Environmental Science and Pollution Research

Ecotoxicology is not normal - A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology

--Manuscript Draft--

Manuscript Number:	ESPR-D-15-00741R1
Full Title:	Ecotoxicology is not normal - A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology
Article Type:	Research Article
Corresponding Author:	Eduard Szöcs University Koblenz-Landau Landau, GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	University Koblenz-Landau
Corresponding Author's Secondary Institution:	
First Author:	Eduard Szöcs
First Author Secondary Information:	
Order of Authors:	Eduard Szöcs
	Ralf B. Schäfer
Order of Authors Secondary Information:	
Abstract:	Counts and proportions are data types often encountered by ecotoxicologists, which are rarely normally distributed. 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. Generalised Linear Models (GLM) allow directly model distributions fitting such data. Here, we compare the performance of parametric methods assuming (1) normality of transformed data, (2) appropriate distributions (Poisson, negativ binomial, binomial) and (3) non-parametric methods. We simulated typical data mimicking low replicated ecotoxicological experiments of two common data types (counts and proportions from counts). We compared the performance of different methods in terms of statistical power and type 1 error for detecting a general treatment effect and determining the lowest observed effect concentration (LOEC). In addition, we outlined differences and advantages of GLMs on a real world mesocosm data set. For counts, we found that the quasi-Poisson model and the negative binomial model in combination with the parametric boostrap had higher statistical power than data transformation. For proportions GLMs performed better, except to determine LOEC at extremly low sample sizes. The compared non-parametric methods had generally lower power. We recommend that counts and proportions from counts should be analysed by making appropriate distributional assumptions and GLMs should become a standard method in ecotoxicology.
Response to Reviewers:	See attached response letter.

Environmental Science and Pollution Research manuscript No.
(will be inserted by the editor)

Ecotoxicology is not normal.

A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology.

Eduard Szöcs · **Ralf B. Schäfer**

Received: date / Accepted: date

Abstract Counts and proportions are data types often encountered by ecotoxicologists, which are rarely normally distributed. 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. Generalised Linear Models (GLM) allow directly model distributions fitting such data. Here, we compare the performance of parametric methods assuming (1) normality of transformed data, (2) appropriate distributions (Poisson, negativ binomial, binomial) and (3) non-parametric methods.

We simulated typical data mimicking low replicated ecotoxicological experiments of two common data types (counts and proportions from counts). We compared the performance of different methods in terms of statistical power and type 1 error for detecting a general treatment effect and determining the lowest observed effect concentration (LOEC). In addition, we outlined differences and advantages of GLMs on a real world mesocosm data set.

For counts, we found that the quasi-Poisson model and the negative binomial model in combination with the parametric boostrap had higher statistical power than data transformation. For proportions GLMs performed better, except to determine LOEC at extremly low sample sizes. The compared non-parametric methods had generally lower power.

We recommend that counts and proportions from counts should be analysed by making appropriate distributional assumptions and GLMs should become a standard method in ecotoxicology.

Eduard Szöcs (✉) and Ralf B. Schäfer
Institute for Environmental Sciences
University Koblenz-Landau
Fortstraße 7,
76829 Landau, Germany
Tel.: +49 06341 280 31552
E-mail: szoecs@uni-landau.de

Keywords Generalized Linear Models · Transformations · Simulation · Power · Type 1 error

1 Introduction

Ecotoxicologists perform various kinds of experiments yielding different types of data. Examples are animal counts in mesocosm experiments (non-negative, integer-valued data) or proportions of surviving animals (data bounded between 0 and 1, discrete). These data are typically not normally distributed. Nevertheless, they are often 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 counts or proportions from counts in a statistically sound way (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 (counts), Warton and Hui 2011; Warton 2005 (proportions from counts)).

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 the predominant endpoint in 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 from counts.

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):

$$y_i^T = log(Ay_i + 1) \tag{1}$$

, where y_i is the measured and y_i^T the transformed abundance of the ith observation. The factor A was chosen in such way that Ay equals 2 for the lowest non-zero abundance value (y).

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

$$\begin{aligned} y_i^T &\sim N(\mu_i, \sigma^2) \\ E(y_i^T) = \mu_i &\text{ and } var(y_i^T) = \sigma^2 \\ \mu_i &= \beta Treatment_i \end{aligned} \tag{2}$$

This model assumes a normal distribution of the transformed abundances. The expected value for each observation i is given by its mean (μ_i) and the variance (σ^2) is constant between treatments. We allow this mean to vary between treatments and β are the coefficients related to these changes in transformed abundances between treatments (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 GLM (GLM_p):

$$\begin{aligned} y_i &\sim P(\mu_i) \\ E(y_i) &= var(y_i) = \mu_i \\ log(\mu_i) &= \beta Treatment_i \end{aligned} \tag{3}$$

This model assumes poisson distributed abundances with mean $\mu_i \geq 0$. The expected value for each observation i is given by its mean. Moreover, this model assumes that mean and variance are equal. We are modelling the mean as a function of treatment membership. However, to avoid negative values of the mean this is done on a log scale. Therefore, β also describes the differences between treatments on a log scale (eqn. 3).

The assumption of equal mean and variance is rarely met with ecological data, which is typically characterized by greater variance than the mean (overdispersion). To overcome this problem a quasi-Poisson model (GLM_{qp}) could be used, which assumes that variance is a linear function of the mean (eqn. 4):

$$var(y_i) = \Theta \mu_i \tag{4}$$

Here, Θ is used to account for additional variation and is known as overdispersion parameter. The quasi-Poisson model is a post hoc method, meaning that first a Poisson model is estimated (eqn. 3) and than the standard errors are scaled by the degree of overdispersion.

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

$$\begin{aligned} y_i &\sim NB(\mu_i, \kappa) \\ E(y_i) = \mu_i &\text{ and } var(y_i) = \mu_i + \mu_i^2/\kappa \\ log(\mu_i) &= \beta Treatment_i \end{aligned} \tag{5}$$

This models assumes that abundances are negative binomially distributed, with a mean of $\mu_i \geq 0$ and a variance $\mu_i + \mu_i^2/\kappa$. Similar to the Poisson model we use a log link between mean and treatments. 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

A binomial variable counts how often an event x occurs in a fixed number of independent trials N (e.g. *"5 out of 10 fish survived"*), with an equal probability of occurrence π between trials. The number of times an event occurs can also be calculated as proportion x/N.

2.2.1 Linear model for transformed data

To accommodate the assumptions for the standard linear model with such proportions, a special arcsine square root transformation (eqn. 6) is suggested (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 total number of exposed animals per treatment. 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

A more natural way to model such data is the binomial distribution with parameters N and π (GLM_{bin}):

$$\begin{aligned} y_i &\sim Bin(N, \pi_i) \\ E(y_i) = \pi_i \times N &\text{ and } var(y_i) = \pi_i(1-\pi_i)/N \\ logit\ (\pi_i) &= \beta Treatment_i \end{aligned} \tag{7}$$

This model assumes that the number of occurences are binomially distributed, where N = number of trials (e.g. exposed animals) and π_i is the probability of occurrences (fish survived), which together give the expected number of occurrences. The variance of the binomial distribution is a quadratic function of the mean. We are modelling the probability of occurrence as function of treatment membership and to ensure that $0 < \pi_i < 1$ we do this on a logit scale (eqn. 7). However, the parameters β of this model are directly interpretable as changes in log odds between treatments.

Similarly to counts, binomial data may also show overdispersion. Methods to deal with overdispersed binomial data are either quasi methods (see above) or Generalized Linear Mixed models (GLMM). However, these are not further investigated in this paper (see Warton and Hui (2011) for a comparison).

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 the parametric bootstrap (Faraway 2006) to assess the significance of the LR. Bootstrapping is computationally very intensive and for this reason we applied it only for the negative binomial models (using 500 bootstrap samples, denoted as GLM_{npb}).

To assess the LOEC we used Dunnett contrasts (Dunnett 1955) 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. We adjusted for multiple testing using the method of Holm (1979).

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

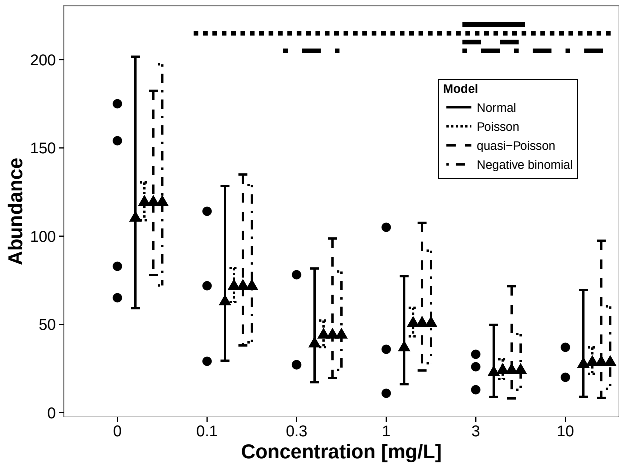

Fig. 1 Data from Brock et al (2015) (dots). Predicted values (triangles) and 95% Wald Z or t confidence intervals from the fitted models (vertical lines) are given beside. Horizontal bars above indicate treatments statistically significant different from the control group (Dunnett contrasts). The data showed overdispersion ($\kappa = 4$) and therefore, the Poisson model underestimates the width of confidence intervals.

treatments (T1 - T5) and one control group (C). Counts were drawn from a negative binomial distribution with overdispersion at all treatments ($\kappa = 4$, eqn. 5). We simulated data sets with different number of replicates (N = {3, 6, 9}) and different abundances in control treatments (μ_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 1000 data sets for each combination of N and μ_C and analysed these using the models outlined in section 2.1.

2.5.2 Binomial data

We simulated data from a commonly used design as described 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 (π_C = {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_C was held constant between groups. For power estimation π_C 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 1000 data sets and analysed these using the models outlined in section 2.2.

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 (detection of an effect when there is none) and power (ability to detect an effect when it is present).

All simulations were done in R (Version 3.1.2) (R Core Team 2014) on an Amazon EC2 virtual Linux server (64bit, 15GB RAM, 8 cores, 2.8 GHz). Source code to reproduce the simulations and paper is available online at `https://github.com/EDiLD/usetheglm`. Moreover, Supplement 2 provides worked examples of the data of Brock et al (2015) and Weber et al (1989).

