Acta Scientiarum

http://periodicos.uem.br/ojs/acta
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v40i1.35300

BIOMETRY, MODELING AND STATISTICS

Data transformation: an underestimated tool by inappropriate use

João Paulo Ribeiro-Oliveira*, Denise Garcia de Santana, Vanderley José Pereira and Carlos Machado dos Santos

*Instituto de Ciências Agrárias, Universidade Federal de Uberlândia, Avenida João Naves de Ávila, 2121, 38400-902, Uberlândia, Minas Gerais, Brazil. *Author for correspondence. E-mail: ribeirooliveirajp@gmail.com*

ABSTRACT. There are researchers who do not recommend data transformation arguing it causes problems in inferences and mischaracterises data sets, which can hinder interpretation. There are other researchers who consider data transformation necessary to meet the assumptions of parametric models. Perhaps the largest group of researchers who make use of data transformation are concerned with experimental accuracy, which provokes the misuse of this tool. Considering this, our paper offer a study about the most frequent situations related to data transformation and how this tool can impact ANOVA assumptions and experimental accuracy. Our database was obtained from measurements of seed physiology and seed technology. The coefficient of variation cannot be used as an indicator of data transformation. Data transformation might violate the assumptions of analysis of variance, invalidating the idea that its use will provoke fail the inferences, even if it does not improve the quality of the analysis. The decision about when to use data transformation is dichotomous, but the criteria for this decision are many. The unit (percentage, day or seedlings per day), the experimental design and the possible robustness of *F*-statistics to 'small deviations' to Normal are among the main indicators for the choice of the type of transformation.

Keywords: assumptions; coefficient of variation; criteria for data transformation; parametric and nonparametric statistics; robustness.

Transformação de dados: uma ferramenta subestimada pelo uso inapropriado

RESUMO. O recurso matemático de mudança de escala dos dados polariza a opinião dos pesquisadores. Para um grupo, a transformação não é recomendada por causar problemas inferenciais e descaracterizar o conjunto de dados, que dificultam a interpretação; enquanto para outro, é considerada necessária para atender as pressuposições dos modelos paramétricos. No entanto, o mau uso da transformação se dá a fim do experimento atingir metas de precisão. Por isto, aqui são abordadas as situações mais frequentes envoltas à transformação dos dados e seu impacto nas pressuposições dos modelos de análise de variância e na precisão experimental. Dados de sete experimentos conduzidos em delineamentos inteiramente casualizados ou em blocos casualizados foram usados como estudo de caso. A ineficácia do coeficiente de variação como indicador da necessidade de transformação foi revelada, como também que a transformação pode violar as pressuposições da análise de variância, desmistificando o entendimento de que seu uso, mesmo quando não melhora a qualidade da análise, não desqualifica as inferências. A decisão de transformar é dicotômica, mas os critérios para esta decisão não são poucos. A unidade das características (porcentagem, dia e plântulas por dia), o delineamento experimental e a possível robustez da estatística *F* a pequenos desvios da Normal estão entre os principais indicadores para a escolha do tipo de transformação.

Palavras-chave: pressuposições; coeficiente de variação; critérios para a transformação; estatística paramétrica e não-paramétrica; robustez.

Introduction

The transformation of a dataset to another mathematical scale remains in scientific publications even after it has received criticism. From the theoretical point of view, it is criticised because the mathematical procedure can modify the original data distribution. From a practical point of view, the problem is that scientists have difficulty in interpretation and discussion of results on scales other than the original. The most interesting point is that the use of data transformation is often not related to theoretical assumptions of the analysis of variance model (ANOVA). However, there is no doubt that, although it is not suitable for every data set (Fernandez, 1992; Quinn & Keough, 2002; Jaeger, 2008; O'Hara & Kotze, 2010), the reasons for its use are greater than for its non-use (Bartlett, 1947; Manly, 1976; Berry, 1987; Keene, 1995; Ahmad, Naing, & Rosli, 2006).

In summary, data transformation must be used to approximate the residuals to the Normal distribution,

to enable homoscedasticity and, specifically for randomised block design, to promote the additive effects between blocks and treatments (Steel & Torrie, 1996; Quinn & Keough, 2002; Sokal & Rohlf, 2012). Data transformations are peculiar operations with several mathematical possibilities (Quinn & Keough, 2002; Osborne, 2010), which when used incorrectly can compromise inferences (Valcu & Valcu, 2011) and, consequently, the interpretation of the results. In view of this, there are many mathematical procedures that indicate the most suitable transformation, such as Box-Cox's transformation (Box & Cox, 1964; Ahmad et al., 2006).

The most used transformations are logarithmic (log or ln), square root and angular. Log or ln are recommended for continuous variables with discrepant values and standard deviations proportional to the mean (Bartlett, 1936; Berry, 1987), while the square root must be considered when variance is proportional to the mean (Bartlett, 1936), which leads to recommendations for cases where there are few variations between variance and mean (O'Hara & Kotze, 2010). Angular scales are applied to variables expressed by a proportion and (or) percentage and, therefore, variance as a quadratic function of proportion (Zubin, 1935; Warton & Hui, 2011). However, as mentioned above, processing of experimental data, regardless of mathematical expressions, is sometimes performed for other purposes, which do not meet the statistical assumptions.

The attempt to reduce the coefficient of variation (*CV*) is a classic example of incorrect use of data transformation (see Souza et al., 2008). The *CV* is a variability measurement considered by some scientists to quantify experimental quality (Pimentel-Gomes, 2000; Oliveira, Muniz, Andrade, & Reis, 2009). From this point of view, a high *CV* in a scientific area can be used to condemn a data set from an experimental trial (Bowman, 2001). The *CV* also might be an undeclared indicator of data transformation for agrarian scientists, users of statistical tools such as ANOVA (Santana, personal communication). However, the *CV* is not associated with the assumptions of ANOVA (Pereira & Santana, 2013), even though changes in data scale can reflect on not only the assumptions but also on the *CV*. The *CV* can be related to high genetic variability of the material, with the sample/plot size (Santana & Ranal, 2004) or presence of zeros in the data set (Couto, Lúcio, Lopes, & Carpes, 2009). Thus, inferences from the *CV* in relation to assumptions of ANOVA, and not by means of statistical tests, may compromise the quality of analyses and inferences.

