The use of statistical tools in field testing for effects of GM plants on non-target organisms (NTOs).

CGM 2012-06

ONDERZOEKSRAPPORT

The use of statistical tools in field testing for effects of genetically-modified (GM) plants on non-target organisms (NTO).

Dr. A.V. Semenov, Microbial Ecology, CEES, University of Groningen

Supervisors:

Prof. dr. ir. J.D. van Elsas, Microbial Ecology, CEES, University of Groningen

Prof. dr. I. Pen, Theoretical Biology / Statistical methods, University of Groningen

Foreword by the chairman of the advisory committee

The assessment of potential adverse effects on non-target organisms (NTOs) is an important part of the mandatory environmental risk assessment that is carried out before the commercial release of an insect-resistant genetically modified (GM) crop on the EU market. Field trials in a variety of set ups are carried out to investigate whether the GM crop adversely affects NTOs.

The statistical analysis of such field trial data is complex, and often only to be correctly interpreted by the few experts that are specialized in the subject. The statistical analysis differs between different field trials and depends, amongst other things, on the design and set up of the field trial. An appropriate statistical analysis is important in order to detect potential adverse effects of GM crops.

In order to obtain more insight in the statistical methods that are used in current field trials and the strengths and weaknesses of these methods, the GMO Office and the Netherlands Commission on genetic modification (COGEM) have commissioned a research project on the statistical analysis of field trials on NTOs

The report that is the result of this research project contains an overview of the most important issues in the design of field trials on NTOs, the most frequently used statistical approaches and common pitfalls. In addition, a checklist is presented that is useful to assess whether field trial data were analysed correctly.

The information in the report provides an important contribution to the assessment of field trial data to detect possible adverse effects on NTOs and provides insights in the conclusions that may be drawn from the presented data.

I hope you find the content of this report of interest,

Dr P.M. Bruinenberg (AVEBE U.A.)

Chairman of the advisory committee

Advisory committee:

Dr. W.F. de Boer, Resource Ecology Group, Wageningen UR

Dr M. Bovers, COGEM staff

Dr P.M. Bruinenberg, AVEBE U.A.

Dr. D.C.M. Glandorf, GMO office

Drs. P.W. Goedhart, Biometris, Wageningen UR

Contents:

Disclaimer

Executive summary

Nederlandse samenvatting

Glossary

1. Introduction

2. Field design

    2.1. Structure of blocks in a trial experiment

    2.2. Field size

    2.3. Sample size

    2.4. Independence of samples and pseudoreplication

    2.5. Environmental parameters to be measured and indicators

3. Statistical power

4. Statistical models for data analysis

    4.1. Fixed effects models (FEM); random effects models (REM) and mixed models (MM)

    4.2. Generalized linear models (GLM) and Generalized linear mixed models (GLMM)

    4.3. Overdispersion

    4.4 Equivalence testing

5. Statistical approaches and real data

6. Checklist for field trials

7. Discussion

8. References

Appendix

Disclaimer:

This report was commissioned by COGEM and the GMO office. The contents of this publication are the sole responsibility of the authors and do not necessarily reflect the views of COGEM/GMO office.

Dit rapport is in opdracht van de Commissie Genetische Modificatie en het Bureau GGO samengesteld. De meningen die in het rapport worden weergegeven zijn die van de auteurs en weerspiegelen niet noodzakelijkerwijs de mening van de COGEM of Bureau GGO.

Executive summary

To fulfill existing guidelines, applicants that aim to place their genetically modified (GM) insect-resistant crop plants on the market are required to provide data from field experiments that address the potential impacts of the GM plants on non-target organisms (NTO's). Such data may be based on varied experimental designs. The recent statistical guidance by EFSA (2010) does not provide clear and structured suggestions for field trials. This report examines existing important field practices in GM plant field testing such as the way of randomization, replication and pseudoreplication. Emphasis is placed on the importance of the plot size used for the field trials in which NTO's are assessed. The report also emphasizes the importance of statistical power and examines the positive and negative sides of various statistical models. Equivalence and difference testing are compared, and the importance of checking the distribution of experimental data is stressed to decide on the final selection of the proper statistical model. While for continuous data (e.g. pH and temperature) classical statistical approaches – e.g. analysis of variance (ANOVA) - are appropriate, for discontinuous data (counts) only generalized linear models (GLM) are shown to be efficient. There is no golden rule as to which statistical test is the most appropriate for any experimental situation. In particular, in experiments in which block designs are used and covariates play a role, chi square tests are not according to statistical rules and GLMs should be used. Finally, the report offers generic advice to risk assessors and applicants that will help in both the setting up of field testing and the interpretation of the data obtained in field testing. The combination of decision trees and a checklist for field trials, which are provided in this report, might help risk assessors in the interpretation of the statistical analyses of field trials and helps them to assess whether these analyses were correctly applied.

The objectives of the study were:

- To review the statistical methods and approaches used by applicants for field releases of (GM) plants in relation to effects on NTO's.
- To facilitate the interpretation of statistical tools used in GM plant field tests by risk assessors.

Nederlandse samenvatting

Aanvragers die hun genetisch gemodificeerde (GG) insectenresistente planten op de markt willen introduceren moeten, om te voldoen aan de geldende richtlijnen, resultaten overleggen van veldexperimenten waarin mogelijke effecten van deze GG planten op niet-doelwitorganismen worden bestudeerd. Deze resultaten zijn vaak verkregen uit experimenten met verschillende statistische bewerkingen. Recente statistische richtsnoeren gepubliceerd door EFSA (2010) geven vooralsnog geen concrete en duidelijke suggesties of aanbevelingen ten aanzien van de te gebruiken technieken. Dit rapport onderzoekt de bestaande praktijk in veldproeven met insectenresistente GM planten, zoals de manier van warring, herhaling en pseudo-herhaling. Het rapport benadrukt het belang van de grootte van de plotjes die gebruikt worden voor het testen van effecten op niet-doelwitorganismen. Verder wordt het belang van een toets op onderscheidend vermogen (statistical power) benadrukt en worden de positieve en negatieve facetten van verschillende statistische modellen onder de loep genomen. Toetsen op equivalentie en verschil worden vergeleken, en daarnaast wordt het belang van een toets aangaande het type datadistributie aangegeven, omdat dit een belangrijke factor is bij het bepalen van het juiste statistische model. Waar voor continue gegevens (bijvoorbeeld pH en temperatuur) de klassieke statistische methoden, zoals variantieanalyse, afdoende zijn, zijn voor discontinue gegevens (tellingen) slechts veralgemeniseerde rechtlijnige statistische modellen (VRMs) verantwoord. In het bijzonder voor proefnemingen waarin blokontwerpen gebruikt worden en covariaties een rol spelen, voldoen chi-kwadraat toetsen niet aan de statistische regels en moeten VRMs worden gebruikt. Uiteindelijk geeft het rapport algemene adviezen aan risicobeoordelaars en aanvragers die van belang zijn voor zowel de proefopzet als de interpretatie van de resultaten verkregen uit de veldproeven. De combinatie van besluitbomen en een checklist met regels voor veldproeven die in dit rapport opgenomen zijn, zullen het werk van de risicobeoordelaars, waar het gaat om de evaluatie van de mogelijke impact van GM planten op niet-doelwitorganismen, faciliteren.

De doelstellingen van deze studie waren:

- Inventarisatie en evaluatie van de statistische methoden die gebruikt worden door aanvragers bij veldproeven waarmee effecten van GM planten op niet-doelwitorganismen worden onderzocht
- Het assisteren van risicobeoordelaars in hun interpretatie van de statistische methoden die gebruikt worden voor veldproeven waarmee effecten van GG planten op niet-doelwitorganismen worden onderzocht.

Glossary:

Analysis of variance (ANOVA): a collection of statistical approaches/models in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation (more details, box 8).

Beta-binomial distribution: a discrete probability distribution which is usually used for proportions but permits more variability than Binomial distribution.

Binomial distribution: a discrete probability distribution which is usually used for proportions (e.g., the mortality of larvae).

Equivalence test: test that allows to prove that a treatment has effects that are indistinguishable from those of another one and hence the two treatments are essentially equivalent. From this, it follows that any difference is of no practical consequence. While the traditional statistical null hypothesis employed is one of equality, the null hypothesis for equivalence testing is one of inequality.

Experimental design: the design of any information-gathering exercise (experiment) in which variation is present.

Exponential family: a family of statistical distributions including the normal, binomial, Poisson, exponential and gamma distributions.

Fixed effects: effects whose levels are experimentally determined or whose interest lies in the specific factors of each level, such as differences among treatments and interactions (more details, box 8).

Generalized linear models (GLM): statistical models that assume errors from the so-called exponential family; the predicted values are determined by continuous predictor variables and by the link function.

Generalized linear mixed models (GLMM): models that combine properties of linear mixed models (which incorporate random effects) and generalized linear models (which handle non-normal data by using link functions and exponential family (e.g. Poisson, binomial) distributions).

Linear mixed model (LMM): a statistical linear model that assumes normally-distributed variation and also includes fixed and random effects. An example is ANOVA incorporating a random effect.

Link function: a continuous mathematical function that defines the response of variables to predictors in a GLM. Applying the link function makes the expected value of the response linear and the expected variances homogeneous.

Maximum likelihood (ML): a statistical framework that estimates the parameters of a model that maximizes the probability of the observed data (the likelihood).

Mixed effects model (mixed model = **MM):** a statistical model containing both fixed and random effects.

Model selection: selection of the best set of candidate statistical models.

Negative Binomial distribution: a discrete probability distribution, which expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate but permits more variability than Poisson distribution.

Overdispersion: the occurrence of more variance in the data than specified by a statistical model.

Poisson distribution: a discrete probability distribution, which expresses the probability of a given number of events occurring in a fixed interval of time and/or space if these events occur with a known average rate. It is usually used for count data (e.g., the number of larvae per plant).

Pseudoreplication: the use of statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent.

Random effects: effects whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific factors of each level (more details, box 8).

Restricted maximum likelihood (REML): restricted maximum likelihood averages over the uncertainty in the fixed-effect parameters. It is used for skewed distribution of data. This in contrast to standard maximum likelihood, which estimates the standard deviations of the random effects assuming that the fixed-effect estimates are precisely correct.

Split-plot design: a blocked experiment, in which the blocks themselves serve as experimental units for a subset of factors.