3 Results

3.1 Case study

The data set showed considerable higher variance then expected by the Poisson model ($\Theta = 22.41$, 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 detecting a general treatment effect, GLM_{nb} and GLM_p showed inflated type 1 error rates, whereas *KW* was conservative at low sample sizes. However, using parametric bootstrap for the negative binomial model (GLM_{npb}) resulted in appropriate type 1 error rates. For detecting a treatment effect, GLM_{npb} and GLM_{qp} exhibited higher power than *LM* and *KW*, the latter having least power (Figure 2). 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%. At small sample sizes (n = 3, 6) and low abundances (μ_C = 2, 4) many of the negative binomial

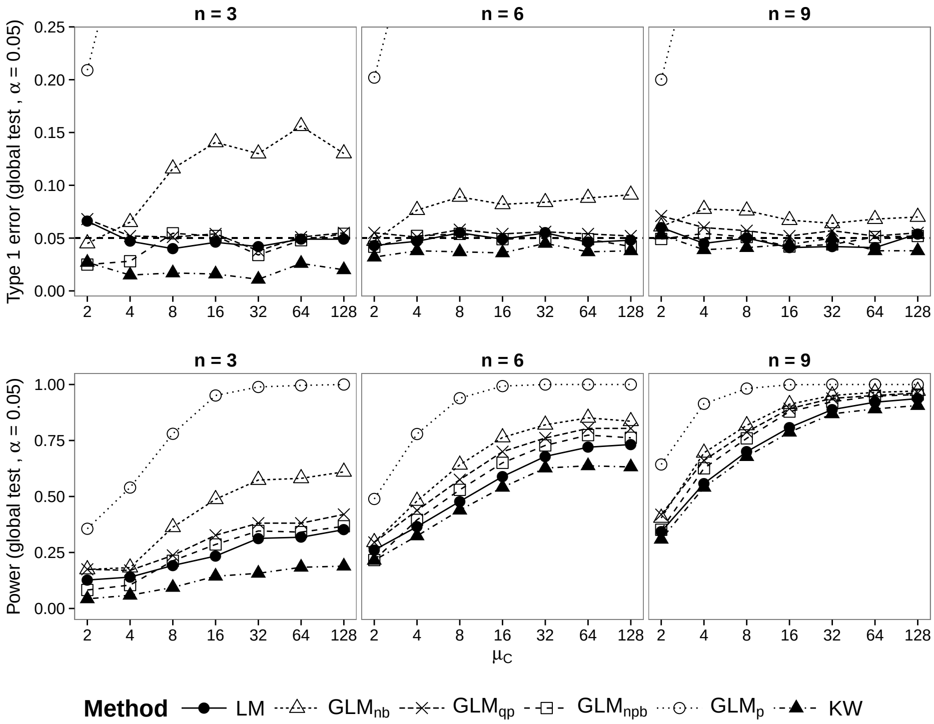

Fig. 2 Count data simulations: Type 1 error (top) and Power (bottom) for the test of a treatment effect. Only type 1 errors <25% are displayed. GLM_p showed type 1 errors >20% in all simulation scenarios. Power levels for models with inflated type I error are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than 85% of GLM_{nb} and GLM_{npb} models did converge. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

models (GLM_{nb} and GLM_{npb}) did not converge to a solution (convergence rate <85% of the simulations, Supplement 1).

For LOEC determination GLM_{nb} and GLM_p showed an increased Type 1 error and all other methods were slightly conservative. The inferences on LOEC generally showed less power. *LM* showed a mean reduction of 20.7% and GLM_{qp} of 24.3 %. Power to detect the LOEC was highest for GLM_{qp}. *LM* and WT showed less power, with WT having no power to detect the LOEC at low sample sizes (Figure 3).

3.2.2 Binomial data

GLM_{bin} showed slightly increased type 1 error rates at low sample sizes and small effect sizes. KW was more conservative than *LM* and GLM_{bin}. In addition, GLM_{bin} exhibited the greatest power for testing the treatment effect. This was especially apparent at low sample sizes (n = 3), with up to 27% higher power compared to LM. However, the differences between methods quickly vanished with increasing samples sizes (Figure 4).

For inference on LOEC we found that all methods were slightly conservative. WT was generally more conservative and GLM_{bin} especially at low effect sizes ($p_E > 0.7$). 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 (n = 3, 6). WT had no power for n = 3 and showed less power in the other simulation runs (Figure 5).

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 and endpoints like the LOEC. Therefore, ecotoxicologists should not base their inferences solely on statistical significance tests, but also on parameter estimates, their uncertainty and importance (Gelman and Stern 2006). Nevertheless, O'Hara and Kotze (2010) showed that *LM* using a log transformation gave unreliable and biased parameter estimates, whereas GLMs performed well with little bias. Bias occurs also when back-transforming means to the original scale, which explains the lower predicted means by *LM* in Figure 1 (Rothery 1988) and should be corrected for (Newman 1993).

This is further highlighted by the fact that for the same model (linear model of transformed data), Brock et al (2015) reported a 10-fold lower LOEC (0.3 mg/L) then found in our

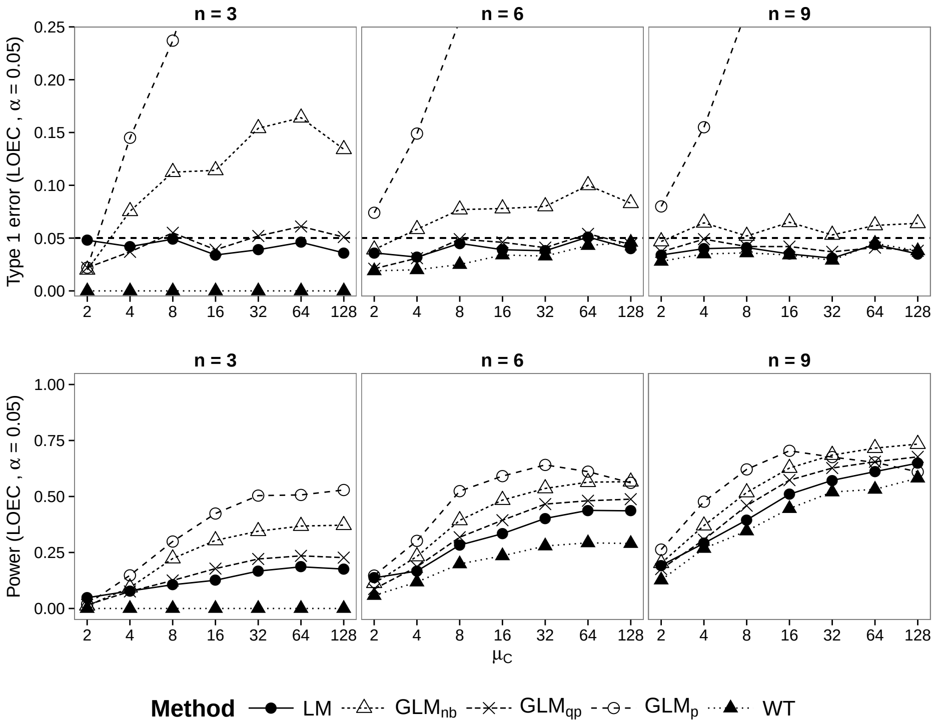

Fig. 3 Count data simulations: Type 1 error (top) and Power (bottom) for determination of LOEC. For clarity only type 1 errors <25% are displayed. Power levels for models with inflated type I error are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than 85% of GLM_{nb} models did converge. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

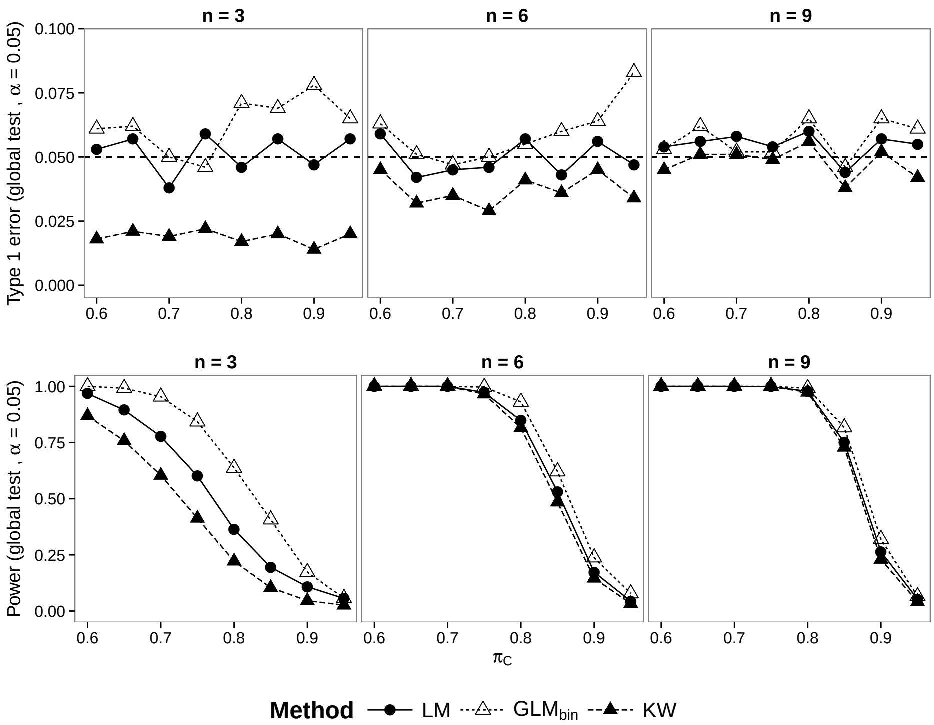

Fig. 4 Binomial data simulations: Type 1 error (top) and power (bottom) for the test of a treatment effect. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

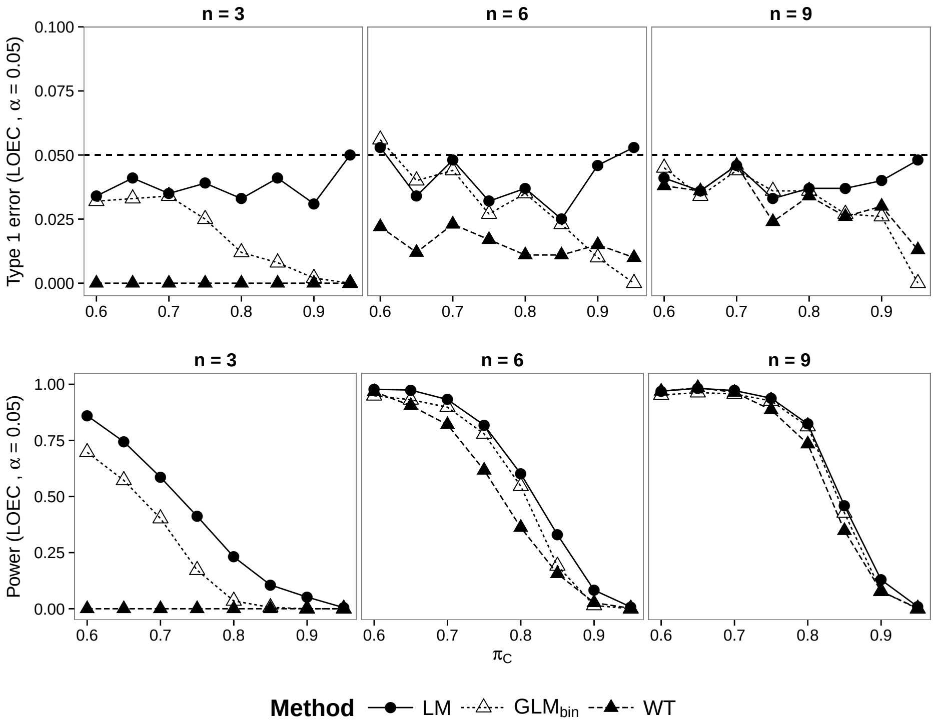

Fig. 5 Binomial data simulations: Type 1 error (top) and power (bottom) for the test for determination of LOEC. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

study (3 mg/L, Figure 1). The reasons are 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). However, this contributed only little to the differences. A much bigger impact had the type of multiple comparison: Brock et al (2015) used a one-sided Williams test (Williams 1972), whereas we used one-sided comparisons to the control (Dunnett contrasts). In contrast to the Williams test, Dunnett contrasts do not assume a monotonic dose-response relationship and allow individual comparisons between treatment groups and the control. However, under monotonicity they have less power, which explains the differences (Jaki and Hothorn 2013). Both types of multiple comparisons are available as multiple contrast tests in a GLM framework. Therefore, our comparison of methods should be independent from the choice of contrast, which is determined by assumptions and research questions.

