Standard Error of the mean, SE ($\bar{x}$) = SD of sample /$\sqrt{n}$ (equation 1)

where 'n' is the number of observations in the sample.

The size of the SE depends on the degree of variation in the sample and on the size of the sample; the larger the sample, the smaller the SE. The SE in turn is used to calculate the confidence interval for the estimated mean. The 95% confidence interval is calculated as:

95% CI = sample mean $\pm$ 1.96 x SE ($\bar{x}$) (equation 2)

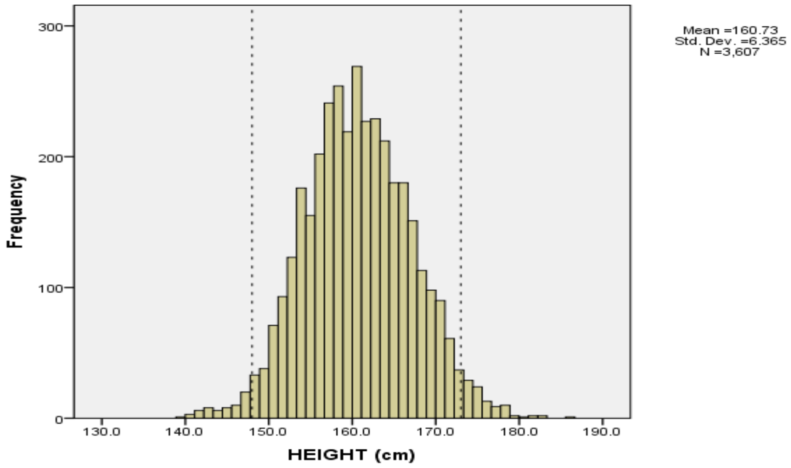
Using 1.96 to define the interval is acceptable when you have a sufficiently large sample (>30) but the situation is different when you have much smaller samples. Then you use a larger value than 1.96, taken from the t-distribution, to define the interval because, when based on small numbers (<30) there is a greater level of uncertainty in the assumptions (further details in the NHS Study Guide: Introduction to Medical Statistics).



Mean = 160.7 cm,
SD = 6.4 cm

Mean +/- 1 SD =
154 – 167

This range will contain
about 68% of the
observations



Mean = 160.7 cm,
SD = 6.4 cm

Mean +/- 1.96 SD =
148 – 173

This range will contain
about 95% of the
observations

Figure 1. The height of 3,607 adult women recorded in the Scottish Health Survey, 1998.

(7) Confidence interval for a proportion (percentage).

A narrow confidence interval means your estimate is precise whereas a wide confidence interval means it is imprecise. A wide confidence interval can occur because it is based on a small number of observations, or that there is simply a large degree of variation in the individual data. To calculate a confidence interval (CI) for a proportion you need:

- The proportion (or percentage) itself
- The number of observations
- The standard error of the proportion (or percentage)
- The number of standard errors needed to encompass the interval chosen (e.g. a 95% CI needs 1.96 SEs)

For a proportion (P), SE (P) = $\sqrt{(P\ (1\text{-}P)/n})$ (equation 3)

95% CI = P ± 1.96 x SE (P) (equation 4)

For a percentage (P%), SE (P%) = $\sqrt{(P\%\ (100\text{-}P\%)/n})$ (equation 5)

95% CI = P% ± 1.96 x SE (P%) (equation 6)

In the audit of emergency admissions above the percentage of inappropriate referrals based on 93 observations was 19%. The 95% CI was calculated as:

SE (P%) = $\sqrt{(\mathrm{P\%}\ (100\text{-}\mathrm{P\%})/\mathrm{n})}$ = $\sqrt{19(100-19)/93}$ = 4%

The 95% CI = P% ± 1.96 x SE (P%)

the upper CI = 19 + (1.96 x 4.0) = 27%

the lower CI = 19 – (1.96 x 4.0) = 11%

(8) The relationship between Confidence Intervals and P-values from tests of statistical significance

A confidence interval (CI) is a range of values in which the true mean for a population is likely to lie. A narrow CI means your estimate is precise. A wide CI means your estimate is imprecise, due either to it being based on a small number of observations, or that there is simply a lot of variation in the data.

A P-value is derived from a test of statistical significance as used in testing a hypothesis. In any experiment comparing, for example, the frequency of observations in two samples there will always be a difference between them. The question is whether the observed difference reflects a true difference between the two populations from which the samples were drawn. Significance tests cannot prove that an observed result is due to a real effect. It can only assess this in terms of **‘the probability of occurrence of a result as extreme, or more extreme than that observed if the null hypothesis were true’.** This is the definition of the P-value (also referred to as the type I error). Effectively, it is the probability of getting the wrong answer! Selecting a P-value of 0.05 to delineate *statistical significance* and the decision to reject the null hypothesis means we are prepared to be wrong 1 chance in 20 (5%).

The P-value has been likened to the ‘probability that the observed effect is due to chance’ but this interpretation is frowned upon by some statisticians.

The first stage is to make up the null hypothesis which states that there is no difference (in whatever you are testing) between the two populations. We then select a sample from each population which we hope is representative of that population and set out to disprove the null hypothesis by applying a significance test to the measurements from each sample. The P-value assesses how likely it would be to observe such a difference between the two samples when in reality there is no such difference between the populations from which the samples were drawn. It is common practice to consider a P-value of less than 0.05 to indicate a *statistically significant result*. The smaller the P-value the stronger is the evidence against the null hypothesis of no difference in the mean value under test between the two populations. A P-value of 0.001 would encourage you to reject the null hypothesis and conclude that there is a real difference between the two populations, but you might be wrong and, in this case, the probability of being wrong is only 0.001 or 1 in 1000.

Authors sometimes write that a significance test revealed a P-value between 0.05 and 0.10 and the test “just failed to reach statistical significance”. This is a poor phrase frowned upon by statisticians. What this actually means is that there is no difference in the mean values of whatever is under test between the groups, or, more likely, that there were too few participants to demonstrate a difference between the groups if one truly exists (Figure 2).

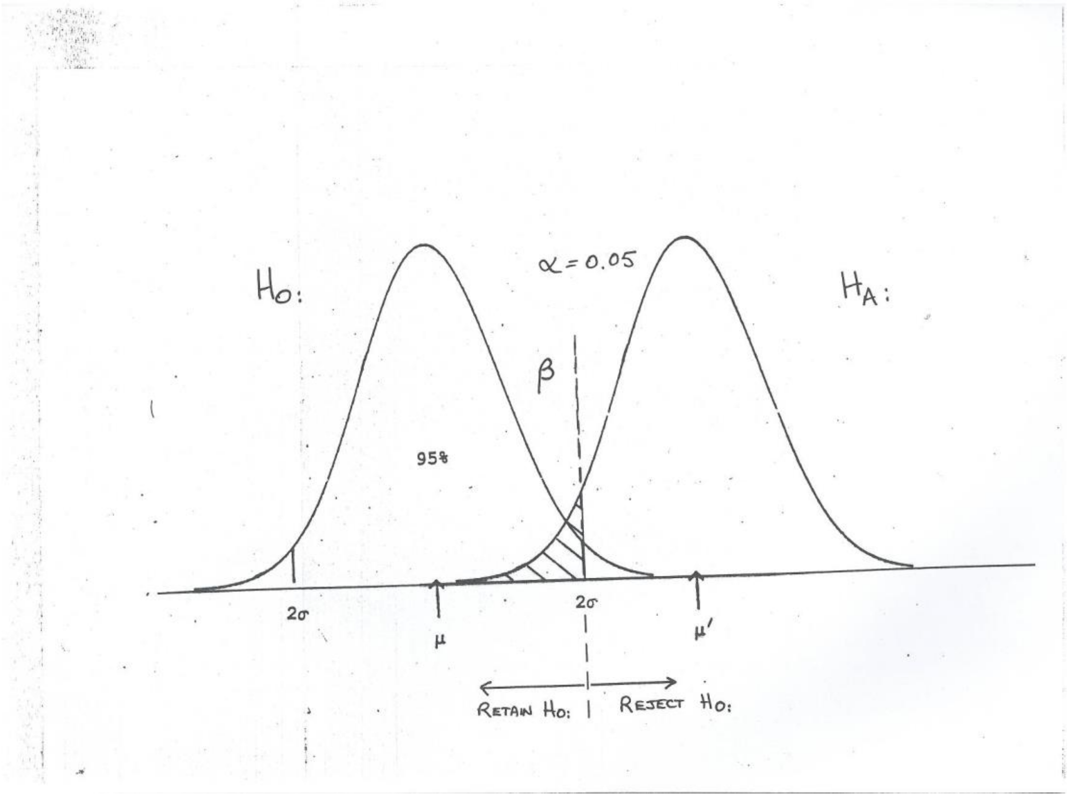


Figure 2. A hypothesis test and statistical considerations comparing two distributions.

H_O= null hypothesis of no difference in the means between the two populations
H_A= alternative hypothesis in favour of a difference between the two populations
μ and μ´ are the mean values for each population
2σ refers to 2 standard deviations.
α = 0.05 is the P-value associated with the decision to reject the null hypothesis in favour of the alternative hypothesis (also referred to as the type I error).
β is the type II error and 1- β is the *power* of the study to detect a difference if one truly exists.
The further apart the distributions are the greater is the difference between the mean values of the two samples, and the greater the likelihood that the difference measured reflects a true difference in mean values of the populations from which each sample was drawn.

Example: Mortality in patients with heart failure
A randomised controlled trial was undertaken to compare the mortality experience of two drugs (A and B) being used to treat patients with heart failure.
Of those patients on drug A, 33% died.
Of those patients on drug B, 38% died.
The difference = 5%. The statistical test comparing proportions gave a P-value of 0.07 (not statistically significant by convention).
The 95% confidence interval of the *difference* in mortality was -1% to +12%.
This *95%* interval includes the value zero so is not significantly different from zero at the *5%* level (P=0.07).
But, in interpreting the confidence interval we are 95% confident that the true difference in mortality between drugs A & B is between:
-1% (A worse than B), or +12% (A better than B)
Interpretation: Drug A is likely to be better than Drug B for reducing mortality in patients with heart failure but the evidence underpinning the inference is weak.

Should we cite P-values or confidence intervals? A P-value will tell you whether or not there is a statistically significant difference between two populations. The confidence interval will provide information about the size of the difference and the strength of the evidence. Confidence intervals provide more information than a P-value alone and both should be cited.

The relationship between P-values and confidence intervals is visually represented in Figure 3.

Statistical inference and extrapolation of results from a sample to a population assumes that the sample is representative of the population from which it is drawn. When this is not the case the conclusions will lack validity and be unreliable.

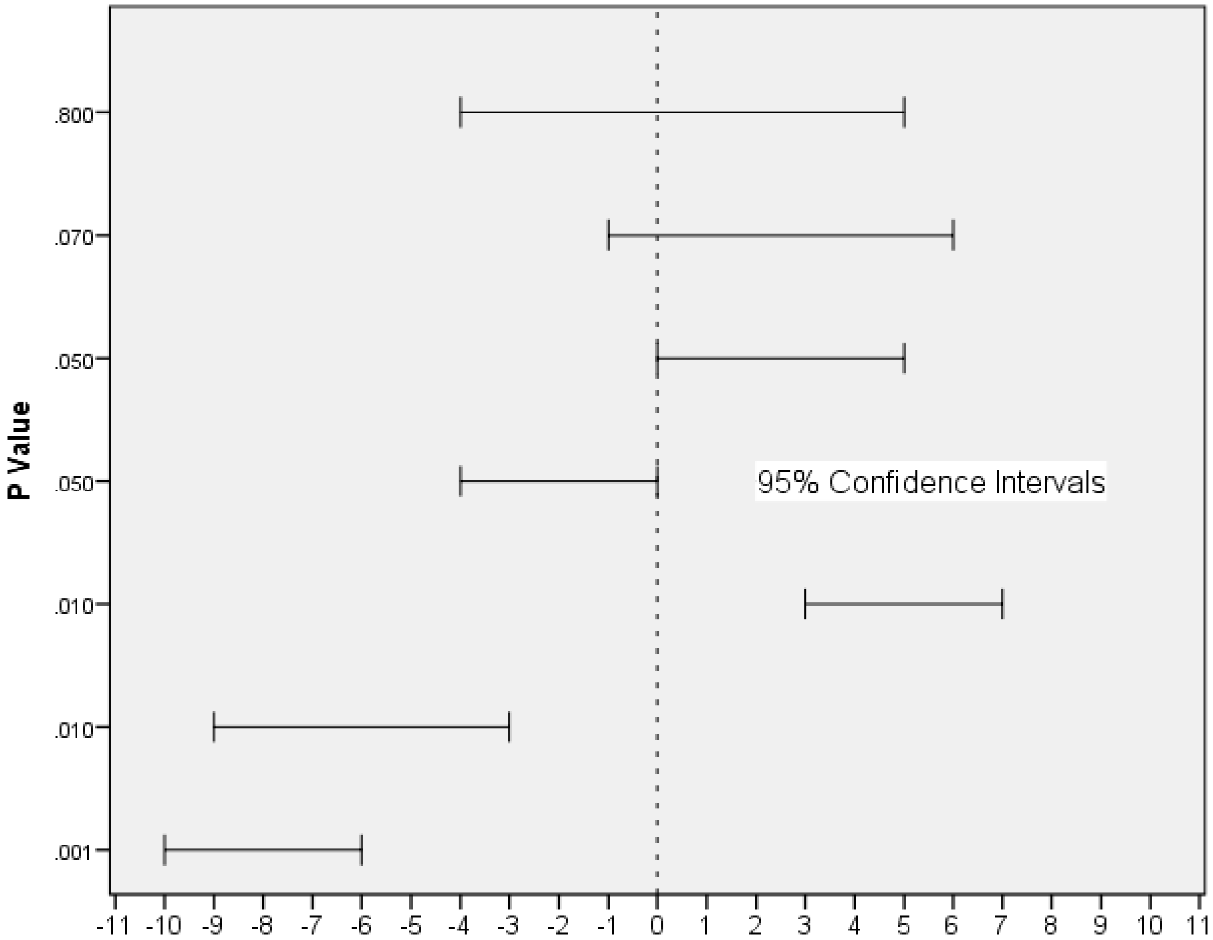


Figure 3. The relationship between P-values and the 95% confidence intervals.

The zero line along the horizontal axis represents 'no difference' between the samples. A P-value greater than 0.05 (**5%**) (e.g. 0.8 as above) will be associated with a **95%** confidence interval that **staggers the zero line**. This implies the null hypothesis cannot be rejected suggesting there is no difference in the populations from which the samples were drawn. A P-value of exactly 0.05 (**5%**) will be associated with a **95%** confidence interval that **starts from zero**. A P-value less than 0.05 will be associated with a 95% confidence interval that **does not include zero**, hence we can reject the null hypothesis in favour of an alternative suggesting that the two populations from which the samples were drawn are significantly different from one another in the characteristic under investigation. In simple terms the smaller the P-value the further away from zero will be one end of the confidence interval. However, we may be wrong in rejecting the null hypothesis. The reality may be that the two populations do not differ but our samples suggest they do. The P-value represents the probability of making the wrong decision.