Statistical power: the probability that a test that is applied will reject the null hypothesis when it is false (i.e. the probability of not committing a Type II error, or making a false negative decision).

T-test: test of the null hypothesis, which implies that the means of two normally distributed populations are equal.

Type I error: occurs when one rejects the null hypothesis when it is true. The probability of a type I error is the level of significance of the test of the hypothesis, and is denoted α (more details, box 6).

Type II error: occurs when one does not reject the null hypothesis when it is false. The probability of a type II error is denoted β (more details, box 6).

Unbalanced data: a distribution of data that is characterized by asymmetry of the probability distribution (e.g. right side tail is longer than the left side tail and the bulk of the values lie to the left of the mean).

1. Introduction

In science, there are five components to a field experiment: hypothesis, experimental design, experimental execution, statistical analysis and interpretation (Hurlbert. 1983). Obviously, the hypothesis is of primary importance, as, if it is not sound by any criterion, even a well-conducted experiment will be of little value.

In experimental work, the primary function of statistics is to show the clarity, conciseness and objectivity with which results are presented and interpreted. Statistical analysis and interpretation are critical aspects of experimentation, however, if any statistical or interpretative errors are made, the data can be reanalyzed. On the other hand, in case of an improper experimental design or errors in the execution of an experiment, the only possible solution is to repeat the experiment. There are several important issues which have to be considered to avoid the use of a wrong experimental design or the application of improper statistical analyses. These issues include the use of sufficient replicates to cover the expected variation, the use of a sufficiently large field (field size), the proper blocking of the experiment, the usage of the appropriate control (comparator), etc.

There is variation in the statistical approaches used by investigators involved in field trials with GM plants to study impact on non-target organisms (NTO's). Generally, the aim of such experiments is to make comparisons between the impacts of GM plants compared to those of their near-isogenic counterparts. Unfortunately, one often observes that such field experiments follow an improper experimental design, examples of which are the use of insufficient replicates or an improper blocking (for instance, a field separated in several unequal parts). In addition, the statistical analyses that are applied are sometimes incorrectly chosen.

In this report, we examine the statistical approaches used in studies on the effects of GM plants on NTOs in the field. These studies are characterized by several specific features, such as the high variation in the abundance of NTOs (in contrast to an analysis of species diversity) with, often, non-normal distributions. We also select some examples which highlight the positive or negative aspects of the approaches followed in each study. We observe that it is sometimes difficult to assess whether the statistical approach used was proper or not, in particular since authors in their Materials and Methods sections often omit details on this aspect (e.g., the number of replicates) (Hoss et al. 2011; Druart et al., 2011; Oliveira-Filho et al., 2011). It is important to state that the experimental design (e.g. the field lay-out, sample size, sampling method, number

of (sub-) samples and replicates and the way in which the treatments are randomized over the experimental units) defines how the data should be analyzed. In other words, an appropriate choice of the experimental design, taking into account all sources of variation and establishing replicate numbers on the basis of these, is primordial. Various NTO studies until today are questionable because there is no well-defined working hypothesis (Yamamori, 2011) as well as statistical analyses. As we deal with field studies, we will first examine the design of experiments in respect of the statistical requirements posed by the scientific question.

2. Field design

Depending on the purpose of the study, any field design for impact analysis should take into account the level of accuracy of the data needed in relation to the observed variability. Therefore, planning of field experiments without a clear prior understanding of the experimental hypothesis and how the results will be analyzed and interpreted may lead to incorrect statistical analysis and thus wrong conclusions. In particular, under/overestimated impacts of e.g. GM cultivars may be found. For instance, a half-field (a field separated in two equal parts, a common example of blocking) design in comparison to paired fields has a high potential for reduction of environmental variability and so of measured impact. The reason is that two halves of a field are more likely to be similar in previous management regime, soil type and surrounding habitat, than sites away from each other (Perry et al., 2003). Care must be taken to avoid interferences between experimental units that are closely together. Established separation distances, or buffer zones, according to agronomical rules (e.g., 50m for rape and 6m for beet; Perry et al., 2003) between half-field units help to minimize interference problems (e.g. bare soil or a reference variety).

2.1. Structure of block in a trial, randomization and replication

There are two fundamentally different cases in which data can be obtained: (1) a designed experiment (control over experimental conditions and ability to vary these conditions) and (2) an observational study (experiments in which the conditions are beyond the control of the experimenter) (Neter et al., 1990). For designed experiments, the main principles of randomization, replication and across-unit homogeneity (blocking) are important. All designed

experiments are usually set up as comparative experiments, in which a change in a variable is to be shown due to a cause (e.g. the presence of a transgene in the GM plant). A properly designed experiment must follow three rules/principles:

1) treatments are randomly allocated to experimental units to neutralize the effects of location (or other uncontrolled factors, e.g. weather effects);

2) treatments are sufficiently replicated to allow an adequate estimation of experimental error variance;

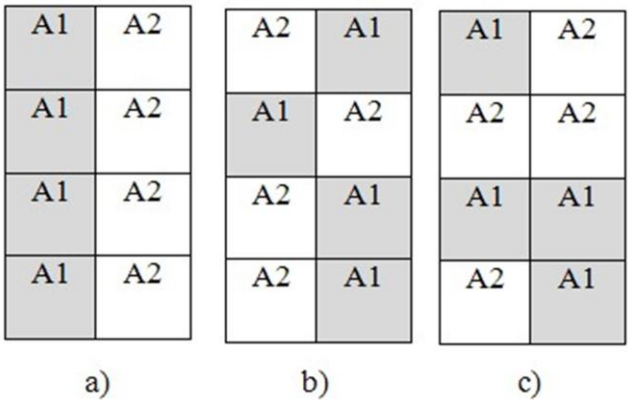
3) experimental units are (e.g. fields) grouped in homogeneous blocks prior to treatment application to minimize the impact of other controllable factors, such as difference in soil composition (Schabenberger & Pierce., 2002).

If a variable by which the experimental units should be blocked is not taken into account, the experimental design will lead to larger errors and thereby make it more difficult to find treatment effects. The resulting design might be inefficient, leading to a large error. Moreover, statistical tests might be lacking power. The third principle is the only one which is negotiable, since a completely randomized design which does not have any block factors can be more efficient than a randomized complete block design. If treatments are replicated but not randomly assigned to experimental units the data should be treated as observational, because the effect of location is not neutralized by randomization.

An observational study produces data where the independent variables are merely observed but not assigned to dependent variables. In observational studies, experiments are included in which the factors of interest (such as the level of herbicides used) have not been randomized or which have not been properly replicated. The conclusions that can be obtained from observational studies are less strict than those of designed experiments. The variables are associated with each other in a data set and often show a large degree of co-linearity or are confounding. Therefore, we cannot conclude that the differences in the values of the independent variables are at the basis of the observed differences. Thus, some NTO studies should be characterized as observational studies (e.g. Yamamori, 2011, absence treatments to with) due to a lack of (controlled) replicates and/or randomization.

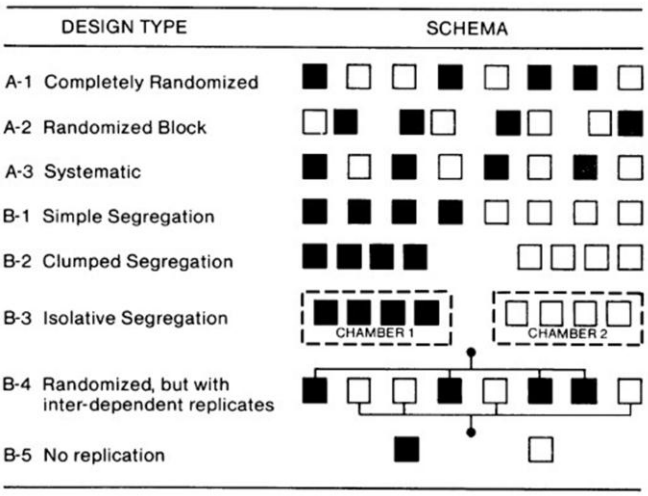
Box 1. Blocking

Linked to the type of randomization, treatment effects may interfere with non-treatment effects. For instance, in a completely randomized experimental block design, all replicates of treatment A1 may lie in the west of the field, whereas those of A2 lie in the East (a). In this case, wind or water flows from a certain direction might cause treatment differences by non-treatment effects. Such randomization might be done with a more balanced arrangement (b). East-west effects can be controlled by blocking (c) with a restriction for each treatment to appear two times in the east and two times in the west.



Box 2. Randomization

Several acceptable (A modes) and unacceptable ways of treatments (B modes) in a two-treatment experiment (shaded and unshaded). Each unit is assumed to have been treated independently of the other units in the same treatment (Hurlbert., 1984)



While blocking (Box 1) eliminates the effects of systematic factors in a certain designed experiment, randomization (Box 2) can neutralize the effects of unknown factors and allows to estimate treatment differences and variance components without systematic bias. On the other hand, replication does not necessarily lead to unbiased calculations of treatment effects (Schabenberger & Pierce., 2002).

2.2. Field size

To determine the right experimental design, measurements of variables (e.g. density of NTOs) should be taken at different locations in the field before the actual experiment is performed (e.g. in a pilot study). It is also possible to use general knowledge on the level of variation and possible distributions, e.g. from other trials or from the literature (Box 3). These measurements will provide information about the level of expected variation at the field, thus indicating an optimal experimental design, as well as the statistical approach to be used. Furthermore, it is important to know how many samples have to be taken for each field assessment, the minimum required sample size, and to determine the size of each sample (e.g. the size of the area to be sampled in one sampling). The researcher has to trade off his efforts in terms of a higher number of samples per experimental unit or lower one in exchange for more experimental units. However, a general rule of experimentation is that it is more efficient to have more experimental units with less samples per unit than less units with more samples. Thus, Clark et al (2006) compared the influence of the herbicide management of GM herbicide-tolerant and conventional crops on weed densities. They showed significant influences of the distances between experimental units (fields) on the final weed densities. They also indicated the effective number of samples which allowed distinguishing the effects of GM in comparison to those of non-GM crops. A high variability in the values of variable indicators is usually counterbalanced by increasing numbers of samples, while, for indicators with lower variability, relatively smaller numbers of samples can be used (Clark et al., 2006). This is dependent on the size of the effect. There is often a trade-off between an increase of the size of one sample (e.g. field size) and increase of the sample number (e.g. replicates).

