

Project Number 289706 Start date of the project: 01/12/2011, duration: 48 months

Deliverable 9.2

Environmental Risk Assessment of Genetically Modified Organisms: Statistical aspects of a protocol for singleenvironment GMO field studies based on a simulation study to investigate properties of difference and equivalence tests

Authors:

Hilko van der Voet and Paul W. Goedhart

Organisation names of lead contractor for this deliverable:

DLO: Wageningen University and Research centre, Plant Research International, Biometris, Wageningen, Netherlands <u>http://www.biometris.nl</u>

March 2014

Dissemination Level: Public

Contents

Abstract	.4
----------	----

D9.2a Statistical aspects of a protocol for single-environment GMO field studies

1	Introduction				
	1.1 Background	6			
	1.2 Relation to work program and overview	7			
	1.3 Models simulating field trial data	7			
	1.4 Statistical analysis methods	8			
	1.5 Power analysis for difference and equivalence testing	8			
2	Summary of conclusions from a simulation study with count data	9			
	2.1 Setup of study	9			
	2.2 Results for the Difference test	9			
	2.3 Results for the Equivalence testing	0			
	2.4 Conclusion from simulation study	0			
3	Statistical elements for a protocol for experimental design and prospective power analysis				
	11				
4	Protocol for statistical analysis	3			
5	Statistical analysis examples	5			
	5.1 One- and two-sided difference and equivalence tests	.5			
	5.2 Difference and equivalence tests on a preliminary AMIGA potato data	.7			
6	References	.9			

D 9.2b Simulation study to investigate properties of difference and equivalence tests

1	ntroduction	22
	.1 Statistical analysis and design for environmental risk assessment	22
	.2 A protocol for the design and analysis of single-environment field trials – Task 9	.3. 23
	.3 Overview of this report	23
2	Setup of simulation study	24
	2.1 Basic setup and simulation distributions	24
	2.2 Parameter values used in the simulation	26
	2.3 Statistical models for analysis	27
3	Results for difference testing	32
	3.1 General remarks on difference testing	32

	3.2	Scaling of the deviance difference for OP, P1, P2, P3 and GM	33
	3.3	Simulated significance level of difference test	36
	3.4	Power of difference test	41
	3.5	Coverage of confidence intervals	45
	3.6	Approximate power of the difference test	46
	3.7	Method of choice for difference test	48
4	Res	ults for equivalence testing	49
	4.1	General remarks on equivalence testing	49
	4.2	Size of equivalence test	49
	4.3	Power of equivalence test	53
	4.4	Approximate power of the equivalence test	53
	4.5	Method of choice for equivalence testing	54
5	Zero	o inflation	55
	5.1	Introduction	55
	5.2	A zero-inflated negative binomial distribution and its non-inflated counterpart	55
	5.3	Size of the LN difference test when there is zero inflation	56
	5.4	Power of the NB difference test for negative binomial data	57
6	Con	clusion	59
7	Refe	erences	60
8	App	pendices	62

Abstract

One of the aims of the EU project AMIGA (Assessing and monitoring the impacts of genetically modified plants on agro-ecosystems) is to provide protocols on how to perform field studies, and on how to analyse data obtained from such studies. Workpackage 9 of AMIGA works on statistical methods relevant for evaluation of non-target effects, and the current deliverable proposes elements for protocols of experimental design and statistical analysis.

For the most common data type in ecological field studies, i.e. count data, a large simulation study was conducted, including multiple ways of simulating count data, and multiple ways of statistical analysis. Four different count distributions were used to simulate count data for a mean count ranging from 0.5 (for rare species) to 100 (for more common species). Different coefficients of variation and different levels of replication, ranging from 4 to 100, were used to simulate data. The ratio of the means of the GM plant and its comparator was set to 1, 0.75, 0.50 and 0.25. A ratio of 1 implies no difference between the GM plant and its comparator. The simulated data were analysed by means of eight different models, such that the most robust model could be selected. Results for difference testing are the simulated size and power of the difference test as well as coverage of confidence intervals. We also describe an approximate fast method to obtain the power of a difference test. A recommendation is given about which difference test is to be preferred. Results for one-sided equivalence testing are the simulated significance level of various methods, the simulated power and a fast way of calculating the power. This also results in a recommendation about which equivalence test is to be preferred. There is a discussion on the problem of zero inflation, i.e. when there are more zeros than predicted by the count distribution.

Based on the results of this simulation study a checklist is proposed regarding the methodology to perform for a prospective power analysis to guide experimental design. Further a protocol is proposed on how to conduct the statistical analysis for both difference tests and equivalence tests. The analytical protocol is summarised in a flow chart. Some simple examples are given. The next step in the AMIGA project will be the implementation of statistical methods in user-friendly software.

The work for Deliverable D9.2 of the AMIGA project is reported in two parts:

- D9.2a describes statistical aspects of a protocol for single-environment GMO field studies.
- D9.2b describes a simulation study to investigate properties of difference and equivalence tests.



Project Number 289706 Start date of the project: 01/12/2011, duration: 48 months

Deliverable 9.2a

Environmental Risk Assessment of Genetically Modified Organisms: Statistical aspects of a protocol for singleenvironment GMO field studies

Authors:

Hilko van der Voet, Paul W. Goedhart

Organisation names of lead contractor for this deliverable:

DLO: Wageningen University and Research centre, Plant Research International, Biometris, Wageningen, Netherlands <u>http://www.biometris.nl</u>

March 2014

Dissemination Level: Public

1 Introduction

1.1 Background

The EFSA Guidance on the environmental risk assessment (ERA) of genetically modified (GM) plants (EFSA 2010) gives broad guidance on the design and analysis of field experiments. The AMIGA research project aims at providing more detailed guidance in the form of *protocols for design and analysis*. This report provides statistical elements for such protocols.

Recently, several papers have been published which aim at providing statistical guidance for ERA field experiments. Perry et al. (2009) noted that the null hypothesis of a GM risk assessment test should be that of non-equivalence, at a level of difference between the GM plant and its comparator which was termed the *Limit of Concern (LoC)*, and which have to be set before the experiments. The LoC was defined as the minimum relevant ecological effect that is deemed biologically significant, and is deemed of sufficient magnitude to cause harm. In Food-feed risk assessment a procedure has been developed to derive LoCs from the variation between reference varieties in the same field trials (van der Voet et al. 2011). Perry et al. (2009) found this approach less appropriate for ERA, but also noted that experience suggests that the direct setting of LoCs is more feasible in ERA than in food-feed risk assessment. The need for a *prospective power analysis* based on the LoC or other treatment effect sizes of interest was stressed.

Goedhart et al. (2013, 2014) summarised *statistical models* that could be useful in the analysis of ERA field experiments. For count data the Poisson distribution is the basic distribution, but it was noted that over-dispersion and/or excess zeroes imply the need for more advanced distributions, such as the over-dispersed Poisson, negative binomial, or Poisson-Lognormal distribution. Similarly, for quantal data, the basic binomial distribution should be replaced by a beta-binomial or binomial-logitnormal distribution. For both count and quantal data, excess zeroes could be handled assuming an additional spike of structural zero results in addition to the other data (which may still contain incidental zero values). Such models can be analysed directly (mixture models) or in a 2-step procedure (hurdle models). Goedhart et al. (2013, 2014) provided a *simulation tool* to generate dummy field trial datasets based on any of these distributions. For the analysis of such data estimation methods based on knowing the right model behind the data can be used. Alternatively, in spite of all the complex modelling options, a simple data transformation followed by normal-theory modelling is also an option for analysis, and it is what is most commonly used in practice.

Semenov et al. (2013) reiterated the same ideas of prospective power analysis, equivalence testing, and choosing statistical models for counts and quantal data. They provided some decision trees and a checklist to assist the interpretation of statistical analyses of field trials.

More specific guidance on sample size calculation for ERA field trials has been given by Perry et al. (2003), Prasifka et al. (2008) and Comas et al. (2013). These studies, however, arrive at very different conclusions. For a twofold change in non-target counts (i.e. +100% or -50%) Perry et al. (2003, Table 6) conclude that 60 replicates provide a power of at least 85%, provided count levels are \geq 5, and the coefficient of variation is \leq 100%. In contrast, for a similar power to detect a -50% change Prasifka et al. (2008, Figures 1-4) found that on average less than 6 replicates would be sufficient in their datasets. Comas et al. (2013, Table 2) seem to need only 3 replicates to attain, with power 80%, an expected

capacity of field tests that not exceeds impacts of 100% relative to the comparator's mean. As another example, the study of Perry et al. (2003) suggests that in many cases it will be very difficult to detect impacts of around 30% with sufficient power, whereas the other two studies suggest this is easily possible in many cases. The differences between the reported studies were analysed and will be reported in a subsequent paper. Here we complement them with results from new simulation studies, which focus on the robustness of results under different models of the reality.

1.2 Relation to work program and overview

This report (D9.2a) and the companion report (D9.2b) describes research in AMIGA Work Package 9, Task 3. This task focuses on *single-environment trials*, and is preparatory for Tasks 9.4 and 9.5 where multi-environment trials are addressed. Several statistical issues regarding data types, equivalence testing and test characteristics can however be better researched in the relatively simple situation of a single-environment trial. This is also relevant because of the emphasis of the EFSA guidance document on single-environment trials. The EFSA guidance document states that "*For field trials, since each field trial at a site on a particular occasion should have sufficient replication to be able to yield a stand-alone analysis if required, this power analysis should relate to a single site"*. Therefore protocols for power analysis and statistical analysis of a single field trial have been developed in this task.

We investigated the applicability of linear models (LM) based on normal distributions for transformed variables relative to generalized linear models (GLM) for typical ERA data. This subtask involves *robustness studies*, e.g. simulating counts according to a range of distributions and analysing the resulting data using a range of analysis methods.

The report describes the development of *protocols for setting sample sizes* in experimental design based on the desired performance of difference and equivalence tests.

The *experimental design protocol* includes a *checklist* which enables a risk assessor to provide full information on the study, for example a list of endpoints and why they are chosen, a description of the chosen experimental design with justification in terms of power, and the sampling strategy. The *statistical analysis protocol* is summarised in a flow chart. This includes justification of the distributional assumptions and the robustness of such assumptions, the generic form of the analysis and why it was chosen, criteria for identification of outliers, the way in which difference and equivalence testing is performed, and the way in which the results of the analysis should be presented. In the further development the protocol will be accompanied by *software* for performing power analysis, for fitting the statistical models and for reporting and displaying the results of the analysis.

1.3 Models simulating field trial data

In Goedhart et al. (2013) four statistical models have been described for simulation of count data: Poisson (P), Over-dispersed Poisson (OP), Negative Binomial (NB), and Poisson-Lognormal (PL). In addition, in this report we also include the Taylor power law model, which has been found to provide adequate descriptions of practical data, with powers often between 1 and 2 (Taylor et al. 1978). The power law only specifies a relation between variance and mean ($V = a\mu^p$), therefore in the simulations a negative binomial distribution was used to generate data.

For simulation of presence/absence data Goedhart et al. (2013) described three statistical models: Binomial (B), Beta-Binomial (BB) and Binomial-Logitnormal (BL).

1.4 Statistical analysis methods

For statistical analysis of data many methods are available. A first possibility is a maximum likelihood analysis corresponding to exactly the simulation model. But it is often convenient to use simple approximation methods, e.g.

- a normal-distribution analysis of transformed data (e.g. log- or square root transform for nonnegative or count data, empirical logit transform for presence/absence data),
- a quasi-likelihood analysis to address over-dispersion,
- an analysis based on a two-part (hurdle) model rather than a mixture distribution.
- an analysis of presence/absence derived from count data (all counts > 0 are reset to 1)

Several method were investigated in a simulation study (see companion report D9.2b and the summary of conclusions in Section 2 of this report). In the protocol of the current report the focus is on two categories of analysis:

- 1. linear models (assuming a normal distribution of errors) after an appropriate transformation of the data;
- 2. generalized linear models (GLMs), which specify transformations for the expected values rather than the data (McCullagh and Nelder, 1989).

1.5 Power analysis for difference and equivalence testing

The parameter of interest in tests is the ratio Q between expected counts for the GMO and the comparator (CMP), or equivalently, the difference D between the log-counts. For a power analysis of the difference test a range of alternative values for D has to be specified.

In an equivalence test the null hypothesis is Q = LoC. For a power analysis of the equivalence test a range of alternative values for Q has to be specified, between 1 and LoC (if LoC>1), between LoC and 1 (if LoC<1), or between a lowe LoC and an upper LoC (if there are concerns in both directions).

For a non-equivalence test the null hypothesis is also D = LoC. But now the alternative values for Q are those above the LoC (if LoC>1), below the LoC (if LoC<1), or both (if there are concerns in both directions).

2 Summary of conclusions from a simulation study with count data

The simulation study is fully described in the companion report on Deliverable 9.2b. here we present a short summary of the setup and conclusions.

2.1 Setup of study

Different models were used for simulating count data, and different methods for analysing the generated data. Almost all datasets show overdispersion in practice, therefore only models allowing for overdispersion were used in the simulation study. The Poisson model itself was not used.

Abbreviation	Description	used for	used for
		simulation	analysis
OP	overdispersed Poisson	Х	Х
NB	negative binomial	Х	Х
PL	Poisson-Lognormal	Х	
P1	Power model with p=1.5	Х	Х
P2	Power model with p=1.7		Х
P3	Power model with p=1.99		Х
GM	Gamma		Х
LN	Log(y+1) transformation		Х
SQ	Sqrt(y) transformation		Х

Table 1. Models and transformations used in simulating and analysing count data

The parameters used in the simulation were varied across a range of means and CV values, as described in the simulation report D9.2b. Data for GMO and comparator counts were simulated under effect sizes (ratios GMO to comparator) 1, 0.75, 0.5 and 0.25, thus focussing on negative effects of the GMO.

Sizes and powers of tests were calculated by repeated simulation of data. In addition, estimation of sizes and powers with the method of Lyles et al. (2007) which does not need the repeated simulation was investigated.

2.2 Results for the Difference test

The main results from the simulation study for the difference test were:

- Comparison of size of difference test for OP, P1 and GM when the LR test statistics is scaled by Pearson's Chi-squared or by the mean deviance
 - scaling by Pearson has somewhat better properties
- 2. Comparison of size of test for LN, SQ, OP, NB, P1, P2, GM
 - LN has generally the best properties
- 3. Power of test for LN, SQ, NB, P1, P2, GM for those settings for which the size is OK
 - > LN has the same power as OP when simulating according to OP
 - LN has marginally smaller power than NB in some case when simulating according to NB; however size of NB is frequently not satisfactory
 - > LN is at least as good as other models when simulating according to PL
 - > LN is at least as good as other models when simulating according to P1
 - > LN is the method of choice for difference testing

- 4. Properties of the (back-transformed) generalized confidence interval, i.e. coverage probabilities, for the LN analysis are identical to those of the t-test. However this is only true for properties under the null-hypothesis of equal means. Coverage of the LN interval deteriorates when the quotient of the two means differs more strongly from one, and when the CV increases
 - > The LN interval approach can be used for difference testing; apparently it cannot always be used for equivalence testing
- 5. The method of Lyles *et al* (2007), using a synthetic dataset, can be used to perform a prospective power analysis for the LN analysis; this is in very good agreement with the simulated power
 - > There is no need to perform a simulation study for a prospective power analysis in the simple situation of a GMO and a comparator.

2.3 Results for the Equivalence testing

The main results from the simulation study for the equivalence tests were:

- 6. Results are based on the estimate of the log(ratio) and its standard error (scaled by Pearson) for GLM-like analysis methods, and on the Generalized CI for the LN and SQ analyses.
- 7. Comparison of size of one-sided equivalence test for LN, SQ, OP, NB, P1, P2, GM for effect sizes 0.75, 0.5 and 0.25. The null-hypothesis is then H_0 : mu1/mu2 \leq effectsize
 - Size of LN is generally bad (conservative as well as progressive)
 - Size of OP seems to be best across the board. However conservative for small means, small levels of replications and large CV values. Occasionally somewhat progressive.
- 8. For effect sizes 0, 0.75 and 0.5, and hypothetical one-sided LOC of 0.5 he power of EQ test is very similar for OP, NB, P1, P2 and GM. Also the probability of "Equivalent more likely than not" is very similar.
- 9. For effect size 0.50 one would expect a probability of 50% for "Equivalent more likely than not". This is generally the case, except for small means combined with small levels of replication.
 - An OP based confidence interval can best be used for equivalence testing. This interval does not always have the correct size.
- 10. The method of Lyles *et al* (2007) can also be used to approximate the power of the one-sided equivalence test using OP. The approximation is less good than for the difference test; it is however good enough as a first approximation especially for larger power values around 0.8. The approximation is not good for data simulated with PL and large CVs possibly because the PL distribution is then very un-similar to the overdispersed Poisson with the same CV.
 - > There is not always a need to perform a simulation study for a prospective power analysis in the simple situation of a GMO and a comparator.

