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

Deliverable 9.1
Report describing ERA datasets and simulation model

Environmental Risk Assessment of Genetically Modified Organisms: Overview of field studies, examples of datasets, statistical models and a simulation tool

Authors:

Paul W. Goedhart[1], Hilko van der Voet[1],
Ferdinando Baldacchino[2], Salvatore Arpaia[2]

Organisation names of lead contractors for this deliverable:

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

[2] ENEA: National Agency for New Technologies,
Energy and Sustainable Economic Development,
Centro Ricerche Trisaia, Rotondella, Italy
http://www.trisaia.enea.it

January 2013

Dissemination Level: Public, http://edepot.wur.nl/455502

Contents

1 Introduction

This report describes

1) an inventory of field studies and statistical lessons learned from existing datasets about ecological data from field studies studying differences between varieties of a crop, and
2) statistical models and a simulation tool that may be useful for designing such field studies and analysing the results.

This introductory chapter first recalls the general objectives of statistical modelling in the AMIGA project. Then the two tasks which are reported in this document are described. Results of Task 9.1 (Inventory of existing datasets) are given in Chapters 2 and 3, results of Task 9.2 (Building a statistical simulation model) are in Chapters 4-7. The simulation tool described in these chapters is available on request (e-mail paul.goedhart@wur.nl). In Chapter 8 some preliminary applications in a simple example are given. This work will be further elaborated in Task 9.3 starting in 2013. Chapter 9 gives the references.

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

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

Existing datasets will be studied to characterise baseline conditions found in different bio-geographic regions, and to typify the variation of genotypes and environments (Task 9.1). Based on these results a simulation model will be built (Task 9.2), which will be used to test various statistical approaches for data analysis in relation to the possible design of experiments (e.g. sample size). Statistical approaches will use both difference and equivalence

testing, and a graphical display of assessment results will be developed (Task 9.3). Also for multi-environment studies appropriate statistical methodology will be developed, including the consideration of genotype by environment interaction (Task 9.4). The statistical methods for analysis and design of field trials for Environmental Risk Assessment that give the best performance will be described in protocols for both single-environment (Task 9.3) and multi-environment studies (Task 9.6).

The current report describes the results of the first two tasks in the WP 9 plan of work:

1.1 An inventory of existing datasets – Task 9.1

In order to develop a robust protocol, it is essential to test it using real life datasets. Such datasets will become available in other WPs of the project in which field experiments will be performed with GM plants and their conventional counterparts (WP4, WP5, WP6, WP8). However, this will only result in real data later in the project. The first task is therefore to collect data of field experiments which have been performed in the past. Such datasets will include endpoints which are envisaged to be used in future ERA experiments.

1.2 Building a statistical simulation model – Task 9.2

The assessment of GM plants includes the use of statistical difference and equivalence testing. The protocol will give guidelines for performing such tests. It is important to know the statistical properties of such tests, for example the power and robustness of a test and whether the test has the correct size. Such and other issues related to the protocol can and must be researched by means of a statistical simulation model. A simulation model will therefore be developed mimicking the type of field experiments as will be used in WP4, WP5 and WP6. The simulation model will generate data for various endpoints having different statistical distributions. Typical environmental data are counts or presence/absence data, and a common approach is to use statistical models based on the normal distribution. For large abundances or replicated presence/absence data this may be a valid approach. However, counts of species with low abundance levels typically follow a Poisson distribution, and presence/absence data usually follows a binomial distribution. Clumping might give rise to over-dispersed distributions such as the negative binomial for counts and the beta-binomial for presence/absence data. The simulation model must also encompass the blocking structure of an experiment and the environmental variation which is manifest on a site-to-site and a year to year basis. The parameters of the simulation model will be guided by analysis of the datasets gathered in Task 9.1, and, if necessary, by expert opinion.

2 Overview of existing ERA datasets

Data collection in experimental fields with genetically modified crops has been conducted for many years and a large variability of experimental designs, sampling techniques, guilds of non-target arthropods and statistical methods have been used (e.g. Marvier et al., 2007).

To summarize the different approaches presented in the scientific literature, we selected 25 field studies among those firstly published, where the detection of possible effects of GM plants on natural enemies was the primary goal of the study (Table 2-1). The papers were published from 1992 until 2005 and several different crops were included in the selected references. The table presents some of the indicators relevant to the experimental design, collection methods and statistical analyses performed on the data. None of the papers provided a power analysis for the experiments described.

Table 2-1: Main characteristics of fields experiment using GM crops

Authors	Functional group	Crop	Measurement endpoint	Dimensions	Experim. design	Statistic. method
Al Deeb et al 2001 J. Econom. Ent.	Predators	Maize	Abundance (visual counts)	40 plants, 2 locations	Completely randomised	ANOVA mixed model
Pilcher et al 1997 Env. Ent.	Predators	Maize	Abundance (visual counts)	2 years, 3 replications (6 plants in each), 3 sampling dates	Randomised blocks	ANOVA
Wold et al 2001 J Ent Sci	Predators	Maize	Abundance (visual counts)	2 years, 4 replications, 6 sampling dates	Completely randomised	ANOVA
Al Deeb & Wilde 2003. Env. Ent.	Predators	Maize	Abundance (visual counts, pitfall traps)	2 years, 8 locations	Completely randomised	ANOVA mixed model
Johnson 1997 Env. Ent.	Parasitoids	Tobacco	Parasitism rate	3 years, 15 sites	Randomised blocks	ANOVA
Orr & Landis 1997. J. Econom. Entomol.	Parasitoids Predators	Maize	Egg fate Parasitism rate Visual counts	3 replications (50 plants), 3 sampling dates	Completely randomised	ANOVA
Riddick et al 1998 Ann.Ent.Soc.Am.	Predators	Potato	Abundance (visual counts, sweep nets, pitfall traps)	2 years, 3 sites	Completely randomised	ANOVA
Johnson & Gould 1992. Env. Ent.	Parasitoids	Tobacco	Parasitism rate	9 replications, 2 years	Randomised blocks	Chi-square
Mascarenhas & Luttrel 1997 Env. Ent.	Parasitoids	Cotton	Host survival	4 replications	Completely randomised	ANOVA

