Ecotoxicology is not normal.

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

Eduard Szöcs, Ralf B. Schäfer

January 9, 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. 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 used, 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

We first give two motivating examples showing that different methods may lead to different conclusions. Then we compare these three types of 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):

$$Y_i^T = log(AY + 1)$$

, where Y is the measured abundance, Y^T the transformed abundance and A = 5.5 (the lowest abundance value in the dataset is 11).

We fit the well known linear model:

$$\begin{aligned} Y^T &\sim N(\mu, \sigma^2) \\ var(Y^T) &= \sigma^2 \\ Y^T &= \alpha + \beta X \end{aligned}$$

This model assumes a normal distributed response with constant variance (σ). Note, that we it parametrised as contrast (βX) 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:

$$\begin{aligned} Y &\sim P(\lambda) \\ log\ (\lambda) &= \mu \\ \mu &= \alpha + \beta X \\ var(Y) &= \lambda \end{aligned}$$

Again, this model is parametrised as contrast to the control group. The expected value 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):

$$\begin{aligned} Y &\sim P(\lambda, \Theta) \\ var(Y) &= \Theta\lambda \end{aligned}$$

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

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

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 models for count data are possible, like the negative binomial with a linear mean variance relationship (also known as NB1) and the poisson inverse gaussian (Hilbe, 2014).

2.1.3 Hypothesis testing

We could test different hypotheses, (i) if is there any effect of the treatment and (i) test single treatments/parameters to determine the LOEC. We used F-tests for the linear 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 Wald t tests (normal and quasi-poisson) and Wald Z tests (Poisson and negative binomial) following general recommendations (Bolker et al., 2009). Moreover, we used parametric bootstrap (we generated 2000 bootstrap samples) to assess the LR and parameters in the negative binomial model (Faraway, 2006). For comparison, we also performed the commonly used non-parametric Kruskal-Wallis test and pairwise Wilcoxon test on untransformed data. P-values from multiple comparisons where adjusted using the method of Holm (1979).

2.1.4 Results

F-test of the linear model (F = 2.57, p = 0.084) and the quasi-Poisson model (F = 2.90, p = 0.061) as well as Kruskal-Wallis test (p = 0.145) did not indicate any treatment related effects. The LR test of the negative binomial model indicated a treatment related effect (LR = 13.99, p = 0.016), whereas parametric bootstrap did not (p = 0.058). Because of overdispersion the Poisson model did not fit to the data and inferences are not valid.

All methods resulted in similar predicted values except the normal model predicting always lower abundances (Figure 1). 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: 3 mg/L, quasi-Poisson: >10mg/L, negative binomial (Wald Z and bootstrap) : 0.3 mg/L).

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 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):

