Ecotoxicology is not normal.

How the use of proper statistical models can increase statistical power in ecotoxicological experiments.

Eduard Szöcs, Ralf B. Schäfer

February 12, 2015

Abstract

Ecotoxicologists are often confronted with non-normally distributed data. To meet the assumptions of normality and heteroscedasticity, the standard procedure has been to either transform the data or use non-parametric methods if this fails. Here, we compare the statistical power of analyses using transformed data or non-parametric methods to analyses using appropriate distributional assumptions, namely Generalised Linear Models (GLM).

We simulated data mimicking ecotoxicological experiments of two common data types (counts and proportions). We compare the performance of methods in terms of statistical power and type 1 error. In addition, we outlined differences and advantages of GLMs on a real world mesocosm data set.

We found that GLMs provide in most cases a gain in statistical power compared to analysis of transformed data or using non-parametric methods. We recommend that non-normal data should be analysed by GLMs and not by transformations or non-parametric methods. GLMs should become a standard method in ecotoxicology.

1 Introduction

Ecotoxicologists perform various kinds of experiments yielding different types of data. Examples are: animal counts in mesocosm experiments (positive, integer-valued data), proportions of surviving animals (data bonded between 0 and 1, continuous) or biomass in growth experiments (positive, continuous data). These data are typically not normally distributed. Nevertheless,

they are usually analysed using methods assuming a normal distribution and variance homogeneity (Wang and Riffel, 2011). To meet these assumptions, data are usually transformed. For example, ecotoxicological textbooks (Newman, 2012) and guidelines (EPA, 2002; OECD, 2006) advise that survival data can be transformed using an arcsine square root transformation. For count data from mesocosm experiments a log(Ay + C) transformation is usually applied, where the constants A and C are either chosen arbitrarily or following general recommendations. For example, van den Brink et al. (2000) suggest to set the term Ay to be 2 for the lowest abundance value (y) greater than zero and C to 1. Moreover, other transformations like the square root or fourth root are commonly applied in community ecology. Note that there has been little evaluation and advice for practitioners, which transformations to use. If the transformed data still do not meet the assumptions (i.e. normality and variance homogeneity), non-parametric tests are usually applied (Wang and Riffel, 2011).

Generalised linear models (GLM) provide a method to analyse such non-normally distributed data (Nelder and Wedderburn, 1972). GLMs can handle various types of data distributions, e.g. Poisson or negative binomial (for count data) or binomial (for proportions); the normal distribution being a special case of GLMs. Despite GLMs being available more than 40 years, ecotoxicologists do not regularly make use of them. Recent studies concluded that data transformations should be avoided and GLMs be used as they have better statistical properties (O'Hara and Kotze, 2010; Warton and Hui, 2011).

Ecotoxicological experiments often involve small sample sizes due to practical constraints. For example, extremely low samples sizes (n <5) are common in many mesocosm studies (Sanderson, 2002; Szöcs et al., 2015). Small sample sizes lead to low power in statistical hypothesis testing, on which many ecotoxiological approaches (e.g. risk assessment for pesticides) rely. Such an endpoint are L/NOEC (Lowest / No observed effect concentration) values. Although their use has been heavily criticized in the past (Laskowski, 1995), they are still regularly used in ecotoxicology (Jager, 2012). Especially in mesocosm studies L/NOEC calculations are used in the majority of mesocosm experiments (Brock et al., 2015; EFSA PPR, 2013).

We explore how GLMs may enhance inference in ecotoxicological studies and compared three types of statistical methods (transformation and normality assumption, GLM, non-parametric tests). We first illustrate differences between statistical methods using a data set from a mesocosm study. Then we further elaborate differences in detecting a general treatment effect and determining the LOEC using simulations of two common data types in ecotoxicology: counts and proportions.

2 Methods

2.1 Models for count data

2.1.1 Linear model for transformed data

To meet the assumptions of the standard linear model, count data usually needs to be transformed. We followed the recommendations of van den Brink et al. (2000) and used a log(Ay + 1) transformation (eqn. 1):

$$\begin{aligned} y_i^T &= log(Ay_i + 1) \\ A &= 2 \ / \ min(y) \quad \text{, for } y > 0 \end{aligned} \tag{1}$$

, where y_i is the measured abundance and y_i^T the transformed abundance.

