Ecotoxicology is not normal.

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

Eduard Szöcs, Ralf B. Schäfer

February 2, 2015

Abstract

Ecotoxicologist are often confronted with non-normally distributed data. To achieve the assumptions of normality and heteroscedasticity is has been a standard procedure to either transform the data or use non-parametric methods if this fails. Here, we argue that using appropriate models, namely Generalised Linear Models (GLM), can enhance statistical power.

We present examples of ecotoxicological studies illustrating the differences and advantages of GLM. Using simulations of two common data types (counts and discrete proportions), we show that GLMs provide a gain in statistical power compared to analysis of transformed data or using non-parametric methods. Moreover, GLMs provide a gain in interpretability of results.

GLMs should become a standard method in ecotoxicology to analyse inherently non-normally distributed data.

1 Introduction

Ecotoxicologists perform various kinds of experiments yielding different types of data. Examples are: animal counts in mesocosm experiments (positive, integer-valued data), proportions of surviving animals (data bonded between 0 and 1, continuous) or biomass in growth experiments (positive, continuous data). These data are typically not normally distributed. Nevertheless, they are usually analysed using methods assuming a normal distribution and variance homogeneity (Wang and Riffel, 2011). To meet these assumptions, data are usually transformed. For example, ecotoxicological textbooks (Newman, 2012) and guidelines (EPA, 2002; OECD, 2006) advise that survival data can be transformed using an arcsine square root transformation. For

count data from mesocosm experiments a log(Ay + C) transformation is usually applied, where the constants A and C are either chosen arbitrarily or following general recommendations. For example, van den Brink et al. (2000) suggest to set the term Ay to be 2 for the lowest abundance value (y) greater than zero and C to 1. Moreover, other transformations like the square root or fourth root are commonly applied in community ecology. Note that there has been little evaluation and advice for practitioners, which transformations to use. If the transformed data still do not meet the assumptions (i. e. normality and variance homogeneity), non-parametric tests are usually applied (Wang and Riffel, 2011).

Generalized linear models (GLM) provide a method to analyse such non-normally distributed data (Nelder and Wedderburn, 1972). GLMs can handle various types of data distributions, e.g. Poisson or negative binomial (for count data) or binomial (for proportions); the normal distribution being a special case of GLMs. Despite GLMs being available more than 40 years, ecotoxicologists do not regularly make use of them.

Recent studies concluded that data transformations should be avoided and GLMs be used as they have better statistical properties (O'Hara and Kotze, 2010; Warton and Hui, 2011). Low sample sizes are common in ecotoxicological studies (Sanderson, 2002; Szöcs et al., 2015) and lead to low power in statistical hypothesis testing, on which many ecotoxiological approaches (e.g. risk assessment for pesticides) rely. Differences between statistical methods may be apparent especially at such low sample sizes.

We explore how GLMs may enhance inference in ecotoxicological studies. We compared three types of statistical methods (transformation and normality assumption, GLM, non-parametric tests) using simulations and demonstrate on example data sets how conclusions can be influenced by the choice of statistical method.