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

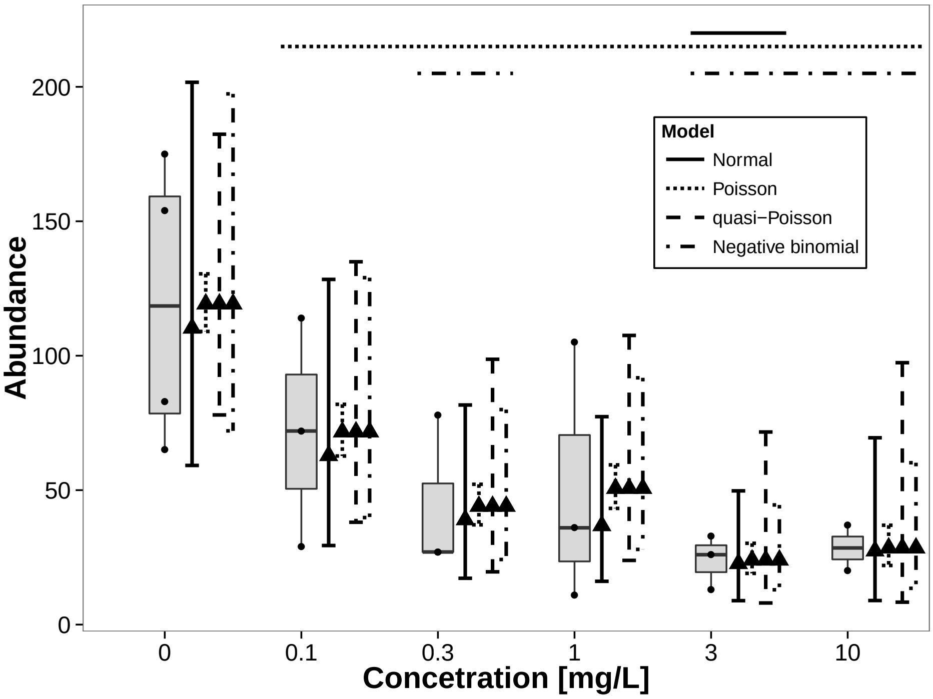

Figure 1: Data from Brock et al. (2014) (boxes + black points) and estimates + 95% Wald Z or t Confidence intervals from the fitted models (vertical lines). Bars above indicate treatments statistically significant different from the control group (Dunnett contrasts).

, where y^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 parameter n and π:

$$Y \sim Bin(n, \pi,)$$
$$logit\ (\pi) = \alpha + \beta X$$
$$var(Y) = \pi(1 - \pi)/n$$

, with n = number of exposed animals and π is the probability of survival. The variance of binomial proportion is a quadratic function of the mean.

2.2.3 Results

For this dataset, both methods yielded to same 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: $\chi^2 = 64.79$, p <0.001). Moreover, both methods identified the highest concentration (521 $\mu g/L$) as LOEC. The coefficients of the binomial model are directly interpretable as change in the odds ratio, whereas this is not possible with the transformed data.

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.

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 ($\mu_C = 2, 4$) many of the negative binomial models (GLM_{nb} and GLM_{pb}) did not converge to a solution (convergence rate <80% of the simulations).

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). The linear model on transformed data, quasi-Poisson GLM and negative binomial GLM (with bootstrap) maintained an appropriate Type I error, with quasi-Poisson having greatest power. The Kruskal-Wallis test showed least power, with low Type I error at small sample sizes. At high sample sizes (n = 9) GLM had greater power than the linear model or the Kruskal test (Figure 2).

The inferences on parameters showed less power to

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

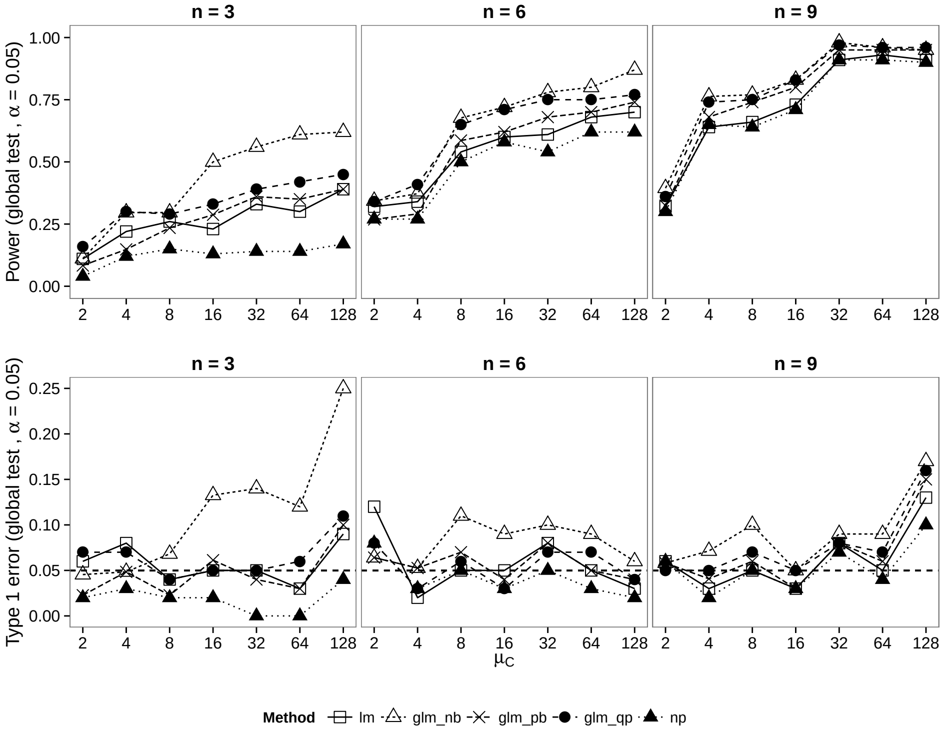

Figure 2: Simulation results for count data. Power (top) and Type I error (bottom) for the test of a treatment effect. Compared methods were: Linear model after log (Ax + 1) transformation (lm), negative binomial GLM with LRT (glm_nb), negative binomial GLM with parametric boostrap (glm_pb), quasi-Poisson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