Then we fitted the linear model to the transformed abundances (hereafter LM):

$$\begin{aligned} y_i^T &\sim N(\mu_i, \sigma^2) \\ y_i^T &= \alpha + \beta x_i \\ var(y_i^T) &= \sigma^2 \end{aligned} \tag{2}$$

This model assumes a normal distributed response with constant variance (σ^2). Note, that we parameterised the model as contrast (βx_i) to the control group (α) so that parameters (β) are directly interpretable as changes from the control group (eqn. 2).

2.1.2 Generalised Linear Models

GLMs extend the normal model by modelling other distributions. Instead of transforming the response variable, the counts could be directly modelled by a Poisson distribution (GLM_p):

$$\begin{aligned} y_i &\sim P(\lambda_i) \\ log(\lambda_i) &= \mu_i \\ \mu_i &= \alpha + \beta x_i \\ var(y_i) &= \lambda_i \end{aligned} \tag{3}$$

Again, this model was parametrised as contrast to the control group. The response variable is linked to the predictors via a log-function to avoid negative fitted values (eqn. 3). The

Poisson distribution assumes that mean and variance are equal - an assumption that is rarely met with ecological data, which is typically characterized by greater variance than the mean (overdispersion). To overcome this problem a quasi-Poisson distribution could be used which introduces an additional overdispersion parameter (Θ) (GLM_{qp}, eqn. 4).

$$y_i \sim P(\lambda_i, \Theta) \tag{4}$$
$$var(y_i) = \Theta\lambda_i$$

The quasi-Poisson model yields to parameter estimates equal to the Poisson model (eqn. 3), but with standard errors scaled by the degree of overdispersion.

Another possibility to deal with overdispersion is to fit a negative binomial distribution (GLM_{nb}, eqn. 5).

$$y_i \sim NB(\lambda, \kappa) \tag{5}$$
$$var(y_i) = \lambda_i + \kappa\lambda_i^2$$

In both cases the parametrisation and link function is equal to the Poisson GLM (eqn. 3). Note, that the quasi-Poisson model assumes a linear mean-variance relationship (eqn. 4), whereas the negative binomial model assumes a quadratic relationship (eqn. 5).

The above described models are most commonly used in ecology (Ver Hoef and Boveng, 2007), although other distributions for count data are possible, like the negative binomial model with a linear mean-variance relationship (also known as NB1) or the poisson inverse gaussian model (Hilbe, 2014).

2.2 Models for binomial data

2.2.1 Linear model for transformed data

To accommodate the assumptions for the standard linear model, a special arcsine square root transformation (eqn. 6) is suggested for such data (EPA, 2002; Newman, 2012):

$$y_i^T = \begin{cases} arcsin(1) - arcsin(\sqrt{\frac{1}{4n}}) & \text{, if } y_i = 1 \\ arcsin(\sqrt{\frac{1}{4n}}) & \text{, if } y_i = 0 \\ arcsin(\sqrt{y_i}) & \text{, otherwise} \end{cases} \tag{6}$$

, where y_i^T are the transformed proportions and n is the number of exposed animals per treatment ($n = 4 \cdot 10 = 40$). The transformed proportions are then analysed using the standard linear model (LM, eqn. 2). Note, that the parameters of the linear model are not directly interpretable due to transformation.

2.2.2 Generalised Linear Models

Data of type *x out of N* can be modelled by a binomial distribution with parameters N and π (GLM_{bin}):

$$\begin{aligned} y_i &\sim Bin(N, \pi_i) \\ logit\ (\pi_i) &= \alpha + \beta x_i \\ var(y_i) &= \pi_i(1 - \pi_i)/N \end{aligned} \tag{7}$$

, where N = number of exposed animals and π is the probability of survival. The variance of the binomial distribution is a quadratic function of the mean (eqn. 7). The parameters β of this model are directly interpretable as changes in log odds compared to the control group. Note, that there are also quasi-binomial models available if the assumed mean-variance relationship is not met.

2.3 Statistical Inference

After model fitting and parameter estimation the next step is statistical inference. Ecotoxicologists are generally interested in two hypotheses: (i) is there any treatment related effect? and (ii) which treatments show a treatment effect (to determine the LOEC)?

Following general recommendations (Bolker et al., 2009; Faraway, 2006), we used F-tests (LM and GLM_{qp}) and Likelihood-Ratio (LR) tests (GLM_p, GLM_{nb} and GLM_{bin}) to test the first hypothesis. However, it is well known that LR test are unreliable with small sample sizes (Wilks, 1938). Therefore, we additionally explored parametric bootstrap (Faraway, 2006) to assess the significance of the LR. Bootstrapping is computationally very intensive, therefore, we applied it only for the negative binomial models (using 500 bootstrap samples, denoted as GLM_{pb}).

