Statistical analysis of no observed effect concentrations or levels in eco-toxicological assays with overdispersed count endpoints

Ludwig A. Hothorn, Felix M. Kluxen
Im Grund 12, D-31867 Lauenau, Germany,
ADAMA Deutschland GmbH, D-51149 Cologne, Germany
ludwig@hothorn.de

January 15, 2020

Abstract

In (eco-)toxicological hazard characterization, the No Observed Adverse Effect Concentration or Level (NOAEC or NOAEL) approach is used and often required despite of its known limitations. For count data, statistical testing can be challenging, due to several confounding factors, such as zero inflation, low observation numbers, variance heterogeneity, over- or under-dispersion when applying the Poisson model or hierarchical experimental designs. As several tests are available for count data, we selected sixteen tests suitable for overdispersed counts and compared them in a simulation study. We assessed their performance considering data sets containing mixing distribution and over-dispersion with different observation numbers. It shows that there is no uniformly best approach because the assumed data conditions and assumptions are very different. However, the Dunnett-type procedure based on most likely transformation can be recommended, because of its size and power behavior, which is relatively better over most data conditions as compared to the available alternative test methods, and because it allows flexible modeling and effect sizes can be estimated by confidence intervals. Related R-code is provided for real data examples.

1 Introduction

While the use of statistical tests have been recently criticized [2], [36], many (eco-)toxicological assays require the derivation of no observed adverse effect concentrations or levels (NOAEC, NOAEL - which are used interchangeable in the current document) by statistical test methods. Sometimes the primary endpoints are count data, such a number of offspring in daphnia aquatic assays, with several replicates per concentration [34], which are not normal-distributed. The variability between these replications should be modeled accordingly, e.g. as a variance component in the mixed effect model or as an over-dispersed discrete distribution.

Recently the use the closure principle computational approach test (CPCAT) assuming the Poisson distribution was proposed [23] for the analysis of count data, e.g. reproduction data in (eco-) toxicology. While a simulation study on power was conducted in order to verify its statistical power [22], the test was not compared to other test methods and only considered limited assumption violations. This is concerning, because the analysis of count data can be challenging, due to several confounding factors:

- A possible zero inflation, e.g. in the control group of a micronucleus assay with near-to-zero counts
- Possible over- or even under-dispersion compared to the standard Poisson model
- An additional variance component between the experimental units (plates, flasks, tanks)
- Possible variance heterogeneity between the concentrations
- Very different value ranges of the counts (either very small values in direction to ordered categorical values including zero, or large values in direction of approximate normal distribution)

In practice, these effects often occur jointly. All of those confounding factors affect the reliability of derived p-values or confidence intervals if not considered in the applied statistical test method, which can result in unreliable hazard characterization estimates.

Further, for count data in assays with small observations numbers (n_i, denoted sample size in biostatistics) a distribution assumption can neither be determined by pre-tests (because of too small sample sizes n_i) nor by mechanistic arguments of the underlying biological process (due to missing relationship between model and distribution type). Therefore and in general, a flexible and robust approach to specific assumptions or data situations would be helpful, instead of a test that is only exactly optimal for one selected assumption. As several tests are available for count data, we selected 16 tests suitable for overdispersed counts and compared them in a simulation study. Here the focus is on modifications of the Dunnett test [4] for the comparison of several treatments to a control group that provides effect size and their simultaneous confidence intervals whenever possible. We assessed their performance considering data sets containing mixing distribution and over-dispersion with different observation numbers.

It shows that there is no uniformly best approach because the assumed data conditions and assumptions are too different. The most likely transformation method [17, 14] can be recommended from the point of view of flexible modeling and robustness. And, because effect sizes can be easily derived, which is generally recommended [15]. We further describe two challenges in more detail when comparing count data, which relate to data generation and the definition of the NOAEC/L.

2 Motivating examples

Three motivating examples with count endpoints were selected and presented by boxplots also contain the raw data [27]. In the left panel, the number of offsprings in a daphnia bioassay [1], in the middle panel the number of micronuclei in a micronucleus assay on phenylethanol [5], and in the right panel the number of revertants in a TA98 Salmonella Ames assay [25] are presented. We see i) counts in a small value range or in a larger one, ii) variance heterogeneity (both increasing or almost zero), iii) approximately symmetrical distributions in designs up to 5 doses and iv) small n_i down to even 3.

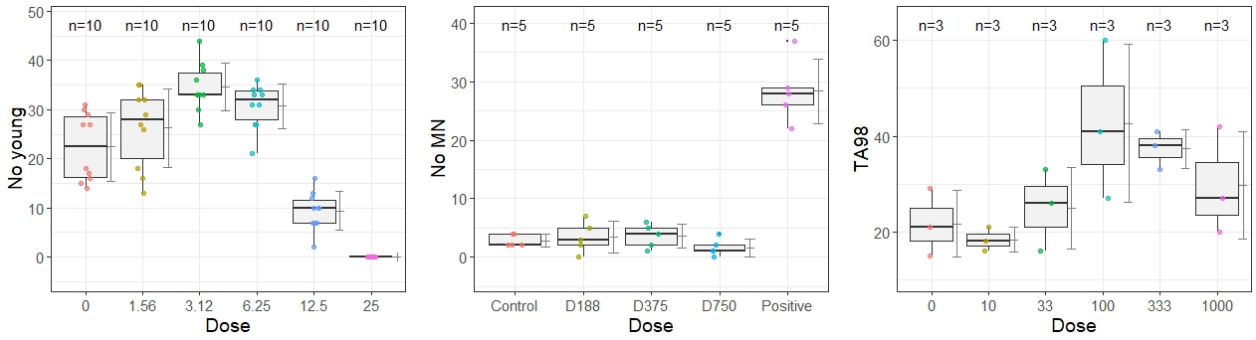

Figure 1: Box plots with superimposed individual values corresponding means with standard deviations of three example data sets from literature. Left) Shows the number of offspring (No young) from a daphnia bioassay [1]. Middle) Shows the number of micronuclei (No MN) of a genotoxicity assay conducted with phenylethanol [5]. Right) Shows the number of revertants (TA98) in on strain of an Ames assay [25].

Not just global overdispersion (i.e. variance higher than Poisson variance) occur in these data, but dose-specific dispersions and/or variance heterogeneity between the concentrations.

3 Approaches and initial considerations

A problem is the unclear definition of the NOAEC in a randomized design, which usually includes several concentrations $NC(C_0 = 0), C_1, ..., C_k$ (where k commonly 3-5). The actually direct method is the estimation of the maximum safe dose $max(i)|\mu_{D_i} \leq (\mu_0 + \delta)(i = 1, 2, ...)$ [13, 32, 33], i.e., by directed non-inferiority tests. This is not used in (eco-)toxicology at all, probably because a consensus for endpoint-specific non-inferiority threshold δ is lacking, i.e. an agreed effect size which is not considered toxicologically relevant. However, also the assumption of NOAEC via $NOAEC = D_{MED} - 1$ is problematic because of the direct control of the false positive error rate, and also because the definition of the minimum effective dose (MED) is blurred: significant or relevant decision, assuming a concentration-response monotonicity into account or not, using hypothesis tests or nonlinear models, modeling concentration qualitatively [9] or quantitatively [29, 3].

3.1 NOAEC definition

Two definitions are used here:

A) Just the concentration before the lowest significant one

$$NOAEC = C_{i-1} : min(C_i)|H^1(\mu_{C_i} - \mu_0) \tag{1}$$

where all higher dose effects are ignored.

