TITLE1

TITLE2

by

EDUARD SZÖCS

from ZĂRNESTI / ROMANIA

Submitted Dissertation thesis for the partial fulfillment of the requirements for a
Doctor of Natural Sciences
Fachbereich 7: Natur- und Umweltwissenschaften
Universität Koblenz-Landau

August 23, 2016

CONTENTS

1 ECOTOXICOLOGY IS NOT NORMAL 1
1.1 Abstract 2
1.2 Introduction 2
1.3 Methods 3
1.3.1 Models for count data 3
1.3.2 Models for binomial data 5
1.3.3 Statistical Inference 6
1.3.4 Case study 6
1.3.5 Simulations 6
1.3.6 Data Analysis 8
1.4 Results 8
1.4.1 Case study 8
1.4.2 Simulations 8
1.5 Discussion 11
1.5.1 Case study 11
1.5.2 Simulations 14
1.6 References 16
A SUPPLEMENTAL INFORMATION FOR: ECOTOXICOLOGY IS NOT NORMAL 19
A.1 Supplementary Tables 19
A.2 Worked R examples 29
A.2.1 Count data example 29
A.2.2 Binomial data example 44
A.2.3 References 50

LIST OF FIGURES

Figure 1.1 Example data from Brock et al. (2015). 7
Figure 1.2 Count data simulations: Type I error (top) and Power (bottom) for the test of a treatment effect. 9
Figure 1.3 Count data simulations: Type I error (top) and Power (bottom) for determination of LOEC. 10
Figure 1.4 Binomial data simulations: Type I error (top) and power (bottom) for the test of a treatment effect. 11
Figure 1.5 Binomial data simulations: Type I error (top) and power (bottom) for the test for determination of LOEC. 12

LIST OF TABLES

Table A.1 Count data simulations - Proportion of models converged 20
Table A.2 Count data simulations - Power to detect a treatment effect. 21
Table A.3 Count data simulations - Power to detect LOEC. 22
Table A.4 Count data simulations - Type 1 error to detect a global treatment effect. 23
Table A.5 Count data simulations - Type 1 error to detect LOEC. . . 24
Table A.6 Binomial data simulations - Power to detect a global treatment effect. 25
Table A.7 Count data simulations - Power to detect LOEC. 26
Table A.8 Binomial data simulations - Type 1 error to detect a global treatment effect. 27
Table A.9 Binomial data simulations - Type 1 error to detect LOEC. . 28

1

ECOTOXICOLOGY IS NOT NORMAL - A COMPARISON OF STATISTICAL APPROACHES FOR ANALYSIS OF COUNT AND PROPORTION DATA IN ECOTOXICOLOGY

Eduard Szöcs[a] & Ralf B. Schäfer[a]

[a]Institute for Environmental Sciences, University Koblenz-Landau, Landau, Germany

Adapted from the article published in 2015 in *Environmental Science and Pollution Research*, 22(18), 13990-13999.

1.1 ABSTRACT

Ecotoxicologists often encounter count and proportion data that are rarely normally distributed. To meet the assumptions of the linear model such data are usually transformed or non-parametric methods are used if the transformed data still violate the assumptions. Generalised Linear Models (GLM) allow to directly model such data, without the need for transformation. Here, we compare the performance of two parametric methods, i.e., (1) the linear model (assuming normality of transformed data), (2) GLMs (assuming a Poisson, negative binomial, or binomially distibuted response), 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 the different methods in terms of statistical power and Type I error for detecting a general treatment effect and determining the lowest observed effect concentration (LOEC). In addition, we outlined differences on a real world mesocosm data set.

For count data, we found that the quasi-Poisson model yielded the highest power. The negative binomial GLM resulted in increased Type I errors, which could be fixed using the parametric bootstrap. For proportions, binomial GLMs performed better than the linear model, except to determine LOEC at extremely low sample sizes. The compared non-parametric methods had generally lower power.

We recommend that counts in one-factorial experiments should be analysed using quasi-Poisson models and proportions from counts by binomial GLMs. These methods should become standard in ecotoxicology.

1.2 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, such data are often analysed using methods that assume a normal distribution and variance homogeneity (M. Wang and Riffel, 2011). To meet these assumptions data are usually transformed. For example, ecotoxicological textbooks (Michael C Newman, 2012) and guidelines (EPA, 2002; OECD, 2006) advise that survival data should 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, 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. Other transformations, like the square root or fourth root transformation, are also commonly applied in community ecology (Anderson et al., 2011). 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 of the linear model, non-parametric tests are usually applied (M. 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 for more than 40 years, ecotoxicologists do not regularly make use of them. Recent studies concluded that the linear model should not be applied on transformed data and GLMs be used as they have better statistical properties (O'Hara and Kotze 2010; Warton 2005 (counts), Warton and Hui 2011 (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 values (Lowest / No observed effect concentration). 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, when appropriately used, inference in ecotoxicological studies and compared three types of statistical methods (linear model on transformed data, 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.

1.3 METHODS

1.3.1 *Models for count data*

1.3.1.1 *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 Brink et al. (2000) and used a log(Ay + 1) transformation (eqn. 1.1):

$$Y_{new\ i} = log(AY_i + 1) \tag{1.1}$$

, where Y_i is the measured and $Y_{new\ i}$ 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):

$$\begin{aligned} Y_{new\ i} &\sim N(\mu_i, \sigma^2) \\ E(Y_{new\ i}) = \mu_i \text{ and } &var(Y_{new\ i}) = \sigma^2 \\ \mu_i &= \beta \times X_i \end{aligned} \tag{1.2}$$

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 (X_i codes the treatments) and β are the estimated coefficients related to these changes in transformed abundances between treatments (eqn. 1.2).

1.3.1.2 *Generalised Linear Models*

GLMs extend the linear model to variables that are not normally distributed. Instead of transforming the response variable, the counts could be directly modeled by a Poisson GLM (GLM_p):

$$\begin{aligned} Y_i &\sim P(\mu_i) \\ E(Y_i) &= var(Y_i) = \mu_i \\ log(\mu_i) &= \beta \times X_i \end{aligned} \tag{1.3}$$

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 modeling the mean as a function of treatment membership (X_i). 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. 1.3).

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 models the variance as a linear function of the mean (eqn. 1.4):

$$var(Y_i) = \phi \mu_i \tag{1.4}$$

Here, ϕ 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. 1.3) and than the standard errors are scaled by the degree of overdispersion (Hilbe, 2014).

Another possibility to deal with overdispersion is to model abundances by a negative binomial distribution (GLM_{nb}, eqn. 1.5):

$$\begin{aligned} Y_i &\sim NB(\mu_i, \kappa) \\ E(Y_i) = \mu_i \text{ and } &var(Y_i) = \mu_i + \mu_i^2/\kappa \\ log(\mu_i) &= \beta \times X_i \end{aligned} \tag{1.5}$$

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. 1.4), whereas the negative binomial model assumes a quadratic relationship (eqn. 1.5).

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).

1.3.2 *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.

1.3.2.1 *Linear model for transformed data*

To accommodate the assumptions for the standard linear model with such proportions, a special arcsine square root transformation (eqn. 1.6) is suggested (EPA, 2002; Michael C Newman, 2012):

$$Y_{new\ i} = \begin{cases} arcsin(1) - arcsin(\sqrt{\frac{1}{4n}}) & \text{, if } Y_i = 1 \\ arcsin(\sqrt{\frac{1}{4n}}) & \text{, if } Y_i = 0 \\ arcsin(\sqrt{Y_i}) & \text{, otherwise} \end{cases} \tag{1.6}$$

, where Y_i are the untransformed proportions, $Y_{new\ i}$ 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. 1.2). Note, that the coefficients of the linear model are not directly interpretable due to transformation.

1.3.2.2 *Generalised Linear Models*

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

$$\begin{gathered} Y_i \sim Bin(N, \pi_i) \\ E(Y_i) = \pi_i \times N \text{ and } var(Y_i) = \pi_i(1 - \pi_i)/N \\ logit\ (\pi_i) = \beta \times X_i \end{gathered} \tag{1.7}$$

This model assumes that the number of occurrences (Y_i) 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 modeling the probability of occurrence as function of treatment membership (X_i) and to ensure that $0 < \pi_i < 1$ we do this on a logit scale (eqn. 1.7). The estimated coefficients (β) of this model are directly interpretable as changes in log odds between treatments.

Non-independent trials (e.g. fish are grouped in aquaria) may lead to overdispersion (Williams, 1982). Methods to deal with overdispersed binomial data are for example 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).

1.3.3 *Statistical Inference*

After model fitting 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 the LR test is 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 LR test of 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).

1.3.4 *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 at one sampling date. A total of 18 mesocosms 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 1.1).

1.3.5 *Simulations*

1.3.5.1 *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 overdispersion at all treatments ($\kappa = 4$, eqn. 1.5). We simulated data sets with different number of

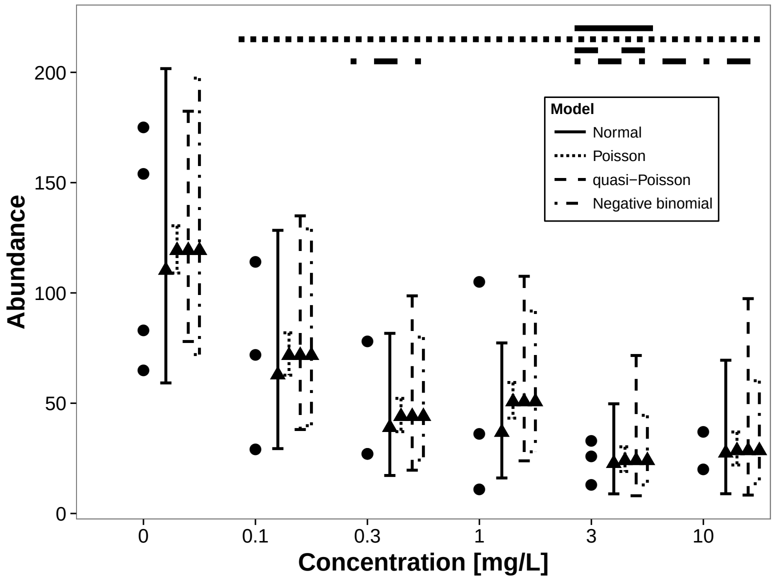


Figure 1.1.: Data from Brock et al. (2015) (dots). Predicted values (triangles) and 95% Wald Z or t confidence intervals from the fitted models (vertical lines) are given beside. Horizontal bars above indicate treatments statistically significant different from the control group (Dunnett contrasts). The data showed considerable overdispersion ($\kappa = 3.91, \phi = 22.41$) and therefore, the Poisson model underestimates the width of confidence intervals.

replicates (N = {3, 6, 9}) and different abundances in control treatments (μ_C = {2, 4, 8, 16, 32, 64, 128}). For Type I error estimation mean abundance was equal between treatments. 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. We generated 1000 data sets for each combination of N and μ_C and analysed these using the models outlined in section 1.3.1.

