Ophthalmic Statistics Note 10: Data Transformations

Catey Bunce * [1,2], John Stephenson[3], Caroline J Doré[4], Nick Freemantle[5]
*Corresponding author

[1]Research &Development, NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK;
[2]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK;

[3]School of Human and Health Sciences, University of Huddersfield, Queensgate, Huddersfield GB-HD1 3DH, Great Britain

[4]Comprehensive Clinical Trials Unit, University College London, Great Britain

[5] Department of Primary Care and Population Health, University College London, Great Britain

* Correspondence to Dr Catey Bunce; c.bunce@ucl.ac.uk

Keywords: medical statistics

Word count: 1286

Data transformations

Introduction

Many statistical analyses in ophthalmic and other clinical fields are concerned with describing relationships between one or more "predictors" (explanatory or independent variables) and usually one outcome measure (response or dependent variable). Our earlier statistical notes make reference to the fact that statistical techniques often make assumptions about data (1, 2). Assumptions may relate to the outcome variable, to the predictor variable or indeed both; common assumptions are that data follow normal (Gaussian) distributions and that observations are independent. It is, of course, entirely possible to ignore such assumptions, but doing so is not good statistical practice and in medicine, poor statistical practice can impact negatively upon patients and the public (3).

One approach when assumptions are not adhered to is to use alternative tests which place fewer restrictions on the data – non-parametric or so-called distribution free methods (2). A more powerful alternative, however, is to transform your data. Whilst your "raw" (untransformed) data may not satisfy the assumptions needed for a particular test, it is possible that a mathematical function or *transformation* of the data will. Analyses may then be conducted on the transformed data rather than the raw data.

Scenario 1: A study to evaluate the accuracy of intraocular lens power estimation in eyes having phacovitrectomy for rhegmatogenous retinal detachment (4) measured the axial length (in mm) of 71 eyes. The raw data (Figure 1a) exhibited a fairly strong positive skew (rather than being symmetric there is an extended tail in the histogram to the right); the same data with a logarithmic transformation applied (Figure 1b) appears much more Normal (less of a tail to the right), and hence the power of a test conducted on transformed data should be greater. A further benefit of this transformation is that it can effectively stabilize variance across all values of the predictor variable, another requirement for a valid regression analysis. We can be more confident of the reliability of parametric procedures such as t-tests, regression and analysis of variance conducted on the transformed data than of the same procedures conducted on the raw, skewed data.

[Figures 1a, 1b to be placed about here]

For data which are very highly skewed, the reciprocal transformation may be useful as an alternative to the logarithmic transformation in reducing divergence of data from Normality and stabilizing variance. This transformation may be appropriate in the analysis of data which relate to the duration of events – for example the time taken to conduct cataract surgery or the number of days between ocular trauma and attendance at Accident & Emergency. A square root transformation may also be effective in reducing mild positive skew. Negatively skewed data (i.e. data that is "piled up" to the right rather than the left) is a less common occurrence, but can also sometimes be dealt with effectively by transformations; in this case we usually first *reflect* the data by subtracting all values from some fixed value before applying the transformation. Significance tests can be performed to assess formally whether the sample data follow a normal distribution before and after transformation (5).

Care must be taken when interpreting analysis of transformed data; results from analyses will be for transformed data, not the raw data. Confidence intervals will therefore relate to confidence around estimates for the transformed data rather than the raw data (6).

Scenario 2: A colleague has conducted an exploratory randomized controlled clinical trial evaluating a novel treatment for ocular trauma in 40 patients; 20 of whom received standard care, and 20 of whom received the novel treatment. The primary outcome is visual acuity in the treated eye 6 months after surgery measured using ETDRS charts at a starting distance of 4 metres. A histogram of visual acuity is highly asymmetric, so that the data appear to violate the assumption of normality required for a t-test. Whilst we might apply a non-parametric test such as a Mann-Whitney test (2), I understand that this may result in a loss of power i.e. it would require a larger sample size to identify statistically significant differences. A logarithmic transformation makes the data much more symmetric, and so we apply a t-test to the transformed data. A colleague asks to see an estimate of the treatment effect. Whilst our analysis does furnish an estimated mean difference and confidence interval, we realise that since the analysis was on logged data, the results presented relate not to raw data (i.e. ETDRS vision) but to logged ETDRS vision which is not the same. We can back-transform the results into the natural units by exponentiation. However, exponentiating the mean of the transformed values gives us the geometric mean of the natural data, which will generally be smaller than the arithmetic mean (6). We can also back-transform differences between means on the log scale which describe the ratio of the geometric means for the two treatments, rather than their natural values, while the back transformed confidence interval will be for this ratio and non-symmetric (7). These ratio properties of the difference in geometric means can be particularly helpful, with a ratio of 2 between treatment groups, for example, indicating that the experimental treatment doubles visual acuity. These ratio relationships are quite often seen with continuous variables where patients start with different severities of condition. This method of analysis can be useful but it is important that the 'currency' in which the results are obtained is clearly understood and reported.