(9) Tests of Statistical Significance: the T-test

The t-test is a parametric test, that is, one applied to data that are Normally distributed with a bell-shape (see Figure 1 above for an example of a Normally distributed variable). The assumptions underlying use of the t-test are:

- the data come from a Normal (Gaussian) distribution
- the samples are not too small
- the samples do not contain outliers (particularly a problem for small samples)
- For comparison of 2 samples, that:

- the samples are of equal or nearly equal size
- the variances are equal or approximately so (but this is not critical).

When comparing two groups the shape of the two distributions should be similar (see Figure 2). If the shapes vary markedly in that one distribution is 'wider' and less 'peaked' than the other the analysis may be unsound. However, there is a modification to the calculations that allow for the difference in shape of the distributions and statistical packages will make this adjustment. In general, the t-test is robust against small deviations from these assumptions.

There are 3 different t-tests available depending on the data being analysed.

(9.1) The One-Sample t-Test

This test is applied to a single distribution. The null hypothesis states that the mean of the sample does not differ significantly from some hypothesized mean. In effect it compares a sample mean with what you think the true mean should be.

Example: haemoglobin concentration at referral in women with colorectal cancer.

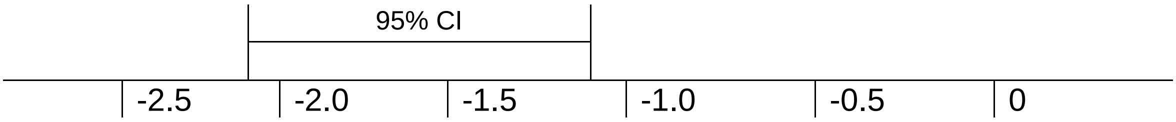
The haematology results were reviewed from 58 women who were subsequently diagnosed with colorectal cancer to see if they had evidence of anaemia on presentation.

The expected reference value for haemoglobin (Hb) in women = 13.6 g/dl.

Mean Hb on presentation =12.0, SD 1.9 g/dl

Mean *difference* between the women's Hb and 13.6 (the reference value) = –1.6 g/dl

The 95%CI of the difference = – 2.1 to –1.1, P<0.001



Note: this **95%** confidence interval does not include the value zero so it is significantly different from zero at the **5%** level, in this case <0.001.

Interpretation: we are 95% confident that the true mean Hb at presentation in women with colorectal cancer is between -2.1 and -1.1 g/dl less than the reference value of 13.6 g/dl.

(9.2) The Two-Sample, Unrelated, Independent Groups t-Test

The test, also called the unmatched t-test, is used to compare two unrelated (independent) samples. The null hypothesis states that the means of the populations from which the samples are drawn do not differ significantly from one another.

Example: haemoglobin concentration at referral in women with colorectal cancer.

This time the results were reviewed from 58 women diagnosed with colorectal cancer to see if those with right-sided disease of the colon had a different degree of anaemia on presentation compared with those with left-sided disease of the colon.

These women represent two independent samples

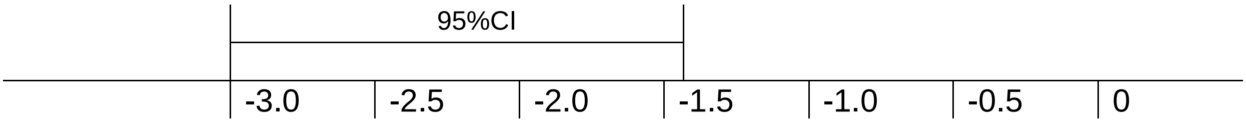
Mean Hb on presentation (n=58) =12.0, SD 1.9 g/dl

Mean Hb on presentation (n=36 with Left-sided disease) =12.9, SD 1.5 g/dl

Mean Hb on presentation (n=22 with Right-sided disease) =10.7, SD 1.6 g/dl

The *difference* in the means of the two groups = –2.2 g/dl

The 95%CI of the difference = – 3.0 to –1.4, P<0.001



Interpretation: we are 95% confident that the true mean Hb in women with right-sided disease is between -3.0 and -1.4 g/dl less than those women with left-sided disease.

(9.3) The Related, Paired Samples t-Test

This test is similar to the one-sample t-test. It involves one group of subjects where each participant has two measurements that are paired (e.g. before and after a drug, a follow-up study, a cross-over design where each participant gets both treatments in turn, or a case-control design where a participant is matched with a single control). The null hypothesis is that the mean difference in the pairs is zero.

Example: change in disease activity in ulcerative colitis over 3 months

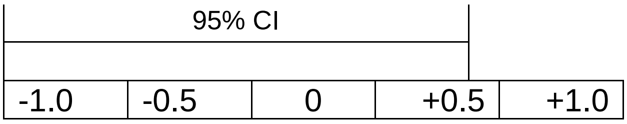
59 patients with ulcerative colitis (UC) were examined and their disease activity assessed. They were re-examined 3 months later to determine the change, if any, in disease activity.

Mean baseline score = 5.9, SD 4.3

Mean follow-up score = 5.6, SD 3.9

Mean difference in paired scores = –0.3 (SD 2.7)

95% CI = –1.0 to + 0.4, P=0.4



This result was not significant (P=0.4) and the 95% confidence interval includes the value zero, so we cannot reject the null hypothesis of no change in disease activity.

Could we have compared the baseline and follow-up results with the two-sample t-test? No, because the two samples are related and not independent. In any case the baseline mean of 5.9 and SD of 4.3 suggests that the data are not Normally distributed, which violates one of the assumptions for using the t-test. The clue here is that, because disease activity can only take a positive value the mean (5.9) minus twice the SD results in a negative value (-2.7). Remember, the mean +/- twice the SD encompasses 95% of the observations in a set of data that are Normally distributed (see Figure 1). The same problem of a negative value for disease activity holds true for the follow-up score. The mean and SD of the *differences* in disease activity was -0.3 and 2.7 so it is not possible from these numbers to assess if the distribution of the differences did conform to a Normal distribution. Use of the paired t-test assumes the differences are Normally distributed and, thankfully, this was the case (Figure 4).

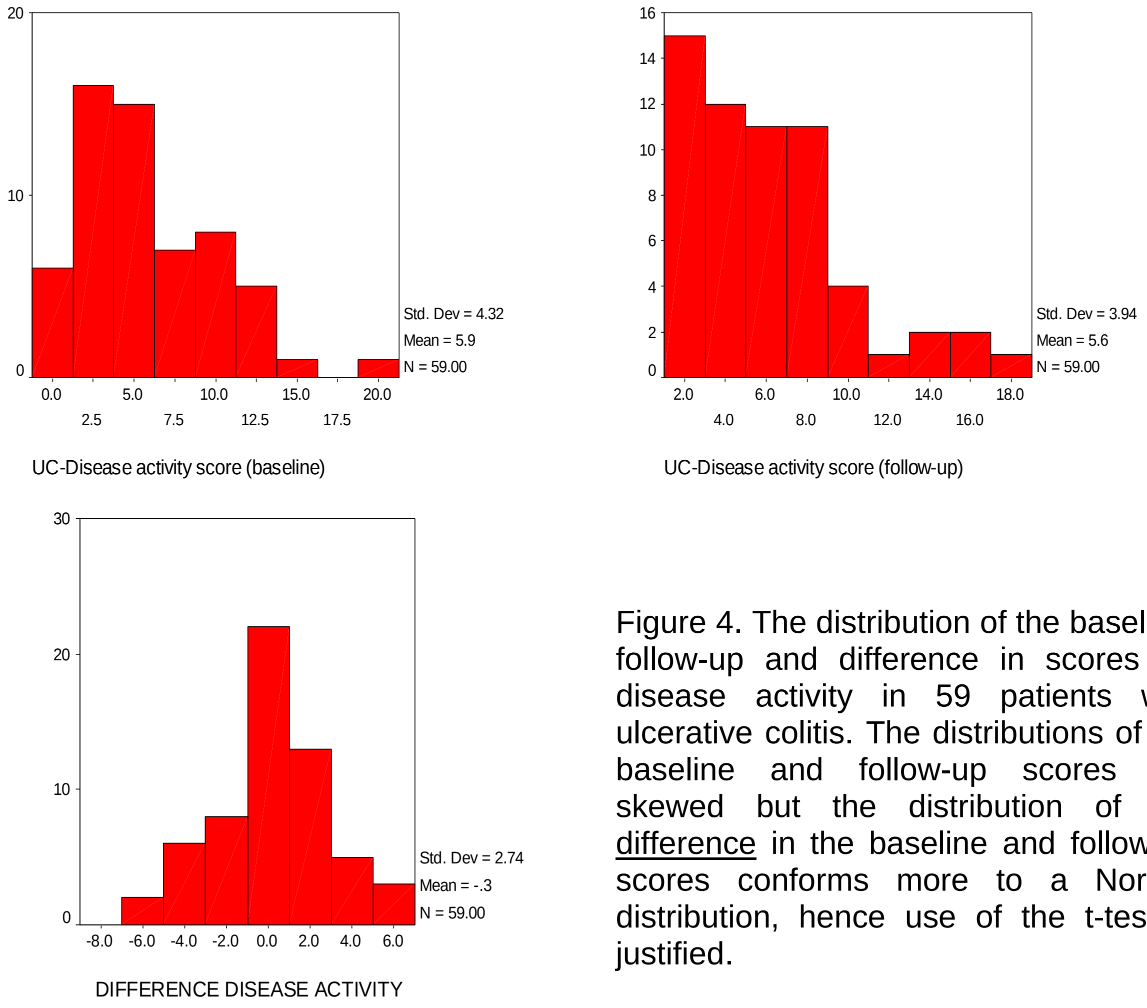
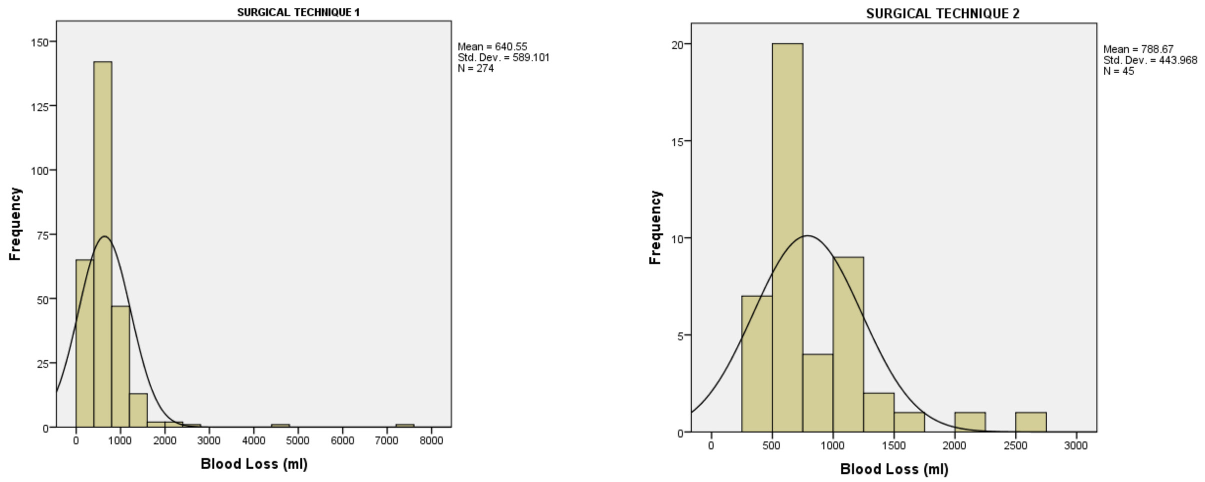


Figure 4. The distribution of the baseline, follow-up and difference in scores for disease activity in 59 patients with ulcerative colitis. The distributions of the baseline and follow-up scores are skewed but the distribution of the difference in the baseline and follow-up scores conforms more to a Normal distribution, hence use of the t-test is justified.

(9.4) Consequences of using the t-test when the assumptions are not met

A study was undertaken comparing the average blood loss associated with a particular operation performed using two techniques. Technique 1 (the usual procedure) was performed on 274 patients with a mean blood loss of 640 ml and an SD of 589 ml. Technique 2 (a new procedure) was performed by the same surgeons on 45 patients with a mean blood loss of 789 ml and an SD of 444 ml. The researchers used a two-sample t-test to compare the data yielding a P-value of 0.108 suggesting the null hypothesis of no difference in the mean level of blood loss between the two techniques could not be rejected. However, use of the t-test (a parametric test) assumes the data are Normally distributed and the researchers had not checked the distribution of the data beforehand. However, just ‘eyeballing’ the summary statistics suggests the data are not Normally distributed as, for each technique, the mean value minus twice the SD yields a negative value for blood loss, which is implausible! This was confirmed when the data were plotted as a histogram (Figure 5).



Surgical Technique 1 Surgical Technique 2

Figure 5. Distributions of blood loss associated with two surgical techniques

When data are not Normally distributed the mean is less reliable as a measure of 'central tendency' when the correct statistic is the median (*See* Study Guide 10, An Introduction to Medical Statistics). Because the data are not Normally distributed the researchers should have used a non-parametric test which, effectively, compares the medians (the t-test compares the mean values). One such test is the Mann-Whitney U test (*See* Study Guide 12 How to Choose a Statistical Test). The summary statistics and results of the statistical tests are shown in Table 5.

Table 5. Consequences of using the wrong test when comparing two groups

Blood Loss (ml):	Technique 1	Technique 2	Statistical Test	
n	274	45	Parametric test	Non-parametric test
Mean	640	789		
Median	500	700	t-test	Mann-Whitney
SD	589	444	P=0.108	P=0.002
Minimum	100	250		
Maximum	7500	2500		
25th Percentile	400	500		
75th Percentile	750	1000		

The correct analysis, using the non-parametric test now revealed that, overall, blood loss was significantly greater using technique 2 (P=0.002).

(10) Tests of Statistical Significance: the Chi-Square Test to Compare Proportions

The chi-square test is a common test to compare proportions (frequencies) in 2 or more groups. Consider the research question: In women eligible for breast screening does a personalised letter from the GP improve uptake? A randomised controlled trial was undertaken in one practice where women eligible for breast screening were randomised into two groups. One group (the intervention) received a letter from their GP encouraging them to attend the breast screening invitation. The other group did not receive the letter (control). The null hypothesis is: ‘in women eligible for breast screening there is no difference in uptake to an invitation to attend mammography from use of a personalised letter from the patient’s GP’.

Of 470 women invited for breast screening 254 (54%) attended. The attendance rates were 51.3% in the control group and 56.8% in the intervention group. Is this difference in proportions statistically significant? The results are summarised in Table 6 which is referred to as a 2 x 2 contingency table.

Table 6. A randomised controlled trial comparing the effects of a personal letter from a GP on the uptake of an invitation to breast screening.

Attended for Mammography:	+Letter (Intervention)	- Letter (Control)	Totals
Yes	134	120	254
No	102	114	216
Totals	236	234	470

In each of the four shaded cells we calculate the number of women *expected* to attend if the letter had no effect on their decision. Overall, 254/470 women attended (54%). The expected number is calculated for each cell as:

Column total x Row total / Overall total (equation 7)

For the upper left cell this becomes: 236 x 254 / 470 = 127.5 and the expected values for the other cells are:

Lower left	= 236 x 216 / 470 = 108.5
Upper right	= 234 x 254 / 470 = 126.5
Lower right	= 234 x 216 / 470 = 107.5

The chi-square statistic, referred to as χ^2, is calculated as:

$= \sum(\text{observed - expected})^2$ / expected (from the 4 cells) (equation 8)

The symbol ∑ refers to ‘sum of’, in this case the sum of the equation for the 4 cells.