1.3.5.2 *Binomial data*

We simulated data from a commonly used design as described in Weber et al. (1989), with 5 treated (T1 - T5) and one control group (C). Proportions were drawn from a Bin(10, π) distribution, with varying probability of survival (π = {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 equal between treatments. For power estimation π was fixed at 0.95 in C and T1 and varied only in treatments T2 - T5. For each combination we simulated 1000 data sets and analysed these using the models outlined in section 1.3.2.

1.3.6 *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 I error (detection of an effect when there is none) and power (ability to detect an effect when it is present) at a significance level of $\alpha = 0.05$.

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).

1.4 RESULTS

1.4.1 *Case study*

The data set showed considerably higher variance then expected by the Poisson model ($\phi = 22.41$ (eqn. 1.4), $\kappa = 3.91$ (eqn. 1.5)). Therefore, the Poisson model did not fit to this data and led to underestimated standard errors and confidence intervals, as well as overestimated statistical significance (Figure 1.1). 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 1.1). 95% confidence intervals (CI) were 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.

1.4.2 *Simulations*

1.4.2.1 *Count data*

For detecting a general treatment effect, GLM_{nb} and GLM_p showed inflated Type I error rates, whereas KW was conservative at low sample sizes. However, using the parametric bootstrap for the negative binomial model (GLM_{npb}), as well as LM and GLM_{qp} resulted in appropriate Type I error rates. For detecting a treatment effect,GLM_{qp} had the highest power, followed by GLM_{npb}, LM and KW, the latter having least power (Figure 1.2). 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

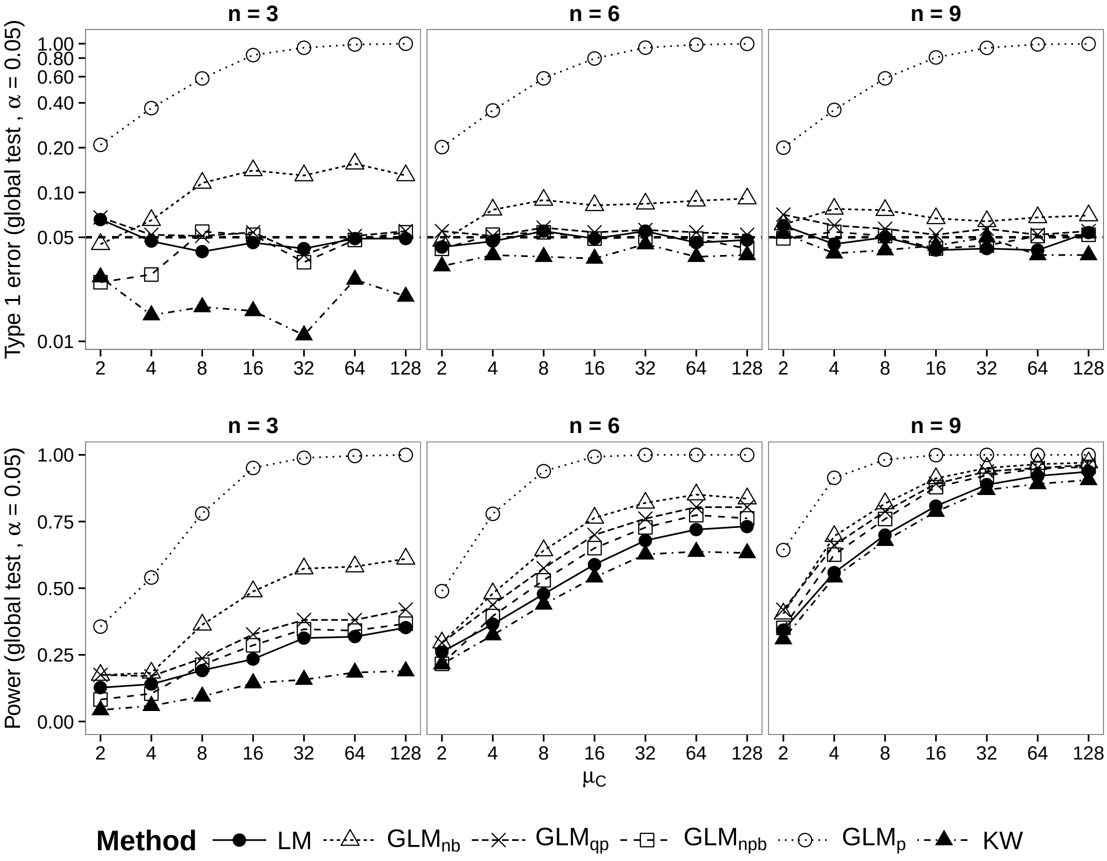


Figure 1.2.: Count data simulations: Type I error (top) and Power (bottom) for the test of a treatment effect. Type I errors are displayed on a logarithmic scale. Power levels for models with inflated Type I errors (GLM_P and GLM_{qp}) are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than 85% of GLM_{nb} and GLM_{npb} models did converge. Dashed horizontal line denotes the nominal I error rate at $\alpha = 0.05$.

converge to a solution (convergence rate <85% of the simulations, Supplement 1).

For LOEC determination GLM_{nb} and GLM_p showed an increased Type I 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 1.3).

1.4.2.2 *Binomial data*

GLM_{bin} showed slightly increased Type I 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

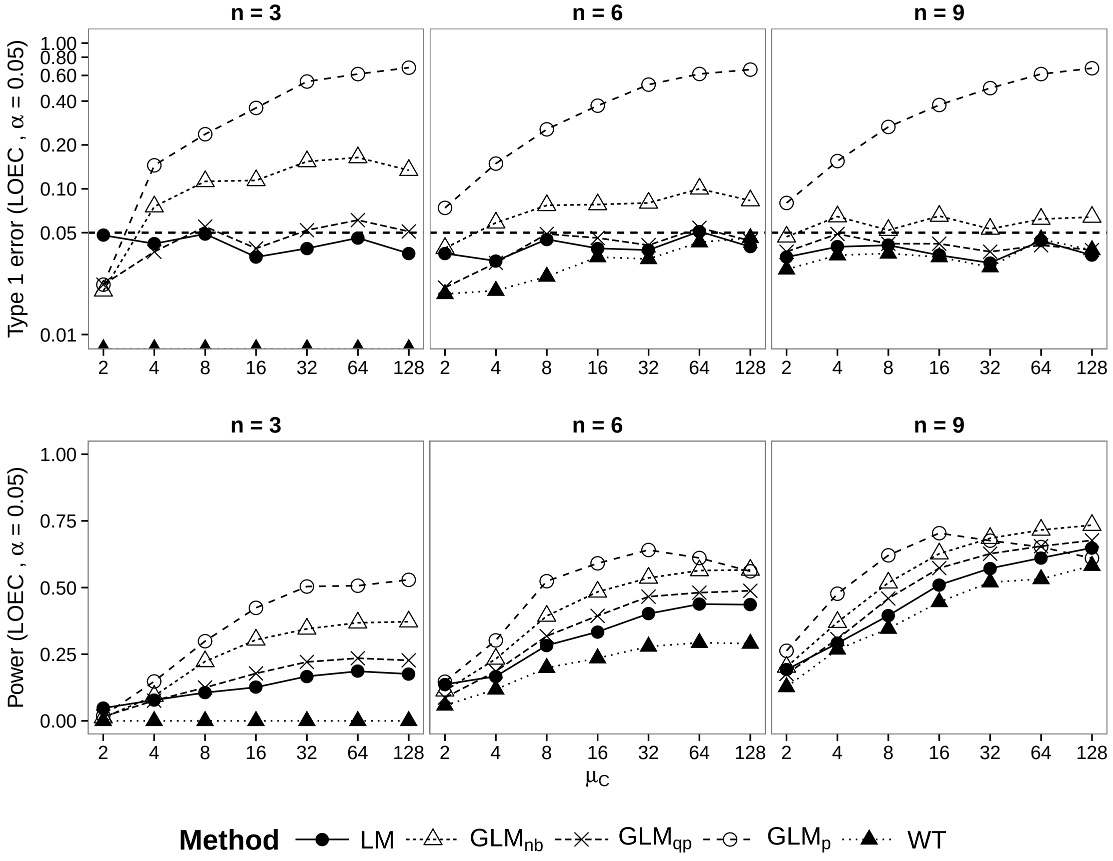


Figure 1.3.: Count data simulations: Type I error (top) and Power (bottom) for determination of LOEC. Type I errors are displayed one a logarithmic scale. Power levels for models with inflated Type I error are shown for completeness. For n = {3, 6} and μ_C = {2, 4} less than 85% of GLM_{nb} models did converge. Dashed horizontal line denotes the nominal Type I error rate at $\alpha = 0.05$.

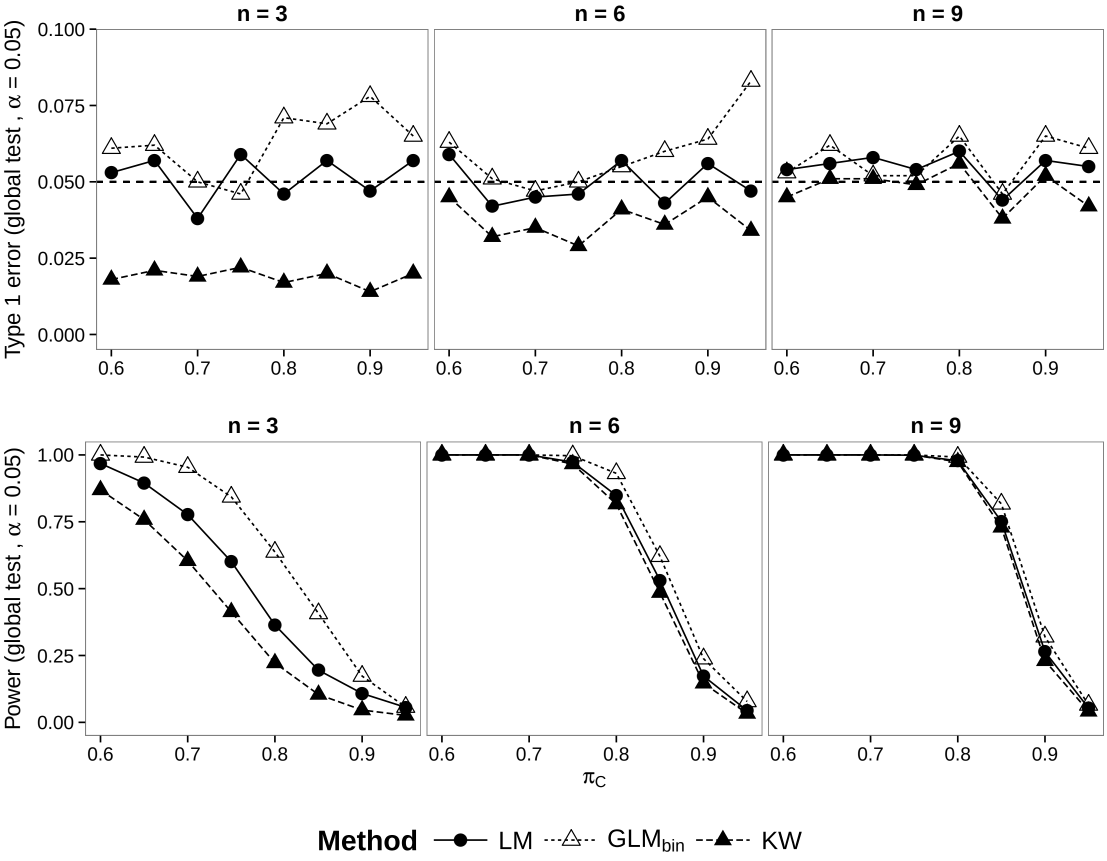


Figure 1.4.: Binomial data simulations: Type I error (top) and power (bottom) for the test of a treatment effect. Dashed horizontal line denotes the nominal Type I error rate at $\alpha = 0.05$.

compared to LM. However, the differences between methods quickly vanished with increasing samples sizes (Figure 1.4).