Overdispersion is common for ecological datasets (Warton 2005) and the case study illustrates the potential effects of overdispersion that is not accounted for: standard errors will be underestimated and significance overestimated (Figures 1). This is also shown by our simulations (Figures 2, 3) where GLM_p showed increased type 1 error rates because of overdispersed simulated data. However, in factorial designs the mean-variance relationship can be easily checked by plotting mean versus variance of the treatment groups (see Supplement 2). 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 using knowledge about the data (e.g. bounds, 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 tools. Therefore, choosing an appropriate model is easier 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 ($n \leq 4$) 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). Statistical power to detect the correct LOEC was even lower (less than 25%), which can be attributed to multiple testing. The low power of all methods to detect significant treatment levels such as the LOEC or NOEC suggests that these endpoints from ecotoxicological studies should be interpreted with caution and underpins their criticism (Laskowski 1995; Landis and Chapman 2011).

Mesocosm studies allow also inferences on community level. For community analyses *GLM for multivariate data* (Warton et al 2012) 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). However, ter Braak and Šmilauer (2014) argue to use data transformations with community data because of their simplicity and robustness. 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 at the population level 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 priori* 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 KW and WT have less power compared to LM. 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.

We found an increased Type-I error for GLM_{nb} at low sample sizes. However, it is well known that the LR statistic is not reliable at small sample sizes (Bolker et al 2009; Wilks 1938). Parametric bootstrap (GLM_{npb}) is a valuable alternative in such situations and maintains appropriate levels (Figure 2). Moreover, 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.

GLM_{qp} showed higher statistical power than GLM_{npb} (Figure 2, bottom). This could be explained by the simpler mean-variance relationship of GLM_{qp} (eqn. 4 and 5), because at small samples sizes, low abundances or few treatment groups it is difficult to determine the mean-variance relationship.

Binomial data are often collected in lab trials, where increasing the sample size may be relatively easy to accomplish. We found notable differences in power to detect a treatment effect for all simulated sample sizes. Similarly, Warton and Hui (2011) also found that GLMs have higher power than arcsine transformed linear models. Though we did not simulate overdispersed binomial data, this should be checked and accounted for. In such situations a GLMM may offer an appealing alternative (Warton and Hui 2011). At low effect sizes GLM_{bin} became conservative with increasing π_C, although this effect lessened as sample size increased (Figure 5). This is because π approaches its boundary and is also known as the *Hauck-Donner effect* (Hauck and Donner 1977). A LR-Test or parametric bootstrap may provide an alternative in such situations (Bolker et al 2009). This can also explain why LM performed better for deriving LOECs at low sample sizes.

GLMs can be fitted with several statistical software packages and many textbooks are available to introduce ecotoxicologists to these models (e.g. Zuur 2013 or Quinn and Keough 2009). We recommend that ecotoxicologists should change their models instead of their data. GLMs should become a standard method in ecotoxicology and incorporated into respective guidelines.

5 Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

References

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

ter Braak CJF, Šmilauer P (2014) Topics in constrained and unconstrained ordination. Plant Ecology DOI 10.1007/s11258-014-0356-5, URL http://link.springer.com/10.1007/s11258-014-0356-5

van den Brink PJ, Hattink J, Brock TCM, Bransen F, 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

Brock TCM, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte HT, Roessink I, Strauss T, Van den Brink PJ (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

Dunnett CW (1955) A Multiple Comparison Procedure for Comparing Several Treatments with a Control. Journal of the American Statistical Association 50(272):1096–1121

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 JJ (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

Gelman A, Stern H (2006) The difference between "significant" and "not significant" is not itself statistically significant. The American Statistician 60(4):328–331, URL http://pubs.amstat.org/doi/abs/10.1198/000313006X152649

Hauck WW, Donner A (1977) Wald's Test as Applied to Hypotheses in Logit Analysis. Journal of the American Statistical Association 72(360):851, DOI 10.2307/2286473, URL http://www.jstor.org/stable/2286473?origin=crossref

Hilbe JM (2014) Modeling Count Data. Cambridge University Press, New York, NY

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6(2):65–70

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

Johnson PCD, Barry SJE, Ferguson HM, Müller P (2014) Power analysis for generalized linear mixed models in ecology and evolution. Methods in Ecology and Evolution DOI 10.1111/2041-210X.12306

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

Landis WG, Chapman PM (2011) Well past time to stop using NOELs and LOELs. Integrated Environmental Assessment and Management 7(4):vi–viii

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 JA, Wedderburn RWM (1972) Generalized Linear Models. Journal of the Royal Statistical Society Series A (General) 135(3):370–384

Newman MC (1993) Regression analysis of log-transformed data: Statistical bias and its correction. Environmental Toxicology and Chemistry 12(6):1129–1133, URL http://onlinelibrary.wiley.com/doi/10.1002/etc.5620120618/abstract

Newman MC (2012) Quantitative ecotoxicology. Taylor & Francis, Boca Raton, FL

OECD (2006) Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application. No. 54 in Series on Testing and Assessment, OECD, Paris

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. Methods in Ecology and Evolution 1(2):118–122

Quinn GP, Keough MJ (2009) Experimental design and data analysis for biologists. Cambridge Univ. Press, Cambridge

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Rothery P (1988) A cautionary note on data transformation: bias in back-transformed means. Bird Study 35(3):219–221, DOI 10.1080/00063658809476992, URL http://www.tandfonline.com/doi/abs/10.1080/00063658809476992

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

Stroup WW (2014) Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. Agronomy Journal DOI 10.2134/agronj2013.0342

Szöcs E, Van Den Brink PJ, Lagadic L, Caquet T, Roucaute M, Auber A, Bayona Y, Liess M, Ebke P, Ippolito A, Ter Braak CJ, Brock CM, Schäfer RB (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: A comparison of methods. Ecotoxicology DOI 10.1007/s10646-015-1421-0

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, 4th edn. Springer, New York

Ver Hoef JM, Boveng PL (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88(11):2766–2772

Wang M, 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 DI (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16(3):275–289

Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92(1):3–10

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

Weber CI, Peltier WH, Norbert-King TJ, Horning WB, Kessler F, Menkedick JR, Neiheisel TW, Lewis PA, Klemm DJ, Pickering Q, Robinson EL, Lazorchak JM, Wymer L, Freyberg RW (1989) Short-term methods for estimating the chronic toxicity of effluents and receiving waters to fresh- water organisms. Tech. Rep. EPA/600/4–89/001, Environmental Protection Agency, Cincinnati, OH: Environmental Monitoring Systems Laboratory

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

Williams DA (1972) The comparison of several dose levels with a zero dose control. Biometrics pp 519–531, URL http://www.jstor.org/stable/10.2307/2556164

Zuur AF (2013) A beginner's guide to GLM and GLMM with R: a frequentist and Bayesian perspctive for ecologists. Highland Statistics, Newburgh

Responses to reviewers

Ms. No. ESPR-D-15-00741
submitted to
Environmental Science and Pollution Research

Eduard Szöcs and Ralf B. Schäfer

March 23, 2015

Dear editor Dr. Schulz and reviewers,

We are thankful for reviewing our manuscript and the comments that helped to improve the paper. We revised the manuscript according to comments of the three reviewers and are re-submitting the manuscript for consideration for publication in Environmental Science and Pollution Research.

In the remainder of this document, we describe the changes that we have made to the paper for resubmission. To assist the assessment of our changes we have submitted two versions of the revised manuscript: one with highlighted changes and another without any highlighting. Note, that we did not highlight changes in citations and figures.

Kind regards,
Eduard Szöcs and Ralf B. Schäfer

Response to Reviewer 1

Comment 1: *"The purpose of the manuscript is highly commendable. Unfortunately, the present manuscript has some major shortcomings: The simulation results, which constitute the key contribution, are far from being as unequivocal as the title may lead one to think. Also, it is disturbing to see a fairly technical (and for ESPR non-standard) formulation of statistical models and yet the authors seem to lack some understanding of the models in some cases."*

Response: We changed the subtitle (see also comment 2) and updated the notation of the models used (see also comments 6, 7, 8, 10, 25, and 43). The simulation scenarios are tailored to experiments and data often encountered in ecotoxicology, which are characterised by small sample sizes and a factorial design (see also comment 13). Therefore, we submitted to ESPR, though, it might be interesting also to a greater audience of non-ecotoxicologists.

Comment 2: *"About the title: Please consider a more informative title such as "Comparison of statistical approaches for analysis of non-normally distributed response in ecotoxicology". Also in view of the fact that the results are not very clearly favouring the generalized linear models (GLMs)."*

Response: We agree and changed the subtitle to
"A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology".

Comment 3: *"p. 2, left column, lines 2–12: The issue of LOEC/NOEC versus regression is interesting but not really relevant to the aim of the manuscript. Consider skipping this part. However, you're right that for small sam- ple sizes it is extremely important to choose the most efficient statistical analysis and the better the model describes the data the more efficient."*

Response: We shortened this section. However, we kept a short introduction to LOEC/NOEC as inference on parameters is also part of our results.

Comment 4: *"p. 2, left column, lines 56–59: You aren't mentioning the issue with back-transformation So, in particular, βx is not additive changes from control on the original scale. To many ecotoxicologists such estimates on a transformed scale aren't directly interpretable as claimed by the authors. A more detailed explanation or at least discussion on how to back-transform is warranted (this is an issue regardless of* *whether GLMs or models for transformed responses are used!).”*

Response: We agree and include transformation bias in our discussion.