To assess the LOEC we used Dunnett contrasts with one-sided Wald t tests (normal and quasi-Poisson models) and one-sided Wald Z tests (Poisson, negative binomial and binomial models). Beside these parametric methods we also applied two, in ecotoxicology commonly used, non-parametric methods: The Kruskal-Wallis test (KW) to test for a general treatment effect and a pairwise Wilcoxon test (WT) to determine the LOEC.

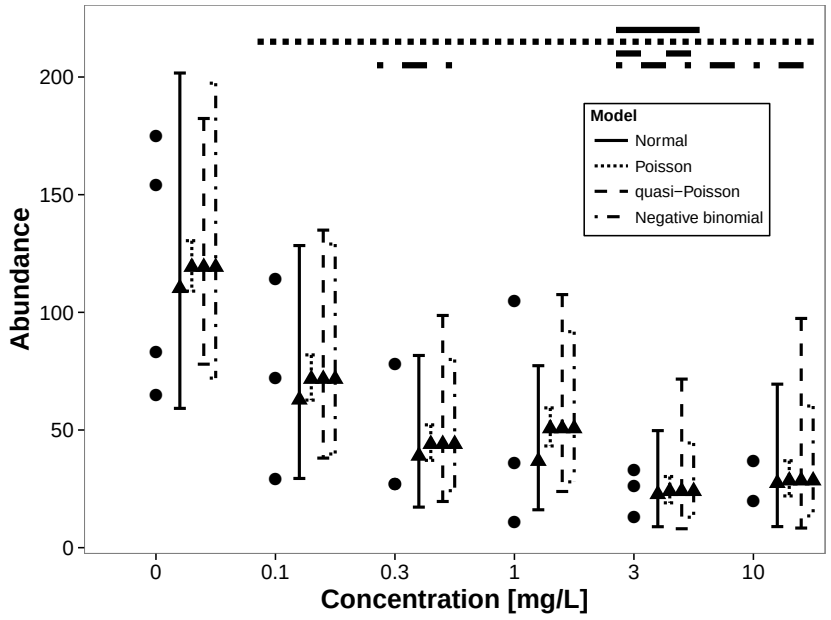

Figure 1: Data from Brock et al. (2015) (dots). Predicted values (triangles) and 95% Wald Z or t confidence intervals from the fitted models (vertical lines) are given beside. Horizontal bars above indicate treatments statistically significant different from the control group (Dunnett contrasts). The data showed considerable overdispersion ($\Theta = 22.41$) and therefore, the Poisson model underestimates the confidence intervals.

2.4 Case study

Brock et al. (2015) presents a typical example of data from mesocosm studies, which we use to demonstrate differences between methods. The data are mayfly larvae counts on artificial substrate samplers were at one sampling date. A total of 18 mesocosm have been sampled from 6 treatments (Control (n = 4), 0.1, 0.3, 1, 3 mg/L (n = 3) and 10 mg/L (n = 2)) (Figure 1).

2.5 Simulations

2.5.1 Count data

To further scrutinise the differences between methods we simulated data sets with known properties. We simulated count data that mimics the data of the case study with five treatments (T1 - T5) and one control group (C). Counts were drawn from a negative binomial distribution with slight over dispersion at all treatments ($\kappa = 0.25$, eqn. 5). We simulated data sets with different number of replicates (N = {3, 6, 9}) and different abundances in control treatments ($\mu_C = \{2, 4, 8, 16, 32, 64, 128\}$). For power estimation, mean abundance in treatments T2 - T5

was reduced to half of control and T1 ($\mu_{T2} = ... = \mu_{T5} = 0.5\ \mu_C = 0.5\ \mu_{T1}$), resulting in a theoretical LOEC at T2. Mean abundance was kept equal between all groups in Type 1 error simulations.

We generated 100 data sets for each combination of N and μ_C and analysed these using the models outlined previously. We did not fit Poisson models because we simulated data with overdispersion.

2.5.2 Binomial data

We simulated data from a commonly used design as in (Weber et al., 1989), with 5 treated (T1 - T5) and a control group (C). Proportions were drawn from a Bin(10, π) distribution, with varying probability of survival ($\pi = \{0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$) and varying number of replicates (N = {3, 6, 9}). For Type 1 error estimation, π was held constant between groups. For power estimation π in C and T1 was fixed at 0.95 and was set to values between 0.6 and 0.95 for the treatments T2 - T5. For each combination we simulated 250 data sets.