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 1.5).

1.5 DISCUSSION

1.5.1 *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 model estimates, their uncertainty and importance (Gelman and Stern, 2006). O'Hara and Kotze (2010)

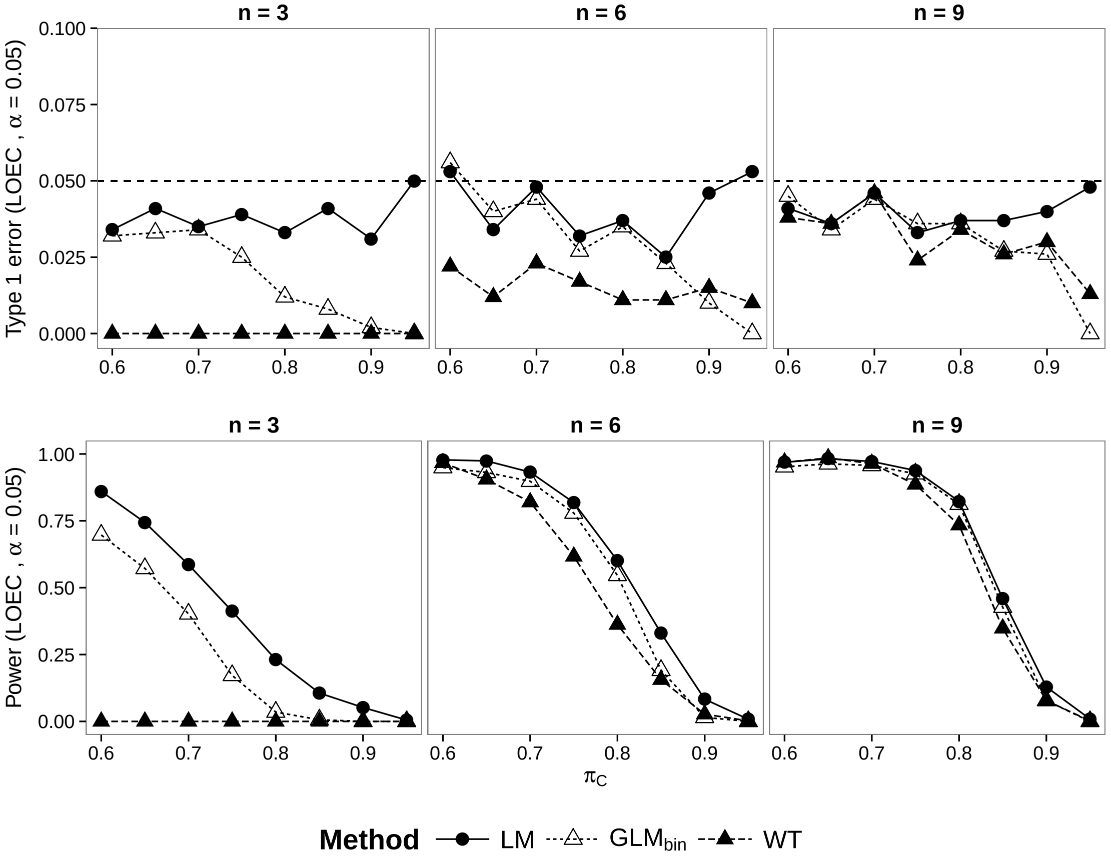


Figure 1.5.: Binomial data simulations: Type I error (top) and power (bottom) for the test for determination of LOEC. Dashed horizontal line denotes the nominal Type I error rate at $\alpha = 0.05$.

showed that the linear model on log transformed data gave unreliable and biased estimates, whereas GLMs performed well with little bias. Bias occurs also when back-transforming fitted means to the original scale, which explains the lower predicted means by LM in Figure 1.1 (Rothery, 1988) and should be corrected for (Michael C. Newman, 1993). When applied to non-transformed data, the linear model would predict identical treatment means as GLMs, because for a categorical predictor the predicted means of the LM and GLM are identical. When applied to non-transformed data, the linear model would result in identical predicted treatment means as GLMs. However, predictions would differ with continuous predictors and GLMs are particularly advantageous in this case.

This is further highlighted by the fact that for the same model (linear model applied to 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 1.1). The reasons are manifold: (i) 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 (Brink et al., 2000). (ii) We adjusted for multiple testing using Holm's (1979) method. (iii) Brock et al. (2015) used a one-sided Williams test (Williams, 1972), whereas we used one-sided comparisons to the control (Dunnett contrasts). The choice of transformation contributed only little to the differences. If the assumptions of Williams test are met it has strictly greater power than Dunnett contrasts (Jaki and L. A. Hothorn, 2013), which explains the differences in the case study. A generalisation of the Williams test as multiple contrast test (MCT) can be used in a GLM framework (T. Hothorn, Bretz, and Westfall, 2008). Nevertheless, such a Williams-type MCT is not a panacea (L. A. Hothorn, 2014) and our simulated semi-concave dose-response relationship is a situation where it fails and likely underestimates the LOEC (Kuiper, Gerhard, and L. A. Hothorn, 2014).

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 1.1). This is also shown by our simulations (Figures 1.2, 1.3) where GLM_p showed increased Type I 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 or by inspecting residual versus fitted values plots (see Supplement 2). Our simulations revealed that the LR test for GLM_{nb} is invalid because of increased Type I errors. This explains why it had the lowest p-value in the case study.

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 modeled can be chosen using knowledge about the data (e.g. bounds, integer or continuous data etc). Knowing what type of data is modeled (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.

1.5.2 *Simulations*

Our simulations showed that GLMs have generally greater power than the linear model applied to transformed data. 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 I error, Figure 1.2). 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 for inferences on the community level. For community analyses *GLM for multivariate data* (Warton, Wright, and Y. Wang, 2012) have been proposed as alternative to Principal Response Curves (PRC) and yielded similar inferences, but better indication of responsive taxa (Szöcs et al., 2015). However, Braak and Šmilauer (2015) 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, 2015).

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., 2015), 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 GLMs can enhance the power at no extra costs.

M. 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, L. A. Hothorn, and Brunner, 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 1.2). 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 1.2, bottom). This could be explained by the simpler mean-variance relationship of GLM_{qp} (eqn. 1.4 and 1.5), because at small samples sizes, low abundances or few treatment groups it is difficult to determine the mean-variance relationship. Our results are similar to Ives (2015), who compared GLMs to LM applied to transformed data for testing regression coefficients. Because of inflated Type I errors for GLM_{nb} and, in the case of multiple explanatory variables in the model, inflated Type I errors of GLM_{qp} he considered the LM on transformed data as most robust and recommended its preferred use. However, we showed that the parametric bootstrap LR test of GLM_{nb} provides appropriate Type I errors and bootstrapping might be an alternative for testing coefficients. Nevertheless, bootstrapping is computationally very intensive and we found no gains in power compared to GLM_{qp} (Figure 1.2). Given the higher power, appropriate Type I errors, stable convergence and reduced bias (O'Hara and Kotze, 2010) we suggest that count data in one factorial experiments should be analysed using the quasi-Poisson model.

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 1.5). 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.

1.6 REFERENCES

Anderson, M. J., T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. B. Kraft, J. C. Stegen, and N. G. Swenson (2011). "Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist." In: *Ecology Letters* 14.1, pp. 19–28.

Bolker, B.M, M.E Brooks, C.J Clark, S.W Geange, J.R Poulsen, M.H.H Stevens, and J.S.S White (2009). "Generalized linear mixed models: a practical guide for ecology and evolution." In: *Trends in Ecology & Evolution* 24.3, pp. 127–135.

Braak, Cajo JF ter and Petr Šmilauer (2015). "Topics in constrained and unconstrained ordination." In: *Plant Ecology* 216.5, pp. 683–696.

Brink, P. J. van den, J. Hattink, T. C. M. Brock, F. Bransen, and E. van Donk (2000). "Impact of the fungicide carbendazim in freshwater microcosms. II. Zooplankton, primary producers and final conclusions." In: *Aquatic Toxicology* 48.2-3, pp. 251–264.

Brock, T. C. M., M. Hammers-Wirtz, U. Hommen, T. G. Preuss, H-T. Ratte, I. Roessink, T. Strauss, and P. J. Van den Brink (2015). "The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems." en. In: *Environmental Science and Pollution Research* 22.2, pp. 1160–1174.

Dunnett, Charles W. (1955). "A Multiple Comparison Procedure for Comparing Several Treatments with a Control." In: *Journal of the American Statistical Association* 50.272, pp. 1096–1121.

EFSA PPR (2013). "Guidance on tiered risk assessment for plant protection products for aquatic organisms in edge-of-field surface waters." In: *EFSA Journal* 11.7, p. 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, Julian James (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. Boca Raton: Chapman & Hall.

Gelman, A. and H. Stern (2006). "The difference between "significant" and "not significant" is not itself statistically significant." In: *The American Statistician* 60.4, pp. 328–331.

Hauck, Walter W. and Allan Donner (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis." In: *Journal of the American Statistical Association* 72.360, p. 851.

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

Holm, Sture (1979). "A simple sequentially rejective multiple test procedure." In: *Scandinavian journal of statistics* 6.2, pp. 65–70.

Hothorn, Ludwig A. (2014). "Statistical evaluation of toxicological bioassays – a review." In: *Toxicol. Res.* 3.6, pp. 418–432.

Hothorn, T., F. Bretz, and P. Westfall (2008). "Simultaneous inference in general parametric models." In: *Biometrical Journal* 50.3, pp. 346–363.

Ives, Anthony R (2015). “For testing the significance of regression coefficients, go ahead and log-transform count data.” In: *Methods in Ecology and Evolution* 6.7, pp. 828–835.

Jaki, Thomas and Ludwig A. Hothorn (2013). “Statistical evaluation of toxicological assays: Dunnett or Williams test—take both.” In: *Archives of Toxicology* 87.11, pp. 1901–1910.

