Statistics Notes

The cost of dichotomising continuous variables

Douglas G Altman, Patrick Royston

Cancer Research UK/NHS Centre for Statistics in Medicine, Wolfson College, Oxford OX2 6UD
Douglas G Altman *professor of statistics in medicine*

MRC Clinical Trials Unit, London NW1 2DA
Patrick Royston *professor of statistics*

Correspondence to: Professor Altman doug.altman@cancer.org.uk

BMJ 2006;332:1080

Measurements of continuous variables are made in all branches of medicine, aiding in the diagnosis and treatment of patients. In clinical practice it is helpful to label individuals as having or not having an attribute, such as being “hypertensive” or “obese” or having ”high cholesterol,” depending on the value of a continuous variable.

Categorisation of continuous variables is also common in clinical research, but here such simplicity is gained at some cost. Though grouping may help data presentation, notably in tables, categorisation is unnecessary for statistical analysis and it has some serious drawbacks. Here we consider the impact of converting continuous data to two groups (dichotomising), as this is the most common approach in clinical research.[1]

What are the perceived advantages of forcing all individuals into two groups? A common argument is that it greatly simplifies the statistical analysis and leads to easy interpretation and presentation of results. A binary split—for example, at the median—leads to a comparison of groups of individuals with high or low values of the measurement, leading in the simplest case to a *t* test or χ^2 test and an estimate of the difference between the groups (with its confidence interval). There is, however, no good reason in general to suppose that there is an underlying dichotomy, and if one exists there is no reason why it should be at the median.[2]

Dichotomising leads to several problems. Firstly, much information is lost, so the statistical power to detect a relation between the variable and patient outcome is reduced. Indeed, dichotomising a variable at the median reduces power by the same amount as would discarding a third of the data.[2 3] Deliberately discarding data is surely inadvisable when research studies already tend to be too small. Dichotomisation may also increase the risk of a positive result being a false positive.[4] Secondly, one may seriously underestimate the extent of variation in outcome between groups, such as the risk of some event, and considerable variability may be subsumed within each group. Individuals close to but on opposite sides of the cutpoint are characterised as being very different rather than very similar. Thirdly, using two groups conceals any non-linearity in the relation between the variable and outcome. Presumably, many who dichotomise are unaware of the implications.

If dichotomisation is used where should the cutpoint be? For a few variables there are recognised cutpoints, such as >25 kg/m^2 to define “overweight” based on body mass index. For some variables, such as age, it is usual to take a round number, usually a multiple of five or 10. The cutpoint used in previous studies may be adopted. In the absence of a prior cutpoint the most common approach is to take the sample median. However, using the sample median implies that various cutpoints will be used in different studies so that their results cannot easily be compared, seriously hampering meta-analysis of observational studies.[5] Nevertheless, all these approaches are preferable to performing several analyses and choosing that which gives the most convincing result. Use of this so called “optimal” cutpoint (usually that giving the minimum P value) runs a high risk of a spuriously significant result; the difference in the outcome variable between the groups will be overestimated, perhaps considerably; and the confidence interval will be too narrow. This strategy should never be used.[6 7]

When regression is being used to adjust for the effect of a confounding variable, dichotomisation will run the risk that a substantial part of the confounding remains.[4 7] Dichotomisation is not much used in epidemiological studies, where the use of several categories is preferred. Using multiple categories (to create an “ordinal” variable) is generally preferable to dichotomising. With four or five groups the loss of information can be quite small, but there are complexities in analysis.

Instead of categorising continuous variables, we prefer to keep them continuous. We could then use linear regression rather than a two sample *t* test, for example. If we were concerned that a linear regression would not truly represent the relation between the outcome and predictor variable, we could explore whether some transformation (such as a log transformation) would be helpful.[7 8] As an example, in a regression analysis to develop a prognostic model for patients with primary biliary cirrhosis, a carefully developed model with bilirubin as a continuous explanatory variable explained 31% more of the variability in the data than when bilirubin distribution was split at the median.[7]

Competing interests: None declared.

[1] Del Priore G, Zandieh P, Lee MJ. Treatment of continuous data as categoric variables in obstetrics and gynecology. *Obstet Gynecol* 1997;89:351-4.

[2] MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. *Psychol Meth* 2002;7:19-40.

[3] Cohen J. The cost of dichotomization. *Appl Psychol Meas* 1983;7:249-53.

[4] Austin PC, Brunner LJ. Inflation of the type I error rate when a continuous confounding variable is categorized in logistic regression analyses. *Stat Med* 2004;23:1159-78.

[5] Buettner P, Garbe C, Guggenmoos-Holzmann I. Problems in defining cutoff points of continuous prognostic factors: example of tumor thickness in primary cutaneous melanoma. *J Clin Epidemiol* 1997;50:1201-10.

[6] Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using “optimal” cutpoints in the evaluation of prognostic factors. *J Natl Cancer Inst* 1994;86:829-35.

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

[8] Royston P, Sauerbrei W. Building multivariable regression models with continuous covariates in clinical epidemiology–with an emphasis on fractional polynomials. *Methods Inf Med* 2005;44:561-71.

Endpiece

Reading and reflecting

Reading without reflecting is like eating without digesting.

Edmund Burke

Kamal Samanta, retired general practitioner,
Denby Dale, West Yorkshire