B) The concentration before the lowest significant concentration where all higher ones must be significant as well- but not necessarily monotonous ordered.

$$NOAEC = D_{i-1} : min(D_i)|H^1(\mu_{D_i} - \mu_0) \bigcap \forall j \in (i+1, ..., k) : H^1(\mu_{D_j} - \mu_0) \tag{2}$$

Both definitions can be extended for effective concentrations, i.e. $H^1(\mu_{C_i} - \mu_0 + \Delta)$ which is commonly not used because of the missing fixation of Δ. One-sided tests (or more appropriate confidence limits) are throughout where the direction of harm is a-priori known.

Note, contrary to the both definitions, the no observed adverse effect level is also defined as *the highest dose tested that causes no adverse effects in a test species in a properly designed and executed study* [6], which becomes relevant when no statistically observed effect levels exist between statistically observed effect levels.

3.2 NOAEC procedures

Following eq(1) definition, the following four hypotheses tests can be performed: i) Closed testing procedure (CTP) [10], ii) Dunnett-type procedure [12], iii) step-down pairwise tests [7], and iv) contrast tests [28].
CTP for eq(1) can be formulated for comparison against control only, forming a complete family of hypotheses [31], which does no need complex partition-type hypotheses. For $k = 3$ and simple tree alternative the hypotheses are: global $[0, 1, 2, 3]$, partition: $[0, 1, 2], [0, 2, 3], [0, 1, 3]$, elementary $[0, 1], [0, 2], [0, 3]$; closed under intersection. While for monotonic order restriction the hypotheses are much simpler: $[0, 1, 2, 3]$, partition: $[0, 1, 2]$, elementary $[0, 1]$ (because the rejection of $[0, 1, 2, 3]$ implies the rejection of $[0, 2, 3], [0, 1, 3]$, etc. under monotonic alternative. Any level α-test can be used for each hypothesis (including different tests for different hypotheses). Reflecting the many-to-one (NC) comparisons, the Helmert contrast seems to be suitable for simple tree alternatives, where under total order restriction two-sample tests $D_{max} - D_0$ (or better related pairwise contrasts when the common degree of freedom is used) are suitable. ANOVA-type tests can be used as well, because of their inherently 2-sided hypotheses formulation they are not too appropriate [22]. A hybrid method robust for both monotone and non-monotone shapes is available as well [38].
Notice, model selection approaches [21] or Bayesian approaches [26] are not considered here , because they follow different testing philosophies, which would not allow an appropriate comparative assessment.

3.3 Possible biased estimations

At least six sources of biased NOAEC estimations exists, e.g. where the estimated NOAEC of one group is affected by the responses in other groups: i) NOAEC depends directly on σ/n_i, both are no standardized (not so much in regulatory assays with at least n_i recommendation and a common variance), ii) NOAEC is an experimental concentration only (no model-specific interpolation), and depends on an appropriate choice of concentrations (number, intervals), iii) NOAEC based on a directional decision (a-priori clear either increase or decrease)- where two-sided testing may cause bias, iv) pooling contrasts/tests [8], [13], v) ignoring concentration-specific variance heterogeneity when using common MSE estimator, particularly in unbalanced designs, and vi) ignoring concentration-specific dispersions where common approaches focusing on a global dispersion parameter only. Figure 2 shows two common scenarios for such a bias. In the left panel $NOAEC = D_1$, but the Williams contrast test [37] finds $MED = D1$ because of the significant pooled $(\mu_1 + \mu_2 + \mu_3)/3$ contrast. In the right panel $NOAEC = D_1$, but the standard Dunnett test revealed $MED = D_3$ because the extreme variance in D_3 increases the common variance estimator and hence D_2 is biased not significant. The related R-code is available in the Appendix.

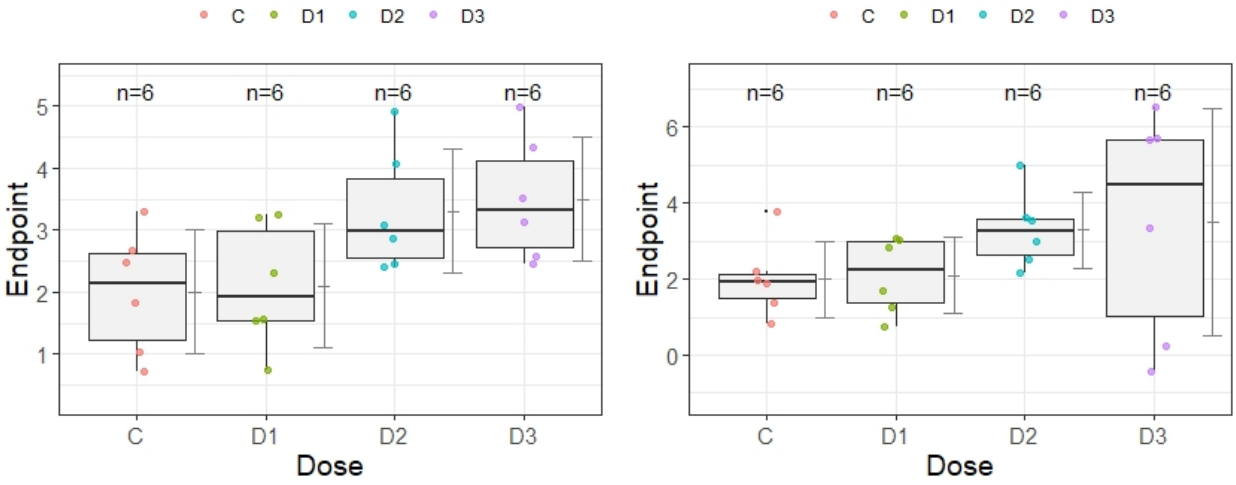

Figure 2: Box plots with superimposed individual values corresponding means with standard deviations of two simulated experiments with biased NOAEC estimation. The left panel shows a plateau-type shape where the Williams test estimates a too small NOAEC due to its pooling contrast issue. The right panel shows an increased variance in the high dose group resulting in a too large NOAEC estimation

4 Procedures comparing overdispersed counts in a design $[NC, C_1, ..., C_k]$

As the standard approach, modifications of the Dunnett test [4] for comparing treatments versus a negative control in the generalized linear mixed effect model (glmm) [16] are commonly used. To model the variation between the experimental units, a quasi-Poisson model can be assumed or a random factor *between units* can be used in the mixed effect model. For the latter different algorithms exist, whereby the selection was made from the point of view of numerical stability at small n_i. Alternatively a negative binomial model, not belonging to the glm family, can be assumed. Notice, these asymptotic approaches may be problematic for small n_i for both numerical stability and controlling the coverage probability. Therefore, a simple Freeman-Tukey transformation was used for the Dunnett test [11], [19] particularly appropriate for small n_i. When the distribution is unknown and not only location effects should be considered, the most likely transformation model [12] can be used for the Dunnett test [14], the similar approach of continuous outcome logistic regression model [24] and the more specifically the cotram approach for count endpoints [30], i.e. most likely transformation methods for counts. A completely different approach is represented by the nonparametric Dunnett-type procedure for relative effect size, allowing any discrete distribution with variance heterogeneity [20]. The CPCAT test, a closed testing principle for parametric bootstrap tests [22] can also be conducted and has been proposed before for the assessment of (eco-)toxicological bioassays.
To account for variance heterogeneity between the concentrations, a sandwich variance estimator can be used when available, for linear models [39] as well as for generalized mixed effect models [35].