Naranjo 2005 Env. Ent.	Predators	Cotton	Diversity	6 years, 3-4 replications	Completely randomised	ANOVA, PCA
Manachini & Lozzia 2002 Boll. zool. agr. bachic.	Soil organisms	Maize	Abundance and diversity (after extraction)	2 separate fields at 8 locations. 50 soil samples	n.a.	ANOVA
Bourguet et al 2002 Env. Biosaf. Res.	Predators Parasitoids	Maize	Abundance Parasitization	2 sites, 4 replications, weekly samplings	Split-plot	ANOVA
Buckelew et al 2000 J. Econ. Entomol.	Predators	Soybean	Abundance (sweep nets)	2 sites, 2 years, weekly samplings	Randomised blocks	ANOVA
Wei-Di et al 2004 Chinese J Agric. Biotec.	Herbivores Predators Parasitoids	Cotton	Abundance and diversity (suction device)	2 years, 3 replications	Completely randomised	ANOVA, Diversity indexes
Jasinsky et al 2003 Environ Entomol	Predators	Soybean Maize	Abundance (sweep nets, sticky traps, soil samples)	24 commercial fields	n.a.	ANOVA
Men et al 2003 Environ Entomol	Herbivores Predators Parasitoids	Cotton	Abundance (sweep nets, visual counts)	3 years, 3 replications, 5 sampling dates	Completely randomised	ANOVA + diversity
Musser & Shelton 2003 J. Econ. Entomol.	Predators	Maize	Abundance Egg predation	2 years 2-10 plants/replication	Randomised block	ANOVA
Reed et al 2001 Ent Exp Appl	Predators	Potato	Abundance (visual counts)	2 years, 6 replications	Latin square	ANOVA
Wolkmar et al 2003 Agricul Ecosys Environ	Predators	Sugar beet	Abundance (pitfall traps)	4 replications	Randomised blocks	ANOVA
WU & Guo 2003 Environ Entomol	Predators	Cotton	Abundance (visual counts)	3 replications	Completely randomised	ANOVA
Duan et al 2004 Environ Entomol	Predators	Potato	Abundance (pitfall traps)	2 years, six replications	Latin square	ANOVA
Wade French et al 2004 Environ Entomol	Predators	Maize	Abundance (pitfall traps)	2 years, commercial fields	n.a.	Canonical correspondence
Candolfi et al 2004 Biocontrol Science and Technology	Predators Herbivores Soil organisms	Maize	Abundance (pitfall traps, yellow traps, frappage)	3 replications (field size)	Completely randomised	Principal response curve, diversity indexes
De La Poza et al 2005 Crop Protection	Predators	Maize	Abundance (visual couints, pitfall traps)	2 locations, 3 years, 3-4 replicates	Completely randomised (split for year and location)	ANOVA
Manachini et al 2004 IOBC/WPRS Bullettin	Soil organisms	Canola	Extraction from soil	3 replications	Completely randomised	Multi variate

However, ecological data collections in agricultural research are routinely conducted with many different scopes and the use of plants resistant to insects is only one of such cases. For an investigation of statistical models for further use in the AMIGA project, the availability of raw data collected in herbaceous crops was necessary. Some data sets available by partners of the project were then screened with the aim of analysing the characteristics of ecological data in different cropping systems. Several functional groups of arthropods were considered, with particular attention to herbivores and their natural enemies.

Case study 1: strawberry. **Source**: Field collection performed by ENEA. **Sampling method**: visual counts. **Functional group studied**: herbivores.
The data set consists of observations conducted for two years (2004 and 2005) in Southern Italy in an experimental field where a collection of 10 strawberry varieties was sampled with the aim of selecting for tolerance/resistance to herbivores (namely, spider mites and aphids). The experimental design was a completely randomized one, with 6 replications (in 2004) and 5 replications (in 2005). Each replicated plot contained 10 plants, all of them were sampled.
Aphys gossypii Glover (the cotton aphid) is a serious pest of several herbaceous crops. Its population was sampled by counting visually the numbers of individuals on one randomly chosen leaf for each plant; samples were taken weekly from 15 March 2004 to 31 May 2004 and from 3 May 2005 to 20 June 2005.
Tetranychus urticae Koch (the red spider mite) is one of the main arthropod pests of strawberry. The data set contains visual counts of spiders present on one randomly chosen leaf on each of the ten plants per plot. Twelve weekly samples were conducted in 2004 between 15 March and 31 May, in 2005 only two samples were conducted (16 and 29 May).