Whilst in scenario 2, we can provide meaningful confidence intervals for differences, other transformations may not be amenable to interpretation, and for this reason, the logarithmic transformation is the most useful (7).

Another common reason for applying a non-linear transformation to numerical data is to improve the linearity of a relationship between variables. A study (8) to identify the incidence and risk factors for developing outer foveal defects in patients undergoing macular hole surgery measured visual acuity on the LogMAR scale pre-operatively and post-operatively after vision stabilisation. The relationship between these variables was seen to be non-linear when plotted on a scatter diagram (Figure 2a). Applying a square root transformation to the predictor variable was reasonably effective in achieving a linear relationship (Figure 2b), allowing a subsequent regression analysis to be conducted based on the relationship:

Post-operative visual acuity = a + b $x\cdot$(Pre-operative visual acuity)$^{0.5}$

where a and b are constants to be estimated.

[Figures 2a, 2b to be placed about here]

Linearising transformations of numerical data may be applied to either predictor or outcome variables as appropriate.

Because of the difficulties associated with the interpretation of transformed data, decisions to transform continuous data should not be taken lightly (9). Transformations may improve distributional characteristics, but rarely result in "perfect" data. Whilst linearity can be regarded as an essential pre-requisite for regression-based procedures, many standard analysis methods are robust to moderate divergences from normality, and hence lack of normality, particularly in large data sets, may weaken but is unlikely to violate integrity of procedures. If doubt remains, researchers are encouraged to seek guidance from an experienced biostatistician.

Lessons learned

- All statistical tests make assumptions. If these assumptions are not reasonable, results of the statistical analysis may be misleading.
- Whilst raw ("untransformed data") may not adhere to assumptions, a mathematical function ("transformation") of raw data may do.
- Transformations will, however, impact upon interpretability of results. If in doubt, consult an experienced biostatistician.

Contributors

JS drafted the paper. CJD, CB and NF critically reviewed and revised the paper. CB redrafted the paper after review. CJD, JS and JF critically reviewed the redraft.

Funding

CB is partly funded by the National Institute of Health Research (NIHR) Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology.

Competing interests

None declared.

References

1. Bunce C, Patel KV, Xing W, Freemantle N, Doré CJ; Ophthalmic Statistics Group. Ophthalmic statistics note 1: unit of analysis. Br J Ophthalmol 2014 Mar; 98(3):408-12

2. Skene SS, Bunce C, Freemantle N, Doré CJ; Ophthalmic Statistics Group. Ophthalmic statistics note 9: parametric vs non-parametric methods for data analysis. Br J Ophthalmol 2016

3. Altman DG. The scandal of poor medical research. BMJ 1994 Jan 29; 308(6924):283-4

4. Rahman R, Bong C, Stephenson J. Accuracy of Intraocular Lens Power Estimation in Eyes Having Phacovitrectomy for Rhegmatogenous Retinal Detachment. *Retina* 2014; 34 (7): 1415-1420.

5. Bland JM, Altman DG. The normal distribution. BMJ 1995 February 4;310:298

6. Bland JM, Altman DG. Transformations, means, and confidence intervals. BMJ 1996 April 27;312:1079

7. Bland JM, Altman DG. The use of transformation when comparing two means. BMJ 1996 May 4;312(7039):1153

8. Rahman R, Oxley L, Stephenson, J. Persistent outer retinal fluid following non-posturing surgery for idiopathic macular hole. *Br J Ophthalmol* 2013; 97 (11): 1451-1454.

