Radiography 17 (2011) e1–e6

Contents lists available at ScienceDirect

Radiography

journal homepage: www.elsevier.com/locate/radi

Review Article

An introduction to inferential statistics: A review and practical guide

Gill Marshall*, Leon Jonker

Faculty of Health, Medical Sciences and Social Care, University of Cumbria, Lancaster LA1 3JD, UK

ARTICLE INFO

Article history:
Received 29 May 2009
Received in revised form
23 November 2009
Accepted 22 December 2009
Available online 29 January 2010

Keywords:
p values
Confidence intervals
Parametric data
Non-parametric data
Evidence based practice
Type I and type II errors

ABSTRACT

Building on the first part of this series regarding descriptive statistics, this paper demonstrates why it is advantageous for radiographers to understand the role of inferential statistics in deducing conclusions from a sample and their application to a wider population. This is necessary so radiographers can understand the work of others, can undertake their own research and evidence base their practice. This article explains *p* values and confidence intervals. It introduces the common statistical tests that comprise inferential statistics, and explains the use of parametric and non-parametric statistics. To do this, the paper reviews relevant literature, and provides a checklist of points to consider before and after applying statistical tests to a data set. The paper provides a glossary of relevant terms and the reader is advised to refer to this when any unfamiliar terms are used in the text. Together with the information provided on descriptive statistics in an earlier article, it can be used as a starting point for applying statistics in radiography practice and research.

Introduction

As health care professionals, radiographers are commonly exposed to statistics; the application of statistics is important to clinical decision making.[1–4] In a study undertaken by Welch and Gabbe of papers from the American Journal of Obstetrics and Gynaecology, more articles, 31.7%, were classified as having inappropriate testing than those 30.3% using appropriate testing.[5] Furthermore in an article by Reznick it was shown that someone only understanding descriptive statistics has only 44.5% access to the data presented, whilst those with a more advanced repertoire (i.e. knowledge of *t*-tests, contingency tables, and some non-parametric tests) increases the access rate to 80.5%.[6] Hence, education of writers of research on how to undertake and present statistical testing must take place such that the work is accessible to readers, which 90% of Reznick's sample concurred with.[6,7] It is found to be most effective in the teaching of these statistics if the learners first understand the concepts behind statistics prior to being taught "how to undertake" them.[8] This is a daunting task but one that is essential in order that e.g. radiographers may gain the best value from statistics.[4,9,10] It is important in teaching statistical testing that it appears relevant to the students e.g. by testing data relevant to clinical decision making.[11]

Many statistical methods used by researchers do not only describe the data but also enable conclusions to be drawn about the populations from which the samples are taken. They can be applied to compare two or more samples with each other to investigate potential differences and they can also be used for studying the relationship between two or more variables. Such statistics are inferential statistics, which are used to infer from the sample group generalisations that can be applied to a wider population. This allows the detection of large or even small, but important differences, in variables or correlations between variables that are relevant to a particular research question.[1–3] It is of paramount importance to ensure that the sample that has been selected is representative and this is determined predominantly by an appropriate sampling method.[2] If, for instance, a researcher wants to test two different courses of radiotherapy as a cancer treatment, by testing a long course versus a short course, then there are different ways to decide on who is selected for this study. There are also different ways to decide who receives which treatment. When deciding who will be sampled (selected) and received (allocated) which treatment, if one wants to be certain that the characteristics of subjects – in this case patients – does not affect the chance of being allocated to a certain group then both random selection and allocation should be applied. Random selection and random allocation, generally also known as randomization, can be achieved by simply tossing a coin or by using a computer programme (www.randomizer.org). What is important here is that the patients are selected, for example from a patient list, by chance and the patients are also allocated to one of the treatments by chance. Often

* Corresponding author. Tel.: +44 1524 384384x2246; fax: +44 1524 384385.
E-mail address: gill.marshall@cumbria.ac.uk (G. Marshall).