2 Motivating examples

2.1 Count data

Brock et al. (2015) provides a typical example data from a mesocosm study of mayfly larvae counts on artificial substrate samplers at one sampling date (Figure 1). 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)). We will use this data to demonstrate the differences between transformations, different GLMs and a non-parametric approach. A full exemplary analysis of this data using R (R Core Team, 2014) can be found in Supplement 2.
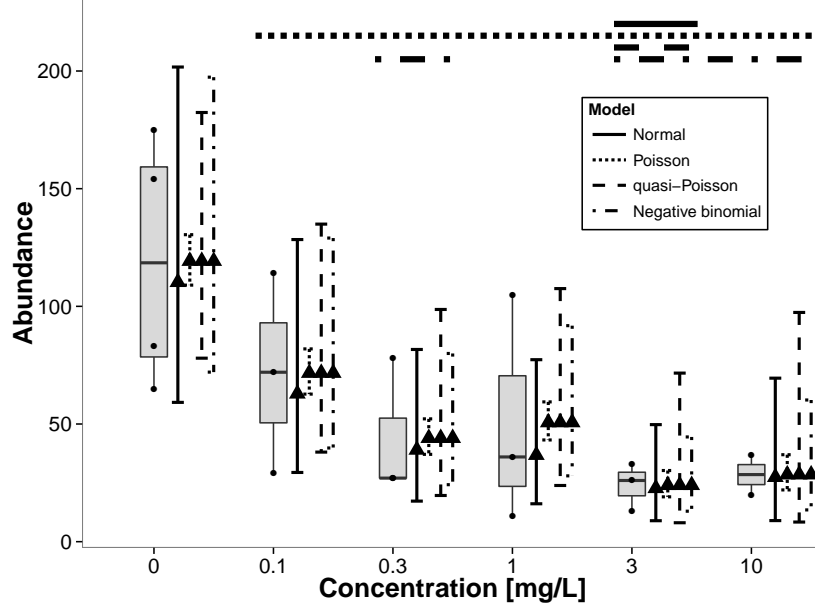
Figure 1: Data from Brock et al. (2015) (boxplots + black points within boxes). Predicted values (triangles) and 95% Wald Z or t Confidence intervals from the fitted models (vertical lines) are given besides the boxplots. Horizontal bars above indicate treatments statistically significant different from the control group (Dunnett contrasts).

2.1.1 The linear model

To meet the assumptions of the standard linear model, we transform the counts following van den Brink et al. (2000) (eqn. 1):

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

, where y_i is the measured abundance, y_i^T the transformed abundance. The lowest non-zero abundance is 11 and A = 2 / 11 = 0.182.

Then we fit the linear model to the transformed abundances:

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

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

directly interpretable as changes from the control group (eqn. 2).

2.1.2 Generalized Linear Models

GLMs are the extension of the normal model, by allowing other distributions of the response variable. Instead of transforming the response variable the counts could be directly modelled by a Poisson distribution:

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

Again, this model is parametrised as contrast to the control group. The response variable is linked via a log-function to the predictors to avoid negative fitted values (eqn. 3). The Poisson distribution assumes that mean and variance are equal - a assumption that is rarely met with ecological data which is typically characterized by greater variance (overdispersion). To overcome this problem a quasi-Poisson distribution could be used which introduces an additional overdispersion parameter (Θ) (eqn. 4).

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

Another possibility to deal with overdispersion is to use a negative binomial distribution which also introduces an additional parameter (eqn. 5).

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

In both cases the parametrisation and link function is the same as in the Poisson GLM (eqn. 3). Note, that the quasi-Poisson model assumes a linear mean-variance relationship (eqn. 4), whereas the negative binomial model assumes a quadratic relationship (eqn. 5). The above described models are most commonly used in ecology (Ver Hoef and Boveng, 2007), although other distributions for count data are possible, like the negative binomial model with a linear

mean variance relationship (also known as NB1) or the poisson inverse gaussian model (Hilbe, 2014).

2.1.3 Hypothesis testing

On this data set, we could test different hypotheses like (i) if there is any effect of the treatment or (ii) test single parameters (treatments) to determine the LOEC. Following general recommendations (Bolker et al., 2009), we used F-tests for the normal and quasi-Poisson models and Likelihood-Ratio (LR) tests for Poisson and negative binomial models to test if there is any treatment related effect. To assess the LOEC we used Dunnett contrasts with one-sided Wald t tests (normal and quasi-Poisson models) and one-sided Wald Z tests (Poisson and negative binomial models).

2.1.4 Results

The Poisson model showed considerable overdispersion and did not fit the data. This leads to underestimated standard errors and confidence intervals and overestimated statistical significance. Therefore, 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) did not indicate any treatment related effects. Whereas the LR test of the negative binomial model indicated a treatment related effect (LR = 13.99, p = 0.016).