We compared 16 statistical test candidates in a simulation study, that cover the described different methods to analyse count data. Specifically the following tests were used:

1. Standard Dunnett test assuming normal distributed homoscedastic errors (Du)
2. Dunnett test modified with a sandwich variance estimator (DuS)
3. Dunnett test for log-transformed counts (DuL)
4. Dunnett test for log-transformed counts and a sandwich variance estimator (DuLS)
5. Dunnett test for Freeman-Tukey transformed counts (FT)
6. Dunnett test for Freeman-Tukey transformed counts and a sandwich variance estimator (DuFS)

7. Dunnett test in glm using Poisson link function (without overdispersion) (DuP)
8. Dunnett test in glm using quasi-Poisson link function (QP)
9. Dunnett test in the negative binomial model (DuN)
10. Dunnett test in the mixed effect model (MM)
11. Dunnett test in the mixed effect model using related sandwich estimator (DuMS)
12. Dunnett test for most likely transformation (MLT)
13. Dunnett test for continuous outcome logistic regression model (CORL)
14. Dunnett test for mlt-cotram (COT)
15. Dunnett test for continuous outcome logistic regression model (COTd)
16. Closure principle computational approach test (CPCAT)

In the appendix, the simulation results of these tests are presented in detail. After an initial study comparing 16 possible test candidates, we compared the most promising and relevant tests in a thorough simulation study, which is detailed in the following.

5 Simulation study

5.1 Size and any-pair power for comparisons vs. control

It is not clear from the start how to generate random count data in the (k+1) sample design with possible overdispersion and/or variance heterogeneity in the high(er) concentrations to characterize the size/power behavior of (eco-)toxicological assays. A first model assumes a mixing distribution of responding and non-responding subjects in the higher concentrations, based on integered normal mixing distributions (MIX). A second model based on mean-parametrized Conway-Maxwell-Poisson distribution allowing concentration specific dispersions [18] (CMP) For a fair comparison to CPCAT, 2-sided tests were considered only and tests with p-value outcome only.

5.1.1 Results

Figure 3 shows two typical simulated scenarios. In the left panel an assay based on integered mixing distribution ($responder + non - responder$) (MIX) with overdispersion in the highest concentration is shown and in the right panel mean-parametrized Conway-Maxwell-Poisson distribution (CMP) in an inhibition assay with underdispersion in the highest concentration is displayed. Both balanced and unbalanced $[NC, C_1, C_2, C_3]$ designs were considered.

In the simulation study, the empirical size and any-pairs power (i.e. at least one contrast, anyone, in the alternative) were estimated from 16 test variants under different balanced and unbalanced designs (with medium and very small n_i) for both distributions (MIX, CPM) with and without over-dispersion.

Appendix 1 shows the size (under the null hypothesis H_0) and power (under certain alternative H_1) estimates for various conditions for these 16 tests. Not surprisingly, there is no uniformly best test for all conditions. Table 1 summarized these estimated for the five preferred tests under different conditions: the Dunnett-tests for Freeman-Tukey transformed counts (FT), GLM with quasipoisson link function (QP), GLMM with Poisson link function (MM), most likely transformation (MLT) and as well the closed testing approach (CPCAT). Both site and power are not too different, i.e. from this point of view these 5 tests are all usable, The power functions for overdispersed CPM distribution in Figure 4 reveals a quite similar pattern under this special balanced $n_i = 10$ design, where MM and QP is more powerful than CPCAT over a wide range of non-centrality (effect size increase against control in a fixed design with constant variance).

Notice, CPCAT is formulated for two-sided hypotheses only, whereas the other approaches can be formulated two-sided, or more appropriate, one-sided for NOAEC estimation. The impact on power

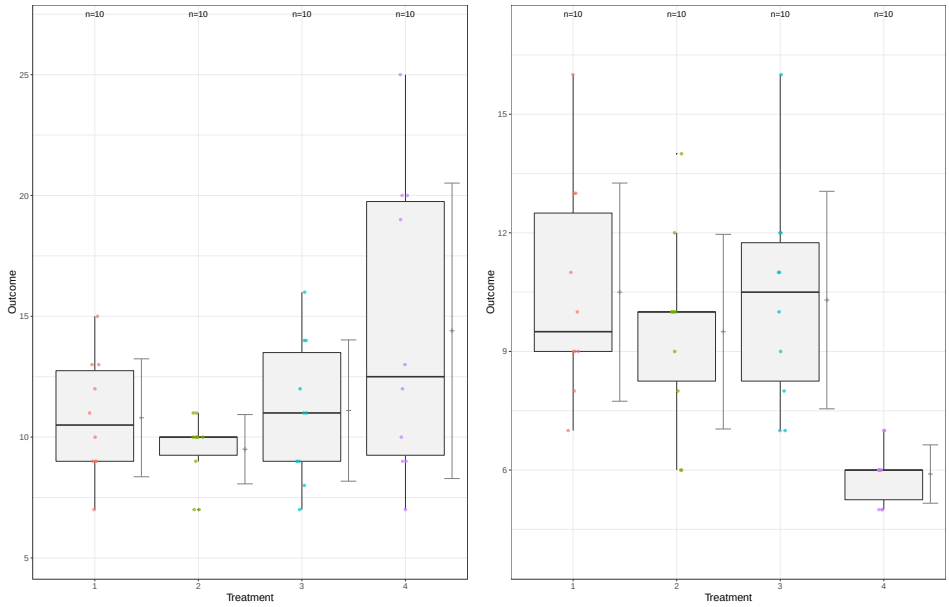

Figure 3: Example data sets in the simulation study under specific alternatives. The left panel shows a mixing distribution with overdispersion in the highest concentration. The right panel shows a mean-parametrized Conway-Maxwell-Poisson distribution with underdispersion in the highest concentration.

is serious, e.g. for a certain point in H_1 (CPM, overdispersed) the power of the two-sided MLT and CPCAT is 0.83, whereas the power increases to 0.90 for one-sided MLT Dunnett-type test.

5.2 Correct NOAEC estimation rates

The correct NOAEC rates for the above selected procedures (FT, QP, MM, MLT, CPCAT) were estimated for relevant concentration-response patterns (for 1000 runs and CPM distribution with overdispersion. The correct estimation rates are presented for selected conditions in Appendix II. Figure 5 shows a mosaicplot with the correct NOAEC estimation rates for decreasing non-centralities, abbreviated by C1, C2,...,C7. For condition C1 the true NOAEC is number 2 estimate, for condition C7 the true NOEAC is number 3 estimate where C2-C6 represent decreasing non-centrality in between. The patterns of the five tests are quite similar, with advantages for QP,MM and MLT.

Distribution	Hypothesis	Dispersion	n	FT	QP	MM	MLT	CPCAT
MIX	H0	-	10	0.06	0.01	0.07	0.08	0.001
	H0	over	10	0.06	0.07	0.02	0.06	0.01
	H1	-	10	0.91	0.87	0.40	0.88	0.16
	H1	over	10	0.50	0.60	0.38	0.51	0.22
CMP	H0	-	10	0.05	0.06	0.04	0.07	0.04
	H0	-	5	0.06	0.08	0.04	0.11	0.03
	H0	over	10	0.05	0.05	0.05	0.06	0.05
	H0	over	5	0.06	0.10	0.07	0.11	0.08
	H1	-	10	0.89	0.91	0.90	0.91	0.85
	H1	over	10	0.84	0.87	0.88	0.85	0.84

Table 1: Size (under H0) and power (under H1) estimations for selected tests: (FT) Dunnett-tests for Freeman-Tukey transformed counts, (QP) generalized linear model with quasipoisson link function, (MM) generalized mixed effect linear model with Poisson link function, (MLT) most likely transformation, (CPCAT) the closed testing approach, under mixing distribution (MIX) or Conway-Maxwell-Poisson distribution (CMP) with and without overdispersion