1078-8174/$ – see front matter © 2010 Published by Elsevier Ltd.
doi:10.1016/j.radi.2009.12.006

randomization is not feasible, for example when certain exclusion criteria apply in a controlled trial or when only patients from a certain region or country are included in a study. Bias can be introduced in instances where the selection or allocation is not truly randomized. Careful consideration of the sample used and the way that subjects are compared with each other is therefore warranted, well before any statistical test is applied to the data generated.[2,11]

Terms used in inferential statistics

Before covering the different tests that can be applied for data sets and measurements, it is essential to introduce some of the common terms used. Inferential statistics measures the significance, i.e. whether any difference e.g. between two samples is due to chance or a real effect, of a test result. This is represented using *p* values. The type of test applied to a data set relies on the sort of data analysed, i.e. binary, nominal, ordinal, interval or ratio data; the distribution of the data set (normal or not); and whether a potential difference between samples or a link between variables is studied. Readers can refresh themselves regarding the terms nominal, ordinal and interval/ratio data and discrete and continuous variables at this point, by consulting the earlier article in this series.

p values and power

A *p* value is the product of hypothesis testing via various statistical tests and is claimed to be significant (i.e. not due to chance) most commonly when the value is 0.05 or less. The value 0.05 is arbitrary; it is simply a convention amongst statistician that this value is deemed the cut off level for significance.[12] The *p* value that is considered significant can be and is often varied by the researchers, and it demonstrates the probability of making a type I error i.e. an error created by rejecting the null hypothesis i.e. that there is no difference in the amount of contrast induced-nephrotoxicity between two contrast agents, when it is in fact true i.e. that there is no difference between the two agents in this regard. Thus it concludes there is a difference, when this is not true, giving a false positive result.[13] The *p* values say nothing about what the size of an effect is or what that effect size is likely to be on the total population. Often, *p* values can be misleading because even the smallest effect e.g. in the use of two contrast agents, can be made significant by increasing the sample size. Equally a large and important difference can prove to be of no significance if a small sample size is used. The latter is a false-negative (Type II) error, where the null hypothesis is accepted when in fact it should be rejected. Many authors now report the confidence interval as well as the *p* value, or indeed on its own, and some authors go as far as saying that a paper reporting only a *p* value is suspect.[7,14]

The type II error is implicated in deciding on the sample size used for a study. The required sample size can actually be calculated before conducting a study. In order to do so, using freeware programmes such as GPower or other available programmes, a number of parameters need to be predetermined. The type of statistical test that is planned needs to be known and also the effect size. The effect size is by how much you anticipate or hypothesize outcome measurements, e.g. survival, side-effect incidence or tumour volume reduction in two types of radiotherapy treatments, to differ between each other. Then the allowed type I error, the *p* value that is considered significant in this study, and the power of the study (1 − type II error), need to be determined. As mentioned previously, the *p* value is often set at 0.05; power can be anything from 0.80 (80%) to 0.99 (99%) depending on requirements. Undertaking such calculations will avoid too many subjects being recruited for a study. A drawback is that a-priori power calculations do not take variation of data into account, although this can be accounted for to some extent by estimating the sample size for a number of effect sizes and power levels.

Confidence interval and standard error

If the total population of e.g. patients suffering contrast-induced nephrotoxicity, were studied, then that statistic would also be the population parameter. The confidence interval is a measure of the researchers' uncertainty in the sample statistic as an estimate of the population parameter, if less than the whole population is studied. This is because the study result is not precisely the effect in the population; consequently it is necessary to estimate a range of values within which the true or precise effect in the whole population probably lies. This is 'probably', because it is never certain that this is in fact the case. Generally, if there is a 95% probability that something is the case, then that is generally accepted as good enough to be accepted. Again, as with the *p* value, confidence interval levels are usually set at 95% by convention. A 95% confidence interval is the estimated range of values within which it is 95% possible or likely that the precise or true population effect lies.[7,13] Confidence intervals are a pivotal tool in evidence-based practice, because they allow study results to be extrapolated into the relevant population. In the calculation of this, three elements are considered:

1) The standard error− whilst standard errors shrink with increasing sample size, the researcher should be seeking to reach an optimal sample size, rather than the maximal sample size. Testing more subjects than required in a clinical trial may not be ethical and in addition it would be a waste of money and resources.
2) The mean and the variability, i.e. standard deviation, of the effect being studied − the less variability in the sample, the more precise the estimate in the population and therefore a narrower range.
3) The degree of confidence required − the more confident someone wants to be in the obtained results, the higher the confidence interval needs to be. In other words, if a 99% confidence interval is desired then the range will have to be wider, to cover the extra data that needs to be covered over and above the arbitrary 95%, to ensure that it is possible to be more confident that the average for the population (the population mean) lies within it.[7,15] Conversely, if a 90% confidence interval is considered sufficient then the range of data required will be narrower, and hence the required sample size will be smaller.

To further illustrate the above, one needs to appreciate the formulae associated with calculating confidence intervals. If a large sample size is used and the distribution of the measurements/observations shows a normal distribution, we can assume that the variance in the sample is representative of the variance in the wider population. The formulaic relationship between the confidence interval (CI) standard error (SE) in such instances is: 95% CI = sample mean ± 1.96 SE. The standard error of a sample equals the standard deviation of the sample divided by the square root of the sample size ($SE = SD/\sqrt{n}$). Therefore, if the standard deviation is large and the sample size small, the standard error will be large, and vice versa. The standard deviation and standard error are very easily mixed up. A standard deviation says something about the variation around the mean of the actual sample, whereas the standard error gives information on the variation one could theoretically expect in the sample mean if the experiment was repeated with a different sample. The confusing part is that the standard error is actually a measurement of variation − it is the variation of

the mean itself (e.g. average percentage of nephrotoxicity cases in a group of patients), not around the mean, that you can expect if you would repeat your experiment with a different group of patients.[16,17]

Data amenable to the use of parametric statistics

Firstly, please note that data should not be subjected both to parametric and non-parametric testing. It is necessary for the researcher to decide which method will be most appropriate given the level and distribution of the data, and the types of statistical tests that are that can be performed (see previous article). If data shows a normal distribution, then tests such as one way analysis of variance (ANOVA) and the paired and unpaired *t*-tests can be applied for ratio and interval data. If parametric statistical testing is used, the following characteristics of the data need to be checked to ensure they fulfil the requirements below. Deciding which test to apply is also subject to personal preference and the researcher's own judgement. Therefore, the pointers below can be used for guidance and are certainly not exhaustive.

Generally speaking data should:

- Be approximately normally distributed. As usually it is not known whether a whole population that is being sampled is normally distributed the sample data must be examined when considering this question. Data with an underlying normal distribution exhibits a bell-shaped frequency distribution, so it is useful to create a frequency histogram of each sample. Most statistical programs can generate frequency histograms in which the distribution of values can be compared with a curve representing the normal distribution. Other ways of determining if samples are normally distributed are beyond the scope of this article but include normal probability plots and tests for normality such as the Kolmogorov–Smirnov test, Lilliefors test and a chi-square goodness-of-fit test.[18] If it is still unclear whether the data are normally distributed after examining these plots and tests then the most appropriate solution may be to conduct a non-parametric test as these are more conservative than parametric tests.
- Exhibit equality of variance. This is also known as uniform variance or homogeneity of variance. In other words: are the standard deviations for the samples similar? There are several tests that can be used for determining equality of variance such as the *F*-test and the Levene test. Knowing the equality of variance will allow the application of an appropriate statistical test.

