Environmental Risk Assessment of Genetically Modified Organisms:

Setup of a simulation study to investigate properties of difference and equivalence tests

Paul W. Goedhart, Hilko van der Voet

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

Note:
Detailed results will be published separately as Deliverable D9.2b on the project website http://www.amigaproject.eu/documents/deliverables/.

Contents

1 Introduction

1.1 Statistical analysis and design for environmental risk assessment

A basic statistical approach to environmental risk assessment (ERA) has been outlined in the EFSA Guidance Document (EFSA, 2010b) and in Perry et al. (2009). However, this approach is not specified in great detail. In the context of developing statistical concepts, methods, software and protocols for environmental risk assessment (ERA) and post-market environmental monitoring (PMEM), our main objectives were:

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

This report describes results of a simulation study to investigate properties of various statistical models, which are used to perform difference and equivalence testing, for analysing count data.

2 Setup of simulation study

2.1 Basic setup and simulation distributions

The most simple trial in which a *GM* plant is compared to its conventional counterpart is a completely randomized field trial with level of replication N. In that simple case there are only two parameters: the mean count of the non-target organism for the *GM* plant (μ_G) and the mean count (μ_C) for the comparator. In practice there might be repeated counts on the same plots, but this is ignored in this simulation study. Goedhart et al (2013, 2014) describe five statistical distributions commonly used to simulate counts: the Poisson distribution, the overdispersed Poisson distribution, the negative binomial distribution, the Poisson-Lognormal distribution and a distribution which follows Taylor's power law. The Poisson distribution was not used in this simulation study because it is generally believed (Perry et al 2003, Duan et al, 2006) that counts of non-target organisms (NTOs) typically have larger variance than according to the Poisson distribution. Table 1 summarizes the four distributions which are used to simulate data, with the dispersion parameter σ^2 as a function of the mean μ and the variation coefficient CV in the last column. There is no statistical distribution associated with Taylor's power law, as it only specifies a relationship between the variance and the mean. Perry et al (2003) used the negative binomial distribution to simulate according to Taylor's power law employing a negative binomial dispersion parameter which follows from equating the variance of the negative binomial to the power law. The same approach is followed here. Using the negative binomial is however somewhat arbitrary, as e.g. the Poisson-Lognormal has the same variance to mean relationship, but has a different distribution.

Table 1: Distributions and values for the dispersion parameter used to simulate data.

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

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

The assumed variability in field testing of NTOs is mostly defined in terms of the coefficient of variation (CV), for example Duan et al (2006), and this convention is also used here. The mean μ_C of the comparator and the coefficient of variation CV define the dispersion parameter σ^2, see Table 1. This same dispersion parameter is then used to generate counts for the comparator and also for the *GM* plant. So for example with μ_C=10 and CV=100%, the negative binomial dispersion parameter equals σ^2=0.9. In case the *GM* plant, in the same simulation, has a mean μ_G=2.5, the corresponding CV value equals $\sqrt{2.5+0.9\times 2.5^2}/2.5=$ 114%. Moreover, a mean μ_G=1 has a corresponding CV=138% in this setting. This somewhat higher CV value than for the comparator reflects the general believe that smaller means are associated with larger CV values. The quotient of the CV value for the *GM* plant and the comparator for each distribution is given below as a function of $Q=\mu_G/\mu_C$.

Overdispersed Poisson simulation distribution

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

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

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

Negative binomial and Poisson-Lognormal simulation distributions

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

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

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

Power law simulation distribution

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

The quotient of the coefficients of variation is given by

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

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

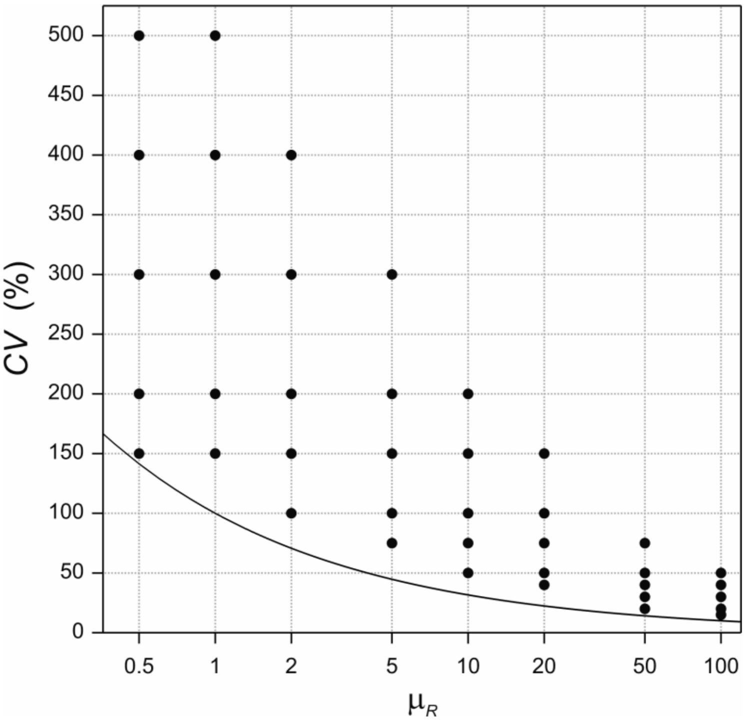
2.2 Parameter values used in the simulation