All methods resulted in similar predicted 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 quasi-Poisson - especially at lower estimated abundances. Accordingly, the LOECs differed (Normal and quasi-Poisson: 3 mg/L, negative binomial: 0.3 mg/L).

Brock et al. (2015) assumed normality after data transformation and reported a LOEC of 0.3 mg/L for this data. The reason for this difference may be twofold: (Brock et al., 2015) used a $log(2\,y+1)$ transformation, whereas we used a $log(0.182\,y+1)$ transformation (van den Brink et al., 2000). Moreover, we applied a one-sided Dunnett test, as the toxic response in a mesocosm experiment may be either decreasing or increasing (due to biological interactions). Brock et al. (2015) used a one-sided Williams test, which is known to have larger power if the assumptions are met (Jaki and Hothorn, 2013). This example demonstrates that the choice of the statistical model and procedure has substantial impact on ecotoxicological inferences, especially when sample sizes are low.

2.2 Binomial data

Weber et al. (1989) studied fathead minnow (*Pimephales promelas*) larval survival data after sodium pentachlorophenol (NaPCP) exposure. This data set is available in Newman (2012) and EPA (2002). At six NaPCP concentrations (0, 32, 64, 128, 256, 512 µg/L) with 4 replications ten fish were exposed and proportions of total number alive at the end were reported. A full exemplary analysis of this data using R (R Core Team, 2014) can be found in Supplement 2.

2.2.1 The linear model after transformation

To accommodate the assumption for the standard linear model the EPA suggests a special arcsine square root transformation (eqn. 6) for such data (EPA, 2002):

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

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

2.2.2 Generalized Linear Models

Data of type *x out of N* can be modelled by a binomial distribution with parameters N and π:

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

, where N = number of exposed animals and π is the probability of survival. The variance of the binomial distribution is a quadratic function of the mean (eqn. 7). The parameters β of this model are directly interpretable as changes in log odds compared to the control group.

We used F-test for the normal and LR test for binomial GLM to determine a general treatment effect. LOEC was assessed using one-sided Dunnett contrasts.

Table 1: Estimated parameters and 95% Confidence Intervals for the binomial data example. Asterisks indicate LOEC as determined using one-sided Dunnett tests.

| | | Model | | |
Treatment		LM		GLM
Control	1.331	(1.169, 1.149)	2.994	(1.523, 4.366)
32 µg/L	-0.147	(-0.376, 0.081)	-1.210	(2.876, 0.456)
64 µg/L	0.041	(-0.188, 0.269)	0.719	(-1.723, 3.161)
128 µg/L	-0.076	(-0.305, 0.152)	-0.747	(-2.505, 1.010)
256 µg/L	-0.221	(-0.450, 0.007)	-1.708	(-3.312, -0.104)
512 µg/L	-0.727	(-0.956, -0.499)*	-3.675	(-5.244, -2.107)*

2.2.3 Results

Both methods lead to same ecotoxicological inferences: The global tests of both methods indicated a effect of NaPCP on larval survival (linear model: F = 13.31, p <0.001; GLM: LR = 64.79, p <0.001). Moreover, both methods identified the highest concentration (512 µg/L) as LOEC.

The coefficients of the binomial model are directly interpretable as change in odds ratio: Compared to the control group, odds in the highest treatment are reduced by a factor of $e^{-3.675} = 0.025$. Such a direct interpretation of parameters is not possible with the transformed data (Table 1).

3 Simulations

We used simulations to compare the methods described above to analyse count and binomial data. Methods were compared in terms of Type I error (maintain a significance level of 0.05 when there is no effect) and power (detect an effect when it is present). We fitted the models and tested hypotheses on the simulated data as described in the motivating examples.

All simulations were done in R (Version 3.1.2) (R Core Team, 2014) on a 64-bit Linux machine with 8 GB and 2.2 GHz. Source code for the simulations is available online at `https://github.com/EDiLD/usetheglm`.