Box 3. Field size

Dispersal rates of NTO's can directly influence the size of the field that is appropriate for adequate statistical analysis.

The relation between most common NTOs and appropriate field size to study the impact of GM plants (based on information from Prof. Dr. M. Schilthuizen, Naturalis)

Dispersal rate	Taxa	Appropriate size of the plots	References
Low	Gastropods Mites Flightless aphids Collembola	25 m^2	Schilthuizen, personal communication; Schilthuizen et al., 1994; Schilthuizen et al., 2005; Lehmitz et al., 2012; Gil et al., 2004; Auclerc et al., 2009.
Moderate	Adult Spiders Soil-dwelling beetles (e.g., ground beetles) Thrips	250 m^2	Schilthuizen, personal communication; Bonte et al., 2008; Boer, 1970; Liebherr, 1988; Morsello et al., 2008.
Fairly high	Bugs Other beetles Winged aphids	2,500 m^2	Schilthuizen, personal communication; Smith King, 1987; Hazell et al., 2005.
High	Bees Butterflies Flies Moths Juvenile spiders	25,000 m^2	Schilthuizen, personal communication; Løjtnant et al., 2012; Slatkin, 1985; Feder et al, 1994; Cameron et al., 2009; Bonte et al., 2008.

Distances (buffers) between fields should be at least the same as the plot diameter.

2.3. Sample size

A general rule for sampling is "the larger the number of samples, the more accurate the estimate of the population mean". However, sample numbers are restricted by various limitations and have to be handled in reasonable time and with reasonable investment of labor and money. The expected variation in the variable to be assayed must first be determined by analysis of a small number of samples, in a pilot experiment. Thus, the mean of the preliminary pool might be required to be within, e.g., 10% of the real population mean, since 10% is considered to be

accurate enough for most purposes (Perry et al., 2009). As indicated before, a general rule of experimentation is that it is more efficient to have more units with less samples than less units with more samples. Depending on the variation within or between plots this tradeoff is made.

Box 4. Replication

The formula $n = Z^2 S^2 / d^2$ (where n – number of replicates; Z – probability [$Z_{0.05} = 1.96$]; S^2 – error variance of samples, and d^2 – margin of error for the plot), allows to calculate a required number of samples.

Example: Number of replicates (n), if the size of the difference is 10 (d) at the 0.05 level of confidence ($Z = 1.96$), the sampling error variance is estimated as 200 (S^2) based on the samples collected:

$n = Z^2 S^2 / d^2$;

n = 3.84 x 200 / 100 = 7.68; The equation provides the minimum sample number as 7.68 (rounded off to 8).

This formula is one of the most commonly used, although it might underestimate n (Kupper et al., 1989). Another formula can be found in Cochran & Cox (paragraph 2.2); it includes a table which is helpful. However, nowadays the number of replicates can be easily calculated with many statistical packages.

2.4. Independency of samples and pseudoreplication

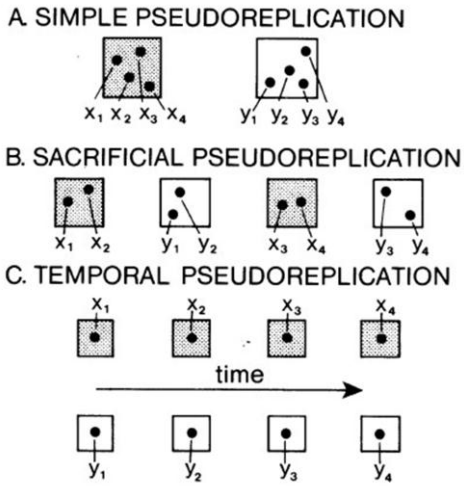
A fundamental assumption of all statistical analyses is that the data obtained from experimental studies represent independent observations of a representative sample from the population of interest. Measurements or observations are independent if the value of each observation is in no way influenced by, or related to, the value of other observations in the sample (LeBlanc, 2004). Pseudoreplication represents a typical violation of the sample independency assumption.

Replication refers to having more than one experimental (or observational) unit within the same treatment, where each unit with the same treatment is called a replicate. Most models for statistical analysis require true replication which permits the estimation of variability within a treatment. Without estimating variability within treatments, it is impossible to perform statistical

inference. True replicates are often confused with repeated measures or with pseudoreplicates. The term pseudoreplication (Hurlbert., 1984) refers to "the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated (though samples may be) or replicates are not statistically independent." The following example illustrates the way this can occur (Box 5). It is sometimes possible to deal with pseudoreplication by using the mean of the subsamples or repeated measures in GLM analysis (discussed below).

Box 5. Pseudoreplication

Let us suppose that we want to test whether cultivar 1 contains bigger larvae than cultivar 2. We take 20 larvae from a plant which represents cultivar 1 and 20 larvae from a plant of cultivar 2, and perform an independent samples t-test with a total sample size of n=40. The problem is that 20 larvae from a plant do not provide 20 pieces of independent information, as it would by taking a single larva from 20 different plants. Therefore the p-value associated with the t-test is meaningless.



The figure represents the three most common types of pseudoreplication. Shaded and unshaded boxes represent experimental units which receive different treatments. Each dot represents a sample or measurement. Pseudoreplication is a consequence (in each example) of statistical testing for a treatment effect by means of procedures which assume that the four data for each treatment have appeared from four independent experimental units. Important remark: example A cannot be analyzed properly, while B can, by taking the means for each unit.

Doing statistical inference using pseudoreplicates rather than true replicates might cause an underestimation of variability. This will result in the fact that confidence intervals are too

small and an inflated probability of a Type I error (falsely rejecting a true null hypothesis) occurs. For NTO field testing it means that the chance to reject the null hypothesis (e.g. of no difference between beetle densities on GM vs. non-GM plants, see box 6) is higher.

Box 6. Error types

Study: the effect of beetle density in a field.

Null hypothesis: There is no difference between two beetle densities

Alternative hypothesis: There is a significant difference between two beetle densities

A **type I error** occurs when one rejects the null hypothesis when it is true. The probability of a type I error is the level of significance of the test of hypothesis, and is denoted by α.

Example of type I error: If the population density is normally distributed with a mean of 180 and a standard deviation of 20, and population density over 225 is considered as environmental risk, what is the probability of a type one error?

z=(225-180)/20=2.25; the corresponding tail area is 0.0122, which is the probability of a **type I error**.

A **type II error** occurs when one rejects the alternative hypothesis (fails to reject the null hypothesis) when the alternative hypothesis is true. The probability of a type II error is denoted by β.

Example of type II error: If the population density has a mean of 300 with a standard deviation of 30, but another population has a mean of 225 or more, what is the probability of a type II error?

z=(225-300)/30=-2.5 which corresponds to a tail area of 0.0062, which is the probability of a **type II error**.

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False positive	Correct outcome True positive
Fail to reject null hypothesis	Correct outcome True negative	Type II error False negative

2.5. Environmental parameters to be measured and indicators

In the light of the complexity of the (terrestrial) environment, it is not possible to measure changes in all environmental parameters (e.g. the abundance of all species that are possibly present). Therefore, where possible, indicators have to be chosen to represent larger groups of organisms or to identify ecological processes that result in important changes over large scales of time and space. The indicators, in most cases organisms that are not *a priori* targeted by the specific GM plant (NTOs), need to respond to different crops at scales appropriate to the experimental design (e.g. 4 replications might be sufficient to detect a 50% difference with a 70% power level (explained below), but only for the most abundant NTO (e.g. arthropods) (Albajes et al., 2012). Sample sizes, in relation to the levels of variability, have to be adequate to test the assumption that there is no significant influence of GM cultivars as compared to non-GM ones. Many aspects of field experiments (e.g. field design and size, replication) have been discussed in the literature (Clark et al., 2003; Perry et al., 2003; Duan et al., 2006), but there is no agreement as to how many replications are needed to detect a difference between a GM crop and its isogenic counterpart, since it depends on the magnitude of the putative difference, the plot size, the variability in the data, the design etc. Most trials that have so far been conducted in the field have shown no effects of GM plants (Vacher et al., 2004; Post et al., 2011). However, these studies often do not include an analysis of the probability (statistical power) to reject the null hypothesis given a certain treatment effect (Cohen, 1988; Albajes et al., 2012). Therefore, it is not clear with what certainty an effect which is present would be detected.

3. Statistical power

The power of a statistical test is related to the probability to distinguish an effect (e.g., of a GM plant in comparison with its isogenic counterpart) as a function of the magnitude of the effect intended to be detected, the variability of the data and the number of values used to calculate the means. Therefore, studies should justify their sample size (size of the sample and number of replicates). An analysis of statistical power (power analysis), as part of the analysis, should be a prerequisite of every study.

Power analyses also provide the confidence that the level of replication is neither too small to detect the effects that are present, nor too great to avoid that unnecessary extra resources are used for trial experiments. Power analysis will yield the value of the probability that the test will reject the null hypothesis when it is actually false.

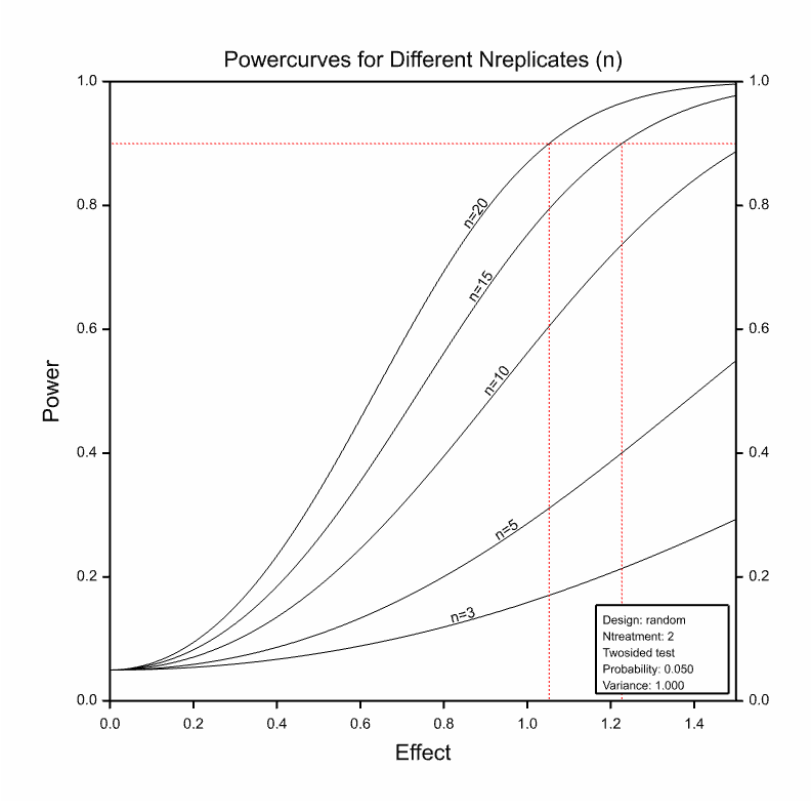


Figure 1. An example of a power curve for different levels of replication.