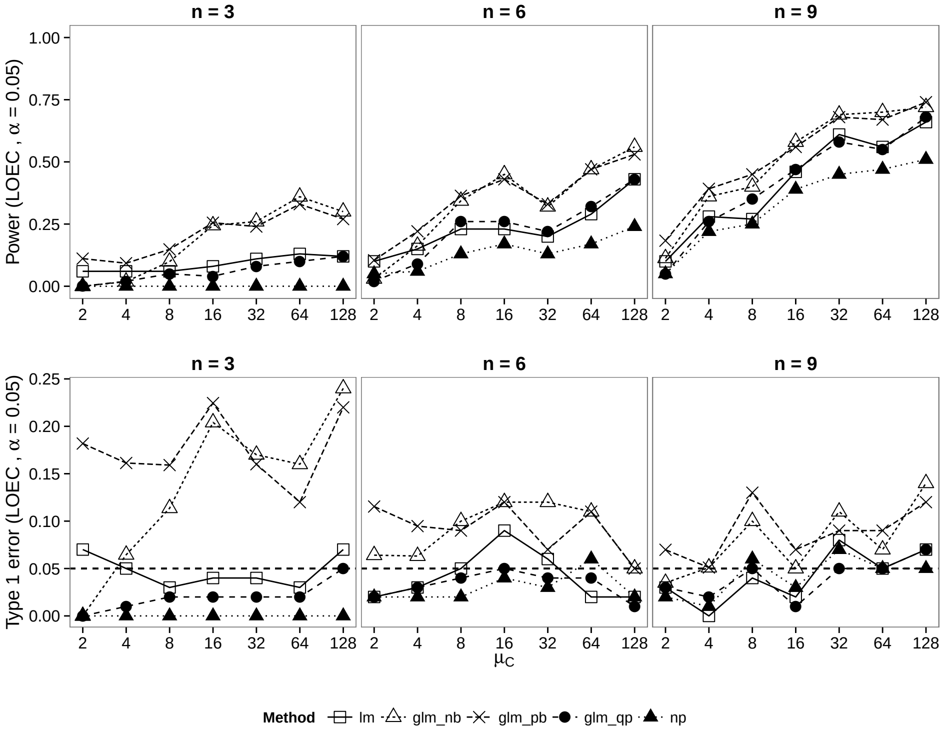

Figure 3: Simulation results for count data. Power (top) and Type I error (bottom) for determination of LOEC. Compared methods were: Linear model after log (Ax + 1) transformation (lm), negative binomial GLM with LRT (glm_nb), negative binomial GLM with parametric boostrap (glm_pb), quasi-Poisson GLM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