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

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

The assumed variability in field testing of NTOs is mostly defined in terms of the coefficient of variation (CV). Duan et al (2006) present graphs with CV values ranging from 25% to 200% with generally low CV values for means larger than 5 and CV values up to 200% for means close to zero. In this study, five different values of CV are used for different values of μ_C as given in Figure 1 and Table 2. Compared to Duan et al (2006) the larger CV values used in this simulation study seem to be at the upper end of what can be expected in practical field trials.

Figure 1: Combinations of comparator means μ_C and coefficients of variation CV. The solid line denotes the coefficient of variation of a Poisson distribution.



Finally the level of replication N must be specified. Environmental risk assessment of *GM* plants is typically performed in experiments with a small number of plots. This is (partly) due to the fact that relatively large plots and large guard rows are required in order to measure effects on NTOs without bias, see Perry et al (2003). It is therefore that such experiments are frequently repeated in different years and different locations such that larger levels of replication are obtained. A range of 4 to 100 for the level of replication N is employed in this study with some emphasis on lower values.

Table 2 summarizes the parameter values which are used in the simulation study. These values result in 1600 parameter combinations. For each combination of the simulation distribution (*OP*, *NB*, *PL* and *P1*) and parameter values 1000 datasets were simulated. Each dataset was analysed using the models given in the next session and an appropriate difference test at the 5% level was performed (details are given below). The proportion of datasets for which the difference test is rejected then gives an estimate of the true significance level (α) of the test when there is no difference, i.e. Q=1, and the power (β) of the test when there is a difference, i.e. Q≠1. These are only estimates of the true size of the test. Suppose that the size of the test is indeed exactly 5%, then with 1000 simulations a 99% prediction interval for the number of times the null hypothesis will be rejected is given by (33, 67) resulting in an interval of 3.3% – 6.7% for the true size. So only when the simulated significance level is outside this interval there is an indication that the true level of the test does not equal 5%.

Table 2: Parameters used in the simulation study.

Parameter	Values used in simulation
Mean μ_C of comparator	0.5, 1, 2, 5, 10, 20, 50, 100
Ratio $Q = \mu_G/\mu_C$	1, 0.75, 0.5, 0.25
Number of replication N	4, 6, 8, 10, 15, 20, 30, 40, 60, 100

μ_C	Coefficient of variation CV for comparator				
	CV-1	CV-2	CV-3	CV-4	CV-5
0.5	150	200	300	400	500
1	150	200	300	400	500
2	100	150	200	300	400
5	75	100	150	200	300
10	50	75	100	150	200
20	40	50	75	100	150
50	20	30	40	50	75
100	15	20	30	40	50

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

2.3 Statistical models for analysis

Fitting the Poisson-Lognormal model by means of maximum likelihood requires (adaptive) Gauss-Hermite integration within an iterative weighted least squares algorithm. This algorithm turned out to fail too frequently for data with small means, small levels of replication and/or small coefficients of variation. Therefor the Poisson-Lognormal model was not used to analyse simulated data. The other models with which each dataset was analysed are summarized in Table 3. All models were fitted using standard facilities in the statistical package GenStat (VSN international, 2013). Details for each analysis model are given below. A difference test for all models can be obtained by comparison of the fit of the model, more specifically the deviance, under the null-hypothesis H_0: $Q = 1$ and the fit of the model under the alternative hypothesis H_1: $Q \neq 1$.

Table 3: Statistical models used to analyse the simulated data.