3.1 Count data

3.1.1 Methods

We simulated count data that mimics count data encountered in mesocosm experiments with five treatments (T1 - T5) and one control group (C). Counts were drawn from a negative binomial distribution with slight over dispersion (dispersion parameter for all treatments: $\kappa = 0.25$). 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 to a theoretical LOEC at T2. Mean abundance was kept equal between all groups in Type I error simulations.

We generated 100 data sets for each combination of N and μ_C. We fitted a linear model after $\log(Ax + 1)$ transformation (LM), negative binomial GLM (GLM_{nb}) and quasi-Poisson GLM (GLM_{qp}) to these datasets. We tested a general treatment effect using F tests (LM and GLM_{qp}) and LR tests (GLM_{nb}). Additionally, we applied parametric bootstrap (500 bootstrap samples; Faraway (2006)) to assess the LR in the negative binomial model (GLM_{pb}) and used a Kruskal-Wallis test on untransformed data as a non-parametric method. To assess the performance in detecting the LOEC (T2 in our simulation design) we used a one-sided Dunnett test for all models and a one-sided pairwise Wilcoxon test as a non-parametric method. `rerun!`

3.1.2 Results

At small sample sizes (n = 3, 6) and low abundances ($\mu_C = 2, 4$) many of the negative binomial models (GLM_{nb} and GLM_{pb}) did not converge to a solution (convergence rate <80% of the simulations, Supplement 1). For this simulation design (reduction in abundance by 50%) a sample size of n = 9 was needed to achieve a power greater than 80%. GLM_{nb} showed an increased Type I error at low sample sizes. However, this decreased with increasing sample sizes (Figure 2, bottom). Using parametric bootstrap (GLM_{pb}) resulted in an appropriate Type 1 error level for the negative binomial model. LM, GLM_{qp} maintained also an appropriate Type I error. The Kruskal-Wallis test showed the least power, with low Type I error at small sample sizes. All GLM showed greater power than LM or the Kruskal test. GLM_{qp} showed up to 17% greater power compared to LM.

