The Journal of Laryngology & Otology (2006), ***120***, *85–86*.

doi:10.1017/S002221510500006X
Printed in the United Kingdom

Editorial

Reviewing the reviews: some thoughts from the *JLO* statistical advisor

L P HUNT BSc (HONS) MBBS MRCS DO-HNS

Authors submitting manuscripts to this and other journals often find the statistical element of their work to be a major stumbling block. Help with statistics is often difficult to obtain, and many studies falter or fail as a result of what appears to be a widespread lack of understanding of elementary statistics.

In the light of this, the *JLO* Editors asked me to write an aide-memoire that, it is hoped, will prove of use to potential authors. This highlights some of the more frequent statistical analysis and presentational issues that I have needed to comment on during the last five years. The list is not exhaustive but it is hoped that it will help those preparing manuscripts to avoid some of the more obvious pitfalls and problems that routinely arise.

Subjects

In describing the subjects included in a study, it is helpful to know whether these were consecutive, and over what period of calendar time. Were any omitted? If so, what were the reasons? Was any information known about the omissions? The same applies to any control group, if required; how (and why) were these selected, and over what period of time?

Statistical hypothesis testing depends on samples being random and representative. If there are omissions, we need to know whether this is likely to bias the overall findings.

If there is any patient drop-out or non-response, perhaps because a questionnaire is not returned, again, we need to know whether this might cause bias. In a pain questionnaire, for example, non-response might occur because the patient has no pain and feels completely OK. Some attempt, therefore, should be made to ascertain the reasons for non-response, perhaps by contacting a random sample of the non-responders; do the non-responders have particular characteristics?

It is important to give a justification for the sample size(s) used, so give a power calculation if this is appropriate.

In a clinical trial, methods of randomization to treatment and control groups should be described carefully, together with other procedures adopted, such as blind assessment.

Summary statistics

In the calculation of percentages, think carefully about the denominators used. State the denominators, where relevant, to avoid inconsistencies.

When quoting means, always accompany these with a measure of dispersion (e.g. standard deviation (SD) or range) or precision (standard error (SE) or 95 per cent confidence intervals (95% CI)); state clearly which have been quoted.

What is the unit of observation? In a study of ears, for example, the unit of observation for some analyses may be the subject, for others the ear. The latter may need special statistical treatment, because results from two ears on the same subject may be correlated; advice from a statistician should be sought. If you choose to analyse specifically the worst ear, the left or right ear, or the average of the two ears, the unit of observation will be the subject.

Statistical inference

Data that are 'paired' and 'unpaired' should be distinguished and analysed accordingly. The analysis of pair-matched study (for example a one-to-one matched case–control study) should take into account the matching, but remember that you cannot compare the groups with respect to any of the variables used in the matching.

Indicate whether one- or two-tailed tests have been used. Since one-tailed tests are seldom used, some justification for these should be given.

If possible, try to give exact p values rather than just writing $p < 0.05$. Three decimal places is usually quite sufficient; write '$p < 0.001$' rather than '$p = 0.000$'.

When comparing groups, as well as the p values, give 95% CIs for reported differences in means or proportions.

Watch out for small expected frequencies that might invalidate a Chi-squared test; groups should be combined only if there are a priori reasons for doing so.

Non-parametric analyses are methods that can be used when the data are not Normally distributed. When using these methods to compare groups, it is usually better to quote medians and ranges rather than means and SDs. (Note that for paired data the median difference is not usually the same as the difference between two medians.) Non-parametric methods are sometimes used to accommodate results that are below the limit of detection; in a Mann–Whitney U-test, for example, such results are given the average of the lowest ranks.

Data can sometimes be rendered Normal by transformation; parametric methods can be used on the transformed values. Summary statistics in these cases are a bit more problematic. If the data have been logarithmically transformed to remove skewness (e.g. triglycerides, IgG) then the means of the logarithms can be back-transformed to give the geometric mean. The SD cannot be back-transformed (although upper and lower 95 per cent confidence limits can be); it may be helpful just to give the range of the untransformed values.

Visual analogue scales

Visual analogue scales (VAS) are often used to 'measure' pain. Details should be given of how they are constructed and used; if the scale runs from 0 to 10, how should the subject interpret each of the two ends of the scale, 0 and 10?

Because VAS scores are constrained to be between 0 and 10, they are often analysed using non-parametric methods (see above). Some authors have suggested the VAS values should be transformed (see Senn[1] for a discussion on using the logit and arc-sine transformations).

Adjustment for baselines

In a parallel group clinical trial, if cases have been randomized to groups then it is not really appropriate to use hypothesis tests to compare the baseline characteristics. Baseline characteristics should still be shown in the report. If concern remains about the groups not being comparable, despite the randomization, then adjustment can be made for such imbalance using analysis of covariance (for a helpful review, see Vickers and Altman[2]).

Repeated measures

Some group comparison studies involve individual response measurements made at a number of times, such as hourly blood levels or daily pain measures; these are called 'repeated measures'.

In a repeated measures design, it is inefficient to compare the groups at each time point, e.g. by separate *t*-tests. Repeated measures analyses of variance have been used extensively for these, but 'time' is not really a factor and, for an individual subject, pairs of results closer in time may be more closely correlated than pairs of results that are further apart; modern mixed model approaches can take into account the correlation structure.

Group comparisons of daily pain scores, however, can be considerably simplified by calculating the area under the curve (AUC), representing the total pain over the period of study; group comparisons can then be made with respect to the AUCs.

Survival analysis

Define carefully the start- and end-points for the analysis (e.g. the start-point may be diagnosis or treatment, the end-point death or relapse/death). Are any cases censored? The reason for censoring must be independent of the end-point. 'Lost to follow ups' are usually regarded as censorings, but intercurrent deaths might, for example, be regarded as censorings for some analyses and end-points for others.

Survival times may sometimes be calculated up to the point in time at which the analysis was carried out (say, for example, the end of 2004), or may sometimes be up to a fixed time relative to treatment in a clinical trial (say, two years after treatment). Incomplete follow up, where the defined end-point was not reached at last assessment, is accommodated in the analysis as a censored observation and it is important that such cases are included.

Kaplan–Meier estimates are time-related. When quoting these in the text, state the time point at which they were calculated. Try to include a SE or 95% CI together with the estimates.

Multivariable analyses

Justify your decision to carry out any multivariable analyses and check carefully that any assumptions made by the methods about the data are met (e.g. proportional hazards in Cox's proportional hazards regression model); if in doubt, seek advice from a statistician.

Make sure there are enough cases; in multiple logistic regression analysis, for example, it has been suggested that coefficients are likely to be unstable if the number of events per variable is less than 10.[3]

In any regression analysis, be wary of the impact of including two or more predictor variables that are highly correlated with each other; this 'collinearity' can sometimes lead to inflated coefficients.

Justify any non-standard statistical procedures used. Briefly describe the methods and give references.

References

1 Senn S. *Cross-over Trials in Clinical Research*. Chichester: John Wiley and Sons, 1993;74–9
2 Vickers AJ, Altman D. Analysing controlled trials with baseline and follow up measurements. *BMJ* 2001:**323**; 1123–4
3 Feinstein A. *Multivariable Analysis: an Introduction*. New Haven and London: Yale University Press, 1996;269

https://doi.org/10.1017/S002221510500006X Published online by Cambridge University Press