Data amenable to the use of non-parametric statistics

Non-parametric statistical tests are used for binary, ordinal or nominal data, and also for interval and ratio data that is not normally distributed or, in specific instances, does not exhibit equality of variance. Examples for detecting a difference in variables are the Mann–Whitney *U* test (non-parametric equivalent of two-sample *t*-test) and Wilcoxon matched pairs test (non-parametric equivalent of paired *t*-test); for detecting correlation the Spearman's rank correlation coefficient can be applied. This group of tests requires fewer assumptions to be met regarding the underlying population distributions.[19] In these tests the testing is carried out on ranked data, even when e.g. interval data is analysed, which may lose information about the magnitude of differences within the scale data.[2] Such tests are considered less powerful than parametric testing, with the reduced power increasing the chance of a type II error.[13,20] Using the previous example of nephrotoxicity once more, when the researcher decides, based on the outcome of a statistical test, that there is no relationship between the incidence of contrast-induced nephrotoxicity and the type of contrast agent when in fact there is but the test did not detect this either because the sample size is too small, the standard deviation of the sample data is too large or the effect size is too small.

Application of inferential statistics

Inferential statistics are very useful for a magnitude of applications, experiments and investigations. In order to illustrate this we can use an example of two participant groups that attended prevalent breast screening and have returned for incident screening, compared with those that attended for prevalent breast screening but did not attend incident screening. If there are differences between survey scores, for answers to the question "was the examination painful?", in the mean score for each of these two groups then there could potentially be three different explanations:

1) experimental manipulation caused a change in the phenomena of interest. The patients who did not return for incident screening felt that the breast examination was painful and therefore they were reluctant to return for follow-up examinations.
2) the samples come from different populations with an underlying difference causing the change in the phenomena of interest. The patients who did not return may have been from a different ethnic, cultural or social background from those who did return for incident screening.
3) the samples are from the same population and the differences in the mean occurred due to chance. This can be determined by applying an appropriate statistical test.

Of the three explanations above, number one is the explanation sought when conducting research – a true effect that links the effect (reduced return for screening) with a cause (examination is painful). Number two in the list can be caused by either a confounder effect or selection bias. The above is an example of a cohort study. In such studies and also in case-control studies and clinical trials for instance, care has to be taken to ensure that two unpaired groups are similar. There are various strategies to minimise the risk of introducing confounding and/or bias.[21] The last explanation, chance, is controlled by applying *p* values and ensuring that a study has sufficient power. For more information on power, please refer to the section on confidence interval and sample size, as well as Altman and Bland's statistics notes.[12]

Inferential statistics can be used to calculate the probability that the populations are different and the experimental intervention did have an effect. If a probability of 0.05 or 0.01 was observed whilst there is a probability that the effect occurred due to chance, i.e. a 5% or 1% probability respectively, then the results differed possibly due to chance alone and not the experimental phenomenon, but clearly it is more likely there is a relationship here. The *p* value will yield this information where if the *p* value is 0.05 or less the difference between the two groups is considered statistically significant thus allowing the researcher to reject the null hypothesis and accept the alternative hypothesis. If the *p* value is greater than 0.05 the null hypothesis is accepted. The level of the *p* value used for probability is chosen at the design stage of the study.

Choosing a statistical test

When selecting a statistical test, the appropriate test to determine the *p* value is based on:

Table 1
Types of statistical tests depending on the type of data, number of samples and distribution of data.

Type of data	Number of samples	Statistical test
Binary	One or paired	McNemar's test
	Two independent samples	Chi squared test or Odds ratio
Nominal	One or paired	Stuart test
	Two independent samples	Chi squared test
Ordinal	One or paired	Wilcoxon test
	Two independent samples	Mann–Whitney *U*-test
Interval/ratio	One or paired	Wilcoxon test (non-normal)
		Paired *t*-test (normal)
	Two independent samples	Mann–Whitney *U*-test (non-normal)
		Unpaired *t*-test[a] (normal)

