Journal of Seed Science
ISSN 2317-1545
www.abrates.org.br/revista

Journal of Seed Science, v.40, n.3, p.281-287, 2018
http://dx.doi.org/10.1590/2317-1545v40n3185259

Why analyze germination experiments using Generalized Linear Models?[1]

Fábio Janoni Carvalho[2*], Denise Garcia de Santana[3], Lúcio Borges de Araújo[3]

ABSTRACT - We compared the goodness of fit and efficiency of models for germination. Generalized Linear Models (GLMs) were performed with a randomized component corresponding to the percentage of germination for a normal distribution or to the number of germinated seeds for a binomial distribution. Lower levels of Akaikes's Information Criterion (AIC) and Bayesian Information Criterion (BIC) combined, data adherence to simulated envelopes of normal plots and corrected confidence intervals for the means guaranteed the binomial model a better fit, justifying the importance of GLMs with binomial distribution. Some authors criticize the inappropriate use of analysis of variance (ANOVA) for discrete data such as copaiba oil, but we noted that all model assumptions were met, even though the species had dormant seeds with irregular germination.

Index terms: AIC, ANOVA assumptions, *Copaifera langsdorffii* Desf, forest species.

Por que analisar experimentos de germinação usando os Modelos Lineares Generalizados?

RESUMO - A qualidade do ajuste e eficiência de modelos de germinação foram comparadas. Modelos Lineares Generalizados (MLGs) foram executados com o componente aleatório correspondendo ao percentual de germinação para a distribuição normal e o número de sementes germinadas para a distribuição binomial. Baixos valores do Critério de Informação de Akaike (AIC) e do Critério de Informação Bayesiano (BIC), ajuste aos envelopes simulados dos Normal *plots* e intervalos de confiança corretos para as médias justificam a importância do uso dos MLGs com distribuição binomial. Alguns autores criticam o uso inapropriado da análise de variância (ANOVA) para dados discretos como a germinação de copaíba, mas todas as pressuposições do modelo foram atendidas, mesmo a espécie possuindo sementes dormentes e germinação irregular.

Termos para indexação: AIC, pressuposições da ANOVA, *Copaifera langsdorffii* Desf, espécies florestais.

Introduction

Analysis of variance (ANOVA) is one of the most applied statistical models in agronomic experiments, including seed science. Based on a normal linear model, the method emphasizes the crucial role of replication, randomization and local control for efficient analyses (Sokal and Rohlf, 1995) and some assumptions need to be met before its execution. The residuals must follow a normal distribution and be independent; the variances need to be homogeneous; and the blocks should have an additive effect when the experiment follows a Randomized Block Design. One of the reasons for the prevalence of ANOVA in seed science is the low residual variability of crop species with decades of plant breeding, generating higher germination standards and contributing for to the non-checking of data to a normal distribution. However, this fact is not a guarantee for the use of ANOVA, and the assumptions always need to be verified. The use of techniques, such as data transformation, allowed the use of ANOVA for several years, despite its rigid assumptions. With confused interpretation and problems in the analysis, data transformation received severe criticism from researchers (Warton and Hui, 2011; Sileshi, 2012; Stroup, 2015).

Generalized Linear Models (GLMs) include distributions with fewer requirements than normal. Therefore ANOVA's assumptions do not necessarily need to be met (Wilson

[1]Submitted on 09/13/2017. Accepted for publication on 05/24/2018.
[2]Coordenação Geral de Pesquisa, Pós-Graduação e Inovação, Instituto Federal de Educação Ciência e Tecnologia do Triângulo Mineiro, 38064-790 - Uberaba, MG, Brasil.
[3]Instituto de Ciências Agrárias, Universidade Federal de Uberlândia, 38400-902 - Uberlândia, MG, Brasil.
*Corresponding author <fabiojanoni@ufu.br>

and Hardy, 2002; Crawley, 2007). GLMs are defined by a probability distribution that belongs to the exponential parametric family (Nelder and Wedderburn, 1972). They are the flexibilization of classic linear models for continuous variables, extending the whole structure for estimation and prediction to models with other distributions, including discrete variables (Dobson and Barnett, 2008).