In this example, χ^2 = 1.42, and P=0.23 (from the stats tables), so a non-significant result. The 95% confidence interval for the difference in proportions is -3.5% (GP letter reduced uptake) to +14.5% (GP letter increased uptake).

Interpretation: In this trial there is no evidence that a personalised letter from the GP would improve the uptake of breast screening among the population of women from which the sample was drawn.

But, **a lack of evidence of an effect is not the same as evidence of no effect** and the 95% confidence interval suggests the letter is more likely to improve uptake than reduce it. The options at this stage include planning a larger study with a power calculation to determine the sample size needed whereby a difference in uptake of about 5% would be statistically significant at P=0.05 or less, assuming the null hypothesis can be rejected to reflect a true, positive effect from the letter.

(11) ‘Negative’ studies

Occasionally a trial comparing a drug with a placebo is described as being ‘negative’ ($P>0.05$) implying that the drug is no better than placebo. The drug comparison study above in patients with heart failure had a P-value of 0.07 and may have been described as a negative trial implying the two drugs were equivalent. However, you should be wary of trials described as ‘negative’ as **lack of evidence of an effect is not the same as evidence of no effect.** For example, a study was undertaken on the relationship between overuse of mobile ‘phones and the development of brain cancer. The results failed to show an association. But, the fact that you do not have evidence to show mobile ‘phones are harmful is not the same as stating they are safe. (See the article by DG Altman and JM Bland. Absence of evidence is not evidence of absence. *BMJ* 1995; 311: 485)

(12) Statistical Versus Clinical Significance

It is very important to distinguish between statistical significance and clinical significance. Be wary of results of very large studies where small changes in clinical outcomes may be reported as highly significant but have little meaning clinically.

(13) Interpretation of Results from a Drug Trial

Consider the question: Is aspirin effective in reducing the incidence of heart attacks? A randomised controlled trial was undertaken in 22,071 men randomised into one of two groups. One group took one aspirin tablet a day, the other group took one placebo tablet a day. The outcome was the number of heart attacks over 1 year. This involved a comparison of proportions and the data are summarised in Table 7.

Table 7. Results from a randomised controlled trial comparing daily aspirin and placebo in the incidence of heart attack

Group:	Heart attack	No Heart attack	n	Attack rate
Placebo	239	10,795	11,034	239/11,034 = **0.0217**
Aspirin	139	10,898	11,037	139/ 11,037 = **0.0126**

Can we reject the null hypothesis that the attack rate is the same in both groups?

Attack rate in placebo group (p_1) = 0.0217,
Attack rate in aspirin group (p_2) = 0.0126,
Difference in attack rates ($p_1 - p_2$) = 0.0091, $P<0.00001$

The significance test: *this part can be omitted*

$z = (p_1 - p_2) / SE(p_1 - p_2)$

$SE(p_1 - p_2) = \sqrt{\ } [(p_1(1-p_1)/n_1) + (p_2(1-p_2)/n_2)]$

$SE(p_1 - p_2) = \sqrt{\ } [(0.0217 (1 - 0.0217) / 11{,}034) + (0.0127 (1-0.0127) / 11{,}037)]$

$= 0.001749$

$z = 0.0091 / 0.001749 = 5.20$, $P<0.00001$

The 95% CI $= p_1 - p_2 \pm t_{(0.05)} \times SE(p_1 - p_2)$

$= 0.0091 \pm 1.96 \times 0.001749 = 0.0091 \pm 0.0034 = 0.0057, 0.0125$

The P-value is <0.00001 so a highly significant result with a less than 1 in 100,000 chance that we would be wrong in rejecting the null hypothesis of no effect of aspirin.

The 95% confidence interval for the difference in proportions is: 0.0057, 0.0125 so we are 95% confident that the *true* difference in attack rate lies between 0.0057 and 0.0125.

We can extract other useful statistics from this comparison of proportions. For example, the difference in proportions ($p_1 - p_2 = 0.0091$) and its confidence interval is small, but consider the *relative* risk which is: 0.0217 / 0.0126 = 1.72

> **Interpretation:** members of the placebo group were 1.72 times more likely to have a heart attack than members of the aspirin group.

The *relative risk reduction* (RRR) is the proportional reduction in rates of adverse events between an experimental and control group.
EER = experimental event rate, CER = control event rate
RRR = |EER-CER| / CER = |0.0126 - 0.0217| / 0.0217 = 0.0091 / 0.0217 = 0.419 or 42%
Note: the '|' lines before and after the term EER-CER indicates that we must ignore the sign of the difference

> **Interpretation:** members of the aspirin group showed a 42% reduction in adverse outcome compared with the placebo group.

The *absolute risk reduction* (ARR) is the absolute (arithmetic) difference in rates of adverse events between the experimental and control group.
ARR = |EER-CER| = 0.0091 or 0.9%
This value is used to calculate the *number needed to treat* (NNT) which is the number of patients who need to be treated to achieve one additional favourable outcome. The NNT is calculated as the reciprocal of the ARR.
NNT = 1 / ARR = 1 / 0.0091 = 110 patients.

> **Interpretation:** we need to give 110 patients aspirin for a year to prevent one heart attack.

In a similar way a study could focus on drug side effects and the results be used to determine the numbers needed to harm (NNH). In fact, for aspirin the NNH is about 400 so that for every 400 patients treated with aspirin for a year we would expect one patient to suffer an adverse effect (? Gastrointestinal bleed or whatever) but to prevent a heart attack in about 4 patients.

The NNT and NNH values are useful statistics to be taken into account with costs when recommending treatments and when communicating with patients on the risks and benefits of a particular medication.

Example (from the introduction):
In the 1990s a survey was undertaken of NHS Board members who held responsibility for commissioning services. A questionnaire was sent with details of 4 rehabilitation programmes. Respondents were told that each programme cost about the same. They were asked to review the information presented on the outcome of each programme and select the best one suitable for funding. The 4 programmes with their associated outcomes were:

Prog 1 – with an absolute reduction in deaths of 3%
Prog 2 – with an increased survival from 84% to 87%
Prog 3 – with reduced death rates by 19%
Prog 4 – 33 patients needed to avoid 1 death

The information presented gave different criteria on outcomes but, in reality, the programmes were identical. 140 board members responded but only 3 identified the summary statistics were from the same programme. The authors concluded that, in this sample, those charged with commissioning services lacked the necessary skills to make informed decisions.
The rationale:

Death rate rehab = 13% (survival 87%)
Death rate control = 16% (survival 84%)
Reduction in death rate = 3%
Proportional reduction in deaths = 3% / 16% = 19%
NNT = 1 / 0.03 (or 100 / 3) = 33

See: Fahey et al *BMJ* 1995; 311: 1056-1059.

(14) How to Make Sense of Data Presented in Graphs

(14.1) Assessing the appropriateness of the scales

Data presented in graphical form can sometimes mislead. For example, a paper was published in which the authors claimed that infant mortality for a country had fallen markedly between 1970 and 1994. They presented a graph (Figure 6) which showed a downward trend. When presented with such a graph your eye is first drawn to the information in the middle, i.e. the declining slope. But you should also look at the scales. The horizontal axis (Year) is appropriate but the vertical axis (IMR, Infant mortality rate) is scaled from 23 to 26 deaths / 1000. Hence, this does represent a decline but the *magnitude* of that decline is only 2 deaths / 1000 over the 24 years, hardly a ‘marked’ fall. This is further seen if the data are re-plotted with zero on the vertical axis (Figure 7).

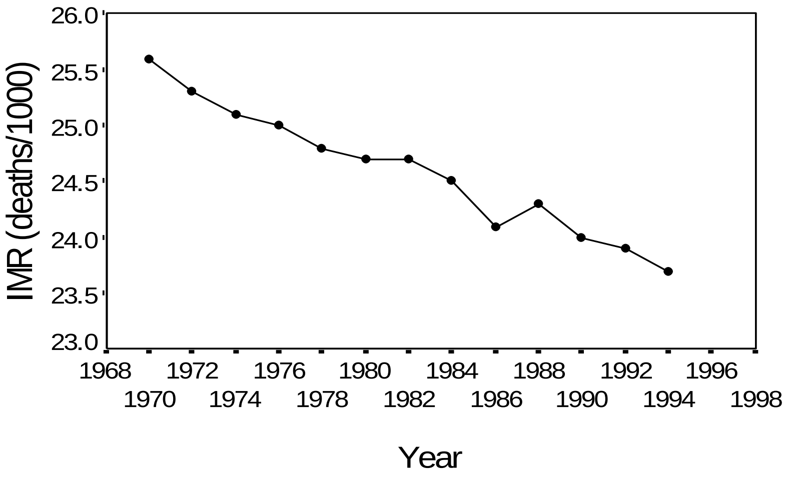


Figure 6. Infant Mortality Rate (deaths / 1000 live births)

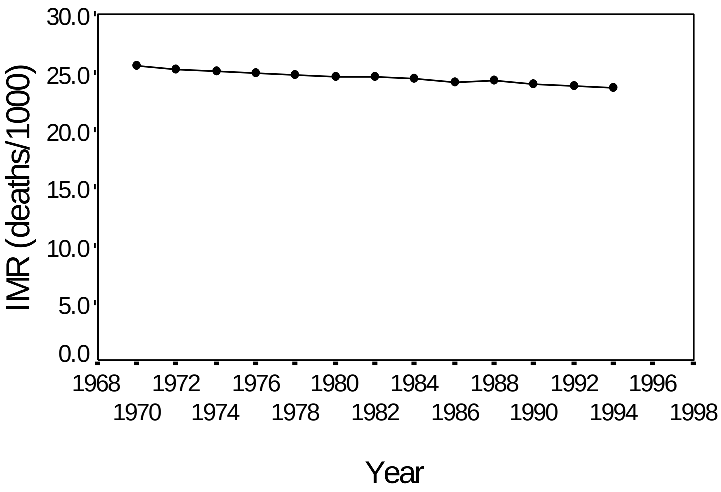


Figure 7. Infant Mortality Rate (deaths / 1000 live births) re-plotted with zero on the vertical axis.

Consider the data plotted in Figure 8 which shows the relationship between % body fat and biceps skin fold thickness in adult males. The biceps thickness is plotted on a linear scale, i.e. equal increments from 0 to 30 mm. However, the correlation with % body fat is not linear but suggestive of a curvilinear relationship. Furthermore, the *spread* of values for the biceps skin fold on the vertical axis increases with increasing % body fat. This variation in spread of values is referred to as being heteroscedastic and suggests a proportional relationship between the spread (variance) of skin fold thickness and % body fat. The analysis of data in this form can be a challenge. However, we can improve the situation. The biceps skin fold data is skewed in distribution (not bell-shaped) and transforming the data by taking logarithms (*see* Glossary) results in a distribution that fits better with the Normal, bell-shaped distribution (Figure 9). Now the relationship between % body fat and the log of the biceps skin fold does appear linear and the spread of the biceps data is approximately equal whatever the value of % body fat, a pattern described as homoscedastic (Figure 10). Compare the pattern of the data as plotted in Figures 8 and 10. The analysis of data as displayed in Figure 10 is more straightforward. The message, again, is to look at the graph's horizontal and vertical scales to know what you have been presented with. Compare the vertical axes in Figures 8 (linear) and 10 (logarithmic). Note the difference in the size of the gap between adjacent marks.

Incidentally, it is possible to transform data from a skewed distribution into a Normal distribution using other mathematical functions such as the reciprocal (1/*x*), or by taking an exponent (x^2, x^3).

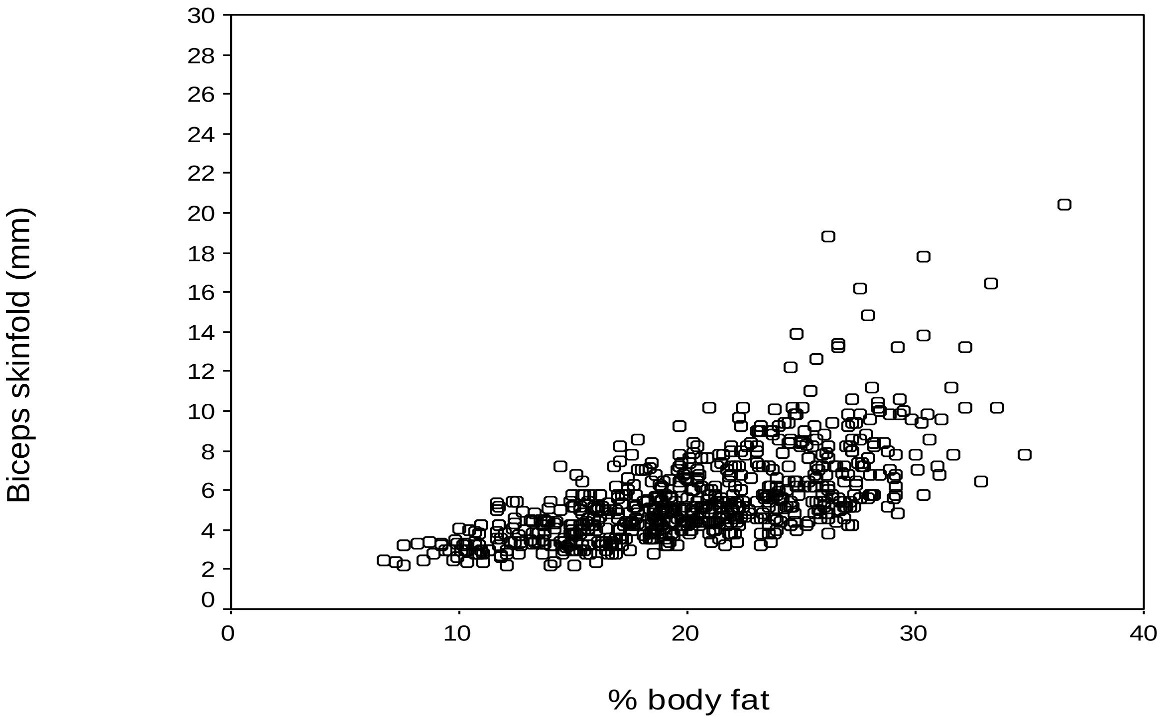


Figure 8. Relationship between %body fat and biceps skin fold thickness (mm) in adult males
(linear vertical scale)