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

LN: Log transformation followed by a t-test

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

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

1. A random draw Chi is generated by means of Chi = (2N-2) S^2/χ_{2N-2} where χ_{2N-2} is a random draw from a Chi-squared distribution with 2N-2 degrees of freedom;
2. N_C is a random draw from a normal distribution with mean X_C and variance Chi/N;
3. N_G is a random draw from a normal distribution with mean X_G and variance Chi/N;
4. Back-transform N_C by means of $N_C = \exp(N_C + Chi/2)$ and similarly N_G. Note that the back-transformation uses the equation for the mean of the lognormal distribution;
5. Subtract 1 from N_C and N_G; this accounts for the addition of 1 before log-transforming the count. This might sometimes result in a negative value for N_C or N_G. Such values are replaced by a small positive value, i.e. by 0.0001.
6. Calculate the ratio N_G/N_C

7. Repeat steps 1-6 many times, e.g. 10.000 or when more precise results need to be obtained 100.000 times. Calculate appropriate percentiles of the large sample which is the generalized confidence interval.

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

SQ: Squared root transformation followed by a t-test

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

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

OP: Overdispersed Poisson by a GLM-like analysis

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

In this model the underlying mean is log-transformed, rather than taking logs of the observed counts. This implies that the logarithm of the ratio of the two means, i.e. $\log(Q)$, is directly estimated in this model. A so-called Wald test statistic (Buse, 1982) can then be used for difference testing. This equals the quotient of the estimate of $\log(Q)$ and its standard error, and this is usually compared to a t-distribution to compensate for the estimation of the dispersion parameter. However it is generally believed that the likelihood ratio statistic has better statistical properties (McCullagh and Nelder, 1989). Moreover the Wald statistics breaks down when either sample only contains zero's since the estimate of $\log(Q)$, and its standard error, then becomes plus or minus infinity. So difference testing is based on the scaled likelihood ratio test. Equivalence testing under this model is however based on the estimate of $\log(Q)$ and its standard error, scaled by Pearson's statistic, which can be used to generate a confidence interval and thus to perform equivalence testing for arbitrary limits of concern. An alternative would have been to calculate a so-called profile likelihood interval but this requires a search algorithm which was considered to be too computer intensive in this simulation study.

NB: Negative binomial model by a GLM-like analysis

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

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

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

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

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

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

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

$$= 2\left[\frac{y^{2-\beta}}{(1-\beta)(2-\beta)} - \frac{y\,\mu^{1-\beta}}{1-\beta} + \frac{\mu^{2-\beta}}{2-\beta}\right] = 2(z1 - z2 + z3)$$

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

```
CALCULATE b1,b2 = 1,2 - power
EXPRESSIO dcalc[1] ; VALUE=!e(vfunction = mu**power)
EXPRESSIO dcalc[2] ; VALUE=!e(z1 = response**b2/(b1*b2))
EXPRESSIO dcalc[3] ; VALUE=!e(z2 = response*mu**b1/b1)
EXPRESSIO dcalc[4] ; VALUE=!e(z3 = mu**b2/b2)
EXPRESSIO dcalc[5] ; VALUE=!e(deviance = 2*(z1-z2+z3))
MODEL     [DISTRIBUTION=calculated ; DCALCULATION=dcalc[] ; \
          LINK=log ; DMETHOD=pearson ; DISPERSION=*] response ; \
          FITTED=fitted ; VFUNCTION=vfunction ; DEVIANCE=deviance
```

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

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

GM: Gamma model using a GLM-like analysis

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

Special cases

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

A. Sample 1 equals $\{0, 0, 0, 0\}$ and sample 2 equals $\{0, 0, 0, 0\}$. In this case there is no information and the deviance under the null model and under the alternative model are both zero for all models. The p-value for the difference test is set to 1 for all analysis models as there is no indication of a difference between the two samples. For the most extreme parameter combination μ_R=0.5, CV=500, Q=0.25, N=4 and the overdispersed Poisson distribution this situation occurs for 570 of the 1000 simulated datasets. For negative binomial, Poisson-LogNormal and Power models these numbers are respectively 511, 287 and 565. Clearly there is also no information for calculating a confidence interval and thus formal equivalence testing cannot be performed. Graphical results for equivalence testing present the proportion of these cases separately. Note that this case can be considered as "equivalent more likely than not".
B. Sample 1 equals $\{0, 0, 0, 0\}$ and sample 2 equals $\{c, c, c, c\}$ where c is some positive value. The deviance under the alternative model equals zero and so no test statistic can be calculated. However this situation is very rare. For the Poisson-LogNormal distribution there are 28 parameter combinations for which this situation occurs with a maximum of 5 out of 1000 such datasets at most. For the other distributions this situation occurs even less. These situations are therefore discarded, i.e. the corresponding p-value is set to missing.
C. Sample 1 equals $\{0, 0, 0, 0\}$ and sample 2 has different values with a positive variance. In this case all the p-values can be calculated in the usual way.
D. The mean of both samples are positive with a zero variance, e.g. $\{1, 1, 1, 1\}$ and $\{3, 3, 3, 3\}$. This is essentially the same as case B although it will occur even rarely. There are only 2 simulated datasets for which this occurs and these are discarded.