2.6 Data Analysis

We analysed the case study and the simulated data using the outlined methods. We compared the methods and models in terms of Type 1 error (maintain a significance level of 0.05 when there is no effect) and power (detect an effect when it is present). All computations were done in R (Version 3.1.2) (R Core Team, 2014) on a Linux machine. Source code for the simulations and analysis of the case study is available online at `https://github.com/EDiLD/usetheglm`.

3 Results

3.1 Case study

The data set showed considerable overdispersion ($\Theta = 22.41$, eqn. 4). Therefore, the Poisson model did not fit to this data and lead to underestimated standard errors and confidence intervals, as well as overestimated statistical significance (Figure 1). In this case, inferences on the Poisson model are not valid and we do not further discuss its results. The normal ($F = 2.57$, $p = 0.084$) and quasi-Poisson model ($F = 2.90$, $p = 0.061$), as well as the Kruskal test ($p = 0.145$) did not show a statistically significant treatment effects. By contrast, the LR test and parametric bootstrap of the negative binomial model indicated a treatment-related effect (LR = 13.99, $p = 0.016$, bootstrap: $p = 0.042$).

All methods predicted similar values, except the normal model predicting always lower abundances (Figure 1). 95% confidence intervals (CI) where most narrow for the negative binomial model and widest for the quasi-Poisson model - especially at lower estimated abundances. Consequently, the LOECs differed (Normal and quasi-Poisson: 3 mg/L, negative binomial: 0.3 mg/L). The pairwise Wilcoxon test did not detect any treatment different from control.

3.2 Simulations

3.2.1 Count data

For our simulation design (reduction in abundance by 50%) a sample size per treatment of n = 9 was needed to achieve a power greater than 80%. For detecting a treatment effect GLM_{nb}, GLM_{pb} and GLM_{qp} exhibited higher power than LM and KW, the latter having least power. Type 1 error rate was inflated for GLM_{nb}, but this could be fixed by using parametric bootstrap. KW was conservative at low sample sizes (Figure 2). At small sample sizes (n = 3, 6) and low abundances (μ_C = 2, 4) many of the negative binomial models (GLM_{nb} and GLM_{pb}) did not converge to a solution (convergence rate <80% of the simulations, Supplement 1).

The inferences on LOEC generally showed less power. For LM this reduction was up to 35% compared to the overall treatment effect (n = 9, μ_C = 64, Figures 2 and 3). The power to detect the LOEC was highest for GLM_{nb} and GLM_{pb}. LM and WT showed less power, with WT having no power to detect the LOEC at low sample sizes. At low sample sizes GLM_{nb} showed an increased Type 1 error and WT was slightly conservative (Figure 3).

3.2.2 Binomial data

GLM_{bin} showed the greatest power for testing the treatment effect. This was especially apparent at low sample sizes (n = 3), with up to 24% higher power compared to LM. KW had the lowest power and was slightly conservative. However, the differences between methods quickly vanished with increasing samples sizes. KW was more conservative than LM and GLM_{bin} (Figure 4).

Inference on LOEC was not as powerful as inference on the general treatment effect. Contrary to the general treatment effect, LM showed the higher power than GLM_{bin} at small sample sizes. However, these differences in power were only apparent at n = 3 and vanished quickly with increasing sample sizes (Figure 5). WT had no power for n = 3 and showed less power in the other simulation runs. LM maintained a Type 1 error level of 0.05 in all simulations. GLM_{bin} was conservative at small effect sizes ($p_E > 0.8$) and WT was generally conservative showing lowered Type 1 error rates (Figure 5).