Comment 5: *”p. 2, right column, line 1: The real advantage of using a generalized linear model for count data shows up in case of many small counts and in the presence of ties. For large counts (e.g., corresponding to a large μ_C) there is practically no difference. However, this difference may not show up in simulations when the sample size is kept between 3 and 9.”*

Response: We agree that for large counts there is no practical difference between methods. However, our simulation design aimed to mimic data sets frequently encountered in ecotoxicology. See also response to comment 13.

Comment 6: *”p. 2, right column, lines 19–21: It is not the response variable that is linked to the linear predictors, it is the mean of the response variable (McCullagh & Nelder, 1989). This is a basic fact of generalized linear models as you also show in Eqn. (3). Please explain Eqn. (3) in more detail (explain the parameters).”*

Response: We agree and rephrased this complete section.

Comment 7: *”p. 2, right column, lines 25–27: There exists no quasi-Poisson distribution!!! You use R terminology, but this is not in this case statistical terminology. Please consult McCullagh & Nelder (1989) to understand that over-dispersion is dealt with by means of an ad hoc scaling of the standard errors that takes place after having fitted the ordinary Poisson GLM. So Eqn. (4) doesn't make sense.”*

Response: We agree and rephrased this complete section.

Comment 8: *”p. 2, right column, Eqn. (5): Have you explained the parameters in the text?”*

Response: We added the missing explanation for eqn. 5. Moreover, we unified the explanations of the models.

Comment 9: *”p. 3, left column, lines 26–42: What about over-dispersion for binomial data? Shouldn't that also be mentioned now that you consider models for over-dispersion for count data.”*

Response: We added that overdispersion might also occur in binomial data and pointed the reader to Warton and Hui (2011).

Comment 10: *"p. 3, left column, line 33: Why not specify the mean next to the variance in all equations? Or skip the variance where not needed (as in this case)."*

Response: We edited sections 2.1 and 2.2 and specified the mean next to the variance.

Comment 11: *"p. 3, right column, Fig. 1: Please be precise in the figure text: By using the Poisson model the width of confidence intervals is underestimated in the presence of overdispersion."*

Response: We agree and stated more precisely that the width of confidence intervals is underestimated.

Comment 12: *"p. 3, right column, line 28: How is multiple testing taken care of? This may also affect the simulation results."*

Response: We adjusted for multiple testing using the method of Holm (1979). We added this information to the manuscript. For the first submission we did not use Holm's method for the binomial data simulations. However, we unified this with the rerun of simulations (see comment 14).

Comment 13: *"p. 4, left column, line 2: The simulations will not be very informative for such small sample sizes (3–9) as all methods will have low power. Suggestion: Do also consider larger sample sizes: 12, 25, 50, 100. Not completely unrealistic any more as experiments tend to get larger. And then instead keep the μ_C small: 2, 4, 8, 16 to get more ties."*

Response: Our simulation design aimed to mimic data sets frequently encountered in ecotoxicology. Sample sizes between 3 and 9 are very common in regulatory ecotoxicology. For example, mesocosm experiments rarely exceed four replicates per treatment due to logistic constraints (Szöcs et al, 2015). Similarly the *Daphnia magna* standard test (OECD, 2004) is usually conducted with four to five replicates. We believe that the current simulation design is relevant for ecotoxicologists and did not simulate larger sample sizes. Moreover, we provide a fully reproducible source code at `https://github.com/EDiLD/usetheglm`, enabling other researchers to easily adapt our simulation design to specific needs.

Comment 14: *"p. 4, left column, lines 10 and 28: In simulation studies it is quite common to use 1000 simulated datasets for each scenario (it would also reduce the sampling variability in the results)."*

Response: We agree and rerun our simulations generating 1000 datasets for all scenarios. Because of the increased computational burden we run the computation in parallel on Amazon EC2 instance (with 15 GB RAM, 2.8 GHz and 8 cores) - taking more then 12 hours for the count data simulations.

Comment 15: *"p. 4, left column, line 36: Please do define the type I error properly."*

Response: We agree and rephrased to:
"[...] Type 1 error (detection of an effect when there is none) [...]"

Comment 16: *"p. 4, right column, lines 26–29: The lack of convergence is not surprising as the simulated datasets are simply too small to fit complex models. Please re-consider the entire concept of the simulation (in particular which scenarios to consider)."*

Response: Our simulations covered scenarios often encountered by ecotoxicologists and therefore, it is like that practicing ecotoxicologists will also encounter such convergence problems. See also response to comment 13.

Comment 17: *"p. 4, right column, lines 31–33: One explanation for the reduced power is certainly the multiple testing issue. Perhaps this point could be addressed?"*

Response: We added this point to the discussion.

Comment 18: *"p. 4, right column, lines 57–58: Did LM really maintain the nominal significance level in all cases? This seems to be a subjective assessment (as the level is above in some cases for $N = 3$)."*

Response: We rerun our analyses using 1000 simulations (see also comment 14) and updated this section.

Comment 19: *"p. 6, right column, lines 43–46: Why didn't you also use the transformation suggested by Brock et al. (2015)? Would be useful to see the comparison."*

Response: As pointed out in the introduction, there is little information and advice which transformation to use. The $log(A\ y + 1)$ is the most commonly used transformation in ecotoxicological mesocosm experiments. We used only the $log(A\ y+1)$ transformation because our paper did not aim to compare different transformations and there are some inconsistencies in Brock et al (2015). They state:

"For the examples presented in this paper, we followed Van den Brink et al. (2000) using the transformation y(x)=ln(ax+1), where x is the measured abundance and the factor 'a' is selected in such a way that the lowest non-zero abundance of the data set is transformed to 1."

Maybe this is just a typographical error, but in the publication of van den Brink et al (2000) it is stated:

"We decided that the factor Ax in the ln(Ax +1) transformation should make 2 by taking the lowest abundance value higher than zero for x."

Moreover, in the supporting material of Brock et al (2015), they used a $log(2\ y+1)$ transformation to which we refer to in our manuscript.

Comment 20: *"p. 6, right column, line 49: Dunnett test is R terminology (yes, it may be found in other recent publications, but it doesn't make it more correct). Please use a term like "comparisons to the control"."*

Response: We agree and changed accordingly.

Comment 21: *"p. 7, right column, lines 7–10: A nice result that is not found in many publication: non-parametric approaches lack power (due to less assumptions being made). However, your results for the parametric models are all quite similar and as it stands the paper does not present a strong case for the use of GLMs; and a more balanced abstract would be appropriate."*

Response: We agree and revised the abstract.

Comment 22: *"p. 7, Simulations: A suggestion: remove all diverging discussion of LOEC/NOEC, GLMs for multivariate data, GLMMs, sample size cal- culation and concentrate on discussing the simulation results. Or, alter- natively, discuss in much more detail that GLMMs may actually offer a difference approach for handling over-dispersion in binomial and count data (instead of considering completely different designs)."*

Response: We agree and pointed to GLMM as an alternative to deal with overdispersion in the methods section and discussion.

Response to Reviewer 2

Comment 23: *"One point to consider in a resubmission is that GLMs aren't for use for all non-normal data - some qualifications are needed of when to use GLMs. Distributionally, GLMs replace the assumption of normality with the assumption that data come from the exponential family, which does not cover all situations. More critically, GLMs replace the assumption of equal variance with an assumption that the mean-variance relationship follows a pre-specified pattern, and violation of this assumption has similar impacts to violations of the equal-variance assumption in a linear model would have. Particular situations where GLMs are useful are binary data (known to be a quadratic mean-variance), count data (especially count data with lots of low counts, which typically cannot be transformed to normality effectively), and proportions constructed from counts. These are the examples highlighted in the text, and are of clear relevance in ecotoxicology, but the argument does not extend much more broadly than these examples - it is not a case of "your data aren't normal so you should use GLMs". It is more accurate to say "if you have binary, counts or proportions from counts you should use GLMs". I think this requires a modest adjustment to the phrasing of the abstract and introduction to better qualify the situations where GLMs are appropriate. e.g. where O'Hara and Kotze and Warton & Hui are mentioned, this should be qualified to refer to counts and proportions (of counts) respectively. I don't think the title needs changing, sufficient qualification elsewhere of the type of non-normality if interest should do the trick."*

Response: We agree and revised our manuscript accordingly. We changed the subtitle (see also comment 2) and the abstract (see also comment 21). The main text is now more specific to counts and proportions from counts.

Comment 24: *"In the first paragraph, "positive" is used where "non-negative" is needed (i.e. zero is a possible value), also change bonded to bounded"*

Response: We fully agree and changed accordingly.

Comment 25: *"Page 2, second column - the parameterisation isn't equal to the Poisson model - the mean model is the same, I think that is what was meant."*

Response: We edited this section (see also comment 6) and corrected this.

Comment 26: *"Page 2, second column - also, strictly speaking, not all the methods mentioned here are GLMs. Negative binomial regression ("NB2") is a GLM if the* *overdispersion parameter is fixed, otherwise it is a generalisation of GLM. Similarly, quasi-Poisson isn't really a GLM, but a related method."*

Response: We agree and mention that quasi-Poisson is a related post hoc method. See also comment 7.

Comment 27: *"Section 2.2 - a description of what is meant by binomial data would be helpful at the start of this section. Perhaps for completeness a description of what is meant by counts would help at the start of section 2.1"*

Response: We added a brief description of a binomial random variable to the start of section 2.2.

Comment 28: *"Section 2.2.2 - it would be useful to also mention GLMMs with a random intercept term. Often with this sort of data there are extra sources of variation from one sample to the next that are not explained by the binomial assumption, and this is the most natural way to account for it."*

Response: We added a reference to GLMM and (Warton and Hui, 2011) to this section.

Comment 29: *"Sections 2.1, 2.2: references to relevant articles or texts would be useful for GLMs, where the reader can find more details at a level appropriate for readership (e.g. a Zuur text? Quinn & Keough?). Also point readers to supp 2 for code and a worked example."*

Response: We are thankful for pointing us to the missing reference to supplement 2. We added this to the methods section and pointed to introductory textbooks at the end of the manuscript.

Comment 30: *"page 3, line 58, the parametric bootstrap"*

Response: We changed accordingly.

Comment 31: *"Page 3, second column, line 30: Wald tests should be used with caution, when a mean estimate is zero (or close to it) they can have strange behaviour (related to the issue of separable data - parameter estimates and ses diverging, test statistics going to zero)."*

Response: We are thankful for this comment and added discussion of the Hauck-

Donner effect.

Comment 32: *"Page 4, second column, lower means: it is noted that the means are lower when data are transformed. This is known as transformation bias, and occurs because the transform linear model is actually estimating something different - it is no longer trying to estimate mean response. It would be helpful to say this and maybe include a useful reference. This also gets at the issue of interpretability - a disadvantage of the transform approach is a loss of interpretability, because we are no longer modelling mean response."*

Response: We added discussion of transformation bias, see also comment 4.