The inferences on LOEC generally showed less power. For LM this reduction was up to 35% (Figures 2 and 3). At low sample sizes GLM_{nb} showed an increased Type 1 error and the pairwise Wilcoxon Test had no power at all to detect the correct LOEC. GLM_{qp} and LM
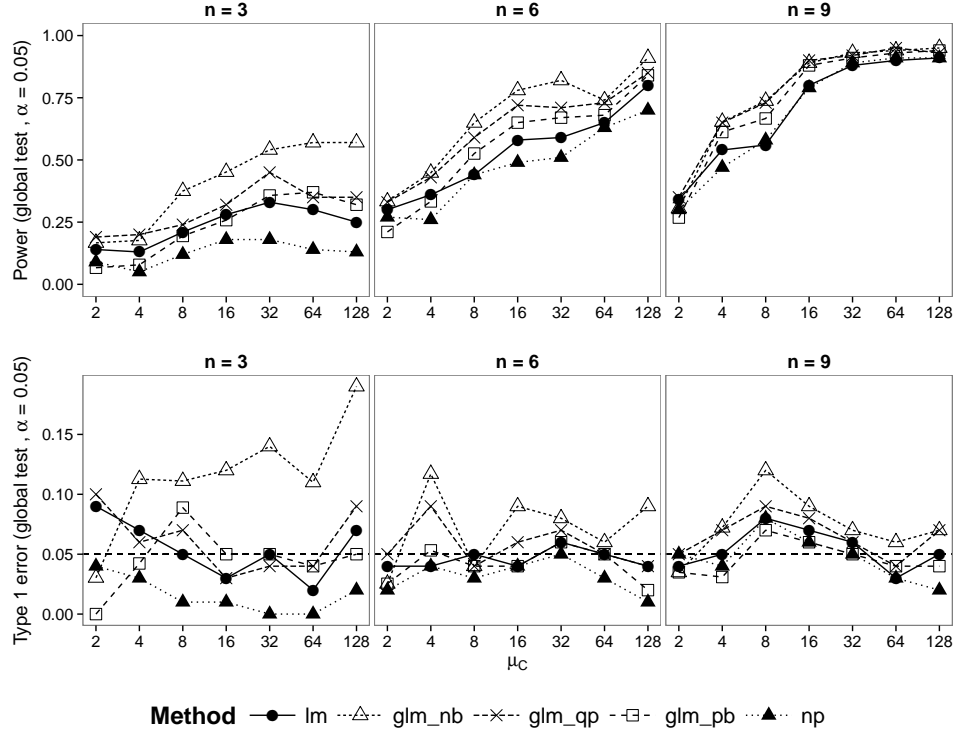
Figure 2: Simulation results for count data. Power (top) and Type I error (bottom) for the test of a treatment effect. Compared methods were: Linear model after $\log(Ax + 1)$ transformation (lm), negative binomial GLM with LRT (glm_nb), negative binomial GLM with parametric boostrap (glm_pb), quasi-Poisson GLM (glm_qp) and Kruskal-Wallis test on untransformed data (np). For n = 3 and $\mu_C = 2, 4$ less then 80% of glm_nb and glm_pb models did converge.
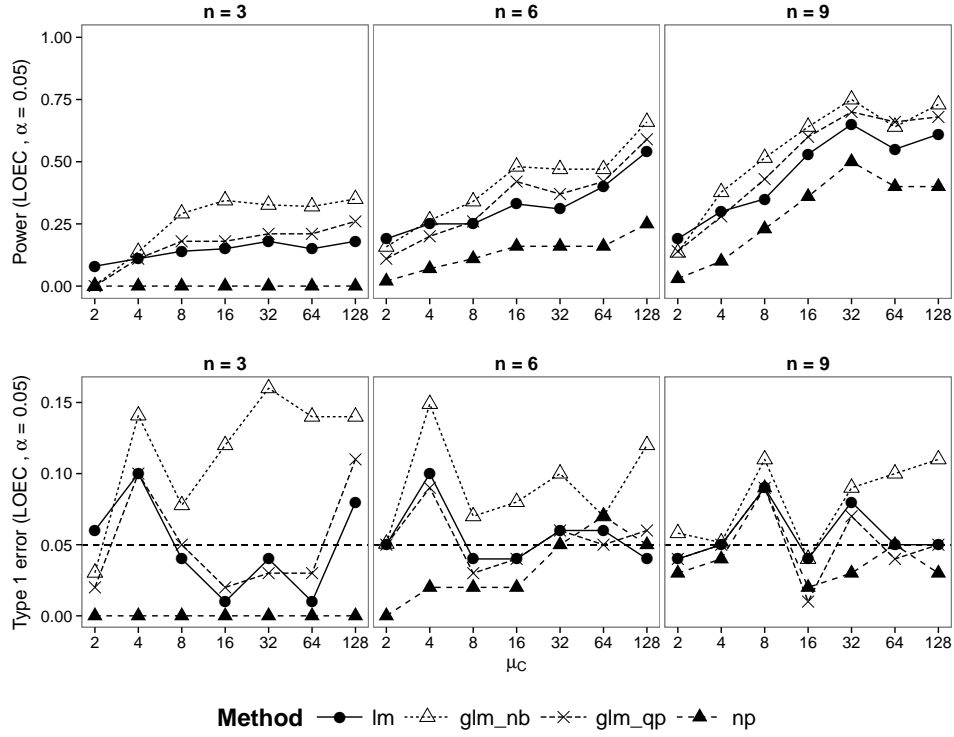
Figure 3: Simulation results for count data. Power (top) and Type I error (bottom) for determination of LOEC. Compared methods were: Linear model after log (Ax + 1) transformation (lm), negative binomial GLM with LRT (glm_nb), negative binomial GLM with parametric boostrap (glm_pb), quasi-Poisson GLM (glm_qp) and pairwise Wilcoxon test on untransformed data (np). For n = 3 and $\mu_C = 2, 4$ less then 80% of glm_nb and glm_pb models did converge.

yielded comparable Type 1 errors, with GLM_{qp} having up to 11% greater power (Figure 3, top).

