Statistical review: frequently given comments

Stian Lydersen

Handling editor Tore K Kvien

Correspondence to
Professor Stian Lydersen, Regional Centre for Child and Youth Mental Health and Child Welfare, Norwegian University of Science and Technology, Olav Kyrres gate 9, P.O. Box 890, MTFS, Trondheim N-7491, Norway; stian.lydersen@ntnu.no

Received 30 June 2014
Revised 20 August 2014
Accepted 13 September 2014
Published Online First
26 September 2014

To cite: Lydersen S. *Ann Rheum Dis* 2015;**74**: 323–325.

ABSTRACT

From 2006 to 2014, I have carried out approximately 200 statistical reviews of manuscripts for *ARD*. My most frequent review comments concern the following:

1. Report how missing data were handled.
2. Limit the number of covariates in regression analyses.
3. Do not use stepwise selection of covariates.
4. Use analysis of covariance (ANCOVA) to adjust for baseline values in randomised controlled trials.
5. Do not use ANCOVA to adjust for baseline values in observational studies.
6. Dichotomising a continuous variable: a bad idea.
7. Student's t test is better than non-parametric tests.
8. Do not use Yates' continuity correction.
9. Mean (SD) is also relevant for non-normally distributed data.
10. Report estimate, CI and (possibly) p value—in that order of importance.
11. *Post hoc* power calculations—do not do it.
12. Do not test for baseline imbalances in a randomised controlled trial.
13. Report actual p values with 2 digits, maximum 3 decimals.
14. Format for reporting CIs.

INTRODUCTION

From 2006 to 2014, I have carried out approximately 200 statistical reviews of manuscripts for ARD. Some errors and weaknesses occur more often than others. The following is a description of 14 of my comments most frequently given to authors. The first 10 points concern choosing an appropriate analysis method, points 11–12 concern avoiding superfluous analyses and points 13–14 concern reporting formats. Some statistical terms are explained in Appendix. I hope this can help authors to avoid these statistical errors and weaknesses in future manuscripts.

1. Report how missing data were handled

Report the amount of missing data in the different variables, and how this was handled in the analysis.[1] Commonly used methods are, from the less to the more complex ones, complete case analysis (disregarding cases with partially missing data), single imputation methods like expectation-maximation imputation, multiple imputation and full information maximum likelihood. Further, in longitudinal studies, mixed models analysis may be appropriate, while 'last observation carried forward' is not unbiased under any sensible assumptions, and should not be used.

2. Limit the number of covariates in regression analyses

Some authors attempt to include too many covariates compared with the number of cases in a regression model, for example, 17 covariates in a study with 64 cases. Traditional rules of thumb state that the ratio of cases per covariate ought to be in the size of order 10. Some authors recommend 15, some 20, others state that 5 is sufficient. In logistic regression and Cox regression, 10 events per variable is usually sufficient[2] and in many situations 5 events per variable is sufficient.[3] Note that in logistic regression this is not the total number of observations, but the smallest of the two outcome groups. Similarly, in Cox regression, only the number of events excluding censored observations is counted as cases in this context.

3. Do not use stepwise selection of covariates

Automated variable selection procedures like stepwise selection used to be very popular. Today an increasing number of analysts criticise such methods. For example,[4] page 419 states: "There are several systematic, mechanical, and traditional algorithms for finding models (such as stepwise and best-subset regression) that lack logical and statistical justification and that perform poorly in theory, simulations and case studies … One serious problem is that the P-values and standard errors … will be downwardly biased, usually to a large degree".

Selection of covariates should be based on the research question at hand and on substantial knowledge such as what is biologically plausible. Chapter 10 'Predictor selection' in the book[5] gives good guidance on this matter.

4. Use analysis of covariance to adjust for baseline values in randomised controlled trials

Consider a randomised controlled trial (RCT) comparing two treatments, where the outcome variable is measured before treatment and after treatment. Testing if there is a significant change (difference) from before to after treatment in each treatment arm separately is not an appropriate analysis method. One can compare the mean change between the treatment arms. But an even better approach is regression with outcome after treatment as dependent variable, and baseline value and treatment group as covariates.[6] This method is often called analysis of covariance (ANCOVA).