Case study 2: cabbage. **Source**: Field collection performed by Wageningen University. **Sampling method**: visual counts. **Functional groups studied**: herbivores and natural enemies.
The data set consists of observations conducted for two years (2008 and 2009) in The Netherlands. Four cultivars of white cabbage were sampled weekly from early June till end of September by randomly choosing 9 plants per plot. Assessment endpoints were herbivores (an aphid and the diamondback moth) and their natural enemies (two parasitoid and two predator species). While the first herbivore is a sap feeder, the second one is a leaf miner.
Brevicoryne brassicae L. (the cabbage aphid) is a pest of Crucifers. The data set is made of visual counts of aphids.
Plutella xylostella L. (the diamondback moth) is a worldwide pest lepidoptera of Crucifers. The data set contains counts of larvae and pupae.
Diaeretiella rapae McIntosh is a hymenoptera parasitoid which has *B. brassicae* as a host. The data set contains visual counts of mummies and parasitized aphids.
Diadegma semiclausum Hellén is a parasitoid of *P. xylostella* larvae; the data set contains the number of pupae of the herbivore pest parasitized by *D. semiclausum*.
Episyrphus balteatus de Geer is a syrphid fly (Diptera) whose larvae are generalist predators on several insect species, and are active predators of aphids. The available data set contains visual counts of larvae and pupae of the predator.

Chrysoperla carnea Stephens is a generalist predator belonging to the family Neuroptera, aphids are among the favourite preys of the species. The data set contains visual counts of eggs and larvae of the predator.
When considering population dynamics at field level, it is expected that trophic relationships between taxa generally lead to correlations between the abundance of herbivores and their natural enemies. However, several possible outcomes can be expected.

Case study 3: potato. **Source**: Field collection performed by ENEA. **Sampling method**: visual counts. **Functional groups studied**: herbivores and natural enemies.
The data set consists of observations conducted for two years (2001 and 2003) in Southern Italy in an experimental field where a genetically modified potato clone resistant to Coleoptera and its isogenic control were sampled. The goal of the study was to detect any possible effects due to the use of the different varieties on arthropod species assemblages. The experimental design was a completely randomized one, with 3 replications per each treatment. Twelve plants per plot were sampled. Assessment endpoints were specific stages of herbivores and natural enemies (19 of such endpoints were selected in 2001 and 23 in 2003). Ten samplings were conducted in 2001and five sampling dates are available for 2003.

Case study 4: pollinators. **Source**: Field collection performed by Wuerzburg University. **Sampling method**: visual observations. **Functional group studied**: pollinators(honeybees). Direct effects of flowering Bt maize plants on honey bee colony development was studied under semi-field conditions. Inside the tents, the only available pollen source for the bees was maize pollen of the different maize varieties. The data set consists of four treatments (including Bt plants, near isogenic line and further conventional maize varieties) distributed according to a randomized block design with 13-14 replicates (represented by hives). Several endpoints were collected (e.g. foraging behaviour, reproductive capacities, colony development, hatched bees, sealed broods, etc.)

Case study 5: soil DNA sequences. **Source**: Field experiment performed by the Institute for Biodiversity Braunschweig. **Sampling method**: DNA extraction from soil. **Species studied**: soil bacteria.
The dataset is based on bacterial DNA sequences and the number of sequences assigned to the respective genus is reported as measurement endpoint. Five treatments (including soil collected from plots cultivated with genetically modified maize) were tested and the analysis was done on two soil samples (replicates) per treatment. The goal of the study was to associate microbial diversity to the different treatments.

The remainder of this report focuses on relatively simple ecological datasets similar to case studies 1-3, where counts or presence/absence data of non-target organisms are available in randomised block designs. Case studies 4 and 5 concern more specialised experimental designs and types of data, and require the development of statistical methods on a case-by-case basis outside the scope of this report.

3 Analysis of some datasets

Two datasets provided by AMIGA partners have been re-analysed, and results are reported in this chapter. In Section 3.1 some statistical characteristics of the ENEA strawberry data are described, in Section 3.2 this is done for the WU cabbage data.

3.1 ENEA strawberry experiments, case study 1

Data were provided by ENEA. Ten different strawberry varieties were assessed for resistance to arthropods without a reference variety. Two taxa were sampled: aphids and spider mite. A completely randomized block design was used with 6 blocks and 10 plants per plot. Taxa on one leaf per plant were weekly sampled. The experiment was conducted in 2004 over a period of 12 weeks from March 15 until May 31, and in 2005 over a period of 8 weeks from May 03 until June 20. The observed counts per plants were summed over the 10 plants for each experimental unit and for each sampling date.

3.1.1 ENEA aphids on strawberry 2004

The mean number of aphids classified by the varieties, which are numbered from 1 to 10, on the 12 sampling dates is given in Table 3-1. It is clear that aphids became more abundant during the growing season although even at the end of the experiment there are still plots with very low numbers of aphids.

Table 3-1: Mean number of aphids for each variety in the ENEA strawberry experiment 2004

Variety	15-3	23-3	30-3	06-4	13-4	20-4	27-4	03-5	11-5	18-5	25-5	31-5
1	3.3	1.5	1.3	2.2	5.0	3.2	8.7	9.8	42.3	27.3	50.3	25.8
2	1.7	0.3	0.5	0.5	1.8	4.0	6.2	8.3	39.3	44.7	54.2	27.8
3	1.3	0.8	1.8	0.5	3.0	3.5	4.8	6.2	33.8	31.2	46.0	24.8
4	5.2	1.7	7.0	4.7	6.8	8.7	5.8	15.3	65.7	36.2	27.5	24.5
5	2.7	2.3	2.3	1.7	15.5	6.0	69.5	65.0	58.7	82.5	58.7	58.5
6	2.2	6.5	4.8	2.8	8.3	4.5	22.0	20.7	74.5	74.8	29.5	25.5
7	3.2	1.5	2.7	1.7	7.2	6.0	12.5	28.8	37.2	47.8	24.0	20.7
8	0.0	0.8	0.3	0.4	2.7	5.2	11.3	9.2	73.7	44.8	51.2	24.3
9	2.3	4.3	2.7	1.5	2.7	2.2	1.7	8.2	10.3	8.8	15.7	8.0
10	2.0	2.5	1.3	1.0	3.5	0.7	1.7	8.0	4.3	2.3	8.7	8.3