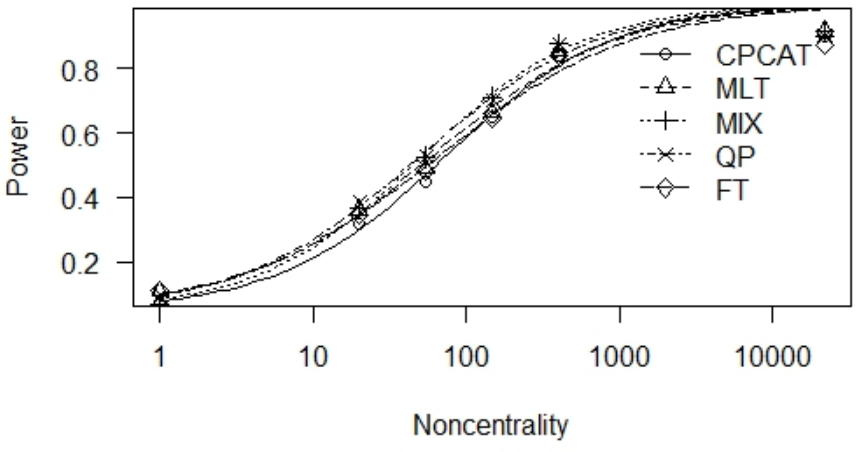

Figure 4: Power curve for selected tests: (FT) Dunnett-tests for Freeman-Tukey transformed counts, (QP) generalized linear model with quasipoisson link function, (MM) generalized mixed effect linear model with Poisson link function, (MLT) most likely transformation, (CPCAT) the closed testing approach. Power is (1-false negative error rate). Non-centrality is effect size increase against control

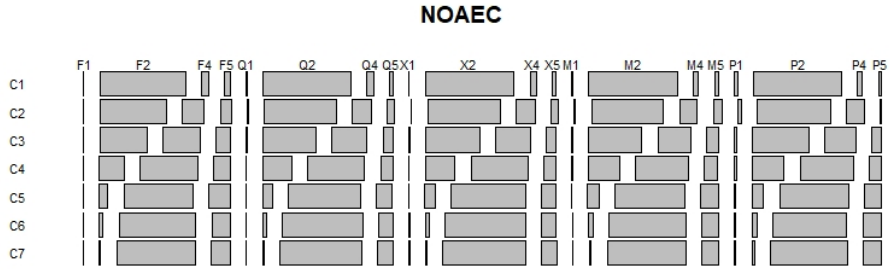

Figure 5: Mosaicplot for correct NOAEC estimation rates for 7 decreasing non-centralities C1,...,C7 for the selected tests FT, QP, MM, MLT, CPCAT .

6 Evaluation of two examples

The daphnia example was analysed with the five selected methods, where decreasing number of young is the relevant direction of harm. The related R-code is given in Appendix II. Table 2 reveals test statistics (t) and multiplicity-adjusted p-value for the five comparisons of the dose groups against control for the five selected tests (FT, QP, MM, MLT, CPCAT).While for FT, MLT and CPCAT the NOAEC is correctly estimated at 6.25, QP and MM show biased estimates due to variance heterogeneity and over-dispersion in the data.

Comparison	FT(t)	FT(p)	QP(t)	QP(p)	MM(t)	MM(p)	MLT(t)	MLT(p)	CPCAT(p)
1.56-0	1.16	0.98	1.58	0.99	1.77	0.99	-1.38	0.99	0.0818
3.12-0	4.69	0.99	4.53	0.99	5.07	0.99	-4.21	0.99	0.0001
6.25-0	3.30	0.99	3.20	0.99	3.59	0.99	-2.75	0.99	0.0013
12.5-0	-5.51	< 0.001	-6.31	< 0.001	-7.07	< 0.001	4.69	< 0.001	< 0.001
25-0	-19.25	< 0.001	-0.01	0.89	-0.04	0.89	6.72	< 0.001	< 0.001

Table 2: Test statistics (t) and adjusted p-value (p) of five selected tests FT, QP, MM, MLT, CPCAT of the daphnia example for the individual comparisons of dose against control

The better representation are simultaneous confidence intervals, e.g. for the MLT approach (adapted to continuous outcome logistic regression modeling) where odds ratio are used as effect size, as demonstrated for the micronucleus assay example without positive control in Figure 6.

The dots representing the point estimators of the odds ratios (OR). The dotted line at 1 is for the null hypothesis of $OR = 1$. None of the lower limits is larger than 1, i.e. neither increase of number of MN in any dose can be concluded.

7 Conclusions

NOAEC/L estimation of count data in (eco-)toxicity assays poses a challenge, due to usually very small observation numbers, with overdispersed counts and an a priori unclear distribution assumption as well as possible variance heterogeneity between concentrations. We demonstrated that there is no uniformly best approach simply because the underlying data conditions and assumptions are very different between study designs. The most likely transformation methods can be recommended from the point of view of flexible modeling and robustness. They perform well in most data conditions. Procedures providing simultaneous confidence intervals, especially their interpretation as compatibility intervals, should be preferred to p values, as they allow the assessment of effect sizes which is recommended by

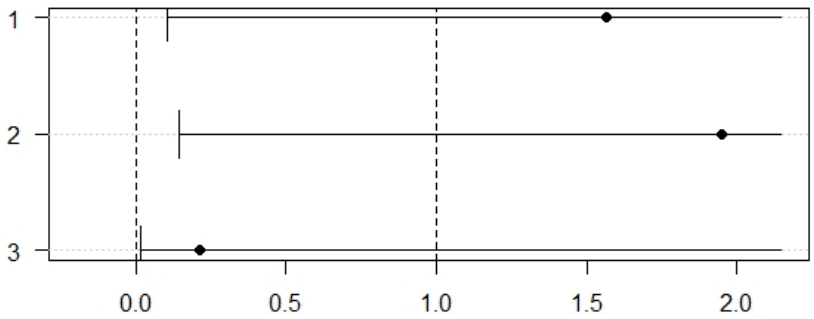

Figure 6: Odds ratios and their simultaneous confidence limits for the micronuceus example estimated by a MLT approach using continuous outcome logistic regression modeling

several authors, e.g. [2, 15]. Our results also show that caution is advised when assays with too small n_i, e.g. $n_i = 3$ are to be evaluated, which is similar to our previous observation for continuous data sets [14].
Corresponding R-code is included in the Appendix III for the examples and can be easily adapted to other data situations.

8 Appendix I: Detailed simulation results

Abbreviated: MIX ...integered normal distributions, H0 ... null hypothesis, H1... alternative hypothesis, d ... using finite df, s... using sandwich estimator.

Table 3: Simulated size (under H0) and power (under H1) for overdispersed integered mixing normal distributions