Conversely, tests and (or) nonparametric models are extensively mentioned in science, when they do not require assumptions as ANOVA. Nonparametric statistics are applicable especially when there are no adjusting residuals to the Normal distribution (Fernandez, 1992; Judice, Muniz, & Carvalheiro, 1999; Pontes & Correntes, 2001; Santana & Ranal, 2004), which seems somewhat contradictory since the nonparametric statistics are based on approximate Normal distributions for large samples (see Zar, 1999). Nonparametric statistics are inefficient for multiple comparisons and, as a consequence, in inferences. This can occur because they do not control Type I and II errors (Lix, Keselman, & Keselman, 1996; Pontes & Correntes, 2001). The problems get worse when the analyses involve models with factorial schemes, which have very limited nonparametric statistical procedures. Some authors have been venturing into alternatives to perform the analysis of experimental data according to parametric statistics, even if they do not meet all the assumptions (Pontes & Correntes, 2001; Ribeiro-Oliveira, Ranal, & Santana, 2013; Ribeiro-Oliveira & Ranal, 2016).

Considering this, we have the following questions. What are the consequences of data transformation on *CV* values? Does the type of data transformation (both with and without adjustments) affect the assumptions for ANOVA? Could a criterion, such as the ones based on the robustness of *F* distribution, complementary to the assumption tests, be useful for decisions about data transformation? Would nonparametric statistics be non-robust to detect differences in treatments in a data set analysis? These questions should be answered by well-founded statistical theories and (or) by data simulation based on populations, but it is possible to obtain evidence by analysis of experimental data. Weaddress the most frequent situations in data transformation, using literature and experimental results, for the purpose of helping users of statistical procedures with general applications, such as ANOVA. In addition, we indicate why and when criteria complementary to assumption tests must be used in a data set.

Material and methods

An experimental data set from seven trials carried out with native and cultivated plants was used as a case study. This data set was used to discuss consequences of data transformation on statistical inferences. The trials were planned and conducted in a completely randomised design (CRD) or randomised block design (RBD). The characteristics studied are examples of measurements commonly used in seed science to determine the germination process and (or)

growth/development in seedlings (young plants).The variables (measurements) used as the case study are originally discrete when they are collected from counting observations. However, the variables were converted from discrete to continuous and expressed by percentage, day or seedlings per day.

The consequences of data transformation on assumptions and the factorial ANOVA model (main effects and interaction), as well as on the *CV* were demonstrated. We used a data set of the time to last germination (Labouriau, 1983) and of germinability in *Guazuma ulmifolia* Lam. seeds. We also used one of germinability in *Enterolobium contortisiliquum* (Vell.) Morong. seeds. The impact of data transformation on assumptions of a CRD model was demonstrated by measurements of seed technology (normal, infected, damaged, and vigorous seedlings, as well as dormant seeds) of *Parkia pendula* (Willd.) Benth. ex Walp. and *Senna macranthera* (DC. ex Collad.) H. S. Irwin & Barneby.

The greater number of demands of the RBD model, regarding assumptions, led us to expand the discussion of the consequences of data transformation. This discussion was performed using a data set of seed germination and seedling emergence of *Zea mays* L. We analysed the time of initial and last germination (Labouriau, 1983), Maguire's Rate for seedling emergence (Maguire, 1962) and seedling vigour obtained in the cold test. These data were also the basis for the discussion of when to use data transformation.

The data of germination time (first and last) and Maguire's Rate were expressed by day and seedling per day, respectively. These data were transformed $k = 1.0$ to square root ($\sqrt{x+k}$, where $k = 0$; $k = 0.5$), and (or) to logarithmic [$\log(x+k)$ and $\ln(x+k)$, where $k = 0$; $k = 1$]. The measurements of seed germination and seedling emergence expressed by percentage, such as germinability, were transformed to arcsine $\sqrt{x/100}$.

The hypothesis that nonparametric tests are prone to type I and II errors was studied using a data set of uncertainty of seed germination (Labouriau, 1983) of *Enterolobium contortisiliquum*. The would-be robustness of ANOVA to small violations of the assumptions (Scheffé, 1959) led us to offer an additional criterion to use data transformation for processing the parametric statistics, even when there are violations in assumptions. This discussion was supported by the same data set used to study the data transformation consequences.

More details about seed physiology measurements including preadsheets to calculate these germination measurements can be obtained in Ranal and Santana (2006), and Ranal, Santana, Ferreira, and Mendes-Rodrigues (2009). More details on seed technology measurements can be obtained in Brasil (2009).

The data were submitted to tests of Kolmogorov-Smirnov and Levene to analyse the adjustment of residuals to the Normal distribution and homogeneity of variances. Data from RBD was a special case. We also verified the assumption of additivity of effects between treatments and blocks, which was tested using the Tukey test. The Levene test was processed using the mean when atested characteristic had residuals adhering to Normal; otherwise, the test was processed using the median (Brown & Forsythe, 1974). For parametric analysis, ANOVA models were used based on the experimental design, while for nonparametric analysis, we used the Kruskal-Wallis test. We tested each null hypothesis of the nonparametric and parametric statistics at 0.05 significance.

Results

In some case studies, reductions in the *CV* are notable when there is data transformation (from here on called 'transformed scale'), especially on the logarithm scale. In the time to last germination of *Guazuma ulmifolia* seeds, the *CV* would be considered (Pimentel-Gomes, 2000 *sense*) very high — 62.03% — when observed by the original scale; high — 25% approximately — when the data were transformed to the square root, with or without adjustment; and low — 15% approximately — when the data were transformed to the logarithmic scale (Table 1).

Data transformation does not necessarily reduce the *CV*. This was observed in the germinability in *Enterolobium contortisiliquum* and *Guazuma ulmifolia* seeds (Table 2). The angular transformation promoted an increase in *CV* values (from 2.26 to 6.64% and from 5.31 to 7.54%), although they remained below 10% for both species (Table 2).