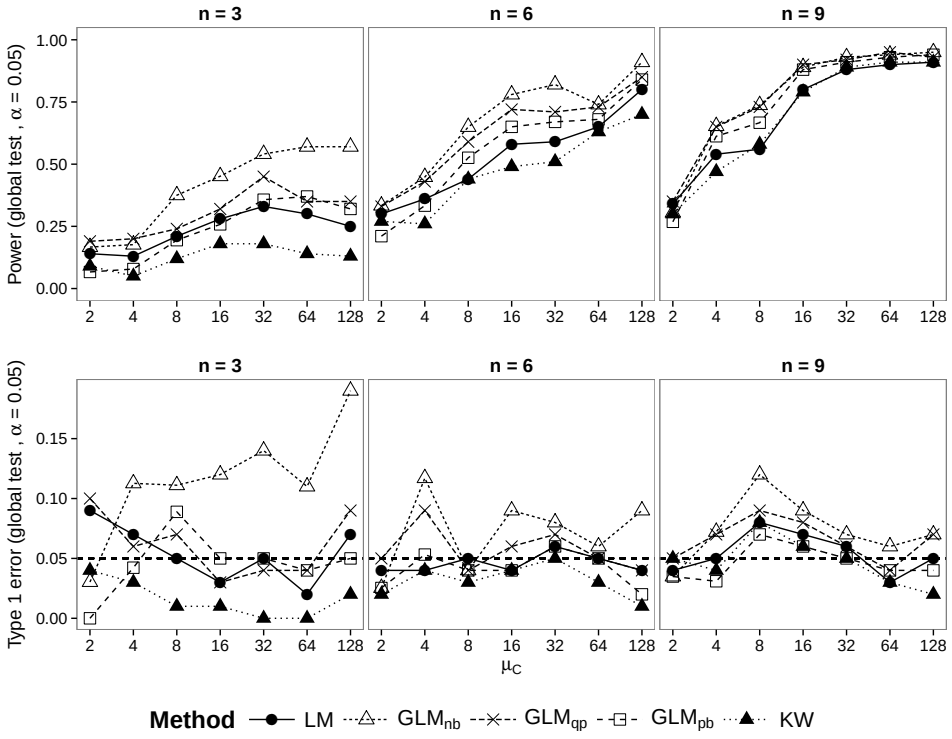

Figure 2: Count data simulations: Power (top) and Type 1 error (bottom) for the test of a treatment effect. For n = 3 and $\mu_C = \{2, 4\}$ less than 80% of GLM_{nb} and GLM_{pb} models did converge. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

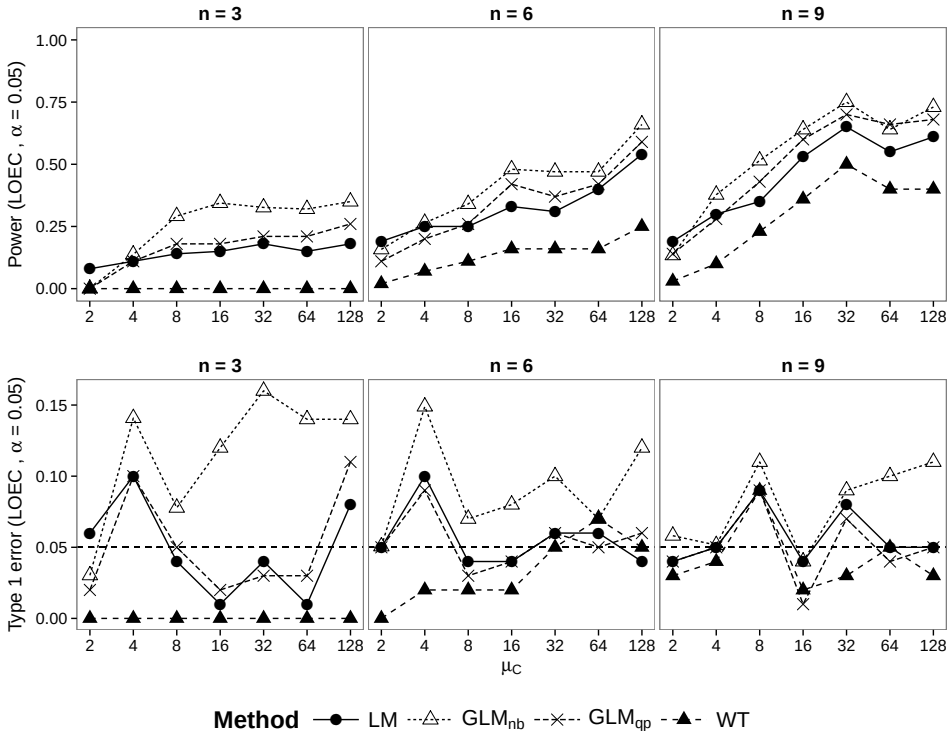

Figure 3: Count data simulations: Power (top) and Type 1 error (bottom) for determination of LOEC. For n = 3 and $\mu_C = \{2, 4\}$ less than 80% of GLM_{nb} and GLM_{pb} models did converge. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