2.4 Conclusion from simulation study

Difference testing for count data can best be done by an LN analysis; based on this analysis a generalized CI can be constructed on the original scale. The method of Lyles *et al* (2007) can be used to approximate the power of this test.

Equivalence testing for count data can best be done by constructing a CI after an OP analysis. This procedure does not have perfect properties. When simulating according to OP, NB or P1 (all using different variants of the negative binomial distribution) the method of Lyles et al. can be used to approximate the power of the one-sided equivalence test.

3 Statistical elements for a protocol for experimental design and prospective power analysis

Attention is required before a field trial is performed to ensure that the experiment will be meaningful to answer research questions. We present relevant points from a statistical viewpoint as a checklist.

Checklist

- 1. Describe all the **questions** the experiment is meant to answer, in words.
- 2. Prepare the **list of endpoints**. This may be divided into a list of primary endpoints (with strict requirements regarding power of tests) and a list of secondary endpoints.
- 3. For each endpoint classify the **measurement type**, e.g. non-negative continuous data, count data or fractions (percentage) data.
- 4. For each primary endpoint to be tested formulate the Limits of Concern (LOCs). For each endpoint one lower and/or one upper LOCs can be set. For non-negative continuous and count data these will typically be ratios of GMO divided by CMP true values. For percentage data ... Make explicit whether equivalence has to be proven (in a formal test at the set significance level) or that it is sufficient to show 'equivalence more likely than not'.
- 5. Describe the research questions in the form of **null hypotheses**, both for difference and equivalence tests.
- 6. Set the **significance levels** (α) for statistical testing. Conventionally the level (size) will be e.g. 0.05. In the TOST approach to equivalence testing (Schuirmann 1987) the significance level for the difference test is twice the significance level for the equivalence test.
- 7. Set the **required power** of the tests to detect differences at specified effect sizes. Typically these effect sizes will be equal to the LoC. Conventional values for power are between 70 and 90%. If equivalence has to be proven, formulate effect sizes for which equivalence would need to be proven using the equivalence test with pre-defined power (e.g. 80% power to proof equivalence at an effect size of 0.75 (-25%) given an LoC of 0.5 (-50%).
- 8. Describe the structure of the proposed **experimental design**, e.g. completely randomized, randomized block, split-plot, incomplete balanced block.
- **9.** Describe the **experimental units** (typically plots or sub-plots), and give details of the **blocking structure** (e.g. 4 main plots per randomized block, each split into 3 sub-plots) and the **treatment structure** (e.g. three types of spraying and four crop varieties). Also describe if interactions should be included.
- 10. Describe whether **repeated measurements** will be taken from the same experimental unit.
- 11. Provide a **model formula** partly specifying how the data will be analysed, using the syntax of one of the common software tools for statistical analysis (SAS, GenStat, R, ...), for example *block/plot/subplot + treatment + variety*. Include terms and a correlation structure for repeated measurements if used. Indicate which factors are random rather than fixed.
- 12. For each primary endpoint provide **prior estimates of central value and variation** for a measurement on one experimental unit. For non-negative continuous and count data the prior estimates for central values will typically be expected values or geometric means, and the prior estimates for variation will typically be coefficients of variation. Such values can be derived from previous experiments or based on expert knowledge.
- 13. For each endpoint specify the simplest **statistical analysis method** that will be used (unless there are unexpected deviations in the execution of the field study or unexpected data). See the statistical analysis protocol for details.
- 14. Based on the replication and the prior estimates **estimate the power of the proposed design as a function of replication,** for the difference test, and if needed also for the equivalence test. In simple cases this can be performed using analytical formulae, in more complex cases this can be found in published results of simulation studies such as performed in the AMIGA project. If not available, a new simulation can be performed to estimate the power.

- 15. From the power curves derive the **replication** of the comparison of GMO to CMP in the proposed design.
- 16. If the calculated minimal replication cannot be realized in practice, the **power is insufficient**. In such case adapt the design or reformulate the research questions.
- 17. Randomise the treatments over the experimental units taking proper account of the design.

4 Protocol for statistical analysis

1. The method of statistical analysis depends on the type of endpoint. For typical ecological endpoints it is recommended to perform both an analysis based on data transformation and normality, and an analysis on the original scale using an appropriate link function.

Endpoint type	data transformation ¹	distribution and link function for GLN	
Positive continuous x	log(x)	gamma, log	
Non-negative continuous x	$log(x+m)$, where $m \le min(x_{+})$	gamma, log	
Positive counts x	log(x)	over-dispersed Poisson, log	
Counts x	log(x+1)	over-dispersed Poisson, log	
Fractions 0 < x/n < 1	logit(x) = log[(x)/(n-x)]	over-dispersed binomial, logit	
Fractions x/n	log[(x+0.5)/(n-x+0.5)]	over-dispersed binomial. logit	

Table 2. Recommended data transformations and GLMs

¹ For data transformation any base of logarithm can be chosen as is considered convenient, e.g. 2, e or 10. Note that the GLM link functions will use the natural logarithm (log_e).

- 2. Analyse the transformed data by linear models: ANOVA if the design is balanced, or by a mixed model (REML) if they are not.
- 3. Analyse the untransformed data by generalized linear models (GLM), or by a generalized linear mixed models (GLMM) is there are additional stochastic terms in the model. Allow for over-dispersion in counts and fractions.
- 4. Check the reasonableness of statistical assumptions, e.g. as follows:
 - a. Outliers: check data points with large standardised residuals. Compare analyses with and without such data points in a sensitivity analysis.
 - b. QQ plot should show approximately a straight line
 - c. Plot residuals vs. fitted values can be used to check if there is heteroscedasticity.
- 5. If statistical assumptions are unreasonable, then an ad-hoc strategy will have to be followed. For example, non-parametric tests may be used. This protocol continues assuming that the model fits sufficiently well.
- 6. From the ANOVA or REML results find estimators of the mean and standard errors of the mean for GMO and CMP. From these distributions back-transform to distributions for the means of GMO and CMP on the original scale (method, see D9.2b report).
- 7. From these back-transformed distributions create a distribution of the ratio GMO vs. CMP, and from this find the generalized confidence limits as 2.5% and 97.5% points for two-sided difference tests, or as 5% and 95% points for two one-sided difference tests. (Note: for visual display it is recommended to calculate and display both limits, even if the test is one-sided.)
- 8. From the GLMM or GLM analysis find the best estimator of the mean, and 5% and 95% confidence limits by a profile likelihood method (see D9.2b report). Back-transform the estimate and the limits by the inverse link function. (Note: for visual display it is recommended to calculate and display both limits, even if the test is one-sided.)
- 9. For each endpoint, plot point estimates and intervals, together with lines for the equality ratio 1, and the LoCs. In most cases plots on a logarithmic scale are advised. Use a recognizable symbol (e.g. an arrowhead) for interval endpoints that represent two one-sided tests (TOST).
- 10. Use the intervals based on the linear models for the difference tests
- 11. Use the intervals based on the generalized linear models for the equivalence tests.



Figure 1. Flow chart to guide statistical analysis (updated from Semenov et al. 2013).

5 Statistical analysis examples

5.1 One- and two-sided difference and equivalence tests

Here we show an example for a situation where there is a concern about decreased levels of a counted organism. We assume that the Limit of Concern has been set to 0.5 for three endpoints, i.e. there is ecological concern if the count level in the GMO plots would be 50% or less of the level in the CMP plots. There is no concern about increased levels. We assume a testing confidence level of 95% throughout.

Results were obtained from programs in GenStat (VSN 2012). In Figure 2 we show data and results for three endpoints, each with 2 x 20 counts (10 for CMP and 10 for GMO). In the joint graph the intervals for difference testing and for equivalence testing are shown together for each endpoint. Note that both intervals have arrowheads indicated that they represent two one-sided tests (TOST). This simply means that these intervals are intended to cover 90% (rather than 95%), with 5% probability of a true ratio below the lower endpoint and 5% probability of a true ratio above the upper endpoint.

For the one-sided difference test the upper limit can be compared to the ratio value 1 (which represents the null hypothesis of equality). In this example the GMO is not significantly different from the CMP for endpoint A, but it is for endpoints B and C. The P values for the one-sided difference test are indicated next to the relevant interval upper limit in the graph.

For the one-sided equivalence test the lower limit can be compared to the ratio value 0.5 (which represents the null hypothesis of border-line non-equivalence). In this example the GMO is equivalent to the CMP for endpoints A and B, but it is non-equivalent more likely than not for endpoint C. The P values for the one-sided equivalence test are indicated next to the relevant interval lower limit in the graph for endpoints A and B. For endpoint C the point estimate is already lower than the LOC, therefore the result of a non-equivalence test is shown. In this case the non-equivalence is not significant, hence the resulting classification as 'non-equivalence more likely than not'.

It can be observed that in this case the two types of interval are reasonably similar, and the same conclusions would have been obtained if only one type of interval had been used for both the difference and the equivalence tests.

In Figure 3 the same data are analysed under a setting of two-sided concern. For the chosen examples the observed ratios are 1 or less, so there is no indication from the data for an increase. The difference intervals now are 95% rather than 90% intervals (and therefore slightly wider), and the P value for the difference test is approximately double the one-sided P value for these endpoints. This is the normal difference between on- and two-sided testing. Note, however, that the equivalence and non-equivalence tests are not influenced (the additional tests w.r.t. the upper LOC are performed, but are irrelevant for these data).



Figure 2. Three examples of count data (n=20) where there is concern for a decreased level. Limit of Concern (LOC) is 0.5 (GMO 50% of CMP, red vertical line). Bi-directed arrows represent 95% confidence intervals corresponding with two one-sided tests (TOST). P values are shown near the arrowheads for the one-sided difference (D) test (black) and the one-sided equivalence (E) or non-equivalence (NE) test (red) that is relevant for LOC<1.

(A) Not significantly decreased and equivalent;

(B) Significantly decreased and equivalent;

(C) Significantly decreased and non-equivalence more likely than not.



Figure 3. Same example as in Figure 2, but now with concern for decrease and increase, and two Limits of Concern, at ratios 0.5 and 2. Bars at the end of the difference interval indicate that this is a two-sided interval. The equivalence TOST interval is unchanged.

5.2 Difference and equivalence tests on a preliminary AMIGA potato data

Counts of non-target organisms were made in a field experiment with three potato varieties, performed in 2013 in Valthermond, the Netherlands, as part of the AMIGA project. Preliminary data (see Table 3) for two of the varieties (a GMO and a comparator) were analysed by the statistical methods proposed in this report.

block	variety	treatment	Predators	Detrivores	Parasitoids	Fungivores	Herbivores
6	CMP	IPM	14	3	5	0	18
3	CMP	IPM	12	10	10	1	23
2	CMP	IPM	19	6	8	0	28
5	CMP	IPM	20	9	7	0	16
7	CMP	IPM	17	6	2	0	21
1	CMP	IPM	16	6	6	0	17
4	CMP	IPM	8	5	1	0	11
6	CMP	NoControl	6	4	4	0	25
5	CMP	NoControl	20	7	4	1	12
1	CMP	NoControl	33	12	8	1	43
4	CMP	NoControl	15	4	9	1	19
3	CMP	NoControl	13	7	4	0	13
7	CMP	NoControl	6	2	0	1	12
2	CMP	NoControl	21	13	13	0	13
4	CMP	WeeklySchedule	23	12	6	0	22
1	CMP	WeeklySchedule	36	6	8	0	35
6	CMP	WeeklySchedule	18	4	4	0	25
3	CMP	WeeklySchedule	15	7	3	0	17
2	CMP	WeeklySchedule	15	6	6	0	27
7	CMP	WeeklySchedule	25	13	10	1	17
5	CMP	WeeklySchedule	17	9	5	0	28
5	GMO	IPM	19	2	4	1	17
4	GMO	IPM	19	7	6	0	20
3	GMO	IPM	25	8	8	0	29
7	GMO	IPM	12	8	9	0	19
2	GMO	IPM	20	6	11	3	12
6	GMO	IPM	17	6	7	1	16
1	GMO	IPM	10	10	7	0	33
5	GMO	NoControl	8	6	6	0	27
2	GMO	NoControl	13	5	8	0	26
1	GMO	NoControl	15	9	5	1	24
7	GMO	NoControl	11	4	1	0	10
6	GMO	NoControl	8	8	4	0	11
4	GMO	NoControl	15	7	3	0	34
3	GMO	NoControl	11	9	7	0	22
3	GMO	WeeklySchedule	12	13	9	1	39
1	GMO	WeeklySchedule	19	12	11	1	34
6	GMO	WeeklySchedule	11	7	1	1	20
5	GMO	WeeklySchedule	13	5	1	0	16
4	GMO	WeeklySchedule	15	8	7	0	23
7	GMO	WeeklySchedule	13	6	6	0	18
2	GMO	WeeklySchedule	18	9	10	0	23

Table 3. AMIGA potato experiment Valthermond, July 2013. Counts per guild, preliminary data(data courtesy Jenny Lazebnik, Wageningen University).

The results are shown in Figure 4. No prior discussion was made on appropriate Limits of Concern, and these were set at 0.5 and 2 for illustration of the method. Two-sided difference tests were performed.

No significant differences were found between the GMO and the CMP. For four of the five guilds equivalence could be proven at the 95% confidence level. For the Fungivores guild the observed numbers were very low (see Table 3). Consequently interval are wider. Equivalence could not be proven, but is still more likely than not. Note that the P value is shown for the equivalence test w.r.t. the nearest LoC, i.e. LoC=0.5 for endpoint A (Predators), and LoC=2 for the other endpoints.

Regarding the methodology, the two intervals are more similar when the observed counts are higher (endpoints A and E) than when they are low (e.g. endpoint D).



Figure 4. Analysis count data sticky traps per guild. AMIGA potato experiment Valthermond, July 2013. (A) Predators, (B) Detrivores, (C) Parasitoids, (D) Fungivores, (E) Herbivores. Limits of Concern set to 0.5 and 2 for illustration of the method only.

6 References

- Comas J, Lumbierres B, Pons X, Albajes R (2013). Ex-ante determination of the capacity of field tests to detect effects of genetically modified corn on nontarget arthropods. Journal of Economic Entomology, 106(4), 1659-1668.
- EFSA (2010). EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. EFSA Journal, 8(11): 1879. [111 pp.], doi:10.2903/j.efsa.2010.1879.
- Goedhart PW, van der Voet H, Baldacchino F, Arpaia S (2013). Environmental Risk Assessment of Genetically Modified Organisms: Overview of field studies, examples of datasets, statistical models and a simulation tool. Deliverable 9.1, AMIGA project, project number 289706.
- Goedhart PW, van der Voet H, Baldacchino F, Arpaia S (2014, in press). A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. Ecology and Evolution.
- Lyles RH, Lin H-M & Williamson JM (2007). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. Statistics In Medicine, 26(7): 1632-1648.
- McCullagh P & Nelder JA (1989). Generalized Linear Models, second edition. Chapman and Hall. London.
- Perry JN, Rothery P, Clark SJ, Heard MS & Hawes C (2003). Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. Journal of Applied Ecology, 40: 17-31.
- Perry JN, ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F & van der Voet, H (2009). Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environmental Biosafety Research, 8: 65-78.
- Schuirmann DJ (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6): 657-680.
- VSN International (2012). GenStat for Windows 15th Edition. VSN International, Hemel Hempstead, United Kingdom. Web page: <u>www.GenStat.co.uk</u>.