	k	n1	n4	s1	s4	DuO	DuS	DuL	DuLS	FT	DuFS	DuP	QP	MM	DuMS	MLT	MITd	MITs	CORL	COT	COTd	NP	CPCAT
MIXMH0221	1	10	10	2	2	0.0	0.078	0.047	0.076	0.043	0.077	0.001	0.056	0.001	0.001	0.065	0.050	0.074	0.055	0.060	0.051	0.032	0.001
MIXMH0241	1	10	10	2	4	0.1	0.092	0.057	0.090	0.055	0.095	0.018	0.068	0.018	0.018	0.062	0.053	0.072	0.079	0.078	0.065	0.037	0.011
MIXH0221	1	10	10	2	2	0.0	0.078	0.047	0.076	0.043	0.077	0.001	0.056	0.001	0.001	0.065	0.050	0.074	0.055	0.060	0.051	0.032	0.001
MIXH0224	1	10	10	2	4	0.1	0.092	0.057	0.090	0.055	0.095	0.018	0.068	0.018	0.018	0.062	0.053	0.072	0.079	0.078	0.065	0.037	0.011
MIXH02222	10	10	2	2	0.1	0.083	0.049	0.083	0.056	0.087	0.001	0.070	0.001	0.001	0.078	0.057	0.079	0.061	0.067	0.048	0.039	0.001	
MIXH02212	1	10	10	2	2	0.0	0.077	0.043	0.074	0.044	0.077	0.001	0.058	0.001	0.001	0.066	0.051	0.071	0.055	0.057	0.045	0.036	0.001
MIX13142	2	10	10	2	2	0.7	0.759	0.621	0.737	0.662	0.755	0.408	0.707	0.407	0.407	0.723	0.688	0.723	0.711	0.707	0.662	0.624	0.200
MIX13132	1	10	10	2	2	0.8	0.876	0.753	0.868	0.789	0.871	0.304	0.815	0.304	0.304	0.843	0.815	0.849	0.805	0.806	0.765	0.746	0.128
MIX13132	1	10	10	2	4	0.5	0.475	0.392	0.410	0.472	0.438	0.355	0.567	0.355	0.354	0.495	0.459	0.354	0.488	0.505	0.459	0.260	0.183
MIX1314221	1	10	10	2	2	0.9	0.911	0.793	0.901	0.835	0.911	0.391	0.866	0.392	0.391	0.875	0.853	0.893	0.852	0.858	0.815	0.812	0.163
MIX1314241	1	10	10	2	4	0.6	0.528	0.430	0.464	0.501	0.496	0.384	0.602	0.382	0.382	0.512	0.461	0.409	0.531	0.538	0.493	0.316	0.223

Table 4: Simulated size (under H0) and power (under H1) for overdispersed CPM distribution

	n1	n4	nu1	nu4	DuO	DuS	DuL	DuLS	FT	DuFS	DuP	QP	MM	DuMS	MLT	MITd	MITs	CORL	COT	COTd	NP	CPCAT
CPM1010111	10	10	1	1	0.044	0.072	0.048	0.067	0.047	0.072	0.045	0.056	0.044	0.043	0.070	0.046	0.073	0.054	0.057	0.045	0.032	0.036
CPM101311	10	10	1	1	0.370	0.403	0.306	0.399	0.349	0.407	0.372	0.376	0.354	0.354	0.405	0.360	0.379	0.375	0.377	0.328	0.263	0.235
CPM101311	10	10	1	1	0.370	0.421	0.323	0.411	0.353	0.422	0.382	0.388	0.357	0.357	0.413	0.368	0.404	0.365	0.370	0.324	0.273	0.252
CPM101010.6	10	10	1	1	0.047	0.077	0.052	0.080	0.048	0.080	0.067	0.053	0.051	0.051	0.063	0.049	0.072	0.061	0.060	0.045	0.038	0.054
CPM101611	10	10	1	1	0.907	0.916	0.844	0.901	0.888	0.917	0.914	0.906	0.903	0.903	0.914	0.895	0.884	0.888	0.890	0.868	0.807	0.849
CPM10100.60.4	10	10	1	0	0.038	0.079	0.043	0.077	0.041	0.081	0.102	0.046	0.060	0.060	0.059	0.044	0.068	0.053	0.057	0.041	0.035	0.070
CPM10200.6	10	10	1	1	0.998	0.996	0.985	0.991	0.996	0.995	0.999	0.999	0.998	0.998	0.992	0.988	0.968	0.988	0.993	0.988	0.948	0.998
CPM10100.6(5T)	10	10	1	1	0.049	0.084	0.049	0.082	0.048	0.086	0.068	0.059	0.054	0.054	0.067	0.053	0.076	0.066	0.067	0.052	0.044	0.055
CPM101011(5T)	10	10	1	1	0.049	0.083	0.050	0.077	0.050	0.081	0.049	0.064	0.045	0.045	0.076	0.060	0.075	0.064	0.068	0.051	0.040	0.041
CPM10160.6	10	10	1	1	0.867	0.845	0.777	0.813	0.837	0.837	0.907	0.873	0.876	0.876	0.853	0.824	0.785	0.815	0.827	0.795	0.669	0.840
CPM1016160.60.6	10	10	1	1	0.920	0.926	0.859	0.910	0.896	0.917	0.979	0.938	0.945	0.945	0.907	0.892	0.892	0.900	0.905	0.883	0.822	0.975
CPM10160.60.6	10	10	1	1	0.823	0.806	0.669	0.776	0.781	0.799	0.905	0.831	0.840	0.840	0.796	0.758	0.742	0.771	0.791	0.739	0.612	0.836
CPM10611.6	10	10	1	2	0.768	0.885	0.806	0.844	0.798	0.876	0.812	0.861	0.800	0.800	0.848	0.800	0.827	0.814	0.833	0.783	0.708	0.709
CPM10761.62 1	10	10	1	2	0.880	0.917	0.854	0.872	0.876	0.904	0.854	0.914	0.848	0.847	0.903	0.879	0.871	0.882	0.887	0.862	0.780	0.831
CPM510100.6	5	5	1	1	0.067	0.150	0.063	0.134	0.064	0.148	0.077	0.096	0.067	0.067	0.109	0.073	0.126	0.081	0.075	0.044	0.065	0.076
CPMH011	5	5	1	1	0.053	0.134	0.048	0.130	0.055	0.132	0.041	0.080	0.039	0.039	0.108	0.070	0.102	0.077	0.075	0.031	0.067	0.028
CPM75311	7	3	1	1	0.050	0.131	0.048	0.120	0.051	0.130	0.043	0.067	0.040	0.040	0.078	0.053	0.094	0.055	0.055	0.025	0.065	0.043
CPMH011	5	5	1	1	0.046	0.118	0.043	0.109	0.048	0.117	0.044	0.068	0.040	0.040	0.082	0.050	0.098	0.070	0.066	0.031	0.048	0.039
CPM10160.6	10	10	1	1	0.867	0.845	0.777	0.813	0.837	0.837	0.907	0.873	0.876	0.876	0.853	0.824	0.785	0.815	0.827	0.795	0.669	0.840
CPM1020111 1	5	5	1	1	0.940	0.941	0.859	0.939	0.911	0.942	0.960	0.942	0.955	0.955	0.935	0.890	0.862	0.755	0.888	0.769	0.795	0.923
CPM510100.6	5	5	1	1	0.067	0.150	0.063	0.134	0.064	0.148	0.077	0.096	0.067	0.067	0.109	0.073	0.126	0.081	0.075	0.044	0.065	0.076
CMP10130.6	10	10	1	1	0.361	0.349	0.277	0.309	0.324	0.344	0.432	0.383	0.376	0.376	0.362	0.329	0.305	0.346	0.350	0.313	0.212	0.318
CMP10140.6	10	10	1	1	0.512	0.490	0.386	0.451	0.470	0.483	0.585	0.526	0.527	0.527	0.490	0.448	0.434	0.455	0.464	0.417	0.318	0.448
CPM10150.6	10	10	1	1	0.700	0.652	0.547	0.615	0.646	0.643	0.773	0.701	0.711	0.712	0.664	0.622	0.580	0.628	0.633	0.590	0.447	0.648
CPM10160.6	10	10	1	1	0.867	0.845	0.777	0.813	0.837	0.837	0.907	0.873	0.876	0.876	0.853	0.824	0.785	0.815	0.827	0.795	0.669	0.840
CPM10200.6	5	5	1	1	0.877	0.863	0.771	0.857	0.843	0.868	0.943	0.900	0.916	0.916	0.847	0.783	0.756	0.666	0.789	0.657	0.688	0.896