Comment 33: *"page 7 first column, line 7: bounds not bonds"*

Response: Thanks for pointing to this typo.

Comment 34: *"Section 4.1 the case study discussion was a little confusing. So is the point that you can get different results by linear models according to transformation, but that GLMs can resolve the problem of choice of transformation? (by replacing it with a decision about mean-variance relationship?) I think a stronger point is that the GLM is usually a better fit to count data, no transformation can make data Gaussian when it has lots of zeros and small counts, but GLM is designed for such data. If diagnostic tools support this argument for your case study they could be included."*

Response: Yes, we find it much easier to choose appropriate distributions than from the virtually infinite possibilities of transformations. We agree, that GLMs give predictions that better fit to the data. But note, that Warton (2005) found that "*a Gaussian model based on transformed abundances fitted data surprisingly well*" (see also comment 59).

Comment 35: *"page 7 column 1 line 27: it is not so much unreliable and biased - they are estimating a different parameter (in an unbiased fashion) - they estimate the mean of log(Ax+1) data, and a problem is that this has no natural interpretation in terms of the original data. It is fair to make the point that it is biased as an estimate of the mean response, but it would be helpful to explain why and how the estimators have problematic interpretations."*

Response: We added discussion of transformation bias, see also comment 4.

Comment 36: *"page 7 column 1 line 50: a priori not a priory*
page 7 column 2 line 37: data are (data being the plural of datum)
page 7 column 2 line 42: GLMs have (or GLM has)"

Response: We are thankful for pointing to these errors and changed accordingly.

Comment 37: *"page 7 column 2 line 44: the higher power in Fig 5 seemed to be through the GLMs being more conservative, rather than due to greater efficiency. This issue could likely be fixed using a parametric bootstrap, as was done previously in figure 2. A related point is that the linear model had the advantage that it seemed to be better at maintaining nominal significance levels at small sample sizes, which should be mentioned. As the authors say however it did not have as good power (and also loses something in interpretability and biased estimation of means). Hence while the GLM has a number of advantages, combining it with param bootstrap at small sample sizes might be worth considering to address the Type I error issue."*

Response: We agree and pointed to parametric bootstrap for situations of low sample sizes or estimates on the boundary of parameter space. See also comment 31

Comment 38: *"Supplement 2 is a useful addition, and it might be worth giving it a little more emphasis in the text to ensure ecotoxicologists wanting to try out new analyses give it a look. A couple of suggestions though: - there is no parametric bootstrap code there at the moment. This would be easy enough to do by using the mvabund package, calling manyglm (even on a univariate response) then anova(..., resamp="monte.carlo") will do a parametric bootstrap. - you can construct residual plots for GLMs on this package, using the plot function on a fitted manyglm object."*

Response: We are thankful for this comment. We refined Supplement 2 to incorporate all described models and methods and point the reader to it in the methods section.

Response to Reviewer 3

Comment 39: *"With its simulations in an ecotox. context, the paper is a fine contribution to make researchers aware of the possibilities and advantages of a generalized linear models (GLM). What is lacking is pointing to the disaster of using GLM in an improper way, notably without adjustment for overdispersion. The paper and ab-*

stract can still be made more useful by given not just the recommendation for GLM but also to give advice and warnings of proper usage. For example, (almost) never use loglinear/Poisson regression without adjustment for overdispersion, be aware that a nicer looking model like the negative binomial may give inflated type I error, a case that is overcome in this paper by bootstrapping. The same applies to binomial/logit regression unless the experiment really consist of independent observations for each 0/1 result (and is not just a count with a predetermined maximum). In this light it is strange that the logistic/binomial case is treated differently from the loglinear/Poisson case! Should this not be repaired or at least given more warning.
(1) Also, if the authors stick to the binomial/logistic without overdispersion, the normal model could be made powerful in the simulations by treating the error variance as known, because it is a known constant for the arcsine transformation.
(2) Also, the simulation now use slight overdispersion whereas the example /typical data has large overdispersion. Add simulations with large overdispersion. The current ones can go to supplementary.
(3) Without the use of a GLM equivalent of the Williams test all the advantage of the use of GLM in terms of power are gone. See the example. Discuss this ambiguity. You can perhaps use a bootstrap test based on (GLM?) monotonic regression or similar. I know some cues/leads in this direction."

Response: We rerun our simulations and now include also the poisson GLM with overdispersed data. This clearly shows the effect of not accounting for overdispersion and is discussed now.
We stick to the binomial simulations without overdispersion, because we don't know if overdispersion is a common phenomena with ecotoxcicological data. However, we point the reader to overdispersion and methods to account for it.
Our simulations mimicked the data of the case study and we clarified this (see also comment 48).
Williams type contrasts are also available in a GLM framework. The choice of which contrasts to use really depends on the assumptions one wants to make and which questions to answer. However, it does not affect our comparisons between methods. We stress this in the revised MS.

Comment 40: *"Page 1 Line 22 right column: a continuous proportion is fine for area cover and so on, but not for proportion of surviving animals where is just "k out of n" turn into a proportion. Discrete therefore. "*

Response: Thank you for pointing to this error. Of course, we meant discrete proportions and changed accordingly.

Comment 41: *"Line 58 right add ref to (Warton 2005)."*

Response: This reference fits very well and we added it.

Comment 42: *"Page 2 Line 40 : , for y¿0 should for y¿= 0 and be on the next line without an indent after formula (if it is not new paragraph)."*

Response: We clarified this equation and moved this part of eqn. 1 to the text: *"The factor A was chosen in such way that Ay equals 2 for the lowest non-zero abundance value (y)."*

Comment 43: *"Line 51 x_i is undefined. The model appear to specify a simple linear regression or a control-one_treatment model. Neither is used in the paper!!! Modify the notation therefore.."*

Response: We clarified this whole section. Eqn. 2 is in fact a simple linear regression, but with dummy coded predictors / treatments.

Comment 44: *"Line 59 Mention backtransformation here, which works fine for the confidence intervals, and not the backtransformed is the median on the original scale. If you want to backtransform to the original mean, use exp(mean+ 0.5*sigma2). (Aitchison & Brown 1969) page 8 (Jongman et al. 1995)page 19. "*

Response: We agree and added a discussion of backtransformation bias. See also comment 4.

Comment 45: *"Page 3 line 17 n= 4.10 = 40 what does the mean here/ what is the purpose?"*

Response: We are sorry for this error. It was a remnant of an earlier version of the manuscript - removed.

Comment 46: *"Line 41 : give more attention to the condition for the binomial: independent observations of each of the successes (k) out of the number of experiments (n). And: thus give attention to overdispersion in this case as well, particularly as you found nasty things for the LR of NB."*

Response: We now point the reader to overdispersed binomial data, though our simulation design does not cover this case.

Comment 47: *"Line 28 right. Ref for Dunnett contrasts."*

Response: We agree and added reference to the original description of Dunnett (1955).

Comment 48: *"Line 59 right. Only slight overdispersion, whereas the example data have large overdispersion. Add large overdispersion."*

Response: We clarified this section. We now specify the overdisperion in terms of κ in both, the case study and the simulation design. The simulated data mimicked the case study and such data is common for mesocosm studies in ecotoxicology.

Comment 49: *"Page 4 L 19 (missing."*

Response: We fixed the wrong citation type.

Comment 50: *"Line 56 left to 2 right. Four p-values: what to believe? Your simulations (although with small overdispersion) show it. The LR of NB cannot be trusted. So the 0.016 is out. The remaining tests all show about the same p-value if you interpret statistics properly. Please note that the distinction between significant and non-signicant (here p = 0.061 and 0.042) in itself is statistically insignificant. (Gelman & Stern 2006)."*

Response: We fully agree with the reviewer and cautioned that p-values should not be overinterpreted by ecotoxicologists (citing Gelman & Stern 2006). Parameter estimates (and their uncertainty) give much more information - on which we also emphasized in Figure 1.

Comment 51: *"Line 4 right : the backtransformed mean gives a median on the original scale, a value that is more interesting than the mean (generally for skew data), and for skew right always smaller than the mean."*

Response: We agree and added a discussion of back-transformation bias. See also comments 4 and 44.

Comment 52: *"Section 3.2 Always start with the Type I error results because the power results are without meaning if the Type I error is inflated. This also applies the order* *of the figs: interchange the two rows of subfigs. "*

Response: We agree and changed the order (first type I error, then power) in the text, as well as in the figures.

Comment 53: *"21 right: remove GLM_nb in this sentence as GLM_nb must be discarded due to inflated Type I error."*

Response: We agree and changed accordingly.

Comment 54: *"L32 right. ", but this." Although you should reorder the results, I mention that the "but this" should at least be "This could fortunately be ..."[and we could not have known for sure (apart from the general warnings about LR) without the simulations] This inflated error for GLM_nb must receive more attention. It is an interesting result of this study: you must to bootstrap when using NB."*

Response: We reordered and edited the respective results and discussion sections and emphasized this.

Comment 55: *"Page 5 Fig. 2 legend. Add that n =100 simulations. Add the interpretation that GLM_nb has inflated type I error, and that its line for power is just added for completeness and should not be interpreted as an estimate of the true power."*

Response: We unified the simulations and generated simulated 1000 data sets for each type of scenario (see comment 14). Moreover, we added
"Power levels for models with inflated type I error are shown for completeness."
to figure captions 2 and 3.

Comment 56: *"Page 6 Fig.5 LM for 0.8 has a high Type I error. Does it deviate significantly from 0.05? [same question for the high Type I for GLM_nb in fig 2]. Use confidence interval for binomial p = 0.05 n = 100."*

Response: We refined our simulations (see comment 14,)

Comment 57: *"Case study : You can check whether the difference between log(2y+1) and log(Ay+1) with A = 0.182 already causes the difference. (it probably does not). The important difference is the use of the Willems test! Without the use of a GLM equivalent of the Williams test all the advantage of the use of GLM in terms of power are gone! See the report of the Williams test below."*

Response: The reviewer is correct that the Williams test has higher power (if the assumptions are met). Williams type contrast are also available in a GLM framework and the choice of contrast to use really depends on the assumptions one wants to make and which questions to answer. This explains the difference between the results of Brock 2015 and us, but it does not affect our comparisons between methods.

Comment 58: *"Page 7 L35 Move ref to Warton to after data."*

Response: We changed accordingly.

Comment 59: *"L45 Add a reference to (ter Braak & Šmilauer 2014) who advocate the use of the transformation approach in a multivariate (dimension reduction) context."*

Response: We agree and added reference to ter Braak & Šmilauer 2014.

Comment 60: *"Line 1 right (like GLMs) ->(like GLMs and the Williams test)"*

Response: We did not investigate the differences between multiple comparison procedures (MCP). We used Dunnett constrasts in all simulations, because only these allow individual comparisons between treatments and the control. For a comparison of MCP see Jaki and Hothorn (2013), who recommend a combination of both, Dunnett and Williams constrasts.

