Ecotoxicology is not normal.

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

Eduard Szöcs, Ralf B. Schäfer

January 14, 2015

Abstract

1 Introduction

In environmental risk assessments (ERA) statistical tests play an important role to evaluate the effects of pesticides. Despite criticism (e.g. Landis and Chapman (2011)) statistics like the No Observed Effect Concentration (NOEC) are still regularly used to report results experiments (Jager, 2012). A critical issue of reporting a NOEC is the statistical power in the underlying experiments, i.e. the ability to detect an effect.

Ecotoxicologists perform various kinds of experiments yielding to different types of data, potentially with very low samples sizes. Examples are: animal counts in mesocosm experiments (positive, integer valued, discrete data), proportions of surviving animals (discrete, bonded between 0 and 1) or biomass in growth experiments (strictly positive data).

Such data are usually analysed by using methods assuming normal distributed data, although these types are inherently not normally distributed (Wang and Riffel, 2011). In order to approximate the normality and variance homogeneity assumptions data is usually transformed. It is advised that survival data can be transformed using an arcsine square root transformation (Newman, 2012; OECD, 2006). For count data from mesocosm experiments a $\log(Ax + 1)$ transformation is usually usually, where the constant A is either chosen arbitrarily or following the recommendation of van den Brink et al. (2000): Ax to be 2 for the lowest abundance value (x) greater than zero. Note, that there has been little evaluation and advice for the practitioners which transformations to use. If the transformed data does not meet the normality assumptions, usually non-parametric tests are applied (Wang and Riffel, 2011).

Generalized linear models (GLM) are a third possibility to analyse such not 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 discrete proportions); the normal distribution being a special case of GLMs. Despite that GLMs were available more than 40 years now, ecotoxicologists do not regularly make use of them.

Recently studies concluded that data transformations should be avoided and GLMs be used as they have better statistical properties (*Do not log-transform count data*, (O'Hara and Kotze, 2010); *The arcsine is asinine*, (Warton and Hui, 2011)). Especially in the light of low sample sizes, which are common in ecotoxicological studies (Sanderson, 2002; Szöcs et al., 2015), differences between statistical methods may be apparent.

We first give two motivating examples showing that different methods may lead to different conclusions. Then we compare three types of statistical methods (transformation and normality assumption, GLM, non-parametric tests) using simulations.

2 Motivating examples

2.1 Count data

Brock et al. (2014) provides a typical example data from a mesocosm study of mayfly larvae counts on artificial substrate samplers at one sampling day (Figure 1). 18 mesocosms have been sampled, with 6 treatments (Control, n = 4; 0.1 mg/L, 0.3mg/L, 1mg/l, 3mg/L, n = 3; 10 mg/L, n = 2). We will use this data to demonstrate the differences between transformations, different GLMs and a non-parametric approach.

2.1.1 The linear model

To fit the standard linear model, we first transform the counts following van den Brink et al. (2000):

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

, where y_i is the measured abundance, y_i^T the transformed abundance and A = 2 / 11 = 0.182.

We fit the well known linear model:

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

This model assumes a normal distributed response with constant variance (σ^2). Note, that we it parametrised as contrast (βx_i) to the control group (α) so that the LOEC can be directly deduced from the estimates of β.

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:

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

Again, this model is parametrised as contrast to the control group. The expected values of y (λ) are linked with a log-function to the predictors, to avoid negative fitted values. The Poisson distribution assumes that the mean and the 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):

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

Another possibility to deal with overdispersion is to use a negative binomial distribution:

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

In both cases the parametrisation and link function is the same as in the Poission GLM. Note, that the quasi-Poisson model assumes a linear mean-variance relationship, whereas the negative binomial model a quadratic relationship. The above described models are most commonly used in ecology, although other distributions for count data are possible, like the negative binomial with a linear mean variance relationship (also known as NB1) or the poisson inverse gaussian (Hilbe, 2014).