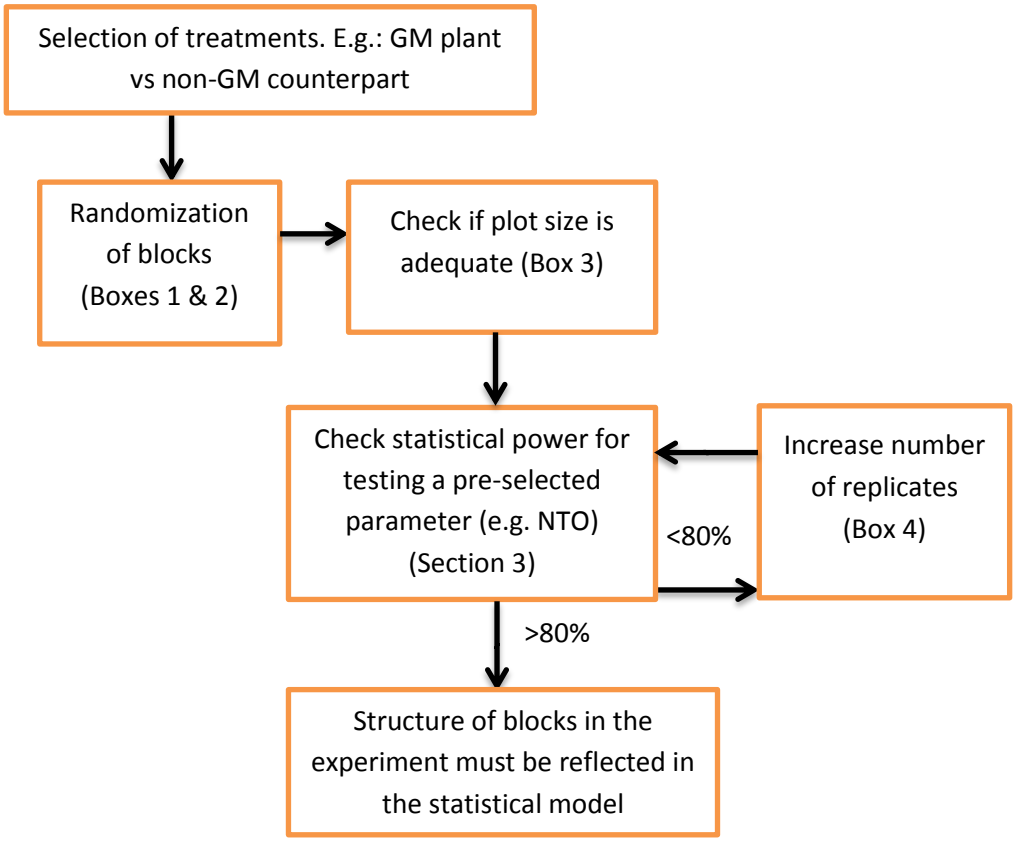
Statistical power is the ability of a test to detect an effect with a given certainty, given that the effect actually exists. At any stage in a study, the available data may be used to derive an estimate of this power for a particular variable by estimating σ^2 (indicates the variance) from an analysis of variance (ANOVA), specifying a desired number of replicates. Given the high expenditure of large experiments, studies of statistical power are vital when such experiments are planned. Power analyses give insight into the degree of replication required to achieve a certain power, usually of around 80% (Perry et al., 2003; Storkey et al., 2008). The power varies in relation to the magnitude of the effect addressed as well as the significance level and whether a one-sided or two-sided test is used. Statistical power analysis is important for difference tests

(null hypothesis of no difference between the impact of a GM plant and a non-GM plant) for each experiment done to support an environmental risk assessment (ERA).

In practice, values of 70% (Prasifka et al., 2008) and 80% (Perry et al., 2003) are commonly used in field trials as the desired statistical power. For many field trials which study NTOs, power analysis has indicated that replication of 20 fields per crop per year over 3 years (in total, 60 replicates) should have adequate power (>80%) to detect differences of 1.5 fold to detect 50% difference (Perry et al., 2003). Albajes et al. (2012) showed that four replications might be sufficient to detect a 50% difference with 70% power (for the most abundant arthropods, such as soil and plant dwelling predatory arthropods and insect parasitoids) (trials had 4 blocks for 3 years, plot size 0.5 ha) (Albajes et al., 2012).

There are several frequently used websites that allow to calculate the power of a test and the minimum required sample sizes, e.g.: http://wise.cgu.edu/powermod/power_applet.asp or http://www.divms.uiowa.edu/~rlenth/Power/.

For data that approximately follow a normal distribution, the power of standard tests (e.g., ANOVA) can be calculated routinely. Based on the importance of statistical power for trial experiments, a simple scheme can help to avoid the most common problems encountered with the set-up of a field trial, as below:



However, data such as counts, e.g. of NTOs (which can have an asymmetric frequency distribution), might require more complex statistical approaches such as computer simulations of data. Such analyses are not discussed due to the high complexity of calculations. In fact, asymmetric distributions are often the case for NTO data (e.g. non-normal distribution).

4. Statistical models for data analysis

Any data set obtained in an experimental study has a particular distribution. It is the distribution of the data (i.e. the dependent variable or the response variable) that dictates what statistical tools are appropriate to use. **Analysis of variance** (ANOVA) is a collection of statistical models, in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation (Box 7). ANOVA can only be used in case NTO data follow a normal distribution, possibly after applying a transformation such as the logarithm. ANOVA provides a statistical test of whether or not the means of several groups are all equal, and therefore generalizes the *t*-test (test of the null hypothesis that the means of two normally distributed populations are equal) to more than two groups. Doing multiple t-tests would result in an increased chance of committing a type I error (Box 6). For this reason, ANOVA is useful in comparing two or more, means. The most common use of ANOVA is a linear relation of the response to the treatments and blocks. ANOVA should follow several assumptions: 1) independency of units (section 2.4); 2) the distributions of the residuals are normal; 3) equality (or "homogeneity") of variances, 4) the variance of data in groups should be similar. In case of skewed distributions of the data, restricted maximum likelihood (REML) has to be used.

There are three classes of models used in ANOVA: Fixed effects models (FEM), random effects models (REM), mixed models (MM). Before any examples are discussed, it is important to understand these models, statistical terms involved and the approaches commonly used in field trials.

Box 7. Analysis of variance (ANOVA)

There are several types of ANOVA. Every type is used for a specific design of the experiment:

- One-way ANOVA is used to test for differences among two or more independent groups. Usually, one-way ANOVA is used to test for differences among at least three groups, since the two-group case can be covered by a t-test. When there are two groups to compare and variances between groups are equal, t-test and ANOVA are equivalent.

- Factorial ANOVA is used when the interaction effects among the treatments has to be studied.

- Repeated measures ANOVA is used when the same subjects are used for each treatment.

- Multivariate analysis of variance (MANOVA) is used when there is more than one response variable.

4.1. Fixed effects models (FEM), random effects models (REM), mixed models (MM)

The statistical tools to be used are commonly aggregated under so-called models, like FEM, MM, GLM and GLMM (see below). Application of a statistical model (e.g. a mixed model) to GM plant field testing is illustrated in Box 9.

Box 8. Fixed and Random effects

Fixed effects are effects whose levels are experimentally determined or whose interest lies in the specific factors of each level, such as differences among treatments and interactions (e.g. concentration of herbicides used).

Random effects are effects whose levels are sampled from a larger population, or whose interest lies in the variation among them rather than the specific factors of each level (e.g time). The parameters of random effects are the standard deviations of variation at a particular level (e.g. among experimental blocks).
The precise definitions of 'fixed' and 'random' are controversial; the status of particular variables depends on experimental design and context.

The distinction of fixed and random effects applies to the unknown model components:
- a fixed effect is an unknown constant (does not vary);
- a random effect is a random variable.

A **FEM** is a statistical model that represents observed quantities (a numerical property that can exist as a magnitude or multitude) in terms of explanatory variables that are treated as if they were fixed. The **FEM** applies to situations in which the experimenter applies one or more treatments to the subjects of the experiment (e.g. using two levels of herbicide vs a control without herbicide) to see if the response variable values change (e.g. the level of beetles vs the control). This is in contrast to **REM** in which explanatory variables might be treated as if they arise randomly. Such model (**REM**) assist in controlling for unobserved heterogeneity when this heterogeneity is constant over time and correlated with independent variables.

An **MM** is a statistical model containing both fixed and random effects. **MMs** are particularly useful in settings where repeated measurements are made on the same statistical parameters or if other sources of random variation (e.g. site-effects) need to be accommodated for. It means that, in most of the cases, **MM** is an appropriate model for NTO studies, since NTO are usually sampled over time in multiple plot replicates.

Box 9. Mixed model example

One of the most common mixed models is the split-plot design. The split-plot design involves two experimental factors, **A** and **B**. Levels of **A** are randomly assigned to whole plots (main plots), and levels of **B** are randomly assigned to split plots (subplots) within each whole plot. An example is where **A** represents levels of herbicide used for large plots of land and **B** represents different crop varieties (e.g. including a GM one) planted in each large plot. Treatment **A** is tested against the between main plot variation (which is typically large), while **B** and the interaction **A*B** are tested against the between subplot variation (which is typically small).

```
proc mixed;
   class A B Block;
   model Y = A B A*B;
   random Block A*Block;
run;
```

The variables **A**, **B**, and **Block** are listed as classification variables in the CLASS statement. The model consist of indicator variables corresponding to the levels of the fixed effects **A**, **B**, and **A*B** while the dependent variable is **Y** (e.g. crop yield). The indicator variables corresponding to the levels of the random effects are **Block** and **A*Block**.