Comment 61: *"Line 10-40 Again: discuss Type I error first. Then no mention of GLM_nb in power."*

Response: We agree and changed accordingly. See also comment 52.

Comment 62: *"Line 18 right. Add after "non-normal data' the point the importance of proper accounting for under and overdispersion when using GLM. What is lacking is pointing to the disaster of using GLM in an improper way, notably without adjustment for overdispersion."*

Response: We rerun our simulations and also included the Poisson GLM. The results for Poisson GLM are now also displayed in Figures 2+3 and we discuss the consequences of failing to account for overdisperion.

Comment 63: *"Line 41 right size of 9 ->size of 6-9 (?)"*

Response: We rerun our simulations and changed this to
for all simulated sample sizes.

Comment 64: *"Line 44 right the transformation ->the LM"*

Response: We changed accordingly.

References

van den Brink PJ, Hattink J, Brock TCM, Bransen F, 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

Brock TCM, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte HT, Roessink I, Strauss T, Van den Brink PJ (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

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6(2):65–70

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

OECD (2004) Test No. 202: Daphnia sp. Acute Immobilisation Test. Organisation for Economic Co-operation and Development, Paris, URL `http://www.oecd-ilibrary.org/content/book/9789264069947-en`

Szöcs E, Van Den Brink PJ, Lagadic L, Caquet T, Roucaute M, Auber A, Bayona Y, Liess M, Ebke P, Ippolito A, Ter Braak CJ, Brock CM, Schäfer RB (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: A comparison of methods. Ecotoxicology DOI 10.1007/s10646-015-1421-0

Warton DI (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16(3):275–289

Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92(1):3–10

Environmental Science and Pollution Research manuscript No.
(will be inserted by the editor)

Ecotoxicology is not normal.

~~How the use of proper statistical models can increase statistical power in ecotoxicological experiments~~A comparison of statistical approaches for analysis of count and proportion data in ecotoxicology.

Eduard Szöcs · Ralf B. Schäfer

Received: date / Accepted: date

Abstract ~~Ecotoxicologists are often confronted with non-normally distributeddata~~Counts and proportions are data types often encountered by ecotoxicologists, which are rarely normally distributed. 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. Generalised Linear Models (GLM) allow directly model distributions fitting such data. Here, we compare the ~~statistical power of analyses using transformed dataor~~ performance of parametric methods assuming (1) normality of transformed data, (2) appropriate distributions (Poisson, negativ binomial, binomial) and (3) non-parametric ~~methods to analyses using appropriate distributional assumptions, namely Generalised Linear Models (GLM)~~ methods.

We simulated ~~data mimicking~~ typical data mimicking low replicated ecotoxicological experiments of two common data types (counts and proportions from counts). We ~~compare~~ compared the performance of different methods in terms of statistical power and type 1 error for detecting a general treatment effect and determining the lowest observed effect concentration (LOEC). 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~~ For counts, we found that the quasi-Poisson model and the negative binomial model in combination with the parametric boostrap had higher statistical power than data transformation. For proportions GLMs performed better, except to determine LOEC at extremly low sample sizes. The compared non-parametric methods ~~.~~ had generally lower power.

We recommend that ~~non-normal data~~ counts and proportions from counts should be analysed by ~~GLMs and not by transformations or non-parametric methods.~~ making appropriate distributional assumptions and GLMs should become a standard method in ecotoxicology.

Keywords Generalized Linear Models · Transformations · Simulation · Power · Type 1 error

1 Introduction

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

Eduard Szöcs (✉) and Ralf B. Schäfer
Institute for Environmental Sciences
University Koblenz-Landau
Fortstraße 7,
76829 Landau, Germany
Tel.: +49 06341 280 31552
E-mail: szoecs@uni-landau.de

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~~ counts or proportions from counts in a statistically sound way (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 (counts), Warton and Hui 2011; Warton 2005 (proportions from counts)).

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 . Especially in mesocosm studies L/NOEC calculations are used in the majority of mesocosm~~ the predominant endpoint in 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 from counts.

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):

$$y_{i}^{T}{}_{i}^{T} = log(Ay_i + 1) A = 2 \; / \; min(y) \quad , \text{for } y > 0 \tag{1}$$

, where y_i is the measured ~~abundance~~ and y_i^T the transformed abundance ~~.~~ of the ith observation. The factor A was chosen in such way that Ay equals 2 for the lowest non-zero abundance value (y).

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

$$y_i^T \sim N(\mu_i, \sigma^2)$$
$$E(y_i^T) = \alpha + \beta x_i \mu_i \text{ and } var(y_i^T) = \sigma^2 \tag{2}$$
$$\mu_i = \beta Treatment_i$$

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 (~~distribution of the transformed abundances. The expected value for each observation i is given by its mean (μ_i) and the variance (σ^2) is constant between treatments. We allow this mean to vary between treatments and β ~~) are directly interpretable as changes from the control group~~ are the coefficients related to these changes in transformed abundances between treatments (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 (GLM_p):

$$y_i \sim P(\lambda \mu_i)$$
$$log(\lambda E(y_i) = var(y_i) = \mu_i \tag{3}$$
$$\mu_i log(\mu_i) = \alpha + \beta x \beta Treatment_i var(y_i) = \lambda_i$$

~~Again, this model was parametrised as contrast to the control group. The response variable is linked to the predictors via a log-function~~ This model assumes poisson distributed abundances with mean $\mu_i \geq 0$. The expected value for each observation i is given by its mean. Moreover, this model assumes that mean and variance are equal. We are modelling the mean as a function of treatment membership. However, to avoid negative ~~fitted values~~ values of the mean this is done on a log scale. Therefore, β also describes the differences between treatments on a log scale (eqn. 3). ~~The Poisson distribution assumes that~~

The assumption of equal 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~~ model (GLM_{qp}) could be used~~which introduces an additional overdispersion parameter (Θ) (GLM_{qp},~~, which assumes that variance is a linear function of the mean (eqn. 4)~~.~~ :

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

Here, Θ is used to account for additional variation and is known as overdispersion parameter. The quasi-Poisson model ~~yields to parameter estimates equal to the Poisson model~~ is a post hoc method, meaning that first a Poisson model is estimated (eqn. 3) ~~, but with standard errors~~ and than the standard errors are scaled by the degree of overdispersion.

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

$$y_i \sim NB(\cancel{\lambda}\mu_i, \kappa)$$

$$\cancel{var}E(y_i) = \mu_i \text{ and } var(y_i) = \cancel{\lambda}\mu_i + \cancel{\kappa\lambda_i^2}\mu_i^2/\kappa \tag{5}$$

$$log(\mu_i) = \beta Treatment_i$$

~~In both cases the parametrisation and link function is equal~~ This models assumes that abundances are negative binomially distributed, with a mean of $\mu_i \geq 0$ and a variance $\mu_i + \mu_i^2/\kappa$. Similar to the Poisson ~~GLM (eqn. 3).~~ model we use a log link between mean and treatments. 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

A binomial variable counts how often an event x occurs in a fixed number of independent trials N (e.g. *"5 out of 10 fish survived"*), with an equal probability of occurrence π between trials. The number of times an event occurs can also be calculated as proportion x/N.

2.2.1 Linear model for transformed data

To accommodate the assumptions for the standard linear model with such proportions, 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 total 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~~ A more natural way to model such data is the binomial distribution with parameters N and π (GLM_{bin}):

$$y_i \sim Bin(N, \pi_i)$$

$$\cancel{logit\ (\pi_i)}E(y_i) = \cancel{\alpha + \beta x_i}\pi_i \times N \text{ and } var(y_i) = \pi_i(1-\pi_i)/N \tag{7}$$

$$logit\ (\pi_i) = \beta Treatment_i$$

This model assumes that the number of occurences are binomially distributed, where N = number of ~~exposed animals and π~~ trials (e.g. exposed animals) and π_i is the probability of ~~survival~~occurrences (fish survived), which together give the expected number of occurrences. The variance of the binomial distribution is a quadratic function of the mean. We are modelling the probability of occurrence as function of treatment membership and to ensure that $0 < \pi_i < 1$ we do this on a logit scale (eqn. 7). ~~The~~ However, 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~~between treatments.

Similarly to counts, binomial data may also show overdispersion. Methods to deal with overdispersed binomial data are either quasi methods (see above) or Generalized Linear Mixed models (GLMM). However, these are not further investigated in this paper (see Warton and Hui (2011) for a comparison).

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 the parametric bootstrap (Faraway 2006) to assess the significance of the LR. Bootstrapping is computationally very intensive and for this reason we applied it only for the negative binomial models (using 500 bootstrap samples, denoted as GLM_{npb}).

To assess the LOEC we used Dunnett contrasts (Dunnett 1955) 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. We adjusted for multiple testing using the method of Holm (1979).

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~~ overdispersion at all treatments ($\cancel{\kappa = 0.25}\ \kappa = 4$, 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 ($\mu_{T2} = \ldots = \mu_{T5} = 0.5\ \mu_C = 0.5\ \mu_{T1}$), resulting in a theoretical LOEC at T2. Mean abundance was kept equal between all groups in Type 1 error simulations.

~~We generated 100~~ We generated 1000 data sets for each combination of N and μ_C and analysed these using the

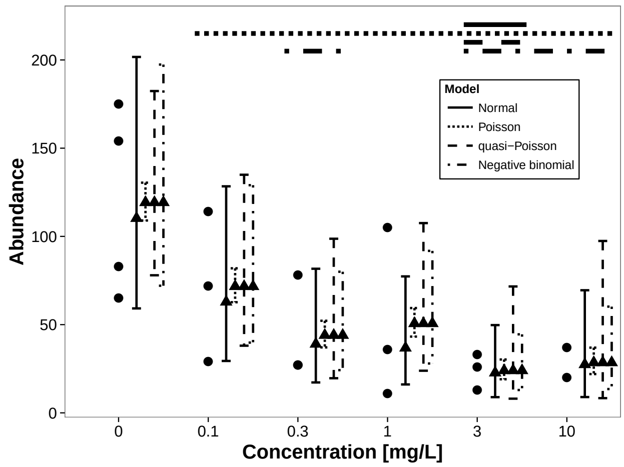

Fig. 1 Data from Brock et al (2015) (dots). Predicted values (triangles) and 95% Wald Z or t confidence intervals from the fitted models (vertical lines) are given beside. Horizontal bars above indicate treatments statistically significant different from the control group (Dunnett contrasts). The data showed ~~considerable~~ overdispersion ($\cancel{\Theta = 22.41}\ \kappa = 4$) and therefore, the Poisson model underestimates the width of confidence intervals.

models outlined ~~previously. We did not fit Poisson models because we simulated data with overdispersion.~~ in section 2.1.