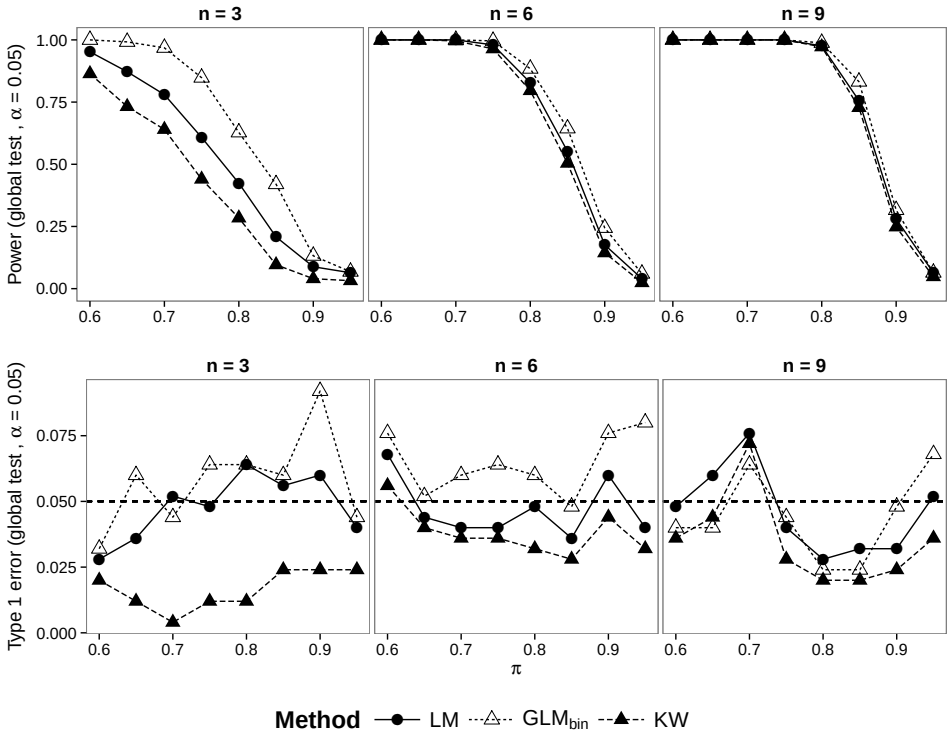

Figure 4: Binomial data simulations: Power (top) and Type 1 error (bottom) for the test of a treatment effect. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

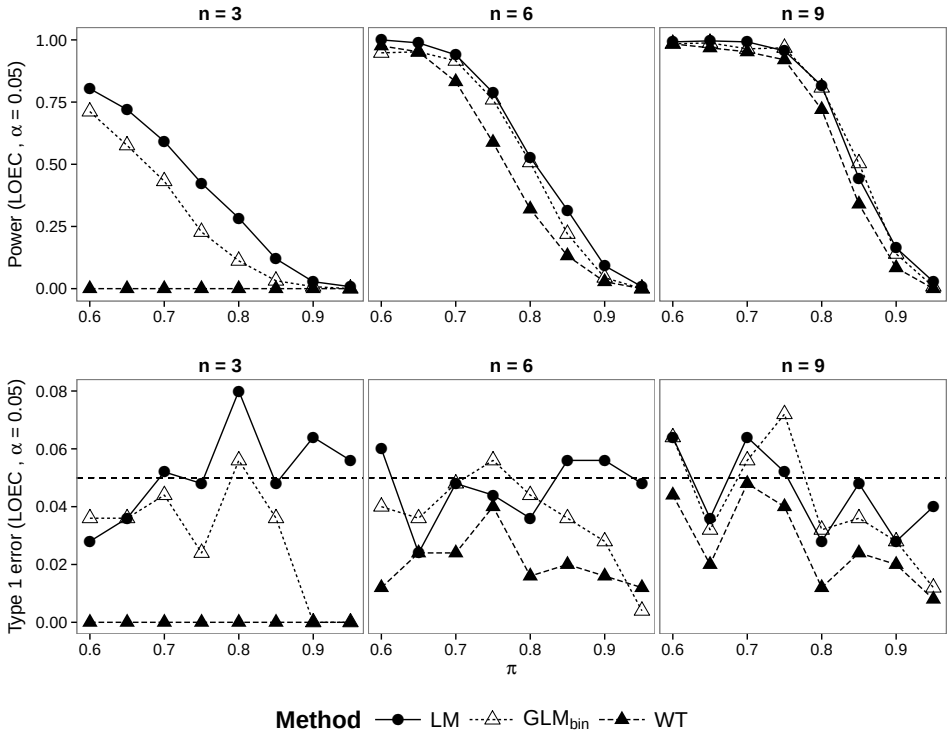

Figure 5: Binomial data simulations: Power (top) and Type 1 error (bottom) for the test for determination of LOEC. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