The "Class Level Information" table lists the levels of all variables. You can check this table to make sure that the data are correct.

Class Level Information

Class	Levels	Values
A	3	1 2 3
B	2	1 2
Block	4	1 2 3 4

Finally, the fixed effects are tested. As we can see, crop varieties (**B**) have significant effect (Pr > F is smaller than 0.05) on crop yield (**Y**), despite 3 different levels of herbicide (**A**) and random effects of blocks (**Block**).

Type III Tests of Fixed Effects

Effect	Num DF	F Value	Pr > F
A	2	4.07	0.0764
B	1	19.39	0.0017
A*B	2	4.02	0.0566

4.2. Generalized linear models (GLM) and generalized linear mixed models (GLMM)

Furthermore, the **GLM** is a flexible generalization of ordinary linear regression and analysis of (co)variance. Sometimes the data can be transformed (e.g. a logarithmic transformation) to stabilize the variance. **GLM**s generalize linear regression by allowing a linear model to be related to the mean of the underlying distribution via a non-linear link function (explained below) and by allowing the magnitude of the variance of each measurement to be a function of the mean. In addition to continuous data (data that are continuous in a selected range, e.g. pH or concentration of dissolved carbon), **GLM**s allow the modeling of discontinuous (count) data (e.g. numbers of beetles counted per area) and proportions as well as the cases when many zeros in a data set are present (which is often the case for NTO data and is a complicating factor for the statistical analysis). While there are other approaches (e.g. Chi-square) to analyze count data, none of them can be efficient and flexible enough as **GLM** or **GLMM**. Thus, Chi-square tests can handle only the most simple tests, e.g. comparison of two treatments without blocking and time continuity, which is rarely possible for properly designed field trials.

GLMs consist of three elements:

1) A probability distribution such as the normal, exponential, binomial, Poisson etc.

2) The linear predictor is the quantity which incorporates information about independent variables (such as temperature, concentration of herbicides) that may have an influence into the model. It is related to the expected value of the data through the link function.

3) The link function (mathematical function that links response variables to predictors) provides the relationship between the linear predictor and the mean of the distribution function. There are many commonly used link functions, and their choice can be somewhat arbitrary. The link function can linearize the expected response value as well as homogenize the (expected) variances.

Finally, the **GLMM** is a particular mixed model. It is an extension of the **GLM,** in which the linear predictor contains random effects (e.g. blocking) in addition to the usual fixed effects

(e.g. level of herbicide). These random effects are usually assumed to have a normal distribution. In the **GLMM**, it is numerically difficult to estimate parameters. Various so-called approximate estimation methods have been developed, but unfortunately none has good properties for all possible models and data sets (e.g. for ungrouped binary data). For this reason, numerical methods involving the Markov Chain Monte Carlo method (Berg, 2004) have increasingly been used as increasing computing power and advances in methods have made them more practical. But these have drawbacks too, since priors have to be specified and there is no common opinion on what prior should be chosen for variance parameters.

4.3. Overdispersion

Overdispersion is the condition by which the variability of data in a data set exceeds the variability expected under a particular probability distribution (Box 10). Thus, data which are normally distributed are never overdispersed. However, overdispersion can occur in GLM in which the mean and variance are functionally dependent. Counts may exhibit more variability than is possible under Binomial or Poisson probability models. McCullagh and Nelder (1989) suggested that overdispersion may be the normal situation in case of environmental studies (including effects of GM plants on non-target organisms) rather than an exception. This might be due to the fact that experimenters resort to a small number of probability distributions to model their data. In most of the cases, this leads to the Binomial or Poisson distributions for counts. It is also important to choose a proper distribution model that permits higher dispersion if necessary, such as a Beta-binomial instead of Binomial model and a Negative Binomial instead of Poisson model. Overdispersion can also appear due to an improper selection of independent variables (e.g. the concentration of herbicides vs environmental parameters) and effects to model the data. Such cases must be solved by altering the set of effects and independent variables and not by selecting a different probability distribution for the data. In many cases (Schabenberger & Pierce., 2002), overdispersion might be covered by addition of random effects and coefficients to the linear predictor of GLM. In general, low levels of overdispersion can be handled very well by inflating the variance function with a fixed factor (which itself is then called the dispersion parameter). This approach turns a GLM into a GLMM.

Box 10.1 GLMM example

The experiment is a split-plot design with different management methods as the whole-plot treatment factor and different crop varieties as the split-plot treatment factor. The whole-plots are arranged in randomized complete blocks. The measurement is the number of NTO of a given species per experimental unit. There are 7 management types (**TRT**), 4 blocks (**BLK**), 4 types of crop varieties (**CRP**) and the response is the number of NTO (**COUNT**). Because it is count data, we will assume a Poisson distribution.

```
class trt blk crp;
model count=trt crp trt*crp/dist=poisson link=log ddfm=satterth;
lsmeans trt crp/diff;
random blk blk*trt;
run;
```

The variables **TRT**, **BLK**, and **CRP** are listed as classification variables in the CLASS statement. The model dependent variable (**COUNT**) is the count information, therefore we use the Poisson distribution (**dist=poisson link=log**) with log link function (a continuous mathematical function that defines the response of variables to predictors). The independent variables on the right side of the equation include the fixed effects, which are **TRT**, **CRP**, and the **TRT* CRP** interaction effect. The random statement specifies the random effects (**BLK**, and whole plot error **BLK*TRT**) that are included in the model.

The "Class Level Information" table lists the levels of all variables. You can check this table to make sure that the data are correct.

Class Level Information

Class	Levels	Values
trt	7	1 2 3 4 5 6 7
blk	4	1 2 3 4
crp	4	1 2 3 4

Box 10.2 GLMM example

The following information is the goodness of fit. The only fit statistics that we focus on is the Pearson Chi-Square/DF. This gives us an estimate of the overdispersion parameter. In this case, the Pearson Chi-Square/DF suggests that there is an overdispersion, since it is too high (more than 1).

Fit Statistics

Pearson Chi-Square	**634.08**
Pearson Chi-Square / DF	**7.55**

Finally, the fixed effects are tested.

Type III Tests of Fixed Effects

Effect	Num DF	F Value	Pr > F
trt	6	3.64	0.0130
crp	3	14.82	0.1882
trt*crp	18	10.48	0.2391

If we run the same model without the ddfm=satterth (assuming that there is no overdispersion) , the Pr > F for **CRP** and **TRT*CRP** become <.0001.

There is a significant treatment effect (**TRT**). Even though both **CRP** and **TRT*CRP** have large F values, neither of them are statistically significant. This is because the probabilities have taken into account the overdispersion.

4.4 Equivalence testing

In most field trials to study the impact of GM crops on NTO's, a difference test is used (difference between GM and non-GM crop). From the statistical point of view, there are several major reasons why this common statistical procedure may require review.

The error of most concern in a difference test is of falsely inferring that no impacts (possibly indicating no hazards) exist where, in reality, there are. Because the traditional statistical null hypothesis is one of equality (no difference between GM and non-GM crop), this error is relatively difficult to estimate and/or set to a desired magnitude. This disadvantage is overcome by the equivalence test, sometimes referred to as a 'proof of safety', since here the null hypothesis is one of inequality, however the error of most concern may not be set easily. The advantage of equivalence testing is therefore that the responsibility is placed back onto those

who wish to demonstrate the safety of GMs, to do high quality, well-replicated experiments with sufficient statistical power (Perry et al., 2009).

Equivalence testing contrasts with other biological experimentation: in the past, the risk assessor tests a null hypothesis of inequality between the GMO and its control, which must be actively disproved if the experimenter is to conclude that the GMO is equivalent to the comparator concerned. By contrast, in the statistical test and in most biological experiments the null hypothesis is one of equality (no difference)(Perry et al., 2009). Equivalence testing is commonly used in biomedical and pharmaceutical statistics: pharmacokinetic parameters of alternative drug formulations have to be shown to be within a factor 1.25 of the value for the reference drug. The null hypothesis of the equivalence test is "there is a difference between the GMO and its reference of a certain minimum size" against the alternative hypothesis: "there is no or only a small difference between the GMO and its reference". Therefore, in this testing procedure, we need a significant result (rejection of the null hypothesis) in order to conclude that the GMO and the reference are equivalent.

The basic comparison of interest focuses on the difference between the GM and the appropriate (usually near-isogenic) comparator. However, equality is clearly neither reasonable nor possible when genetically modified insect-resistant (GMIR) systems (*e.g. Bt* maize crops) are compared with a near-isogenic variety managed without the insecticides that would be typically applied conventionally. Therefore, it is sensible to also consider extra comparators that help to place differences between the GM and its comparator into context (Perry et al., 2009), such as including a non-transgenic treatment managed with conventional insecticides.

For example, Marvier et al. (2007) reported a meta-analysis of 42 field experiments that indicated that non-target invertebrates were generally more abundant in *Bt* cotton and *Bt* maize fields than in non-transgenic fields managed conventionally with insecticides, but in comparison with insecticide-free control fields, certain NTO taxa were less abundant in the *Bt* crop fields (Perry et al., 2009).

The general idea is that a comparative risk assessment can establish equivalence between the GMO and its non-GM counterpart for characteristics other than the intended effects of the genetic modification. There are only a couple of examples when applicants have been using equivalence tests for field trials. Thus, Oberdoerfer et al. (2005) applied equivalence tests using fixed but arbitrary equivalence limits of ± 20% average values. In a later paper (Hothorn and

Oberdoerfer, 2008), the fixed value was described as rigid and not reflecting the difference in variability between components, and component-specific safety ranges were proposed to be proportional to the variance of the concurrent control in the same field trials. This method ignores the amount of background variation found between commercial varieties.

However, a successful test needs equivalence limits, which are still very difficult to select for NTO field studies (van de Voet et al., 2011). Equivalence limits could be estimated from field studies with concurrent reference varieties, which are typically the same studies in which also the GMO and its non-GM counterpart are tested. Therefore, van der Voet (2011) suggested to have a two-step procedure, at least in principle. The first step is the setting of equivalence limits, the second is their use for assessing equivalence.