[a] In addition to testing the distribution of the data, the equivalence of variance should also be assessed. The outcome of these assessments will influence the choice of *t*-test applied – with an unequal variance the unequal variance *t*-test should be applied.[24]

a) the level of the data – binary, nominal, ordinal, interval/ratio.
b) the number of different groups in the investigation.
c) whether the data was collected from independent groups i.e. fifty patients who were hydrated peri-procedurally for MRI contrast agent-enhanced examinations, and fifty other patients who had not have the hydration regime prior to their contrast agent-enhanced examinations.
d) the distribution of the data, i.e. are parametric assumptions justified.[4]
e) if the test designed to investigate a correlation or difference.

Table 1 below is adapted from Edmondson to help researchers decide on applying the appropriate test for studying differences between samples.[22] Please note the tests summarised here are in essence for two sets of variables only. Examples are a group receiving a placebo versus a group receiving a medicine, or the volume of a tumour prior and following radiation therapy treatment in a group of cancer patients. Where there are more than two variables or the group/s have been subjected to more than two treatments the analysis of variance ANOVA test will be more appropriate. This would apply if, say, a placebo group is compared to a group of patients receiving medicine A and a third group receiving medicine B. Although Table 1 is made as complete as possible, various different tests may be more appropriate depending on the sample type, size and characteristics. An overview of what regression or correlation analysis method to apply with certain types of data can be found elsewhere.[23]

There is a plethora of different statistical tests available to analyse data. As a result it is difficult for inexperienced researchers to decide which test is most appropriate. Because it is impossible to cover all the different statistical tests, what follows is a description of some of the commonly used tests.

Pearson product moment coefficient

This is a way to quantify the relationship between two variables and it relates to ratio/interval data. It tests for correlation between two samples not difference. Correlation coefficients are represented as *r* and can be positive meaning that in a related way that one variable increases in relationship to another or negative, i.e. that one variable decreases in relationship to another. They vary from +1, meaning a perfect positive relationship, to −1, which is a perfect negative or inverse relationship. If the *r* value is 0, it implies that the two variables are unrelated, i.e. variable B does not change, in a predictable linear fashion, when variable A changes. What is considered a strong or good correlation is a point of debate and there are no set rules that determine the strength of a relationship, although suggestions have been made in the past.[25] The reason why there are no set rules here is because correlation depends on the context and the sample size. As with other inferential statistical tests, a *p* value can be produced to determine the significance of an obtained correlation.

It is important though not to jump to the conclusions when a correlation coefficient shows a relationship, even when shown to be statistically significant. Association does not imply causation, as other factors may be involved. One of the most recent and high-profile examples of making unsupported conclusions is probably the non-causal relationship between the MMR vaccine and autism.[26] There is a similar correlation test for when at least one of the data sets is ordinal, called Spearman's rho.

The *t*-test

This is a parametric test to compare the means of two samples which can be related e.g. the same group of patients is used in a cross-over study in which pain is measured before and after a treatment, using an analogue visual pain rating scale, or they can be unrelated samples e.g. where one sample of patients is fully briefed about a diagnostic examination, including being given a printed fact sheet whilst, the second is not and subsequently the patients in these samples are asked how painful the examination was, again using an analogue visual pain rating scale. The *t*-test requires the input of interval/ratio data and the assumption that the data is suitable to having parametric statistics data (see above) applied to it, must be satisfied. The computed *t* value for a test will give the *p* value if you are using one of the common packages for statistics. This can then be compared with the arbitrary threshold *p* value set to consider an outcome statistically significant. Although the software handles many things on behalf of the researcher, what does have to be put in manually is whether the *t*-test should be one-sided or two-sided. A simplified way of explaining the difference is that one-sided is only used if one knows in which direction the effect of an intervention or comparison is likely to go. For example, if the introduction of a diagnostic test is certain to improve detection rates. Often though, it is not known in what direction an experiment will move – that is why the experiment or study is done in the first place. In such a case, and if you are uncertain, the *t*-test should be two-sided. The chance of the occurrence of a negative of positive effect is taken into account.