It is common to analyze germination expressed as percentage instead of as the number of germinated seeds. The number of germinated seeds represents the original variable that is discrete and follows all binomial distribution criteria: fixed number of seeds, independence of germination, only two possible results (seed germinates or does not germinate) and, constant chance of germination (Lee et al., 2006). Based on this distribution, other statistical approaches, such as GLMs, can be used without data transformation and possibly guarantee a better fit (Dobson and Barnett, 2008).

Since ANOVA is a particularization of GLMs, we will use a classic experiment involving germinating seeds of copaiba oil (*Copaifera langsdorffii* Desf.), a native specie with a wide geographical distribution in the Brazilian territory, to show the relation between both methods of analysis (GLMs and ANOVA). This experiment also aimed to compare the goodness of fit and efficiency for the proposed models expressed as seed germination percentage and number of germinated seeds.

Material and Methods

A study with seeds from 13 individuals of copaiba oil was chosen to represent a classic germination experiment involving a species distributed in almost all Brazilian biomes (Atlantic and Amazon forests, Cerrado, Caatinga and Pantanal), with extensive genetic variability and dormant seeds. Copaiba oil seeds were arranged in a completely randomized design with four replications (r=4) of 25 seeds per plot (n=25) in a factorial scheme 4 x 3, with the first factor being the methods to overcome dormancy. The second factor was three samples differentiated by their physiological quality (high, intermediate and low quality), totaling 48 experimental plots (t=48). The methods to overcome dormancy consisted of disinfection of seeds with 0.05% of sodium hypochlorite for 5 minutes (1), soaking the seeds in water for 24 hours (2) and for 48 hours (3). One treatment was used as a control group (4). Seeds were distributed alternately in two sheets of filter paper, covered with another two to make the rolls. The rolls were distributed in a BOD germinator according to the experimental design and incubated at 25 °C under continuous fluorescent white light. The number of germinated seeds was quantified 35 days after sowing and the criterion adopted was normal seedlings.

The GLMs were performed with a randomized component corresponding to the percentage of germination for normal distribution or to the number of germinated seeds for binomial distribution; a systematic component corresponding to methods, samples and interaction in the form of a linear structure for both distributions; and a link function. It should be noted that the application of GLMs can be extended to any species and germination experiments with a factorial structure, such as germinability, normal seedlings, abnormal seedlings, dead seeds among others. Linear combination of the effects was given by:

$\eta_i = \sum_{j=1}^{p} x_j \beta_j = x_i^T \beta$ or $\eta = X\beta$; where $X = (x_1, x_2, ..., x_n)^T$ is the matrix of the model composed of explanatory components (methods, samples and interaction), x_i^T is the i^{th} row of experimental matrix X, $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the vector of parameters and $\eta = (\eta_1, \eta_2, ..., \eta_n)^T$ is the linear predictor.

The linear predictor η appears in the linear model as the sum of each term of the p parameters and it was obtained by transforming the percentage of germination or the number of germinated seeds with their respective link functions. The expected value of y was obtained by applying the inverse of the link function in η:

$\eta = \mu_p + \alpha_i + \delta_j + \gamma_j$; where μ_p is the mean of the predictor, α_i the effect of the i^{th} method, δ_j the effect of the j^{th} sample and γ_{ij} the effect of the interaction of the i^{th} method and the j^{th} sample.

The identity link function was applied to the percentage of germination where $g(\mu_i) = \mu_i$. This model requires residuals normality and independence, tested by the Shapiro-Wilk (Shapiro and Wilk, 1965) and Durbin-Watson test (Durbin and Watson, 1950), respectively, as well as variance homogeneity, checked by Levene's test (Levene, 1960).

The normal model was defined by: $y = X\beta + \varepsilon$ for $E(Y_i) = \mu_i = x_i^T \beta \quad Y_i \sim N(\mu_i, \sigma^2)$; where $y = (Y_1, Y_2, ..., Y_N)$ is the vector composed of the percentage of germination, $X = (x_1, x_2, ..., x_n)^T$ is the matrix of the model composed of explanatory components (methods, samples and interaction), $\beta = (\beta_1, \beta_2, ..., \beta_p)^T$ is the vector of parameters and $\varepsilon = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_N)$ is the vector composed of residuals.

The logit link function was applied for the number of germinated seeds with the binomial model defined

as: $g(\pi_i) = x_i^T \beta$ and $E(Y_i) = n\pi_i$; where π_i is the vector composed of germinated seed proportions, x_i^T the i^{th} row of experimental matrix X and n number of seeds per replication.