Project Number 289706

Start date of the project: 01/12/2011, duration: 48 months

Deliverable 9.2b Protocol for difference and equivalence testing for ERA

Environmental Risk Assessment of Genetically Modified Organisms:

Simulation study to investigate properties of difference and equivalence tests

Authors:

Paul W. Goedhart, Hilko van der Voet

Organisation names of lead contractors for this deliverable:

DLO: Wageningen University and Research centre, Plant Research International, Biometris, Wageningen, Netherlands <u>http://www.biometris.nl</u>

March 2014

Dissemination Level: Public

Contents D9.2b

1	Intr	oduction
	1.1	Statistical analysis and design for environmental risk assessment
	1.2	A protocol for the design and analysis of single-environment field trials – Task 9.3. 23
	1.3	Overview of this report
2	Setu	up of simulation study
	2.1	Basic setup and simulation distributions
	2.2	Parameter values used in the simulation
	2.3	Statistical models for analysis
3	Res	ults for difference testing
	3.1	General remarks on difference testing
	3.2	Scaling of the deviance difference for <i>OP</i> , <i>P1</i> , <i>P2</i> , <i>P3</i> and <i>GM</i>
	3.3	Simulated significance level of difference test
	3.4	Power of difference test
	3.5	Coverage of confidence intervals
	3.6	Approximate power of the difference test
	3.7	Method of choice for difference test
4	Res	ults for equivalence testing
	4.1	General remarks on equivalence testing
	4.2	Size of equivalence test
	4.3	Power of equivalence test
	4.4	Approximate power of the equivalence test
	4.5	Method of choice for equivalence testing
5	Zer	o inflation
	5.1	Introduction
	5.2	A zero-inflated negative binomial distribution and its non-inflated counterpart55
	5.3	Size of the LN difference test when there is zero inflation
	5.4	Power of the <i>NB</i> difference test for negative binomial data
6	Cor	clusion
7	Ref	erences
8	App	bendices

1 Introduction

1.1 Statistical analysis and design for environmental risk assessment

A basic statistical approach to environmental risk assessment (ERA) has been outlined in the EFSA Guidance Document (EFSA, 2010b) and in Perry et al. (2009). However, this approach is not specified in great detail. The aim of the statistics work package is to make the EFSA guidelines workable, practical and to fill in the gaps. This will result in a protocol which will provide risk assessors with a step-by-step approach for both design and statistical analysis of field trials. Statistical consideration of the EFSA for the safety evaluation of genetically modified organisms (EFSA, 2010a) will be incorporated in this protocol. Work package 9 will develop statistical concepts, methods, software and protocols for environmental risk assessment (ERA) and post-market environmental monitoring (PMEM). Main objectives are:

- to develop appropriate statistical methods to handle Genotype by Environment interaction in studies over multiple bio-geographic regions and under varying agronomical conditions. This is expected to be a major issue in the context of European ERA;
- to introduce equivalence testing as a main approach for ERA in addition to difference testing, and to establish protocols for experimental design based on acceptable test characteristics;
- to develop statistical approaches for handling data sets with many low counts and presence/absence data, as often encountered in ERA. Current practice is to use models based on normal distributions but this may not be appropriate;
- to implement methods in software for practical use;
- to provide protocols and draft texts for guidelines. The protocol will provide risk assessors with a set of evaluated, standardized and harmonized sampling and testing methods for environmental risk assessment;
- to provide guidelines for multivariate statistical approaches appropriate for PMEM.

Existing datasets will be studied to characterise baseline conditions found in different biogeographic regions, and to typify the variation of genotypes and environments (Task 9.1). Based on these results a simulation model will be built (Task 9.2), which will be used to test various statistical approaches for data analysis in relation to the possible design of experiments (e.g. sample size). Statistical approaches will use both difference and equivalence testing, and a graphical display of assessment results will be developed (Task 9.3). Also for multi-environment studies appropriate statistical methodology will be developed, including the consideration of genotype by environment interaction (Task 9.4). The statistical methods for analysis and design of field trials for Environmental Risk Assessment that give the best performance will be described in protocols for both single-environment (Task 9.3) and multienvironment studies (Task 9.6).

Tasks 9.1 (overview of existing ERA datasets) and 9.2 (simulation model for ERA data) are described in Goedhart et al (2013, 2014). This report describes results of a simulation study to investigate properties of various statistical models, which are used to perform difference and equivalence testing, for analysing count data.

1.2 A protocol for the design and analysis of single-environment field trials – Task 9.3.

This task is preparatory for Tasks 9.4 and 9.5 where multi-environment trials are addressed. Several statistical issues regarding data types, difference testing, equivalence testing and test characteristics can however be better researched in the relatively simple situation of a single-environment trial. This is also relevant because of the emphasis of the EFSA guidance document on single-environment trials. The EFSA guidance document states that "For field trials, since each field trial at a site on a particular occasion should have sufficient replication to be able to yield a stand-alone analysis if required, this power analysis should relate to a single site". Therefore protocols for power analysis and statistical analysis of a single field trial will be developed in this task. To develop such protocols it is important to known the statistical properties of various tests which are used in practice, for example the power and robustness of a test and whether the test has the correct significance level. This can best be researched by means of a simulation model. This reports describes such a simulation study.

1.3 Overview of this report

The simulation model developed in Task 9.2 was used to generate count data for the simple, but important, situation in which a field study is conducted to compare a GM plant with its conventional counterpart. It is assumed that a completely randomized experiment is used and that a single count, without excess zeros, of a non-target organism is available for each experimental unit. Chapter 2 describes the setup of the simulation study. Four different count distributions were used to simulate count data for a mean count ranging from 0.5 (for rare species) to 100 (for more common species). Different coefficients of variation and different levels of replication, ranging from 4 to 100, were used to simulate data. The ratio of the means of the GM plant and its comparator was set to 1, 0.75, 0.50 and 0.25. A ratio of 1 implies no difference between the GM plant and its comparator. The simulated data were analysed by means of eight different models, such that the most robust model could be selected. Chapter 3 describes the results obtained for difference testing; this includes the simulated size and power of the difference test as well as coverage of confidence intervals. It also compares an approximate fast method to obtain the power of a difference test. Finally a recommendation is given about which difference test is to be preferred. Chapter 4 deals with one-sided equivalence testing and describes the simulated significance level of various methods, the simulated power and a fast way of calculating the power. This also results in a recommendation about which equivalence test is to be preferred. Chapter 5 shortly deals with the problem of zero inflation, i.e. more zeros than predicted by the count distribution

2 Setup of simulation study

2.1 Basic setup and simulation distributions

The most simple trial in which a GM plant is compared to its conventional counterpart is a completely randomized field trial with level of replication N. In that simple case there are only two parameters: the mean count of the non-target organism for the GM plant (μ_G) and the mean count (μ_c) for the comparator. In practice there might be repeated counts on the same plots, but this is ignored in this simulation study. Goedhart et al (2013, 2014) describe five statistical distributions commonly used to simulate counts: the Poisson distribution, the overdispersed Poisson distribution, the negative binomial distribution, the Poisson-Lognormal distribution and a distribution which follows Taylor's power law. The Poisson distribution was not used in this simulation study because it is generally believed (Perry et al 2003, Duan et al, 2006) that counts of non-target organisms (NTOs) typically have larger variance than according to the Poisson distribution. Table 4 summarizes the four distributions which are used to simulate data, with the dispersion parameter σ^2 as a function of the mean μ and the variation coefficient CV in the last column. There is no statistical distribution associated with Taylor's power law, as it only specifies a relationship between the variance and the mean. Perry et al (2003) used the negative binomial distribution to simulate according to Taylor's power law employing a negative binomial dispersion parameter which follows from equating the variance of the negative binomial to the power law. The same approach is followed here. Using the negative binomial is however somewhat arbitrary, as e.g. the Poisson-Lognormal has the same variance to mean relationship, but has a different distribution.

Distribution	Abbreviation	Mean	Variance	Dispersion parameter
				σ^2 as a function of <i>CV</i>
Overdispersed Poisson	OP	μ	$\sigma^2 \mu$	$\mu (CV/100)^2$
Negative Binomial	NB	μ	$\mu + \sigma^2 \mu^2$	$(CV/100)^2 - 1/\mu$
Poisson-Lognormal	PL	μ	$\mu + \sigma^2 \mu^2$	$(CV/100)^2 - 1/\mu$
Power model ($p=1.5$)	<i>P1</i>	μ	$\sigma^2 \mu^{1.5}$	$\mu^{0.5} (CV/100)^2$

 Table 4:
 Distributions and values for the dispersion parameter used to simulate data.

The variance function of the Power model is more generally given by $Var = \sigma^2 \mu^p$ in which p is some power. In this simulation study p=1.5 was chosen because this results in a variance function nicely in between the variance function for the overdispersed Poisson on the one hand and the negative binomial and Poisson-Lognormal on the other hand.

The assumed variability in field testing of NTOs is mostly defined in terms of the coefficient of variation (*CV*), for example Duan et al (2006), and this convention is also used here. The mean μ_C of the comparator and the coefficient of variation *CV* define the dispersion parameter σ^2 , see Table 4. This same dispersion parameter is then used to generate counts for the comparator and also for the *GM* plant. So for example with μ_C =10 and *CV*=100%, the negative binomial dispersion parameter equals σ^2 =0.9. In case the *GM* plant, in the same simulation, has a mean μ_G =2.5, the corresponding *CV* value equals $\sqrt{2.5 + 0.9 \times 2.5^2}/2.5 = 114\%$. Moreover, a mean μ_G =1 has a corresponding *CV*=138% in this setting. This somewhat

higher *CV* value than for the comparator reflects the general believe that smaller means are associated with larger *CV* values. The quotient of the *CV* value for the *GM* plant and the comparator for each distribution is given below as a function of $Q = \mu_G/\mu_C$.

Overdispersed Poisson simulation distribution

The overdispersed Poisson distribution requires a dispersion parameter σ^2 which is larger than or equal to 1, where the limiting value of 1 results in an ordinary Poisson distribution. The quotient of the variation coefficients is given by

$$\frac{CV_G}{CV_C} = \sqrt{\frac{\sigma^2/\mu_G}{\sigma^2/\mu_C}} = \sqrt{\frac{\mu_C}{\mu_G}} = \sqrt{\frac{1}{Q}}$$

This implies that with Q = 0.25 the *GM* plant has a *CV* value which is twice as large as the *CV* of the comparator, irrespective of the value of μ_c .

Negative binomial and Poisson-Lognormal simulation distributions

The negative binomial and Poisson-Lognormal distributions both require a dispersion parameter σ^2 which is larger than 0. The quotient of the variation coefficients is given by a more complicated formula:

$$\frac{CV_G}{CV_C} = \sqrt{\frac{(\mu_G + \sigma^2 \mu_G^2)/\mu_G^2}{(\mu_C + \sigma^2 \mu_C^2)/\mu_C^2}} = \sqrt{1 + \frac{1 - Q}{Q \ \mu_C \ (CV/100)^2}}$$

This will be close to 1 for large CV values and for large values of μ_C .

Power law simulation distribution

For simulating according to the Power model, first the following equation is solved for τ : $\sigma^2 \mu^p = \mu + \tau \mu^2$; subsequently data are simulated according to a negative binomial distribution with dispersion parameter τ . Note that the equation is separately solved for the comparator, with mean μ_c , and for the GMO with mean $\mu_G = Q\mu_c$. This might results in a combination of parameter values which is not allowed. Suppose, as an example, $\mu_c=9$, $\mu_G=1$ and CV=50%. The dispersion parameter of the Power model with p=1.5 is then given by $\sigma^2=0.75$. However the equation for μ_c : $1+\tau 1^2 = 0.75*1^{1.5}$ cannot be solved for positive τ .

The quotient of the coefficients of variation is given by

$$\frac{CV_G}{CV_C} = \sqrt{\frac{\sigma^2 \mu_G^p / \mu_G^2}{\sigma^2 \mu_C^p / \mu_C^2}} = Q^{0.5p-1}$$

This implies that with Q = 0.25 and p=1.5 the *GM* plant has a *CV* value which is $\sqrt{2}$ as large as the *CV* of the comparator.

2.2 Parameter values used in the simulation

Depending on the NTO at hand, mean counts can be very small but can also be quite large. A range of 0.5 to 100 for the mean μ_c of the comparator is therefore employed.

Rather than focusing on the difference between μ_c and μ_g , it is more natural to focus on the ratio $Q = \mu_g/\mu_c$ of the two means. Generally accepted values in field testing for Q are between 0.5 and 0.25 (Comas et al, 2012). We used values 1, 0.75, 0.5 and 0.25. The value of 1, i.e. no difference between the comparator and the *GM* plant, is specifically meant to examine whether the difference test attains its nominal α -level. The other values of Q assume that the *GM* plant has a negative effect on the mean count.

The assumed variability in field testing of NTOs is mostly defined in terms of the coefficient of variation (*CV*). Duan et al (2006) present graphs with *CV* values ranging from 25% to 200% with generally low *CV* values for means larger than 5 and *CV* values up to 200% for means close to zero. In this study, five different values of *CV* are used for different values of μ_c as given in Figure 5 and Table 5. Compared to Duan et al (2006) the larger *CV* values used in this simulation study seem to be at the upper end of what can be expected in practical field trials.

Figure 5: Combinations of comparator means μ_C and coefficients of variation *CV*. The solid line denotes the coefficient of variation of a Poisson distribution.



Finally the level of replication N must be specified. Environmental risk assessment of GM plants is typically performed in experiments with a small number of plots. This is (partly) due to the fact that relatively large plots and large guard rows are required in order to measure effects on NTOs without bias, see Perry et al (2003). It is therefore that such experiments are frequently repeated in different years and different locations such that larger levels of

replication are obtained. A range of 4 to 100 for the level of replication N is employed in this study with some emphasis on lower values.

Table 5 summarizes the parameter values which are used in the simulation study. These values result in 1600 parameter combinations. For each combination of the simulation distribution (*OP*, *NB*, *PL* and *P1*) and parameter values 1000 datasets were simulated. Each dataset was analysed using the models given in the next session and an appropriate difference test at the 5% level was performed (details are given below). The proportion of datasets for which the difference test is rejected then gives an estimate of the true significance level (α) of the test when there is no difference, i.e. Q=1, and the power (β) of the test when there is a difference, i.e. $Q\neq 1$. These are only estimates of the true size of the test. Suppose that the size of the test is indeed exactly 5%, then with 1000 simulations a 99% prediction interval for the number of times the null hypothesis will be rejected is given by (33, 67) resulting in an interval of 3.3% – 6.7% for the true size. So only when the simulated significance level is outside this interval there is an indication that the true level of the test does not equal 5%.

Parameter		Values used in simulation			
Mean μ_C of c	comparator	0.5, 1, 2, 5	5, 10, 20, 50,	100	
Ratio $Q = \mu_0$	$_{\rm g}/\mu_{\rm C}$	1, 0.75, 0.5	6, 0.25		
Number of re	eplication N	4, 6, 8, 10,	, 15, 20, 30, 4	10, 60, 100	
		Coefficient of	variation CV f	or comparator	
μ_C	<i>CV</i> -1	<i>CV</i> -2	<i>CV</i> -3	<i>CV</i> -4	<i>CV</i> -5
0.5	150	200	300	400	500
1	150	200	300	400	500
2	100	150	200	300	400
5	75	100	150	200	300
10	50	75	100	150	200
20	40	50	75	100	150
50	20	30	40	50	75
100	15	20	30	40	50

Table 5:	Parameters	used in	the	simulation	study.
----------	------------	---------	-----	------------	--------

Data were simulated using the statistical package GenStat (VSN international, 2013).

2.3 Statistical models for analysis

Fitting the Poisson-Lognormal model by means of maximum likelihood requires (adaptive) Gauss-Hermite integration within an iterative weighted least squares algorithm. This algorithm turned out to fail too frequently for data with small means, small levels of replication and/or small coefficients of variation. Therefor the Poisson-Lognormal model was not used to analyse simulated data. The other models with which each dataset was analysed are summarized in Table 6. All models were fitted using standard facilities in the statistical package GenStat (VSN international, 2013). Details for each analysis model are given below. A difference test for all models can be obtained by comparison of the fit of the model, more specifically the deviance, under the null-hypothesis $H_0: Q = 1$ and the fit of the model under the alternative hypothesis $H_1: Q \neq 1$.