Johnson, Paul C. D., Sarah J. E. Barry, Heather M. Ferguson, and Pie Müller (2015). “Power analysis for generalized linear mixed models in ecology and evolution.” In: *Methods in Ecology and Evolution* 6.2, pp. 133–142.

Konietschke, Frank, Ludwig A. Hothorn, and Edgar Brunner (2012). “Rank-based multiple test procedures and simultaneous confidence intervals.” In: *Electronic Journal of Statistics* 6, pp. 738–759.

Kuiper, Rebecca M., Daniel Gerhard, and Ludwig A. Hothorn (2014). “Identification of the Minimum Effective Dose for Normally Distributed Endpoints Using a Model Selection Approach.” In: *Statistics in Biopharmaceutical Research* 6.1, pp. 55–66.

Landis, Wayne G and Peter M Chapman (2011). “Well past time to stop using NOELs and LOELs.” In: *Integrated Environmental Assessment and Management* 7.4, pp. vi–viii.

Laskowski, R. (1995). “Some good reasons to ban the use of NOEC, LOEC and related concepts in ecotoxicology.” In: *Oikos* 73.1, pp. 140–144.

Nelder, J. A. and R. W. M. Wedderburn (1972). “Generalized Linear Models.” In: *Journal of the Royal Statistical Society. Series A (General)* 135.3, pp. 370–384.

Newman, Michael C. (1993). “Regression analysis of log-transformed data: Statistical bias and its correction.” In: *Environmental Toxicology and Chemistry* 12.6, pp. 1129–1133.

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

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

O’Hara, Robert B. and D. Johan Kotze (2010). “Do not log-transform count data.” In: *Methods in Ecology and Evolution* 1.2, pp. 118–122.

Quinn, Gerry P. and Michael J. Keough (2009). *Experimental design and data analysis for biologists*. Cambridge: Cambridge Univ. Press.

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

Rothery, P. (1988). “A cautionary note on data transformation: bias in back-transformed means.” In: *Bird Study* 35.3, pp. 219–221.

Sanderson, Hans (2002). “Pesticide studies.” In: *Environmental Science and Pollution Research* 9.6, pp. 429–435.

Stroup, Walter W (2015). “Rethinking the analysis of non-normal data in plant and soil science.” In: *Agronomy Journal* 107.2, pp. 811–827.

Szöcs, Eduard, Paul J. Van den Brink, Laurent Lagadic, Thierry Caquet, Marc Roucaute, Arnaud Auber, Yannick Bayona, Matthias Liess, Peter Ebke, Alessio Ippolito, Cajo J. F. ter Braak, Theo C. M. Brock, and Ralf B. Schäfer (2015).

"Analysing chemical-induced changes in macroinvertebrate communities in aquatic mesocosm experiments: a comparison of methods." In: *Ecotoxicology* 24.4, pp. 760–769.

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

Ver Hoef, Jay M. and Peter L. Boveng (2007). "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?" In: *Ecology* 88.11, pp. 2766–2772.

Wang, M. and M. Riffel (2011). "Making the right conclusions based on wrong results and small sample sizes: interpretation of statistical tests in ecotoxicology." In: *Ecotoxicology and Environmental Safety* 74.4, pp. 684–92.

Warton, David I. (2005). "Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data." In: *Environmetrics* 16.3, pp. 275–289.

Warton, David I. and Francis K. C. Hui (2011). "The arcsine is asinine: the analysis of proportions in ecology." In: *Ecology* 92.1, pp. 3–10.

Warton, David I., Stephen T. Wright, and Yi Wang (2012). "Distance-based multivariate analyses confound location and dispersion effects." In: *Methods in Ecology and Evolution* 3.1, pp. 89–101.

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

Wilks, Samuel S. (1938). "The large-sample distribution of the likelihood ratio for testing composite hypotheses." In: *The Annals of Mathematical Statistics* 9.1, pp. 60–62.

Williams, D. A. (1972). "The comparison of several dose levels with a zero dose control." In: *Biometrics*, pp. 519–531.

Williams, D. A. (1982). "Extra-Binomial Variation in Logistic Linear Models." In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 31.2, pp. 144–148.

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

A

SUPPLEMENTAL INFORMATION FOR: ECOTOXICOLOGY IS NOT NORMAL - A COMPARISON OF STATISTICAL APPROACHES FOR ANALYSIS OF COUNT AND PROPORTION DATA IN ECOTOXICOLOGY

A.1 SUPPLEMENTARY TABLES

Table A.1.: Count data simulations - Proportion of models converged. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_p = Poisson model

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p
3.00	2.00	1.00	0.33	1.00	1.00
3.00	4.00	1.00	0.53	1.00	1.00
3.00	8.00	1.00	0.79	1.00	1.00
3.00	16.00	1.00	0.94	1.00	1.00
3.00	32.00	1.00	0.99	1.00	1.00
3.00	64.00	1.00	1.00	1.00	1.00
3.00	128.00	1.00	1.00	1.00	1.00
6.00	2.00	1.00	0.63	1.00	1.00
6.00	4.00	1.00	0.85	1.00	1.00
6.00	8.00	1.00	0.98	1.00	1.00
6.00	16.00	1.00	1.00	1.00	1.00
6.00	32.00	1.00	1.00	1.00	1.00
6.00	64.00	1.00	1.00	1.00	1.00
6.00	128.00	1.00	1.00	1.00	1.00
9.00	2.00	1.00	0.76	1.00	1.00
9.00	4.00	1.00	0.95	1.00	1.00
9.00	8.00	1.00	1.00	1.00	1.00
9.00	16.00	1.00	1.00	1.00	1.00
9.00	32.00	1.00	1.00	1.00	1.00
9.00	64.00	1.00	1.00	1.00	1.00
9.00	128.00	1.00	1.00	1.00	1.00

Table A.2.: Count data simulations - Power to detect a treatment effect. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_{qp} = Poisson model, np = pairwise Wilcoxon test.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p	np	NA
3.00	2.00	0.13	0.17	0.17	0.08	0.36	0.04
3.00	4.00	0.14	0.18	0.17	0.10	0.54	0.06
3.00	8.00	0.19	0.36	0.24	0.21	0.78	0.09
3.00	16.00	0.23	0.49	0.33	0.29	0.95	0.14
3.00	32.00	0.31	0.57	0.38	0.35	0.99	0.16
3.00	64.00	0.32	0.58	0.38	0.34	1.00	0.18
3.00	128.00	0.35	0.61	0.42	0.37	1.00	0.19
6.00	2.00	0.26	0.30	0.29	0.22	0.49	0.21
6.00	4.00	0.36	0.48	0.44	0.40	0.78	0.32
6.00	8.00	0.48	0.64	0.57	0.53	0.94	0.44
6.00	16.00	0.59	0.76	0.70	0.65	0.99	0.54
6.00	32.00	0.68	0.82	0.76	0.73	1.00	0.63
6.00	64.00	0.72	0.85	0.80	0.77	1.00	0.64
6.00	128.00	0.73	0.84	0.80	0.76	1.00	0.63
9.00	2.00	0.34	0.40	0.42	0.35	0.64	0.31
9.00	4.00	0.56	0.69	0.66	0.63	0.91	0.54
9.00	8.00	0.70	0.82	0.79	0.76	0.98	0.68
9.00	16.00	0.81	0.91	0.89	0.88	1.00	0.79
9.00	32.00	0.89	0.95	0.94	0.92	1.00	0.87
9.00	64.00	0.92	0.96	0.95	0.95	1.00	0.89
9.00	128.00	0.94	0.97	0.96	0.95	1.00	0.91

Table A.3.: Count data simulations - Power to detect LOEC. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_p = Poisson model, np = pairwise Wilcoxon test.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p	np
3.00	2.00	0.05	0.01	0.02	0.02	0.00
3.00	4.00	0.08	0.09	0.08	0.15	0.00
3.00	8.00	0.11	0.22	0.12	0.30	0.00
3.00	16.00	0.13	0.30	0.18	0.42	0.00
3.00	32.00	0.17	0.35	0.22	0.50	0.00
3.00	64.00	0.19	0.37	0.23	0.51	0.00
3.00	128.00	0.18	0.37	0.23	0.53	0.00
6.00	2.00	0.14	0.11	0.09	0.15	0.06
6.00	4.00	0.17	0.23	0.19	0.30	0.12
6.00	8.00	0.28	0.39	0.32	0.52	0.20
6.00	16.00	0.33	0.48	0.39	0.59	0.23
6.00	32.00	0.40	0.54	0.47	0.64	0.28
6.00	64.00	0.44	0.56	0.48	0.61	0.29
6.00	128.00	0.44	0.57	0.49	0.56	0.29
9.00	2.00	0.19	0.20	0.18	0.26	0.13
9.00	4.00	0.29	0.37	0.31	0.48	0.27
9.00	8.00	0.40	0.52	0.46	0.62	0.35
9.00	16.00	0.51	0.63	0.57	0.70	0.45
9.00	32.00	0.57	0.69	0.63	0.68	0.52
9.00	64.00	0.61	0.72	0.66	0.65	0.53
9.00	128.00	0.65	0.73	0.68	0.61	0.58

Table A.4.: Count data simulations - Type 1 error to detect a global treatment effect. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_{pb} = negative binomial model with parametric boostrap, GLM_p = Poisson model, np = Kruskal-Wallis test.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_{pb}	GLM_p	np
3.00	2.00	0.07	0.04	0.02	0.07	0.21	0.03
3.00	4.00	0.05	0.07	0.03	0.05	0.37	0.01
3.00	8.00	0.04	0.12	0.05	0.05	0.58	0.02
3.00	16.00	0.05	0.14	0.05	0.05	0.84	0.02
3.00	32.00	0.04	0.13	0.03	0.04	0.94	0.01
3.00	64.00	0.05	0.16	0.05	0.05	0.99	0.03
3.00	128.00	0.05	0.13	0.05	0.06	1.00	0.02
6.00	2.00	0.04	0.05	0.04	0.06	0.20	0.03
6.00	4.00	0.05	0.08	0.05	0.05	0.36	0.04
6.00	8.00	0.06	0.09	0.05	0.06	0.58	0.04
6.00	16.00	0.05	0.08	0.05	0.05	0.80	0.04
6.00	32.00	0.06	0.08	0.05	0.06	0.94	0.04
6.00	64.00	0.05	0.09	0.05	0.05	0.98	0.04
6.00	128.00	0.05	0.09	0.04	0.05	1.00	0.04
9.00	2.00	0.06	0.06	0.05	0.07	0.20	0.05
9.00	4.00	0.04	0.08	0.05	0.06	0.36	0.04
9.00	8.00	0.05	0.08	0.05	0.06	0.58	0.04
9.00	16.00	0.04	0.07	0.04	0.05	0.81	0.04
9.00	32.00	0.04	0.06	0.04	0.06	0.94	0.05
9.00	64.00	0.04	0.07	0.05	0.05	0.99	0.04
9.00	128.00	0.05	0.07	0.05	0.06	1.00	0.04

Table A.5.: Count data simulations - Type 1 error to detect LOEC. N = sample sizes, μ_C = mean abundance in control, LM = Linear model after transformation, GLM_{nb} = negative binomial model, GLM_{qp} = quasi-Poisson model, GLM_p = Poisson model, np = pairwise Wilcoxon.