4 Discussion

4.1 Case study

The outlined case study demonstrates that the choice of the statistical model and procedure can have substantial impact on ecotoxicological inferences. This is further highlighted by the fact that Brock et al. (2015) reported a LOEC of 0.3 mg/L for this data assuming normality of transformed data, whereas we report a value of 3 mg/L (Figure 1). The reasons are be manifold: (Brock et al., 2015) used a $log(2\ y+1)$ transformation, whereas we used a $log(A\ y+1)$ transformation, where A = 2 / 11 = 0.182 (van den Brink et al., 2000). Furthermore, Brock et al. (2015) used a one-sided Williams test which assumes a monotonic dose-response relationship. In contrast, we used a one-sided Dunnett test, which does not assume monotonicity and allows individual comparisons between treatment groups and the control. Although under monotonicity the Williams test has larger power then Dunnett test (Jaki and Hothorn, 2013), this does not affect the comparisons between models.

Moreover, the case study illustrates the potential effects of overdispersion that is not accounted for: standard error will be underestimated and significance overestimated (Figure 1). However, in factorial designs the mean-variance relationship can be easily checked by plotting mean versus variance of the treatment groups (see supplemental material). In the introduction we pointed out that there is little advice how to choose between the plenty of possible transformations - how do GLMs simplify this problem? The distribution modelled can be chosen by the nature of the data giving a statistically sound model reflecting its properties (e.g. bonds, integer or continuous data etc.). Knowing what type of data is modelled (see Methods section), the model selection process can be completely guided by the data and diagnostic plots. Therefore, choosing an appropriate model is more sound and straightforward than choosing between possible transformations.

4.2 Simulations

Our simulations showed that generally GLMs have greater power than data transformations. However, the simulations also suggest that the power at the population level in common mesocosm experiments is low. For common samples sizes and a reduction in abundance of 50% we found a low power to detect any treatment-related effect (<50% for methods with appropriate Type 1 error, Figure 2). Additionally, O'Hara and Kotze (2010) showed that using a log transformation gave unreliable and biased parameter estimates. Statistical power to detect the correct LOEC was even lower (less than 30%). This suggests that population level NOECs reported from mesocosm experiments should be interpreted with caution and underpins the criticism of

NOEC.

Mesocosm studies allow also inferences on community level. For community analyses *GLM for multivariate data* have been proposed as alternative to Principal Response Curves (PRC) and yielded to similar inferences, but better indication of responsive taxa (Szöcs et al., 2015; Warton et al., 2012). Although our simulations covered only simple experimental designs at the population level, findings may also extend to more complex situations. Nested or repeated designs with non-normal data could be analysed using Generalised Linear Mixed Models (GLMM) and may have advantages with respect to power (Stroup, 2014).

To counteract the problems with low power Brock et al. (2015) proposed to take the Minimum Detectable Difference (MDD), a method to assess statistical power *a posteriori*, for inference into account. However, *a priory* power analyses can be performed easily using simulations, even for complex experimental designs (Johnson et al., 2014), and might help to design, interpret and evaluate ecotoxicological studies. Moreover, Brock et al. (2015) proposed that statistical power of mesocosm experiments can be increased by reducing sampling variability through improved sampling techniques and quantification methods, though they also caution against depleting populations through more exhaustive sampling. As we showed, using appropriate statistical methods (like GLMs) can enhance the power at no extra costs.

Wang and Riffel (2011) advocated that in the typical case of small sample sizes (n <20) and non-normal data, non-parametric tests perform better than parametric tests assuming normality. In contrast, our results showed that the often applied Kruskal test and pairwise Wilcoxon test have equal or less power compared to tests assuming normality after data transformation. Moreover, GLMs always performed better than non-parametric tests. Though more powerful non-parametric tests may be available (Konietschke et al., 2012), these are focused on hypothesis testing and do not provide estimation of effect sizes. Additionally to testing, GLMs allow the estimation and interpretation of effects that might not be statistically significant, but ecologically relevant. Therefore, we advise using GLMs instead of non-parametric tests for non-normal data.

At small sample sizes and low abundances a significant amount of negative binomial models did not converge. We used an iterative algorithm to fit these models (Venables and Ripley, 2002) and other methods assessing the likelihood directly may perform better. Moreover, the Likelihood-Ratio test gave an increased Type-I error for these models, where the non reliability of the LR statistic for small sample sizes has long been reported (Bolker et al., 2009; Wilks, 1938). We found that parametric bootstrap (GLM_{pb}) provides a valuable alternative in such situations (Figure 2). At small samples sizes, low abundances or few treatment groups it is

difficult to determine the mean-variance relationship. GLM_{qp} assumes a simpler, linear mean-variance relationship, which might explain the higher power compared to GLM_{pb} at small sample sizes (Figure 2, top).