The observed sums were analysed for each time point separately. The standard analysis of count data employs a Poisson distribution. Such an analysis assumes that the variance equals the mean. A preliminary analysis reveals that there is more variation than according to a Poisson distribution. This phenomenon is known as over-dispersion. It is then convenient to assume, as an approximation, that the variance is proportional to the mean. The proportionality factor is then known as the over-dispersion factor. Here the over-dispersion factor is estimated by means of the Pearson statistic. Alternative analyses methods include the use of the negative binomial distribution or the lognormal distribution. The latter analysis amounts to first taking the logarithm and then doing a normal analysis of variance. However zero observations cannot be log-transformed and therefore, when there are observed zeroes, 0.5 was added before taking logs. The negative binomial distribution assumes that the variance equals $\mu + \omega\mu^2$, and the lognormal distribution assumes that the standard error is

proportional to the mean (or equivalently the variance is proportional to μ^2). Residual plots for the over-dispersed Poisson, negative binomial and lognormal analysis for each sampling date are generally satisfactory and do not indicate a clear preference for either analysis. Of special interest is the estimate of the dispersion parameter. These are given in Table 3-2 for each sampling date. The over-dispersion parameter for the Poisson becomes quite large for later sampling dates, indicating considerable over-dispersion.

Table 3-2: Estimates of dispersion parameter for strawberry aphids data 2004

Date	Poisson Over-dispersion	Negative binomial parameter ω	LogNormal Variance
15-3	3.95	1.26	1.276
23-3	3.07	0.96	1.120
30-3	3.59	0.78	1.030
06-4	1.69	0.31	0.781
13-4	5.66	0.66	1.230
20-4	5.62	0.86	1.337
27-4	8.70	0.42	0.745
03-5	11.77	0.60	1.183
11-5	21.98	0.57	1.189
18-5	23.78	0.51	0.886
25-5	15.14	0.38	0.796
31-5	15.29	0.46	0.790

3.1.2 ENEA aphids on strawberry 2005

Data for the 6th block are missing and therefore this block was omitted from the analysis. The mean number of aphids classified by the varieties on the 8 sampling dates is given in Table 3-3.

Table 3-3: Mean number of aphids for each variety in the ENEA strawberry experiment 2004

Treat	03-5	10-5	16-5	24-5	31-5	06-6	13-6	20-6
1	10.2	25.4	27.6	16.2	76.8	48.2	17.2	3.8
2	4.6	6.6	27.2	22.8	38.0	12.4	6.6	2.8
3	28.0	80.4	120.4	91.4	197.2	122.8	74.2	12.0
4	8.6	27.2	23.8	29.4	79.0	39.0	22.0	10.8
5	12.8	50.8	60.4	65.4	114.2	85.4	27.2	12.6
6	9.0	38.0	89.2	73.2	80.8	40.2	15.0	6.2
7	64.2	72.4	63.8	63.4	90.8	60.6	21.8	6.2
8	8.0	32.2	27.6	28.2	57.0	18.6	16.4	8.8
9	18.8	39.2	88.0	80.0	113.4	182.2	124.0	7.0
10	20.4	56.2	45.8	23.2	47.8	30.4	10.6	1.2

Residual plots for the Poisson are not satisfactory with increasing residuals with increasing fitted values. The plots for the negative binomial and the lognormal distribution are generally satisfactory. The estimates of the dispersion parameter is given in Table 3-4.

Table 3-4: Estimates of dispersion parameter for strawberry aphids data 2005

Date	Poisson Over-dispersion	Negative binomial parameter ω	LogNormal Variance
03-5	33.21	1.45	2.330
10-5	57.90	1.01	1.972
16-5	69.96	0.82	1.492
24-5	71.93	0.99	1.974
31-5	125.56	0.82	1.361
06-6	107.82	0.91	1.562
13-6	51.19	0.81	1.438
20-6	8.59	0.60	0.942

Again there is heavy over-dispersion as compared to the Poisson distribution.

3.1.3 ENEA mites on strawberry 2004

There are 19 out of 60 plots which do not have any mites at all during the sampling period. Up till May 3 there are at most 5 plots with a positive number of mites, and on the last four sampling dates there are respectively 11, 8, 20 and 25 plots (of totally 60 plots) with positive numbers of mites. Some counts are quite large, again revealing over-dispersion. Analysis of the counts for the last two sampling dates (25-5 and 31-5) gives a Poisson over-dispersion parameters of 3.48 and 2.26 respectively, while the index parameter (ω) of the negative binomial distribution is estimated as 1.75 and 0.77. The large over-dispersion factor for the experiment on 25-5 is mainly due to a count of 110 for treatment 5 in the first block. Residual plots are not very informative due to the large number of zeroes. This experiment might indicate that there are more zero observations then is predicted by a count distribution.

3.1.4 ENEA mites on strawberry 2005

Mites were only sampled on May 16 and 29 and again counts for the 6^{th} block are completely missing. On May 16 the number of mites is low, with the notable exception of two plots, while at May 29 mites are observed on every plot. The over-dispersion factor for the Poisson analysis equals 6.95 and 13.84 respectively, while the index parameter (ω) of the negative binomial distribution is estimated as 1.75 and 0.37. The value of 1.75 is due to two large counts. The residual plots for the negative binomial distribution is more satisfactory than for the Poisson distribution, especially for the first sampling data.