Chi squared test for independence

This test for independence, also known as Pearson's Chi squared test, shows whether there is a relationship between two variables and is normally used for data suitable for non-parametric statistical testing. A few important assumptions for this test to be suitable for testing data are that the sampling is random, the sample size is large enough and the observations are independent. Frequency

counts of data, let's say a comparison of the number of women and men being who need an MRI scan but have either claustrophobia or not, can be analysed to see if the observed numbers differ significantly (arbitrary level would be p value of 0.05) from the expected frequencies. The computer will calculate the expected frequencies and subsequently a chi squared value χ^2 which must be equal to or exceed the critical value deemed statistically significant.[22] The formula for this is: (observed frequency°−°expected frequency)^2/ expected frequency. In order to obtain the p value for a Chi squared test, the Chi squared statistic can be looked up and compared to p values in a table. The degree of freedom applicable to the data is important in performing this test; this is calculated using the formula (r°−°1)(c°−°1). For the example above there would two columns (male, female) and two rows (claustrophobic, not claustrophobic), hence one degree of freedom (2°−°1)(2°−°1).

Notes regarding statistical tests

Whilst several tests are mentioned in this article it is beyond the scope of the article to describe them all. However, the background information to inferential statistics and Table 1 should help the reader choose the appropriate statistical test for difference when the experiment involves two different samples only.

Calculating a test result is rarely undertaken by hand nowadays, and instead is done using statistical packages such as Excel, Minitab, Supastat and the comprehensive and commonly used SPSS (Statistical Package for Social Sciences) some of which can easily be downloaded.

Checklist for understanding inferential statistics or choosing what sort of data to collect

- Are statistics performed just to describe the data − if so, only descriptive statistics should be used
- Is the data suitable for parametric or non-parametric statistical testing?
- Are you looking for a test to show correlation or difference? These will require different tests from those testing for a difference between groups, interventions or treatments.
- Are statistics used to draw inferences from the data? If so, descriptive statistics alone will be inadequate
- Are the variables discrete or continuous?
- Is the data binary, nominal, ordinal or ratio/interval data?
- What p value or confidence interval needs to be achieved in order to apply a study's findings to the wider population?
- Does the data consist of the same sample of subjects e.g. pre and post mammography (dependent/paired groups) or is the data from independent/unpaired groups
- What is the p value (type I error) set for declaring statistical significance − usually 0.05 or 0.01 − and is this quoted with a confidence interval rather than on its own?
- What power (1 − type II error) should be achieved? Usually this ranges from 0.80 to 0.99 depending on the situation.
- Is the data normally distributed or not?
- How many groups will be analysed c.q. compared with each other?

Conclusion

Inferential statistics provide a way of inferring from the data trends that can be applied to a wider sample. They need to be presented clearly in a clinically relevant way in order that it can be assimilated by readers, such that it can be used to evidence base their practice, narrowing the theory-practice gap. Statistics can be difficult to apply, especially to those without experience, where the selection and undertaking of the appropriate statistical test may appear daunting. It is essential that the type of data collected and its analysis is appropriate so that the research question can be answered. However, if the article above is read, the checklist is used as guidance and some of the references are consulted, the undertaking of statistics should be more successful and reading a scientific paper may be made less tasking.