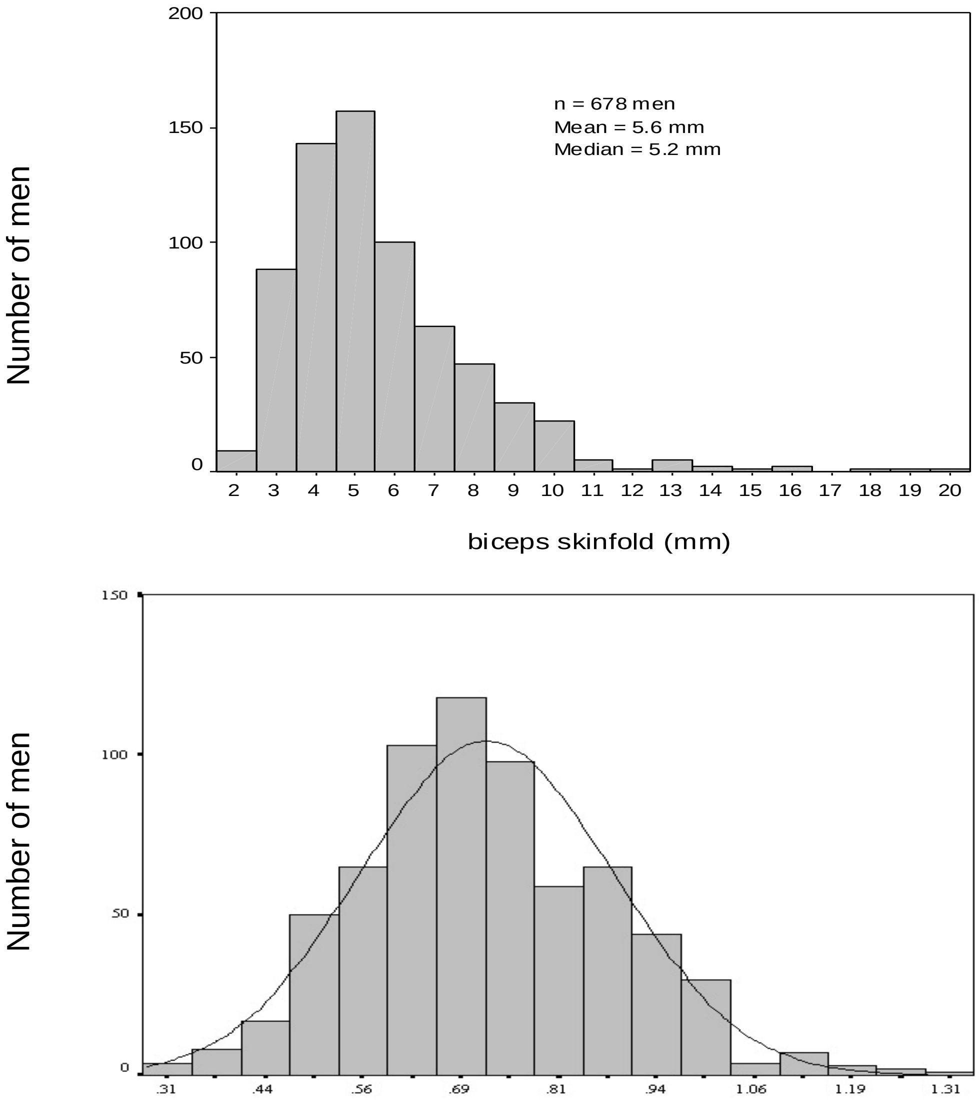


Figure 9. The distribution of the biceps skin fold (upper panel) and after transforming by taking logarithms (lower panel).

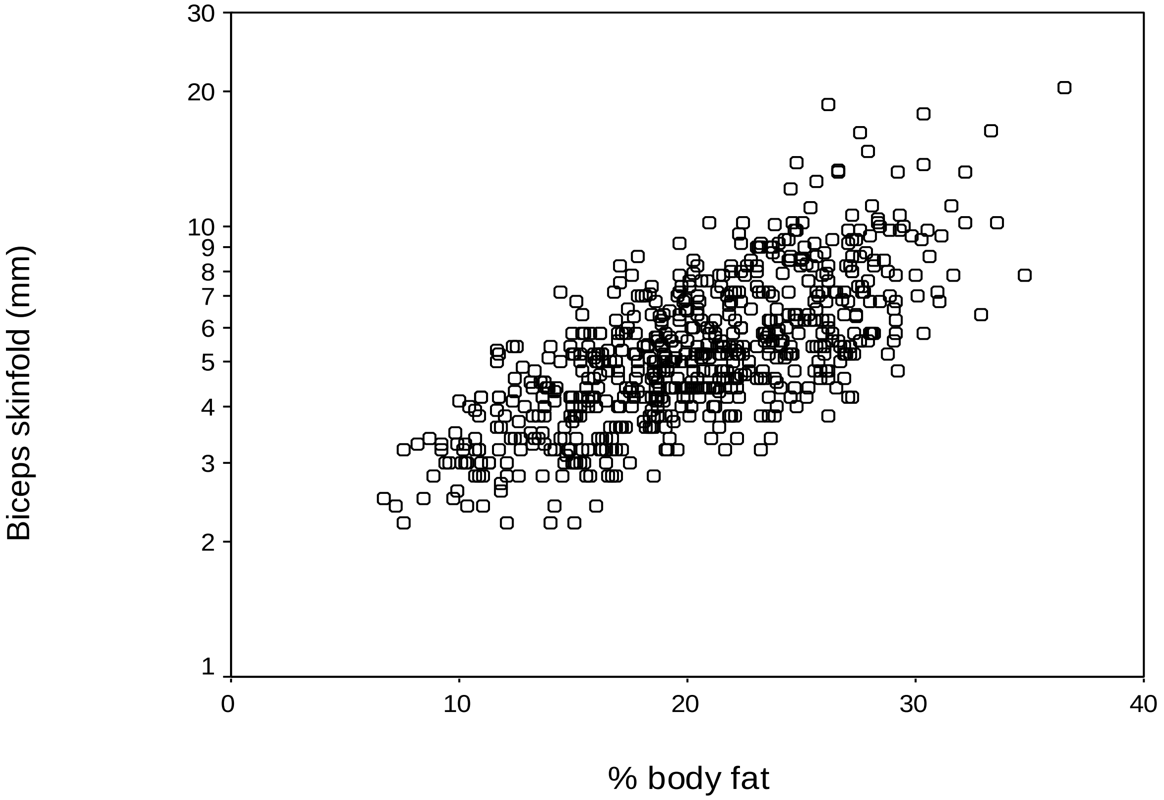


Figure 10. Relationship between %body fat and biceps skin fold thickness (mm) in adult males
(logarithmic vertical scale)

(14.2) Checking for outliers

Check each graph for the presence of outliers (Figure 11). Is it clear how they have been dealt with in the analysis? The presence of one or more outliers can have a marked effect in the analysis on, for example, estimates of regression slopes and correlation coefficients obtained.

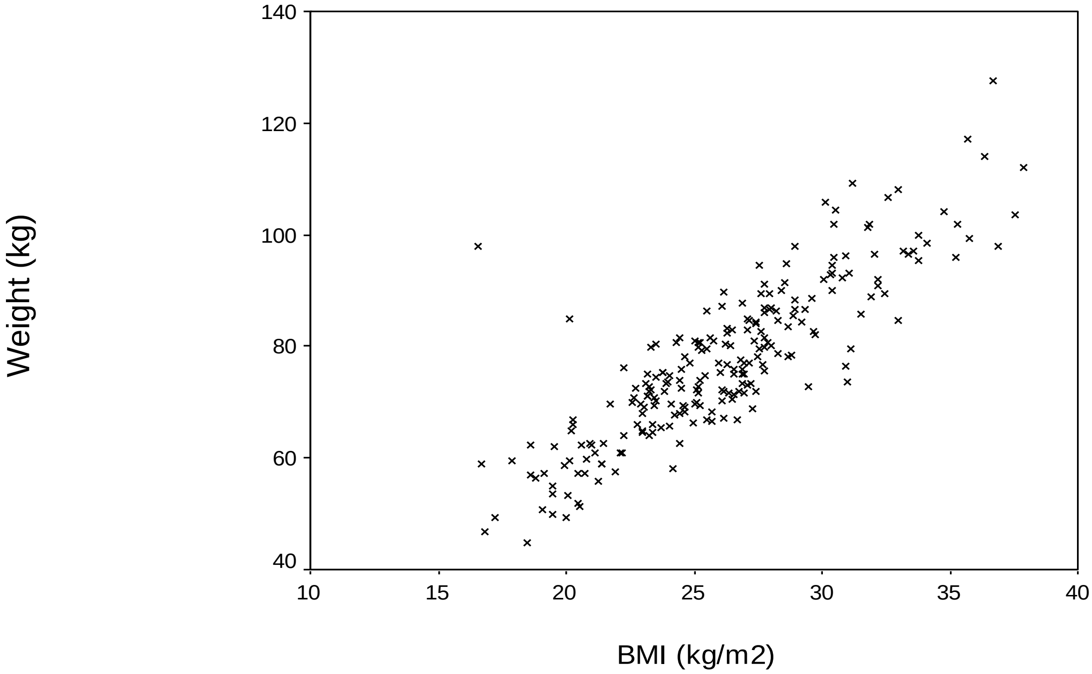


Figure 11. Relationship between Weight (kg) and Body Mass Index (BMI, kg/m^2) in adult males. Note the two outliers.

(14.3) Line graphs, scatterplots, histograms and bar charts

Data can be visually presented in other forms, for example as line graphs and bar charts. Here is an example of some research that used a variety of charts for representing the results.

In the 1990s a National Service Framework on diabetes was released which stated that GPs should actively seek to identify patients with diabetes. An audit was undertaken by the Public Health department in one area asking each practice to provide data on the number of their patients registered with diabetes. However, because, at that time the uptake of computerisation was known to be patchy amongst the practices it remained uncertain just how reliable the data would be.

A data collection scheme was set up to study the local prevalence of diabetes and to identify those practices which had evidence of poor-quality record keeping. Data were collected from five large sentinel practices which had good evidence of high-quality electronic record keeping and use of their clinical software. These five practices represented about 12% of the resident population. The data were pooled and the prevalence of diabetes was calculated for each 5-year age group, by gender and plotted in a line graph (Figure 12). These rates were then applied to the age-sex registers of each practice in that area to provide estimates of the *expected* number of patients with diabetes. This number was then compared in a scatterplot with the *actual* number of patients each practice had returned (Figure 13). Most practices laid along the 'line of identity' where the number of patients recorded was close to the number expected. However, some practices were lying well away from the line of identity.

The difference between the number recorded and that expected for each practice was plotted as a histogram (Figure 14). A few practices had very many more patients than expected whilst others had markedly fewer patients than expected. Remember, these figures were adjusted for age and gender but not for any other features known to influence the prevalence of diabetes (e.g. South Asian ethnicity).

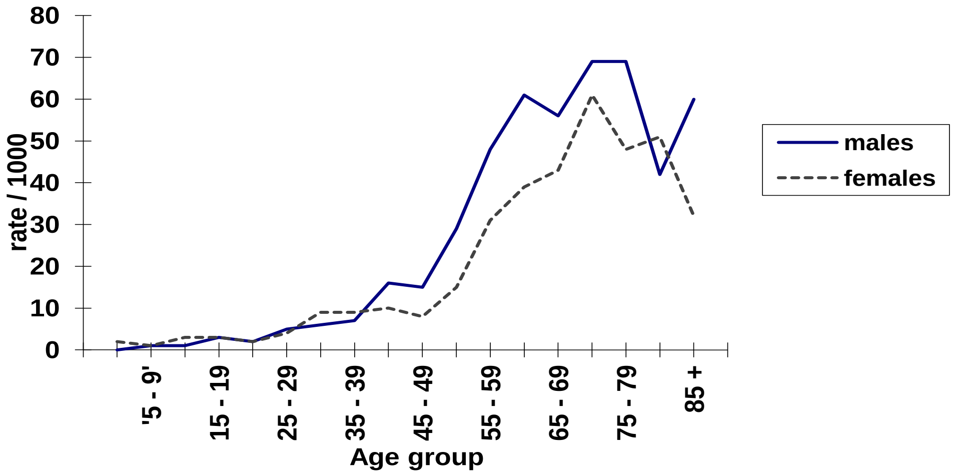


Figure 12. The prevalence of diabetes by age and gender.

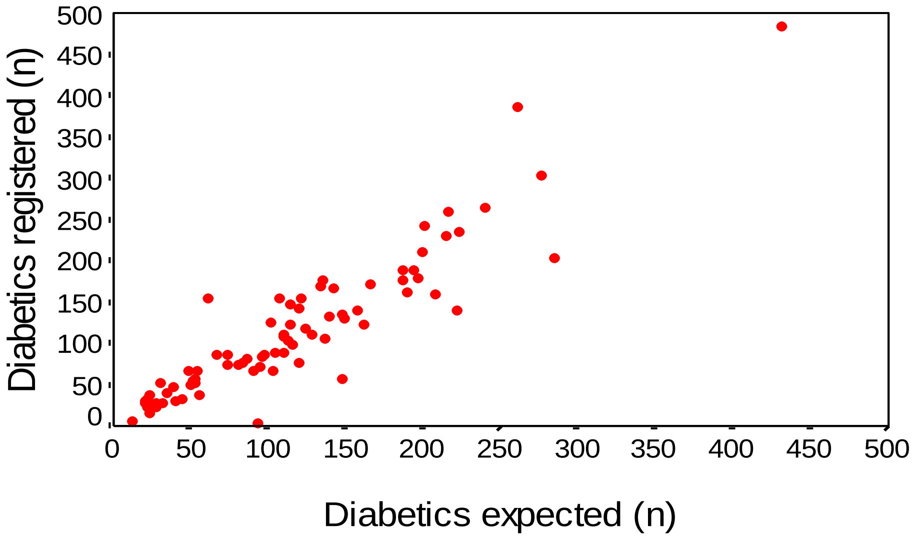


Figure 13. The number of registered patients with diabetes compared with that expected for each practice.

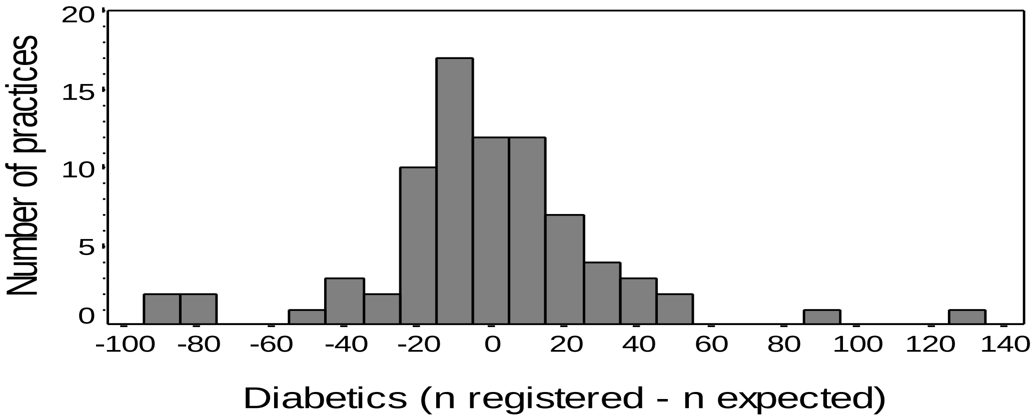


Figure 14. The difference between the number of registered patients with diabetes from that expected for each practice.

Another way of graphing the data was to plot the *standardised morbidity ratio* (SMR) for each practice, that is, the number of patients with diabetes registered divided by the number expected, multiplied by 100. This is a *notional* rate where a value of 100 indicates the practice had registered the exact number expected. The practices were then sorted in order from lowest to highest SMR values and a bar chart generated where each bar represented a practice (Figure 15).