3.1.5 Conclusion Strawberry experiment

Counts of aphids and mites generally have a large over-dispersion as compared to the Poisson distribution. The 2005 experiment reveals that the Poisson assumption, i.e. a variance which is proportional to the mean, is not realistic for aphids. An analysis employing the lognormal distribution seems generally acceptable although such an analysis has the disadvantage that zero counts do not fit in naturally. Furthermore it may seem unlikely that varieties with very small means would have the same proportionality factor for the variance as varieties with large means. The negative binomial distribution on the other hand scales naturally between a Poisson distribution for small means and something similar to the lognormal distribution for large means. The index parameter ω of the negative binomial distribution ranges between 0.31 and 1.75 for the data analysed here. The mites data obtained in 2004 has many zeroes,

some of which might not have happened by chance. It is therefore important. A simulation model should therefore have the ability to simulate excess zeroes.

3.2 WU datasets White Cabbage

A dataset on white cabbage was made available by Wageningen University. Four different non-GM cultivars of white cabbage were assessed in a randomized block design (Kos 2012, Chapter 4): Christmas Drumhead (CD) and Badger Shipper (BS) (Centre for Genetic Resources, CGN, Wageningen, The Netherlands), representing older, open pollinated, cultivars, and Lennox (Len) and Rivera (Riv) (Bejo Zaden BV, Warmenhuizen, The Netherlands), representing more recently cultivated, commercially grown, F1 hybrids.

In two study years, 2008 and 2009, during 14 weeks, from week 23 (early June) until week 36 (early September), central plants of each plot were monitored weekly for the presence of the following insect species: non-mining caterpillars and pupae of *P. xylostella* and pupae of its parasitoid *Diadegma semiclausum* Hellén (Hymenoptera: Ichneumonidae); colony size of *B. brassicae* aphids and other aphids; mummies (pupae of the parasitoid inside the host integument) of the aphid parasitoid *Diaeretiella rapae* McIntosh (Hymenoptera: Braconidae); larvae and pupae of the predator *Episyrphus balteatus* de Geer (Diptera: Syrphidae) and eggs, larvae and pupae of the predator *Chrysoperla carnea* Stephens (Neuroptera: Chrysopidae).

We consider here the count data for the species for which data were available in both years: *P. xylostella* larvae and pupae (Pxl and Pxp), *B. brassicae* aphids (Bb), other aphids (Oa), *Chrysoperla carnea* eggs and larvae (Cce, Ccl), *Episyrphus balteatus* larvae and pupae (Ebl, Ebp), and other predatory hoverflies larvae and pupae (Ophl and Ophp). In 2008 the nine central plants of each plot were monitored every week. In 2009 seven plants were monitored in weeks 23-29, and six plants in weeks 30-36. Counts were summed over plants per plot.

3.2.1 Summary statistics

The mean abundances per week are very different (Table 3-5). Relatively high mean numbers (10-75) are found for Pxl, Bb and Oa, but mean numbers are below 1 for the other responses. There are also large differences between cultivars, e .g. Bb mean abundance varies between 1.5 on Len and 67 on CD.

Table 3-5: Mean abundances 2008-2009, cabbage data

cultivar	Pxl	Pxp	Bb	Oa	Cce
Riv	6.36	0.64	1.49	21.24	0.31
Len	7.84	0.62	5.70	39.43	0.62
CD	12.60	0.95	67.16	181.50	1.14
BS	13.23	0.93	11.60	58.74	0.82
Mean	10.01	0.78	21.49	75.23	0.72
cultivar	Ccl	Ebl	Ebp	Ophl	Ophp
Riv	0.5223	0.0268	0.2411	0.0000	0.0937
Len	0.9062	0.0580	0.3214	0.0000	0.0938
CD	0.6125	0.2741	1.2036	0.3098	0.6625
BS	0.6464	0.0634	0.4527	0.0732	0.1821
Mean	0.6719	0.1056	0.5547	0.0958	0.2580

Note that summed over the season (14 weeks) all mean abundances are 14 times as large, and therefore always above 1.

There are also large differences between years (Table 3-6). Oa, Ccl and Ebp are far more abundant in 2008, while Pxl and Bb are far more abundant in 2009. But this also varies per cultivar, e.g. Bb numbers on Riv are almost equal in both years.

Table 3-6: Mean abundances per year, cabbage data

Year 2008					
cultivar	Pxl	Pxp	Bb	Oa	Cce
Riv	1.91	0.5536	1.45	39.18	0.18
Len	2.43	0.4643	1.84	74.66	0.53
CD	1.78	0.6250	3.11	332.36	0.51
BS	2.97	0.7857	2.06	97.93	0.38
cultivar	Ccl	Ebl	Ebp	Ophl	Ophp
Riv	1.0268	0.0089	0.4821	0.0000	0.1875
Len	1.7946	0.0446	0.6339	0.0000	0.1875
CD	1.1429	0.0536	2.3661	0.5804	1.2946
BS	1.2411	0.0268	0.8750	0.1250	0.3214
Year 2009					
cultivar	Pxl	Pxp	Bb	Oa	Cce
Riv	10.82	0.7179	1.54	3.31	0.44
Len	13.24	0.7679	9.55	4.20	0.72
CD	23.42	1.2661	131.21	30.64	1.78
BS	23.49	1.0661	21.14	19.54	1.25
cultivar	Ccl	Ebl	Ebp	Ophl	Ophp
Riv	0.0179	0.0446	0.0000	0.0000	0.0000
Len	0.0179	0.0714	0.0089	0.0000	0.0000
CD	0.0821	0.4946	0.0411	0.0393	0.0304
BS	0.0518	0.1000	0.0304	0.0214	0.0429

Zero counts are common. The group below 0 in Figure 3-1 left represents the frequencies of count 0 for the 10 endpoints. Less frequent are counts between 1 and 10 (between 0 and 1), 10 and 100 (between 1 and 2), 100 and 1000 (between 2 and 3), and above 1000 (above 3). Table 3-7 shows the fraction of the 32 plots that have a positive count.