Binomial data is often collected in lab trials, where increasing the sample size is easy to accomplish. We found notable differences in power to detect a treatment effect up to a sample size of 9. Similarly, Warton and Hui (2011) also found that GLM have higher power than arcsine transformed linear models. Nevertheless, for deriving LOECs the transformation performed better at low sample sizes (n = 3) (Figure 5).

We recommend that non-normal data should be analysed by GLMs and not by transformations or non-parametric methods. To improve the power to detect effects, GLMs should become a standard method in ecotoxicology and incorporated into respective guidelines.

References

Bolker, B., Brooks, M., Clark, C., Geange, S., Poulsen, J., Stevens, M., and White, J. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3):127–135.

Brock, T. C. M., Hammers-Wirtz, M., Hommen, U., Preuss, T. G., Ratte, H.-T., Roessink, I., Strauss, T., and Van den Brink, P. J. (2015). The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environmental Science and Pollution Research*, 22(2):1160–1174.

EFSA PPR (2013). Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters. *EFSA Journal*, 11(7):3290.

EPA (2002). *Methods for Measuring the Acute Toxicity of Effluents and Receiving Waters to Freshwater and Marine Organisms*. U.S. Environmental Protection Agency.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Chapman /& Hall/CRC texts in statistical science series. Chapman /& Hall/CRC, Boca Raton.

Hilbe, J. M. (2014). *Modeling Count Data*. Cambridge University Press, New York, NY.

Jager, T. (2012). Bad habits die hard: The NOEC's persistence reflects poorly on ecotoxicology. *Environmental Toxicology and Chemistry*, 31(2):228–229.

Jaki, T. and Hothorn, L. A. (2013). Statistical evaluation of toxicological assays: Dunnett or Williams test—take both. *Archives of Toxicology*, 87(11):1901–1910.

Johnson, P. C. D., Barry, S. J. E., Ferguson, H. M., and Müller, P. (2014). Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*.

Konietschke, F., Hothorn, L. A., and Brunner, E. (2012). Rank-based multiple test procedures and simultaneous confidence intervals. *Electronic Journal of Statistics*, 6:738–759.

Laskowski, R. (1995). Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology. *Oikos*, 73(1):140–144. Times Cited: 35.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384.

Newman, M. C. (2012). *Quantitative ecotoxicology*. Taylor & Francis, Boca Raton, FL.

OECD (2006). *Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application*. Number 54 in Series on Testing and Assessment. OECD, Paris.

O'Hara, R. B. and Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, 1(2):118–122.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sanderson, H. (2002). Pesticide studies. *Environmental Science and Pollution Research*, 9(6):429–435.

Stroup, W. W. (2014). Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. *Agronomy Journal*.

Szöcs, E., Van Den Brink, P. J., Lagadic, L., Caquet, T., Roucaute, M., Auber, A., Bayona, Y., Liess, M., Ebke, P., Ippolito, A., Ter Braak, C. J., Brock, C. M., and Schäfer, R. B. (2015). Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: A comparison of methods. *Ecotoxicology*.

van den Brink, P. J., Hattink, J., Brock, T. C. M., Bransen, F., and van Donk, E. (2000). Impact of the fungicide carbendazim in freshwater microcosms. II. Zooplankton, primary producers and final conclusions. *Aquatic Toxicology*, 48(2-3):251–264.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Springer, New York, fourth edition.

Ver Hoef, J. M. and Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? *Ecology*, 88(11):2766–2772.

Wang, M. and Riffel, M. (2011). Making the right conclusions based on wrong results and small sample sizes: interpretation of statistical tests in ecotoxicology. *Ecotoxicology and Environmental Safety*, 74(4):684–92.

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

Warton, D. I., Wright, S. T., and Wang, Y. (2012). Distance-based multivariate analyses confound location and dispersion effects. *Methods in Ecology and Evolution*, 3(1):89–101.

Weber, C. I., Peltier, W. H., Norbert-King, T. J., Horning, W. B., Kessler, F., Menkedick, J. R., Neiheisel, T. W., Lewis, P. A., Klemm, D. J., Pickering, Q., Robinson, E. L., Lazorchak, J. M., Wymer, L., and Freyberg, R. W. (1989). Short-term methods for estimating the chronic toxicity of effluents and receiving waters to fresh- water organisms. Technical Report EPA/600/4–89/001, Environmental Protection Agency, Cincinnati, OH: Environmental Monitoring Systems Laboratory.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.