The *deviances* were estimated according to data distribution for each factor (method and sample) and interaction, as well as for null and saturated models, whence residual *deviance* derives. Inferences of the *deviance* analysis for binomial distribution were based on chi-square statistic, because the dispersion parameter is known. All the analysis were performed with software R version 3.4.1, considering the significance level of α=0.01 for all applied tests.

Results and Discussion

The non-significant effect for method and the significant effect for sample and interaction reveals equal inference about the sources of variation for both models (binomial and normal) (Table 1). Nevertheless, the p-value associated with mean effects and interaction were different between models. For example, probability for the method factor was 0.166 at the normal and 0.300 at the binomial distribution, which showed discrepancies between models. It is worth mentioning that a normal distribution could only be used because residuals from the germination percentage converged to this distribution (W = 0.976; P = 0.424) and the variances were homoscedastic (F = 2.517; P = 0.018).

The D^2 values from both models were slight equal, with 91.97% and 89.39% of the deviance explained by the insertion of the factors and interaction, for the normal and binomial model, respectively. The Sample variation was the factor with more deviation in the germination process, justified by the large germination difference between the three samples.

Similarities in inferences can lead to a misunderstanding that both models could be applied for germination of copaiba oil seeds. This similarity is apparent and must be contested. It is also necessary to consider the combination of descriptive and graphic measurements to assess the goodness of fit from models. The lowest values of AIC (205.87) and BIC (266.95) from the binomial compared to the normal distribution indicated that the first distribution was more generalist than the second (Table 1). Therefore, generalist models explain all the data extensions more efficiently.

Normal plot graphs complemented what was previously observed by AIC and BIC, and presented better adjustment of studentized residuals in the binomial model, with fewer points distant from the linear line, including the points close to and far from the mean (Figure 1a). Better accommodation of values in the simulated envelopes was also verified in the binomial model. For Cook's distance, all values associated were lower than 0.25 (Figure 1b) revealing no *outliers* (sensu

Table 1. Analysis of *deviance* (ANODEV) to seed germination of copaiba oil (*Copaifera langsdorffii* Desf.), for normal distribution with identity link and binomial distribution with logit link planned in a completely randomized design factorial scheme (method, sample and interaction).

Source of variation	*df*	*df* dif.	Normal distribution/identity link (ANOVA)				
			Deviance	*Deviance* Difference	*% of ED*	*F*	*P*
Null	0	47	27456.0				
Method	3	44	329.0	27127.0	1.20	1.79	0.166
Sample	2	42	22483.0	4644.0	81.89	183.62	< 0.001
Interaction	6	36	2440.0	2204.0	8.89	6.64	< 0.001
Saturated	11	36	2204.0		91.97*		
		AIC	345.91				
		BIC	370.23				
Source of variation	*df*	*df* dif.	Binomial distribution/logit link				
			Deviance	*Deviance* Difference	*% of ED*		*P*
Null	0	47	349.10				
Method	3	44	3.67	345.44	1.05		0.300
Sample	2	42	275.40	70.04	78.89		< 0.001
Interaction	6	36	33.01	37.03	9.46		< 0.001
Saturated	11	36	37.03		89.39*		
		AIC	205.87				
		BIC	266.95				

df: degrees of freedom; F: statistic of Snedecor distribution; ED: Explained deviance.*Also called D^2.