5. Do not use ANCOVA to adjust for baseline values in observational studies

In an observational study, on the other hand, use of ANCOVA cannot be generally recommended[7] (page 126). In fact, ANCOVA can produce different conclusions than analysing a score difference (after score minus before score), a phenomenon also known as Lord's paradox.[8] A central issue is that in a randomised trial, the treatment is applied after measuring the baseline score. Hence the treatment cannot have affected the baseline score. In an observational study, the exposure may also have been present before the baseline score was measured. Then, ANCOVA would generally introduce bias. See also ref 9.

6. Dichotomising a continuous variable: a bad idea

Avoid dichotomising continuous variables if possible.[10–12] Dichotomising implies loss of information and hence loss of statistical power. Moreover, dichotomizing a covariate implies that the effect of that covariate is a step-function changing only at the threshold. In reality, most effects are smooth functions of the covariate. However, sometimes it can be sensible to dichotomise according to some predefined clinical threshold. Data-driven categorisation such as above/below the median of the observations is never a good idea. The same arguments are valid for categorising into more than two categories, although the harm is then somewhat less than by dichotomising.

7. Student's t test is better than non-parametric tests

Student's t test has major advantages over non-parametric tests such as the Wilcoxon test[13]: First, the method allows to compute a CI for the mean of interest, not only a p value. Second, Student's t test is more powerful, particularly in small samples.[14] A widespread misunderstanding is that Student's t test should not be used in small samples. Third, Student's t test is readily generalised into regression analysis and other analyses.

Student's t test is rather robust to deviations from normality[15] as long as there are no residuals extremely distant, say much more than 4–5 SDs, from zero. Visual inspection of Q-Q plots is well suited to detect such deviations. Visual inspection of P-P plots is *not* suited for detecting such deviations. When the data deviate substantially from the normal distribution, one can for example, use bootstrapping to obtain CIs and p values.[16] Bootstrapping has been available in standard statistical software for several years, and is an underused technique in many applications of statistics.

8. Do not use Yates' continuity correction

Many methods have been proposed for testing equality of two proportions. A traditional recommendation is to use Pearson's asymptotic χ^2 test without Yates' correction in 'large' samples, say all expected cell counts are at least five, else, use a small sample method such as Fisher's exact test. Some authors use Pearson's test with Yates' correction. But Yates' correction should be regarded as a historic curiosity from the time before computers were commonly available, and it should never be used.[17 18] Similarly, the version of Yates correction for CIs should never be used.[19] Further recommendations are given in refs 20 and 21.

9. Mean (SD) is also relevant for non-normally distributed data

The mean and SD are meaningful descriptive statistics for data following all types of continuous distributions and sometimes even for ordinal data, not only the normal distribution. A widespread misunderstanding is that one *must* use other measures such as median and IQR if data do not follow the normal distribution. In fact, the mean and SD have several favourable properties. For example, the mean and SD from different studies can readily be combined in a possible later meta-analysis. This is not the case for the quantile-related measures.

10. Report estimate, CI and (possibly) p value—in that order of importance

p Values are overused and overemphasised in medical research as well as many other applied sciences. This problem is well described in a recent article in *Nature*[22] and its accompanying editorial.[23] Sometimes authors report *only* the p value, for example: "Patients exposed to E were more likely than the unexposed to develop the disease D (p=0.04)". The 'Vancouver'-guidelines http://www.icmje.org/recommendations/browse/manuscript-preparation/preparing-for-submission.html#d state the following: "When possible, quantify findings and present them with appropriate indicators of measurement error or uncertainty (such as confidence intervals). Avoid relying solely on statistical hypothesis testing, such as p values, which fail to convey important information about effect size and precision of estimates".

