HOW TO ANALYSE A DATASET

Dr Declan O'Regan, Consultant Radiologist, Imperial College London.

Introduction

Many people who are new to research find statistical analysis daunting – and poor methodology is widespread despite its clear importance in determining the conclusions of a study. In this section we will summarise the important statistical considerations for collecting, interpreting and presenting your data.

Study design

Firstly, consult a medical statistician before you start to collect any data. A study's design is more important than the analysis as poorly designed research can never be salvaged. When designing your study consider what the inputs are, for instance an intervention or exposure to a risk factor, and what the outcome measures will be. Then you can plan the time sequence in which these will be studied.

The most powerful design is a randomised controlled study in which the treatment arms are concurrent and balanced for prognostic risk factors. In crossover studies each patient acts as their own control; however the effects of the first treatment may carry over into the second. A cohort study follows subjects over a long period of time to identify the relationship between a risk factor and a disease. Retrospective studies begin with the diseased cases and attempt to identify risk factors compared to a control group, but this design is prone to confounding factors and biases. Cross-sectional studies collect data at a defined time to assess the prevalence of disease. The level of evidence provided by your research will depend on the study design you have chosen (Figure 1).

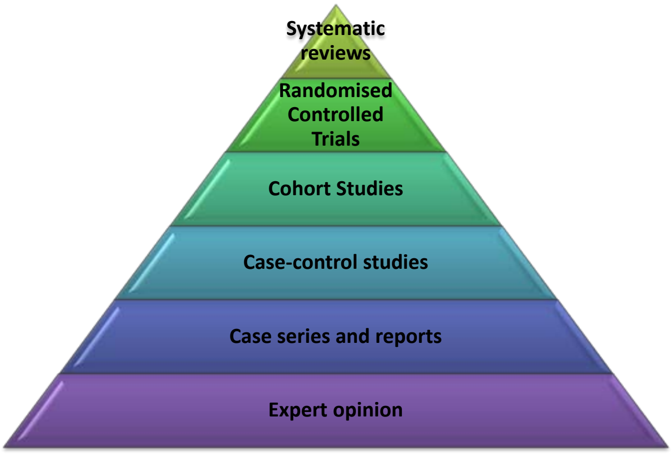

Figure 1: *Research provides different levels of confidence depending on the study design shown by the position on this pyramid.*

Many studies are too small to detect even large effects. Studies that are under-powered may not be able to answer the hypothesis and do not benefit patients. You should try to collect estimates of the variance of your outcome measure from the literature or using pilot data to calculate the sample size needed for the expected effect.

At this stage, before any data has been collected, you should have defined your study design, planned the blinding and randomisation methods, performed a sample size calculation and understand what your main hypothesis and outcome variable are.

Analysing your data

Read the journal's guidance for authors carefully, but do not rely on previously published analyses to justify your own. If your study is well-designed it should be already clear what the main hypothesis is. You should give concise details of your statistical methods and understand what assumptions are made. In choosing what sort of test to use consider the following questions:

> *What are the independent and dependent variables?*
> The dependent variable is the phenomenon we want to explain or predict (outcome) and the independent variable represents the predictor or causal factor (input).
>
> *What is the scale of measurement of the study variables?*
> Categorical – mutually exclusive, but not ordered eg disease state. Ordinal – where the order matters but not the difference in value eg pain score. Interval – continuous data where there is a mathematical relationship between values eg Hounsfield units.
>
> *How many samples/groups are in the design?*
> How many samples are being compared and are they paired or unpaired.
>
> *Have I met the assumptions of the statistical test selected?*
> If your data is categorical or ordinal then use non-parametric tests and if using continuous data which are normally distributed use parametric tests.

A summary of common statistical tests is shown in Table 1. This will help you to decide what the appropriate statistical test is depending on the type of data and the design of the study.

Table 1: *A summary of common statistical tests used in medical research*

	Categorical	**Ordinal**	**Interval**
Agreement between observers	Kappa	Weighted Kappa	Intraclass correlation
Agreement between techniques			Bland Altman plot
Association between variables		Spearman rank correlation	Linear regression Pearson correlation
"Before and after" in same patients	McNemar's Test	Wilcoxon signed rank test	Paired t-test
Compare variables in independent groups	Chi-squared test Fischer's exact test	Mann Whitney U test	Unpaired t-test
Compare repeated measures in the same patients		Friedman test	Repeated measures analysis of variance
Compare three or more independent groups	Chi-squared test	Kruskal-Wallis test	One way analysis of variance

Statistical tips

Many people misunderstand what the P-value means. If the P value is 0.03, that means that there is a 3% chance of observing a difference as large as you observed even if the two population means are identical. Once you have set a threshold P value for statistical significance, every result is either statistically significant or is not statistically significant and so it is often more informative to report the “effect size” and “confidence intervals” for the difference in means. Never conclude that there is no difference or relationship because it is not significant. Quote all your P values in full to one significant figure, eg P=0.51, P=0.003 and avoid phrases like “failed to achieve statistical significance” or commenting on “trends”.

You should consider if there are clustering effects in your data when more than one lesion per subject is measured. A generalised estimating equation, or similar approach, should be used to account for the correlations between outcomes.

If there is more than one outcome measure consider using the Bonferroni method to correct for the number of tests being performed. If there is more than one independent variable a multiple regression analysis can be used. Serial measurements can be compared using a summary statistic such as area under the curve, or a method such as repeated measures analysis of variance (Anova).

Correlation is not the same as linear regression, but the two are related. Linear regression finds the line that best predicts Y from X, whereas correlation quantifies how well X and Y vary together. Also, correlation and accuracy are not synonymous. Even if a new test is highly correlated with the old one they may not be the same. A Bland-Altman plot will reveal how far apart two sets of measurements are from the mean of those measurements (Figure 2).