Analysis model	Abbreviation	Type of difference test
Log transformation	LN	t-test
Squared-root transformation	SQ	t-test
Overdispersed-Poisson	OP	scaled deviance difference
Negative binomial	NB	deviance difference
Power model $p=1.5$	<i>P1</i>	scaled deviance difference
Power model $p=1.7$	P2	scaled deviance difference
Power model <i>p</i> =1.99	<i>P3</i>	scaled deviance difference
Gamma model	GM	scaled deviance difference

 Table 6:
 Statistical models used to analyse the simulated data.

LN: Log transformation followed by a t-test

The count data are log-transformed after the addition of 1 to prevent taking the logarithm of zero. The simple two-sample t-test is then applied to the log transformed counts. The log transformation stabilizes the variance for distributions with a standard deviation which is proportional to the mean, or $Var(Y) \propto \mu^2$. This transformation therefore seems appropriate for the negative binomial and the Poisson-lognormal distribution with means that are not too small.

The two-sample t-test employs an estimate of the difference between the *GM* plant and the comparator on the transformed logarithmic scale. This difference is however a quantity that is not easy to interpret, especially when the underlying means μ_G and μ_C are small. Instead interest is in the ratio $Q = \mu_G/\mu_C$. The so-called generalized confidence interval approach can be applied to provide an interval for the ratio of two lognormal means, see Krishnamoorthy & Mathew (2003) and Chen and Zou (2006). According to these authors such an interval has excellent coverage probabilities. This approach uses the fact that, assuming that the log-transformed counts follow a normal distribution, the residual mean square follows a scaled Chi-squared distribution and that the two sample means follow a normal distribution which is independent of the Chi-squared distribution. A simulation approach is then used to generate a large sample for the ratio of the two lognormal means, accounting for the addition of 1. Percentiles of this large sample then define a confidence interval. More specifically, with X_C and X_G the two sample means on the log-transformed scale, S^2 the estimate of the variance on the transformed scale and 2N-2 the number of degrees of freedom for S^2 , a large sample for the ratio Q is generated in the following way

- 1. A random draw *Chi* is generated by means of *Chi* = $(2N-2) S^2 / \chi_{2N-2}$ where χ_{2N-2} is a random draw from a Chi-squared distribution with 2*N*-2 degrees of freedom;
- 2. N_C is a random draw from a normal distribution with mean X_C and variance Chi/N;
- 3. N_G is a random draw from a normal distribution with mean X_G and variance Chi/N;
- 4. Back-transform N_c by means of $N_c = \exp(N_c + Chi/2)$ and similarly N_G . Note that the back-transformation uses the equation for the mean of the lognormal distribution;

- 5. Subtract 1 from N_c and N_G ; this accounts for the addition of 1 before log-transforming the count. This might sometimes result in a negative value for N_c or N_G . Such values are replaced by a small positive value, i.e. by 0.0001.
- 6. Calculate the ratio N_G/N_C
- 7. Repeat steps 1-6 many times, e.g. 10.000 or when more precise results need to be obtained 100.000 times. Calculate appropriate percentiles of the large sample which is the generalized confidence interval.

The generalized confidence interval can be used for difference testing as well as for equivalence testing.

SQ: Squared root transformation followed by a t-test

The squared root transformation is frequently used as an alternative for the log transform, and a simple t-test is also performed on squared root transformed counts. This transformation stabilizes the variance when the variance is proportional to the mean, or $Var(Y) \propto \mu$. This transformation is therefore especially appropriate for the overdispersed Poisson distribution.

The generalized confidence interval approach can also be employed to obtain an interval for the ratio on the original scale. The only modification to the seven steps described for the *LN* analysis is the back-transformation in step 4. For the squared root transform this is given by $N_C = N_C^2 + Chi$ which employs the well-known relation $Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2$ where \mathbb{E} denoted taking the expectation. Step 5 has to be skipped.

OP: Overdispersed Poisson by a GLM-like analysis

There does not seem to be standard software to fit the overdispersed Poisson distribution by means of maximum likelihood. However, a common way to analyse overdispersed counts is to use the quasi-likelihood approach of McCullagh and Nelder (1989). This amounts to fitting the ordinary log-linear model, which employs the Poisson distribution and a log-link, and to scale standard errors of parameter estimates by means of the squared root of an estimate of the dispersion parameter. This is the approach which is followed here. A scaled likelihood ratio statistic is obtained by calculating the scaled deviance difference of the model under H_0 and H_1 . Scaling can be done by the mean deviance or by Pearson's Chi-squared statistic, both under H_1 , and both methods are compared. The scaled likelihood ratio statistic is compared with a F distribution with 1 and 2N-2 degrees of freedom to obtain a p-value.

In this model the underlying mean is log-transformed, rather than taking logs of the observed counts. This implies that the logarithm of the ratio of the two means, i.e. log(Q), is directly estimated in this model. A so-called Wald test statistic (Buse, 1982) can then be used for difference testing. This equals the quotient of the estimate of log(Q) and its standard error, and this is usually compared to a t-distribution to compensate for the estimation of the dispersion parameter. However it is generally believed that the likelihood ratio statistic has better statistical properties (McCullagh and Nelder, 1989). Moreover the Wald statistics breaks down when either sample only contains zero's since the estimate of log(Q), and its standard error, then becomes plus or minus infinity. So difference testing is based on the scaled likelihood ratio test. Equivalence testing under this model is however based on the

estimate of log(Q) and its standard error, scaled by Pearson's statistic, which can be used to generate a confidence interval and thus to perform equivalence testing for arbitrary limits of concern. An alternative would have been to calculate a so-called profile likelihood interval but this requires a search algorithm which was considered to be too computer intensive in this simulation study.

NB: Negative binomial model by a GLM-like analysis

The negative binomial regression model, with logarithmic link, is fitted to the counts by means of maximum likelihood. The likelihood ratio test is calculated and compared to a Chi-squared(1) distribution. The dispersion parameter of the negative binomial distribution was bounded to the interval [0.001, 1000] to avoid numerical problems.

The estimate of log(Q) and its standard error is used for equivalence testing.

P1, P2 and P3: Power Law model by a GLM-like analysis

The Power model is defined by a variance-to-mean relationship and there is no true statistical distribution associated with it. Therefore, like the overdispersed Poisson model, quasi likelihood is used to fit the model. The quasi deviance *D* can be obtained by employing its definition, see McCullagh & Nelder (1989):

$$D(y;\mu) = 2\int_{\mu}^{y} \frac{y-t}{Var(t)} dt$$

For Taylors Power Law, i.e. $Var(t) = t^{\beta}$, the quasi deviance becomes

$$D(y;\mu) = 2\int_{\mu}^{y} \frac{y-t}{t^{\beta}} dt = 2\left[\frac{yt^{1-\beta}}{1-\beta} - \frac{t^{2-\beta}}{2-\beta}\right]_{\mu}^{y} = 2\left[\frac{y^{2-\beta}}{1-\beta} - \frac{y^{2-\beta}}{2-\beta}\right] - 2\left[\frac{y\mu^{1-\beta}}{1-\beta} - \frac{\mu^{2-\beta}}{2-\beta}\right]$$
$$= 2\left[\frac{y^{2-\beta}}{(1-\beta)(2-\beta)} - \frac{y\mu^{1-\beta}}{1-\beta} + \frac{\mu^{2-\beta}}{2-\beta}\right] = 2(z1-z2+z3)$$

The model is fitted using GenStats facilities for generalized linear models with non-standard variance functions. The GenStat directives for defining the model are as follows, where 'response' is the observed count, 'power' is the value of p in the variance function and 'z1', 'z2' and 'z3' are the three terms between squared brackets in the equation above.

```
CALCULATE b1,b2 = 1,2 - power

EXPRESSIO dcalc[1] ; VALUE=!e(vfunction = mu**power)

EXPRESSIO dcalc[2] ; VALUE=!e(z1 = response**b2/(b1*b2))

EXPRESSIO dcalc[3] ; VALUE=!e(z2 = response*mu**b1/b1)

EXPRESSIO dcalc[4] ; VALUE=!e(z3 = mu**b2/b2)

EXPRESSIO dcalc[5] ; VALUE=!e(deviance = 2*(z1-z2+z3))

MODEL [DISTRIBUTION=calculated ; DCALCULATION=dcalc[] ; \

LINK=log ; DMETHOD=pearson ; DISPERSION=*] response ; \

FITTED=fitted ; VFUNCTION=vfunction ; DEVIANCE=deviance
```

To obtain a test-statistic the deviance difference can be scaled by the mean deviance or Pearson's test statistic, both under H_1 . The test statistic was compared to a F distribution with 1 and 2*N*-2 degrees of freedom. The power model was fitted with a fixed power *p* of 1.5, of 1.7 and of 1.99, and these are denoted by *P1*, *P2* and *P3* respectively. Note that a power p=2is not allowed by the model as this implies division by zero.

A confidence interval is obtained for the estimate of log(Q) and its standard error, scaled by Pearson's statistic and using a t-distribution, and this is used for equivalence testing.

GM: Gamma model using a GLM-like analysis

The final analysis is by means of the Gamma distribution employing a log-link. Since the gamma distribution cannot handle zero observations, zeroes were replaced by 0.001. Again the deviance difference was scaled by the mean deviance or Pearson's chi-squared and compared with a F distribution with 1 and N-2 degrees of freedom to obtain a p-value. Also a confidence interval is obtained for the estimate of log(Q) and its standard error, scaled by Pearson's statistic and using a t-distribution, and this is used for equivalence testing.

Special cases

For small means and small levels of replication sample means can easily become zero for a simulated dataset. When both sample means equal zero, or more generally when both variances within samples equal zero, the analysis according to the log-transformation cannot be performed because the residual mean square equals zero. Some decision has to be taken to deal with such situations. Consider therefore the case with 4 observations of the comparator and 4 observations for the *GM* plant, with obvious generalizations to more observations. The four cases below are then special.

- A. Sample 1 equals {0, 0, 0, 0} and sample 2 equals {0, 0, 0, 0}. In this case there is no information and the deviance under the null model and under the alternative model are both zero for all models. The p-value for the difference test is set to 1 for all analysis models as there is no indication of a difference between the two samples. For the most extreme parameter combination $\mu_R=0.5$, CV=500, Q=0.25, N=4 and the overdispersed Poisson distribution this situation occurs for 570 of the 1000 simulated datasets. For negative binomial, Poisson-LogNormal and Power models these numbers are respectively 511, 287 and 565. Clearly there is also no information for calculating a confidence interval and thus formal equivalence testing cannot be performed. Graphical results for equivalence testing present the proportion of these cases separately. Note that this case can be considered as "equivalent more likely than not".
- B. Sample 1 equals {0, 0, 0, 0} and sample 2 equals {c, c, c, c} where c is some positive value. The deviance under the alternative model equals zero and so no test statistic can be calculated. However this situation is very rare. For the Poisson-LogNormal distribution there are 28 parameter combinations for which this situation occurs with a maximum of 5 out of 1000 such datasets at most. For the other distributions this situation occurs even less. These situations are therefore discarded, i.e. the corresponding p-value is set to missing.

- C. Sample 1 equals $\{0, 0, 0, 0\}$ and sample 2 has different values with a positive variance. In this case all the p-values can be calculated in the usual way.
- D. The mean of both samples are positive with a zero variance, e.g. {1, 1, 1, 1} and {3, 3, 3, 3}. This is essentially the same as case B although it will occur even rarely. There are only 2 simulated datasets for which this occurs and these are discarded.

3 Results for difference testing

3.1 General remarks on difference testing

A key element in environmental risk assessment it to test whether the GM plant is different from its conventional counterpart. The aim of a statistical difference test is to reject the null hypothesis of no difference between the GM plant and its comparator. A significant difference test is then a "proof of difference", but this does not state that the difference is biologically relevant and constitutes a true hazard to the environment. Poorly designed experiments with low levels of replication may have low statistical power of finding a true difference. So the absence of a significant difference is not a proof that there is no difference, or "absence of evidence is not evidence of absence" (Altman and Bland, 1995). There are two possible types of errors for a difference test. A type I error occurs when the null hypothesis of no difference is falsely rejected when it is actually true. In that case the incorrect conclusion is drawn that the GM plant is different from its comparator. A type II error on the other hand occurs when the null hypothesis is not rejected although it is untrue. Typically the probability of a type I error, also known as the size of the test or α , is set to some pre-described small value such as 5%, implying that in 5% of all tests the null hypothesis of no difference is falsely rejected. Given the size of the test, the probability of a type II error depends on the true difference, the level of variation and the level of replication. Note that the power of a test, frequently denoted by β , equals one minus the probability of a type II error.

The size of tests based on the normal distribution, such as the t-test, is exact. However tests based on other distributions, like the Poisson and the negative binomial, depend on asymptotic (meaning large levels of replication) arguments and are therefore not exact. This implies that a test, which is supposed to have a size of say 5%, might in practice have a different size. When the actual size of the test is larger than α the test is said to be progressive, when it is smaller the test is said to be conservative. Progressive tests are considered to be specifically bad because the null hypothesis of no difference is falsely rejected more often than the predescribed α level. Frequently the true underlying distribution of counts is not known. We might for instance falsely analyse data according to the Poisson distribution while in practice the data follow the negative binomial distribution or vice versa. This is particularly likely to happen when counts are small, as encountered frequently in ERA experiments, because then it is hard to discriminate between probability models. This ignorance about the true underlying distribution might result in difference tests to become even more progressive or conservative.

The power of a difference test based on the normal distribution can be calculated exactly. For non-normal distributions, small sample properties of difference tests are not straightforward. A simulation approach for sample size calculations for a difference test is employed by many authors, e.g. Shieh (2001) and Hrdličková (2006) for the Poisson distribution, Shieh (2001)

and Demidenko (2008) for the binomial distribution, Aban et al (2009) and Friede and Schmidli (2010) for the negative binomial distribution. A general practical approach to computing power for non-normal distributions is given by Lyles et al (2007).

In the remainder of this chapter simulation results of various properties of the difference tests are presented. All results presented are for a two-sided test of no difference with a significance level α =5%. Detailed results are given in a separate document with Appendices.

3.2 Scaling of the deviance difference for *OP*, *P1*, *P2*, *P3* and *GM*

When data are analysed by means of the overdispersed Poisson, Power or Gamma model the likelihood ratio statistic can be scaled by means of the mean deviance or by means of Pearson's chi-squared, both for the full model. The simulated significance level of these two variants of the test statistic for specific parameter combinations is given in Figure 6 and Figure 7 when data are simulated by means of the negative binomial distribution with coefficients of variation as given by *CV*-1 and *CV*-3, and in Figure 8 and Figure 9 when data are simulated by means of the Poisson-Lognormal distribution. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed α =0.05. The red lines denote values 0.033 and 0.067 which provide a range that could be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK and such values are denoted by open circles. Values outside this range are denoted by filled circles, while values larger than 0.096 are given by triangles. Results for all parameter combinations are given in Appendix 1 A-D.

Overdispersed Poisson (OP) as analysis model

For small *CV* values (Figure 6 and Figure 8) and the overdispersed Poisson distribution as analysis model the size of both test statistics is good for values of $\mu \ge 2$. For smaller values of μ more replications are needed to attain the correct size. Scaling by means of Pearson's chi-squared seems to have the edge over scaling by means of the mean deviance. For larger *CV* values (Figure 7 and Figure 9) the size of the both test statistics is generally bad for $\mu \le 2$. For larger replication levels and larger μ scaling by means of Pearson's chi-squared results in a better size than scaling by means of the mean deviance.

Power(1.5) (P1) as analysis model

For small *CV* values (Figure 6 and Figure 8) and the Power(1.5) analysis model, scaling by means of the mean deviance generally gives a conservative test for smaller values of μ , while scaling by means of Pearson's chi-squared has correct size, except for small values of μ and low level of replication *N*. For larger *CV* values (Figure 7 and Figure 9) both test statistics are progressive for small values of μ even for large replication levels *N*. For larger μ and simulating according to the negative binomial scaling by means of the mean deviance has better size than scaling by means of Pearson's chi-squared. However when data are simulated by means of the Poisson-LogNormal this is the other way around

Gamma (GM) as analysis model

For small *CV* values (Figure 6 and Figure 8) and the Gamma analysis model, scaling by means of the mean deviance is very conservative, while scaling by means of Pearson's chisquared generally has the correct size. For larger *CV* values (Figure 7 and Figure 9) both test statistics perform badly for values $\mu \leq 5$. For larger means scaling by means of Pearson does have the edge especially when simulating according to the negative binomial distribution.