2.1.3 Hypothesis testing

On this data, we could test different hypotheses like (i) if is there any effect of the treatment or (ii) test single parameters (treatments) to determine the LOEC. 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) and one-sided Wald Z tests (Poisson and negative binomial) following general recommendations (Bolker et al., 2009).

2.1.4 Results

The Poisson model showed considerable overdispersion and did not fit to the data. Therefore, inferences are not valid and we do not further discuss it's 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 ($LR = 13.99$, $p = 0.016$).

see supplement

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 determined LOECs differed (Normal and quasi-Poisson: 3 mg/L, negative binomial: 0.3 mg/L).

Brock et al. (2014) 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., 2014) 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. (2014) 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 might have tremendous impact on ecotoxicological inferences, especially

when sample sizes are low.

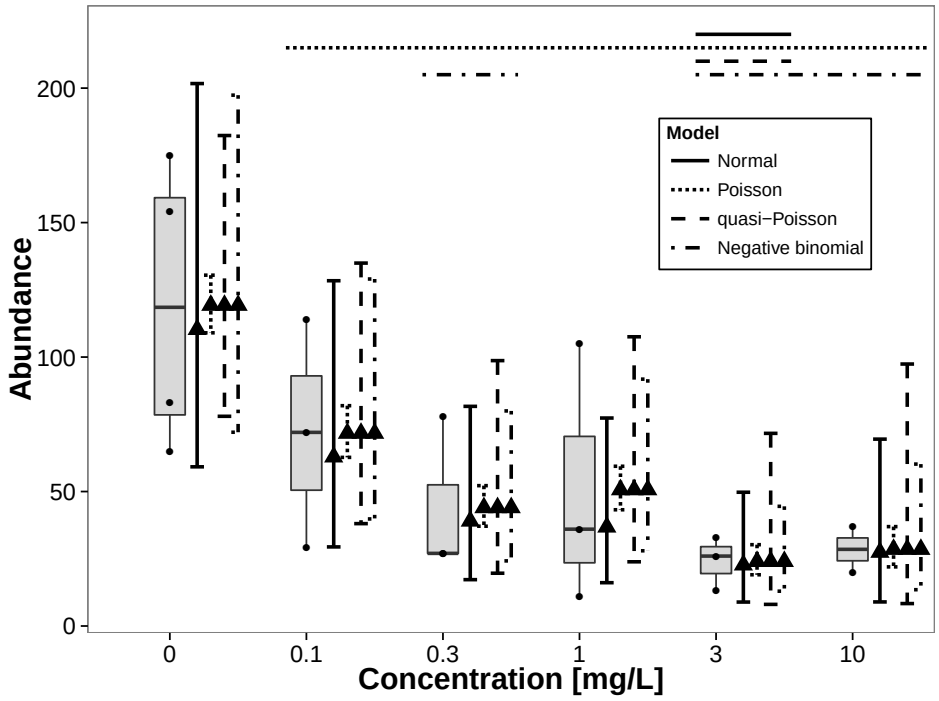

Figure 1: Data from Brock et al. (2014) (boxes + black points) and estimates + 95% Wald Z or t Confidence intervals from the fitted models (vertical lines). Bars above indicate treatments statistically significant different from the control group (Dunnett contrasts).

2.2 Binomial data

Weber et al. (1989) provides fathead minnow *Pimephales promelas* larval survival data after sodium pentachlorophenol (NaPCP) exposure. This data was also exemplary analysed in Newman (2012). 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 reported.

2.2.1 The linear model after transformation

To accommodate the assumption for the standard linear model the EPA suggests a arcsine square root transformation for such kind of data EPA (2002):

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

, where y_i^T are the transformed proportions and N is the number of exposed animals. The transformed proportions are then analysed using the standard linear model (see above).

2.2.2 Generalized Linear Models

Data of type *x out of n* are typically modelled by a binomial distribution with parameters N and π:

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

, with 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.

2.2.3 Results

For this dataset, both methods yielded to same ecotoxicological inferences: The global tests of both methods indicated a strong 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. However, the coefficients of the binomial model are directly interpretable as change in the odds ratio, whereas this 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 compare 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 example.