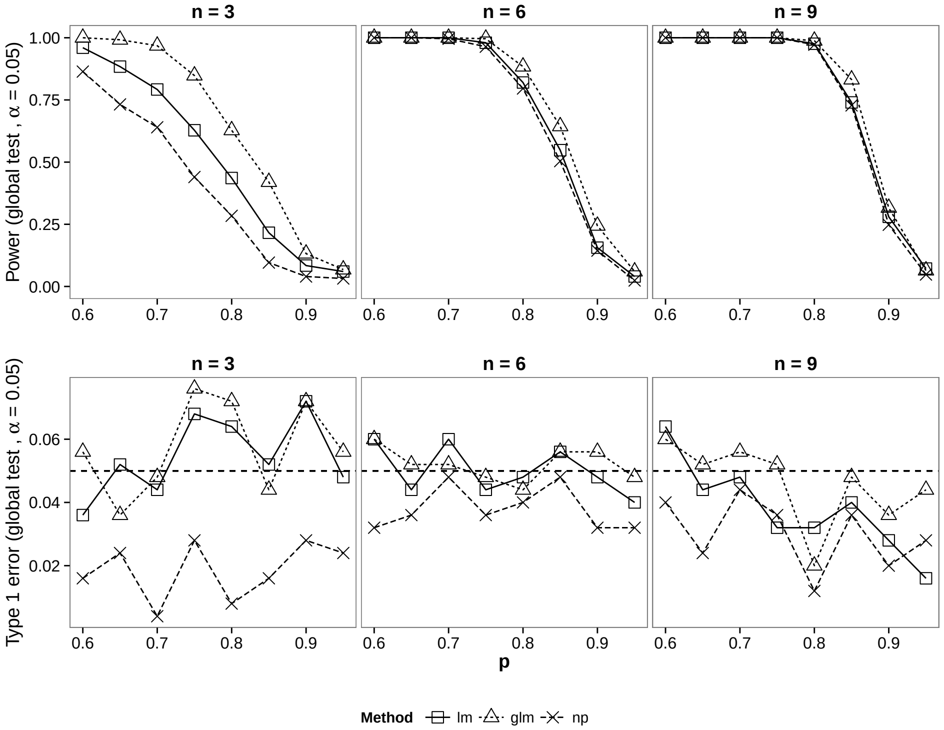

Figure 4: Simulation results for binomial data. Power (top) and Type I error (bottom) for the test of a treatment effect. Compared methods were: Linear model after arcsine square root transformation (lm), binomial GLM with LRT (glm) and Kruskal-Wallis test on untransformed data.

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, binomial 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

4 Discussion

How to choose models? MV plot,...

Multivar

GLMM

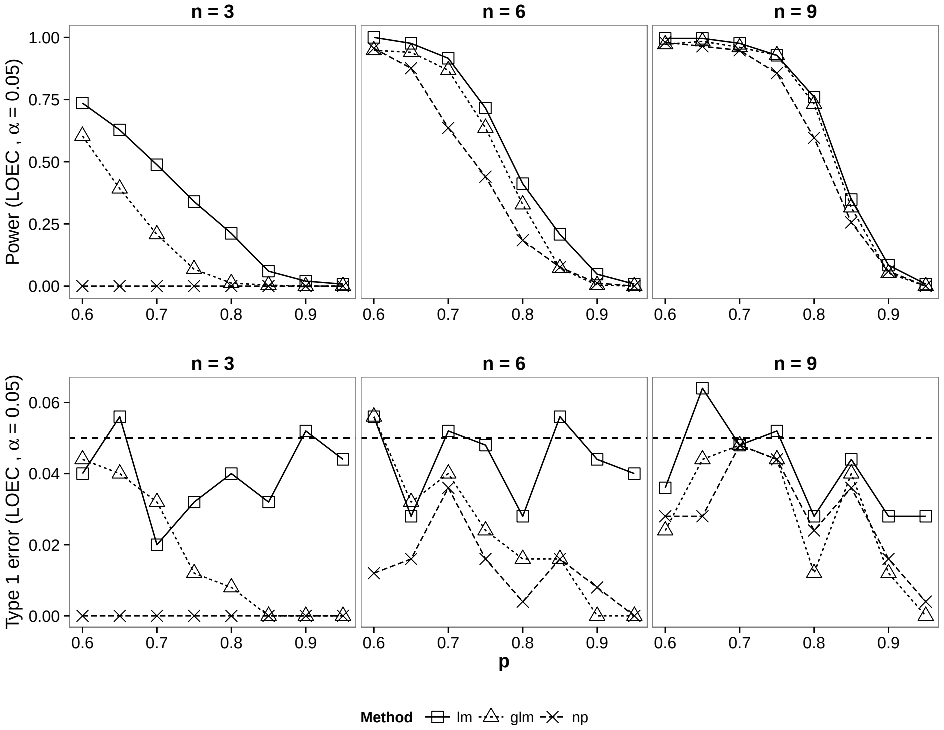

Figure 5: Simulation results for binomial data. Power (top) and Type I error (bottom) for determination of LOEC. Compared methods were: Linear model after arcsine square root transformation (lm), binomial GLM with LRT (glm) and Kruskal-Wallis test on untransformed data.

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*.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70.

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

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.

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.

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.