Journal of Research in Weed Science

Journal homepage: www.jrweedsci.com

Review Article

Relevance of Data Transformation Techniques in Weed Science

Prithwiraj Dey [a], Pramit Pandit [b,*]

a Department of Agronomy, G.B. Pant University of Agriculture & Technology, Pantnagar, India.

b Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, India.

ARTICLE INFORMATION

Received: 25 June 2019
Revised: 29 July 2019
Accepted: 30 July 2019
Available online: 3 August 2019
DOI: 10.26655/JRWEEDSCI.2020.1.8

KEYWORDS

Angular transformation
Herbicidal efficacy
Logarithmic transformation
Square root transformation

ABSTRACT

In the field of weed science, data transformation techniques are of frequent use while evaluating investigating weed count data. Even after having its criticism, data transformation still remains as a very popular technique because the reasons for its use are quite greater than its non-use. Depending on the functional relationship existing between mean and variance of the weed count data, suitable transformations like logarithmic, square root and angular, should be used.

Introduction

Data transformation techniques have been employed frequently by the researchers in weed science (Ahrens et al. 1990). As the interpretation of data based on analysis of variance (ANOVA) is valid only under some certain assumptions, data transformation techniques play a vital role if there is any departure from these assumptions (Rangaswamy, 2018). From the theoretical point of view, there exists a criticism that the mathematical procedure can modify the original data distribution. However, from a practical point of view, the problem is that scientists face difficulty in interpretation and discussion of results on scales other than the original (Ribeiro-Oliveira et al. 2018). Even after having its criticism, data transformation remains as a very commonly used technique because though it is not proper for every data set (Quinn and Keough, 2002; Jaeger,

Corresponding author, email: pramitpandit@gmail.com (Pramit Pandit).

2008; O'Hara and Kotze, 2010), the reasons for transforming data are much more than not to use it all (Bartlett, 1947; Keene, 1995; Ahmad et al. 2006).

Violation of Assumptions

Additive nature of treatment effects and block (environmental) effects, independence of experimental errors and normality of the study variate are necessary assumptions for validity of the inferences made from ANOVA (Rangaswamy, 2018). Statistical tests like t-test, F-test, z-test etc. also require the assumption of independence of errors and normality of character under study (Anderson and McLean, 1974; Draper and Hunter, 1969). Generally, the assumption of variance homogeneity is violated along with the assumption of normality. Normal probability plot (Wilk and Gnanadesikan, 1968; Abrahams and Keve, 1971), Shapiro - Wilk's test (Shapiro and Wilk, 1965; Royston, 1992), D'Augstino's test (D'Agostino and Stephens, 1986) etc. can be applied to the validity of the normality assumption. In fixed effects ANOVA, little to moderate departures from normality are not considered of great concern as F test is just slightly affected by non-normality. However, in case of random effects ANOVA, there is severe effect of normality violation. In order to test homogeneity of variances, the mean and variance for each treatment across the replications would be computed. The equality of variance only then can be tested by Bartlett's χ^2 test. If the Bartlett's χ^2 test rejects the hypothesis of equality of variances based on sample evidence, it can be inferred that the variances are heterogeneous. This heterogeneity of variances can be categorized into two classes, viz. i) where there exists a functional relationship between variance and mean and ii) where there does exist any functional relationship between variance and mean. Data transformation is the suitable tool only for the aforesaid first kind of variance heterogeneity (Andrews, 1971; Chou et al. 2018), where data distribution is non-normal, whereas, partitioning of error is the remedial measure for the second kind.

In practical, there are several situations where serious violation of these assumptions are observed making the inference based on these statistical techniques invalid (Sakia, 1992). Under such circumstances, available options (Graybill, 1976) are:

(i) Ignore the violation of the assumptions and proceed with the analysis as if all assumptions are satisfied.

(ii) Decide what is the correct assumption in place of the one that is violated and use a valid procedure that takes into account the new assumption.

(iii) Design a new model that has important aspects of the original model and satisfies all the assumptions, e.g. by applying a proper transformation to the data or filtering out some suspect data point which may be considered outlying.

(iv) Use a distribution-free procedure that is valid even if various assumptions are violated.

However, majority of the scholars are observed to prefer (iii) i.e. data transformation techniques over the other alternatives (Thoeni, 1969; Hoyle, 1973). Hence, in the experiments conducted for observing the herbicide efficiency for controlling weeds, where very often violation of these assumptions are observed practically, there is a necessity to detect the departures and apply the appropriate remedial measures in order to make the interpretations valid.