3.2 Binomial data

3.2.1 Methods

We simulated data from a design as described in the motivating example, with 5 treated (T1 - T5) and a control group (C). Proportions were drawn from a Bin(10, π) distribution, with varying probability of success ($\pi = \{0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95\}$) and varying number of replicates (N = $\{3, 6, 9\}$). For Type I error estimation π was held constant between groups. For power estimation π in C and T1 was set to 0.95 and π in T2 - T5 varied between 0.6 and 0.95).

We simulated 250 data sets for each combination and analysed them using the linear model
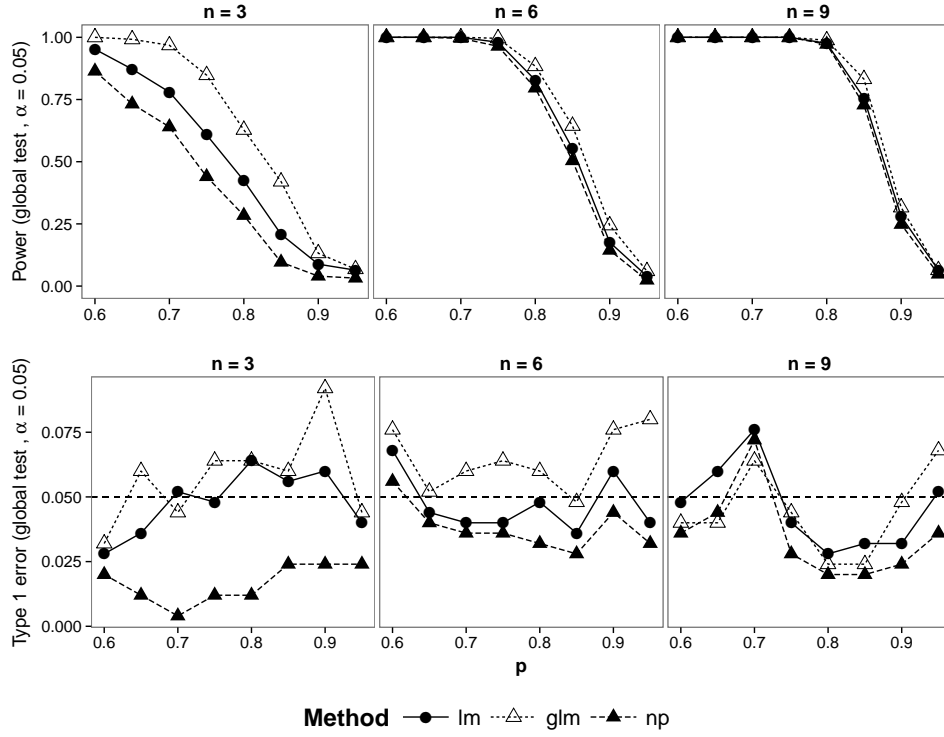
Figure 4: Simulation results for binomial data. Power (top) and Type I error (bottom) for the test of a treatment effect. Compared methods were: Linear model after arcsine square root transformation (lm), binomial GLM with LRT (glm) and Kruskal-Wallis test on untransformed data (np).

after arcsine transformation (LM), binomial GLM (GLM) and Kruskal-Wallis test. Moreover, we compared the methods' ability to determine the LOEC (T2 in our simulation design) by comparing inferences on model parameters and a pairwise Wilcoxon test.

3.2.2 Results

Binomial GLM showed the greatest power for testing the treatment effect, while maintaining an appropriate Type I error level. This was especially apparent at low sample sizes (n = 3), with up to 24% higher power. Kruskal-Wallis test had the lowest power and a low Type I error rate. However, the difference between methods quickly vanished with increasing samples sizes (Figure 4).