9. Osborne, Jason (2002). Notes on the use of data transformations. Practical Assessment, Research & Evaluation, 8(6). Retrieved June 10, 2016 from http://PAREonline.net/getvn.asp?v=8&n=6

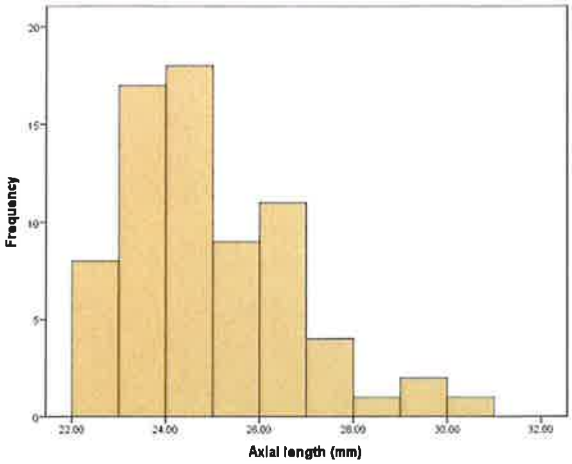

Figure 1a: distribution of axial length values

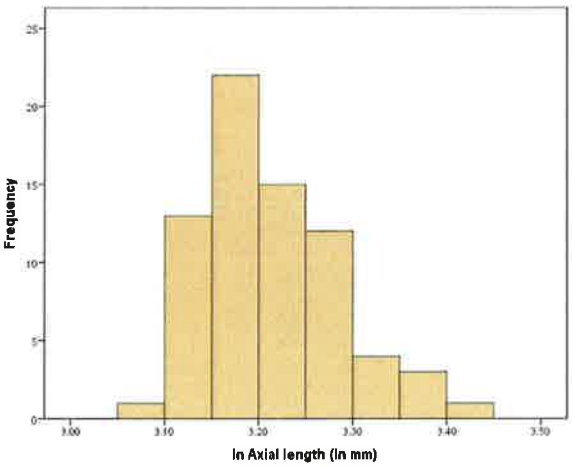

Figure 1b: distribution of log-transformed axial length values

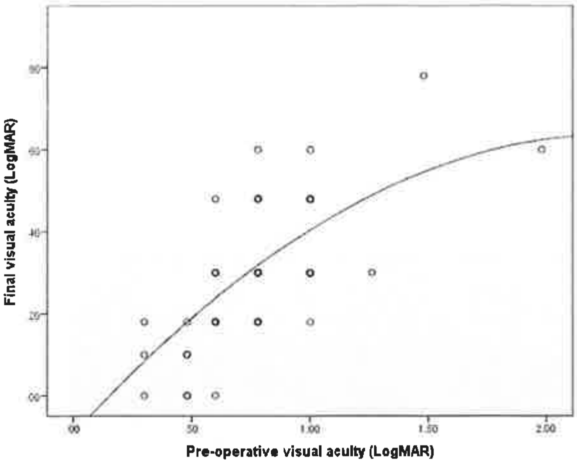

Figure 2a: relationship between raw pre- and post-operative visual acuity scores

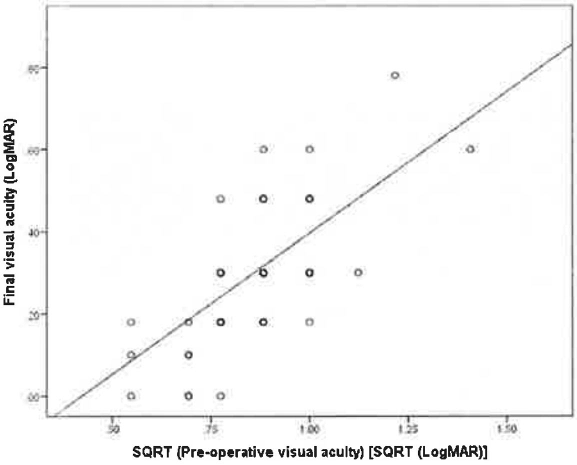

Figure 2b: relationship between transformed pre- and post-operative visual acuity scores