Table 1: Estimated parameters and Confidence Intervals for the binomial data example. Bold values indicate LOEC as determined using Dunnett contrasts and adjustment for multiple testing.

	Model			
Parameter	LM		GLM	
Control	1.331	(1.180, 1.481)	2.994	(1.523, 4.366)
32 µg/L	-0.147	(-0.360, 0.066)	-1.21	(2.876, 0.456)
64 µg/L	0.041	(0.172, 0.254)	0.719	(-1.723, 3.161)
128 µg/L	-0.076	(0.289, 0.137)	-0.747	(-2.505, 1.010)
256 µg/L	-0.221	(-0.434, -0.008)	-1.708	(-3.312, -0.104)
512 µg/L	**-0.727**	**(-0.941, -0.514)**	**-3.675**	**(-5.244, -2.107)**

All simulations were done in R (Version 3.1.2) on a 64-bit Linux machine with 8 GB and 2.2 GHz. Exemplary analysis of data in the motivating example can be found in the supplement. Source code for the simulations is available online at `https://github.com/EDiLD/usetheglm`.

git repo currently private - tidy

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: $\kappa = 3.91$). We simulated datasets with different number of replicates (N = {3, 6, 9}) and different abundances in control treatments ($\mu_{\mathrm{C}} = \{2, 4, 8, 16, 32, 64, 128\}$). For power estimation mean abundance in treatments T2 - T5 was reduced to half of control and T1 ($\mu_{\mathrm{T2}} = ... = \mu_{\mathrm{T5}} = 0.5\,\mu_{\mathrm{C}} = 0.5\,\mu_{\mathrm{T1}}$), resulting to a theoretical LOEC at T2. For Type I error estimation mean abundance was kept equal between all groups.

For each combination we generated 100 datasets. For each dataset we tested the treatment effect using 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). 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.

3.1.2 Results

For this simulation design (reduction in abundance by 50%) a sample size of n = 9 was needed to achieve a power greater then 80%. For small sample sizes (n = 3, 6) and low abundances (μ_C = 2, 4) many of the negative binomial models (GLM_{nb} and GLM_{pb}) did not converge to a solution (convergence rate <80% of the simulations).

Supplem
Conver-
gence
table.

GLM_{nb} showed an increased Type I error at low sample sizes for the test of treatment effect. However, this decreased to an acceptable limit with increasing sample sizes (Figure 2, bottom). LM, GLM_{qp} and GLM_{pb} maintained an appropriate Type I error, with GLM_{qp} having greatest power. The Kruskal-Wallis test showed least power, with low Type I error at small sample sizes. At bigger sample sizes (n = 9) all GLM had higher power than LM or the Kruskal test (Figure 2).