Table 3-7: Fraction of plots with a positive count, cabbage data

year-week	Pxl	Pxp	Bb	Oa	Cce	Ccl	Ebl	Ebp	Ophl	Ophp
2008-23	*0*	*0*	0.59	**1**	*0*	*0*	*0*	*0*	*0*	*0*
2008-24	*0*	*0*	*0*	0.72	0.41	*0*	*0*	*0*	*0*	*0*
2008-25	0.22	0.03	*0*	0.78	0.50	*0*	*0*	0.09	*0*	*0*
2008-26	0.87	0.06	*0*	0.88	0.41	0.06	*0*	0.16	*0*	*0*
2008-27	**1**	0.41	0.50	**1**	0.16	0.41	*0*	0.25	*0*	*0*
2008-28	**1**	0.87	0.97	**1**	0.16	0.69	*0*	0.69	0.03	0.06
2008-29	**1**	0.94	0.97	**1**	*0*	0.75	0.06	0.75	0.03	0.25
2008-30	0.87	0.84	**1**	**1**	0.03	0.72	0.19	0.78	0.09	0.19
2008-31	0.41	0.47	0.63	**1**	*0*	0.25	*0*	0.62	0.28	0.38
2008-32	0.22	0.16	0.62	**1**	0.03	0.47	0.09	0.56	0.31	0.53
2008-33	0.09	0.06	0.37	0.87	*0*	0.31	0.06	0.19	0.22	0.22
2008-34	0.03	0.09	0.28	0.72	*0*	*0*	*0*	0.13	0.06	0.06
2008-35	0.03	*0*	0.09	0.66	*0*	0.03	*0*	0.09	0.25	*0*
2008-36	0.06	0.03	0.09	0.59	*0*	0.03	*0*	0.06	0.19	0.16
year-week	Pxl	Pxp	Bb	Oa	Cce	Ccl	Ebl	Ebp	Ophl	Ophp
2009-23	**1**	0.12	0.16	0.87	0.06	*0*	*0*	*0*	*0*	*0*
2009-24	0.62	0.94	0.91	*0*	*0*	*0*	*0*	*0*	*0*	*0*
2009-25	**1**	0.66	0.09	0.72	*0*	*0*	*0*	*0*	*0*	*0*
2009-26	**1**	0.78	0.31	0.91	0.12	*0*	*0*	*0*	*0*	*0*
2009-27	**1**	0.47	0.62	**1**	0.12	0.09	0.19	0.03	*0*	*0*
2009-28	**1**	0.41	0.62	0.87	0.28	0.03	0.25	0.03	*0*	0.03
2009-29	**1**	0.50	0.75	0.75	0.47	0.09	0.28	0.03	0.06	*0*
2009-30	0.84	0.22	0.78	0.47	0.66	0.03	0.12	0.09	*0*	*0*
2009-31	0.50	0.28	0.69	0.25	0.50	0.03	0.22	*0*	*0*	0.06
2009-32	0.22	0.12	0.56	0.12	0.38	0.16	0.16	*0*	0.03	0.03
2009-33	0.06	*0*	0.53	0.06	0.41	0.06	0.03	0.03	0.06	0.03
2009-34	0.03	0.06	0.53	0.03	0.28	*0*	0.06	0.03	*0*	0.06
2009-35	0.09	*0*	0.44	0.03	0.09	0.03	0.09	*0*	0.03	*0*
2009-36	0.12	*0*	0.47	*0*	*0*	*0*	0.16	*0*	*0*	*0*
Mean All Weeks	0.51	0.30	0.49	0.65	0.18	0.15	0.07	0.17	0.06	0.07

The row “Mean All Weeks” shows the mean fraction of positive counts for the ten endpoints (on average 27% across the ten endpoints). Only in 18 of 280 combinations (bold, 6%) all 32 plots had a positive count. In 87 cases (italics, 31%) all 32 plots had a zero count. The distribution of the fractions in the table is given graphically in Figure 3-1, right.

The conclusion is that for an analysis of these or similar data it is essential to have a method that can handle zero counts properly.

Figure 3-1: Distribution of log-transformed counts and of fraction of plots with positive counts. Colours represent the ten endpoints.