Inference on LOEC was not as powerful as inference on the general treatment effect. Contrary to the global test, LM showed the highest power for small sample sizes, while maintaining a Type 1 error level of 0.05. GLM had less power and showed a lower Type 1 error rate, especially with decreasing effect size. The pairwise Wilcoxon test had no power at all for n = 3 and showed less
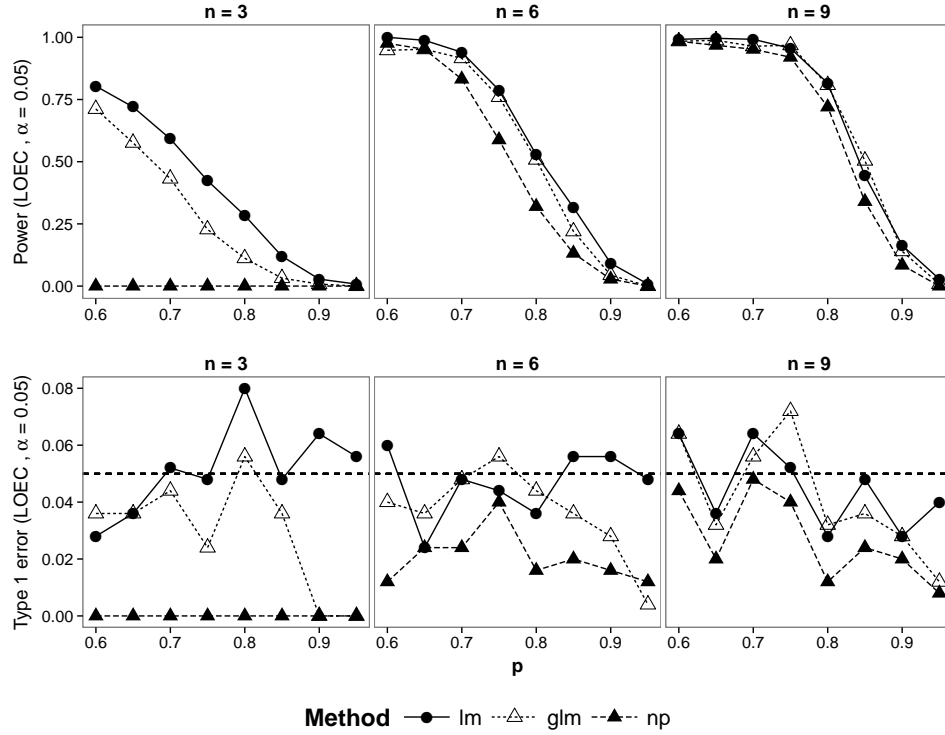
Figure 5: Simulation results for binomial data. Power (top) and Type I error (bottom) for determination of LOEC. Compared methods were: Linear model after arcsine square root transformation (lm), binomial GLM with LRT (glm) and a pairwise Wilcoxon test on untransformed data (np).

power in the other simulation runs. Differences in power to detect a LOEC was only apparent at n = 3 and vanished quickly with increasing sample sizes (Figure 5).

4 Discussion

Ecotoxicological experiments are often planned with small sample sizes due to practical constraints. For example, extremely low samples sizes (n <5) are common in cases (Sanderson, 2002; Szöcs et al., 2015). Statistical power is crucial for the determination of LOEC/NOEC values. Although the use of LOEC/NOEC has been heavily criticized in the past (Landis and Chapman, 2011), they are still regularly used in ecotoxicology (Jager, 2012). Especially in mesocosm studies NOEC calculations are used in the majority of mesocosm experiments (Brock et al., 2015; EFSA PPR, 2013). To counteract the problems with low power Brock et al. (2015) proposed to take the Minimum Detectable Difference (MDD), a method to assess statistical power *a posteriori*, for inference into account. Our results suggest that power in common mesocosm

experiments is low. For common samples sizes and a reduction in abundance of 50% we found an unacceptably low power to detect any treatment related effect (<50% for methods with appropriate Type 1 error, Figure 2). Additionally, O'Hara and Kotze (2010) showed that using a log transformation gave unreliable and biased parameter estimates. Statistical power to detect the correct LOEC was even worse, with power less than 30%. This suggests that NOECs reported from mesocosm experiments should be interpreted with caution and underpins the critics on NOEC. *A priory* power analyses can be performed easily using simulations, even for complex experimental designs (Johnson et al., 2014), and might help to design, interpret and evaluate ecotoxicological studies.