The novelty of the combined approach is the simultaneous assessment of both differences and similarities. To detect unintended effects, it is optimal to study the differences between the GMO and its non-GM counterpart. However, to assess similarities and equivalence, a characterization of natural biological variability is needed. Thus, van der Voet & Perry (2011) proposed that the GMO can be viewed relative to the background variation shown by common varieties (e.g. commercial varieties) used as references. However, this approach was suggested in relation to the food/feed safety of GMO's (compositional analyses). This approach to include several common varieties is not feasible for most NTO studies and there does not seem to be a clear need to include them.

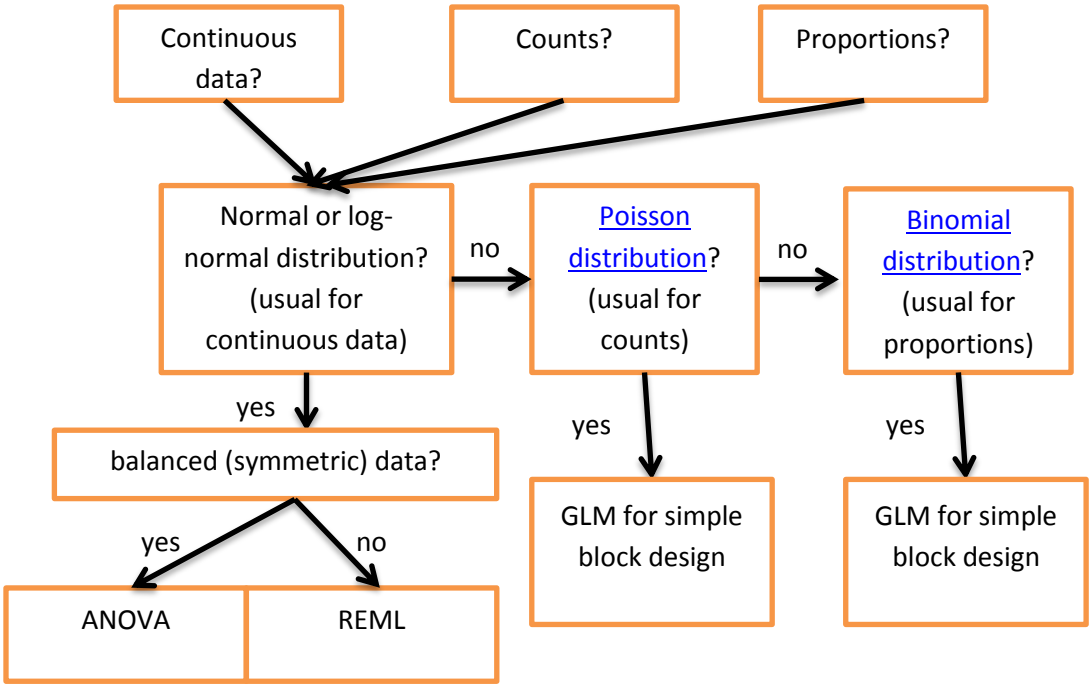
Van der Voet & Perry (2011) suggested that statistical methodology should not be focused exclusively on either differences or equivalences, but should provide a better understanding within which the conclusions of both types of assessment (Box 6; Section 4.4) are allowed (van der Voet & Perry, 2011). Both approaches are complementary: statistically significant differences may point at biological changes caused by the genetic modification, but may not be relevant from the viewpoint of ERA. Equivalence assessments may identify differences that could be larger than normal natural variation, but such cases may or may not be cases where there is an indication of true biological change caused by the genetic modification. On the other hand, Ward et al., 2012 suggested that comparisons with traditional equivalence testing are not very helpful, because with a traditional equivalence test, the focus is on whether the difference between two specific treatments is less than a pre-specified amount. A situation could arise in which two different submissions with very similar profiles (e.g., GM, comparator

and environment) could result in different conclusions because the respective sets of reference varieties led to different sets of equivalence limits (Ward et al., 2012).

5. Statistical approaches and real data

In this section, various field experiments are discussed and any correct and incorrect use of experimental design or statistical analysis is highlighted (using Checklist for field trials). This may assist both the investigators performing new experiments and risk assessors that have to assess the (biological) relevance of results of field trials with GM plants.

A first example among others is offered by a study by Hoss et al (2011). They assessed the possible influence of GM maize (expressing the insecticidal Cry3Bb1 protein), as compared to non-GM maize, on the abundance of free-living soil nematodes. While the experimental design (randomized complete block design) was reasonable for such a comparison (**1., 2., 3., 4. in Chapter 6 "Checklist for field trials"**), the statistical approach was very rudimentary for its purpose, as explained in the following: An ANOVA with maize cultivar as the fixed factor and block as a random factor (**6.3. in "Checklist for field trials"**) was carried out to test for differences in the measured parameters (nematode counts) between the two cultivars. Since the measurements encompassed discontinuous (count) data, the use of a GLM on the basis of a Poisson distribution would have been appropriate (Premise **6.2.2. in "Checklist for field trials" is violated**). Count/Poisson-distributed data have the property that the variance increases with the mean, which violates the ANOVA assumption of homogeneous variances. Thus, applying ANOVA to such data can lead to very inaccurate p values. Moreover, an analysis of the statistical power of the test was not performed (**5. in "Checklist for field trials"**). Based on this mishap, we propose the following initial structure of statistical considerations that guide us to the appropriate test (see decision tree below).



Another case of the many examples of an improper use of statistics can be found in Oliveira Filho et al. (2011). These authors studied the influence of pest control agents on non-target invertebrates in soil. For comparison of the treatments, they transformed the count data of invertebrates to percentages of a maximum (to distinguish relative changes), after which differences in the percentages between the treatments were evaluated by one-way ANOVA. The authors did not check in their data whether the variance increased with the mean. As in the above, a GLM (for proportions) instead of ANOVA (for continuous data) should have been used (Premise **6.2.3. in "Checklist for field trials" is violated**).

A similar problem can be found in Post et al. (2011). These authors studied the effects of transgenic blight-resistant American chestnut, as compared to a non-GM variant, on insect herbivores in a completely randomized block design (**4. in "Checklist for field trials"**). Although it was appropriate to use one-way ANOVA for comparisons of the growth rates (continuous data) of insect herbivores (**6.2.1. in "Checklist for field trials"**), the use of one-way ANOVA to compare the counts of the insect herbivores was not (**6.2.2. in "Checklist for field trials" is violated**). Again, the use of a GLM on the basis of the Poisson distribution would have been appropriate. As we can see from these few examples, usage of the simplified version of the above decision tree or Checklist for field trials assists us using the selection of the proper statistical analyses.

A farm-scale study (Spain) was initiated in 2000 to assess the potential impacts of Bt maize on the abundance and diversity of predatory arthropods (de la Poza et al., 2005). The experimental setup was a randomized block design (**4. in "Checklist for field trials"**) involving three treatments, each with four (Lleida) or three (Madrid) replicates. The treatments were: (1) Bt transgenic maize, (2) the isogenic hybrid without herbicide treatment and (3) the isogenic hybrid with imidacloprid insecticide seed treatment. In the combined statistical analyses of variance, a split–split–plot model was used, in which year and location were considered as the main plots. Subplots were the treatment (3 treatments) and sub-subplots were the sampling dates. All factors, except blocks, were considered fixed and crossed with each other, except, again, for blocks that were nested within locations and years (**6.3. in "Checklist for field trials"**). A priori comparisons of the means among treatments within a given environment (year per location combinations) were performed with the adjusted least square means, using standard t-tests. To normalize the original data, these were transformed by square root transformation (SQRT) prior to analysis (**6. in "Checklist for field trials"**). This study can be characterized as correct.

In the case of Druart et al. (2011), who studied the influence of pesticide drift and transfer on non-target snails in soil, the experimental design (**1., 2., 3., 4., in "Checklist for field trials"**) and statistical analysis were adequate. Differences in snail mass or shell diameter were assessed by a linear mixed-effects model with zone as the fixed explanatory variable and with microcosm as the random variable (**6.3. in "Checklist for field trials"**). The mortality for each treatment between each zone and mortality between treatments in all zones pooled were assessed by a binomial GLM (see the tree above) (**6.2.3. in "Checklist for field trials"**), resulting in an appropriate statistical analysis.

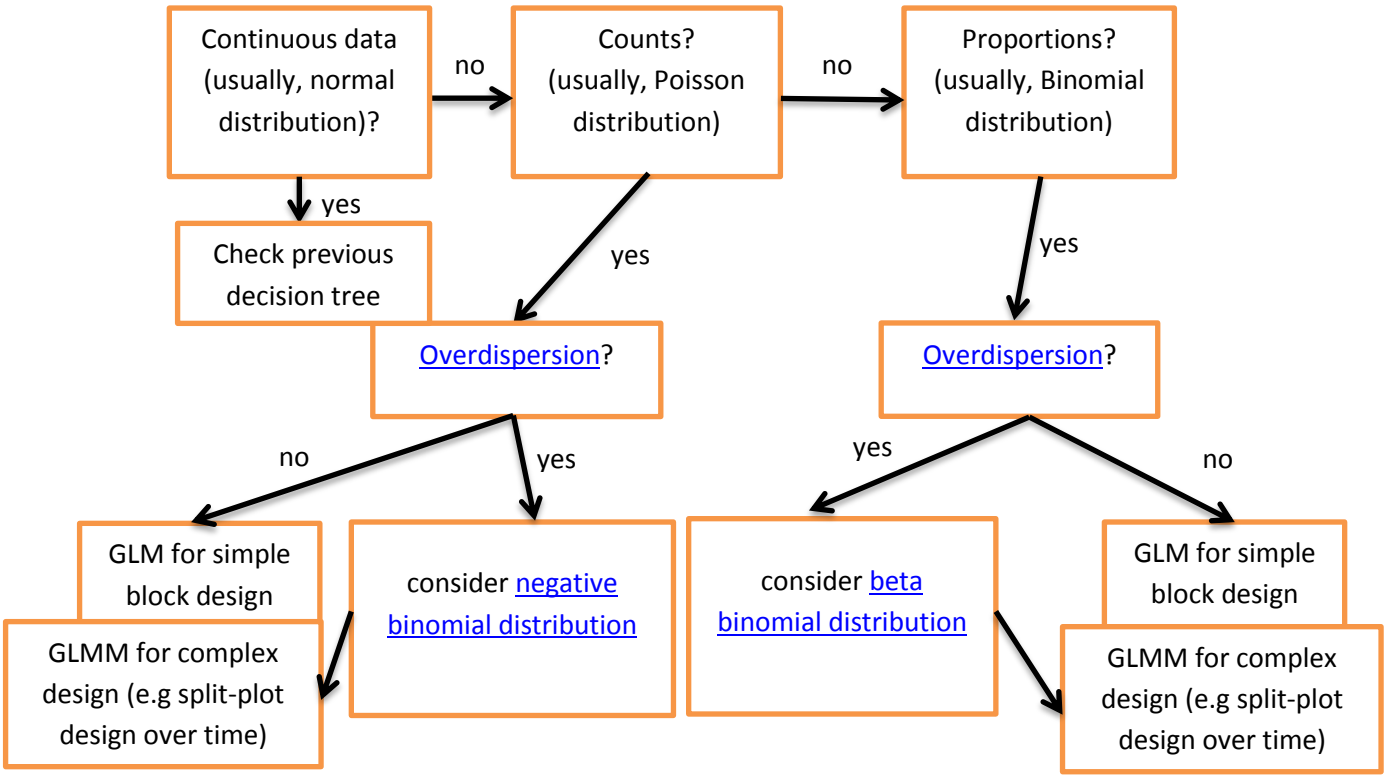
Two studies of Rauschen et al. (2008 and 2010) provide good examples of how statistical analyses might be improved for similar types of studies. In both studies, a completely randomized block design with eight replicates each was used (**4. in "Checklist for field trials"**). For the first study "impact of Bt-corn MON88017 in comparison to three conventional lines on herbivorous plant bugs (Heteroptera: Miridae)", they used a GLM using different soil parameters (no significant effect was found) and maize lines as explanatory variables while no extra statistical information was provided. In the second one, Diabrotica-resistant Bt maize DKc5143 event MON88017 had no impact on the field densities of the leafhopper *Zyginidia scutellaris* (as compared to non-transgenic maize) (Rauschen et al., 2010). The statistical analysis of the second