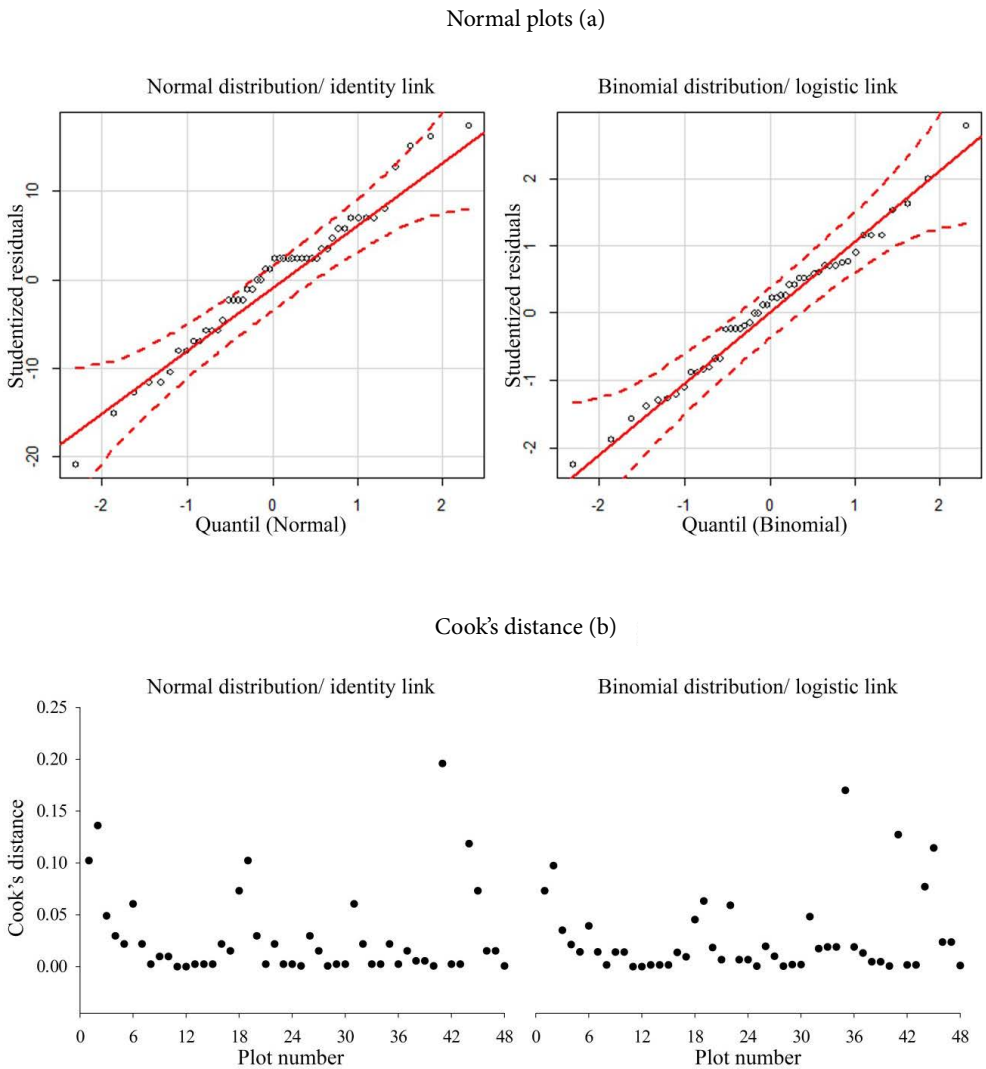

Figure 1. Normal plots including simulated envelopes of 95% confidence interval (a) and Cook's distance for discrepant data diagnosis (b) of copaiba oil (*Copaifera langsdorffii* Desf.) for normal and binomial distributions.

Cook and Weisberg, 1982). In the copaiba experiment, Cook's distance did not prove effective to distinguish the models because the data did not presented *outliers*.

The confidence intervals (CI) for the means of binomial model were more reliable than those from normal model, where treatments with larger variation had higher CI, differing from the normal model that sets the same CI for all means (Figure 2). It occurred because standard error for binomial model is calculated as $\sqrt{\pi(1-\pi)/n}$, where π is the germination seed proportion. In the normal model, standard error is calculated as $\sigma/\sqrt{n}$, where σ is the standard deviation of the sample, not considering the treatment variation but the experimental variation. Another advantage of the binomial model is that the estimated means were in the range of 0 to 100, because it is a discrete distribution. Normal follows a continuous distribution, varying from -∞ to +∞, and the treatments with germinations close to zero and 100, can have their values extrapolated, as example the methods 2 and 3 for the third sample (Figure 2).

Lower levels of AIC and BIC combined and data adherence to simulated envelopes of Normal plots guaranteed the binomial model a better fit, justifying the importance to select models with better adjustment to the data through GLMs. However, the purpose of this research was not only to increase the successful list of binomial models for categorical data (Jaeger, 2008; Sileshi, 2012), but also to propose a critical analysis of why these suitability models are sometimes ignored by researchers and why ANOVA is still prevalent for germination data.