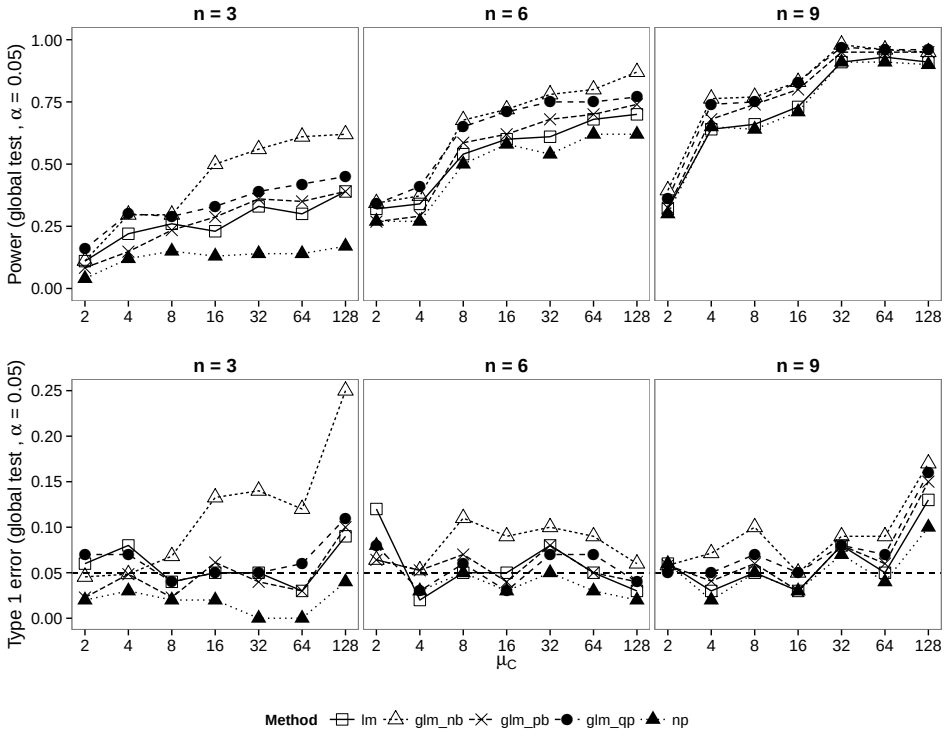

Figure 2: Simulation results for count data. Power (top) and Type I error (bottom) for the test of a treatment effect. Compared methods were: Linear model after log (Ax + 1) transformation (lm), negative binomial GLM with LRT (glm_nb), negative binomial GLM with parametric boostrap (glm_pb), quasi-Poisson GLM (glm_qp) and Kruskal-Wallis test on untransformed data (np). For n = 3 and μ_C = 2, 4 less then 80% of glm_nb and glm_pb models did converge.

The inferences on parameters showed up to 87% less power compared to the test of a general treatment effect (excluding $\mu_C = \{2, 4\}$ due to convergence problems and the pairwise wilcox test; Figures 2 and 3). The pairwise Wilcoxon Test had no power at all to detect the correct LOEC. GLM_{nb} and GLM_{pb} showed an increased Type 1 error at low sample sizes and GLM_{qp} beeing slightly conservative. GLM_{pb} and LM yielded to comparable power.