References

1. Allison JJ, Calhoun JW, Wall TC, Spettell CM, Fargason CA, Weissman NW, et al. Optimal reporting of health care process measures: inferential statistics as help or hindrance. *Manag Care Q* 2000;**8**(4):1–10.
2. Botti M, Endacott R. Clinical research quantitative data collection and analysis. *Int Emerg Nurs* 2008;**16**(2):197–204.
3. Gouveia-Oliveira A. Medical decision making. *Acta Med Port* 1996;**9** (10–12):391–6.
4. Rickard CM. Statistics for clinical nursing practice, an introduction. *Aust Crit Care* 2008;**21**(4):216–9.
5. Welch G, Gabbe S. Review of statistics usage. *Am J Gynecol* 1997;**176** (5):1138–41.
6. Reznick RK, Dawson-Saunders E, Folse JR. A rationale for teaching of statistics to surgical residents. *Surgery* 1987;**101**(5):611–7.
7. Bolton J. Clinicians and the "S-Word". *Clin Chiropractic* 2006;**9**:88–91.
8. Hong E, O'Neil H. Instructional strategies to help learners build relevant mental models of inferential statistics. *J Educ Psychol* 1992;**84**:150–9.
9. Redmond A, Keenan A. Understanding statistics putting *P* values into perspective. *J Am Podiat Med Assocn* 2002;**92**(S):237–305.
10. Justham D, Timmons S. An evaluation of using a web-based statistics test to teach statistics to post-registration nursing students. *Nurse Educ Today* 2005;**25** (2):156–63.
11. Simpson J. Teaching statistics to non specialists. *Sat Med* 2007;**14**(2):199–208.
12. Altman DG, Bland JM. Absence of evidence is not evidence of absence. *BMJ* 1995;**311**:485.
13. Polit D, Beck C. *Nursing, research, appraising evidence for nursing practice*. 7th ed. Philadelphia: Walters Kluner Lippincott, Williams and Wilkins; 2008.
14. Matthews JNS, Altman DG. Interaction 2: compare effect sizes not *P* values. *BMJ* 1996;**313**:808.
15. Rowntree D. *Statistics without tears: a primer for non-mathematicians*. UK: Penguin; 1981. p 96.
16. Altman DG, Bland JM. Standard deviations and standard errors. *BMJ* 2005;**331**:903.
17. Curran-Everett D. Explorations in statistics: standard deviations and standard errors. *Adv Physiol Educ* 2008;**32**:203208.
18. Henderson AR. Testing experimental data for univariate normality. *Clin Chim Acta* 2006;**366**(1–2):112–29.
19. Banks S. Analysis of quantitative research data: part 2. *Prof Nurse* 1999;**14** (10):710–3.
20. Parahoo K. *Nursing research, principles, preass and issues*. 2nd ed. Basingstoke: Palgrave Macmillan; 2006.
21. Jager KJ, Zoccali C, MacLeod A, Dekker FW. Confounding: what is it and how to deal with it. *Kidney Int* 2008;**73**:256–60.
22. Edmondson A. *Advanced biology: statistics*. Oxford: Oxford University Press; 1996.
23. Bland M. *An introduction to medical statistics*. 3rd ed. Oxford: Oxford University Press; 2000.
24. Ruxton G. The unequal variance t test is an underused alternative to student's t test and Mann–Whitney *U* test. Advance Access Publication. *Behav Ecol* 2006;**17**(4):688–90.
25. Cohen J. *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, NJ: Erlbaum; 1988.
26. Taylor B, Miller E, Farrington C, Petropoulos M, Favot-Mayaud I, Li J, et al. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association. *Lancet* 1999;**353**:2026–9.

Glossary

Continuous variables: the data represents an underlying continuum where there are potentially an infinite number of values (such as weight, length and volume).

Descriptive statistics: statistics that can be used descriptively to illustrate the characteristics of a group of observations i.e. the raw data. Often this includes determining a measure of average (central tendency) and spread (dispersion); the mode, median or mean of a data set and the frequency distribution.

Discrete variables: when the data is classified into discrete non-overlapping variables, or whole numbers. An example is the number of ultrasound scans a person has undergone.

Data: numbers or measurements collected as a result of measurements. They could be counts or frequencies or actual numerical values or scores.