The original data of time to last germination of *Guazuma ulmifolia* seeds adhere to the residuals of a Normal distribution. This lack of adherence was solved using the transformed scale (both with and without adjustment) for square root and logarithm (Table 3). Homoscedasticity was achieved in the original scale and maintained after data transformation independent of the transformed scale. Note the poor relevance of the adjustments (here $k = 0.5$ and $k = 1$) in relation to statistics and the significance of tests of Levene and (or) Kolmogorov-Smirnov. As a consequence, they also have low relevance in relation to inferences based on residuals distribution and variance.

Table 1. Values of *F* statistics and associated probabilities of analysis of variance (ANOVA) in a factorial scheme for the time to last germination (t_l) in *Guazuma ulmifolia* Lam. seeds in the original and transformed (square root and logarithmic) scales.

Source of variation[1]	Original	Data transformation		
		$\sqrt{x}$	$\sqrt{x+0.5}$	$\sqrt{x+1.0}$
		F (*P*): Statistics of Snedecor and probability		
Factor 1	0.407 (0.669)	0.493 (0.616)	0.489 (0.618)	0.486 (0.620)
Factor 2	0.461 (0.635)	0.737 (0.488)	0.730 (0.491)	0.722 (0.495)
Factor 1* Factor 2	1.442 (0.247)	1.529 (0.222)	1.526 (0.223)	1.524 (0.223)
CV(%)/adjective	62.03/ Very high	25.78/ High	25.19/ High	24.63/ High
Source of variation[1]		Log transformation		
	Log x	$\log(x+1)$	Ln x	$\ln(x+1)$
		F (*P*): Statistics of Snedecor and probability		
Factor 1	0.658 (0.526)	0.637 (0.537)	0.658 (0.526)	0.637 (0.537)
Factor 2	1.167 (0.326)	1.127 (0.339)	1.168 (0.326)	1.127 (0.339)
Factor 1* Factor 2	1.610 (0.201)	1.600 (0.203)	1.610 (0.201)	1.600 (0.203)
CV(%)/adjective	15.12/ Medium	14.18/ Medium	15.12/ Medium	14.18/ Medium

[1]Source of variation of a factorial model in a completely randomised design; p > 0.05 indicates non-significant effect; *CV*: Coefficient of variation / adjectives Pimentel-Gomes (2000) *sense*.

Table 2. Values of *F* statistics and associated probabilities of analysis of variance (ANOVA) for the germinability seeds of *Enterolobium contortisiliquum*(Vell.) Morong and *Guazuma ulmifolia* Lam. in the original and angular scales.

Source of variation[1]	*Enterolobium contortisiliquum*		*Guazuma ulmifolia*	
	F (*P*): Statistics of Snedecor and probability		*F* (*P*): Statistics of Snedecor and probability	
	Original	Arcsine $\sqrt{x/100}$	Original	Arcsine $\sqrt{x/100}$
Factor 1	0.413 (0.666)	0.834 (0.445)	5.363 (0.011)	3.228 (0.055)
Factor 2	0.634 (0.538)	0.722 (0.495)	0.991 (0.384)	1.149 (0.332)
Factor 1* Factor 2	0.074 (0.990)	0.085 (0.986)	0.366 (0.831)	0.491 (0.743)
CV(%)/adjective	2.26/ Low	6.64/Low	5.31/Low	7.54/Low

[1]The data set used was recorded from a completely randomised design. $p > 0.05$ indicates non-significant effect; *CV*: Coefficient of variation / adjectives Pimentel-Gomes (2000) *sense*.

Table 3. Values of statistics and associated probabilities for inferences regarding Normal distribution of residuals and homoscedasticity for the time to last germination (t_l) and germinability in *Guazuma ulmifolia* Lam. seeds and for the germinability of *Enterolobium contortisiliquum* (Vell.) Morong. seeds in the original and transformed scales (square root and logarithmic).

Characters[1]	Scale	Levene		Kolmogorov-Smirnov	
		F (*P*)	Homogeneous variance	*K-S* (*P*)	Normal residuals
		Guazuma ulmifolia			
Time to last germination (day)	Original	0.622 (0.752)	Yes	0.168 (0.012)	No
	$\sqrt{x}$	2.134 (0.067)	Yes	0.142 (0.063)	Yes
	$\sqrt{(x+0.5)}$	2.158 (0.065)	Yes	0.143 (0.061)	Yes
	$\sqrt{(x+1.0)}$	2.180 (0.062)	Yes	0.143 (0.059)	Yes
	$\log x$	1.102 (0.392)	Yes	0.117 (0.268)	Yes
	$\log(x+1.0)$	1.165 (0.355)	Yes	0.124 (0.174)	Yes
	$\ln x$	1.102 (0.392)	Yes	0.017 (0.268)	Yes
	$\ln(x+1.0)$	1.165 (0.355)	Yes	0.124 (0.174)	Yes
Germinability (%)	Original	2.932 (0.017)	No	0.083 (0.761)	Yes
	Arcsine $\sqrt{x/100}$	2.779 (0.022)	No	0.094 (0.589)	Yes
		Enterolobium contortisiliquum			
Germinability (%)	Original	2.167 (0.064)	Yes	0.141 (0.069)	Yes
	Arcsine $\sqrt{x/100}$	1.868 (0.108)	Yes	0.154 (0.031)	No

[1]*F* (*P*): Statistics and probabilities of Snedecor for Levene test; *K-S* (*P*): Statistics and probabilities of Kolmogorov-Smirnov; p > 0.05 indicates normality of residual distribution and/or homoscedasticity. The data set used was recorded from a completely randomised design. The bold values demonstrate cases of violations of ANOVA assumptions.

The transformation did not affect the significance of the factorial model (i.e., main effects and interaction) used to study the time to last germination of *Guazuma ulmifolia* seeds (Table 1), as well as the germinability of *Enterolobium contortisiliquum* seeds (Table 2). Contrary to this, when the angular transformation was applied on germinability of *G. ulmifolia* seeds, the factor 1 was not significant at 0.05 (Table 2).

Angular transformation, which is recommended for data expressed in percentage, did not solve the heterogeneity observed in the original scale of the germinability in *Guazuma ulmifolia* seeds (Table 3). This type of transformation resulted in loss of adherence of residuals to Normal for germinability in *Enterolobium contortisiliquum* seeds, which had been observed in the original scale. This high lights the low coefficients of variation (*CV*) in the original and transformed scales for the germination of both species even with assumptions violated (Table 2). This proves the inability of this measurement to predict violations of assumptions.