Figure 7: Size of the test when the deviance difference is scaled by means of the mean deviance and by means of Pearson's chi-squared. Data are simulated by the negative binomial distribution with *CV*-3 values.



Figure 8: Size of the test when the deviance difference is scaled by means of the mean deviance and by means of Pearson's chi-squared. Data are simulated by the Poisson-LogNormal distribution with *CV*-1 values.



Figure 9: Size of the test when the deviance difference is scaled by means of the mean deviance and by means of Pearson's chi-squared. Data are simulated by the Poisson-LogNormal distribution with *CV*-3 values.



Conclusion

Scaling of the deviance difference by means of Pearson statistic seems to have somewhat better properties especially when the coefficient of variation is small. This conclusion is not only based on Figure 6 to Figure 9 but also on the results presented in Appendix 1 A-D. Therefore in subsequent comparisons the deviance difference will be scaled by means of Pearson's chi-squared for analysis according to the overdispersed-Poisson, the Power models and the Gamma model.

3.3 Simulated significance level of difference test

Having decided that scaling of the deviance difference by means of Pearson's statistic for *OP*, *P1*, *P2*, *P3* and *GM* generally has better properties than scaling by means of the mean deviance, the size of all analysis methods can be compared. Full details of the size of the difference test for all parameter combinations and simulation distributions are given in Appendix 1 E-H. Results for the *P3* model, with power 1.99, are not displayed since they are very similar to the results for the Gamma (*GM*) model. Results for specific combinations are given in Figure 10 to Figure 13.
Figure 10: Size of the difference test under various analysis models for data simulated by the negative binomial distribution with *CV*-1 values.



Figure 11: Size of the difference test under various analysis models for data simulated by the negative binomial distribution with *CV*-3 values.



Figure 12: Size of the difference test under various analysis models for data simulated by the Poisson-LogNormal distribution with *CV*-1 values.



Figure 13: Size of the difference test under various analysis models for data simulated by the Poisson-LogNormal distribution with *CV*-3 values.



The size of the LN and SQ analysis is extremely good for all parameter combinations, except for small values of μ combined with large coefficients of variations CV and low levels of replication N. In such cases the LN and SQ tests are conservative. The GLM-like models result in sometimes progressive test especially for small means in combination with a large coefficient of variation. Among the GLM-like models there is no clear winner although the OP analysis seems to outperform the other GLM models somewhat, especially when data are simulated according to the Poisson-Lognormal distribution.

The simulated significance level for all parameter combinations and simulation distributions is summarized in Figure 14 to Figure 17. The symbols in Figure 14 to Figure 17 have the following meaning: *open circle* denotes that the test is conservative for lower levels of replication and has the correct size for larger replication; *closed circle* denotes that the test has correct size for all replication levels; *cross* means that the test is mainly progressive; *number* denotes that the test has correct size for levels of replication larger than the plotted number. These plots can be used to quickly check for which parameter combination, and for which level of replication, the difference test has correct size. These plots clearly indicate, once again, that the *LN* and *SQ* analysis models have superior size. The best alternative, especially for for larger means and smaller coefficients of variation is the *OP* analysis model.

	LN transform	Overdispersed Poisson	Power(1.5)					
CV-5 –		★ ★ 60 20 30 ●	★ ★ 60 60 30 ●					
CV-4 -		★ ★ 60 40 20 ● ●	★ ★ 60 40 20 15 ●					
CV-3 –		60 60 20 15 •	100 60 20 20 🔶 🔶 🔶					
CV-2 -	····O··O·•	15-20-30	15 40 30					
CV-1 -		10-30-	1030⊙●●●●●					
	SQ transform	Negative Binomial	Gamma					
CV-5 –	·····©····©····••··•••·••	★ 60 60 20 15 20 8 15	★ ★ ★ 60 60 40 •					
CV-4 -	•••••••••	40 30 20 20 20 20 8 8	★ ★ 40 20 15 ● ●					
CV-3 –			60 60 20 20 •• •					
CV-2 -			15 20 30 .					
CV-1 -	10	15 30 10 15 20 20 8 8	·····6····6····•					
	i i i i i i i i 0.5 1 2 5 10 20 50 100	i i i i i i i 0.5 1 2 5 10 20 50 100	i i i i i i i 0.5 1 2 5 10 20 50 100					
	μ	μ	μ					

Figure 14: Summary of size of difference test; data simulated by Overdispersed Poisson (see text for explanation of symbols)



Figure 15: Summary of size of difference test; data simulated by Negative Binomial (see text for explanation of symbols)

Figure 16: Summary of size of difference test; data simulated by Poisson-Lognormal (see text for explanation of symbols)





Figure 17: Summary of size of difference test; data simulated by Power(1.5) model (see text for explanation of symbols)

There is a large body of literature about the robustness of the two-sample t-test against departures from normality, two early references are Pearson and Adyanthāya (1929) and Gayen (1950). Miller (1986) summarizes the literature by noting that in case the skewness of the two samples is equal and so is the kurtosis then "the kurtosis parameters have little effect on the t statistic and when the sample sizes are approximately equal the skewness parameters cancel each other approximately". So assuming that the two samples follow an identical distribution, the t-test is very robust against departures from normality. This does not imply that the t-test can also be applied to the counts themselves. Taking logs or a squared root makes the count distribution more symmetric giving less departures from normality which results in better properties of the t-test.

Conclusion

The simulated size of the t-test after a log or squared-root transformation, models *LN* and *SQ* respectively, is close to its nominal level, except for small means and large *CV* values. The other analysis models are progressive for small means and larger coefficients of variation. In other cases the *OP* analysis seems to outperform the other GLM-like models.

3.4 Power of difference test

The power of the difference test for all parameter combinations is given in Appendices I-L for effects size Q=0.75 (black), Q=0.50 (red) and Q=0.25 (green). Results for specific combinations are given in Figure 18 to Figure 21. Each small plot has a range of 0 to 1 along the y-axis. The grey horizontal lines denote power values of 0.25, 0.50 and 0.75. Values for

progressive tests, i.e. when the simulated size of the test is larger than 0.067, are not displayed. This once again shows that the LN and SQ tests are never progressive.

When data are simulated according to the overdispersed Poisson distribution (Appendix 1 I) there is very little difference between the power of the various analysis models. However the *LN* and *SQ* method seem to have a somewhat larger power for larger values of μ .

When data are simulated according to the negative binomial distribution (Appendix 1 J, Figure 18 and Figure 19) the *LN* test occasionally has slightly smaller power than the other tests. An example is given in Figure 19 for μ =10 and μ =20. The GLM-like models have very similar power.

When data are simulated according to the Poisson-Lognormal distribution (Appendix 1 K, Figure 20 and Figure 21) the power of the *LN* and *SQ* tests is as least as good as for the other models.

When data are simulated according to the Power P1 model (Appendix 1 L), once again the power of the LN and SQ tests is as least as good as for the other models.

Conclusion

The power of the LN and SQ approach is generally very similar to the power of the other analysis methods. In some cases the power of LN and SQ is marginally larger, in other cases it is marginally lower. Because there is very little difference between the power of the various models, other properties, like the size of tests, of the various models should be decisive as to which method is to be preferred.

Figure 18: Power of the difference test for effects *Q*=0.75 (black), 0.50 (red) and 0.25 (green) under various analysis models for data simulated by the negative binomial distribution with *CV*-1 values



Figure 19: Power of the difference test for effects *Q*=0.75 (black), 0.50 (red) and 0.25 (green) under various analysis models for data simulated by the negative binomial distribution with *CV*-3 values



Figure 20: Power of the difference test for effects *Q*=0.75 (black), 0.50 (red) and 0.25 (green) under various analysis models for data simulated by the Poisson-Lognormal distribution with *CV*-1 values



Figure 21: Power of the difference test for effects *Q*=0.75 (black), 0.50 (red) and 0.25 (green) under various analysis models for data simulated by the Poisson-Lognormal distribution with *CV*-3 values



3.5 Coverage of confidence intervals

After each analysis a 95% two-sided confidence interval can be constructed for the ratio $Q = \mu_G/\mu_C$ of the two means. This uses the generalized confidence interval approach for the *LN* and *SQ* models. For the GLM-like models the confidence interval is constructed in the usual way employing the estimate of log(*Q*) and its standard error. One minus the simulated coverage probabilities of these intervals are given in Appendix 1 M-P. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed α =0.05. The red lines denote values 0.033 and 0.067 which provide a range that could be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK and such values larger than 0.096 are given by triangles. Note that in this case large values denote a confidence interval that is too small (i.e. has smaller coverage probability than 95%), while small values indicate a confidence interval that is too wide (i.e. has larger coverage probability than 95%).

The LN and SQ generalized confidence interval can be used to test the null hypothesis of equal means. For LN both the simulated significance level and the simulated power of the generalized confidence interval are identical to those of the t-test. This can, for the simulated significance level of the LN analysis when data are simulated according to the overdispersed Poisson, be seen by comparing the first columns in Appendix 1 E with the top graphs in Appendix 1 M1. Similarly for SQ the second columns in Appendix 1 E can be compared with the top graphs in Appendix 1 M3. It then turns out that the generalized confidence interval for SQ has a slightly lower size significance level than the corresponding t-test for small means combined with small levels of replication and larger CV values. Similar comparisons can be made when data are simulated by the other three distributions. Results for the simulated power of the difference test employing the generalized confidence interval are not shown, but these are for the LN analysis also identical to the power of the t-test.

However the properties of the *LN* and *SQ* generalized confidence interval are only good when testing the null hypothesis of equal means. Coverage of the *LN* interval deteriorates when the ratio $Q = \mu_G/\mu_C$ of the two means becomes smaller, e.g. the bottom graphs in Appendix 1M1 for Q=0.75 and Appendix 1 M2 for Q=0.50 and Q=0.25. Coverage of the *SQ* interval is even worse for values $Q \neq 1$, see Appendix 1 M3 and M4. So it appears that the generalized confidence interval of *LN* and *SQ* can be used for difference testing, but it cannot be used for equivalence testing.

The coverage probability of the *OP* interval, when simulating according to the overdispersed Poisson, is generally better than those of the other GLM-like models (Appendix 1 M5-12). For smaller values of μ and larger values of *CV* the *OP* interval is too wide indicating that the corresponding test is conservative. Intervals for *NB*, *P1* and *GM* can be too short or too wide depending on the parameter combination.

When data are simulated by means of the negative binomial distribution (Appendix 1 N) the *OP* interval generally has the better properties; the *OP* interval is almost never too small. However for Q=0.25 the *OP* interval is somewhat too wide, while the *NB* and *P1* intervals then have a better coverage except when the *CV* is large in which case these intervals can become too small. The *P1* interval seems to have somewhat better coverage than the *NB* interval.

When data are simulated according to the Poisson-Lognormal distribution (Appendix 1 O) the *NB*, *P1* and *GM* interval can be too small especially for smaller μ and larger *CV* values. For other values the *P1* interval seems to have the edge over the *NB* and *GM* intervals. The *OP* interval is, once again, somewhat too wide for Q=0.25.

Results for the Power model (Appendix 1 P) are similar to those for the Poisson-Lognormal.

Conclusion

The *LN* and *SQ* generalized confidence intervals have the same properties as the t-test for difference testing, although the *SQ* interval has a somewhat lower simulated significance level for some parameter combinations. However these intervals do not have good coverage probability for $Q \neq 1$, especially not for small values of *Q*. In such cases the *LN* interval has a less worse coverage probability than the *SQ* interval. The *OP* interval is almost never too small (meaning that the coverage is not smaller than 95%). It can be too wide though especially for Q=0.25 in combination with a simulation distribution other than overdispersed Poisson. In such cases the *P1* interval seems to be the method of choice although *P1* has the disadvantage that interval can be too small when simulating according to the overdispersed Poisson, and also for smaller μ and larger *CV* values for the other simulation distributions.

3.6 Approximate power of the difference test

Lyles et al (2007) describe a general method to approximate the power of a difference test for generalized linear models. Their approach makes use of a single 'expanded' dataset based on the response distribution. This expanded dataset is then analysed using an appropriate model and the power of the test statistic, either Wald or likelihood ratio, can then be calculated employing a Chi-squared distribution with a non-centrality parameter which can be easily calculated.

This approximate method is compared with the simulated power of the *LN* analysis. The approximate method consists of the following steps:

- 1. Create an 'expanded' dataset for the simulation distribution at hand. First choose the possible outcomes $y_{C1}, ..., y_{CK}$ for a mean value μ_C of the distribution and calculate the corresponding probabilities $w_{C1}, ..., w_{CK}$. The sum of these probabilities should then be close to one. Do the same for a mean value of μ_G giving possible outcomes $y_{G1}, ..., y_{GL}$ with probabilities $w_{G1}, ..., w_{GL}$, with again a sum close to one. Then simply stack the two vectors of possible outcomes and also the two vectors of corresponding probabilities, denote these as *Y* and *W*. Also create an indicator vector *X* with a zero for the first set of possible outcomes and a one for the second set. This results in the 'expanded' dataset consisting of *Y*, *W*, and *X* which are of equal length.
- 2. The mean and variance of both log-transformed samples are calculated employing $Mean_C = \sum_i w_{Ci} \log(y_{Ci} + 1)$ and $Var_C = \sum_i w_{Ci} [\log(y_{Ci} + 1) Mean_C]^2$ and

similarly for $Mean_G$ and Var_G . An estimate of the residual variance on the transformed scale is then given by $Var = 0.5(Var_C + Var_G)$.

- 3. The 'expanded' dataset is analysed by means of a weighted regression of log(Y + 1) on X with weights W/Var and fixed residual variance equal to 1. This results in an estimate of the regression coefficient β for X along with a standard error *se*.
- 4. The non-centrality parameter is then given by $\delta = N (\beta/se)^2$ where *N* is the number of replications. The same non-centrality parameter is obtained by calculating *N* times the difference between the residual sums of squares of the weighted regression model without *X* and the residual sums of squares of the model with *X*.
- 5. The power is calculated in the following way. A critical value F_{crit} is obtained from the F distribution with 1 and 2*N*-2 degrees of freedom. i.e. $P(F_{1,2N-2} > F_{crit}) = \alpha$. The approximate power is then calculate by means of the non-central χ_1 distribution with non-centrality parameter δ , i.e. by means of $P(\chi_1(\delta) > F_{crit})$.

A crucial step is the calculation of the residual variance *Var* on the transformed scale. The non-centrality parameter δ is proportional to the number of replications *N*, so there is no need for stacking the two vectors *N* times as is proposed by Lyles et al (2007). This implies that a single 'expanded' dataset can be used for all levels of replication *N* instead of a separate 'expanded' dataset for each level of replication.

In the implementation of this approach it was found that it might be numerically more stable to use weights N_0W/Var where N_0 is some fixed large number, e.g. 100. This is because units with very small weights, in this case with very small probabilities, are sometimes discarded when fitting a regression model. The non-centrality parameter is then given by $\delta = (N/N_0) (\beta/se)^2$.

The approximate power is calculated for all four distributions and compared with the simulated power. Graphical results are given in Appendix 1 Q1, R1, S1 and T1. Each small plot has a range of 0 to 1 along the y-axis. The grey horizontal lines denote power values of 0.25, 0.50 and 0.75. Simulated powers are given by the dots for Q=0.75 (black), Q=0.50 (red) and Q=0.25 (green), while the approximate power is given by the lines. Across the board there is very good agreement between the two methods. For low power values and smaller numbers of replications the approximate method of Lyles can be somewhat too small, but such low power values are hardly of interest.

The same approach can be followed for the *SQ* analysis, see Appendix 1 Q2, R2, S2 and T2, except that the squared root transformation is used instead of the log transformation. Also in this case there is very good agreement between the simulated power and the approximated power.

The same approximate method can be applied for an analysis according to one of the other models. For parameter combinations with a simulated significance level which is not (too) progressive, the two methods agree closely when analysing with a negative binomial for all four simulation distributions (Appendix 1 Q4, R4, S4 and T4). For an analysis with the power *P1* model (Appendix 1 Q5, R5, S5 and T5) the approximation is good when data are simulated according to the negative binomial or the power model especially for larger power values. When simulating with the overdispersed Poisson or the Poisson-Lognormal

distribution the method of Lyles sometimes gives less good results for the *P1* analysis. For an analysis with the *OP* model ((Appendix 1 Q3, R3, S3 and T3) the approximation is frequently less good, except for larger means with low *CV* levels.