Desirable Properties of the Transformed Variate

However, a constant variance is not the only condition we seek and precautions are still necessary when using analysis of variance with the transformed variate (Bartlett, 1947). In the ideal case (Bartlett and Kendall, 1946), the transformed variate should satisfy the following:

1. The variances of the transformed variate should remain unaffected by changes in the means. This is also called the variance stabilizing transformation.

2. It should be normally distributed.

3. The transformed scale should be one for which real effects are linear and additive.

4. The transformed scale should be done for which an arithmetic average from the sample is an efficient estimate of true mean.

Transformation of Data

If the relation between the variance of observations and the mean is known, then this information can be utilized in selecting the form of the transformation (Dolby, 1963). We now elaborate on this point and show how it is possible to estimate the form of the required transformation from the data. Box-Cox transformation (Box and Cox, 1964) is a power transformation of the original data. Let y_{ut} is the observation pertaining to the u^{th} plot, and then the power transformation implies that we use y_{ut}'s as

$$y_{ut}^{*} = y_{ut}^{\lambda}$$

Box and Cox (1964) have shown how the transformation parameter λ in $y_{ut}^{*} = y_{ut}^{\lambda}$ may be simultaneously estimated with the other model parameters (overall mean and treatment effects) using the method of maximum likelihood estimation. This is considered as a very general transformation. The particular cases of this transformation for different values of λ are given in Table 1 (Montgomery et al. 2017).

Table 1. Box-Cox Transformations for different values of λ.

Value of λ	Name of the transformation
1	No transformation
1/2	Square Root
0	Log
-1/2	Reciprocal Square Root
-1	Reciprocal

Among the transformations employed in biological fields, the most used transformations are logarithmic, square root and angular (Ribeiro-Oliveira et al. 2018). These transformations are usually associated with the type of non-normal data (Zar, 2014). Under such circumstances, data transformation is the most appropriate remedial measure. With the help of this technique, the original weed count data can be converted to a new scale resulting into a new data set, which is expected to satisfy the variance homogeneity principle (Montgomery et al. 2017). As common transformation scale is applied to all the observations, the comparative values between treatments remain unaltered, keeping the comparisons between them valid. Though not so popular, Atkinson (1985) and Piepho (2003) had mentioned other kinds of transformations.

Logarithmic Transformation

This transformation is suitable for the data where the variance is proportional to square of the mean (Montgomery et al. 2017) or the CV (coefficient of variation) is constant or where effects are multiplicative. When data range is wide in herbicidal experiments conducted for controlling weeds, these conditions are usually found. For such cases, it is appropriate to analyse log X instead of X (actual data). When small values or zeros are involved in the data set, log (X+1), log (2X+1) or log (X+3/8) should be used in place of log X. This transformation is effective specifically in case of normalising a positively skewed distribution. It is also helpful to achieve additivity (Zar, 2014; Rangaswamy, 2018).

Square-Root Transformation

While variance is proportional to the mean, the square root transformation must be considered (Bartlett, 1936; Dean and Voss, 1999; Zar, 2014; Montgomery et al. 2017), which leads to recommendations for cases where there exist few variations between variance and mean (O'Hara and Kotze, 2010). In other words, when statistical data is consisted of integers i.e. whole numbers, like number of weeds per plot, homogeneous conditions will often lead to variation in these numbers x following the Poisson distribution (Montgomery, 2013; Gupta and Kapoor, 2014). Since

for such a distribution the variance is exactly equal to the mean, that to stabilize the variance we must work on the square root scale. When very small numbers are involved, the use of $\sqrt{(x+0.5)}$ is recommended instead of $\sqrt{x}$, especially when zeros are occurring among the observed numbers (Rangaswamy, 2018). Variance of Poisson variate on transformed scale (Bartlett, 1947) is given in Table 2.

Table 2. Variance of Poisson variate on transformed scale.

Mean on Original Scale	$\sqrt{x}$	$\sqrt{(x+0.5)}$
0.0	0.000	0.000
0.5	0.310	0.102
1.0	0.402	0.160
2.0	0.390	0.214
3.0	0.340	0.232
4.0	0.306	0.240
6.0	0.276	0.245
9.0	0.263	0.247
12.0	0.259	0.248
15.0	0.256	0.248