One of the reasons for misuse of modern statistics may

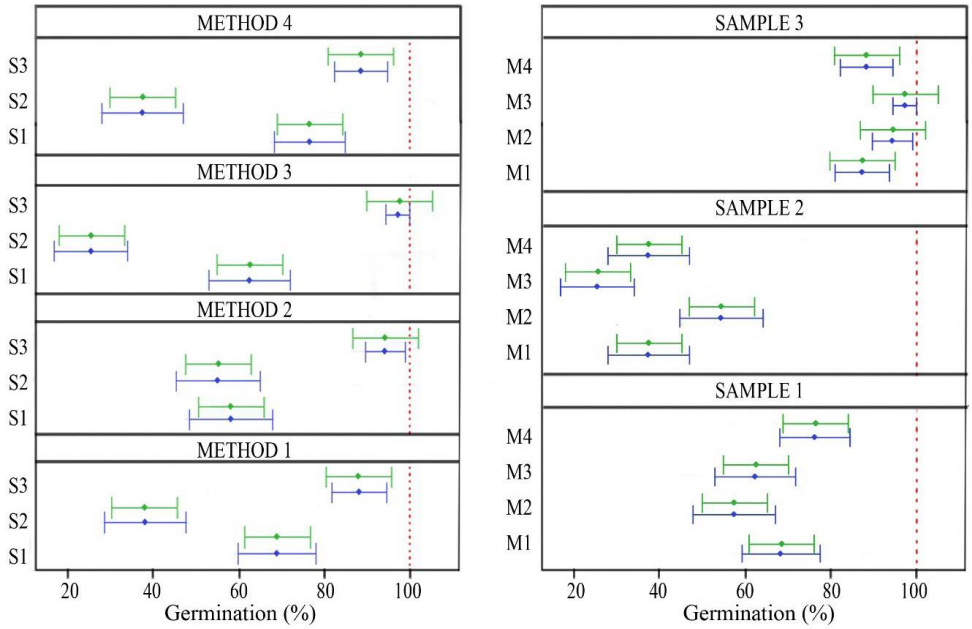

Figure 2. Germination means with confidence intervals of copaiba oil (*Copaifera langsdorffii* Desf.) of three samples submitted by four methods to overcome seed dormancy. Green lines represent the means from normal model and blue lines from the binomial model.

lie in the fact that among the models that are part of GLMs, a normal distribution and identity link are only appropriate if ANOVA assumptions are met. ANOVA is part of GLMs, so it is reasonable to assume that Fisher had already applied modern statistics since 1925. Most of the published articles performing GLMs are not very precise, since the assumptions and model requirements were not checked. Generalization does not suggest unconcern for checking assumptions, but a distribution flexibility to ensure the best model for the data at hand (Crawley, 2007; Dobson and Barnett, 2008).

The non checking of ANOVA assumptions is a recurrent problem in seed science. In germination articles published between 2000 and 2011, only 20% of those checked the normality assumption (Sileshi, 2012). However, this problem is not a peculiarity of germination experiments and recurs in several areas of biological, agricultural and environmental sciences. The affirmation that the normal distribution is not prevalent in forest species (Austin, 1987; Biondini et al., 1988) was debunked by the copaiba oil results. Some authors criticize the inappropriate use of ANOVA for discrete data such as copaiba oil (Jaeger, 2008; Sileshi, 2012), but we note that all model assumptions were met, even though the species had dormant seeds with irregular germination. Moreover, when the number of observations is large, binomial distribution tends to a normal distribution. In this context, the angular transformation of data was discarded, minimizing any criticism regarding this approach. The meeting of all assumptions enabled the use of ANOVA to analyze the germination of the species.

One of the reasons for ANOVA's prevalence may be the vast literature on scientifically recognized *Post hoc* tests (Tukey, SNK, Scott-Knott, Duncan, Dunnett, among others) unlike binomial, which is restricted to orthogonal contrasts. It is unquestionable that contrasts are efficient in the comparison of means, but they are difficult to interpret when compared to classical tests. Futhermore, ANOVA specialized in more complex factor models, such as split plot in time and space, additional treatments, among others, which are easily found by researchers in statistical programs. In this context of GLMs, these factor structures will probably require more modeling, which would limit the autonomy of researchers.