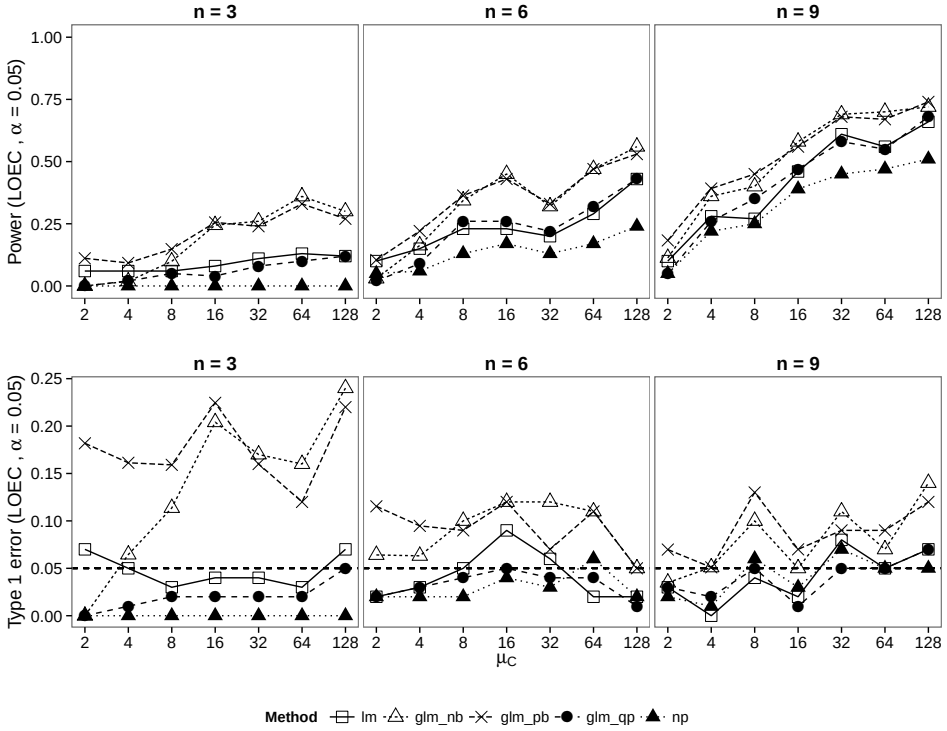

Figure 3: Simulation results for count data. Power (top) and Type I error (bottom) for determination of LOEC. Compared methods were: Linear model after log (Ax + 1) transformation (lm), negative binomial GLM with LRT (glm_nb), negative binomial GLM with parametric boostrap (glm_pb), quasi-Poisson GLM (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.

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 datasets for each combination and analysed them using the linear model 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

At low samples sizes the binomial GLM showed greatest power for testing the treatment effect, while maintaining an appropriate Type I error level. Kruskal-Wallis test had lowest power and a low Type I error rate. However, the difference between methods quickly vanished with increasing samples sizes (Figure 4).

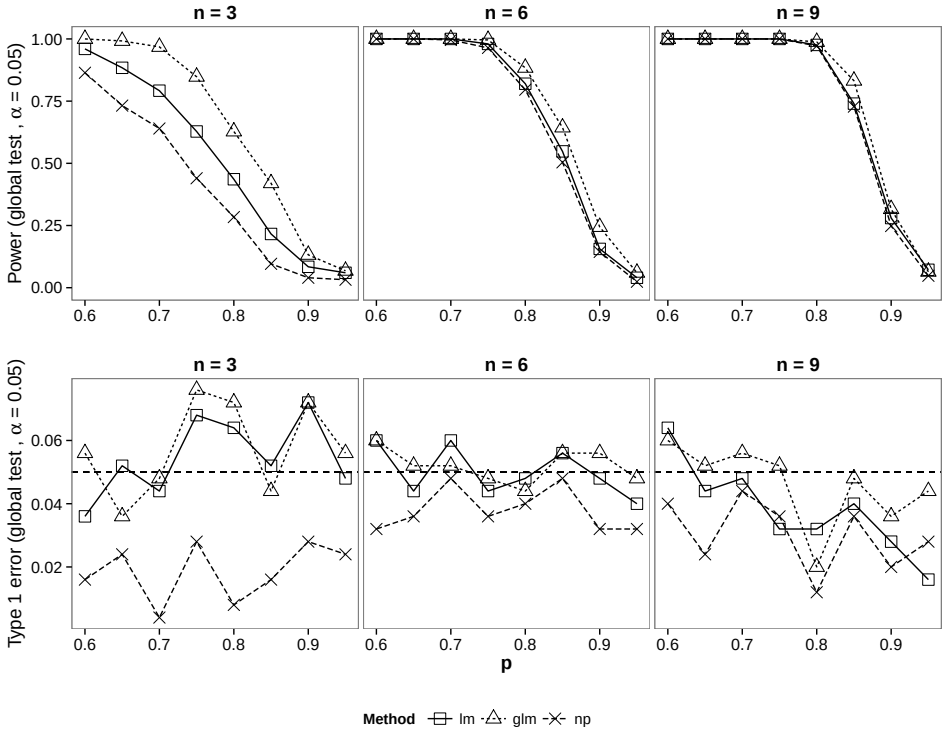

Figure 4: Simulation results for binomial data. Power (top) and Type I error (bottom) for the test of a treatment effect. Compared methods were: Linear model after arcsine square root transformation (lm), binomial GLM with LRT (glm) and Kruskal-Wallis test on untransformed data.

Inference on parameters was not as powerful as inference on the general treatment effect. For

small sample sizes LM showed highest power, while maintaing a Type 1 error level of 0.05. GLM had less power and showed a low Type 1 error rate, especially with decreasing effect size. The pairwise Wilcoxon test had no power at all for n = 3. The differences in power to detect a LOEC vanished quickly with increasing sample sizes (Figure 5).