Likert scale: is a scale commonly used in questionnaires, and is the most widely used scale in survey research. It is used for ordered category data. When responding

to a Likert questionnaire item, respondents specify their level of agreement to a statement. The scale is named after Rensis Likert who published a report describing its use.[15]

Normal distribution: when a large number of measurements are made at random of one particular variable, the results usually fall into a pattern. Most of the measurements will lie close to the mean value, with few values lying at the extremes. When a frequency distribution is plotted a familiar bell-shaped curve is produced which represents a normal or Gaussian distribution. There are a number of tests that can be used if data exhibits a normal distribution, including Spearman's rho, Pearson's product moment, binomial sign test, Wilcoxon signed ranks, the *t*-test, Chi squared, Mann–Whitney, Analysis of Variance.

Null hypothesis: this hypothesis proposes that in the population there is no difference when comparing two or more treatments or interventions in a study. It assumes that any difference or relationship found between two sets of data is not significant, and can thus be expressed by chance. If a statistically significant difference is observed between e.g. two interventions then the null hypothesis can be rejected.

Probability: the probability or p of an event happening is expressed as a decimal number from 1 to 0 e.g. the probability of a coin falling heads is one out of two $p = 0.5$.

Inferential statistics: statistics that can be used to infer from the sample group generalisations that can be applied to a wider population.

Interval data: is stronger data than nominal or ordinal data. It can be achieved by the use of a calibrated scale to provide quantitative measurements. The difference between interval and ratio data is that ratio data has a true zero, thus interval data is weaker data than ratio data. An example of interval data is temperature measured in Celcius, where the zero is an arbitrary measurement on the scale.

Nominal data: is the least robust data which categorizes, but does not hierarchically rank data into mutually exclusive categories (eye colour, for instance).

Non-parametric tests: are used for ordinal or nominal data e.g. the Wilcoxon matched pairs test and the chi squared test.

Ordinal data: is where the data has a clear order or hierarchy but not on a calibrated scale. The Likert scale is an example of this type of data.

Parametric tests: are selected for data that is normally distributed.

Ratio data: is the strongest data, and has a true zero value. It can be achieved by the use of a calibrated scale to provide quantitative measurements. Height and length are examples of ratio data; if something is twice as long then this is measurable and it has a meaning.

Related samples: these are where data is collected from the samples before and after the treatment, such as the level of pain a patient has before and after a clinical intervention. This is also called matched/paired data.

Significance: an event is said to be significant if the probability of it occurring purely by chance is low.

Standard deviation (SD): a measurement of the spread of data around the mean of a population, a data set, or sample. It is the square root of the mean variance and is expressed in the same unit as the mean. A low standard deviation indicates that all the scores or measures tend to be more similar to each other and therefore to the mean. A high standard deviation indicates the opposite, where scores or measurements differ more and are further spread from the mean. Typically, when a large sample is used, two thirds of data lie within 1 SD of the mean and 95% of data lie within 2 SD of the mean.

Standard error (SE): estimated standard deviation of a statistic, most often the sample mean (then known as SEM); this is a hypothetical figure. For example, the sample mean is the usual estimator of a population mean. The standard error gives an indication of how likely it is for the same mean to be obtained if the experiment was repeated. The 95% confidence interval (CI) can be calculated with the formula: 95% CI = mean $\pm$ 1.96*SE.

Statistical significance: the likelihood or probability that a result or finding is caused by something else than chance only, thereby allowing rejection of the null hypothesis. Often, if the probability of it occurring by chance is less than one in twenty, i.e. $p < 0.05$, this is considered statistically significant.

Type I errors: occur when a researcher claims a result to be significant when the differences have occurred due to random events. The null hypothesis is rejected when it is actually true.

Type II errors: occur when the null hypothesis is not rejected, when it is actually false.

Unrelated samples: compares two samples that are fundamentally different from each other. For example, where one sample of patients is fully briefed about their illness including being given a printed fact sheet whilst the second is not. These are also known as independent samples.

Variable: a characteristic which can take different values, e.g. sex, weight.