N	μ_C	LM	GLM_{nb}	GLM_{qp}	GLM_p	np
3.00	2.00	0.05	0.02	0.02	0.02	0.00
3.00	4.00	0.04	0.08	0.04	0.14	0.00
3.00	8.00	0.05	0.11	0.06	0.24	0.00
3.00	16.00	0.03	0.11	0.04	0.36	0.00
3.00	32.00	0.04	0.15	0.05	0.55	0.00
3.00	64.00	0.05	0.16	0.06	0.61	0.00
3.00	128.00	0.04	0.13	0.05	0.68	0.00
6.00	2.00	0.04	0.04	0.02	0.07	0.02
6.00	4.00	0.03	0.06	0.03	0.15	0.02
6.00	8.00	0.04	0.08	0.05	0.26	0.03
6.00	16.00	0.04	0.08	0.05	0.37	0.03
6.00	32.00	0.04	0.08	0.04	0.52	0.03
6.00	64.00	0.05	0.10	0.05	0.61	0.04
6.00	128.00	0.04	0.08	0.04	0.66	0.05
9.00	2.00	0.03	0.05	0.04	0.08	0.03
9.00	4.00	0.04	0.06	0.05	0.15	0.04
9.00	8.00	0.04	0.05	0.04	0.27	0.04
9.00	16.00	0.04	0.07	0.04	0.38	0.03
9.00	32.00	0.03	0.05	0.04	0.49	0.03
9.00	64.00	0.04	0.06	0.04	0.61	0.04
9.00	128.00	0.04	0.06	0.04	0.67	0.04

Table A.6.: Binomial data simulations - Power to detect a global treatment effect. N = sample sizes, p_E = probability in effect treatments, LM = Linear model after transformation, *GLM* = binomial model, np = Kruskal-Wallis test.

N	p_E	LM	*GLM*	np
3.00	0.60	0.97	1.00	0.87
3.00	0.65	0.90	0.99	0.76
3.00	0.70	0.78	0.95	0.60
3.00	0.75	0.60	0.84	0.41
3.00	0.80	0.36	0.64	0.22
3.00	0.85	0.20	0.41	0.10
3.00	0.90	0.11	0.17	0.05
3.00	0.95	0.06	0.06	0.03
6.00	0.60	1.00	1.00	1.00
6.00	0.65	1.00	1.00	1.00
6.00	0.70	1.00	1.00	1.00
6.00	0.75	0.97	1.00	0.97
6.00	0.80	0.85	0.93	0.82
6.00	0.85	0.53	0.62	0.48
6.00	0.90	0.17	0.24	0.15
6.00	0.95	0.04	0.08	0.03
9.00	0.60	1.00	1.00	1.00
9.00	0.65	1.00	1.00	1.00
9.00	0.70	1.00	1.00	1.00
9.00	0.75	1.00	1.00	1.00
9.00	0.80	0.98	0.99	0.97
9.00	0.85	0.75	0.82	0.73
9.00	0.90	0.26	0.32	0.23
9.00	0.95	0.05	0.07	0.04

Table A.7.: Count data simulations - Power to detect LOEC. N = sample sizes, p_E = probability in effect treatments, LM = Linear model after transformation, *GLM* = binomial model, np = pairwise Wilcoxon.

N	p_E	LM	*GLM*	np
3.00	0.60	0.86	0.70	0.00
3.00	0.65	0.74	0.57	0.00
3.00	0.70	0.59	0.40	0.00
3.00	0.75	0.41	0.17	0.00
3.00	0.80	0.23	0.04	0.00
3.00	0.85	0.11	0.01	0.00
3.00	0.90	0.05	0.00	0.00
3.00	0.95	0.01	0.00	0.00
6.00	0.60	0.98	0.95	0.97
6.00	0.65	0.97	0.93	0.91
6.00	0.70	0.93	0.90	0.82
6.00	0.75	0.82	0.78	0.62
6.00	0.80	0.60	0.55	0.36
6.00	0.85	0.33	0.19	0.16
6.00	0.90	0.08	0.01	0.03
6.00	0.95	0.01	0.00	0.00
9.00	0.60	0.97	0.95	0.97
9.00	0.65	0.98	0.96	0.98
9.00	0.70	0.97	0.96	0.96
9.00	0.75	0.94	0.93	0.89
9.00	0.80	0.82	0.81	0.73
9.00	0.85	0.46	0.43	0.35
9.00	0.90	0.13	0.08	0.08
9.00	0.95	0.01	0.00	0.00

Table A.8.: Binomial data simulations - Type 1 error to detect a global treatment effect. N = sample sizes, p = probability, LM = Linear model after transformation, GLM = binomial model, np = Kruskal-Wallis test.

N	p	LM	GLM	np
3.00	0.60	0.05	0.06	0.02
3.00	0.65	0.06	0.06	0.02
3.00	0.70	0.04	0.05	0.02
3.00	0.75	0.06	0.05	0.02
3.00	0.80	0.05	0.07	0.02
3.00	0.85	0.06	0.07	0.02
3.00	0.90	0.05	0.08	0.01
3.00	0.95	0.06	0.07	0.02
6.00	0.60	0.06	0.06	0.04
6.00	0.65	0.04	0.05	0.03
6.00	0.70	0.04	0.05	0.04
6.00	0.75	0.05	0.05	0.03
6.00	0.80	0.06	0.06	0.04
6.00	0.85	0.04	0.06	0.04
6.00	0.90	0.06	0.06	0.04
6.00	0.95	0.05	0.08	0.03
9.00	0.60	0.05	0.05	0.04
9.00	0.65	0.06	0.06	0.05
9.00	0.70	0.06	0.05	0.05
9.00	0.75	0.05	0.05	0.05
9.00	0.80	0.06	0.07	0.06
9.00	0.85	0.04	0.05	0.04
9.00	0.90	0.06	0.07	0.05
9.00	0.95	0.06	0.06	0.04

Table A.9.: Binomial data simulations - Type 1 error to detect LOEC. N = sample sizes, p = probability, LM = Linear model after transformation, *GLM* = binomial model, np = pairwise Wilcoxon.

N	p_E	LM	*GLM*	np
3.00	0.60	0.03	0.03	0.00
3.00	0.65	0.04	0.03	0.00
3.00	0.70	0.04	0.03	0.00
3.00	0.75	0.04	0.03	0.00
3.00	0.80	0.03	0.01	0.00
3.00	0.85	0.04	0.01	0.00
3.00	0.90	0.03	0.00	0.00
3.00	0.95	0.05	0.00	0.00
6.00	0.60	0.05	0.06	0.02
6.00	0.65	0.03	0.04	0.01
6.00	0.70	0.05	0.04	0.02
6.00	0.75	0.03	0.03	0.02
6.00	0.80	0.04	0.04	0.01
6.00	0.85	0.03	0.02	0.01
6.00	0.90	0.05	0.01	0.01
6.00	0.95	0.05	0.00	0.01
9.00	0.60	0.04	0.04	0.04
9.00	0.65	0.04	0.03	0.04
9.00	0.70	0.05	0.04	0.05
9.00	0.75	0.03	0.04	0.02
9.00	0.80	0.04	0.04	0.03
9.00	0.85	0.04	0.03	0.03
9.00	0.90	0.04	0.03	0.03
9.00	0.95	0.05	0.00	0.01

A.2 WORKED R EXAMPLES

A.2.1 *Count data example*

A.2.1.1 *Introduction*

In this example we will analyse data from (Brock et al., 2015). The data are count of mayfly larvae in Macroinvertebrate Artificial Substrate Samplers in 18 mesocosms at one sampling day. There are 5 treatments and one control group.

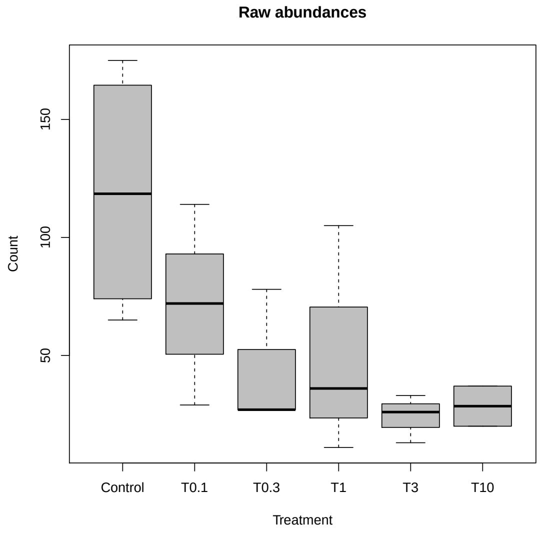
First, we load the data, bring it to the long format and remove NA values.

```
df <- read.table(header = TRUE,
                 text = ’Control  T0.1 T0.3  T1  T3  T10
                          175 29  27  36  26  20
                          65  114 78  11  13  37
                          154 72  27  105 33  NA
                          83  NA  NA  NA  NA  NA’)
require(reshape2)
dfm <- melt(df, value.name = ’abu’, variable.name = ’treatment’)
dfm <- dfm[!is.na(dfm[’abu’]), ]
head(dfm)

##   treatment abu
## 1   Control 175
## 2   Control  65
## 3   Control 154
## 4   Control  83
## 5      T0.1  29
## 6      T0.1 114
```

This results in a table with two columns - one indicating the treatment and one with the measured abundances.

Let’s have a first look at the data:

```
boxplot(abu ~ treatment, data = dfm, xlab = ’Treatment’,
        ylab = ’Count’, col = ’grey75’, main = ’Raw abundances’)
```



We clearly see a treatment related response. Moreover, we may note that variances are increasing with increasing abundances.