The loss of homogeneity in the abnormal seedlings and dormant seeds in *Parkia pendula*, and in the damaged seedlings in *Senna macranthera* (Table 4) confirms that the angular transformation can be an unsatisfactory tool. For the dormant seeds in *S. macranthera*, the loss of adherence of residuals to Normal was also a consequence of this scale.

Three case studies (damaged seedlings of *Parkia*

pendula and *Senna macranthera*, and dormant seeds in *S. macranthera* in Table 4) demonstrated that data transformation can solve the problem of one assumption but violate another. This would be sufficient to refute data transformation, but the *F* statistic is robust to non-Normal distribution (*sense* Scheffé, 1959). In this context, homoscedasticity would be the priority assumption in deciding when the scientist must transform the data set. Thus, considering robustness, the transformation would be suitable for dormant seeds of *S. macranthera*.

The decision to transform can be based on several criteria related to multiple possibilities that allow meeting assumptions as observed in the dataset of seed technology characters of *Parkia pendula* and *Senna macranthera* (Table 4). The indication for data transformation was given in normal and infected seedlings of *P. pendula*, when it would solve the problem of an assumption without violating another (Table 4). Hard seeds in *P. pendula* and hard seeds and infected seedlings of *S. macranthera* demonstrated that when transformation does not meet the assumptions, violated in the original scale, it is preferable not to perform it. Independently of the criteria, it should be noted that the data discussed in this paragraph were obtained from experiments subject to completely randomised design and, therefore, were dependent on observance of only two assumptions (normality of residuals and homogeneity of variances).

The decision to transform the data is more complex for randomised block design, which possesses one other assumption (additivity effect between blocks and treatments); it was observed in the data set of *Zea mays*. This example contains three situations that can be found in data transformation (Table 5), i.e., (*i*) the violation of an assumption with the use of a transformation (time to first emergence), (*ii*) the recommendation for transformation by meeting one or more assumptions (time to final emergence and Maguire's Rate), and (*iii*) the violation of an assumption when meeting another assumption (seedling vigour). Thus, it is interesting to transform data when one or more assumptions, violated in original scale (time to final emergence and Maguire's rate), are met; or, when data transformation meets additivity to blocks and treatments, but does not impact homoscedasticity (seedling vigour) (Table 5).

Several researchers could have doubts in relation to the robustness criterion, but nonparametric tests are not a way to solve problems of assumption violations, as noted by a data set of uncertainty in seed germination of *Enterolobium contortisiliquum*. This case study revealed an interesting case of Type II error. Although there were Normal distribution and heteroscedasticity in the data set, the probability of ANOVA ($p < 0.05$) diverged from the probability of Kruskal-Wallis's statistics ($p > 0.05$). As a consequence, no differences were found among treatments analysed by the nonparametric statistics (Table 6).

Table 4. Values of statistics and probabilities for inferences regarding the Normal distribution of residuals and homoscedasticity for seed technology characters of *Parkia pendula* (Willd.) Benth. ex Walp and *Senna macranthera* (DC. ex Collad.) H. S. Irwin & Barneby. and analysed in the original and angular scales.

Characters[1]	Scale	Levene		Kolmogorov-Smirnov		Data transformation
		F (*P*)	Homogeneous variances	*K-S* (*P*)	Normal residuals	
		Parkia pendula				
Normal seedlings (%)	Original	1.414 (0.166)	Yes	0.123 (0.009)	No	Yes
	Arcsine $\sqrt{x/100}$	1.359 (0.194)	Yes	0.083 (0.258)	Yes	
Infected seedlings (%)	Original	1.834 (0.047)	No	0.347 (0.000)	No	Yes
	Arcsine $\sqrt{x/100}$	1.576 (0.104)	Yes	0.347 (0.000)	No	
Damaged seedlings (%)	Original	1.688 (0.074)	Yes	0.126 (0.007)	No	No
	Arcsine $\sqrt{x/100}$	4.260 (0.000)	No	0.101 (0.068)	Yes	
Dormant seeds (%)	Original	1.366 (0.191)	Yes	0.458 (0.000)	No	No
	Arcsine $\sqrt{x/100}$	2.162 (0.017)	No	0.458 (0.000)	No	
Non-imbibed seeds %)	Original	9.994 (0.000)	No	0.306 (0.000)	No	No
	Arcsine $\sqrt{x/100}$	6.672 (0.000)	No	0.292 (0.000)	No	
		Senna macranthera				
Normal seedlings (%)	Original	1.444 (0.214)	Yes	0.134 (0.069)	Yes	No
	Arcsine $\sqrt{x/100}$	1.135 (0.370)	Yes	0.096 (0.475)	Yes	
Infected seedlings (%)	Original	1.880 (0.094)	Yes	0.175 (0.003)	No	No
	Arcsine $\sqrt{x/100}$	2.056 (0.067)	Yes	0.150 (0.024)	No	
Damaged seedlings %)	Original	1.213 (0.324)	Yes	0.225 (0.000)	No	No
	Arcsine $\sqrt{x/100}$	3.543 (0.004)	No	0.100 (0.399)	Yes	
Dormant seeds (%)	Original	4.499 (0.001)	No	0.136 (0.061)	Yes	Yes
	Arcsine $\sqrt{x/100}$	0.809 (0.612)	Yes	0.170 (0.005)	No	(robustness *sense*)
Non-imbibed seeds (%)	Original	6.258 (0.000)	No	0.325 (0.000)	No	No
	Arcsine $\sqrt{x/100}$	6.126 (0.000)	No	0.325 (0.000)	No	

[1]*F* (*P*): Statistics and probabilities of Snedecor for Levene test; *K-S* (P): Statistics and probabilities of Kolmogorov-Smirnov; $p > 0.05$ indicates normality of residual distribution and/or homoscedasticity to data set. The data set used was recorded from a completely randomised design. The bold values demonstrate cases of violations of ANOVA assumptions.