2.5.2 Binomial data

We simulated data from a commonly used design as ~~in~~ described 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 ($\cancel{\pi}\ \pi_C$ = $\{0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$) and varying number of replicates (N = $\{3, 6, 9\}$). For Type 1 error estimation, $\cancel{\pi}\ \pi_C$ was held constant between groups. For power estimation $\cancel{\pi}\ \pi_C$ 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~~ 1000 data sets and analysed these using the models outlined in section 2.2.

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~~ detection of an effect when there is ~~no effect~~none) and power (ability to detect an effect when it is present). ~~All computations~~

All simulations were done in R (Version 3.1.2) (R Core Team 2014) on ~~a Linux machine~~an Amazon EC2 virtual Linux server (64bit, 15GB RAM, 8 cores, 2.8 GHz). Source code ~~for~~ to reproduce the simulations and ~~analysis of the~~

~~case study~~ paper is available online at https://github.com/EDiLD/usetheglm. Moreover, Supplement 2 provides worked examples of the data of Brock et al (2015) and Weber et al (1989).

3 Results

3.1 Case study

The data set showed considerable ~~overdispersion~~ higher variance then expected by the Poisson model ($\Theta = 22.41$, 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%.~~ detecting a general treatment effect, GLM_{nb} and GLM_p showed inflated type 1 error rates, whereas *KW* was conservative at low sample sizes. However, using parametric bootstrap for the negative binomial model (GLM_{npb}) resulted in appropriate type 1 error rates. For detecting a treatment effect~~GLM_{nb}~~, GLM_{npb} 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). 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%. At small sample sizes (n = 3, 6) and low abundances (μ_C = 2, 4) many of the negative binomial models (GLM_{nb} and GLM_{npb}) did not converge to a solution (convergence rate <~~80~~85% of the simulations, Supplement 1).

For LOEC determination GLM_{nb} and GLM_p showed an increased Type 1 error and all other methods were slightly conservative. 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~~ showed a mean reduction of 20.7% and GLM_{qp} of 24.3 %. Power to detect the LOEC was highest for ~~GLM_{nb} and GLM_{npb}~~GLM_{qp}. *LM* and ~~WT~~ *WT* showed less power, with ~~WT~~ *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 slightly increased type 1 error rates at low sample sizes and small effect sizes. *KW* was more conservative than *LM* and GLM_{bin}. In addition, GLM_{bin} exhibited the greatest power for testing the treatment effect. This was especially apparent at low sample sizes (n = 3), with up to ~~24~~27% 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).

For inference on LOEC we found that all methods were slightly conservative. *WT* was generally more conservative and GLM_{bin} especially at low effect sizes ($p_E > 0.7$). 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~~, 6). *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).

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 and endpoints like the LOEC. Therefore, ecotoxicologists should not base their inferences solely on statistical significance tests, but also on parameter estimates, their uncertainty and importance (Gelman and Stern 2006). Nevertheless, O'Hara and Kotze

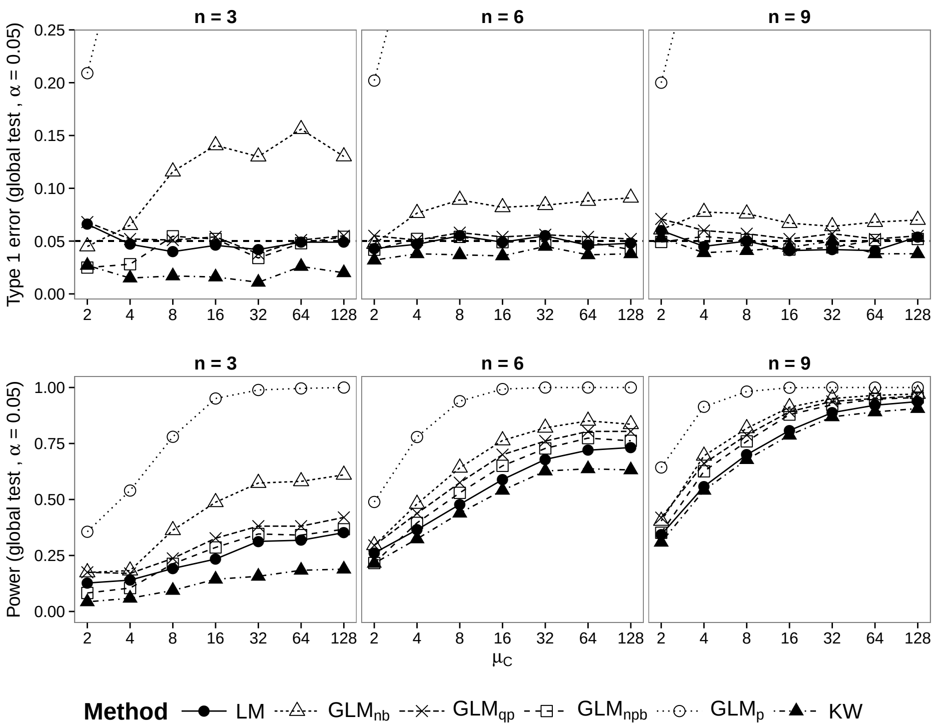

Fig. 2 Count data simulations: ~~Power~~ Type 1 error (top) and ~~Type 1 error~~ Power (bottom) for the test of a treatment effect. Only type 1 errors <25% are displayed. GLM_p showed type 1 errors >20% in all simulation scenarios. Power levels for models with inflated type I error are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than ~~80~~85% of GLM_{nb} and GLM_{npb} models did converge. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

(2010) showed that *LM* using a log transformation gave unreliable and biased parameter estimates, whereas GLMs performed well with little bias. Bias occurs also when back-transforming means to the original scale, which explains the lower predicted means by *LM* in Figure 1 (Rothery 1988) and should be corrected for (Newman 1993).

This is further highlighted by the fact that for the same model (linear model of transformed data), Brock et al (2015) reported a 10-fold lower LOEC (0.3 mg/L) then found in our study (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,~~ However, this contributed only little to the differences. A much bigger impact had the type of multiple comparison: Brock et al (2015) used a one-sided Williams test ~~which assumes a monotonic dose-response relationship. In contrast, we used a~~ (Williams 1972), whereas we used one-sided ~~Dunnett test, which does not assume monotonicity and allows~~ comparisons to the control (Dunnett contrasts). In contrast to the Williams test, Dunnett contrasts do not assume a monotonic dose-response relationship and allow individual comparisons between treatment groups and the control~~, but has under monotonicity less power~~. However, under monotonicity they have less power, which explains the differences (Jaki and Hothorn 2013). Both types of multiple comparisons are available as multiple contrast tests in a GLM framework. Therefore, our comparison of methods should be independent from the choice of contrast, which is determined by assumptions and research questions.

~~Moreover,~~ Overdispersion is common for ecological datasets (Warton 2005) and the case study illustrates the potential effects of overdispersion that is not accounted for: standard ~~error~~ errors will be underestimated and significance overestimated (~~Figure~~ Figures 1). This is also shown by our simulations (Figures 2, 3) where GLM_p showed increased type 1 error rates because of overdispersed simulated data. However, in factorial designs the mean-variance relationship can be easily checked by plotting mean versus variance of the treatment groups (see ~~supplemental material~~Supplement 2). 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~~ using knowledge about the data (e.g. ~~bonds~~bounds, 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~~tools.

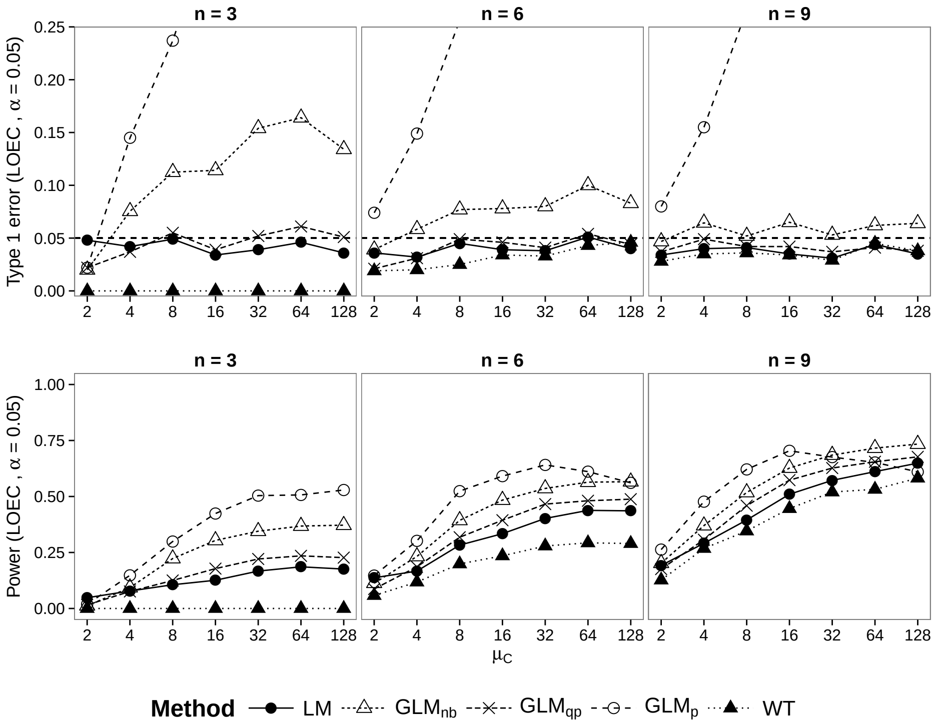

Fig. 3 Count data simulations: ~~Power~~ Type 1 error (top) and ~~Type 1 error~~ Power (bottom) for determination of LOEC. For clarity only type 1 errors <25% are displayed. Power levels for models with inflated type I error are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than ~~80~~85% of GLM_{nb} ~~and GLM_{npb}~~ models did converge. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

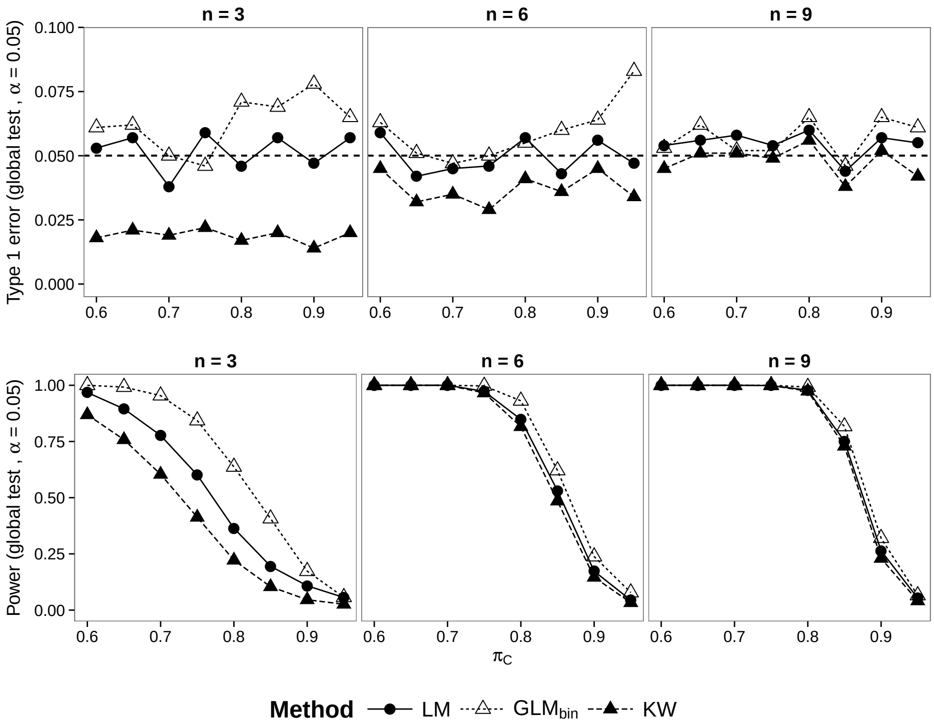

Fig. 4 Binomial data simulations: ~~Power~~ Type 1 error (top) and ~~Type 1 error~~ power (bottom) for the test of a treatment effect. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