A.2.1.2 *Assuming a normal distribution of transformed abundances*

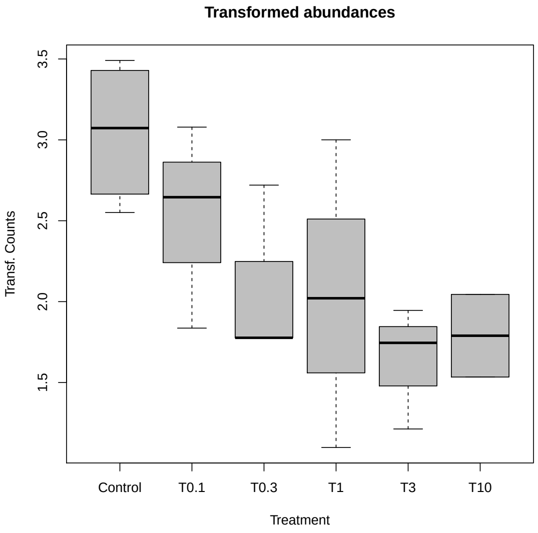
A.2.1.3 *Data transformation*

Next we transform the data using a ln(Ax + 1) transformation. A is chosen so that the term Ax equals two for the lowest non-zero abundance. We add these transformed abundances as extra column to our table.

```
A <- 2 / min(dfm$abu[dfm$abu != 0])
A

## [1] 0.1818182

dfm$abu_t <- log(A * dfm$abu + 1)
head(dfm)

##   treatment abu    abu_t
## 1   Control 175 3.490983
## 2   Control  65 2.550865
## 3   Control 154 3.367296
## 4   Control  83 2.778254
## 5      T0.1  29 1.836211
## 6      T0.1 114 3.078568
```

It looks like the transformation does a good job in equalizing the variances:

```
boxplot(abu_t ~ treatment, data = dfm,
        xlab = 'Treatment', ylab = 'Transf. Counts',
        col = 'grey75', main = 'Transformed abundances')
```



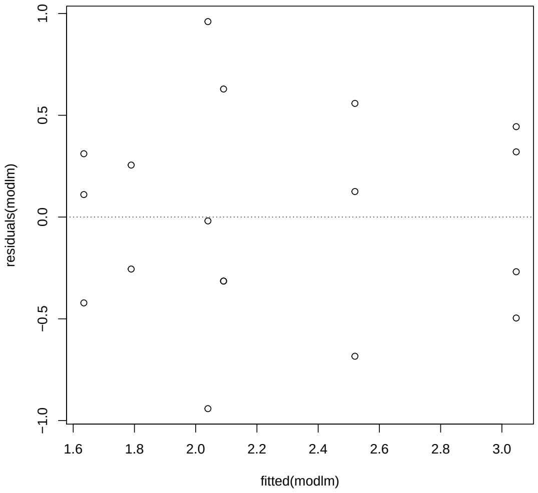
A.2.1.4 *Model fitting*

The model from eqn. 2 can be easily fitted using the `lm()` function:

```
modlm <- lm(abu_t ~ treatment, data = dfm)
```

The residuals vs. fitted values diagnostic plot show no problematic pattern, though it might be difficult to decide with such a small sample size

```
plot(residuals(modlm) ~ fitted(modlm))
abline(h = 0, lty = 'dotted')
```



The `summary()` gives the estimated coefficients with standard errors and Wald t tests:

```
summary(modlm)
```

```
##
## Call:
## lm(formula = abu_t ~ treatment, data = dfm)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94133 -0.31454  0.04576  0.31813  0.96033
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     3.0468     0.2970  10.260 2.71e-07 ***
## treatmentT0.1  -0.5267     0.4536  -1.161  0.26814
## treatmentT0.3  -0.9558     0.4536  -2.107  0.05682 .
## treatmentT1    -1.0069     0.4536  -2.220  0.04646 *
## treatmentT3    -1.4121     0.4536  -3.113  0.00897 **
## treatmentT10   -1.2575     0.5144  -2.445  0.03089 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.5939 on 12 degrees of freedom
## Multiple R-squared:  0.5167,Adjusted R-squared:  0.3154
## F-statistic: 2.566 on 5 and 12 DF,  p-value: 0.08406
```

A.2.1.5 *Inference on general treatment effect*

Or, if you want to have the ANOVA table with an F-test:

```
summary.aov(modlm)

##             Df Sum Sq Mean Sq F value Pr(>F)
## treatment    5  4.526  0.9052   2.566 0.0841 .
## Residuals   12  4.233  0.3528
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

From this output we might infer that we cannot detect any treatment effect (F = 2.566, p = 0.084).

A.2.1.6 *Inference on LOEC*

Let’s move on to the LOEC determination. This can be easily done using the multcomp package (T. Hothorn, F. Bretz, and P. Westfall, 2008):

Here we perform a one-sided (`alternative = ’less’`) using Dunnett contrasts of treatment (`mcp(treatment=’Dunnett’)`). Moreover, we adjust for multiple testing using Holm’s method (`test = adjusted(’holm’)`):

```
require(multcomp)
summary(glht(modlm, linfct = mcp(treatment = 'Dunnett'),  alternative = 'less'),
        test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = abu_t ~ treatment, data = dfm)
##
## Linear Hypotheses:
##                      Estimate Std. Error t value Pr(<t)
## T0.1 - Control >= 0  -0.5267     0.4536  -1.161 0.1341
## T0.3 - Control >= 0  -0.9558     0.4536  -2.107 0.0697 .
## T1 - Control >= 0    -1.0069     0.4536  -2.220 0.0697 .
## T3 - Control >= 0    -1.4121     0.4536  -3.113 0.0224 *
## T10 - Control >= 0   -1.2575     0.5144  -2.445 0.0618 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

Here only treatment 3 mg/L shows a statistically significant difference from control and is the determined LOEC. The column 'Estimate' gives the estimated difference in means between treatments and control and 'Std. Error' the standard errors of these estimates.

To determine the LOEC we could also use a Williams type contrast (Frank Bretz, Torsten Hothorn, and P. H. Westfall, 2010).

Here I use a step-up Williams contrast. First we need to define a contrast matrix (see also ?contrMat()):

```
# observations per treatment
n <- tapply(dfm$abu_t, dfm$treatment, length)
k <- length(n)
CM <- c()
for (i in 1:(k - 1)) {
  help <- c(-1, n[2:(i + 1)] / sum(n[2:(i + 1)]), rep(0 , k - i - 1))
  CM <- rbind(CM, help)
}
rownames(CM) <- paste("C", 1:nrow(CM))
CM

##              T0.1
## C 1 -1 1.0000000 0.0000000 0.0000000 0.0000000 0.0000000
## C 2 -1 0.5000000 0.5000000 0.0000000 0.0000000 0.0000000
## C 3 -1 0.3333333 0.3333333 0.3333333 0.0000000 0.0000000
## C 4 -1 0.2500000 0.2500000 0.2500000 0.2500000 0.0000000
## C 5 -1 0.2142857 0.2142857 0.2142857 0.2142857 0.1428571
```

Then we supply this contrast matrix to `glht()`:

```
summary(glht(modlm, linfct = mcp(treatment = CM),
             alternative = ’less’),
        test = adjusted(’holm’))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: lm(formula = abu_t ˜ treatment, data = dfm)
##
## Linear Hypotheses:
##          Estimate Std. Error t value Pr(<t)
## C 1 >= 0  -0.5267     0.4536  -1.161 0.1341
## C 2 >= 0  -0.7413     0.3834  -1.934 0.0771 .
## C 3 >= 0  -0.8298     0.3569  -2.325 0.0576 .
## C 4 >= 0  -0.9754     0.3429  -2.845 0.0295 *
## C 5 >= 0  -1.0157     0.3367  -3.016 0.0268 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- holm method)
```

This indicates a LOEC at 3 mg/L.

If we do not adjust for multiple testing (`test = adjusted(’none’)`), we end up with the same NOEC (0.1 mg/L) as Brock et al. (2015):

```
summary(glht(modlm, linfct = mcp(treatment = CM),
             alternative = ’less’),
        test = adjusted(’none’))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: User-defined Contrasts
##
##
## Fit: lm(formula = abu_t ˜ treatment, data = dfm)
##
## Linear Hypotheses:
```

```
##           Estimate Std. Error t value  Pr(<t)
## C 1 >= 0  -0.5267     0.4536  -1.161 0.13407
## C 2 >= 0  -0.7413     0.3834  -1.934 0.03855 *
## C 3 >= 0  -0.8298     0.3569  -2.325 0.01921 *
## C 4 >= 0  -0.9754     0.3429  -2.845 0.00739 **
## C 5 >= 0  -1.0157     0.3367  -3.016 0.00537 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- none method)
```

Note, this multiple contrast test is different from the original Williams test (Williams, 1972) used by (Brock et al., 2015). See Frank Bretz (1999) for a comparison.

A.2.1.7 *Assuming a Poisson distribution of abundances*

A.2.1.8 *Model fitting*

We are dealing with count data, so a Poisson GLM might be a good choice. GLMs can be fitted using the `glm()` function and here we fit the model from eqn. 3:

```
modpois <- glm(abu ~ treatment, data = dfm, family = poisson(link = ’log’))
```

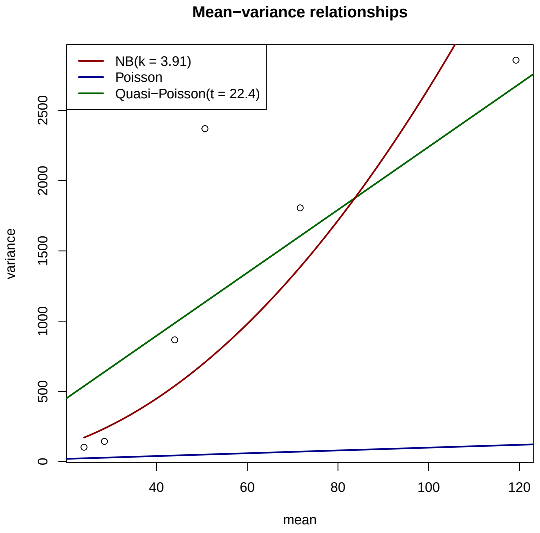
Here `family = poisson(link = ’log’)` specifies that we want to fit a poisson model using a log link between response and predictors.

The `summary` gives the estimated coefficients, standard errors and Wald Z tests:

```
(sum_pois <- summary(modpois))

##
## Call:
## glm(formula = abu ~ treatment, family = poisson(link = "log"),
##     data = dfm)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7625  -2.7621  -0.8219   2.7172   6.6602
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     4.78122    0.04579 104.423  < 2e-16 ***
## treatmentT0.1 -0.50920    0.08214  -6.199 5.69e-10 ***
## treatmentT0.3 -0.99703    0.09835 -10.138  < 2e-16 ***
## treatmentT1   -0.85595    0.09314  -9.190  < 2e-16 ***
## treatmentT3   -1.60317    0.12643 -12.680  < 2e-16 ***
## treatmentT10  -1.43132    0.14014 -10.213  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 604.79  on 17  degrees of freedom
## Residual deviance: 273.77  on 12  degrees of freedom
## AIC: 387.63
##
## Number of Fisher Scoring iterations: 5
```

But is a poisson distribution appropriate here? A property of the poisson distribution is that its variance is equal to the mean. A simple diagnostic would be to plot group variances vs. group means:

```
require(plyr)
# mean and variance per treatment
musd <- ddply(dfm, .(treatment), summarise,
              mu = mean(abu),
              var = var(abu))
musd

##   treatment        mu      var
## 1   Control 119.25000 2857.583
## 2      T0.1  71.66667 1806.333
## 3      T0.3  44.00000  867.000
## 4        T1  50.66667 2370.333
## 5        T3  24.00000  103.000
## 6       T10  28.50000  144.500

# plot mean vs var
plot(var ~ mu, data = musd,
     xlab = ’mean’, ylab = ’variance’, main = ’Mean-variance relationships’)
# poisson
abline(a = 0, b = 1, col = ’darkblue’, lwd = 2)
# quasi-Poisson
abline(a = 0, b = 22.41, col = ’darkgreen’, lwd = 2)
# negative binomial
curve(x + (x^2 / 3.91), from = 24, to = 119.25, add = TRUE,
      col = ’darkred’, lwd = 2)
legend(’topleft’,
       legend = c(’NB(k = 3.91)’, ’Poisson’, ’Quasi-Poisson(t = 22.4)’),
       col = c(’darkred’, ’darkblue’, ’darkgreen’),
       lty = c(1,1, 1),
       lwd = c(2,2, 2))
```



I also added the assumed mean-variance relationships of the Poisson, quasi-Poisson and negative binomial models (see below). We clearly see that the variance increases much more than would be expected under the poisson distribution (the data is overdispersed). Moreover, we could check overdispersion from the `summary`: If the ratio of residual deviance to degrees of freedom is >1 the data is overdispersed.

```
sum_pois$deviance / sum_pois$df.residual

## [1] 22.81412
```

A.2.1.9 *Apply quasi-Poisson to deal with overdispersion*

The plot above suggests that the variance may increasing stronger than the mean and a quasi-Poisson or negative binomial model might be more appropriate for this data.

