amiga

Assessing and Monitoring
the Impacts of Genetically modified plants on Agro-ecosystems

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 single-environment 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
http://www.biometris.nl

March 2014

Dissemination Level: Public, http://edepot.wur.nl/455503

Contents

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 three documents:

- D9.2a (this report) 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.
- For practical reasons the extensive Appendices to the D9.2b report, which contain the detailed results in graphical form, are contained in a separate document.

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 ≥5, and the coefficient of variation is ≤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 non-negative 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.

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

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

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:

1. 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 ≤ 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.

Table 2. Recommended data transformations and GLMs

Endpoint type	data transformation[1]	distribution and link function for GLM
Positive continuous x	log(x)	gamma, log
Non-negative continuous x	log(x+m), where m ≤ 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

[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.

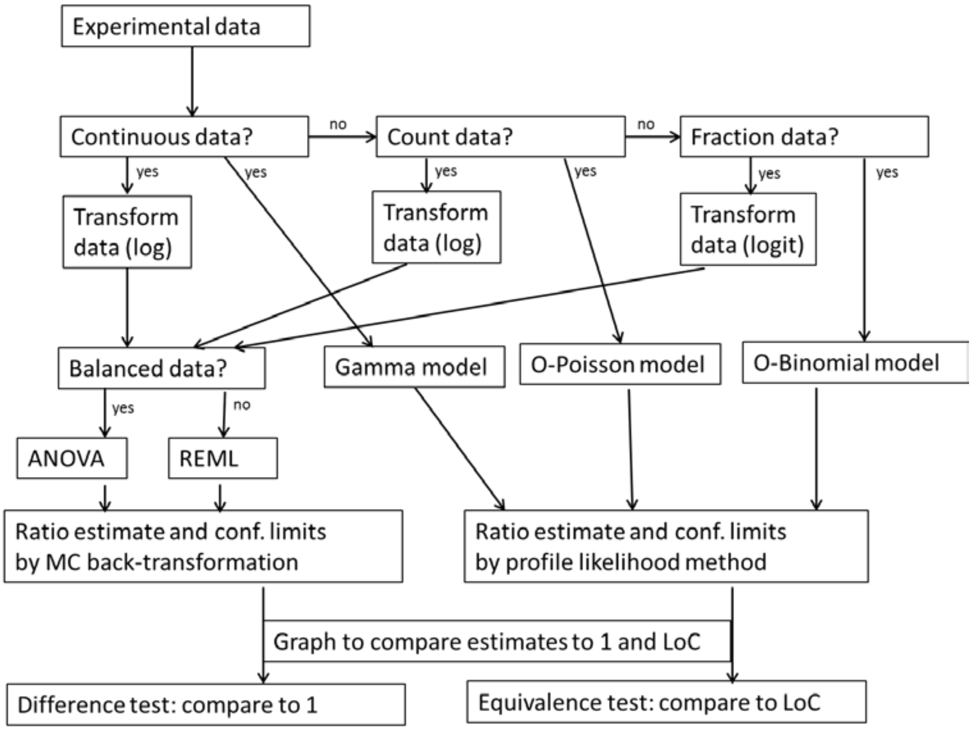
Experimental data
Continuous data?
no
Count data?
no
Fraction data?
yes
yes
yes
yes
yes
yes
Transform data (log)
Transform data (log)
Transform data (logit)
Balanced data?
Gamma model
O-Poisson model
O-Binomial model
yes
no
ANOVA
REML
Ratio estimate and conf. limits by MC back-transformation
Ratio estimate and conf. limits by profile likelihood method
Graph to compare estimates to 1 and LoC
Difference test: compare to 1
Equivalence test: compare to LoC

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).

A		B		C	
CMP	GMO	CMP	GMO	CMP	GMO
64	20	66	20	66	12
18	16	21	16	21	8
23	38	30	38	30	23
14	8	21	8	21	10
22	5	29	5	29	15
14	17	21	17	21	1
64	12	71	12	71	7
17	21	24	21	24	24
17	30	24	30	24	11
44	13	51	13	51	15
27	24	34	24	34	9
12	24	19	24	19	8
3	33	10	33	10	7
10	27	17	27	17	32
6	21	13	21	13	28
7	27	14	27	14	12
18	15	25	15	25	11
23	51	30	51	30	27
59	13	66	13	66	7
41	21	48	21	48	18