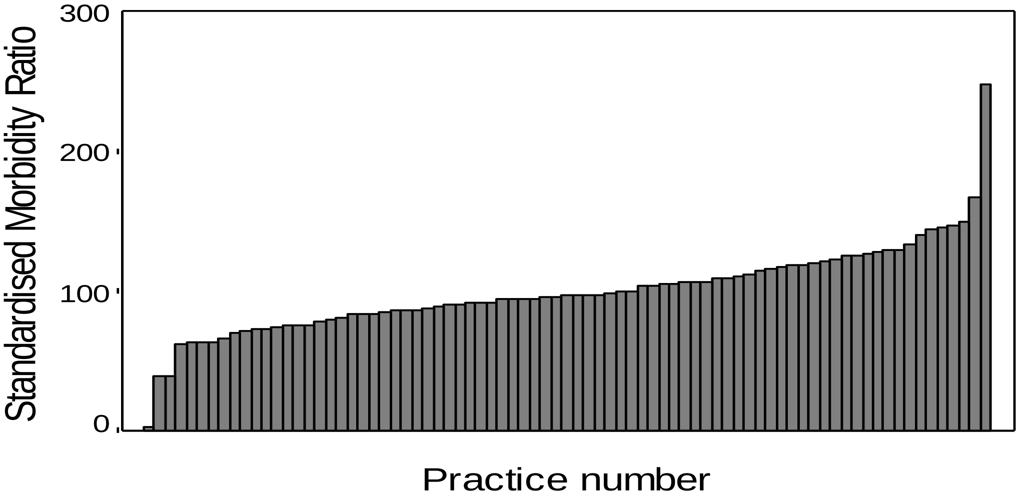


Figure 15. The standardised morbidity ratio for each practice. A value of 100 indicates the number of patients registered is equal to the number expected based on that practice's age-sex register.

At one end of the scale a practice had hardly any patients with diabetes recorded on its computer (SMR much less than 100). At the other end of the scale a practice had an SMR of about 260, or 2-3 times the number expected. What could be the reasons for this variation? How should we interpret these findings?

In these circumstances it is appropriate to first look for technical reasons. In the low recording practices is there evidence of under-recording, or the wrong codes being used for recording diabetes. The practices may not be using their clinical systems to best effect. Are these, perhaps, single-handed practices?

In the higher recording practices is there evidence of over-recording, perhaps with patients recorded with diabetes when there was only a suspicion of the condition, or a family history of it recorded? Again, are the correct codes being used?

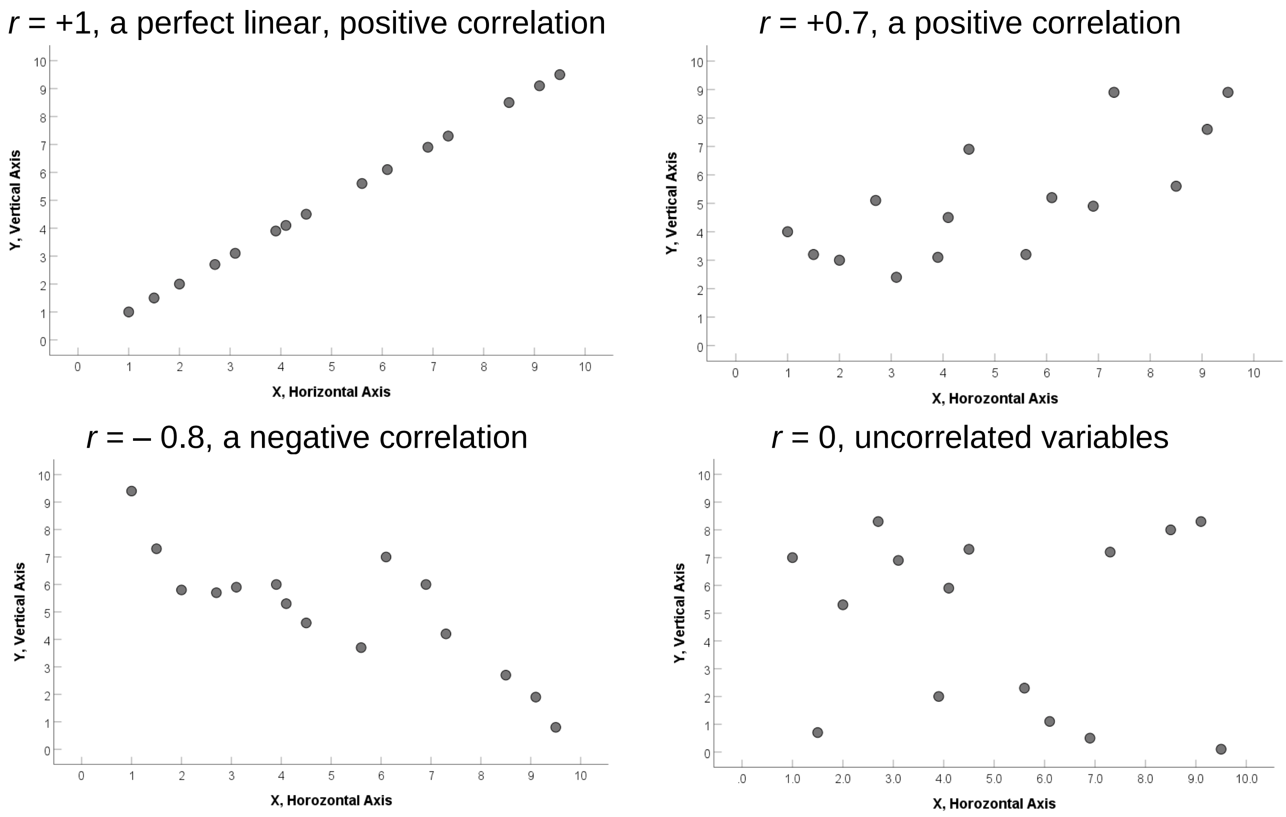
In interpreting the SMR the assumption is that practices in the middle with values around 100 are ‘doing it right’. The data are adjusted for age and gender. However, it may be that the high recording practices have a different ethnic mix of patients. They may, for example, have a disproportionate number of South Asians who are known to be at higher risk of developing diabetes.

However, there is another reason. The estimates from this study suggested a prevalence of diabetes of 1.9% in males and 1.6% in females (all ages). At this time, it was accepted that diabetes was under-recognised and it may be that the practice with an SMR of about 260 had an active screening programme in place to check for diabetes (and pre-diabetes). Hence, their results may reflect the true prevalence of diabetes in the community and all the other practices had got it wrong!

The results of these exercises seldom provide clear cut answers but are useful in identifying the need for further investigations, including individual practice visits to explore differences in results and possible lessons to be learned.

(15) How to Make Sense of the Linear Correlation Coefficient

The relationship between two variables can be visualised in a plot of one variable against the other (*see* Figures 6, 7, 8, 10, 11 as examples). The strength of *linear* relationships, where the data plotted in a scattergram appears to fit around a straight line, can be quantified as the *Pearson correlation coefficient* (*r*) which is dimensionless (no units) and takes the value from – 1 to +1. A negative correlation implies that one variable decreases as the other one increases. A positive correlation implies that both variables increase together. A correlation coefficient of +1 occurs when both variables are perfectly correlated positively. A correlation coefficient of –1 occurs when both variables are perfectly correlated negatively. Some examples are given in Figure 16.



r = +1, a perfect linear, positive correlation

r = +0.7, a positive correlation

r = – 0.8, a negative correlation

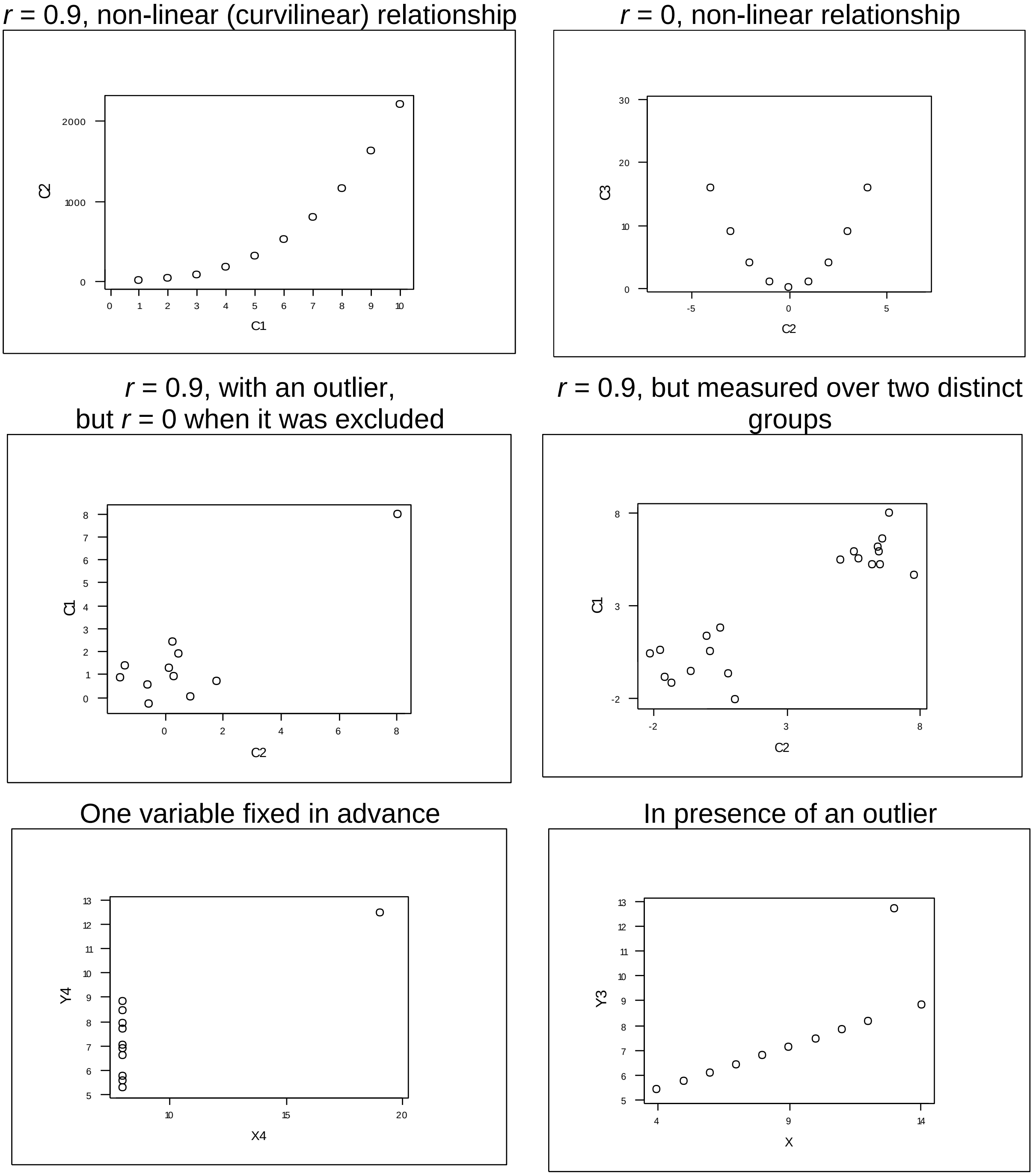
r = 0, uncorrelated variables

Figure 16. Examples of typical correlations and the associated correlation coefficient (*r*)

The Pearson correlation coefficient should not be used:

- if the relationship is non-linear
- in the presence of outliers
- when the variables are measured over more than one distinct group
- when one of the variables is fixed in advance
- for assessing agreement

Examples of the inappropriate use of the correlation coefficient are given in Figure 17.



r = 0.9, non-linear (curvilinear) relationship

r = 0, non-linear relationship

r = 0.9, with an outlier, but *r* = 0 when it was excluded

r = 0.9, but measured over two distinct groups

One variable fixed in advance

In presence of an outlier

Figure 17. Inappropriate use of the correlation coefficient

The Pearson correlation coefficient should not be used when assessing agreement, that is, when trying to decide, for example, how close two sets of measurements are to one another, or when comparing two observers reading the same radiographs. Consider a study to compare two instruments that measure the same characteristic. A series of samples are split and assessed by each instrument. The results are tabulated and the correlation coefficient calculated (Table 8). The table displays the results under three circumstances, (1) where the two instruments have the same calibration which shows perfect consistency and a perfect correlation between them, (2) where the calibration differs for one instrument which records a value that is double that for the other instrument, and (3) where one instrument has a zero error, that is the instrument does not read zero correctly. For examples (2) and (3) the consistency is adrift but the two series of values still retain a perfect correlation.

Table 8. The fallibility of relying on the correlation coefficient when assessing agreement

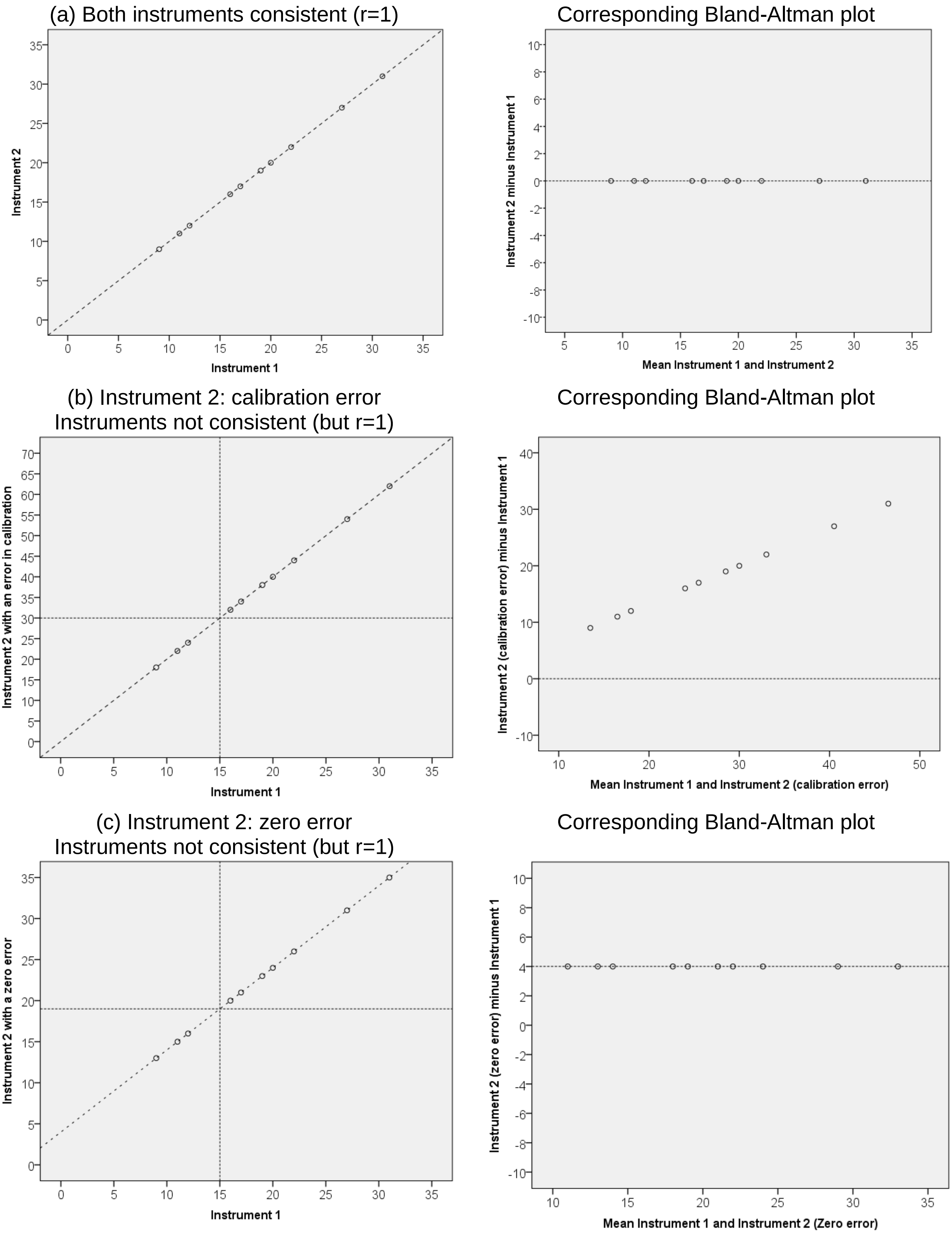
Sample	*Result Instrument 1*	*(1) Result Instrument 2*	*(2) Result Instrument 2 **	*(3) Result Instrument 2 ***
1	12	12	24	16
2	16	16	32	20
3	9	9	18	13
4	31	31	62	35
5	17	17	34	21
6	22	22	44	26
7	11	11	22	15
8	20	20	40	24
9	19	19	38	23
10	27	27	54	31
Consistency:	Both instruments consistent		Instruments not consistent	
Correlation coefficient:	+ 1 (perfect correlation)		+ 1 (perfect correlation)	
			* calibration error	** zero error

The correct method comparison is to use a Bland-Altman plot (formerly called the Oldham Plot, as published by Peter Oldham, a statistician working at the MRC Pneumoconiosis unit at Penarth, South Wales in the 1950s). The difference between the two readings is plotted against the mean of the two readings (Figure 18).