11. *Post hoc* power calculations—do not do it

Post hoc power calculations are futile, although it has been recommended by some journals. Power is the probability of rejecting the null hypothesis in a (future) study. Once the study has been conducted, this probability is either 1 (if the null hypothesis was rejected) else 0. *Post hoc* power is fundamentally flawed.[24] After the study, meaningful quantifications of uncertainty are CIs and p values.[24 25]

12. Do not test for baseline imbalances in a RCT

When reporting a RCT, it is recommended to show a table with baseline demographic and clinical characteristics for each treatment group. But testing for baseline imbalances in a properly randomised trial is futile, although reported in some medical journal articles. Such testing is discouraged by the CONSORT guidelines.[26] Assuming that randomisation has been done properly, we can expect 5% of the baseline variables to differ significantly between the groups (at level 5%), see also refs 27 and 28.

13. Format for reporting CIs

Commonly used separators between confidence limits are comma(,), semicolon(;) and hyphen(-). The comma and hyphen should be avoided, since they resemble a decimal separator, a thousands separator, or a minus sign. A good choice is to use 'to', for example, (0.16 to 0.25), as recommended by refs 29 and 30· The same advice applies for other intervals, such as IQR and minimum to maximum values.

14. Report actual p values with 2 digits, maximum 3 decimals

Avoid reporting p values as n.s. or $p<0.05$ or $p<0.01$. The exception is extremely small p values, which ought to be reported as, for example, $p<0.001$. A much used recommendation is to report p values with up to 2 significant digits and maximum 3 decimals, such as $p=0.12$, $p=0.035$, $p=0.006$ and $p<0.001$.

Competing interests None.

Provenance and peer review Commissioned; externally peer reviewed.

REFERENCES

1 Bjørnstad JF, Lydersen S. Missing Data. In: Veierød M, Lydersen S, Laake P, eds. *Medical statistics in clinical and epidemiological research*. Oslo: Gyldendal Akademisk, 2012:429–61.

2 Peduzzi P, Concato J, Kemper E, *et al*. A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996;49:1373–9.

3 Vittinghoff E, McCulloch CE. Relaxing the rule of ten events per variable in logistic and Cox regression. *Am J Epidemiol* 2007;165:710–18.

4 Rothman KJ, Greenland S, Lash TL. *Modern epidemiology*. 3rd edn., thoroughly rev. and updated ed. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.

5 Vittinghoff E. *Regression methods in biostatistics linear, logistic, survival, and repeated measures models*. 2nd edn. New York: Springer, 2012.

6 Vickers AJ, Altman DG. Statistics notes: analysing controlled trials with baseline and follow up measurements. *BMJ* 2001;323:1123–4.

7 Fitzmaurice GM, Laird NM, Ware JH. *Applied longitudinal analysis*. 2nd edn. Hoboken, NJ: Wiley, 2011.

8 Lord FM. A paradox in the interpretation of group comparisons. *Psychol Bull* 1967;68:304–5.

9 Glymour MM, Weuve J, Berkman LF, *et al*. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. *Am J Epidemiol* 2005;162:267–78.

10 Altman DG, Royston P. The cost of dichotomising continuous variables. *BMJ* 2006;332:1080.

11 Ravichandran C, Fitzmaurice GM. To dichotomize or not to dichotomize? *Nutrition* 2008;24:610–11.

12 Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006;25:127–41.

13 Altman DG, Bland JM. Practice Statistics Notes Parametric v non-parametric methods for data analysis. *BMJ* 2009;338:a3167.

14 Bland JM, Altman DG. Practice Statistics Notes Analysis of continuous data from small samples. *BMJ* 2009;338:a3166.

15 Skovlund E, Fenstad GU. Should we always choose a nonparametric test when comparing two apparently nonnormal distributions? *J Clin Epidemiol* 2001;54:86–92.

16 Storvik G. Bootstrapping. In: Veierød M, Lydersen S, Laake P, eds. *Medical statistics in clinical and epidemiological research*. Oslo: Gyldendal Akademisk, 2012:402–28.