Conclusion

When two-sample count data are analysed by means of the LN or SQ model the method of Lyles et al (2007) approximates the power very well for all four simulation distributions. In such a case there is no need to perform a simulation study to approximate the power.

3.7 Method of choice for difference test

The simulated size of the t-test after a log or squared-root transformation, models *LN* and *SQ* respectively, is close to its nominal level, except for small means and large *CV* values where the test is conservative. The other analysis models are progressive for small means, small levels of replication and larger coefficients of variation. In other cases the *OP* analysis seems to outperform the other GLM-like model.

The power of the *LN* and *SQ* approach is generally very similar to the power of the other analysis methods. In some cases the power of *LN* and *SQ* is marginally larger, in other cases it is marginally lower. In those case where the *LN* and *SQ* analysis are conservative (small means, small levels of replication and larger coefficients of variation), the power is so low that it is hardly worthwhile to perform such experiments. In other words for parameter combinations with sufficient power the size of the *LN* and *SQ* tests is close to its nominal level.

The *LN* generalized confidence interval has the same properties as the t-test for difference testing, with respect to the simulated significance level and with respect to the simulated power. This is also true for the *SQ* interval although the simulated size using this interval is smaller than that of the corresponding t-test for small means combined with low replication and larger *CV* values. The *LN* and *SQ* intervals do not have good coverage probability for ratios $Q \neq 1$. This is especially the case for the *SQ* interval and for values of *Q* which are well away from one.

The method of Lyles et al (2007) can be used to approximate the power of the difference test. This approximation is very accurate for the LN and SQ analysis.

The *LN* or *SQ* analysis therefor seems to be the method of choice for all simulation distributions. They have good size for all relevant parameter combinations, their power is comparable to the other analysis methods, a generalized confidence interval has good properties when it is used for difference testing, and an approximate quick method can be employed for a prospective power analysis. Because the *LN* generalized confidence interval has somewhat better properties than the *SQ* interval, the *LN* analysis method seems to be the method of choice.

4 Results for equivalence testing

4.1 General remarks on equivalence testing

A difference test aims to reject the null hypothesis of no difference, i.e. in the current setting to reject the hypothesis that Q=1. Poorly designed experiments with low levels of replication may have low statistical power of finding a true difference. An equivalence test on the other hand employs a null hypothesis of non-equivalence, i.e. that the ratio Q is smaller, or larger, than some pre-described equivalence limit, also called limit of concern (*LOC*). Rejection of the non-equivalence hypothesis implies that the ratio is larger than the *LOC* and this can be regarded as a "proof of safety". The advantage of equivalence testing is therefore that the onus is placed back on to those who wish to demonstrate the safety of GMOs to do high quality, well-replicated experiments with sufficient statistical power (Perry et al, 2009). Note that both the difference and equivalence test can be implemented by constructing a confidence interval for the ratio of the means of the *GM* plant and its comparator. When there is both an lower and an upper Limit of Concern, the two one-sided tests (TOST) approach of Schuirmann (1987) for equivalence testing can be used.

In the sequel results for a one-sided equivalence test, with significance level 5%, are given where the limit of concern is smaller than one. The null hypothesis is thus $H_0: Q \le LOC$ with alternative hypothesis $H_1: Q > LOC$. Different limits of concern are used in different sections.

All results are based on the generalized confidence interval for LN and SQ and on the ordinary interval for $\log(Q)$ for the GLM-like models where the standard error is scaled by Pearsons Chi-squared if appropriate. An alternative would be to use a likelihood ratio interval.

4.2 Size of equivalence test

The simulated size of the one-sided equivalence test is available for those *LOC* values which are equal to the ratio Q. For Q=1 the equivalence test equals the one-sided difference test; results for the simulated size of the two-sided difference test are already given in Section 0. Results for values Q=0.75, Q=0.50 and Q=0.25 are given in Appendix 2 A-D. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed $\alpha=0.05$. The red lines denote values 0.033 and 0.067 which provide a range that could be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK and such values are denoted by open circles. Values outside this range are denoted by filled circles, while values larger than 0.096 are given by triangles.

The LN and SQ generalized confidence interval have a generally bad simulated significance level, especially for smaller values of Q. This is in accordance with findings in section 3.5, and the LN and SQ interval will further not be discussed. Furthermore the P1, P2 and GMinterval have very similar simulated sizes; only the P1 analysis method will therefore be considered in the sequel.

For the *OP*, *NB* and *P1* intervals and Q=0.50, Appendices 2 A-D are summarized in Figure 22 to Figure 25. For Q=0.25 the appendices are summarized in Figure 26 to Figure 29. The

symbols in these figures have the following meaning: *open circle* denotes that the test is conservative for lower levels of replication and has the correct size for larger replication; *closed circle* symbolizes that the test has correct size for all replication levels; *cross* means that the test is mainly progressive; *number* denotes that the test has correct size for levels of replication smaller than or equal to the plotted number; / means that for larger levels of replication the test is progressive or that the test is progressive for some other replications; \ indicates that the test is progressive for small replication and has the correct size for larger replication.





Figure 23: Summary of size of equivalence test for *Q=LOC=*0.5; data simulated by Negative Binomial (see text for explanation of symbols)



Figure 24: Summary of size of equivalence test for *Q=LOC=*0.5; data simulated by Poisson-Lognormal (see text for explanation of symbols)

	Overdispersed Poisson					Negative Binomial						Power(1.5)								
CV-5 -	/ / *	6	* `	/		1	/	x	× >	(X	×	×	/		X	x		- ×		
CV-4 –	• / ×	/	× /			1	/	×	×				/		×`					
CV-2 -	•••••	••	•••••		•	0	/	x :	*				•	/`				•	,	
Cv-1 -	 					1	1					-		1						
	0.5 1 2	⁵ μ	10 20	50	100	0.5	1	2	5 1 μ	0 20) 50	100	0.5	1	2	5 1 μ	0 20	50) 10)0



Figure 25: Summary of size of equivalence test for *Q=LOC=*0.5; data simulated by Power(1.5) model (see text for explanation of symbols)

Figure 26: Summary of size of equivalence test for *Q=LOC=0.25*; data simulated by Overdispersed Poisson (see text for explanation of symbols)



Figure 27: Summary of size of equivalence test for *Q=LOC=0.25*; data simulated by Negative Binomial (see text for explanation of symbols)



Figure 28: Summary of size of equivalence test for *Q=LOC=0.25*; data simulated by Poisson-Lognormal (see text for explanation of symbols)





Figure 29: Summary of size of equivalence test for *Q=LOC=0.25*; data simulated by Power(1.5) model (see text for explanation of symbols)

For Q=0.50 and data simulated by means of the overdispersed Poisson, the intervals according to *OP* and *NB* have a better simulated significance levels than the intervals according to *P1* (Figure 22). For data simulated by means of the negative binomial distribution, there is not much difference between the analysis methods (Figure 23). For Poisson-Lognormal data the *OP* interval has the edge over the other intervals; the *NB* interval is only good for low coefficients of variation (Figure 24). For data simulated according to the Power(1.5) model, the *OP* interval seems to outperform the other analysis models somewhat (Figure 25).

For Q=0.25 and data simulated by means of the overdispersed Poisson, the intervals according to *OP* have a better simulated significance levels than the intervals according to *NB* or *P1* (Figure 26). For data simulated by means of the negative binomial distribution, the *P1* interval performs best while the *OP* interval can be somewhat conservative for larger values of μ (Figure 27). For Poisson-Lognormal data the *P1* interval has the edge over the *OP* interval; the *NB* interval is only good for low coefficients of variation (Figure 28). For data simulated according to the Power(1.5) model, again the *P1* interval has the edge over the *OP* interval (Figure 29).

Conclusion

The *LN* and *SQ* generalized confidence intervals cannot be recommended for equivalence testing. The *P1*, *P2* and *GM* intervals have very similar simulated significance levels. The size of the *NB* interval is particularly worse than that of the other intervals for data simulated according to the Poisson-Lognormal model. The *NB* interval is not better than *OP* and *P1* for the other simulation distributions. For Q=0.5 the *OP* interval seems to have the edge over the *P1* (and this also the case for the *P2* and *GM*) intervals. However for Q=0.25 it is the other way around because the *OP* interval is then somewhat more conservative for certain parameter combinations.

So with respect to size the *P1* (or *P2* or *GM*) and *OP* intervals can be recommended for equivalence testing. The size of both these interval is only problematic for smaller means and larger coefficients of variation.

4.3 Power of equivalence test

Appendices 2 E-H display the power of the one-sided equivalence test for a hypothetical one-sided limit of concern LOC=0.5. Each small plot has a range of 0 to 1 along the y-axis. The red horizontal lines denote power values of 0.25, 0.50 and 0.75. The simulated probability to reject the null-hypothesis of non-equivalence is given by the black dots and the dark grey area in the plot. The probability to decide that "equivalence is more likely than not" is given by the grey area. The red points denote the cumulative probability to reject the null hypothesis or to decide that equivalence is more likely than not. Finally the light grey area denotes the simulated probability that all observations equal zero; these are only present for low values of μ in combination with large CV values. The light grey area can also considered to represent a decision that "equivalence is more likely than not", if that is the case the green dots denote the cumulative probability of equivalence or equivalence more likely than not. In the sequel the "strict test" stands for equivalence while the "liberal tests" stands for the cumulative probability than not.

There are separate plots for Q=1, Q=0.75, Q=0.5 and Q=0.25. Because the tests based on the *LN* and *SQ* intervals are generally progressive (section 4.2) the power for these tests is larger than for the other models which have a more correct size. The power for the *P1*, *P2* and *GM* tests are very similar. Restricting to those parameter combinations for which the *P1* power is larger than 0.5, the difference for the strict test is maximally 0.01 between *P1*, *P2* and *GM*, while for the liberal test the difference is maximally 0.08. For the same subset the difference between *P1* and *OP* is maximally 0.017 for the strict test of equivalence and 0.027 for the liberal test.

A special case is an effect size Q=0.5 in combination with a limit of concern LOC=0.5. For such cases it is expected that the liberal test will be rejected with a probability of 50%. This is indeed the case, see e.g. Appendix 2 E9-12. Only for small means μ with large *CV* values there is some deviation from the 50% probability.

It is interesting to see that for an effect size Q=0.75 and small means μ , combined with low replication levels, there is still some probability to reject the liberal hypothesis, i.e. there is a probability of around 25% to decide for "equivalence more likely than not", see e.g. Appendix 2 E13-16.

Conclusion

In the preceding section it was found that the intervals based on *P1* (or *P2* or *GM*) and *OP* have the best simulated significance levels. Here it is shown that these intervals results in very similar power for power values that matter, i.e. values larger than 0.5.

4.4 Approximate power of the equivalence test

The method of Lyles et al (2007), used in section 3.6 for difference testing, can also be used to approximate the power of equivalence tests. The relevant calculation are, in addition to those presented in section 3.6, as follows. A critical value t_{crit} is obtained from Students t-distribution with 2*N*-2 degrees of freedom. i.e. $P(t_{2N-2} > t_{crit}) = \alpha$. Furthermore a test

statistic *T* is calculated by $T = \sqrt{N} (\beta - LOC)/se$. The power of the equivalence test is then approximated by means of the upper normal probability $P(u > t_{crit} - T)$.

Results are presented in Appendices 2 I-L, only for the *OP* interval. The different colours represent different limits of concern: LOC=0.75 (black), LOC=0.50 (red), LOC=0.25 (green), LOC=0.10 (blue). The dots denote the simulated values, while the lines represent the approximate values. The pages are for different values of the effect size *Q* as given in the title of the page. When data are simulated according to the overdispersed Poisson distribution (Appendix 2 I) the approximation is very good especially for larger power values. The same holds for data simulated by means of the negative binomial distribution (Appendix 2 J) and the Power(1.5) distribution (Appendix 2 L). However for the Poisson-Lognormal distribution the approximation is not good (Appendix 2 K).

Conclusion

The method of Lyles et al (2007) can be used to approximate the power of a one-sided equivalence test when using the *OP* interval, except when the simulation distribution is Poisson-Lognormal. For the other simulation distributions there is no need to perform a simulation study to approximate the power.

4.5 Method of choice for equivalence testing

The *LN* and *SQ* generalized confidence intervals should, in general, not be used for equivalence testing because they are too progressive, i.e. they result in too many rejections of the null hypothesis of non-equivalence.

The simulated significance level of the *OP* and *P1* (or *P2* or *GM*) intervals outperform that of the *NB* interval when data are simulated by the Poisson-Lognormal distribution. There are only small differences between the power of the *OP* and *P1* (or *P2* or *GM*) intervals for power values that matter.

It is thus hard to discriminate between the *OP* and *P1* intervals. Since the *OP* analysis method is more generally used and widely accepted, as opposed to the maybe more esoteric *P1* analysis, the *OP* analysis is recommended.

The method of Lyles et al (2007) can be used to approximate the power of the one-sided equivalence test using the *OP* interval, except when the simulation distribution is Poisson-Lognormal. This approximate method might also work for the *P1* interval but this was not investigated.

5 Zero inflation

5.1 Introduction

In practice the number of zero observations can be larger than predicted by the count distribution. This is termed excess-zeros or zero-inflation. Examples of situations with excess-zeros are given by Cunningham and Lindenmayer (2005), Sileshi (2008) and Lewin et al (2010). Failure to account for zero inflation in a statistical analysis may results in biased estimation of environmental effects of GM plants. Goedhart (2013, 2014) describes the common way to model zero-inflation.

Having a lot of zero observations in itself does not necessarily mean that a zero-inflated model is needed. For instance the negative binomial distribution with a large coefficient of variation and a not too large mean is capable of generating many zeros along with some large observations. As an example, 10 samples of size 10 are given below which are simulated by means of a negative binomial distribution with mean μ =5 and coefficient of variation *CV*=300. Clearly many zeros can be accompanied by few large observations.

_										
	0	0	0	0	0	0	0	0	0	9
	0	0	0	0	0	0	6	7	22	39
	0	0	0	0	0	0	0	4	13	14
	0	0	0	0	0	1	3	13	13	25
	0	0	0	0	0	0	1	1	3	9
	0	0	0	0	0	1	1	1	5	7
	0	0	0	0	0	0	4	7	12	23
	0	0	0	0	0	0	0	4	41	61
	0	0	0	0	0	0	0	3	4	60
	0	0	0	0	0	0	0	0	0	20

Consequently it can be hard, especially for small samples sizes, to distinguish between a zero inflated distribution and an ordinary non-inflated distribution.

5.2 A zero-inflated negative binomial distribution and its non-inflated counterpart.

Consider a zero-inflated negative binomial distribution with parameters π , μ and dispersion parameter σ^2 . Note that this distribution has mean $(1 - \pi)\mu$. To see whether such a distribution can be distinguished from a non-inflated negative binomial distribution a large number of observations, 10000, are simulated from the zero-inflated distribution. The noninflated negative binomial distribution was then fitted to this large sample yielding fitted probabilities. This was done for $\pi = 0.5$ and a variety of means μ and dispersion parameters σ^2 . Results are given in Figure 30. From this it seems clear that it will only be possible to discriminate between the two distribution for large μ and small dispersion σ^2 . For other values an ordinary negative binomial distribution, with the same overall mean, can be used instead.

Figure 30: Theoretical zero-inflated negative binomial cumulative distribution (black line) and fitted non-inflated negative binomial cumulative distribution (red circles) for various means μ and dispersion parameters σ^2 .