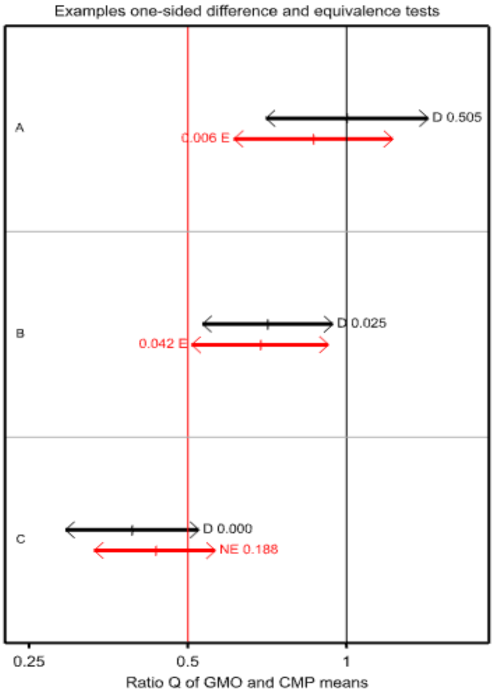

Figure 2. Three examples of count data (n=20) where there is concern for a decreased level. Limit of Concern (LOC) is 0.5 (GMO 50% of CMP, red vertical line). Bi-directed arrows represent 95% confidence intervals corresponding with two one-sided tests (TOST). P values are shown near the arrowheads for 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.

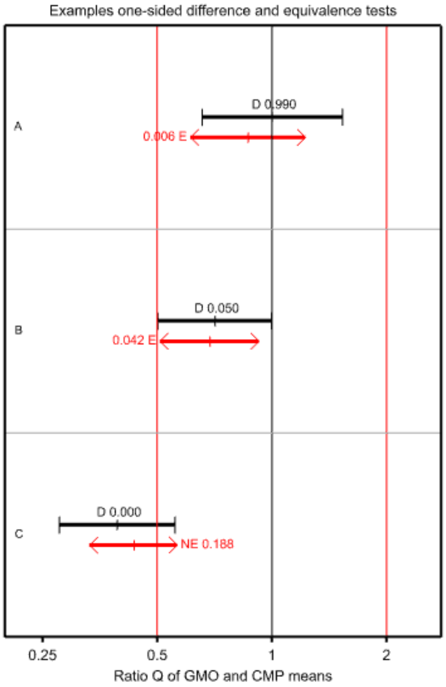

Figure 3. Same example as in Figure 2, but now with concern for decrease and increase, and two Limits of Concern, at ratios 0.5 and 2. Bars at the end of the difference interval indicate that this is a two-sided interval. The equivalence TOST interval is unchanged.

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.

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

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

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).

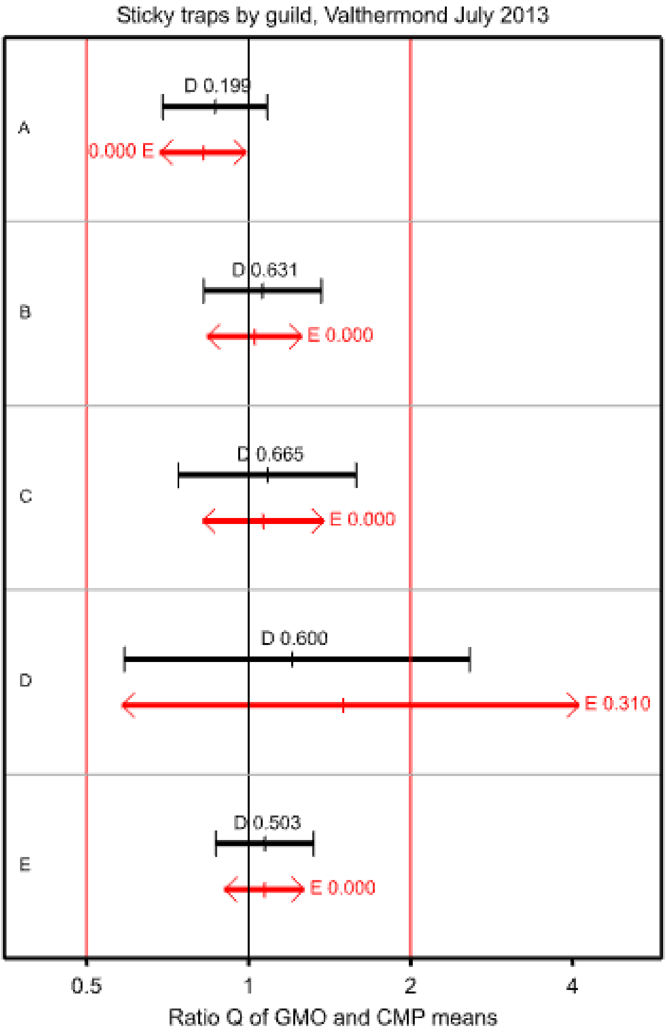

Figure 4. Analysis count data sticky traps per guild. AMIGA potato experiment Valthermond, July 2013. (A) Predators, (B) Detrivores, (C) Parasitoids, (D) Fungivores, (E) Herbivores. Limits of Concern set to 0.5 and 2 for illustration of the method only.

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: www.GenStat.co.uk.