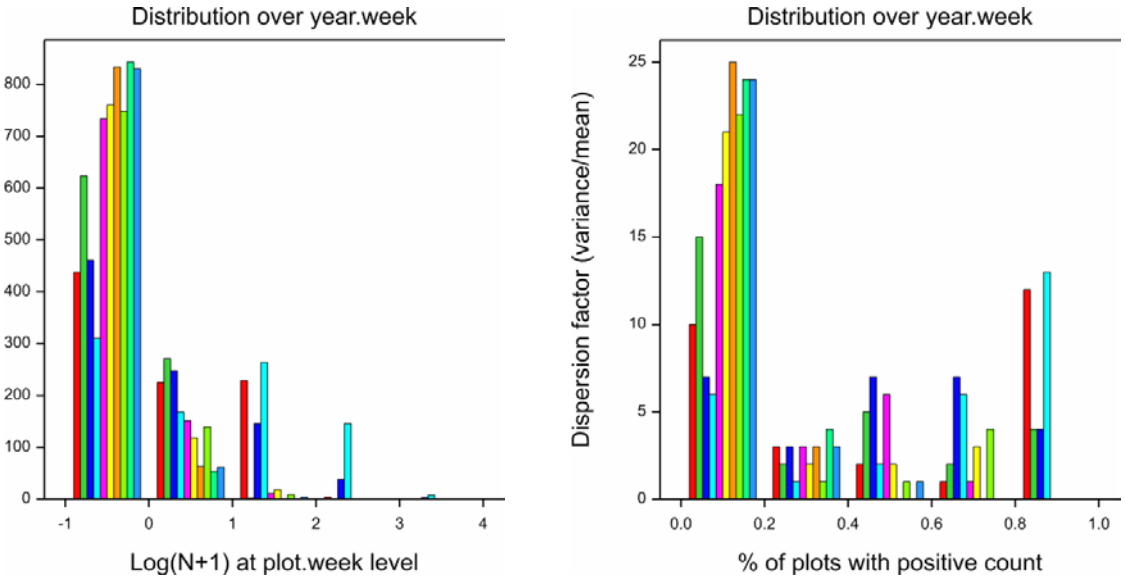


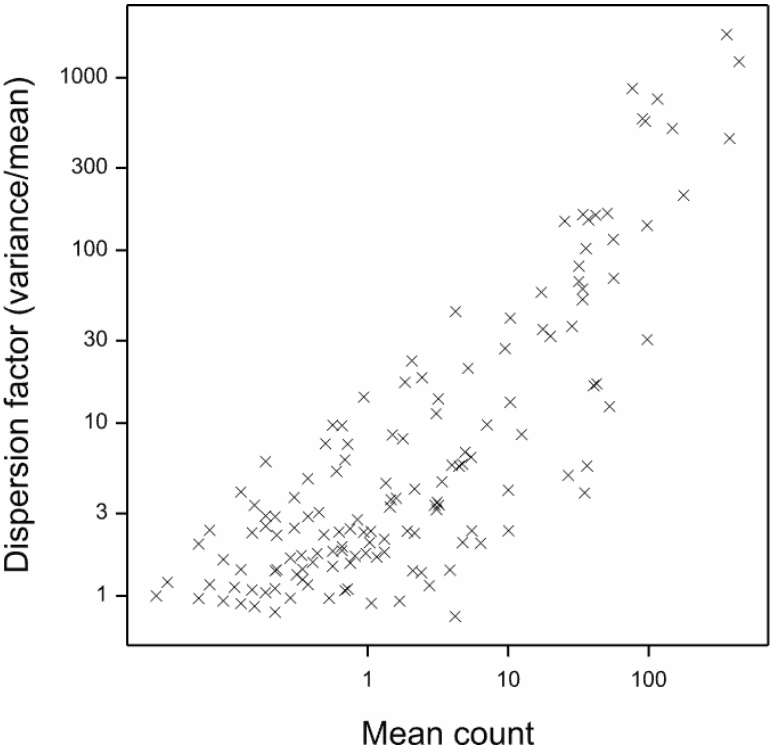
Table 3-8 shows the over-dispersion factors (variance/mean) , each factor based on 32 plots. The results are also plotted against the means (excluding zero means) in Figure 3-2.

Table 3-8: Over-dispersion factors (variance/mean) per year.week, cabbage data

year-week	Pxl	Pxp	Bb	Oa	Cce	Ccl	Ebl	Ebp	Ophl	Ophp
2008-23	-	-	27.0	31	-	-	-	-	-	-
2008-24	-	-	-	41	17.3	-	-	-	-	-
2008-25	1.3	1.0	-	103	8.1	-	-	2.9	-	-
2008-26	1.4	1.0	-	116	2.3	3.3	-	0.9	-	-
2008-27	2.4	1.0	1.1	209	9.7	4.5	-	1.6	-	-
2008-28	4.1	1.4	1.4	446	1.4	3.4	-	3.4	1.0	1.0
2008-29	2.0	1.1	2.0	1240	-	6.3	1.6	5.8	2.0	23.0
2008-30	3.2	0.9	2.4	1785	4.0	5.6	1.1	4.6	0.9	14.2
2008-31	2.8	1.1	1.8	509	-	7.6	-	3.3	2.3	18.4
2008-32	1.4	0.9	1.7	69	1.0	4.2	0.9	2.1	1.9	2.0
2008-33	1.4	1.0	1.5	60	-	1.8	1.0	1.7	1.3	0.8
2008-34	1.0	0.9	1.8	57	-	-	-	0.9	1.0	1.0
2008-35	2.0	-	0.9	35	-	1.0	-	1.4	1.2	-
2008-36	1.0	1.0	0.9	868	-	1.0	-	1.6	1.7	0.9