9 Appendix II: Correct NOAEC estimation rates

Table 5: Correct NOEAC estimation rates (NF_j for FT-test, NQ_j for QP-test, NM_j for MM test, ML_j for MLT-test, CP_j for CPCAT-test

	n4	mu2	mu3	mu4	nu1	nu4	NF1	NF2	NF3	NF4	NF5	NQ1	NQ2	NQ3	NQ4	NQ5	NM1	NM2	NM3	NM4	NM5	ML1	ML2	ML3	ML4	ML5	CP1	CP2	CP3	CP4
1	10	10	10	16	1.0	0.6	2	14	12	785	187	3	20	18	812	153	3	17	15	825	145	7	23	18	778	174	10	28	25	762
1	10	10	16	16	1.0	0.6	6	862	722	71	54	9	887	777	65	31	6	892	782	63	31	13	890	755	53	37	40	876	778	52
1	10	16	16	16	1.0	0.6	659	82	50	32	37	718	70	47	26	24	736	66	45	19	15	713	70	42	22	24	886	25	17	7
1	10	10	16	16	1.0	0.6	6	862	722	71	54	9	887	777	65	31	6	892	782	63	31	13	890	755	53	37	40	876	778	52
1	10	10	15	16	1.0	0.6	5	664	579	219	107	9	721	652	188	75	8	726	656	188	70	12	716	630	176	90	31	731	667	174
1	10	10	14	16	1.0	0.6	7	473	428	369	144	12	524	488	357	100	9	544	511	347	92	10	532	487	332	115	25	558	534	319
1	10	10	13	16	1.0	0.6	3	251	239	572	170	4	297	283	569	125	4	300	286	574	117	6	312	295	536	136	24	319	303	539
1	10	10	12	16	1.0	0.6	4	83	79	695	208	6	99	96	726	158	3	105	104	741	144	4	111	107	698	176	19	118	116	688
1	10	10	11	16	1.0	0.6	3	35	33	765	185	3	41	40	800	141	2	39	38	800	145	5	48	47	762	165	14	54	54	731