A.2.1.10 *Model fitting*

Fitting a quasi-Poisson model (eqn. 4) is straight forward:

```
modqpois <- glm(abu ~ treatment, data = dfm, family = 'quasipoisson')
```

The summary gives the estimated coefficients:

```
summary(modqpois)

##
## Call:
## glm(formula = abu ~ treatment, family = "quasipoisson", data = dfm)
##
## Deviance Residuals:
```

```
##     Min       1Q   Median       3Q      Max
## -6.7625  -2.7621  -0.8219   2.7172   6.6602
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.7812     0.2168  22.058 4.43e-11 ***
## treatmentT0.1  -0.5092     0.3889  -1.309   0.2149
## treatmentT0.3  -0.9970     0.4656  -2.142   0.0534 .
## treatmentT1    -0.8560     0.4409  -1.941   0.0761 .
## treatmentT3    -1.6032     0.5985  -2.679   0.0201 *
## treatmentT10   -1.4313     0.6634  -2.157   0.0519 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for quasipoisson family taken to be 22.41055)
##
##     Null deviance: 604.79  on 17  degrees of freedom
## Residual deviance: 273.77  on 12  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

, with the dispersion parameter $\Theta = 22.41055$. Note, that the coefficients estimates are the same as from the Poisson model, only the standard errors are scaled/wider.

A.2.1.11 *Inference on general treatment effect*

An F-test can be performed using `drop1()`:

```
drop1(modqpois, test = ’F’)

## Single term deletions
##
## Model:
## abu ~ treatment
##           Df Deviance F value  Pr(>F)
## <none>         273.77
## treatment  5   604.79  2.9019 0.06059 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Here we would reject that there is treatment effect (at alpha = 0.05).

A.2.1.12 *Inference on LOEC*

The LOEC can be determined with `multcomp`:

```
summary(glht(modqpois, linfct = mcp(treatment = ’Dunnett’),
             alternative = ’less’),
        test = adjusted(’holm’))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: glm(formula = abu ~ treatment, family = "quasipoisson", data = dfm)
##
## Linear Hypotheses:
##                      Estimate Std. Error z value Pr(<z)
## T0.1 - Control >= 0  -0.5092     0.3889  -1.309 0.0952 .
## T0.3 - Control >= 0  -0.9970     0.4656  -2.142 0.0619 .
## T1 - Control >= 0    -0.8560     0.4409  -1.941 0.0619 .
## T3 - Control >= 0    -1.6032     0.5985  -2.679 0.0185 *
## T10 - Control >= 0   -1.4313     0.6634  -2.157 0.0619 .
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- holm method)
```

, which determines 3 mg/L as LOEC.

A.2.1.13 *Assuming a negative binomial distribution of abundances*

A.2.1.14 *Model fitting*

To fit a negative binomial GLM (eqn. 5) we could use `glm.nb()` from the MASS package (Venables and Ripley, 2002):

```
require(MASS)
modnb <- glm.nb(abu ~ treatment, data = dfm)
```

The estimated coefficients:

```
summary(modnb)

##
## Call:
## glm.nb(formula = abu ~ treatment, data = dfm, init.theta = 3.905898474,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2554  -0.8488  -0.3020   0.5954   1.5899
```

```
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)      4.7812     0.2571  18.596  < 2e-16 ***
## treatmentT0.1   -0.5092     0.3951  -1.289  0.19746
## treatmentT0.3   -0.9970     0.3988  -2.500  0.01241 *
## treatmentT1     -0.8560     0.3975  -2.153  0.03130 *
## treatmentT3     -1.6032     0.4066  -3.943 8.05e-05 ***
## treatmentT10    -1.4313     0.4601  -3.111  0.00186 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(3.9059) family taken to be 1)
##
##     Null deviance: 39.057  on 17  degrees of freedom
## Residual deviance: 18.611  on 12  degrees of freedom
## AIC: 181.24
##
## Number of Fisher Scoring iterations: 1
##
##
##                Theta:  3.91
##            Std. Err.:  1.37
##
##  2 x log-likelihood:  -167.238
```

, with $\kappa = 3.91$.

A.2.1.15 *Inference on general treatment effect (LR-test)*

For an LR-Test we need to first fit a reduced model:

```
modnb.null <- glm.nb(abu ~ 1, data = dfm)
```

, so that the dispersion parameter κ is re-estimated for the reduced model. Then we can compare these two models with a LR-Test:

```
anova(modnb, modnb.null, test = 'Chisq')

## Likelihood ratio tests of Negative Binomial Models
##
## Response: abu
##       Model    theta Resid. df    2 x log-lik.   Test    df LR stat.
## 1         1 1.861577        17       -181.2281
## 2 treatment 3.905898        12       -167.2383 1 vs 2     5 13.98985
##    Pr(Chi)
## 1
## 2 0.015674
```

, which suggests a treatment related effect on abundances.

A.2.1.16 *Inference on general treatment effect (parametric bootstrap)*

To test the LR statistic using paramtric bootstrap, we use two custom functions:

The first function `myPBrefdist` generates a boostrap sample and return the LR statistic for this sample:

```
#’ PB of LR statistic
#’ @param m1 Full model
#’ @param m0 reduced model
#’ @param  data data used in the models
#’ @return LR of boostrap
# generate reference distribution
myPBrefdist <- function(m1, m0, data){
  # simulate from null
  x0 <- simulate(m0)
  # refit with new data
  newdata0 <- data
  newdata0[ , as.character(formula(m0)[[2]])] <- x0
  m1r <-  try(update(m1, .~., data = newdata0), silent = TRUE)
  m0r <- try(update(m0, .~., data = newdata0), silent = TRUE)
  # check convergence (otherwise return NA for LR)
  if(inherits(m0r, "try-error") | inherits(m1r, "try-error")){
    LR <- ’convergence error’
  } else {
    if(!is.null(m0r[[’th.warn’]]) | !is.null(m1r[[’th.warn’]])){
      LR <- ’convergence error’
    } else {
      LR <- -2 * (logLik(m0r) - logLik(m1r))
    }
  }
  return(LR)
}
```

The second one (`myPBmodcomp`) repeats `myPBrefdist` many time and returns a p-value:

```
#’ generate LR distribution and return p value
#’ @param m1 Full model
#’ @param m0 reduced model
#’ @param data data used in m1 and m0
#’ @param npb number of bootstrap samples
#’ @return p-value of boostrapped LR values
myPBmodcomp <- function(m1, m0, data, npb){
  ## calculate reference distribution
  LR <- replicate(npb, myPBrefdist(m1 = m1, m0 = m0, data = data),
                  simplify = TRUE)
  LR <- as.numeric(LR)
  nconv_LR <- sum(!is.na(LR))
  ## original stats
  LRo <- c(-2 * (logLik(m0) - logLik(m1)))
  ## p-value from parametric bootstrap
  p.pb <- mean(c(LR, LRo) >= LRo, na.rm = TRUE)
  return(list(nconv_LR = nconv_LR, p.pb = p.pb))
}
```

Sounds complicated, but we can easily apply this to the negativ binomial model using:

```
set.seed(1234)
myPBmodcomp(modnb, modnb.null, data = dfm, npb = 500)

## $nconv_LR
## [1] 499
##
## $p.pb
## [1] 0.042
```

Here, we specify to generate 500 bootstrap samples (`npb = 500`). Of these 500 samples, 499 converged (`nconv_LR`) (one did not and throws some errors) and gives a p-value of 0.042.

A.2.1.17 *Inference on LOEC*

This is similar to the other parametric models:

```
summary(glht(modnb, linfct = mcp(treatment = 'Dunnett'),  alternative = 'less'),
        test = adjusted('holm'))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: glm.nb(formula = abu ~ treatment, data = dfm, init.theta = 3.905898474,
##     link = log)
##
## Linear Hypotheses:
##                      Estimate Std. Error z value   Pr(<z)
## T0.1 - Control >= 0  -0.5092     0.3951  -1.289 0.098731 .
## T0.3 - Control >= 0  -0.9970     0.3988  -2.500 0.018615 *
## T1 - Control >= 0    -0.8560     0.3975  -2.153 0.031300 *
## T3 - Control >= 0    -1.6032     0.4066  -3.943 0.000201 ***
## T10 - Control >= 0   -1.4313     0.4601  -3.111 0.003727 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- holm method)
```

which suggests a LOEC at the 0.3 mg/l treatment.

A.2.1.18 *Non-parametric methods*

A.2.1.19 *Kruskal-Wallis Test*

We can use the Kruskal-Wallies test to check if there is a difference between treatments:

```
kruskal.test(abu ~ treatment, data = dfm)

##
##  Kruskal-Wallis rank sum test
##
## data:  abu by treatment
## Kruskal-Wallis chi-squared = 8.219, df = 5, p-value = 0.1446
```

A.2.1.20 *Pairwise Wilcoxon test*

To determine the LOEC we could use a Pairwise Wilcoxon test. The built-in `pairwise.wilcox.test()` compares by default all levels (Tukey-contrasts). We are only interested in a subset of these comparisons (Dunnett-contrast).