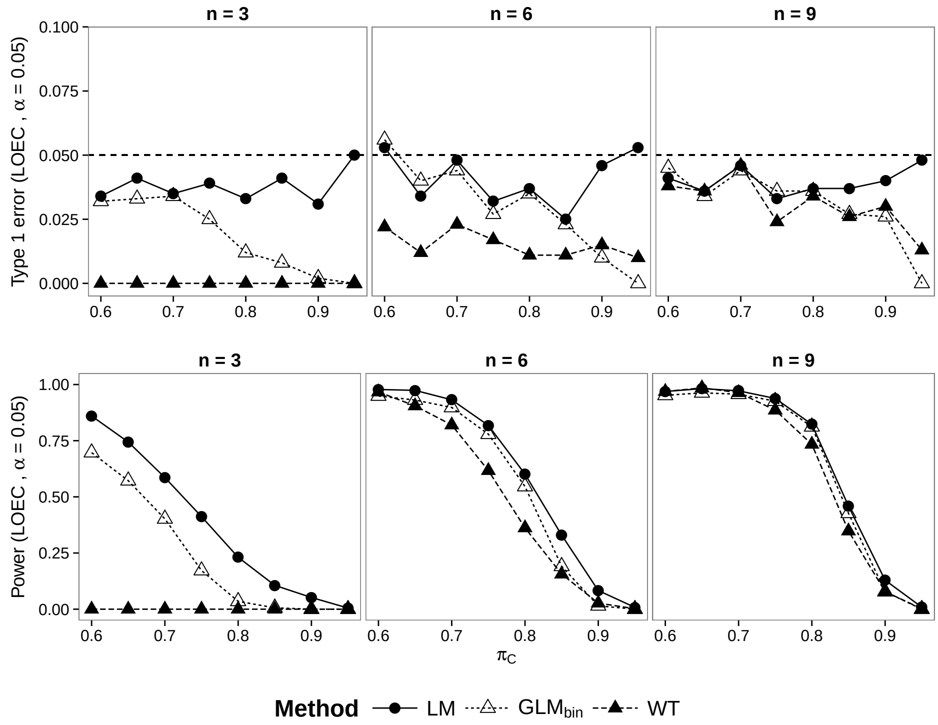

Fig. 5 Binomial data simulations: ~~Power~~ Type 1 error (top) and ~~Type 1 error~~ power (bottom) for the test for determination of LOEC. Dashed horizontal line denotes the nominal Type 1 error rate at $\alpha = 0.05$.

Therefore, choosing an appropriate model is ~~more sound and straightforward~~ easier 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 $(n \leq 4)$ 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, 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~~ 25%), which can be attributed to multiple testing. The low power of all methods to detect significant treatment levels such as the LOEC or NOEC suggests that these endpoints from ecotoxicological studies should be interpreted with caution and underpins ~~the criticism of NOEC~~ their criticism (Laskowski 1995; Landis and Chapman 2011).

Mesocosm studies allow also inferences on community level. For community analyses *GLM for multivariate data* (Warton et al 2012) 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). However, ter Braak and Šmilauer (2014) argue to use data transformations with community data because of their simplicity and robustness. 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 at the population level 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~~ priori* 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~~ *KW* and *WT* have less power compared to ~~tests assuming normality after data transformation~~*LM*. Moreover, ~~GLMs~~ *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~~ We found an increased Type-I error for GLM_{nb} at low sample sizes. However, it is well known that the LR statistic is not reliable at small sample sizes ~~and~~ (Bolker et al 2009; Wilks 1938). Parametric bootstrap (GLM_{npb}) is a valuable alternative in such situations and maintains appropriate levels (Figure 2). Moreover, 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 . We found that parametric bootstrap (~~

GLM_{qp} showed higher statistical power than GLM_{npb} ~~) provides a valuable alternative in such situations~~ (Figure 2~~). At~~ , bottom). This could be explained by the simpler mean-variance relationship of GLM_{qp} (eqn. 4 and 5), because 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_{npb} at small sample sizes (Figure 2, top).~~

Binomial data ~~is~~ are often collected in lab trials, where increasing the sample size ~~is~~ may be relatively easy to accomplish. We found notable differences in power to detect a treatment effect ~~up to a sample size of 9.~~ for all simulated sample sizes. Similarly, Warton and Hui (2011) also found that ~~GLM~~ GLMs have higher power than arcsine transformed linear models. ~~Nevertheless, for deriving LOECs the transformation performed better~~ Though we did not simulate overdispersed binomial data, this should be checked and accounted for. In such situations a GLMM may offer an appealing alternative (Warton and Hui 2011). At low effect sizes GLM_{bin} became conservative with increasing π_C, although this effect lessened as sample size increased (Figure 5). This is because π approaches its boundary and is also known as the *Hauck-Donner effect* (Hauck and Donner 1977). A LR-Test or parametric bootstrap may provide an alternative in such situations (Bolker et al 2009). This can also explain why *LM* performed better for deriving LOECs at low sample sizes~~(n = 3)(Figure 5).~~ .

GLMs can be fitted with several statistical software packages and many textbooks are available to introduce ecotoxicologists to these models (e.g. Zuur 2013 or Quinn and Keough 2009). 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,~~ ecotoxicologists should change their models instead of their data. GLMs should become a standard method in ecotoxicology and incorporated into respective guidelines.

5 Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.

References

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

ter Braak CJF, Šmilauer P (2014) Topics in constrained and unconstrained ordination. Plant Ecology DOI 10.1007/s11258-014-0356-5, URL http://link.springer.com/10.1007/s11258-014-0356-5

van den Brink PJ, Hattink J, Brock TCM, Bransen F, 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

Brock TCM, Hammers-Wirtz M, Hommen U, Preuss TG, Ratte HT, Roessink I, Strauss T, Van den Brink PJ (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

Dunnett CW (1955) A Multiple Comparison Procedure for Comparing Several Treatments with a Control. Journal of the American Statistical Association 50(272):1096–1121

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 JJ (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

Gelman A, Stern H (2006) The difference between “significant” and “not significant” is not itself statistically significant. The American Statistician 60(4):328–331, URL http://pubs.amstat.org/doi/abs/10.1198/000313006X152649

Hauck WW, Donner A (1977) Wald's Test as Applied to Hypotheses in Logit Analysis. Journal of the American Statistical Association 72(360):851, DOI 10.2307/2286473, URL http://www.jstor.org/stable/2286473?origin=crossref

Hilbe JM (2014) Modeling Count Data. Cambridge University Press, New York, NY

Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics 6(2):65–70

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

Johnson PCD, Barry SJE, Ferguson HM, Müller P (2014) Power analysis for generalized linear mixed models in ecology and evolution. Methods in Ecology and Evolution DOI 10.1111/2041-210X.12306

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

Landis WG, Chapman PM (2011) Well past time to stop using NOELs and LOELs. Integrated Environmental Assessment and Management 7(4):vi–viii

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 JA, Wedderburn RWM (1972) Generalized Linear Models. Journal of the Royal Statistical Society Series A (General) 135(3):370–384

Newman MC (1993) Regression analysis of log-transformed data: Statistical bias and its correction. Environmental Toxicology and Chemistry 12(6):1129–1133, URL http://onlinelibrary.wiley.com/doi/10.1002/etc.5620120618/abstract

Newman MC (2012) Quantitative ecotoxicology. Taylor & Francis, Boca Raton, FL

OECD (2006) Current Approaches in the Statistical Analysis of Ecotoxicity Data: A Guidance to Application. No. 54 in Series on Testing and Assessment, OECD, Paris

O'Hara RB, Kotze DJ (2010) Do not log-transform count data. Methods in Ecology and Evolution 1(2):118–122

Quinn GP, Keough MJ (2009) Experimental design and data analysis for biologists. Cambridge Univ. Press, Cambridge

R Core Team (2014) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Rothery P (1988) A cautionary note on data transformation: bias in back-transformed means. Bird Study 35(3):219–221, DOI 10.1080/00063658809476992, URL http://www.tandfonline.com/doi/abs/10.1080/00063658809476992

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

Stroup WW (2014) Rethinking the Analysis of Non-Normal Data in Plant and Soil Science. Agronomy Journal DOI 10.2134/agronj2013.0342

Szöcs E, Van Den Brink PJ, Lagadic L, Caquet T, Roucaute M, Auber A, Bayona Y, Liess M, Ebke P, Ippolito A, Ter Braak CJ, Brock CM, Schäfer RB (2015) Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: A comparison of methods. Ecotoxicology DOI 10.1007/s10646-015-1421-0

Venables WN, Ripley BD (2002) Modern Applied Statistics with S, 4th edn. Springer, New York

Ver Hoef JM, Boveng PL (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88(11):2766–2772

Wang M, 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 DI (2005) Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics 16(3):275–289

Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. Ecology 92(1):3–10

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

Weber CI, Peltier WH, Norbert-King TJ, Horning WB, Kessler F, Menkedick JR, Neiheisel TW, Lewis PA, Klemm DJ, Pickering Q, Robinson EL, Lazorchak JM, Wymer L, Freyberg RW (1989) Short-term methods for estimating the chronic toxicity of effluents and receiving waters to fresh- water organisms. Tech. Rep. EPA/600/4–89/001, Environmental Protection Agency, Cincinnati, OH: Environmental Monitoring Systems Laboratory

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

Williams DA (1972) The comparison of several dose levels with a zero dose control. Biometrics pp 519–531, URL http://www.jstor.org/stable/10.2307/2556164

Zuur AF (2013) A beginner's guide to GLM and GLMM with R: a frequentist and Bayesian perspctive for ecologists. Highland Statistics, Newburgh

Supplement 1
Click here to download Supplementary Material: supp1.pdf

Supplement 2