10 Appendix III: R-code of selected tests

```
## ----datMN, echo=FALSE,results='hide', warning=FALSE, message=FALSE----------------------------
mn <- structure(list(group = structure(c(1L, 1L, 1L, 1L, 1L, 2L, 2L,
2L, 2L, 2L, 3L, 3L, 3L, 3L, 3L, 4L, 4L, 4L, 4L, 4L, 5L, 5L, 5L,
5L, 5L), .Label = c("Control", "D188", "D375", "D750", "Positive"
), class = "factor"), animal = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 22, 23, 24),
    MN = c(4, 2, 4, 2, 2, 3, 5, 7, 2, 0, 5, 6, 1, 4, 2, 2, 4,
    1, 1, 0, 26, 28, 22, 37, 29)), .Names = c("group", "animal",
"MN"), row.names = c(NA, 25L), class = "data.frame")


## ----datDapnhia, echo=FALSE, result='asis',  warning=FALSE-------------------------------------
daphnia <-
structure(list(`Concentration ` = c(0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 1.56, 1.56, 1.56, 1.56, 1.56, 1.56, 1.56, 1.56, 1.56, 1.56,
3.12, 3.12, 3.12, 3.12, 3.12, 3.12, 3.12, 3.12, 3.12, 3.12, 6.25,
6.25, 6.25, 6.25, 6.25, 6.25, 6.25, 6.25, 6.25, 12.5, 12.5,
12.5, 12.5, 12.5, 12.5, 12.5, 12.5, 25, 25, 25, 25,
25, 25, 25, 25, 25), Adults = c(1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1, 2, 3, 4, 5, 6, 7, 8, 9,
10, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10), Number_Young = c(27, 30,
29, 31, 16, 15, 18, 17, 14, 27, 32, 35, 32, 26, 18, 29, 27, 16,
35, 13, 39, 30, 33, 33, 36, 33, 33, 27, 38, 44, 27, 34, 36, 34,
31, 27, 33, 21, 33, 31, 10, 13, 7, 7, 7, 10, 10, 16, 12, 2, 0,
0, 0, 0, 0, 0, 0, 0, 0)), class = "data.frame", row.names = c(NA,
60L))


## ----caseE17aa, echo=FALSE, result="asis",  warning=FALSE-------------------------------------
daphnia$Conc<-as.factor(daphnia$Concentration)
library(dispmod)
data(salmonellaTA98)
salmonellaTA98$Dose<-as.factor(salmonellaTA98$x)


## ----sensa1, echo=FALSE, result="asis", fig.cap='Box-plots count endpoints', fig.subcap=c('Daphnia', 'MN', 'TA98'), out.width='.28\\linewidth',fig.align='center', fig.pos="H
library(toxbox)
boxclust(data=daphnia, outcome="Number_Young", treatment="Conc",  xlabel="Dose", ylabel="No young", psize=1.5, hjitter=0.1, vlines="fg", white=TRUE, printN=FALSE)
boxclust(data=mn, outcome="MN", treatment="group",  xlabel="Dose", ylabel="No MN", psize=1.5, hjitter=0.1, vlines="fg", white=TRUE, printN=FALSE)
boxclust(data=salmonellaTA98, outcome="y", treatment="Dose",  xlabel="Dose", ylabel="TA98", psize=1.5, hjitter=0.1, vlines="fg", white=TRUE, printN=FALSE)


## ----bias1, echo=FALSE, result="asis", fig.cap='Box-plots  for biased NOEAC', fig.subcap=c('Williams', 'Variance hetero'), out.width='.28\\linewidth',fig.align='center', fig
library(SimComp) # laoding the selected CRAN package
set.seed(127711) # selecting a seed allows reproducibility
control <- ermvnorm(n=6 ,mean=2, sd=1)
dose1 <- ermvnorm(n=6 ,mean=2.1, sd=1)
dose2 <- ermvnorm(n=6 ,mean=3.3, sd=1)
dose3 <- ermvnorm(n=6 ,mean=3.5, sd=1)
dose <- factor(rep(c("C","D1","D2","D3"), c(6,6,6,6)))
endpoint <- c(control, dose1, dose2, dose3)
datTA <- data.frame(endpoint, dose)
library(toxbox)
boxclust(data=datTA, outcome="endpoint", treatment="dose",  xlabel="Dose", ylabel="Endpoint", psize=1.5, hjitter=0.1, vlines="fg", white=TRUE, printN=FALSE)

library(multcomp)
mod1<-lm(endpoint~dose, data=datTA)
M1<-summary(glht(mod1, linfct = mcp(dose = "Williams"), alternative="greater"))
M2<-summary(glht(mod1, linfct = mcp(dose = "Dunnett"), alternative="greater"))

control <- ermvnorm(n=6 ,mean=2, sd=1)
dose1 <- ermvnorm(n=6 ,mean=2.1, sd=1)
dose2 <- ermvnorm(n=6 ,mean=3.3, sd=1)
dose3a <- ermvnorm(n=6 ,mean=3.5, sd=3)
dose <- factor(rep(c("C","D1","D2","D3"), c(6,6,6,6)))
endpoint1 <- c(control, dose1, dose2, dose3a)
datTA1 <- data.frame(endpoint1, dose)
library(toxbox)
boxclust(data=datTA1, outcome="endpoint1", treatment="dose",  xlabel="Dose", ylabel="Endpoint", psize=1.5, hjitter=0.1, vlines="fg", white=TRUE, printN=FALSE)

mod2<-lm(endpoint1~dose, data=datTA1)
M3<-summary(glht(mod2, linfct = mcp(dose = "Williams"), alternative="greater"))
M4<-summary(glht(mod2, linfct = mcp(dose = "Dunnett"), alternative="greater"))


## ----pow1, echo=FALSE, result="asis", fig.cap='Power curves', out.width='.48\\linewidth',fig.align='center', fig.pos="H", cache=TRUE, warning=FALSE----
CPCAT<-rbind(c(0,3,4,5,6, 10),c(0.076,  0.318, 0.448, 0.648, 0.840, 0.896))
MLT<-rbind(c(0,3,4,5,6, 10),c(0.11,  0.362, 0.490, 0.664, 0.853, 0.916))
MIX<-rbind(c(0,3,4,5,6, 10),c(0.067,  0.367, 0.527, 0.711, 0.876, 0.916))
QP<-rbind(c(0,3,4,5,6, 10),c(0.096,  0.383, 0.526, 0.701, 0.873, 0.90))
FT<-rbind(c(0,3,4,5,6, 10),c(0.11,  0.344, 0.483, 0.643, 0.837, 0.87))
power<-as.data.frame(rbind(t(CPCAT), t(MLT), t(MIX), t(QP), t(FT)))
power$ID<-as.factor(c(rep("CPCAT", 6),rep("MLT", 6), rep("MIX", 6), rep("QP", 6), rep("FT", 6) ))
colnames(power)<-(c("noncentrality","prob", "ID"))
library(drc)
mypow <- drm(prob~exp(noncentrality), curveid=ID, data = power, fct = LL.4(fixed=c(NA, 0.05, 1, NA)))
plot(mypow, ylim=c(0.1,0.95), type="all", xlab="Noncentrality")


## ----corr1, echo=FALSE, cache=TRUE, result='hide', warning=FALSE-------------------------------
C1<-c(6, 6, 862, 722, 71 , 54 , 9 , 887 , 777 , 65 , 31 , 6 , 892 , 782 , 63 , 31 , 13 , 890 , 755 , 53 , 37 , 40 , 876 , 778 , 52 , 23)
C2<-c(5, 5 , 664 , 579 , 219 , 107 , 9 , 721 , 652 , 188 , 75 , 8 , 726 , 656 , 188 , 70 , 12 , 716 , 630 , 176 , 90 , 31 , 731 , 667 , 174 , 5)
C3<-c(4 , 7 , 473 , 428 , 369 , 144 , 12 , 524 , 488 , 357 , 100 , 9 , 544 , 511 , 347 , 92 , 10 , 532 , 487 , 332 , 115 , 25 , 558 , 534 , 319 , 86)
C4<-c(3, 3 , 251 , 239 , 572 , 170 , 4 , 297 , 283 , 569 , 125 , 4 , 300 , 286 , 574 , 117 , 6 , 312 , 295 , 536 , 136 , 24 , 319 , 303 , 539 , 113)
C5<-c(2 , 4 , 83 , 79 , 695 , 208 , 6 , 99 , 96 , 726 , 158 , 3 , 105 , 104 , 741 , 144 , 4 , 111 , 107 , 698 , 176 , 19 , 118 , 116 , 688 , 165)
C6<-c(1, 3 , 35 , 33 , 765 , 185 , 3 , 41 , 40 , 800 , 141 , 2 , 39 , 38 , 800 , 145 , 5 , 48 , 47 , 762 , 165 , 14 , 54 , 54 , 731 , 179)
C7<-c(0 , 2 , 14 , 12 , 785 , 187 , 3 , 20 , 18 , 812 , 153 , 3 , 17 , 15 , 825 , 145 , 7 , 23 , 18 , 778 , 174 , 10 , 28 , 25 , 762 , 177)

CC<-rbind(C1,C2,C3, C4, C5, C6, C7)
colnames(CC)<-c("NC", " F1", " F2", " F3", "F4", "F5","Q1", "Q2", "Q3", "Q4", "Q5","X1", "X2", "X3", "X4", "X5", "M1", "M2", "M3", "M4", "M5", "P1", "P2", "P3", "P4", "P5")
CCC<-CC[, -1]


## ----corr2, echo=FALSE, result="asis", fig.cap='Mosaicplot correct NOEAC rates', out.width='.68\\linewidth',fig.align='center', fig.pos="H", cache=TRUE, warning=FALSE----
```

```
CC2<-CCC[, c(-3,-8, -13, -18, -23)]
mosaicplot(CC2, main="NOAEC", dir=c("h", "v"),las=1)


## ----Lehmann1, echo=FALSE, result="asis", cache=TRUE, warning=FALSE----------------------------
# Ask Prof. Lehmann for the CPCAT test R-Code

## ----exaDaph1, echo=TRUE, result="asis", cache=FALSE, warning=FALSE, message=FALSE-------------
library(multcomp); library(mlt); library(sandwich);  library(lme4); library(merDeriv); library(ggplot2)
daphnia$FT<-sqrt(daphnia$Number_Young)+ sqrt(daphnia$Number_Young+1)
da1 <-lm(FT~Conc, data=daphnia)
da2 <-glm(Number_Young~Conc, data=daphnia,family=quasipoisson(link = "log"))
da3 <- glmer(Number_Young~Conc+ (1|Adults), data=daphnia, family=poisson(link="log"), verbose=FALSE); VCC<-vcov(da3)
yvar <- numeric_var("Number_Young", support = quantile(daphnia$Number_Young, prob = c(.001, .999)));
bstorder<-5 # order bernstein polynomial
    yb <- Bernstein_basis(yvar, ui = "increasing", order =bstorder)
    ma <- ctm(yb, shifting = ~ Conc, todistr = "Normal", data = daphnia)
    m_mlt<-mlt(ma, data = daphnia)
    K <- diag(length(coef(m_mlt)))
    rownames(K) <- names(coef(m_mlt))
    K <- K[-(1:6),] # 1:6 for order 5
      SID<-daphnia[, c(1,3)]
      SID$Poissondaten<-SID$Number_Young
      SID$Gruppenzugeh?rigkeit<-as.factor(SID$Conc)
      sid<-droplevels(SID[, c(3,4)])
      Lehp<-CPCAT(sid)
      duF<-summary(glht(da1, linfct = mcp(Conc = "Dunnett"), alternative="less", vcov = sandwich))
      duQ<-summary(glht(da2, linfct = mcp(Conc = "Dunnett"), alternative="less"))

    duMS<-summary(glht(da3, linfct = mcp(Conc ="Dunnett"), alternative="less",vcov=VCC))
        mlt<-summary(glht(m_mlt, linfct = K, alternative="greater"))
DUF<-fortify(duF);  DUQ<-fortify(duQ); DUMS<-fortify(duMS); MLT<-fortify(mlt)
pvals<-cbind(DUF [, c(5,6)], DUQ[, c(5,6)], DUMS[, c(5,6)], MLT[, c(5,6)], Lehp)
colnames(pvals)<-c("FT(t)", "FT(p)", "QP(t)", "QB(p)", "MM(t)", "MM(p)", "MLT(t)", "MLT(p)", "Leh(p")
rownames(pvals)<-c("1.56-0", "3.12-0", "6.25-0", "12.5-0", "25-0")
PVALS<-as.data.frame(pvals)


## ----bsrf1, echo=FALSE, results='asis', message=FALSE, warning=FALSE----------------------------
library(xtable)
tabroo<-xtable(PVALS, digits=c(0,2,6, 2,6, 2,6, 2,6, 6))
print(tabroo, size="small")


## ----danh32a, echo=TRUE, result="asis", cache=TRUE, warning=FALSE------------------------------
library("tram")
mnd<-mn[mn$group!="Positive", ]
COC<-Colr(MN~group, data=mnd) # COLR
CC<-glht(COC, linfct = diag(length(coef(COC))))
ccMLT<-1/exp(confint(CC)$confint)# OR and sCI


## ----danh32, echo=FALSE, result="asis", fig.cap='Confidence intervals for odds ratios',
out.width='.38\\linewidth',fig.align='center', fig.pos="H", cache=TRUE, warning=FALSE----
library(MCPAN)
plotCII(estimate=ccMLT[, 1], lower=ccMLT[, 3],  alternative="greater", lines=c(0,1))
```

References

[1] *Short-term Methods for Estimating the Chronic Toxicity of Effluents and Receiving Waters to Freshwater Organisms, Fourth Edition; U.S. Environmental Protection Agency Office of Water (4303T), Washington, DC 20460,EPA-821-R-02-013; Section 13: Test methods Daphnid, Ceriodaphnia Dubia, Survival and Reproduction Test Methods 1002.0; Table 4.*

[2] V. Amrhein, S. Greenland, and B. McShane. Retire statistical significance. *Nature*, 567(7748):305–307, March 2019.