study was more detailed and appropriate. The densities of insects per plot (n = 8 per sampling date) were analyzed separately for each sampling date using GLM (although this is not a proper way, since it is possible to inflate the Type I error) and assuming a negative binomial distribution (**6. in "Checklist for field trials"**). A parameter modeling the Poisson variability of the counts was estimated from the data (**6.2.2. in "Checklist for field trials"**). First, the abundance of a species in a given plot was expressed in a log-linear dependency of the block effects. Secondly, different soil parameters were measured at the field site as well. Thirdly, GLM with negative binomial assumptions were fitted including the block and variety effects only. Based on these model fits, all pairwise comparisons among the maize varieties were performed.

There are also examples in which no field replicates were organized or a tested field was unequally separated between GM and non-GM plants (Yamamori., 2011), or no control was used (Farinos et al., 2003) (**1., 2., 3., 4. in "Checklist for field trials" are violated**).

There are certainly studies with a proper statistical analysis (in our opinion) even if a relatively complex experimental design was used. Thus, Stoleson et al. (2011) studied the responses of bird communities to an operational herbicide treatment over time. They used a randomized block design (**4. in "Checklist for field trials"**), in which half of each 8-ha block received herbicide and the other half acted as control. As for the statistical analyses, they used generalized linear mixed models to model the effects of year, site, herbicide treatment and cutting sequence on vegetation and avian target variables (**6. in "Checklist for field trials"**). In all models, they considered site as a random effect, and year, herbicide treatment, and cutting sequence as fixed effects (**6.3. in "Checklist for field trials"**). Shannon indices were modeled using a Beta distribution, while other diversity indices were modeled with a normal distribution, vegetation covers were modeled using a lognormal distribution, whereas bird abundances were modeled using a Poisson distribution. All models used the restricted maximum-likelihood method. While standard maximum likelihood (used for classical ANOVA to fit parameters) estimates the standard deviations of the random effects assuming that the fixed-effect estimates are precisely correct, restricted maximum likelihood averages over some of the uncertainty in the fixed-effect parameters to adjust the denominator degrees of freedom.

Based on the discussed weaknesses and strengths of the methods used in the above described examples, we improve the most common part (count and proportion data for NTO) of the decision tree (see below):



From all previous examples we can conclude that, when applying any significance tests for abundance data, we often face the problem of uncertainty concerning the true effect (or the width of confidence intervals, in case of equivalence tests), becoming large for low abundances (power analysis can provide insight into the size of the problem), small numbers of replications and large residual variation. Counting of individuals scattered randomly in the observational windows might yield data following the Poisson distribution. However, large residual variation in abundance data often occurs due to the clustering of individuals, termed extra-Poisson variation or overdispersion (McCullagh and Nelder 1989; Rauschen et al., 2009). Therefore, the probability to find a significant difference in the case that the GM treatment has no effect on the abundance of a species (Type I error) may be high if rare species or species with high variability in local or temporal abundance or activity are investigated with a low number of replications. On the other hand, several types of NTO (e.g. collembola) might be characterized as types for which

relatively low number of replications (due to their high density) are sufficient. Thus, planning trials with a sufficient number of replications, based on available prior information concerning the mean abundance and variability of the observations is an important issue. For this purpose, we suggest to analyze available datasets of a certain NTO (might be obtained from other field trials) concerning its mean abundance and variability. In complex cases, it is important to simulate abundance data for different choices of mean abundance, variability and experimental design.

---

6. Checklist for field trials

The use of proper statistics in field studies is complex, as we have seen in the foregoing. Hence, we decided to develop a checklist that provides guidance in the use of statistical principles related to field testing of GM crops. Most of the mistakes discussed in Section 5 can be avoided if the rules below are considered.

Checklist:

1. Statement of a hypothesis: in any field test, a hypothesis has to be formulated. Since a hypothesis is a statement of the presumed relationship between variables, it must be properly stated. The hypothesis suggests a particular relationship between variables and it therefore narrows the problem to one that is specific and researchable. This makes the specification of independent and dependent variables relatively easy.

2. Definition of variables: In order to observe whether the hypothesized relationship between variables exist, the latter must be clearly defined. Definition of the variables in a trial experiment allows everyone (both the experimenter and the regulator) to know what is being studied and facilitates interpretation of the results.

3. Specification of sample: The experimenter must clearly define which biological parameters (e.g. NTO) are studied and how, e.g.:

Were all possible or a specific set of NTOs studied?
Were samples randomly selected?
Was the sample only one organism or many?
Were organisms made up in groups?

These clarifications will help to determine the generalizations that were made, the data collection procedures that were selected and the statistical analysis that was employed.

4. Experimental design: The experimental design chosen should allow the experimenter to test the hypothesis. In the design, the experimenter should have provided answers to the following question and considerations:

Were the treatments blocked? If not, then use completely randomized design. If yes and only one variable was studied, then use a randomized block design. If there was more than one variable, use a factorial randomized block design.

When the experimental design is selected, the following questions have to be answered positively:

4.1. Were the treatments (blocks) properly randomized? (Box 1).

4.2. Were the treatments (blocks) properly replicated? (Box 2).

4.3. Was the field size appropriate for a certain NTO? (Box 3). Justification should be provided.

4.4. Was the sample size appropriate for a certain NTO? (Box 4). Sample size calculation (or justification) should be provided.

4.5. Was true replication performed and pseudoreplication avoided? (Box 5). How were subsamples pooled?

5. Statistical power: statistical power is the probability that the test applied will reject the null hypothesis when the null hypothesis is indeed false. It also provides the confidence that replication is neither too small to detect effects that are present, nor too great to avoid that unnecessary extra resources are used for trial experiments (Box 6). Values of 70% (Prasifka et al., 2008) and 80% (Perry et al., 2003) are commonly used in field trials as the desired statistical power.

6. Statistical analysis: After the data have been collected, the experimenter must assess the relationships between independent and dependent variables. Most of these assessments are based on statistical analyses.

6.1. Type of null hypothesis: This hypothesis, denoted H_0, should be capable of being proven false using a test of observed data. The null hypothesis typically corresponds to a general or default position. A set of data can only reject a null hypothesis or fail to reject it. Test of difference (H_0: $\mu_1 = \mu_2$) or equivalence test (H_0: $\mu_1 - \mu_2 > \sigma$ or H_0: $\mu_1 - \mu_2 < -\sigma$).

6.2. What types of data were analyzed?

6.2.1. If data are **continuous** (e.g. pH) then consider the normal or log-normal distribution (i.e. use a log transformation) and subsequently use ANOVA (Box 7) for balanced or REML for unbalanced (asymmetric) data. Check the residual plot.

6.2.2. If data are **counts** (e.g. numbers of larvae), then GLM with Poisson distribution and log-link function are used. Either use a GLM for simple block design or a GLMM for designs such as split-plot design. In case of overdispersion, use a quasi-likelihood approach (i.e. variance proportional to the mean). An alternative way to model overdispersion is by using the negative binomial distribution. In case of simple experimental design (e.g. absence of random factors), Chi-square test is possible.

6.2.3. If data are **proportions** (e.g. the mortality of larvae), use GLM with binomial distribution and proper link function (e.g. logit). Either use a GLM for simple block design or a GLMM for designs such as split-plot design. In case of overdispersion of the data, use a quasi-likelihood approach. An alternative way to model overdispersion is by using the beta-binomial distribution. Overdispersion should not be used for 0/1 data as overdispersion is then not possible.

6.3. Fixed and Random effects: these are the types of dependent variables in statistical analyses (Box 8). Check how the fixed and random effects were selected.

6.4. Overdispersion: is the condition by which the variability of the data exceeds that expected under a certain probability distribution (data which are normally distributed are never overdispersed).

- Check for overdispersion (the occurrence of more variance in the data than predicted by a statistical model) , especially for data that obey the Poisson (counts) and binomial distributions (proportions) (Box 10).
- In some cases, distribution models might have to be changed to Beta-binomial instead of Binomial and Negative Binomial instead of Poisson to deal with overdispersion. Overdispersion might also be addressed by the addition of random effects and coefficients to the linear predictor of GLM. This approach turns GLM to GLMM.