Angular Transformation

Variables expressed by a proportion and (or) percentage are suitable for the application of angular transformation (Zar, 2014) so that variance can be expressed as a quadratic function of the proportion (Warton and Hui, 2011). The distribution of percentages is binomial (Dean and Voss, 1999; Montgomery, 2013; Gupta and Kapoor, 2014; Montgomery et al. 2017) and this transformation makes the distribution normal. It is also known as 'arcsine' or 'inverse sine' transformation. In the agronomical experiments conducted for weed science, usually number of a particular weed species is converted to proportion and (or) percentage of total weed counts. Since the role of this transformation is not properly understood, there is a tendency to transform any percentage using angular transformation. It should be noted that only that percentage data that are derived from count data as described earlier should be transformed (Rangaswamy, 2018). The angular transformation is given as (Bartlett, 1947),

$$g(x) = sin^{-1}\sqrt{x}$$

However, as mentioned above, the transformations of experimental data, regardless of mathematical expressions, are sometimes performed for other purposes, where statistical

assumptions are not met (Ribeiro-Oliveira et al. 2018). A classic instance of improper employment of data transformation is the attempt to decrease the coefficient of variation (Souza et al. 2008). Oliveira et al. (2009) had considered coefficient of variation as an index of experimental quality. However, Pereira and Santana (2013) mentioned that any assumption on coefficient of variation is not necessary to make the outcomes from ANOVA valid. Presence of zeros in the data set is also associated with coefficient of variation (Couto et al. 2009). A relation between samples per plot size with coefficient of variation can also be drawn.

On the contrary, non-parametric models can be extensively used to avoid the assumptions needed for parametric ANOVA. Ribeiro-Oliveira et al. (2018) mentioned that non-parametric methods are applicable, especially when there are no residuals adjusting to the Gaussian distribution (Judice et al. 1999). This seems to be quite conflicting because normal approximation for large samples is considered as a basis of the nonparametric statistics (Zar, 1999). Because of the inability of non-parametric tests to minimise type I and type II errors (Lix et al. 1996), these tests are considered inefficient compared to its parametric analogous in inferential statistics (Ribeiro-Oliveira et al. 2013; Ribeiro-Oliveira and Ranal, 2016).

In order to study the functional relationship, especially if it is non-linear in nature, between predictors (eg. Weather parameters like rainfall, maximum temperature, minimum temperature etc.) and response variables (eg. Number of weeds), it should be remembered that if transformation is only made on the response variable, model parameters will lose their biological meaning along with distorting the functional relationship as a whole (Onofri et al., 2010). In order to avoid this problem, Carroll and Ruppert (1988) suggested to transform both regressor and regressed variables. Streibig (1988) have mentioned numerous instances of application of this approach in the field of weed science.

Conclusion

Data transformation techniques are widely used in biological fields especially in weed science when necessary assumptions are not satisfied. While evaluating the efficacy of herbicide treatments, suitable transformations like logarithmic, square root and angular should be very carefully used depending on the functional relationship existing between mean and variance of the weed count data.

Conflicts of Interest

No conflicts of interest have been declared.

References

Abrahams S.C, Keve E.T. 1971. Normal probability plot analysis of error in measured and derived quantities and standard deviations. Acta Crystallogr. A. 27: 157-165.

Ahmad W.M.A.W, Naing N.N, Rosli N. 2006. An approached of Box-Cox data transformation to biostatistics experiment. Statistika. 6: 1-6.

Ahrens W.H, Cox D.J, Girish B. 1990. Use of the Arcsine and Square Root Transformations for Subjectively Determined Percentage Data. Weed Sci. 38: 452-458.

Atkinson A.C. 1985. Plots, Transformations and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis, Oxford University Press, Oxford.

Anderson V.L, McLean R.A. 1974. Design of Experiments: A realistic approach. Marcel Dekker Inc., New York.

Andrews D.F. 1971. A note on the selection of data transformations. Biometrika. 58: 249-254.

Bartlett M.S. 1936. The square root transformation in analysis of variance. J. R. Stat. Soc. 3: 68-78.

Bartlett M.S. 1947. The use of transformations. Biometrics. 3: 39-52.

Bartlett M.S, Kendall D.G. 1946. The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. J. R. Stat. Soc. 7(Suppl.): 128.

Box G.E.P, Cox D.R. 1964. An analysis of transformations. J. R. Stat. Soc. 26: 211-252.

Carroll R.J, Ruppert D. 1988. Transformation and Weighting in Regression, Chapman and Hall, London.

Chou Y.M, Polansky A.M, Mason R.L. 2018. Transforming Non-Normal Data to Normality in Statistical Process Control. J. Qual. Technol. 30: 133-141.

Couto M.R.M, Lúcio A.D, Lopes S.J, Carpes R.H. 2009. Transformações de dados em experimentos com abobrinha italiana em ambiente protegido. Ciência Rural. 39: 1701-1707.

D'Agostino R.B, Stephens M.A. 1986. Goodness-of-fit Techniques, Marcel Dekkar Inc., New York.

Dean A, Voss D. 1999. Design and Analysis of Experiments, Springer-Verlag New York, Inc., New York.