ANOVA remained absolute in germination experiments because of the lack of quality indicators for the model. The coefficient of variation prevailed and still prevails as the main indicator of experimental precision, but it does not bear any relation to the goodness of fit. The coefficient of variation supremacy is a reflection of Snedecor and Cochran's approach (1967) that used the measurements proposed by Karl Pearson in 1895 to compare the relative variation of different crops. As a consequence of this widespread use, reference values were determined for various crops, and when lower than 15%, the experiments were considered to have

high experimental precision, an insufficient adjective to make any assumptions regarding the goodness of fit.

It is important in our discussion to report overdispersion in the analysis of discrete data. Overdispersion occurs because the mean and variance components of a GLM are related and depends on the same parameter that is being predicted through the independent vector. For binomial data, overdispersion occurs when the discrepancies between the observed responses and their predicted values are larger than what the model would predict. There is no overdispersion in ordinary linear regression because variance is estimated independently of the mean function (Agresti, 2012).

It is expected that residual *deviance* is approximately equal to the residual degrees of freedom, applying GLMs with a known scale parameter (the case for binomial and Poisson distributions). binomial data of copaiba oil germination did not express overdispersion because the relation between residual deviance and residual degrees of freedom was approximately equals to one (Table 1), eliminating any problem for the model.

Overdispersion makes the standard errors obtained from the model incorrect and may be seriously underestimated. Consequently, we may incorrectly assess the significance of individual regression parameter. Interpretation of the model will be incorrect and any predictions will be too inaccurate. A number of different models and associated estimation methods have been proposed to overcome overdispersion (Collett, 1991; Piepho, 1999).

With the normal distribution, estimates of the mean and variance require distinct calculations; with the binomial, a single calculation, the estimate of p, determines both the mean and variance. In general, for germination percentages higher than 50%, a normal distribution of the observations will be left-skewed; for percentages lower than 50%, the distribution will be right-skewed. The skewness increases as the probability approaches 0 or 100%, giving unbiased predicted values for normal distribution (Stroup, 2015). Forest species have a large germination variation and cultivated species have high germination standards reached by plant breeding, which makes the binomial distribution more reliable for germination studies.

Descriptive measurements to analyze goodness of fit are not new. AIC and BIC indicated that the binomial model is more efficient than the normal for copaiba oil seed germination. Meeting assumptions is necessary, but not a satisfactory reason for the use and application of ANOVA to germination data, which is the main justification to apply GLMs. Despite their convergence to point out binomial as the most appropriate distribution, AIC and BIC indicators are divergent. When the models are more complex and explain several variables with different degrees of interaction, studies have shown AIC to be more reliable than BIC. When the models are simpler, BIC is preferable (Nylund et al., 2007; Yang and Yang, 2007; Vrieze, 2012). Furthermore, both criteria pointed that the binomial performed better adjustment but sometimes one statistical inference may differ.

It is necessary to recognize the importance of each measurement and decide which one will be considered to evaluate the goodness of fit. Other germination experiments may fit more properly to a normal distribution model, hence the importance of using these inferential criteria. Graphic diagnostics are also important, but they depend on "a clinical look", which tends to be idiosyncratic, so these graphics should be applied jointly.

Graphic diagnostics take on more significance when different link functions are being used in a same distribution or when different distributions are being compared. For the germination variable, residual normality check (normal distribution) and its representation as the number of germinated seeds (binomial distribution) guarantee reliability for the analyst that these distributions are consistent. Other link functions can be adjusted for germination, which ensure the creation of new models. It is important to report that link functions were fixed for these distributions in our research, because the scope was only to compare model fitting to different data distribution.

We used canonical functions of their respective distributions because they simplify the model themselves (Myers et al., 2002; Crawley, 2007). If a tested model does not show a good adjustment and the graphical analysis demonstrates irregularity, new models with different link functions need to be tested, observing their own peculiarities. For example, a probit link function could be also used with binomial distribution to adjust the germination (Jaeger, 2008). Goodness of fit for different link functions in the same distribution could be verified with residual *deviance*, where the lowest values would indicate the best model.

Conclusions

This research demonstrates that Generalized Linear Models can be applied efficiently in seed science. This methodology can fit different statistical cases and allows the researcher to make new inferences, when other distributions from the exponential family could be considered. GLM with binomial approach performs better models for germination data.