Table 5. Values of statistics and probabilities for inferences regarding Normal distribution of residuals, homoscedasticity and additivity effects on seed germination and early growth of seedlings in *Zea mays* L. analysed in the original and transformed scales (square root, with and without adjustments, and angular transformation) scales.

		Levene		Kolmogorov-Smirnov		Tukey		
Characters[1]	Scale	*F* (*P*)	Homogeneous variance	*K-S* (*P*)	Normal residuals	*F'* (*P*)	Additivity effect	Data transformation
Time to first germination (day)	Original	0.246 (0.999)	Yes	0.119 (0.003)	No	0.465 (0.498)	Yes	No
	$\sqrt{x}$	0.213 (1.000)	Yes	0.105 (0.016)	No	9.579 (0.003)	No	
Time to last germination (day)	Original	0.415 (0.978)	Yes	0.059 (0.200)	Yes	13.662 (0.000)	No	Yes
	$\sqrt{x}$	0.304 (0.996)	Yes	0.088 (0.084)	Yes	3.484 (0.066)	Yes	
Maguire's rate (seedlings day^{-1})	Original	0.000 (1.000)	Yes	0.392 (0.000)	No	6.374 (0.014)	No	Yes
	$\sqrt{(x+0.5)}$	0.229 (0.999)	Yes	0.085 (0.121)	Yes	1.846 (0.179)	Yes	
Vigorous seedlings (%)	Original	3.994 (0.000)	No	0.093 (0.051)	Yes	12.210 (0.001)	No	Yes (robustness *sense*)
	Arcsine $\sqrt{x/100}$	2.355 (0.006)	No	0.949 (0.001)	No	3.598 (0.062)	Yes	

[1]*F* (*P*): Statistics and probabilities of Snedecor for Levene test; *K-S* (P): Statistics and probabilities of Kolmogorov-Smirnov; *F'* (P): Statistics and probabilities of Snedecor for Tukey test; p > 0.05 indicates normality of residual distribution and/or homoscedasticity to data set. The data set used was recorded from a randomised block design. The bold values demonstrate cases of violations of ANOVA assumptions.

Table 6. Values of statistics and probabilities of a parametric Analysis of Variance – ANOVA – and a nonparametric analysis of Kruskal-Wallis test for a data set of uncertainty on seed germination of *Enterolobium contortisiliquum*(Vell.) Morong.

	Assumptions		Parametric and nonparametric tests	
	Statistics	Inference	ANOVA (Parametric)	Kruskal-Wallis (Nonparametric)
Characters[1]	*Enterolobium contortisiliquum*			
Uncertainty (Bits)	*K-S*= 0.129 *P* = 0.130	Normal residuals	*F* =2.34 *P* = 0.046	*H* =15.60 *P* = 0.052
	F'= 2.453 *P* = 0.039	Homogeneous variance	There is at least one difference between treatments	There are no differences between treatments

[1]*K-S* and *F'*: Statistics of Kolmogorov-Smirnov and Statistics of Snedecor for Levene test; $p < 0.05$ indicates non-Normal residuals and heteroscedasticity, respectively; *F; H*: Statistics of Snedecor and Kruskal-Wallis tests, respectively; $p < 0.05$ indicates significant differences.

Discussion

In the time to last germination of *Guazuma ulmifolia* seeds, data transformation may impute good experimental accuracy, especially on the logarithmic scale ($14.18 \leq CV \leq 15.12\%$), in relation to the original scale (low precision according to $CV = 62.03\%$). However, evaluating the experimental precision by transformed data (Souza et al., 2008; Oliveira et al., 2009) can generate a false perception of 'efficiency'. This result has a practical impact on some scientific areas of basic advances.

The logarithmic functions are more effective in data scale flattening than other types of data transformation, confirming its recommendation for data sets with high variability (O'Hara & Kotze, 2010; Lúcio et al., 2012). Logarithms in base 10 (log) and natural/Naperian logarithm (ln) are similar mathematical functions and, therefore, have the same effects on the data set. It is expected that changes in the data scale (from original to transformed) have an impact on the estimate of the mean, as well as the mean square error. As a consequence, changing the data scale could also affect the *CV* value (Judice et al., 1999). However, we demonstrate that the way this occurs in the data set (either increasing or decreasing) is not predictable. *CV* reductions are the most common cases when changing the data scale, but there are case studies with *CV* increments. We demonstrated a case of a *CV* increment in the germinability of seeds of *Enterolobium contortisiliquum* and *Guazuma ulmifolia*. The *CV* increment was also reported in the literature, although not discussed. As an example, the *CVs* of germinability of seeds of *E. contortisiliquum*, *Mimosa caesalpiniaefolia* and *Peltogyne confertiflora* also increased after data transformation (Pereira & Santana, 2013), despite being classified as low (Pimentel-Gomes, 2000 *sense*).

There are cases of transformation scales, such as square root, where mathematical function restrictions for data sets with zero or negative values are solved using adjustments (Berry, 1987; Yamamura, 1999). The lack of criteria for the choice of adjustment is another subjective factor that restricts the recommendation to use data transformation. In addition to generating uncertainty in the scientists' decisions (Yamamura, 1999), the use of adjustments can put the efficiency of ANOVA at risk (Fernandez, 1992; Osborne, 2010). We observe that adjustments $k = 0.5$ and $k = 1.0$ do not exert any impact on inferences of Normal distribution and homogeneity in the time to last germination in *Guazuma ulmifolia* seeds. Using

the same data set, we observed minimum differences in the values of statistics and probabilities of Kolmogorov-Smirnov and Levene tests, indicating that adjustments do not have a significant impact on the assumptions of ANOVA.

Even though the relation of variable nature and data transformation is not studied here, we highlight that it is usual for data expressed as a percentage (such as seed germination) to be transformed into an angular scale in an attempt to approximate the data set or the residuals to the Normal distribution. This is due to the consensus that non-Normal is a rule in biological data, as discussed in two extensive reviews of germination data and viability of seeds made by Sileshi (2012) and by Valcu & Valcu (2011). For the germination of *Parkia pendula* and *Senna macranthera* seeds, this tendency in non-adjustment of residuals to Normal prevailed not only in the original scale but also in the transformed scale. Thus, we demonstrate that general instructions (without any statistical test) should be avoided. Many of these recommendations were made when statistical software, which currently facilitate the checking of assumptions, were not widely used. The data set of *Zea mays* enables us to contest the idea that data sets of cultivated species (such as maize) have no problems with Normal distribution.