Dolby J.L.1963. A quick method for choosing a transformation. Technometrics. 5: 317-326.

Draper N.R, Hunter W.G. 1969. Transformations: Some examples revisited. Technometrics. 11: 23-40.

Graybill F.A. 1976. The Theory and Applications of the Linear Model, Duxbury Press, London.

Gupta S.C, Kapoor V.K. 2014. Fundamentals of Mathematical Statistics, Sultan Chand & Sons, New Delhi.

Hoyle M.H. 1973. Transformations: An introduction and a bibliography. Int. Stat. Rev. 41: 203-223.

Jaeger T.F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. J. Mem. Lang. 59: 434-446.

Judice M.G, Muniz J.A, Carvalheiro R. 1999. Avaliação do coeficiente de variação na experimentação com suínos. Ciênc. agrotec. 23: 170-173.

Keene O.N. 1995. The log transformation is special. Stat. Med. 14: 811-819.

Lix L.M, Keselman J.C, Keselman H.J. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. Rev. Educ. Res. 66: 579-619.

Montgomery D.C. 2013. Design and Analysis of Experiments, John Wiley & Sons, Inc, USA.

Montgomery D.C, Peck E.A, Vining G.G. 2017. Introduction to Linear Regression Analysis, Wiley India Pvt. Ltd, New Delhi.

O'Hara R.B, Kotze D.J. 2010. Do not log-transform count data. Methods Ecol. Evol. 1: 118- 122.

Oliveira R.L, Muniz J.A, Andrade M.J.B, Reis R.L. 2009. Precisão experimental em ensaios com a cultura do feijão. Ciênc. agrotec. 33: 113-119.

Onofri A, Carbonell E.A, Piepho H.P, Mortimer A.M, Cousens R.D. 2010. Current statistical issues in Weed Research. Weed Res. 50: 5-24.

Pereira V.J, Santana D.G. 2013. Coefficient of variation of normal seedlings obtained from the validation of methods for the seed germination testing of 20 species belonging to the family Fabaceae. J. Seed Sci. 35: 161-170.

Piepho H.P. 2003. The folded exponential transformation for proportions. Stat. 52: 575-589.

Quinn G.P, Keough M. 2002. Experimental design and data analysis for biologists, Cambridge University Press, Cambridge.

Rangaswamy R. 2018. A Textbook of Agricultural Statistics, New Age International(P) Limited Publishers, New Delhi.

Ribeiro-Oliveira J.P, Ranal M.A. 2016. Sample size in studies on the germination process. Botany. 94: 103-115.

Ribeiro-Oliveira J.P, de Santana D.G, Pereira V.J, dos Santos C.M. 2018. Data transformation: an underestimated tool by inappropriate use. Acta Sci. Agron. 40.

Ribeiro-Oliveira J.P, Ranal M.A, Santana D.G. 2013. A amplitude amostral interfere nas medidas de germinação de Bowdichia virgilioides Kunth. Ciênc. Florest. 23: 623-634.

Royston P. 1992. Approximating the Shapiro-Wilk W-Test for non-normality. Stat. Comput. 2: 117 - 119.

Sakia R.M. 1992. The Box-Cox Transformation Technique: A Review. J. R Stat. Soc. 41: 169-178.

Shapiro S.S, Wilk M.B. 1965. An analysis of variances test for normality (Complete Samples). Biometrika. 52: 591-611.

Souza R.A, Hungria M, Franchini J, Chueire L.M.O, Barcellos F.G, Campo RJ. 2008. Avaliação qualitativa e quantitativa da microbiota do solo e da fixação biológica donitrogênio pela soja. Pesqui. Agropecu. Bras. 43: 71-82.

Streibig J.C. 1988. Herbicide bioassay. Weed Res**.** 28: 479-484.

Thoeni H. 1969. A table for estimating the mean of a lognormal distribution. J. Am. Stat. Assoc. 64: 632-636.

Warton D.I, Hui F.K.C. 2011. The arcsine is asinine: the analysis of proportions in ecology. Ecol. 92: 3-10.

Wilk M.B, Gnanadesikan R. 1968. Probability plotting methods for the analysis for the analysis of data. Biometrika. 55: 1-17.

Zar J.H. 1999. Biostatistical analysis, Pearson Education India, India.

Zar J.H. 2014. Biostatistical Analysis, Pearson India Education Services Pvt. Ltd, India.

Cite this article as: Prithwiraj Dey, Pramit Pandit. 2020. Relevance of Data Transformation Techniques in Weed Science. *Journal of Research in Weed Science*, 3(1), 81-89. DOI: 10.26655/JRWEEDSCI.2020.1.8