References

AGRESTI, A. *Categorical Data Analysis.* New York: Wiley, 2012. 721p.

AUSTIN, M.P. Models for the analysis of species response to environmental gradient. *Vegetation*, v.69, p.35-45, 1987. https://link.springer.com/article/10.1007/BF00038685

BIONDINI, M.E.; MIELKE, P.W.; BERRY, K.J. Data-dependent permutation techniques for the analysis of ecological data. *Vegetation*, v.75, p.161-168, 1988. https://www.ndsu.edu/pubweb/~biondini/vita/0120_Vegetatio_1988_Biondini.pdf

CRAWLEY, M.J. *The R book.* England: Wiley, 2007. 942p.

COLLETT, D. *Modelling binary data.* London: Chapman and Hall, 1991. 408p.

COOK, R.D.; WEISBERG, S. *Residuals and influence in regression.* New York: Chapman and Hall, 1982. 229p.

DOBSON, A.J.; BARNETT, A.G. *An introduction to Generalized Linear Models.* New York: Chapman and Hall, 2008. 320p.

DURBIN, J.; WATSON, G.S. Testing for serial correlation in least squares regression. *Biometrika*, v.37, p.409-428, 1950. http://www.jstor.org/stable/2332391?origin=JSTOR-pdf&seq=1#page_scan_tab_contents

JAEGER, T.F. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, v.59, p.434-446, 2008. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2613284/

LEE, Y.; NELDER, J.A.; PAWITAN, Y. *Generalized Linear Models with random effects.* New York: Chapman and Hall, 2006. 416p.

LEVENE, H. Contributions to probability and statistics. In OLKIN, I. et al. *Essays in honor of Harold Hotelling.* California: Stanford University Press, 1960. p.278-292.

MYERS, R.H.; MONTGOMERY, D.C.; VINING, G.G. *Generalized Linear Models with applications in engineering and the sciences.* New York: John Wiley and Sons Press, 2002. 496p.

NELDER, J.A.; WEDDERBURN, R.W.M. Generalized linear models. *Journal of the Royal Statistical Society*, v.135, p.370-384, 1972. http://www.jstor.org/stable/2344614

NYLUND, K.L.; ASPAROUHOV, T.; MUTHÉN, B.O. Deciding on the number of classes in latente class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling*, v.14, p.535-569, 2007. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.9619&rep=rep1&type=pdf

PIEPHO, H.P. Analysing disease incidence data from designed experiments by generalized linear mixed models. *Plant Pathology*, v.48, p.668-674, 1999. http://onlinelibrary.wiley.com/doi/10.1046/j.1365-3059.1999.00383.x/abstract

SHAPIRO, S.S.; WILK, M.B. An Analysis of variance test for normality. *Biometrika*, v.52, p.591-611, 1965. http://www.jstor.org/stable/2333709

SILESHI, G.W. A critique of current trends in the statistical analysis of seed germination and viability data. *Seed Science Research*, v.22, p.145-159, 2012. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN1/22643

SNEDECOR, G.N.; COCHRAN, W.G. *Statistical methods.* Ames: Iowa State University Press, 1967. 593p.

SOKAL, R.R.; ROHLF, F.J. *Biometry:* the principles and practice of statistics in biological research. New York: W.H. Freeman, 1995. 776p.

STROUP, W.W. Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*, v.107, n.2, p.811-827, 2015. http://lira.pro.br/wordpress/wp-content/uploads/2015/06/stroup-2015.pdf

VRIEZE, S.I. Model selection and psychological theory: A discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). *Psychological Methods,* v.17, p.228-243, 2012. https://doi.org/10.1037/a0027127

WARTON, D.; HUI, F. The arcsine is asnine: the analysis of proportions in ecology. *Ecology,* v.92, p.3-10, 2011. https://doi.org/10.1890/10-0340.1

WILSON, K.; HARDY, I.C.W. Statistical analysis of sex ratios: an introduction. In: HARDY, I.C.W. *Sex ratios: Concepts and Research Methods*. Cambridge: Cambridge University Press, 2002. p.48-92.

YANG, C.; YANG, C. Separating latent classes by information criteria. *Journal of Classification*, v.24, p.183-203, 2007. https://link.springer.com/article/10.1007%2Fs00357-007-0010-1