5.3 Size of the LN difference test when there is zero inflation

To investigate whether the t-test after a logarithmic transformation is also has correct size when there is zero inflation an additional small simulation study was performed. More specifically it is studied whether the simulated level of the t-test is close to its assumed level when the distribution of the two samples is identical and zero-inflated. This was done by simulating from 7 different count distributions: Poisson, overdispersed Poisson with dispersion parameter 3, 4 and 8, and negative binomial with dispersion parameter 1, 2 and 4. Note that, instead of specifying a coefficient of variation, in this simulation study the dispersion parameter itself is specified. Mean λ values of 2, 4, 10, 20, 40 and 80 were employed with an additional zero-inflation probability of δ =0.5. The mean of a zero-inflated distribution equals $\mu = (1 - \delta)\lambda$, implying that mean μ values of 1, 2, 5, 10, 20 and 40 are used here. For each parameter combination 1000 datasets are simulated and a two-sided t-test was performed on the log transformed counts. The simulated significance levels are given in Figure 31. Even in this case the simulated level of the t-test, using the log transformed count, is good except for small levels of replication in combination with a large overdispersion. In such cases the t-test is generally conservative rather than progressive, with the exception of the overdispersed Poisson distribution with small levels of replication and large means μ . So even in this case the simulated significance level of the *LN* analysis is generally good.

Figure 31: Simulated level of the *LN* test when data are simulated according to a zero-inflated distribution. The green line represent the theoretical 5% level. Open circles when the simulated level is within the expected range as given by the red lines.



5.4 Power of the NB difference test for negative binomial data

To evaluate the effect of excess zeros on the power of the ordinary likelihood ratio test a separate simulation with the excess zero negative binomial distribution was executed. Again a single trial without blocking with a single measurement was assumed. Furthermore a multiplicative ratio Q=2 was used between the GM plant and the comparator. The excess zero probability was set to $\delta = 0, 0.1, 0.2$ and 0.5. The mean $(1 - \delta)\mu$ of the zero inflated distribution was set to 1, 5 and 40 ensuring that the means of the distributions are identical for different values of δ . The data were analysed with the negative binomial distribution as if there were no excess zeros. The power for different levels of replication are given in Figure 32. This indicates that for small means and small excess zero probabilities the power is not much affected. However for larger means there can be a considerable decline of the power. For an excess probability of $\delta = 0.5$ and larger means the resulting distribution has a spike at zero in combination with larger values with not very much in between. In such a situation the estimate of the dispersion parameter becomes very large so as to "catch" both the zeros and the larger observations. Consequently the distinction between the means of the comparator and the GMO disappears resulting in very low power values. In such a case the data should possibly be analysed by means of an excess zero distribution.

Figure 32: Power of a difference test with $\alpha = 0.05$ for negative binomial data with overdispersion parameter $\omega = 0.25$ and additional excess zeros with probability $\delta = 0$ (black), 0.1 (red), 0.2 (blue) and 0.5 (green). The comparator has mean $\mu(1 - \delta)$ and the GM plant has a mean of $2\mu(1 - \delta)$.



6 Conclusion

An important note is that the conclusions below pertain to the situation in which a GM plant is compared with a single counterpart in a completely randomized field experiment with a single count as response. It is however likely that

For difference testing the *LN* or *SQ* method seems to be the method of choice with excellent size for parameter combinations with sufficient power. The power of these tests is generally comparable to that of the other models. So even when data are simulated according to say the overdispersed Poisson distribution, it is still best to perform a difference test on the log or squared root transformed counts. The difference test can probably best be communicated by a confidence interval as this visualizes the result of the difference test. When this is indeed the case, the *LN* method has the advantage over the *SQ* method because the *LN* generalized confidence interval has somewhat better properties. However this interval can and should not be used for equivalence testing as it only has good properties under the null hypothesis of no difference. An approximate method, employing an expanded dataset, is available to quickly calculate the power of the *LN* test making a simulation study superfluous.

For equivalence testing the situation is less clear cut. Two competing methods, *OP* and *P1*, perform equally well with respect to size and power of the one-sided equivalence test. However since the *OP* analysis is more generally used and widely accepted, the *OP* analysis is recommended. It must be considered though that the size of the equivalence test is somewhat problematic for smaller means and lager coefficients of variation. Figure 22 to Figure 29 might be used to provide a guideline for which parameter combinations the *OP* equivalence test has the correct size. Also for the one-sided equivalence test a fast method to calculate the power is available, except when the simulation distribution is Poisson-Lognormal.

Zero inflation, i.e. more zeros than predicted by the count distribution, can be a problem. However for small sample sizes it might be difficult to discriminate between a zero-inflated distribution and a non-inflated distribution. A small simulation study suggests that, for the negative binomial distribution, it is only possible to discriminate between the two for large means and small coefficients of variation. Another simulation study indicates that the power will be heavily affected for larger mean counts combined with a large excess zero probability.

7 References

- Aban IB, Cutter GR & Mavinga N (2009). Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. Computational Statistics and Data Analysis, 53(3): 820-833.
- Altman D & Bland JM (1995). Absence of evidence is not evidence of absence. British Medical Journal, 311: 485.
- Buse, A (1982). The likelihood ratio, Wald and Lagrange multiplier tests: an expository note. The American Statistician, 36(3), 153-157.
- Chen Y-H & Zou X-H (2006). Interval estimates for the ratio and difference of two lognormal means. Statistics in Medicine, 25, 4099-4113.
- Comas J, Lumbierres B, Pons X, Albajes R (2013). Ex-ante determination of the capacity of field tests to detect effects of genetically modified corn on nontarget arthropods. Journal of Economic Entomology, 106(4), 1659-1668.
- Cunningham RB & Lindenmayer DB (2005). Modeling count data of rare species: some statistical issues. Ecology, 86(5): 1135 1142.
- Demidenko E (2008). Sample size and optimal design for logistic regression with binary interaction. Statistics in Medicine, 27(1): 36-46.
- Duan JJ, Head G, Jensen A & Reed G (2004). Effects of Transgenic Bacillus thuringiensis Potato and Conventional Insecticides for Colorado Potato Beetle (Coleoptera: Chrysomelidae) Management on the Abundance of Ground-Dwelling Arthropods in Oregon Potato Ecosystems. Environmental Entomology, 33(2): 275-281.
- EFSA (2010a). EFSA Panel on Genetically Modified Organisms (GMO). Statistical considerations for the safety evaluation of GMOs. EFSA Journal, 8(1), 1250. [59 pp.],

doi:10.2903/j.efsa.2010.1250

- EFSA (2010b). EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. EFSA Journal, 8(11): 1879. [111 pp.], doi:10.2903/j.efsa.2010.1879.
- Friede T & Schmidli H (2010). Blinded Sample Size Re-estimation with Negative Binomial Counts in Superiority and Non-inferiority Trials. Methods of Information in Medicine, 49(6): 618-624.
- Gayen AK (1950). Significance of difference between the means of two non-normal samples. Biometrika, 38, 219-247.
- Goedhart PW, Van der Voet H, Baldacchino F & Arpaia S (2013). Environmental Risk Assessment of Genetically Modified Organisms: Overview of field studies, examples of datasets, statistical models and a simulation tool. Deliverable 9.1, AMIGA project, project number 289706.
- Goedhart PW, van der Voet H, Baldacchino F & Arpaia S (2014). A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. Ecology and Evolution. Accepted.

- Hrdličková Z (2006) Comparison of the power of the tests in one-way ANOVA type model with Poisson distributed variables. Environmetrics, 17(3): 227-237.
- Krishnamoorthy K & Mathew T (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of Statistical Planning and Inference, 115, 103-121.
- Lewin WC, Freyhof J, Huckstorf V, Mehner T & Wolter C (2010). When no catches matter: Coping with zeros in environmental assessments. Ecological Indicators, 10(3): 572 583.
- Lyles RH, Lin H-M & Williamson JM (2007). A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. Statistics In Medicine, 26(7): 1632-1648.
- McCullagh P & Nelder JA (1989). Generalized Linear Models, second edition. Chapman and Hall. London.
- Miller RG (1986). Beyond ANOVA, Basics of applied statistics. Wiley. New York.
- Pearson ES & Adyanthāya NK (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika, 21, 259-286.
- Perry JN, Rothery P, Clark SJ, Heard MS & Hawes C (2003). Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. Journal of Applied Ecology, 40: 17-31.
- Perry JN, ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F & van der Voet, H (2009). Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environmental Biosafety Research, 8: 65-78.
- Schuirmann DJ (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6): 657-680.
- Shieh G (2001). Sample Size Calculations for Logistic and Poisson Regression Models. Biometrika, 88(4): 1193-1199.
- Sileshi G (2008). The excess zero problem in soil animal count data and the choice of appropriate models for statistical inference. Pedobiologica, 52(1): 1 17.
- VSN International (2012). GenStat for Windows 15th Edition. VSN International, Hemel Hempstead, United Kingdom. Web page: <u>www.GenStat.co.uk</u>.

8 Appendices

Description of Appendices

Appendix 1 Properties of Difference Tests

Appendix 1 A:	Size of difference test for Overdispersed Poisson distribution	67
Appendix 1 B:	Size of difference test for Negative Binomial distribution	71
Appendix 1 C:	Size of difference test for Poisson-LogNormal distribution	75
Appendix 1 D:	Size of difference test for Power(1.5) distribution	79
		0.2
Appendix I E:	Size of difference test for Overdispersed Poisson distribution	83
Appendix 1 F:	Size of difference test for Negative Binomial distribution	87
Appendix 1 G:	Size of difference test for Poisson-LogNormal distribution	91
Appendix 1 H:	Size of difference test for Power(1.5) distribution	95
Appendix 1 I:	Power of difference test for Overdispersed Poisson distribution	99
Appendix 1 J:	Power of difference test for Negative Binomial distribution	103
Appendix 1 K:	Power of difference test for Poisson-LogNormal distribution	107
Appendix 1 L:	Power of difference test for Power(1.5) distribution	111
Appendix 1 M:	Coverage of intervals for Overdispersed Poisson distribution	115
Appendix 1 N	Coverage of intervals for Negative Binomial distribution	127
Appendix 1 O:	Coverage of intervals for Poisson-LogNormal distribution	139
Appendix 1 P:	Coverage of intervals for Power(1.5) distribution	151
Appendix 1 Q:	Power according to Lyles for Overdispersed Poisson distribution	163
Appendix 1 R:	Power according to Lyles for Negative Binomial distribution	168
Appendix 1 S:	Power according to Lyles for Poisson-LogNormal distribution	173
Appendix 1 T:	Power according to Lyles for Power(1.5) distribution	178

64

Appendix 2 Properties of Equivalence Tests

Appendix 2 A:	Size of equivalence test for Overdispersed Poisson distribution	183
Appendix 2 B:	Size of equivalence test for Negative Binomial distribution	195
Appendix 2 C:	Size of equivalence test for Poisson-LogNormal distribution	207
Appendix 2 D:	Size of equivalence test for Power(1.5) distribution	219

Appendix 2 E:	Power of equivalence test for Overdispersed Poisson distribution	231
Appendix 2 F:	Power of equivalence test for Negative Binomial distribution	247
Appendix 2 G:	Power of equivalence test for Poisson-LogNormal distribution	263
Appendix 2 H:	Power of equivalence test for Power(1.5) distribution	279
Appendix 2 I:	Power according to Lyles for Overdispersed Poisson distribution	295
Appendix 2 J:	Power according to Lyles for Negative Binomial distribution	299
Appendix 2 K:	Power according to Lyles for Poisson-LogNormal distribution	303
Appendix 2 L:	Power according to Lyles for Power(1.5) distribution	307

Description of Appendices

All Appendices

The distribution used to simulate the data is given in the main heading of each page. The abbreviated simulation distribution is also given in the title of each separate plot along with the mean μ of the distribution in parenthesis. The columns across the page either represent 1/ different analyses methods with which the data are analysed, or 2/ different means μ for which results are displayed. In the first case the abbreviated analysis method is also given in the title of each plot after the line (-) symbol. The abbreviations are as follows:

- LN analyses after a Log transformation (1 added to the data before transformation)
- SQ analyses after a Squared Root transformation
- OP Overdispersed Poisson (used for simulation and analyses)
- NB Negative Binomial (used for simulation and analyses)
- P1 Power model with p=1.5 in the variance function (used for simulation and analyses)
- P2 Power model with p=1.7 in the variance function (used for analysis only)
- GM analysis with a Gamma distribution (0.001 added to the data)
- PL Poisson-LogNormal distribution (used for simulating data only)

So for example the title "OP(0.5) - LN" denotes that the data are simulated with an Overdispersed Poisson distribution with mean μ =0.5 and are analysed after a Logarithmic transformation of the data.

The five separate small plots along the y-axis display results for five different values of the coefficient of variation (*CV*) of the simulation distribution. Note that these five *CV* values are different for different values of the mean μ . In some plots the real *CV* values are given, in others they are coded by *CV*-1 (small) to *CV*-5 (large). Results in each plot are for the 10 different replication levels 4, 6, 8, 10, 15, 20, 30, 40, 60 and 100 which are given along the x-axis of each small plot.

Appendix 1 A - D: Size of difference test: scaling by the mean deviance or by Pearson

Displays the simulated size (i.e. significance level) of the difference test with α =0.05 when the test statistic (i.e. the deviance difference) is scaled by the mean deviance or Pearson's statistic of the full model. The two scaling's are denoted by respectively "dev" and "pear" in the heading of each plot. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed significance level α =0.05. Red lines are drawn at values 0.033 and 0.067, and these provide a range which could be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK; such values are depicted by open circles. Values outside this range are denoted by filled circles, while values larger than 0.096 are given by triangles.

Appendix 1 E - H: Size of difference test: comparison different analysis methods

Displays the simulated size of the difference test with α =0.05 for seven different analysis methods. For the analysis models OP, P1, P2 and GM the test statistic (i.e. the deviance difference) is scaled by Pearson's statistic. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed significance level α =0.05. Red lines are drawn at values 0.033 and 0.067, and these provide a range which could

be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK; such values are depicted by open circles. Values outside this range are denoted by filled circles, while values larger than 0.096 are given by triangles.

Appendix 1 I – L: Power of difference test: comparison of different analysis methods

Displays the power of the difference test for an effect size of Q=0.75 (black dots and line), Q=0.5 (red dots and line) and Q=0.25 (green dots and line). For the analysis models OP, P1, P2 and GM the test statistic (i.e. the deviance difference) is scaled by Pearson's statistic. Each small plot has a range of 0 to 1 along the y-axis. The grey horizontal lines denote power values of 0.25, 0.50 and 0.75. Values for progressive tests, i.e. when the simulated size as given in the corresponding Appendix 1 E-H is larger than the upper limit of 0.067, are not displayed.

Appendix 1 M - P: One minus Coverage Probability of confidence intervals

Displays one minus the coverage probability of 95% confidence intervals after an LN, SQ, Op, NB, P1 and GM analysis. The LN and SQ intervals are generalized confidence intervals constructed by simulation. The other confidence intervals are based on the parameter estimate for the ration of the two means and its standard error. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed α =0.05. The red lines denote values 0.033 and 0.067 which provide a range that could be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK; such values are denoted by open circles. Values outside this range are denoted by filled circles, while values larger than 0.096 are given by triangles. The latter values indicate that the interval has a larger coverage probability than 5%

Appendix 1 Q - T: Power of LN difference test using the method of Lyles et al

Displays the power of the LN difference test for an effect size of Q=0.75 (black dots and line), Q=0.5 (red dots and line) and Q=0.25 (green dots and line). The dots denote the simulated power, employing 1000 datasets, and the lines are generated by means of the method of Lyles et al which used a single synthetic dataset. Each small plot has a range of 0 to 1 along the y-axis. The grey horizontal lines denote power values of 0.25, 0.50 and 0.75. Simulated powers for progressive tests, i.e. when the simulated size as given in the corresponding Appendix 1 E-H is larger than the upper limit of 0.067, are not displayed.

Appendix 2 A - D: Size of one-sided equivalence test

Displays the size of the one-sided equivalence test with α =0.05 for a limit of concern *L* which is equal to the effect size *Q*=0.75, *Q*=0.5 and *Q*=0.25. Each small plot has a range of 0 to 0.1 along the y-axis. The green line is halfway each small plot and denotes the assumed α =0.05. The red lines denote values 0.033 and 0.067 which provide a range that could be expected when 1000 datasets are simulated. So simulated sizes within the red lines are OK; such values are depicted by open circles. Values outside this range are denoted by filled circles, while values larger than 0.096 are given by triangles.