In Figure 18 (a) the two instruments are consistent; the difference between their readings is zero for each pair of observations and the Bland-Altman plot has a flat line across the graph crossing the vertical axis at ‘0’.

In Figure 18 (b) the two instruments are not consistent; the difference between their readings increases with the mean. A value of 15 on instrument 1 coincides with a value of 30 for instrument 2. The Bland-Altman plot shows a positive association between the difference and the mean of the paired observations.

In Figure 18 (c) the two instruments are not consistent; the difference between their readings is consistent (=4, Table 8) but does not increase with the mean. A value of 15 on instrument 1 coincides with a value of 19 for instrument 2. The Bland-Altman plot shows a flat line across the graph but this time it intersects the vertical axis at a value of 4 (the zero error).



(a) Both instruments consistent (r=1)

Corresponding Bland-Altman plot

(b) Instrument 2: calibration error
Instruments not consistent (but r=1)

Corresponding Bland-Altman plot

(c) Instrument 2: zero error
Instruments not consistent (but r=1)

Corresponding Bland-Altman plot

Figure 18. Correlation and Bland-Altman plots for the examples in Table 8.

(16) Method comparison and Repeatability Studies

The data from method comparison studies should be analysed with the Bland-Altman approach. Similarly, within-person repeatability studies, comparing the results in the same person but between different times, should also use the technique. An example is given in Figure 19 from a study in children looking at the repeatability of a new measurement of bone density at the heel (*os calcis*). Fifty-three children had their bone density measured twice on the same day. The difference in repeated measures was unrelated (that is, not correlated) with the mean. In other words, the scatter of the differences did not show an association with the mean of the repeated measurements. In this example, the mean and standard deviation of the difference between repeated measures was 0.0001 and 0.015 g/cm^2; respectively. The 95% limits of agreement between repeat measures were -0.029 to +0.029 g/cm^2.

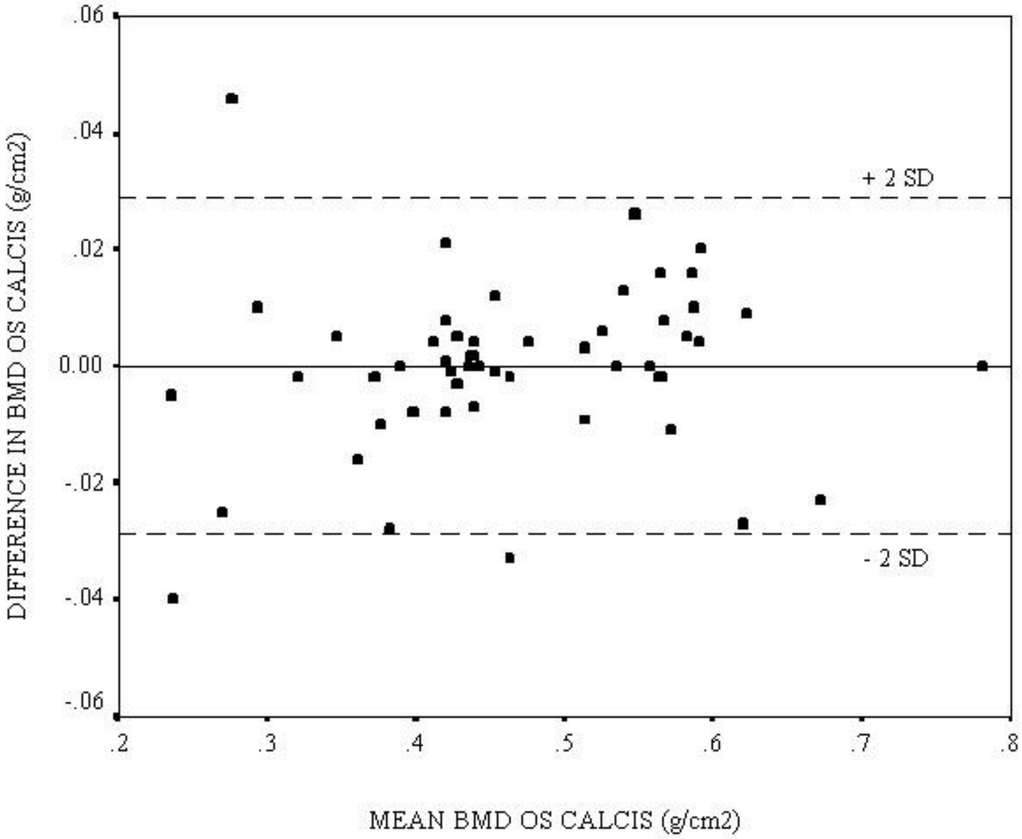


Figure 19. Bland-Altman plot of the difference between repeat measures of bone density at the heel in 53 children.
Reference: Chinn et al. *Arch Dis Child* 2005; **90**: 30-35

For further detail see Bland JM, Altman DG Measuring agreement in method comparison studies *Stat Methods Med Res* 1999; 8; 135-160.

(17) How to Make Sense of an Odds Ratio

The '*odds*' is the ratio of the number of times an event occurs to the number of times it does not occur from a given number of chances. It is used to quantify the 'risk' of something happening. The '*odds ratio*' (OR) is a comparison of odds between two groups to quantify the '*relative risk*' of something happening. If the odds are the same in the two groups the odds ratio is 1.

Consider the road accident statistics from Scotland in 2007 when there were 282 fatalities (Table 9). Suppose we wish to calculate the relative risk of being killed if involved in a motorcycle accident compared to a car accident.

Table 9. The number of casualties and fatalities on Scotland's road, 2007.

	Casualties	Fatalities	% deaths
Car	9953	160	1.6
Pedestrian	2682	61	2.3
Motorcycle	1039	40	3.8
Other	??	21	-
Totals	13,674 ++	282	-

What are the odds of being fatally injured in an accident when riding a motorcycle compared with being in a car?

Table 10. The number of fatal and non-fatal casualties in car and motorcycle accidents, Scotland, 2007.

	Fatal	Non-fatal	Number of casualties
Motorcycle	40	999	1039
Car	160	9793	9953
Totals	200	10792	10992

For motorcyclists 40 out of 1039 casualties were fatal, which means that 999 were non-fatal (Table 10). The odds of being fatally injured if a motorcyclist is therefore 40/999.

For car occupants160 out of 9953 casualties were fatal, which means that 9793 were non-fatal. The odds of being fatally injured if a car occupant is therefore 160/9793.

The ratio of odds = (40/999) / (160/9793) = 2.45 and the 95% CI = 1.72 to 3.48

Interpretation: we are 95% confident that, when involved in a traffic collision a motorcyclist has a relative risk of being fatally injured that is between 1.7 and 3.4 times greater than that of a car occupant.

Detailed calculations to derive the 95% CI of an OR *this part can be omitted*

The ratio of odds = (40/999) / (160/9793) = 2.45 and $\log_e$ (OR) = 0.896

SE ($\log_e$ (OR)) = $\sqrt{}$ (1/40 + 1/160 + 1/999 + 1/9793) = 0.179869

95%CI = 0.896 +/- 1.96 × 0.179869 = 0.5435, 1.248

Take antilogs to get the 95%CI which is 1.72 to 3.48

Odds ratios are commonly derived from case control studies. One such study looked at the relationship between maternal BMI and the risk of stillbirth. Here, a 'case' was a mother who had a stillbirth. She was matched with a woman who had not had a stillbirth ('control'). Cases and controls were matched for characteristics known to be associated with stillbirth, such as parity, age and gestation. Women were grouped according to their BMI and the odds ratio calculated for the relative risk of a stillbirth for women in each BMI category compared to the reference category of 18.5 – 25 kg/m^2. An odds ratio of 1 implies no difference in risk between cases and controls. The results showed that the risk of a stillbirth increased with increasing levels of BMI (Figure 20).

Interpretation: Where the **95%** confidence interval crosses the odds ratio (OR) line of 1.0 the result is not statistically different at the **5%** level (P>0.05). Where the 95% confidence interval does not cross the OR line of 1.0 the result is significantly different between cases and controls with P<0.05. This was the case for women with a BMI between 25 and <30, between 30 and <35, between 35 and <40 and if >=40. The increasing risk associated with an increasing degree of maternal obesity is good evidence of a causal relationship between maternal BMI and the risk of stillbirth.

Note: the length of each confidence interval is not equal on either side of the mean OR (the filled in circle in Figure 20). This is because the calculation of the confidence intervals involves use of a logarithmic function (see the example calculation of a 95% CI for an odds ratio above if you really need to understand this more).

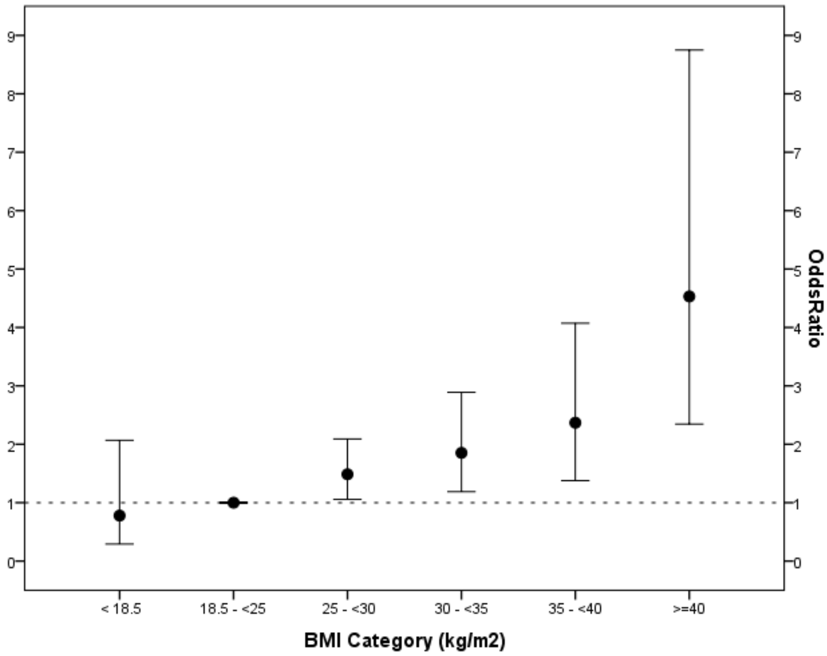


Figure 20. Odds ratio and 95% confidence interval for risk of stillbirth against maternal Body Mass Index (BMI)

(18) Run Charts and Control Charts

A run chart shows the change in a variable or outcome over the course of a period of time. Examples include weekly DNA (‘did not attend’) rates in outpatient clinics, monthly surgical infection rates in a particular hospital and annual stillbirth rates in a maternity unit. They are simple plots, involve no statistics and can be created prospectively as time progresses (Figure 21). A run chart can be created from historical data and include the average value of the measure in question (Figure 22).

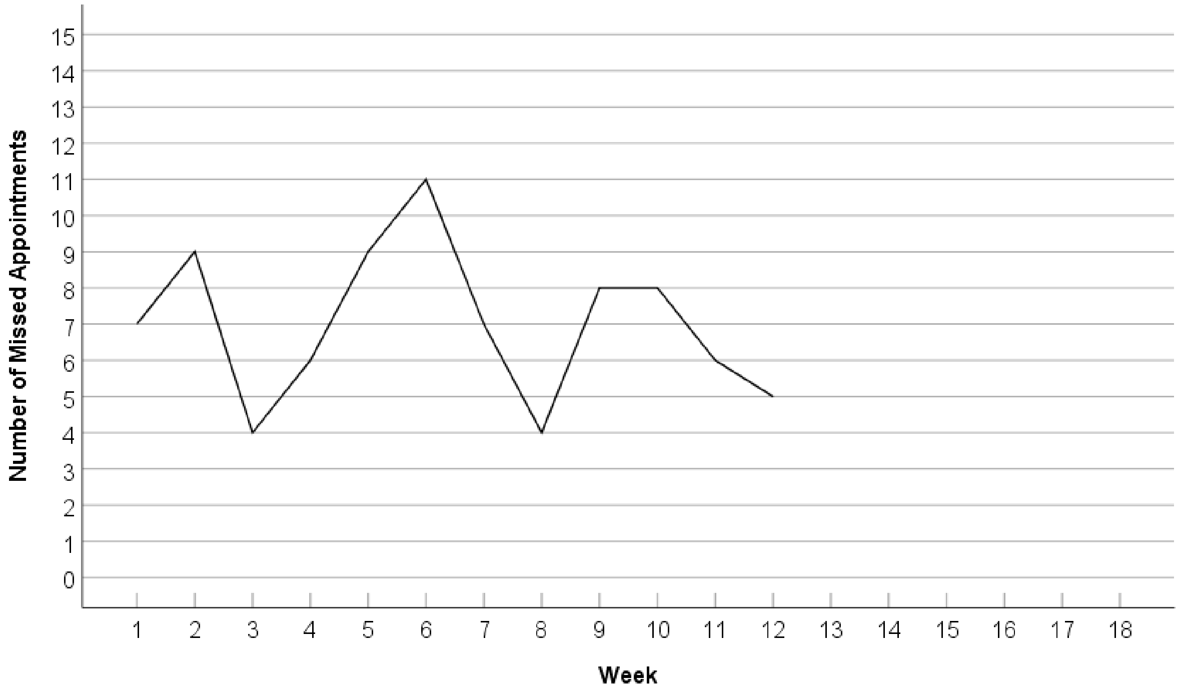


Figure 21. A simple run chart showing the number of missed appointments at an outpatient clinic. The chart can be updated simply each week.