2.4 General remarks on difference testing

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

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

The power of a difference test based on the normal distribution can be calculated exactly. For non-normal distributions, small sample properties of difference tests are not straightforward. A simulation approach for sample size calculations for a difference test is employed by many authors, e.g. Shieh (2001) and Hrdličková (2006) for the Poisson distribution, Shieh (2001) and Demidenko (2008) for the binomial distribution, Aban et al (2009) and Friede and Schmidli (2010) for the negative binomial distribution. A general practical approach to computing power for non-normal distributions is given by Lyles et al (2007).

3 References

Aban IB, Cutter GR & Mavinga N (2009). Inferences and power analysis concerning two negative binomial distributions with an application to MRI lesion counts data. Computational Statistics and Data Analysis, 53(3): 820-833.

Altman D & Bland JM (1995). Absence of evidence is not evidence of absence. British Medical Journal, 311: 485.

Buse, A (1982). The likelihood ratio, Wald and Lagrange multiplier tests: an expository note. The American Statistician, 36(3), 153-157.

Chen Y-H & Zou X-H (2006). Interval estimates for the ratio and difference of two lognormal means. Statistics in Medicine, 25, 4099-4113.

Comas J, Lumbierres B, Pons X, Albajes R (2013). Ex-ante determination of the capacity of field tests to detect effects of genetically modified corn on nontarget arthropods. Journal of Economic Entomology, 106(4), 1659-1668.

Demidenko E (2008). Sample size and optimal design for logistic regression with binary interaction. Statistics in Medicine, 27(1): 36-46.

Duan JJ, Head G, Jensen A & Reed G (2004). Effects of Transgenic Bacillus thuringiensis Potato and Conventional Insecticides for Colorado Potato Beetle (Coleoptera: Chrysomelidae) Management on the Abundance of Ground-Dwelling Arthropods in Oregon Potato Ecosystems. Environmental Entomology, 33(2): 275-281.

Friede T & Schmidli H (2010). Blinded Sample Size Re-estimation with Negative Binomial Counts in Superiority and Non-inferiority Trials. Methods of Information in Medicine, 49(6): 618-624.

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. Available at http://www.amigaproject.eu/documents/deliverables/

Goedhart PW, van der Voet H, Baldacchino F & Arpaia S (2014). A statistical simulation model for field testing of non-target organisms in environmental risk assessment of genetically modified plants. *Ecology and Evolution*, 4: 1267–1283. http://dx.doi.org/10.1002/ece3.1019.

Hrdličková Z (2006) Comparison of the power of the tests in one-way ANOVA type model with Poisson distributed variables. Environmetrics, 17(3): 227-237.

Krishnamoorthy K & Mathew T (2003). Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals. Journal of Statistical Planning and Inference, 115, 103-121.

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.

Pearson ES & Adyanthāya NK (1929). The distribution of frequency constants in small samples from non-normal symmetrical and skew populations. Biometrika, 21, 259-286.

Perry JN, Rothery P, Clark SJ, Heard MS & Hawes C (2003). Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. Journal of Applied Ecology, 40: 17-31.

Perry JN, ter Braak CJF, Dixon PM, Duan JJ, Hails RS, Huesken A, Lavielle M, Marvier M, Scardi M, Schmidt K, Tothmeresz B, Schaarschmidt F & van der Voet, H (2009). Statistical aspects of environmental risk assessment of GM plants for effects on non-target organisms. Environmental Biosafety Research, 8: 65-78.

Shieh G (2001). Sample Size Calculations for Logistic and Poisson Regression Models. Biometrika, 88(4): 1193-1199.

VSN International (2012). GenStat for Windows 15th Edition. VSN International, Hemel Hempstead, United Kingdom. Web page: www.GenStat.co.uk.