7. Discussion

This report summarizes the most important statistical considerations with respect to the field testing of GM insect-resistant crop plants in relation to their effects on NTO's. For applicants and risk assessors alike, it is important to carefully consider the following items:

1. Objective of the study and required experimental set-up (see 3),
2. Field size and its implication for NTO impact testing,
3. Experimental set-up, including design, randomization and replication,
4. Statistical power testing,
5. Normality of the data distribution,
6. Overdispersion of the data and implications for statistics,
7. Difference versus equivalence testing.

Thus, it is essential for the experimenters to plan field experiments in a proper way. Without a clear prior understanding of the experimental hypothesis and how the results will be analyzed and interpreted, an incorrect statistical analysis may be applied, which will lead to improper or wrong conclusions. Next to considering the importance of proper randomization and replication, avoiding pseudoreplication, an experimenter has to pay extra care to the specific rules that reign GM plant field testing, such as the size of the plots and consequently the size of the field required, to test for impacts on NTO's. A common facet of current field trials is the fact that fields are too small to adequately assess the impact of NTOs in a sound manner. Moreover, the statistical power of the design has to be checked routinely in order to be sure that the experimental trials are appropriate (e.g. the number of replicates and the sample number per replicate). Considering existing field experiments, there are many examples (Section 5) in which improper statistical analyses were applied. The most common mistake is that discontinuous data (counts) are analyzed by analysis of variance (ANOVA). This in spite of the fact that only generalized linear models (GLM) are found to be efficient enough for analysis of these kinds of

data. There is a Chi-square test that has been in use to analyze count data, but it is not efficient and flexible enough as compared to **GLM** or **GLMM**. Chi-square tests can handle only the most simple tests, e.g. comparison of two treatments without blocking and time, which is rarely possible for properly designed GM plant field trials. Recently, equivalence rather than difference testing was proposed as an appropriate approach to deal with NTO impact data from GM plant field trials (Perry et al., 2003). There is no *a priori* scientific justification for either of the two approaches, and hence it can be argued that usually difference testing is as appropriate as equivalence testing, if both the experimental design and statistical analyses are justified. Moreover, it is difficult to analyze discontinuous data by equivalence testing for non-professional statisticians.

The recent statistical guidance by EFSA (2010) does not provide clear and structured suggestions for field trials. Hence, we decided to launch a combination of decision trees (in particular the modified version of the tree proposed by Bolker et al., 2009) next to the here proposed "Checklist for field trials". This as a guide into the difficult field of statistics of GM plant field trials. These tools can assist risk assessors to interpret whether the correct statistical analyses are used in field trials to study effects of GM plants on NTO's. In particular, the "Checklist for field trials" will highlight possible improper designs or errors in the statistical analyses of already performed experiments. Especially this checklist is relevant for risk assessors. Several important cases are, however, not included, such as the presence of repeated measurements and analysis of multiple experiments at once (for most of them, very specific knowledge of statistics is essential). The original decision tree by Bolker (2009) allows assessing these (common) as well as several other specific cases (Appendix) such as low densities of NTO's with several random effects involved. However, it requires additional knowledge in statistics which is not included in the report. The use of the proposed decision trees and the "Checklist for field trials" offers a reasonable approach and solution to avoid improper statistical analyses. In particular, we would like to stress that they highlight ways to overcome or avoid the many common severe problems in the final interpretative results of treatment comparisons in field trials. Only professional statisticians are able to provide the certainty that all steps of a chosen statistical approach are taken in a proper way, since minor details might lead to incorrect final conclusions. However, such professionals are still in debate about the most proper statistical

approaches for assessments of GM and reference plant varieties, in particular when it comes to NTO impact assessments (Ward et al., 2012).

8. References

Albajes A. et al., 2012. Field trials to assess risks of transgenic crops for non-target arthropods: Power analysis and surrogate arthropods in Spain. GMOs in Integrated Plant Production. v73., p1-7.

Auclerc, A. 2009. Experimental assessment of habitat preference and dispersal ability of soil springtails. Soil Biology and Biochemistry. v41., p1596-1604.

Berg B. A. 2004. Markov Chain Monte Carlo Simulations and Their Statistical Analysis. Singapore, World Scientific.

Boer, P.J. den, 1970. On the significance of dispersal power for populations of carabid-beetles (Coleoptera, carabidae). Oecologia, 4: 1-28.

Bolker B.M. et al., 2009. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in Ecology and Evolution. v24., p127-135.

Bonte D. et al., 2008. Thermal conditions during juvenile development affect adult dispersal in a spider. PNAS. v105., p 17000-17005.

Cameron P. et al., 2009. Dispersal of potato tuber moth estimated using field application of Bt for mark-capture techniques. Entomologia Experimentalis et Applicata. v132., p99-109.

Clark S.J. et al., 2006. Farm Scale Evaluations of spring-sown genetically modified herbicide-tolerant crops: a statistical assessment. Proc. R. Soc. B. v273., p237-243.

Duan J.J. et al., 2006. Statistical power analysis of 2-year field study and design of experiments to evaluate non-target effects of genetically modified *Bacillius thuringiensis* corn. Ecological Entomology. v31., p521-531.

Druart C. et al., 2011. Snails as indicators of pesticide drift, deposit, transfer and effects in the vineyard. Science of the Total Environment. v409., p4280-4288.

Farinos G.P. et al., 2008. Diversity and seasonal phenology of aboveground arthropods in conventional and transgenic maize crops in Central Spain. Biological Control. v44., p362-371.

Feder, J.L., S.B. Opp, B. Wlazlo, K. Reynolds, W. Go & S. Spisak, 1994. Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. Proc. Natl. Acad. Sci. USA, 91: 7990-7994.

Gil L. et al., 2004. Dispersion of flightless adults of the Asian lady beetle, Harmonia axyridis, in greenhouses containing cucumbers infested with the aphid Aphis gossypii: Effect of the presence of conspecific larvae. Entomologia Experimentalis et Applicata. V112., p1-6.

Hazell S. et al., 2005. Competition and dispersal in the pea aphid: Clonal variation and correlations across traits. Ecological Entomology. v30., p 293-298.

Hoss S. et al., 2011. Assessing the risk posed to free-living soil nematodes by a genetically modified maize expressing the insecticidal Cry3Bb1 protein. Science of the Total Environment. v409., p2674-2684.

Hurlbert S.H. Pseudoreplication and the design of ecological field experiments. Ecological Monographs. v54., 187-211.

Kupper L.L. et al., 1989. How appropriate are sample size formulas? Am. Stat. v43.,101-105.

la Poza M., et al., 2005. Impact of farm-scale Bt maize on abundance of predatory arthropods in Spain. Crop protection. v24., p677-684.

LeBlanc. 2004. Statistics: Concepts and Applications for Science. Jones and Bartlett Publishers, Canada. McCullagh P. & Nelder F. 1989. Generalized linear models. 2nd edition. Chapman and Hall, New York.

Lehmitz, R. et al., 2012. Active dispersal of oribatid mites into young soils. Applied Soil Ecology. v55., p10-19.

Liebherr, J.K., 1988. Gene flow in ground beetles (Coleoptera: Carabidae) of differing habitat preference and flight-wing development. Evolution, 42: 129-137.

Løjtnant et al., 2012. Modelling Gene Flow between Fields of White Clover with Honeybees as Pollen Vectors. Environmental Modeling and Assessment. v17., p 421-430.

Morsello S. et al., 2008. Temperature and precipitation affect seasonal patterns of dispersing tobacco thrips, Frankliniella fusca, and onion thrips, Thrips tabaci (Thysanoptera: Thripidae) caught on sticky traps. Environmental Entomology. v37., p79-86.

Oliveira-Filho E.C. et al., 2011. Susceptibility of non-target invertebrates to Brazilian microbial pest control agents. Ecotoxicology. v20., p1354-1360.

Perry J.N. et al., 2003. Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. Journal of Applied Ecology. v40., p 17-31.

Perry J.N. et al., 2009. Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environ. Biosafety Res. v8., p 65-78.

Post K.H. and Parry D. 2011. Non-target effect of transgenic Blight-resistant American Chestnut (Fagales: Fagaceae) on insect Herbivores. Transgenic Plants & Insects. v40., p955-963.

Rauschen S. et al., 2010. *Diabrotica*-resistant Bt-maize DKc5143 event MON88017 has no impact on the field densities of the leafhopper *Zyginidia scutellaris*. Environ. Biosafety Res. v9., p87-99.

Rauschen S. et al., 2008. Impact of Bt-corn MON88017 in comparison to three conventional lines on *Trigonotylus caelestialium* (Kirkaldy) field densities. Transgenic Res.

Smith King, P., 1987. Macro- and microgeographic structure of a spatially subdivided beetle species in nature. Evolution, 41: 401-416.

Slatkin, M., 1985. Gene flow in natural populations. Annu. Rev. Ecol. Syst., 16: 393-430.

Storkey J. et al., 2008. Providing the evidence base for environmental risk assessments of novel farm management practices. Environmental Science & Policy. v11., p 579-587.

Schabenberger O. & Pierce F. 2002. Contemporary statistical models for the plant and soil science. CRC press.

Schilthuizen, M. B.J. Scott, A.S. Cabanban & P.G. Craze, 2005. Population structure and coil dimorphism in a tropical land snail. *Heredity*, 95: 216-220.

Schilthuizen, M. & M. Lombaerts, 1994. Population structure and levels of gene flow in the Mediterranean land snail *Albinaria corrugata* (Pulmonata: Clausiliidae). *Evolution*, 48: 577-586.

Stoleson S.H. et al., 2011. Ten-year response of bird communities to an operational herbicide-shelterwood treatment in a northern hardwood forest. Forest Ecology and Management. v262., p1205-1214.

Yamamori M. 2011. 2011. Outcrossability of *Brassica napus* L. and *B. rapa* L. in an experimental field. JARQ. v45., p173-179.

Van der Voet H. et al., 2011. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. BMC Biotechnology. v11., p1-20.

Van Elsas et al., 2002. Methods for Sampling soil microbes. In: Manual of Environmental microbiology. Second edition. P 1138.

Ward et al., 2012. Comments on the paper "A statistical assessment of differences and equivalences between genetically modified and reference plant varieties" by van de Voet et al. 2011. BMC Biotechnology. v12., 13.

***Appendix**. Decision tree for the selection of the proper statistical approach (Bolker et al., 2009). This in terms of examining the initial data, the estimates and their confidence intervals, testing of hypotheses, selection of the best models and evaluation of the differences in goodness-of-fit among models.*