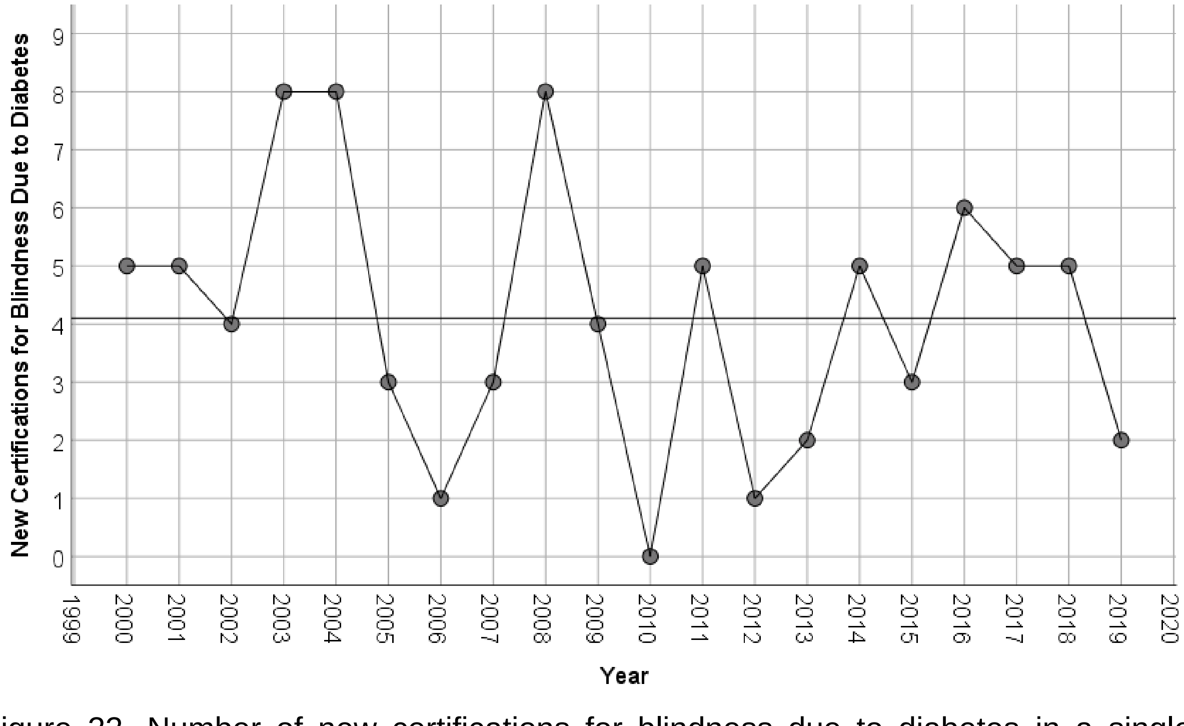


Figure 22. Number of new certifications for blindness due to diabetes in a single Scottish Health Board, 2000 – 2019. The chart includes the average over the period (4.1 per year).

Control charts (also called Shewhart charts) are similar to run charts in that they monitor trends in real time but include a measure of statistical variability and are used in assessing quality control. They help determine when a process can be considered ‘out of control’, particularly early so that special measures can be introduced to prevent further deterioration in the process. The degree of variability can be separated into ‘common cause’ and ‘special cause’ variation. Common cause variation is attributed to the usual, natural changes expected in a process whereas special cause variation suggests the process is out of control, having been influenced by some unusual activity. Examples include hospital acquired infections, patient satisfaction surveys, falls surveillance, hospital mortality rates etc.

Control charts include upper and lower limits describing the statistical variation to be expected (typically, as 2 or 3 standard deviation limits). Rules are needed to define the circumstances when the process is considered out of control. Use of 2 standard deviations as a limit to define ‘out of control’ may result in too many false alerts, whereas use of 3 standard deviations may be over-cautious and miss important events.

There are 7 different versions of control charts depending on the type of data (attribute or continuous). Attribute data refer to discrete, countable events such as the number of surgical complications, the number of prescription errors etc. Continuous data relate to non-discrete measures such as waiting times, length of hospital stay etc.

Example: Each month staff on a medical ward monitored the number of falls and created a control chart of this number as a fraction of bed occupancy using the number of patient-days. The chart was reviewed each month to identify ‘special cause variation’.

Month	Patient-days	N of falls	Fraction (= falls per patient-day)
1	1048	1	1/1048 = 0.000954
2	896	4	4/896 = 0.004464
3	918	3	3/918 = 0.003268
4	995	4	4/995 = 0.004020 …..etc

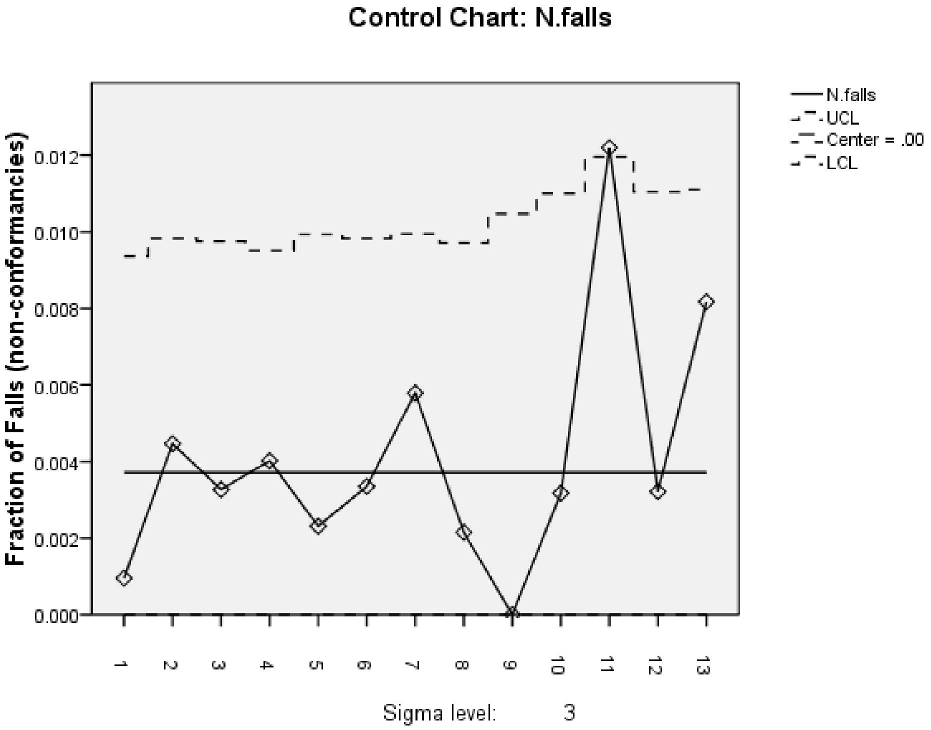


Figure 23. Number of falls as a fraction of the number of patient-days on a ward over 13 months. The solid horizontal line represents the average fraction. The upper dotted line represents 3 standard deviations (sigma level=3). All but one point is within the normal, expected variation (common cause). The fraction for Month 11 exceeded the upper limit and indicated a ‘special cause’. This was picked up early and changes in practice made in month 12 to reduce the incidence of falls.

Another example: It is generally recommended that laboratory workers should not spend more than 2 hours per day engaged in manual, repetitive pipetting. A review of times spent pipetting was done for a hospital laboratory worker engaged in a process to screen biological material.

Month	Days at work	Days where pipetting time was =>2 hours	Proportion
Jan	5	0	0/5 = 0
Feb	20	7	7/20 = 0.35
Mar	23	5	5/23 = 0.22
Apr	19	8	8/19 = 0.42
May	24	8	8/24 = 0.33 … etc

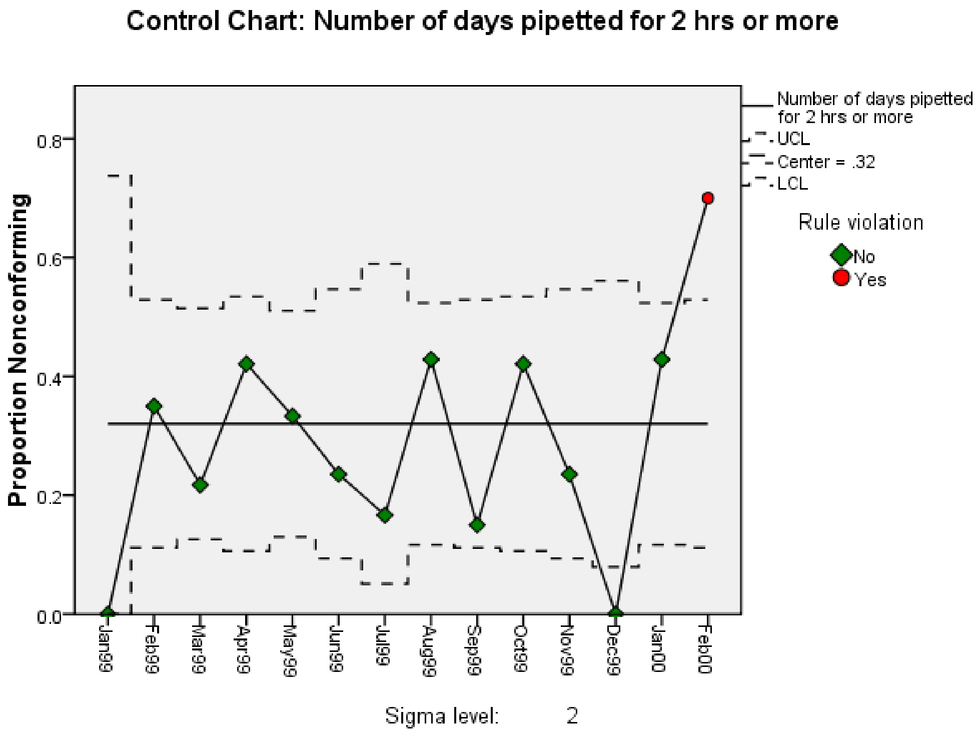


Figure 24. Number of days a laboratory worker was engaged in manual pipetting for more than the recommended 2 hours a day as a proportion of the number of days worked each month, January 1999 to February 2000. The solid horizontal line represents the average proportion. The upper dotted line represents 2 standard deviations (sigma level=2). In February 2000 the laboratory worker experienced wrist pain associated with excessive manual pipetting that occurred because two colleagues who shared the work were on sick leave (special cause). The analysis confirmed that the proportion actually exceeded 3 standard deviations (rule violations).

(19) Funnel Plots

Funnel plots are used to identify publication bias in meta-analyses undertaken as part of a systematic review where the outcomes from multiple, randomised controlled trials are combined in a single assessment of an intervention. Publication bias (more correctly *non-reporting bias*) represents a threat to the interpretation of an analysis when it identifies a failure of publication of papers that, in general, have shown a smaller effect or a negative effect of the intervention under investigation. The effect size of each study is plotted as an odds ratio on the horizontal axis against a measure reflecting that study's sample size or precision (standard error), on the vertical axis. The plot will resemble a funnel in the absence of publication bias. A distorted funnel is evidence that important studies are missing (non-reporting bias). A dependence on published papers that show only a positive effect will overestimate the overall impact of the intervention under review (*see* "Funnel Plots as used in meta-analyses" in Further Reading below for more detail of their use in systematic reviews).

Funnel plots are also used to compare performance between health care providers seeking to identify outliers in an outcome of interest. For example, in 2012 the annual report on maternity outcomes amongst Scottish Health Boards identified NHS Fife as an outlier in the statistics on stillbirths. A funnel plot was created showing the average rate of stillbirth for each Health Board over a 5-year period plotted against the average number of births over the same period (Figure 25).

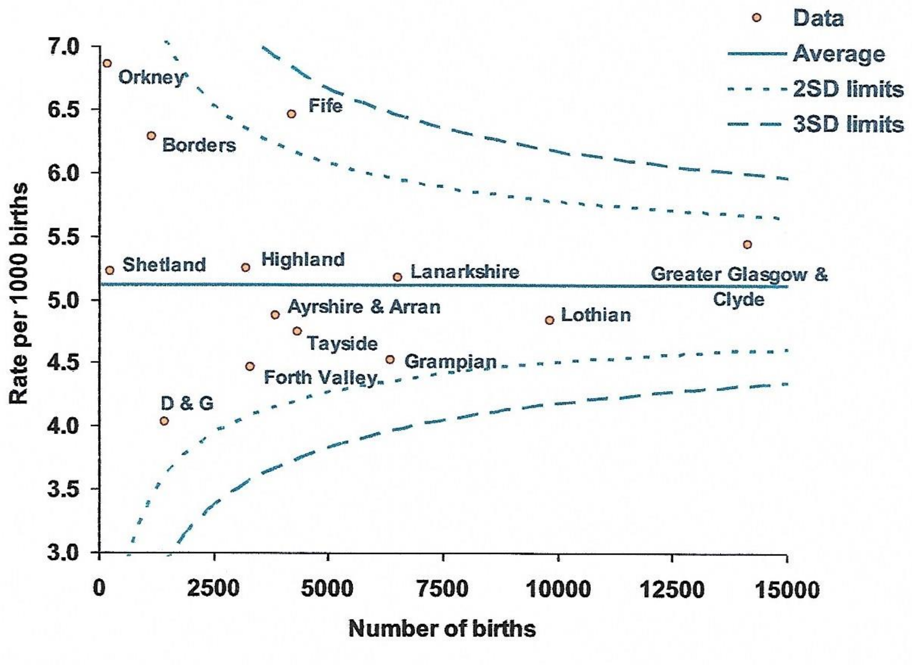


Figure 25. Average Stillbirth Rate (stillbirths/1000 births), 2008 – 2012, by NHS Health Board in Scotland. NHS Fife was an outlier compared to the other Boards, lying more than 2 standard deviations above the Scottish average.
Source: Scottish Perinatal and Infant Mortality and Morbidity Report, 2012. Healthcare Improvement Scotland, published March 2014.

The average stillbirth rate across Scotland is about 5 / 1000 births, or about 1 in 200 births. The number of births varies considerably between Boards and, as expected, is larger in Boards with larger populations. These Boards would also expect to have a larger number of stillbirths but the *rate* of stillbirths should be similar to the average expected. However, the variability, as reflected in the standard deviation, will vary with the number of births, being relatively greater when the number of births is low. Hence, the lines delineating 2 and 3 standard deviations in Figure 25 are curved, or ‘funnelled’. For example, compare the high variability (wider ‘funnel’) of Orkney and Shetland (where the average number of births is low) with the lower variability (narrower ‘funnel’) of Lothian and Greater Glasgow & Clyde (where the average number of births is much higher).

Stillbirth rates in Health Boards vary year to year. The average rate, 2008 – 2012, across the whole of Scotland was 5.08 stillbirths / 1000 births. The highest average rates were for Orkney, Fife and Borders, all above 6 /1000 births (Table 11, Figure 26). The simple bar chart of average rates in Figure 26 identifies these Boards as having excessive mortality. However, stillbirths are, fortunately, rare events and rates based on small numbers can be misleading (*see* section 3 above). The funnel plot allows a comparison of Boards taking into account the difference in variability arising from a difference in the average number of births between Boards. Hence, Orkney and Borders were not *statistical* outliers though Fife was, due mainly to particular high rates in 2008 and 2010 (Table 11).