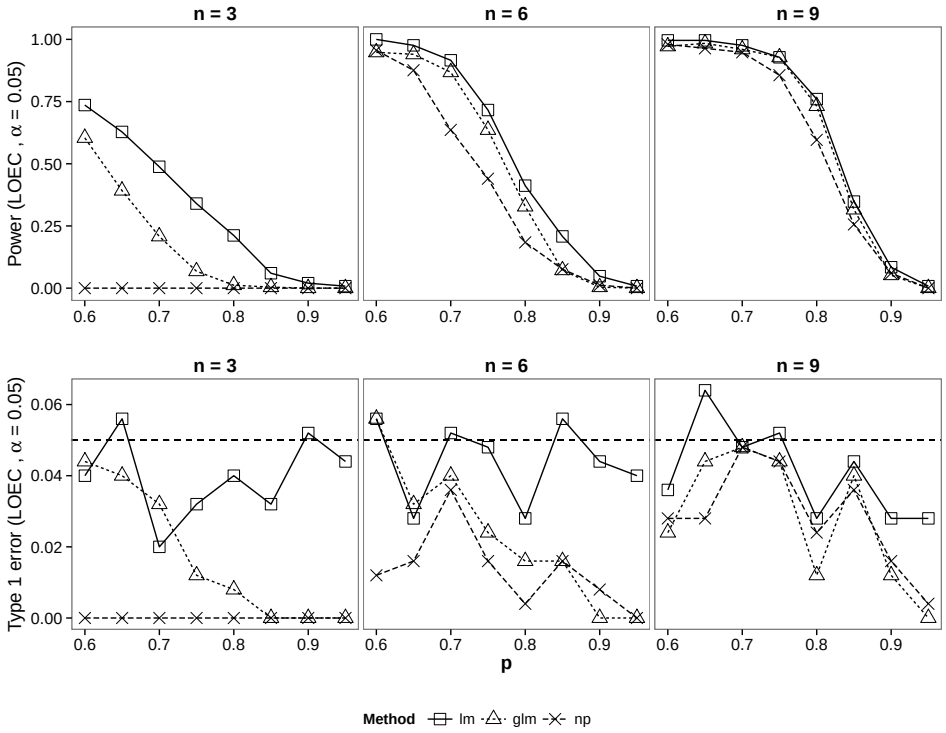

Figure 5: Simulation results for binomial data. Power (top) and Type I error (bottom) for determination of LOEC. Compared methods were: Linear model after arcsine square root transformation (lm), binomial GLM with LRT (glm) and Kruskal-Wallis test on untransformed data.

4 Discussion

Statistical hypothesis tests are commonly used to make ecotoxicological inferences. Often these inferences are based on experiments with small sample sizes due to practical contraints. Überleit

We showed for two common data types in ecotoxicology, that using an appropriate model yielded higher statistical power, than trying to meet the assumptions of normality and variance homogeneity using transformations. How should one choose

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).

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. GLM always performed better than non-parametric tests. However, there might be more powerful non-parametric tests available (Konietschke et al., 2012). Non-parametric statistics are focussing on testing, but not on estimation of effects. GLM allows the estimation, additional to testing, and the 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 instead use GLMs.

Extremely low samples sizes (n <5) are common in mesocosm experiments (Sanderson, 2002; Szöcs et al., 2015). We simulated a reduction of 50% in abundance and with such low sample sizes power to detect a treatment effect was unacceptably low (<50% for methods with appropriate Type 1 error, Figure 2). This is even worse for detecting the correct LOEC with power less than 15%. 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., 2014; EFSA PPR, 2013). Our results suggest, that

How to choose models? MV plot,...

robust S for lower T1 error in glm n

Kritik a NOEC

5 Conclusions

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. (2014). The minimum detectable difference (MDD)

and the interpretation of treatment-related effects of pesticides in experimental ecosystems. *Environmental Science and Pollution Research*.

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.

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.

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.

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*. submitted.

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.

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.