17 Haviland MG. Yates's Correction for Continuity and the Analysis of 2×2 Contingency-Tables. *Stat Med* 1990;9:363–7.

18 Hirji KF. *Exact analysis of discrete data*. Boca Raton: Chapman & Hall, 2006.

19 Fagerland MW, Lydersen S, Laake P. Recommended confidence intervals for two independent binomial proportions. *Stat Met Med Res* 2011. In press.

20 Lydersen S, Fagerland M, Laake P. Tutorial in biostatistics: recommended tests for association in 2×2 tables. *Stat Med* 2009;28:1159–75.

21 Lydersen S, Langaas M, Bakke Ø. The exact unconditional z-pooled test for equality of two binomial probabilities: optimal choice of the berger and boos confidence coefficient. *J Stat Comput Simulation* 2012;82:1311–16.

22 Nuzzo R. Statistical errors. *Nature* 2014;506:150–2.

23 Editorial: Number crunch. The correct use of statistics is not just good for science—it is essential. *Nature* 2014;506:131–2.

24 Hoenig JM, Heisey DM. The abuse of power: the pervasive fallacy of power calculations for data analysis. *Am Statist* 2001;55:19–24.

25 Bacchetti P. Peer review of statistics in medical research: the other problem. *BMJ* 2002;324:1271–3.

26 CONSORT guidelines. 20-4-2010. Ref Type: Internet Communication.

27 Fayers PM, King M. A highly significant difference in baseline characteristics: the play of chance or evidence of a more selective game? *Qual Life Res* 2008;17:1121–3.

28 Pocock SJ, Assmann SE, Enos LE, *et al*. Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Stat Med* 2002;21:2917–30.

29 Altman DG, Machin D, Bryant TN, *et al*. *Statistics with confidence*. 2nd edn. BMJ Books, 2000.

30 Lang TA, Secic M. How to report statistics in medicine annotated guidelines for authors, editors, and reviewers. 2nd edn. New York: American College of Physicians, 2006.

APPENDIX

Some statistical terms used in this article are explained below.

Expectation-maximation imputation of missing data: Missing values in the data set are estimated as their expected values, given all the observed values in the data set. This results in a complete data set with singly imputed values. Single imputation can be an acceptable procedure if there is a low proportion of missing values.

Multiple imputation of missing data: Several complete data sets, for example, m=20 data sets, are created. The missing values are in principle drawn randomly from their expected distributions, given all the observed values in the data set. Subsequently, analysis results from each of the m complete data sets are combined to give estimates, CIs and p values taking the variability within and between the imputed data sets into account.

Last observation carried forward: Consider a longitudinal study where the patients are scheduled to visit the clinic at certain time points. If data are missing at a time point, data from the last available time point are filled in. For example, the scheduled visits are at 1 month, 2 months, 3 months and 6 months, and a patient missing data at 3 months gets the values from 2 months filled in also at 3 months. If data are also missing at 6 months, the same values are carried forward 6 months as well.

Stepwise selection of covariates: From a given set of candidate covariates for a regression analysis, only those fulfilling a given data-driven criterion are included in the final analysis. A common criterion is that the p value must be below a threshold. Several variants of stepwise selection exist, including forward selection, backwards elimination and all subsets regression.

ANCOVA (analysis of covariance): In this context, ANCOVA means regression analysis with outcome after treatment (or exposure) as dependent variable, including baseline value and treatment group as covariates. Some authors use ANCOVA in a slightly different meaning.

Yates' continuity correction: An adjustment which can be used in the calculation of Pearson's χ^2 statistic for 2×2 tables, and in a few other applications. This adjustment has been recommended for small samples.

Q-Q plot: A plot of the observed values versus the expected values under an assumed probability distribution, usually the normal distribution. If the graph is close to a straight line, the data agree well with the assumed probability distribution.

P-P plot: A plot of the observed cumulative probabilities versus the expected ones under an assumed probability distribution, usually the normal distribution. If the graph is close to a straight line, the data agree well with the assumed probability distribution.