Moreover, Brock et al. (2015) proposed that statistical power of mesocosm experiments can be increased by reducing sampling variability by better sampling and quantification methods. But they also caution to avoid depleting populations by increasing sampling efficiency. As we showed, using appropriate statistical methods (like GLMs) can enhance the power at no extra costs.

It has been 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 (Wang and Riffel, 2011). In contrast, our results showed that the often applied Kruskal test and pairwise Wilcoxon test have equal or less power compared to tests assuming normality after data transformation. Moreover, GLMs always performed better than non-parametric tests. However, there might be more powerful non-parametric tests available (Konietschke et al., 2012) which we did not investigate. Non-parametric statistics are focused on testing but not on estimation of effects. Additionally to testing, GLMs allow the estimation and interpretation of effects that might not be statistically significant, but ecologically relevant. Therefore, we do not advise to use non-parametric tests for non-normal data but GLMs instead.

At small sample sizes and low abundance 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 directly the likelihood may perform better. Moreover, the Likelihood-Ratio test gave increased Type-I error for these models. It is well known that the LR statistic is unreliable for small sample sizes (Bolker et al., 2009; Wilks, 1938) and we found that parametric bootstrap provides a valuable alternative in such situations. At small samples sizes and / or low abundances it might be hard to decide which mean-variance relationship fits best. The quasi-Poisson models assumes a simpler, linear mean-variance relationship, which might explain why it performed best for our simulated data sets.

Binomial data is often collected in lab trials, where increasing sample size is easier to accom-

plish. We found notable differences in power to detect a treatment effect up to a sample size of 9. Similarly, Warton and Hui (2011) also found that GLM have higher power than arcsine transformed linear models. However, for deriving LOECs the normal model performed better at low sample sizes. At samples sizes greater than three we found no power differences for detecting the correct LOEC. The interpretation of binomial GLMs is much simpler than for the arcsine transformed data. Their power is is equal or even higher at sample sizes greater then three. Therefore, we recommend to use binomial GLM instead of the arcsine transformation.

In the introduction we pointed out that there is little advice how to choose from the plenty of possible transformations. How do GLMs simplify this problem? First of all, the distribution modelled should be chosen to give a statistically sound model. Proportions are bounded between 0 and 1 and could be modelled using a binomial distribution. Counts are positive discrete values and should be modelled with a discrete distribution. In a factorial design the mean-variance relationship can be easily checked with diagnostic plots. Moreover, it should be checked for overdispersion. Standard error will be underestimated and significance overestimated, if not accounted for (Figure 1). The model selection process can be guided by the data and diagnostic plots. Therefore, it is much more sound than choosing between transformations.

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

5 Conclusions

Statistical hypothesis tests are commonly used in environmental risk assessments to make inferences on pesticide effects. The choice how we treat, model and test the data can have massive impacts on the conclusions we draw from experiments especially at low sample sizes. We showed for two common data types in ecotoxicology, that using appropriate models resulted in higher statistical power than trying to meet the assumptions of normality and variance homogeneity using transformations. Therefore, we cannot recommend the current practice to either transform the data or use non-parametric approaches if data is not normally distributed. We recommend to use models that fit to the data. GLMs should become a standard method in ecotoxicology and guidelines need to be updated accordingly.

References

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

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

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

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

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

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

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

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

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

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

Landis, W. G. and Chapman, P. M. (2011). Well past time to stop using NOELs and LOELs. *Integrated Environmental Assessment and Management*, 7(4):vi–viii.

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

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

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

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

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

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

Stroup, W. W. (2014). Rethinking the analysis of non-normal data in plant and soil science. *Agronomy Journal*.

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

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

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

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

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

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

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

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

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