Appendix 2 E - H: Power of one-sided equivalence test

Displays the power of the one-sided equivalence test for a hypothetical limit of concern of L=0.5 and $\alpha=0.05$ for effect sizes Q=1, Q=0.75, Q=0.5 and Q=0.25 as given in the mean heading of each page. Each small plot has a range of 0 to 1 along the y-axis. The red horizontal lines denote power values of 0.25, 0.50 and 0.75. The dark grey area under the black dots denotes the simulated probability of rejecting the null-hypothesis of non-equivalence, i.e. the probability of concluding that the GMO and the comparator are equivalent. The grey area under the red dots denotes the probability of concluding "equivalence more likely than not". The light-grey area under the green dots denotes the probability that the GMO and the comparator samples only consists of zeroes. This can also be viewed as "equivalence more likely than not".

Appendix 2 I - L: Power of OP equivalence test using the method of Lyles et al

Displays the power of the OP one-sided equivalence test, with α =0.05, for effect sizes Q=1, Q=0.75, Q=0.5 and Q=0.25 as given in the mean heading of each page. Limits of concern are L=0.75 (black dots and lines), L=0.5 (red dots and lines), L=0.25 (green dots and lines) and L=0.1 (blue dots and lines). The dots denote the simulated power, employing 1000 datasets, and the lines are generated by means of the method of Lyles et al which used a single synthetic dataset. Each small plot has a range of 0 to 1 along the y-axis. The grey horizontal lines denote power values of 0.25, 0.50 and 0.75.



Appendix 1 A1: Size of difference test for Overdispersed Poisson



Appendix 1 A2: Size of difference test for Overdispersed Poisson



Appendix 1 A3: Size of difference test for Overdispersed Poisson



Appendix 1 A4: Size of difference test for Overdispersed Poisson



Appendix 1 B1: Size of difference test for Negative Binomial



Appendix 1 B2: Size of difference test for Negative Binomial


Appendix 1 B3: Size of difference test for Negative Binomial



Appendix 1 B4: Size of difference test for Negative Binomial



Appendix 1 C1: Size of difference test for Poisson-LogNormal



Appendix 1 C2: Size of difference test for Poisson-LogNormal



Appendix 1 C3: Size of difference test for Poisson-LogNormal



Appendix 1 C4: Size of difference test for Poisson-LogNormal



Appendix 1 D1: Size of difference test for Power(1.5)



Appendix 1 D2: Size of difference test for Power(1.5)



Appendix 1 D3: Size of difference test for Power(1.5)



Appendix 1 D4: Size of difference test for Power(1.5)



Appendix 1 E1: Size of difference test for Overdispersed Poisson



Appendix 1 E2: Size of difference test for Overdispersed Poisson



Appendix 1 E3: Size of difference test for Overdispersed Poisson



Appendix 1 E4: Size of difference test for Overdispersed Poisson



Appendix 1 F1: Size of difference test for Negative Binomial



Appendix 1 F2: Size of difference test for Negative Binomial



Appendix 1 F3: Size of difference test for Negative Binomial



Appendix 1 F4: Size of difference test for Negative Binomial



Appendix 1 G1: Size of difference test for Poisson-LogNormal



Appendix 1 G2: Size of difference test for Poisson-LogNormal



Appendix 1 G3: Size of difference test for Poisson-LogNormal



Appendix 1 G4: Size of difference test for Poisson-LogNormal



Appendix 1 H1: Size of difference test for Power(1.5)



Appendix 1 H2: Size of difference test for Power(1.5)



Appendix 1 H3: Size of difference test for Power(1.5)



Appendix 1 H4: Size of difference test for Power(1.5)



Appendix 1 I1: Power of difference test for Overdispersed Poisson



Appendix 1 I2: Power of difference test for Overdispersed Poisson



Appendix 1 I3: Power of difference test for Overdispersed Poisson



Appendix 1 I4: Power of difference test for Overdispersed Poisson



Appendix 1 J1: Power of difference test for Negative Binomial



Appendix 1 J2: Power of difference test for Negative Binomial



Appendix 1 J3: Power of difference test for Negative Binomial



Appendix 1 J4: Power of difference test for Negative Binomial



Appendix 1 K1: Power of difference test for Poisson-LogNormal



Appendix 1 K2: Power of difference test for Poisson-LogNormal


Appendix 1 K3: Power of difference test for Poisson-LogNormal



Appendix 1 K4: Power of difference test for Poisson-LogNormal



Appendix 1 L1: Power of difference test for Power(1.5)



Appendix 1 L2: Power of difference test for Power(1.5)



Appendix 1 L3: Power of difference test for Power(1.5)



Appendix 1 L4: Power of difference test for Power(1.5)



Appendix 1 M1: Coverage of LN interval for Overdispersed Poisson



Appendix 1 M2: Coverage of LN interval for Overdispersed Poisson



Appendix 1 M3: Coverage of SQ interval for Overdispersed Poisson



Appendix 1 M4: Coverage of SQ interval for Overdispersed Poisson



Appendix 1 M5: Coverage of OP interval for Overdispersed Poisson



Appendix 1 M6: Coverage of OP interval for Overdispersed Poisson



Appendix 1 M7: Coverage of NB interval for Overdispersed Poisson



Appendix 1 M8: Coverage of NB interval for Overdispersed Poisson



Appendix 1 M9: Coverage of P1 interval for Overdispersed Poisson



Appendix 1 M10: Coverage of P1 interval for Overdispersed Poisson



Appendix 1 M11: Coverage of GM interval for Overdispersed Poisson



Appendix 1 M12: Coverage of GM interval for Overdispersed Poisson



Appendix 1 N1: Coverage of LN interval for Negative Binomial



Appendix 1 N2: Coverage of LN interval for Negative Binomial



Appendix 1 N3: Coverage of SQ interval for Negative Binomial



Appendix 1 N4: Coverage of SQ interval for Negative Binomial



Appendix 1 N5: Coverage of OP interval for Negative Binomial



Appendix 1 N6: Coverage of OP interval for Negative Binomial



Appendix 1 N7: Coverage of NB interval for Negative Binomial



Appendix 1 N8: Coverage of NB interval for Negative Binomial



Appendix 1 N9: Coverage of P1 interval for Negative Binomial



Appendix 1 N10: Coverage of P1 interval for Negative Binomial



Appendix 1 N11: Coverage of GM interval for Negative Binomial



Appendix 1 N12: Coverage of GM interval for Negative Binomial



Appendix 1 O1: Coverage of LN interval for Poisson-LogNormal



Appendix 1 O2: Coverage of LN interval for Poisson-LogNormal



Appendix 1 O3: Coverage of SQ interval for Poisson-LogNormal



Appendix 1 O4: Coverage of SQ interval for Poisson-LogNormal



Appendix 1 O5: Coverage of OP interval for Poisson-LogNormal



Appendix 1 O6: Coverage of OP interval for Poisson-LogNormal


Appendix 1 O7: Coverage of NB interval for Poisson-LogNormal



Appendix 1 O8: Coverage of NB interval for Poisson-LogNormal



Appendix 1 O9: Coverage of P1 interval for Poisson-LogNormal



Appendix 1 O10: Coverage of P1 interval for Poisson-LogNormal



Appendix 1 O11: Coverage of GM interval for Poisson-LogNormal



Appendix 1 O12: Coverage of GM interval for Poisson-LogNormal



Appendix 1 P1: Coverage of LN interval for Power(1.5)



Appendix 1 P2: Coverage of LN interval for Power(1.5)



Appendix 1 P3: Coverage of SQ interval for Power(1.5)



Appendix 1 P4: Coverage of SQ interval for Power(1.5)



Appendix 1 P5: Coverage of OP interval for Power(1.5)



Appendix 1 P6: Coverage of OP interval for Power(1.5)



Appendix 1 P7: Coverage of NB interval for Power(1.5)



Appendix 1 P8: Coverage of NB interval for Power(1.5)



Appendix 1 P9: Coverage of P1 interval for Power(1.5)



Appendix 1 P10: Coverage of P1 interval for Power(1.5)



Appendix 1 P11: Coverage of GM interval for Power(1.5)



Appendix 1 P12: Coverage of GM interval for Power(1.5)



Appendix 1 Q1: Power of LN & Lyles difference test for Overdispersed Poisson



Appendix 1 Q2: Power of SQ & Lyles difference test for Overdispersed Poisson



Appendix 1 Q3: Power of OP & Lyles difference test for Overdispersed Poisson



Appendix 1 Q4: Power of NB & Lyles difference test for Overdispersed Poisson



Appendix 1 Q5: Power of P1 & Lyles difference test for Overdispersed Poisson



Appendix 1 R1: Power of LN & Lyles difference test for Negative Binomial



Appendix 1 R2: Power of SQ & Lyles difference test for Negative Binomial



Appendix 1 R3: Power of OP & Lyles difference test for Negative Binomial



Appendix 1 R4: Power of NB & Lyles difference test for Negative Binomial



Appendix 1 R5: Power of P1 & Lyles difference test for Negative Binomial



Appendix 1 S1: Power of LN & Lyles difference test for Poisson-LogNormal



Appendix 1 S2: Power of SQ & Lyles difference test for Poisson-LogNormal



Appendix 1 S3: Power of OP & Lyles difference test for Poisson-LogNormal



Appendix 1 S4: Power of NB & Lyles difference test for Poisson-LogNormal



Appendix 1 S5: Power of P1 & Lyles difference test for Poisson-LogNormal



Appendix 1 T1: Power of LN & Lyles difference test for Power(1.5)



Appendix 1 T2: Power of SQ & Lyles difference test for Power(1.5)



Appendix 1 T3: Power of OP & Lyles difference test for Power(1.5)


Appendix 1 T4: Power of NB & Lyles difference test for Power(1.5)



Appendix 1 T5: Power of P1 & Lyles difference test for Power(1.5)



Appendix 2 A1: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.75



Appendix 2 A2: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.75



Appendix 2 A3: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.75



Appendix 2 A4: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.75



Appendix 2 A5: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.5



Appendix 2 A6: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.5



Appendix 2 A7: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.5



Appendix 2 A8: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.5



Appendix 2 A9: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.25



Appendix 2 A10: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.25



Appendix 2 A11: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.25



Appendix 2 A12: Size of equivalence test for Overdispersed Poisson; LOC = Q = 0.25



Appendix 2 B1: Size of equivalence test for Negative Binomial; LOC = Q = 0.75



Appendix 2 B2: Size of equivalence test for Negative Binomial; LOC = Q = 0.75



Appendix 2 B3: Size of equivalence test for Negative Binomial; LOC = Q = 0.75



Appendix 2 B4: Size of equivalence test for Negative Binomial; LOC = Q = 0.75



Appendix 2 B5: Size of equivalence test for Negative Binomial; LOC = Q = 0.5



Appendix 2 B6: Size of equivalence test for Negative Binomial; LOC = Q = 0.5







Appendix 2 B8: Size of equivalence test for Negative Binomial; LOC = Q = 0.5



Appendix 2 B9: Size of equivalence test for Negative Binomial; LOC = Q = 0.25



Appendix 2 B10: Size of equivalence test for Negative Binomial; LOC = Q = 0.25



Appendix 2 B11: Size of equivalence test for Negative Binomial; LOC = Q = 0.25



Appendix 2 B12: Size of equivalence test for Negative Binomial; LOC = Q = 0.25



Appendix 2 C1: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.75



Appendix 2 C2: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.75



Appendix 2 C3: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.75



Appendix 2 C4: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.75



Appendix 2 C5: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.5



Appendix 2 C6: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.5







Appendix 2 C8: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.5



Appendix 2 C9: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.25



Appendix 2 C10: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.25


Appendix 2 C11: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.25



Appendix 2 C12: Size of equivalence test for Poisson-LogNormal; LOC = Q = 0.25



Appendix 2 D1: Size of equivalence test for Power(1.5); LOC = Q = 0.75



Appendix 2 D2: Size of equivalence test for Power(1.5); LOC = Q = 0.75







Appendix 2 D4: Size of equivalence test for Power(1.5); LOC = Q = 0.75



Appendix 2 D5: Size of equivalence test for Power(1.5); LOC = Q = 0.5



Appendix 2 D6: Size of equivalence test for Power(1.5); LOC = Q = 0.5



Appendix 2 D7: Size of equivalence test for Power(1.5); LOC = Q = 0.5



Appendix 2 D8: Size of equivalence test for Power(1.5); LOC = Q = 0.5



Appendix 2 D9: Size of equivalence test for Power(1.5); LOC = Q = 0.25



Appendix 2 D10: Size of equivalence test for Power(1.5); LOC = Q = 0.25



Appendix 2 D11: Size of equivalence test for Power(1.5); LOC = Q = 0.25



Appendix 2 D12: Size of equivalence test for Power(1.5); LOC = Q = 0.25



Appendix 2 E1: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 1



Appendix 2 E2: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 1



Appendix 2 E3: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 1







Appendix 2 E5: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.75



Appendix 2 E6: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.75



Appendix 2 E7: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.75







Appendix 2 E9: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.5



Appendix 2 E10: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.5















Appendix 2 E14: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.25







Appendix 2 E16: Power of equivalence test for Overdispersed Poisson; LOC=0.5; Q = 0.25







Appendix 2 F2: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 1



Appendix 2 F3: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 1







Appendix 2 F5: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.75



Appendix 2 F6: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.75


Appendix 2 F7: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.75







Appendix 2 F9: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.5



Appendix 2 F10: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.5



Appendix 2 F11: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.5







Appendix 2 F13: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.25



Appendix 2 F14: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.25



Appendix 2 F15: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.25



Appendix 2 F16: Power of equivalence test for Negative Binomial; LOC=0.5; Q = 0.25







Appendix 2 G2: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 1















Appendix 2 G6: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.75



Appendix 2 G7: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.75







Appendix 2 G9: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.5



Appendix 2 G10: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.5



Appendix 2 G11: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.5



Appendix 2 G12: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.5







Appendix 2 G14: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.25



Appendix 2 G15: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.25



Appendix 2 G16: Power of equivalence test for Poisson-LogNormal; LOC=0.5; Q = 0.25



Appendix 2 H1: Power of equivalence test for Power(1.5); LOC=0.5; Q = 1



Appendix 2 H2: Power of equivalence test for Power(1.5); LOC=0.5; Q = 1



Appendix 2 H3: Power of equivalence test for Power(1.5); LOC=0.5; Q = 1



Appendix 2 H4: Power of equivalence test for Power(1.5); LOC=0.5; Q = 1



Appendix 2 H5: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.75



Appendix 2 H6: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.75







Appendix 2 H8: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.75



Appendix 2 H9: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.5



Appendix 2 H10: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.5


Appendix 2 H11: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.5



Appendix 2 H12: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.5



Appendix 2 H13: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.25



Appendix 2 H14: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.25



Appendix 2 H15: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.25



Appendix 2 H16: Power of equivalence test for Power(1.5); LOC=0.5; Q = 0.25



Appendix 2 I1: Power of OP & Lyles equivalence test for Overdispersed Poisson; Q=1



Appendix 2 I2: Power of OP & Lyles equivalence test for Overdispersed Poisson; Q=0.75



Appendix 2 I3: Power of OP & Lyles equivalence test for Overdispersed Poisson; Q=0.5



Appendix 2 I4: Power of OP & Lyles equivalence test for Overdispersed Poisson; Q=0.25



Appendix 2 J1: Power of OP & Lyles equivalence test for Negative Binomial; Q=1



Appendix 2 J2: Power of OP & Lyles equivalence test for Negative Binomial; Q=0.75



Appendix 2 J3: Power of OP & Lyles equivalence test for Negative Binomial; Q=0.5



Appendix 2 J4: Power of OP & Lyles equivalence test for Negative Binomial; Q=0.25



Appendix 2 K1: Power of OP & Lyles equivalence test for Poisson-LogNormal; Q=1



Appendix 2 K2: Power of OP & Lyles equivalence test for Poisson-LogNormal; Q=0.75



Appendix 2 K3: Power of OP & Lyles equivalence test for Poisson-LogNormal; Q=0.5



Appendix 2 K4: Power of OP & Lyles equivalence test for Poisson-LogNormal; Q=0.25



Appendix 2 L1: Power of OP & Lyles equivalence test for Power(1.5); Q=1



Appendix 2 L2: Power of OP & Lyles equivalence test for Power(1.5); Q=0.75



Appendix 2 L3: Power of OP & Lyles equivalence test for Power(1.5); Q=0.5



Appendix 2 L4: Power of OP & Lyles equivalence test for Power(1.5); Q=0.25