Table 11. The Stillbirth Rate, 2008 – 2010, by NHS Health Board.

NHS Board	2008	2009	2010	2011	2012	Average
Scotland	**5.4**	**5.3**	**4.9**	**5.1**	**4.7**	**5.08**
Ayrshire & Arran	4.3	4.6	5.2	4.4	5.9	4.88
Borders	7.0	7.7	6.0	5.4	5.2	6.26
Dumfries & Galloway	4.9	5.3	5.5	0.7	3.6	4.00
Fife	6.9	5.5	8.3	5.6	5.9	6.44
Forth Valley	5.2	3.3	4.8	5.0	4.0	4.46
Grampian	4.9	4.5	4.0	4.1	5.1	4.52
Greater Glasgow & Clyde	5.6	5.4	5.0	6.6	4.5	5.42
Highland	4.2	7.2	3.7	4.8	6.3	5.24
Lanarkshire	5.2	6.6	5.6	4.4	3.9	5.14
Lothian	4.8	4.8	4.9	4.6	5.1	4.84
Orkney	13.8	10.0	0	0	9.9	6.74
Shetland	0	14.1	0	12.2	0	5.26
Tayside	7.4	5.2	3.0	5.2	2.6	4.68
Western Isles	0	0	0	0	0	0

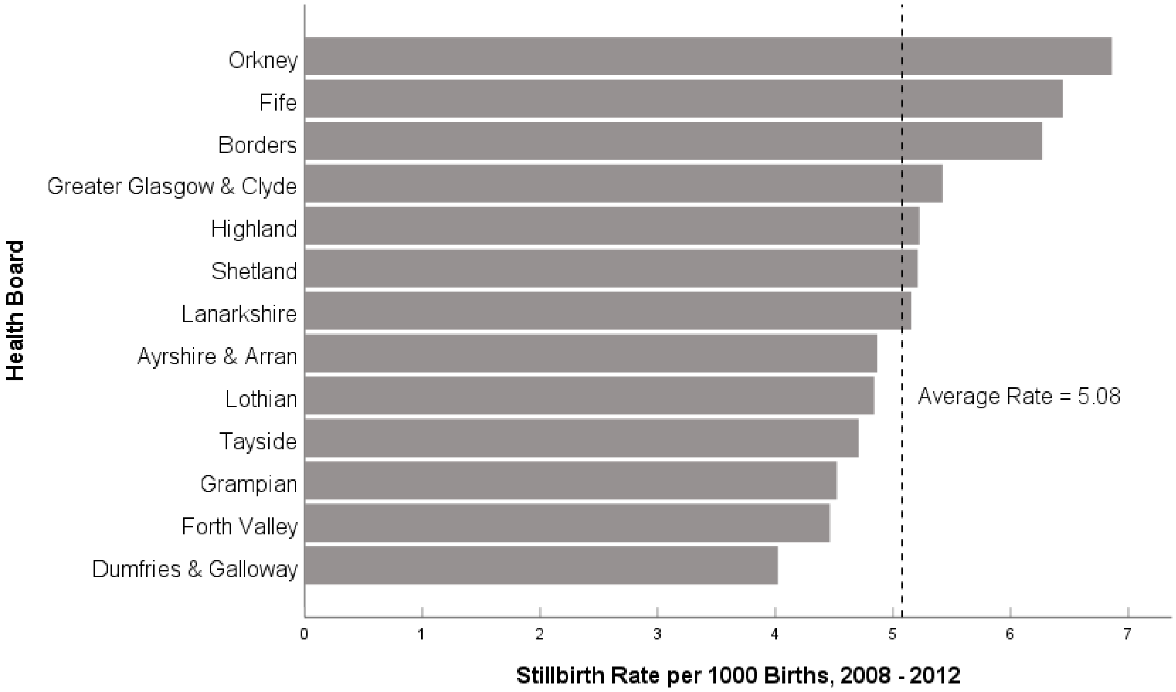


Figure 26. Bar chart of the average stillbirth rate (stillbirths/1000 births) by Health Board, Scotland, 2008 – 2012.

(20) Common Pitfalls in Published Statistics

Generally, it is safe to assume that a paper that includes a statistician amongst the authors will be robust in its study design, analysis of the data and interpretation of the findings. When this is not the case you should carefully read the methods and results sections to look for common mistakes. For example,

(a) Is there a power analysis to justify the choice of the sample size?

(b) Have the authors made a statement about the treatment of missing data or outliers?

(c) Have the authors checked the distribution of the data and used appropriate tests of significance? Parametric tests such as the t-test are used for data that are Normally distributed (bell-shape). Non-parametric tests such as the Mann Whitney test are used for data that are not bell-shaped in distribution (some details of this and other non-parametric tests are in the NHS Fife Study Guide How to choose a statistical test). If using a t-test do the data meet all the assumptions? (*see section 9, page 12 above*)

(d) If the data require a paired analysis (e.g. a before and after study) has the appropriate t-test been used? Remember, use of a paired analysis tests the hypothesis that the mean *change* does not differ from zero. Although the initial and final values may not be Normally distributed it is often the case that the change is Normally distributed and use of a parametric, paired t-test is safe.

(e) Have the authors used a *two-tailed* or *one-tailed* test of significance? A two-tailed test will test for changes in either direction whereas a one-tailed test only tests for an effect in one direction. Generally, you should use a two-tailed test and not assume the effect of some intervention will only ever be in one direction. As an example, a leaflet designed to allay fears in women from receiving an invitation to have a repeat smear test might actually increase anxiety! If, however, the effect of an intervention can only be in one direction then you should use a one-tailed test. For example, in a study of an intervention to increase fertility in infertile couples you can only increase fertility, not reduce it.

(f) In comparing change over time in two or more groups have the authors adjusted for any initial differences between groups?

(g) In a randomised controlled trial have the authors used an *intention to treat analysis?* This is where the patients' data are analysed in the groups to which they were originally assigned.

(h) Have the authors presented confidence intervals along with P-values?

(i) Have the authors reported correlation coefficients? If so, is use of the *linear* correlation coefficient justified? Remember, correlation does not imply causation so have the authors been circumspect in interpreting statistical associations between variables?

(j) Have the authors adjusted for multiple testing using, for example, the Bonferroni correction (*see* Glossary)? The more tests / comparisons you run on a set of data the more likely you are to obtain some spurious findings just by chance alone.

(k) Have the authors reported subgroup analyses and, if so, are they justified and appropriately powered?

(21) Summary

Making sense of numbers can be challenging if you lack the necessary confidence. Whatever work you do as a health professional you will be presented with numerical information and be expected to understand it. The simple message is to be very careful when given such data and 'think beyond the numbers'. Alternative explanations may exist so be cautious in blindly accepting the authors' interpretation and decide for yourself if the numbers justify the conclusions. Remember, no amount of clever statistics can salvage a badly designed, biased study. Hopefully, this account will help you develop that confidence needed. Now go back to the statements made in the introduction (page 2) to see if you have a better grasp of the content!

(22) Further Reading

A-Z of Medical Statistics. Pereira Maxwell F. 1998, Arnold.

An Introduction to Medical Statistics. 3rd ed. Martin Bland, 2000, Oxford Medical Publications.

Essential Medical Statistics. 2nd ed. Betty Kirkwood & Jonathan Sterne, 2003, Blackwell Scientific Publications.

Essential Statistics for Medical Examinations. 2nd ed. Brian Faragher and Chris Marguerie, 2005, PASTEST

Funnel Plots as used in meta-analyses. *See* Page MJ, Higgins JPT, Sterne JAC. Chapter 13: Assessing risk of bias due to missing results in a synthesis. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA (editors). Cochrane Handbook for Systematic Reviews of Interventions version 6.0 (updated July 2019). Available from www.training.cochrane.org/handbook.

Interpreting Statistical Findings. A guide for health professional and students. Walker J, Almond P. 2010. Open University Press.

Medical Statistics at a Glance. 4th ed. Aviva Petrie & Caroline Sabin, 2019, Blackwell Publishing.

Practical Statistics for Medical Research. 2nd ed. Douglas G Altman, 2011, Chapman and Hall.

Statistical Questions in Evidence-Based Medicine. Martin Bland & Janet Peacock, 2000, Oxford Medical Publications.

The Art of Statistics. Learning from data. David Spiegelhalter, 2019, Pelican Books.

(23) Glossary Sources: adapted from A-Z of Medical Statistics. Pereira Maxwell, and Medical Statistics at a Glance. 3rd ed. Aviva Petrie & Caroline Sabin (*see Further reading*).

Bonferroni correction

A procedure for adjusting the P-value in a statistical analysis involving multiple significance testing. When testing, for example, 20 different measures between two groups it is likely that at least one measure will differ statistically at the 5% level by chance alone and may not represent a true difference between those groups.

Chi-squared test

A significance test for comparing two or more proportions from independent groups. The observed proportion in each group is compared with the expected proportion based on a null hypothesis.

Confidence interval, CI

A range of values in which the true mean for a population is likely to lie. It usually has a proportion assigned to it (for example 95%) to give it an element of precision.

Continuous variable

A numerical variable which can theoretically take any value within a given range (for example, height, weight, blood pressure).

Control Chart

A tool used for quality control in which a measure reflecting a process is plotted against time with statistical limits imposed on the chart to identify unusual causes of variation in performance.

Correlation coefficient (Pearson's)

A measure of the linear association (a straight line in a scatter plot) between quantitative or ordinal variables.

Data cleaning

The process of trying to find errors in the data set.

Database

A systematised collection of data that can be accessed and manipulated by a stats package such as SPSS.

Degrees of freedom

A concept used with statistical tests that refers to the number of sample values that are free to vary. In a sample, all but one value is free to vary, and the degrees of freedom is then N-1. For example, consider a set of four values with the mean of 5 and a sum of 20. If you are asked to 'invent' the individual four values then you are only 'free' to invent three of them as the fourth must ensure the sum adds to 20 (note, it can be a negative number).

Effect size

A standardised estimate of the treatment effect calculated by dividing the estimated difference between two groups by the standard deviation of the measurements (means or proportions). In the context of power calculations the effect size is the same as the standardised difference (*see below*).

Funnel Plot

A tool to identify publication bias in meta-analyses (where the effect size of each study is plotted as an odds ratio against a measure reflecting that study's sample size or precision) and to identify statistical outliers when comparing performance between health care providers.

Frequency distribution

A display of data values from the lowest to the highest, along with a count of the number of times each value occurred.

Heteroscedasticity

Unequal variances between two or more subgroups

Histogram

A graphic display of data frequency using rectangular bars with heights equal to the frequency count.

Homoscedasticity

Equality of variances within two or more subgroups

Hypothesis

A statement of the relationship between 2 or more study variables. *See Null Hypothesis*

Logarithm

The logarithm of a number is the exponent (power) to which another fixed value, the base, must be raised to produce that number. For example, the logarithm of 1000 to base 10 is 3 because 10 to the power of 3 (10^3 = 10 x 10 x 10) is 1000

Margin of error

A term used by pollsters to estimate the error from a survey of opinions. In this account it is a range of values equivalent to twice the standard error on either side of the estimated population mean. It is equivalent to the 95% confidence interval.

Mean

The average value or measure of central tendency. The mean is obtained by dividing the sum of values by the total number of values.

Median

Middle value when data are ordered. The value that splits the sample in two equal parts.

Meta-analysis

A statistical analysis whereby results from individual studies in a systematic review are combined to produce an overall effect of interest.

Mode

The value that occurs most frequently.

Non-parametric

Refers to data and tests of significance which makes no assumptions about the distribution of the data. Data that are skewed in distribution (to the right or left) are described as non-parametric.

Normal (Gaussian) distribution

A continuous probability distribution that is bell-shaped and symmetrical; its parameters are the mean and variance.

Null Hypothesis, H_O

The statement that assumes there is no difference between two populations being compared, or no relationship or association between two variables in a population. An experiment may be undertaken to see if H_O can be rejected in favour of an alternative hypothesis, H_A.

Outlier

Values in a set of observations which are much higher, or lower, than the ‘average’ and lie well away from the rest of the data (in the tail of the distribution).

Parameter

A measurable characteristic of a population (e.g. average and standard deviation of blood pressure for a group of individuals).

Parametric

Refers to data in which the distribution is bell-shaped (Normal or Gaussian). Statistical tests that rely on data being distributed this way are called parametric tests.

Power

The probability of rejecting the null hypothesis when it is false.

Power calculation

Refers to a way of calculating the number of subjects needed for the results of a study to be considered statistically significant.

Protocol

A full written description of all aspects of a study – the ‘recipe’.

Publication Bias

The tendency for journals to preferentially publish papers citing mainly positive (statistically significant) findings.

P-value

See Significance Level

Regression coefficient

The slope of the line of best fit in a plot between two variables. It represents the increase in an outcome variable from a unit increase in the predictor variable. For example, in a plot of total lung capacity against height in women the regression coefficient is 6.60 litres/metre which means that for every increase in one metre in height the lung capacity increases by 6.60 litres.

Significance level (P-Value)

In the context of significance tests, the P-value represents the probability that a given difference (or a difference more extreme) is observed in a study sample (between means, proportions etc) when in reality such a difference does not exist in the population from which the sample was drawn. In effect it’s the probability of getting a wrong answer by deciding that two populations differ in some way when in fact they do not. In statistical parlance, it is the probability of rejecting a null hypothesis of no difference between two populations when in fact the null hypothesis is true.

Spreadsheet

A computer program (e.g. Excel) that allows easy entry and manipulation of figures, equations and text. It displays multiple cells that together make up a grid consisting of rows and columns, each cell containing either text or numeric values or a formula that defines how the contents of that cell is to be calculated. Spreadsheets are frequently used for financial information because of their ability to re-calculate the entire sheet automatically after a change to a single cell is made.

Standard deviation, SD

A measure of variability of data. The standard deviation is the average of the deviation of individual values from the mean measured in the same units as the mean.

Standard error (of the mean), SE

A measure of precision of the sample mean. Estimates of a population *mean* value will vary from sample to sample. The distribution of these values is called the sampling distribution. The SE is the ‘standard deviation’ of this distribution.

Standard score (z-score)

Refers to how many standard deviations away from the mean a particular score is located.

Standardised difference

A ratio equal to what is considered the clinically important treatment difference divided by the standard deviation of the measure in question.

T-test

A statistical test used to determine if the means of 2 groups are significantly different.

Type I error (alpha error)

The probability of making the wrong choice by rejecting a null hypothesis when it is true. In other words, a type I error occurs when it is concluded that a difference between groups is not due to chance when in fact it is (reject a true null hypothesis).Also relates to the significance level (P-value).

Type II error (beta error)

The probability of making the wrong choice by accepting a null hypothesis when it is false. In other words, a type II error occurs when it is concluded that differences between groups were due to chance when in fact they were due to the effects of the independent variable (accept a false null hypothesis).This probability becomes smaller with increasing sample size.

Variable

Any quantity that varies (e.g. blood pressure).

Variance

A measure of variability of data equal to the square of the standard deviation.

Z-score

A standard score, expressed in terms of standard deviations from the mean.