[3] F. Bretz, J. C. Pinheiro, and M. Branson. Combining multiple comparisons and modeling techniques in dose-response studies. *Biometrics*, 61(3):738–748, September 2005.

[4] C. W. Dunnett. A multiple comparison procedure for comparing several treatments with a control. *J Am Stat Assoc*, 50(272):1096–1121, 1955.

[5] G. Engelhardt. In vivo micronucleus test in mice with 1-phenylethanol. *Archives Of Toxicology*, 80(12):868–872, December 2006.

[6] Hayes et al. Report on statistical issues related to oecd test guidelines (tgs) on genotoxicity. Technical report, OECDNo. 198, 2014.

[7] L. Hothorn and W. Lehmacher. A simple testing procedure control versus k treatments for one-sided ordered-alternatives, with application in toxicology. *Biometrical Journal*, 33(2):179–189, 1991.

[8] L. A. Hothorn. Multiple comparisons and multiple contrasts in randomized dose-response trials-confidence interval oriented approaches. *Journal of Biopharmaceutical Statistics*, 16(5):711–731, 2006.

[9] L. A. Hothorn. How to deal with multiple treatment or dose groups in randomized clinical trials? *Fundam Clin Pharm*, 21(2):137–154, 2007.

[10] L. A. Hothorn, M. Neuhauser, and H. F. Koch. Analysis of randomized dose-finding-studies: Closure test modifications based on multiple contrast tests. *Biometrical Journal*, 39(4):467–479, 1997.

[11] L. A. Hothorn, K. Reisinger, T. Wolf, A. Poth, D. Fieblingere, M. Liebsch, and R. Pirow. Statistical analysis of the hen's egg test for micronucleus induction (het-mn assay). *Mutation Research-Genetic Toxicology and Environmental Mutagenesis*, 757(1):68–78, September 2013.

[12] L.A. Hothorn. *Statistics in Toxicoloyg- using R.* Chapman Hall, 2016.

[13] L.A. Hothorn and D Hauschke. Identifying the maximum safe dose: a multiple testing approach. *J Biopharm Statist.*, 2000.

[14] L.A. Hothorn and F. Kluxen. Robust multiple comparisons against a control group with application in toxicology. *arXiv:1905.01838*, 2019.

[15] L.A. Hothorn and R. Pirow. Use compatibility intervals in regulatory toxicology. *Reg. Tox. Pharm*, 2020(submitted).

[16] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical J*, 50(3):346–363, 2008.

[17] T. Hothorn, L. Most, and P. Buhlmann. Most likely transformations. *Scandinavian Journal of Statistics*, 45(1):110–134, March 2018.

[18] A. Huang. Mean-parametrized conway–maxwell–poisson regression models for dispersed counts. *Statistical Modelling*, 2017.

[19] T. Jaki, A Kitsche, and L. A Hothorn. Statistical evaluation of toxicological assays with zero or near-to-zero proportions or counts in the concurrent negative control group: A tutorial. *JP J Biostatistics*, 0:0, 2014.

[20] F. Konietschke and L. A. Hothorn. Evaluation of toxicological studies using a nonparametric shirley-type trend test for comparing several dose levels with a control group. *Statistics in Biopharmaceutical Research*, 4(1):14–27, 2012.

[21] R. M. Kuiper, D. Gerhard, and L. A. Hothorn. Identification of the minimum effective dose for normally distributed endpoints using a model selection approach. *Statistics in Biopharmaceutical Research*, 6(1):55–66, February 2014.

[22] R. Lehmann, J. Bachmann, B. Karaoglan, J. Lacker, G. Lurman, C. Polleichtner, H. T. Ratte, and M. Ratte. The cpcat as a novel tool to overcome the shortcomings of noec/loec statistics in ecotoxicology: a simulation study to evaluate the statistical power. *Environmental Sciences Europe*, 30:50, December 2018.

[23] R. Lehmann, J. Bachmann, D. Maletzki, C. Polleichtner, H. T. Ratte, and M. Ratte. A new approach to overcome shortcomings with multiple testing of reproduction data in ecotoxicology. *Stochastic Environmental Research and Risk Assessment*, 30(3):871–882, March 2016.

[24] T Lohse, S Rohrmann, D Faeh, and T Hothorn. Continuous outcome logistic regression for analyzing body mass index distributions. *F1000Research*, 2017, 6:1933 (doi: 10.12688/f1000research.12934.1).

[25] B. H. Margolin, N. Kaplan, and E. Zeiger. Statistical analysis of the Ames Salmonella-Microsome test. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences*, 78(6):3779–3783, 1981.

[26] M. Otava, Z. Shkedy, L. A. Hothorn, W. Talloen, D. Gerhard, and A. Kasim. Identification of the minimum effective dose for normally distributed data using a bayesian variable selection approach. *Journal of Biopharmaceutical Statistics*, 27(6):1073–1088, 2017.

[27] P. Pallmann and L. A. Hothorn. Boxplots for grouped and clustered data in toxicology. *Archives of Toxicology*, 90(7):1631–1638, July 2016.

[28] S. J. RUBERG. Contrasts for identifying the minimum effective dose. *Journal of the American Statistical Association*, 84(407):816–822, September 1989.

[29] F. Schaarschmidt and L.A. Hothorn. library(tukeytrend). 2018.

[30] S. Siegfried and T. Hothorn. Count transformation models: The cotram package.

[31] E. Sonnemann. General solutions to multiple testing problems. *Biometrical Journal*, 50(5, SI):641–656, OCT 2008.

[32] A. C. Tamhane, C. W. Dunnett, J. W. Green, and J. D. Wetherington. Multiple test procedures for identifying the maximum safe dose. *Journal of the American Statistical Association*, 96(455):835–843, September 2001.

[33] A. C. Tamhane and B. R. Logan. Multiple test procedures for identifying the minimum effective and maximum safe doses of a drug. *Journal of the American Statistical Association*, 97(457):293–301, March 2002.

[34] K. Toyota, N. A. McNabb, D. D. Spyropoulos, T. Iguchi, and S. Kohno. Toxic effects of chemical dispersant corexit 9500 on water flea daphnia magna. *Journal of Applied Toxicology*, 37(2):201–206, February 2017.

[35] T. Wang and E. C. Merkle. merderiv: Derivative computations for linear mixed effects models with application to robust standard errors. *Journal of Statistical Software*, 87(CN1):1–16, November 2018.

[36] R. L. Wasserstein, A. L. Schirm, and N. A. Lazar. Moving to a world beyond "p ¡ 0.05". *American Statistician*, 73:1–19, 2019.

[37] D.A. Williams. A test for differences between treatment means when several dose levels are compared with a zero dose control. *Biometrics*, 27(1):103–117, 1971.

[38] M. J. Wolfsegger, G. Gutjahr, W. Engl, and T. Jaki. A hybrid method to estimate the minimum effective dose for monotone and non-monotone dose-response relationships. *Biometrics*, 70(1):103–109, March 2014.

[39] A. Zeileis. Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9), August 2006.