We believe that the scientific discussion about residuals adjustment to Normal and the consequences of data transformation on this adjustment is unnecessary. This idea is based on the supposed robustness of the *F*-statistic to 'small deviations' from the Normal distribution (Box, 1953; Driscoll, 1996; Faraway, 2006; Kikvidze & Moya-Laraño, 2008; Schmider, Ziegler, Danay, Beyer, & Bühner, 2010). We used 'robustness' to indicate the non-requirement of the Normal, but this word has a subjective origin and interpretation. Robustness opens up possibilities for an alternative criterion of data transformation that enables the use of parametric statistics, even when there are violations of assumptions. Thinking about robustness, data transformation was processed for dormant seeds in *Senna macranthera* to ensure homogeneity of variances. Based on the same criterion, we performed data transformation for seedling vigour in *Zea mays* to ensure the effects of additivity of blocks and treatments. In both cases, we used data transformation even in the face of loss of Normal distribution. That decision can be questioned by low probabilities associated with statistics of Kolmogorov-Smirnov test ($p < 0.01$), which could have a conflict with the idea of "small deviations". However, we highlight the lack of numeric and nominal limits in the literature of how much or what are "small deviations" from normality. Although not observed in our data sets, the meeting of two assumptions would be an intuitive criterion for data transformation.

The loss of Normal distribution would be enough to not recommend data transformation (Rice & Gaines, 1989; Ahrens, Cox, & Budhwar, 1990; Warton & Hui, 2011; Sileshi, 2012). However, we have no knowledge of recommendations for this and other situations mentioned here. We also observed that when data transformation is suggested, the models require specific statistical knowledge, reducing the independence of the scientist in relation to the models of analysis of variance. A more robust alternative for the angular scale, for example, are the Generalised Linear Models (GLM), which enable the analysis of the data with other probability distributions (other than Normal) (Nelder & Wedderburn, 1972; McCullagh, 1984). These models have a high power of comparison of results and ease of interpretation (Warton & Hui, 2011). It is important to remember that ANOVA is a particular case of GLM used for data sets based on Normal distribution. Thus, we cannot ignore the requirement of residuals adjusted to the Normal distribution even when using modern statistical models, as GLM. In view of this relevance, "data normalisation" before applying the *F* statistic is a very common practice in medical and biological sciences (Valcu & Valcu, 2011).

The decision to perform data transformation when the variances are heterogeneous is unquestionable, since several authors report the implications of heteroscedasticity for the *F* distribution and, consequently, the increase in Type II error (Ahmad et al*.,* 2006; Moder, 2007). Type II error can affect the ANOVA, but its greater implication is in nonparametric statistics. This is the probable cause of similar effects found for the different treatments analysed by nonparametric statistics (Kruskal-Wallis test), which was used to analyse an uncertainty of seed germination of *Enterolobium contortisiliquum*. The same data set was also studied by ANOVA, demonstrating different groups for the treatments. Although singular, this case study highlights the problem of inferences from nonparametric tests (Box, 1953; Moder, 2010) and the efficiency of ANOVA even for heterogeneous variances. We know (and emphasise) that critiques on nonparametric statistics must be performed by data simulations. However, we used a case study to demonstrate if the inferences from nonparametric statistics can be robust in relation to parametric ones. This justifies 'non-orthodox' alternatives for the use of data transformation.

There are authors using generic indicators to perform data transformation. They may be hiding what led them to perform data transformation (you can see it in a rapid search in the materials and methods of published articles for important journals). It is possible that in some cases the worst decision was made in relation to data transformation, i.e., to transform all variables analysed in the study, although some of them had no assumption violations. We invalidate the argument that when data transformation does not solve the violations of assumptions it does not compromise the inferences. As a case study, we used seedling development of *Parkia pendula*. In this data set, if we had transformed the data based on infected seedlings (to solve the problem of heteroscedasticity), we would have promoted heteroscedasticity in the other development measurements and undermined the *F* statistic analysis of damaged seedlings and dormant seeds.

The coefficient of variation (*CV*) has been used by agrarian scientists as a tool to validate experimental trial results, as noted by Bowman (2001). This can produce a false perceptive that the *CV* is associated with non-adherence of residuals to Normal and heteroscedasticity and non-additive effects between blocks and treatments. Actually, a *CV* increase can be related to an increase in variance heterogeneity because the math expression used to calculate the *CV* is indirectly based on variance error between treatments. However, this is not always true! From this point of view, our findings are a contribution to agrarian science. We show that the *CV* is not an indicator for data transformation. In contrast, the *CV* as an indicator may be dangerous for the quality of statistical inferences. Thus, the use of classifications of the *CV* (Pimentel-Gomes, 2000 sense) as an undeclared and empirical indicator of data processing can be a serious problem for data analyses involving ANOVA. It is important to note that the *CV* classification may be used to detect uncontrolled experimental variations, as suggested in its theoretical conception (see Pimentel-Gomes, 2000), although we did not analyse this.

The use of the *CV* classification as a criterion of data transformation was easily contradicted by the germinability in *Guazuma ulmifolia* seeds. In this sense (Pimentel-Gomes, 2000 sense), values above 20% (high and very high) became an undeclared indicator for data transformation, while values below 20% (low and medium) led the scientist not to use it. For germinability in *Guazuma ulmifolia* seeds the low experimental *CV* (5.31%) was not able to predict heterogeneous variances ($F = 2.932$; $p = 0.017$). Non-Normal residuals of germination of *Enterolobium contortisiliquum* seeds on the transformed scale ($K\text{-}S = 0.154$; $p = 0.031$) were not detected by the *CV*, whose value was 6.64%. The validity of this criterion was also considered inappropriate for germination of *Acacia polyphylla*, *E. contortisiliquum*, *Mimosa caesalpiniaefolia* Benth. and *Peltogyne confertiflora* (Mart. ex Hayne) Benth., which even with low *CV*s (< 10%) had at least one problem in the assumption analysed by inference tests (Pereira & Santana, 2013). These results do not exclude the possibility of the coefficient of variation and the assumptions of the model being affected by the same factors, but it is not possible to establish a relationship of cause-and-effect between a descriptive measurement of variability (as *CV*) and a statistics inferential test, as Kolmogorov-Smirnov and Levene. Therefore, we believe that the relationship among homoscedasticity, residuals adherence to Normal and *CVs* (from 15 to 18%) of germination of *Mimosa scabrella* Benth., *Dalbergia miscolobium* Benth. and *Ormosia arborea* (Vell.) Harms (Pereira & Santana, 2013) must be of causal nature.