Avoid using bar charts with error bars. A scatter plot or a box-and-whisker plot is more informative and easier to understand (Figure 3).

Journals often require authors to comply with the STARD guidelines when evaluating diagnostic accuracy and so these should be reviewed before the study is designed.

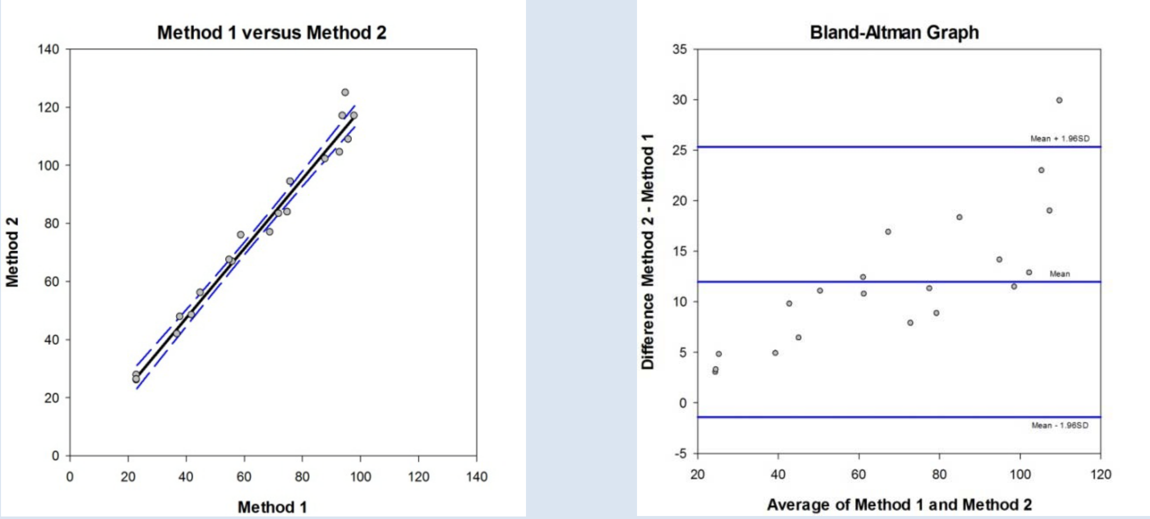

Figure 2. *Two tests may be highly correlated (left), but show a bias that varies with magnitude (right)*

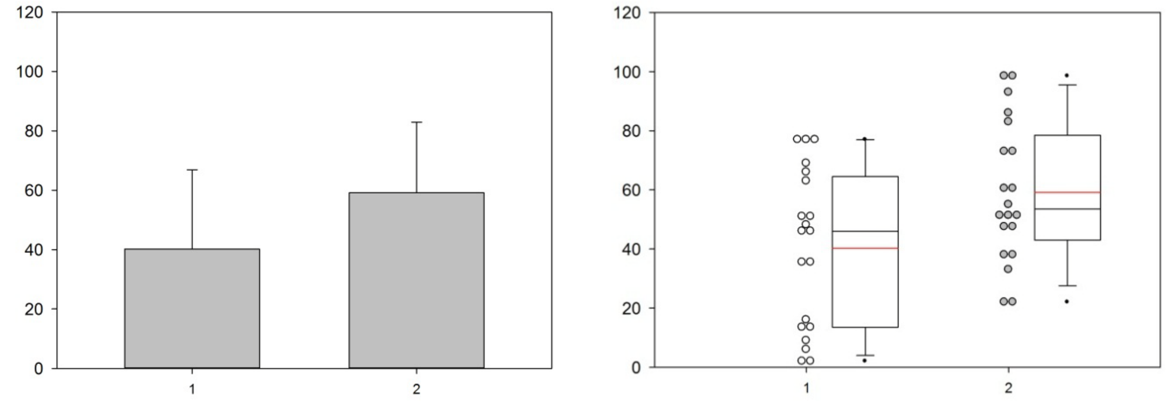

Figure 3. *Bar charts reduce the amount of information communicated and it may be unclear how the error bars have been calculated (left). Dot-plots or box-and-whisker plots make it very clear what the distribution of values is in different groups (right).*

Summary

This is not intended to be an exhaustive guide to medical statistics, but as a starting point for a logical approach to evaluating your data. There are many in-print and on-line resources available to guide you in your statistical analysis and some of these are given in the following bibliography, however none of these is a substitute for expert statistical advice!

Bibliography

Bland M. An introduction to medical statistics. 3rd ed. Oxford: Oxford University Press; 2000
Greenhalgh T. How to read a paper: the basics of evidence-based medicine. 4th ed. ed. Chichester: Wiley-Blackwell. 2010
Harris MD, Taylor G. Medical statistics made easy. 2nd ed. Bloxham: Scion; 2008
Campbell MJP, Walters SJ, Machin D. Medical statistics : a textbook for the health sciences. 4th ed. Chichester: Wiley; 2007.

References

Levine D, Bankier AA, Halpern EF. Submissions to Radiology: Our Top 10 List of Statistical Errors1. Radiology. 2009;253:288-290.
Altman DG. Poor-quality medical research: what can journals do? JAMA. 2002;287:2765-276

Web resources

http://resources.bmj.com/bmj/readers/statistics-at-square-one
A good online resource for advice on medical statistics
www.elsevier.com/framework_**reviewers**/PDFs/Statistics.pdf
Concise summary of common errors identified by referees
http://www-users.york.ac.uk/~mb55/
Prof Martin Bland's personal webpages provides an extensive resource of statistical discussion.
http://en.wikipedia.org/wiki/List_of_statistical_packages
A list of statistical software, including freeware applications.
http://www.stard-statement.org/
Standards for the Reporting of Diagnostic accuracy studies