Therefore, we use a custom function, which is a wrapper around `wilcox.exact()` from the exactRankTests package:

```
#' pairwise wilcox.test with dunnett contrasrs
#' @param y numeric; vector of data values
#' @param g factor; grouping vector
#' @param dunnett logical; if TRUE dunnett contrast, otherwise Tukey-contrasts
#' @param padj character; method for p-adjustment, see ?p.adjust.
#' @param ... other arguments passed to wilcox.exact {exactRankTests}
pairwise_wilcox <- function(y, g, dunnett = TRUE, padj = 'holm', ...){
  if(!require(exactRankTests)){
    stop('Install exactRankTests package')
  }
  tc <- t(combn(nlevels(g), 2))
  # take only dunnett comparisons
  if(dunnett){
    tc <- tc[tc[ , 1] == 1, ]
  }
  pval <- numeric(nrow(tc))
  # use wilcox.exact (for tied data)
  for(i in seq_len(nrow(tc))){
    pval[i] <- wilcox.exact(y[as.numeric(g) == tc[i, 2]],
                            y[as.numeric(g) == tc[i, 1]], exact = TRUE,
                            ...)$p.value
  }

  # adjust p-values
  pval <- p.adjust(pval, padj)
  names(pval) = paste(levels(g)[tc[,1]], levels(g)[tc[,2]], sep = ’ vs. ’)
  return(pval)
}
```

Here, we use one-sided Dunnett contrasts and adjust p-values using Holm’s method:

```
pairwise_wilcox(y = dfm$abu, g = dfm$treatment,
                dunnett = TRUE, p.adj = ’holm’, alternative = ’less’)

## Control vs. T0.1 Control vs. T0.3   Control vs. T1   Control vs. T3
##        0.2285714        0.2285714        0.2285714        0.1428571
##  Control vs. T10
##        0.2285714
```

This indicates no treatment effect at no level of concentration.

A.2.2 *Binomial data example*

A.2.2.1 *Introduction*

Here we will show how to analyse binomial data (*x out of n*). Data is provided in Newman (2012) (example 5.1, page 223) and EPA (2002). Ten fathead minnow (*Pimephales promelas*) larvals were exposed to sodium pentachlorophenol (NaPCP) and proportions of the total number alive at the end of the exposure reported.

First we load the data:

```
df <- read.table(header = TRUE, text = ’conc A B C D
0 1 1 0.9 0.9
32 0.8 0.8 1 0.8
64 0.9 1 1 1
128 0.9 0.9 0.8 1
256 0.7 0.9 1 0.5
512 0.4 0.3 0.4 0.2’)
df
```

```
##    conc   A   B   C   D
## 1     0 1.0 1.0 0.9 0.9
## 2    32 0.8 0.8 1.0 0.8
## 3    64 0.9 1.0 1.0 1.0
## 4   128 0.9 0.9 0.8 1.0
## 5   256 0.7 0.9 1.0 0.5
## 6   512 0.4 0.3 0.4 0.2
```

The we do some house-keeping, reformat the data and convert concentration to a factor:

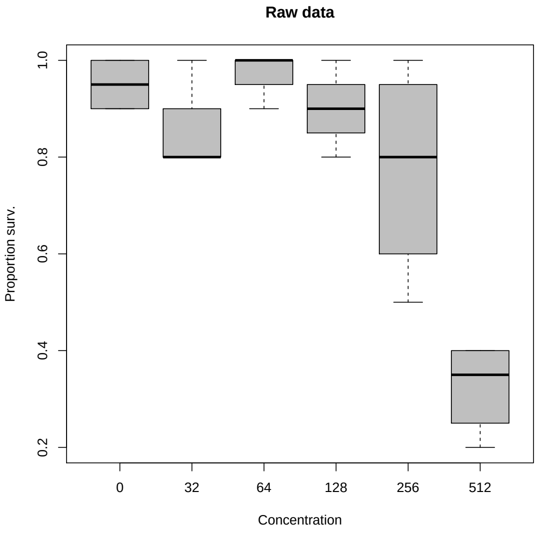
```
require(reshape2)
# wide to long
dfm <- melt(df, id.vars = 'conc', value.name = 'y', variable.name = 'tank')
# conc as factor
dfm$conc <- factor(dfm$conc)
```

So after data cleaning the data looks like

```
head(dfm)

##   conc tank   y
## 1    0    A 1.0
## 2   32    A 0.8
## 3   64    A 0.9
## 4  128    A 0.9
## 5  256    A 0.7
## 6  512    A 0.4
```

Let's have a first look at the data:

```
boxplot(y ~ conc, data = dfm,
        xlab = 'Concentration', ylab = 'Proportion surv.',
        main = 'Raw data', col = 'grey75')
```



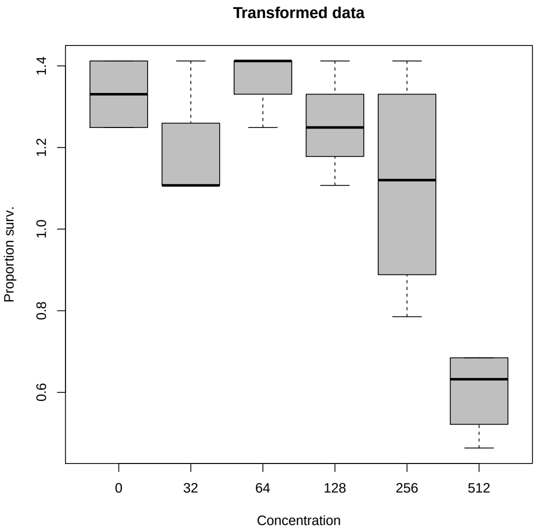
This plot indicates a strong effect at the highest concentration.

A.2.2.2 *Assuming a normal distribution of transformed proportions*

First, we arcsine transform (eqn. 6) the proportions:

```
dfm$y_asin <- ifelse(dfm$y == 1,
                     asin(1) - asin(sqrt(1/40)),
                     ifelse(dfm$y == 0,
                            asin(sqrt(1/40)),
                            asin(sqrt(dfm$y))
                            )
                     )
```

```
boxplot(y_asin ~ conc, data = dfm,
        xlab = 'Concentration', ylab = 'Proportion surv.',
        main = 'Transformed data', col = 'grey75')
```



Then, like in the count data example we fit the model using `lm()`:

```
modlm <- lm(y_asin ~ conc, data = dfm)
```

The summary gives the estimated coefficients:

```
summary(modlm)

##
## Call:
## lm(formula = y_asin ~ conc, data = dfm)
##
## Residuals:
##       Min       1Q   Median       3Q      Max
## -0.32401 -0.08149 -0.00527  0.08150  0.30261
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.33053    0.07693  17.295 1.16e-12 ***
## conc32      -0.14717    0.10880  -1.353   0.1929
## conc64       0.04074    0.10880   0.374   0.7124
## conc128     -0.07622    0.10880  -0.701   0.4925
## conc256     -0.22113    0.10880  -2.032   0.0571 .
## conc512     -0.72735    0.10880  -6.685 2.86e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.1539 on 18 degrees of freedom
## Multiple R-squared:  0.7871,Adjusted R-squared:  0.7279
## F-statistic: 13.31 on 5 and 18 DF,  p-value: 1.612e-05
```

The F-test suggests a treatment related effect:

```
drop1(modlm, test = ’F’)

## Single term deletions
##
## Model:
## y_asin ~ conc
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>              0.42613 -84.746
## conc    5    1.5753 2.00142 -57.621  13.308 1.612e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

And the LOEC is at the highest concentration:

```
summary(glht(modlm, linfct = mcp(conc = ’Dunnett’), alternative = ’less’),
        test = adjusted(’holm’))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: lm(formula = y_asin ~ conc, data = dfm)
##
## Linear Hypotheses:
##              Estimate Std. Error t value   Pr(<t)
## 32 - 0 >= 0  -0.14717    0.10880  -1.353    0.289
## 64 - 0 >= 0   0.04074    0.10880   0.374    0.644
## 128 - 0 >= 0 -0.07622    0.10880  -0.701    0.493
## 256 - 0 >= 0 -0.22113    0.10880  -2.032    0.114
## 512 - 0 >= 0 -0.72735    0.10880  -6.685 7.14e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- holm method)
```

A.2.2.3 *Assuming a binomial distribution*

The binomial model with a logit link (eqn. 7) between predictors and response can be fitted using the `glm()` function:

```
modglm <- glm(y ~ conc , data = dfm, family = binomial(link = ’logit’),
              weights = rep(10, nrow(dfm)))
```

Here the weights arguments, indicates how many fish where exposed in each treatment (N=10, eqn .7).

The summary gives the estimated coefficients:

```
summary(modglm)

##
## Call:
## glm(formula = y ~ conc, family = binomial(link = "logit"), data = dfm,
##     weights = rep(10, nrow(dfm)))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8980  -0.5723   0.0000   0.7869   2.2578
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   2.9444     0.7255   4.059 4.94e-05 ***
## conc32       -1.2098     0.8499  -1.423   0.1546
## conc64        0.7191     1.2458   0.577   0.5638
## conc128      -0.7472     0.8967  -0.833   0.4047
## conc256      -1.7077     0.8183  -2.087   0.0369 *
## conc512      -3.6753     0.8002  -4.593 4.37e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 88.672  on 23  degrees of freedom
## Residual deviance: 23.889  on 18  degrees of freedom
## AIC: 72.862
##
## Number of Fisher Scoring iterations: 5
```

To perform a LR-test we can used the `drop1()` function:

```
drop1(modglm, test = ’Chisq’)

## Single term deletions
##
## Model:
## y ~ conc
##        Df Deviance     AIC    LRT  Pr(>Chi)
## <none>      23.889  72.862
## conc    5   88.672 127.645 64.783 1.243e-12 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Also with the binomial model the LOEC is at the highest concentration:

```
summary(glht(modglm, linfct = mcp(conc = ’Dunnett’), alternative = ’less’),
        test = adjusted(’holm’))

##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Dunnett Contrasts
##
##
## Fit: glm(formula = y ~ conc, family = binomial(link = "logit"), data = dfm,
##     weights = rep(10, nrow(dfm)))
##
## Linear Hypotheses:
##                Estimate Std. Error z value   Pr(<z)
## 32 - 0 >= 0   -1.2098     0.8499  -1.423   0.2319
## 64 - 0 >= 0    0.7191     1.2458   0.577   0.7181
## 128 - 0 >= 0  -0.7472     0.8967  -0.833   0.4047
## 256 - 0 >= 0  -1.7077     0.8183  -2.087   0.0738 .
## 512 - 0 >= 0  -3.6753     0.8002  -4.593 1.09e-05 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
## (Adjusted p values reported -- holm method)
```

A.2.3 *References*

Bretz, Frank (1999). "Powerful modifications of Williams' test on trend." PhD thesis.

Bretz, Frank, Torsten Hothorn, and Peter H. Westfall (2010). *Multiple comparisons using R*. London: Chapman /& Hall.

Brock, T. C. M., M. Hammers-Wirtz, U. Hommen, T. G. Preuss, H-T. Ratte, I. Roessink, T. Strauss, and P. J. Van den Brink (2015). "The minimum detectable difference (MDD) and the interpretation of treatment-related effects of pesticides in experimental ecosystems." en. In: *Environmental Science and Pollution Research* 22.2, pp. 1160–1174.

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

Hothorn, T., F. Bretz, and P. Westfall (2008). "Simultaneous inference in general parametric models." In: *Biometrical Journal* 50.3, pp. 346–363.

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

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

Williams, D. A. (1972). "The comparison of several dose levels with a zero dose control." In: *Biometrics*, pp. 519–531.