Non-controlled variations, especially problems in experimental conduct, are noted as a main cause of problems with variances. As much as these flaws could compromise experiments, they do not cause heteroscedasticity in a data set. The major generators of heterogeneous variance for ANOVA models are the presence of zeros in the data set and the choice of treatments with previously expected discrepant answers (the so-called effect of scale), as seeds in viability extremes, i.e., 10 and 90% of germination (Bartlett, 1936; Ahrens et al., 1990; Sakia, 1992; Lúcio, Couto, Trevisan, Martins, & Lopes, 2010). The presence of zeros was the main cause of data heteroscedasticity of hard seeds of *Parkia pendula* and *Senna macranthera*, and the reason for these data to stay heteroscedastic even when transformed.

Criticised in the literature by statisticians, data transformation was considered by many authors as aberrant, inappropriate, out dated and, ironically, a real panacea (Sakia, 1992; Wilcox, 1998; Sileshi, 2007; Wartun & Hui, 2011; Osborne, 2010; Sileshi, 2012). These severe criticisms need to be reconsidered because (*i*) data transformation has been proposed by recognised scientists because of recommendations of these and other techniques widely applied in modern statistics (Bartlett, 1947; Box & Cox, 1964; 1982); and (*ii*) although the authors predicted that the technique would not solve all problems in adjusting residuals to Normal or of heteroscedasticity, there are several beneficial aspects regarding data transformation. Thus, the criticism that it does not solve assumption problems of the model in certain situations does not hold,

because data transformation limitations have been provided since 1947. We and other scientists recognise the problems in the interpretation of results in transformed scale (Ahrens et al., 1990; Fernandez, 1992; Sakia, 1992; Osborne, 2010), but its contribution to data sets with outliers is undeniable (Berry, 1987; Sakia, 1992).

Data transformation seems to be a good statistical tool for data sets from plant studies that have problems in adjustment of residuals to Normal, which is one of the priority principles to use parametric models prevailing in all science areas. However, to be an efficient statistical instrument, it is necessary to check the model assumptions when using data transformation (Osborne, 2010). We did not find, in the verified literature, any reference that data transformation, when assumptions are checked, is detrimental to the data set and (or) statistical inferences. In contrast, there is evidence that data transformation, when used properly, can increase the power of ANOVA (Levine & Dunlap, 1982). However, in the literature, cases of misuses till prevail, often encouraged by an inefficient empirical indicator that underestimates potential use of data transformation. We recommend the use of an additional criterion to promote data transformation, prioritising the transformation to perform parametric analysis.

Conclusion

The *CV* as an indicator of data transformation may be dangerous for the quality of statistical inferences. Data transformation can affect the assumptions for ANOVA, but the adjustments have poor relevance. The criterion based on robustness of *F* distribution can be useful for decisions about data transformation. Nonparametric statistics are less sensitive to detect differences in treatments for the increase in Type I error and Type II error.

Acknowledgements

We are grateful to the CAPES for the scholarship provided to the first author as Post-doctoral Researcher (PNPD); to Mr. Roger Hutchings for the English review of the manuscript.

References

Ahmad, W. M. A. W., Naing, N. N., & Rosli, N. (2006). An approached of Box-Cox data transformation to biostatistics experiment. *Statistika: Forum Teoridan Aplikasi Statistika*, *6*(2), 1-6.

Ahrens, W. H., Cox, D. J., & Budhwar, G. (1990). Use of the arcsine and square root transformations for subjectively determined percentage data. *Weed Science*, *38*(4-5), 452-458.

Bartlett, M. S. (1936). The square root transformation in analysis of variance. *Journal of Royal Statistical Society*, *3*(1), 68-78.

Bartlett, M. S. (1947). The use of transformations. *Biometrics*, *3*(1), 39-52.

Berry, D. A. (1987). Logarithmic transformations in ANOVA. *Biometrics*, *43*(2), 439-456.

Bowman, D. T. (2001). Common use of the CV: a statistical aberration in crop performance trials. The *Journal of Cotton Science*, *5*(2), 137-141.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, *40*(3-4), 318-335.

Box, G. E. P., & Cox D. R. (1964). An analysis of transformations. *Journal of Royal Statistical Society*, *26*(2), 211-252.

Box, G. E. P., & Cox, D. R. (1982). An analysis of transformations revisited, rebutted. *Journal of the American Statistical Association*, 77(377), 209-210.

Brasil. (2009). *Regras para análise de sementes*. Brasília, DF: MAPA.

Brown M. B., & Forsythe A. B. (1974). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69*(346), 364-367.

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*(6), 1701-1707.

Driscoll, W. C. (1996). Robustness of the ANOVA and Tukey-Kramer statistical tests. *Computer and Industrial Engineering*, *31*(1-2), 265-268.

Faraway, J. J. (2006). *Extending the linear model with R*: Generalised linear, mixed effects and nonparametric regression models. Boca Raton, FL: Chapman and Hall.

Fernandez, G. C. J. (1992). Residual analysis and data transformations: important tools in statistical analysis. *HortScience*, *27*(4), 297-300.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory Language*, *59*(4), 434-446.

Judice, M. G., Muniz, J. A., & Carvalheiro, R. (1999). Avaliação do coeficiente de variação na experimentação com suínos. *Ciência e Agrotecnologia*, *23*(1), 170-173.

Keene, O. N. (1995). The log transformation is special. *Statatiscs in Medicine*, *14*(8), 811-819.

Kikvidze, Z., & Moya-Laraño, J. (2008). Unexpected failures of recommended tests in basic statistical analyses of ecological data. *Web Ecology*, *8*(1), 67-73.

Labouriau, L. G. A. (1983). *Germinação de sementes*. Washington, D.C.: Secretaria Geral da Organização dos Estados Americanos.

Levine, D. W., & Dunlap, W. P. (1982). Power of the *F* test with skewed data: Should one transform or not? *Psychological Bulletin*, *92*(1), 272.

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. *Review of Educational Research*, *66*(4), 579-619.

Lúcio, A. D., Couto, M. R. M., Trevisan, J. N., Martins, G. A. K., & Lopes, S. J. (2010). Excesso de zeros nas variáveis observadas: estudo de caso em experimento com brócolis. *Bragantia*, *69*(4), 1035-1046.

Lúcio, A. D., Schwertner, D. V., Haesbaert, F. M., Santos, D., Brunes, R. R., Ribeiro, A. L. P., & Lopes, S. J. (2012). Violação dos pressupostos do modelo matemático e transformação de dados. *Horticultura Brasileira*, *30*(3), 415-423.

Maguire, J. D. (1962). Speed of germination-aid in selection and evaluation for seedling emergence and vigour. *Crop Science*, *2*(2), 176-177.

Manly, B. F. J. (1976). Exponential data transformations. *Journal of the Royal Statistical Society*, *25*(1), 37-42.

McCullagh, P. (1984) Generalised linear models. *European Journal of Operational Research*, *16*(3), 285-292.

Moder, K. (2007). How to keep the type I error rate in ANOVA if variances are heteroscedastic. *Austrian Journal of Statistics*, *36*(3), 179-188.

Moder, K. (2010). Alternatives to *F*-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling*, *52*(4), 343-353.

Nelder, J. A., & Wedderburn, R. W. M. (1972) Generalised linear models. *Journal of the Royal Statistical Society*, *135*(3), 370-384.

O'Hara, R. B., & Kotze, D. J. (2010).Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 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. *Ciencia e Agrotecnologia*, *33*(1), 113-119.

Osborne, J. W. (2010). Improving your data transformations: applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, *15*(12), 1-9.

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. *Journal of Seed Science*, *35*(2), 161-170.

Pimentel-Gomes, F. (2000). *Curso de estatística experimental.* Piracicaba, SP: Editora da Universidade de São Paulo.

Pontes, A. C. F., & Correntes, J. E. (2001). Comparações múltiplas não-paramétricas para o delineamento com um fator de classificação simples. *Revista de Matemática Estatística*, *19*(1), 179-197.

Quinn, G. P., & Keough, M. (2002).*Experimental design and data analysis for biologists*. Cambridge, UK: Cambridge University Press.

Ranal, M. A., & Santana, D. G. (2006). How and why to measure the germination process?. *Brazilian Journal of Botany*, *29*(1), 1-11.

Ranal, M. A., Santana, D. G., Ferreira, W. R., & Mendes-Rodrigues, C. (2009). Calculating germination measurements and organizing spreadsheets. *Brazilian Journal of Botany*, *32*(4), 849-855.

Ribeiro-Oliveira, J. P., & Ranal, M. A. (2016). Sample size in studies on the germination process. *Botany*, *94*(2), 103-115.

Ribeiro-Oliveira, J. P., Ranal, M. A., & Santana, D. G. (2013). A amplitude amostral interfere nas medidas de germinação de *Bowdichia virgilioides* Kunth.? *Ciência Florestal*, *23*(4), 623-634.

Rice, W. R., & Gaines, S. D. (1989). One-way analysis of variance with unequal variances. *Proceedings of the National Academy of Sciences*, *86*(21), 8183-8184.

Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *Journal of the Royal Statistics Society*, *41*(2), 169-178.

Santana, D. G., & Ranal, M. A. (2004). *Análise da germinação*: um enfoque estatístico. Brasília, DF: UnB.

Scheffé, H. (1959). *The analysis of variance*. New York, NY: Willey.

Schmider, E., Ziegler, M., Danay, E., Beyer, L., & Bühner, M. (2010). Is it really robust? Reinvestigating the robustness of ANOVA against violations of the normal distribution assumption. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(4), 147-151.

Sileshi, G. W. (2007). Evaluation of statistical procedures for efficient analysis of insect, disease and weed abundance and incidence data. *East African Journal of Science*, *1*(1), 1-9.

Sileshi, G. W. (2012). A critique of current trends in the statistical analysis of seed germination and viability data. *Seed Science Research*, *22*(3), 1-15.

Sokal, R. R., & Rohlf, F. J. (2012). *Biometry: the principles and practice of statistics in biological research.* New York, NY: W. H. Freeman and Co.

Souza, R. A., Hungria, M., Franchini, J., Chueire, L. M. O., Barcellos F. G., & Campo, R. J. (2008). Avaliação qualitativa e quantitativa da microbiota do solo e da fixação biológica do nitrogênio pela soja. *Pesquisa Agropecuária Brasileira*, *43*(1), p. 71-82.

Steel, R. G. D., & Torrie, J. H. (1996). *Principles and procedures of statistics*. New York, NY: McGraw Hill Book Company Inc.

Valcu, M., & Valcu, C. M. (2011). Data transformation practices in biomedical sciences. *Nature Methods*, *8*(2), 104-105.

Warton, D. I., & Hui, F. K. C. (2011). The arcsine is asinine: the analysis of proportions in ecology. *Ecology*, *92*(1), 3-10.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist*, *53*(3), 300-314.

Yamamura, K. (1999). Transformation using (x + 0.5) to stabilize the variance of populations. *Researches on Population Ecology*, *41*(3), 229-234.

Zar, J. H. (1999). *Biostatistical analysis*. India: Pearson Education India.

Zubin, J. (1935). Note on a transformation function for proportions and percentages. *Journal of Applied Psychology*, *19*(2), 213-220.

Received on February 8, 2017.
Accepted on June 26, 2017.

License information: This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

(c) 2018. This work is licensed under https://creativecommons.org/licenses/by/3.0/ (the “License”). Notwithstanding the ProQuest Terms and conditions, you may use this content in accordance with the terms of the License.