year-week	Pxl	Pxp	Bb	Oa	Cce	Ccl	Ebl	Ebp	Ophl	Ophp
2009-23	5.7	0.9	6.1	10	2.9	-	-	-	-	-
2009-24	0.9	0.8	8.6	-	-	-	-	-	-	-
2009-25	3.9	2.3	7.6	7	-	-	-	-	-	-
2009-26	5.0	3.5	44.4	36	2.9	-	-	-	-	-
2009-27	12.5	1.7	147.8	140	4.8	0.9	1.1	1.0	-	-
2009-28	16.9	1.5	160.2	52	8.6	1.0	1.0	1.0	-	1.0
2009-29	16.5	1.8	164.3	32	3.7	0.9	1.2	1.0	1.0	-
2009-30	13.2	3.0	66.1	21	5.7	1.2	2.5	1.1	-	-
2009-31	2.4	1.8	81.2	14	11.4	1.2	2.3	-	-	1.2
2009-32	1.7	1.1	161.8	5	3.6	1.0	1.4	-	1.2	1.2
2009-33	1.2	-	150.7	10	2.4	1.2	2.4	1.2	1.2	1.2
2009-34	1.2	1.2	755.5	6	2.4	-	2.3	1.2	-	1.2
2009-35	1.1	-	579.5	6	3.7	1.2	2.5	-	1.2	-
2009-36	2.2	-	559.9	-	-	-	1.0	-	-	-

Figure 3-2: Relation between dispersion factor and mean count, cabbage data.



Generally over-dispersion factors are larger than one. In general factors increase with the mean. The largest over-dispersion is found for the most abundant species, Bb and Oa. Over-dispersion could be due to differences between cultivars; therefore the over-dispersion per cultivar for these two species (each over-dispersion factor now based on 8 values) is given in Table 3-9. This shows that over-dispersion remains a n important issue also at the level of plots with the same cultivar.

Table 3-9: Over-dispersion factors (variance/mean) per year.week.cultivar, cabbage data

year-week	species Bb				species Oa			
	Riv	Len		Riv	Len		Riv	Len
2008-23	13.1	33.5	19.5	5.0	7.8	10.2	35.4	22.6
2008-24	-	-	-	-	52.5	14.2	38.8	5.8
2008-25	-	-	-	-	43.4	195.3	79.5	5.6
2008-26	-	-	-	-	21.3	138.5	99.5	26.8
2008-27	0.9	0.7	2.0	0.9	19.1	169.8	183.2	40.5
2008-28	2.4	0.7	2.3	0.6	37.9	157.0	371.2	125.2
2008-29	1.9	1.5	1.1	3.1	41.6	52.5	1276.7	63.9
2008-30	2.0	1.5	0.9	4.8	30.8	26.5	2369.5	116.4
2008-31	1.0	0.5	0.8	3.8	30.3	7.5	493.3	583.2
2008-32	0.6	0.6	1.2	2.0	2.6	8.4	29.2	27.6
2008-33	-	0.6	1.1	1.5	1.5	9.1	25.1	30.3
2008-34	1.0	1.0	1.1	2.0	1.8	13.3	38.9	13.1
2008-35	-	-	0.9	1.0	-	1.6	14.9	16.1
2008-36	-	-	0.9	1.0	0.9	2.3	832.1	25.7
year-week	species Bb				species Oa			
	Riv	Len	CD	BS	Riv	Len	CD	BS
2009-23	4.0	-	5.7	-	20.0	1.8	8.2	8.2
2009-24	10.0	8.7	8.7	8.2	-	-	-	-
2009-25	-	10.0	5.0	1.0	3.5	9.4	6.3	7.4
2009-26	1.8	60.0	12.9	48.8	10.8	18.7	23.1	35.2
2009-27	8.3	97.4	178.5	26.9	11.6	16.9	116.4	12.8
2009-28	3.6	154.4	95.2	225.9	2.9	3.5	20.4	28.8
2009-29	9.2	94.1	109.1	266.9	1.1	3.6	20.0	6.3
2009-30	8.0	19.6	34.2	24.6	2.7	-	23.1	6.9
2009-31	1.2	2.7	20.7	4.2	-	-	6.0	6.0
2009-32	-	4.9	127.6	15.7	-	-	5.1	4.9
2009-33	1.2	1.2	87.4	9.0	-	-	6.0	12.0
2009-34	6.0	-	505.8	10.4	-	-	6.0	-
2009-35	-	-	376.8	10.8	-	-	6.0	-
2009-36	-	-	345.3	26.0	-	-	-	-

3.2.2 Fitting some possible models

There are many possible statistical models for count data. The work in this chapter was guided by making the following short inventory of models (a fuller treatment of statistical distributions for counts can be found in Chapter 0). In this list Var stands for variance, μ for the mean of the distribution, and ~ indicates proportionality.

1. Data transformation followed by an analysis based on the normal distribution
    a. Square root transformation. Var ~ μ
    b. Logarithmic transformation. Var ~ μ^2
       Zeroes present a problem for the log transform. Some common approaches:
        i. Analyse log(N+1) or log(N+0.5)
        ii. Replace zeroes by 0.5
2. Generalised Linear Models (GLMs)
    a. Poisson. Var = μ
    b. Poisson corrected for over-dispersion. Var ~ μ
    c. Negative Binomial. Var = $\mu + k\mu^2$
    d. Gamma. Var ~ μ^2. Zeroes present a problem for the Gamma distribution. A possible approach is to replace zeroes by 0.5
3. Zero-excess models
    a. Zero-Inflated models, mixture of spike at zero and distribution
        i. Zero-Inflated Poisson (ZIP)
        ii. Zero-Inflated Negative Binomial (ZINB)
    b. Two-part, conditional (or hurdle) models
        i. Binomial + Truncated Poisson (excluding the zeroes)
        ii. Binomial + Truncated Negative Binomial (excluding the zeroes)
4. Generalised Linear Mixed Models (GLMMs)
    a. Modelling correlated counts
    b. Modelling random effects
5. Combinations of the above
6. Binomial model on simplified data (data reduced to absence/presence)

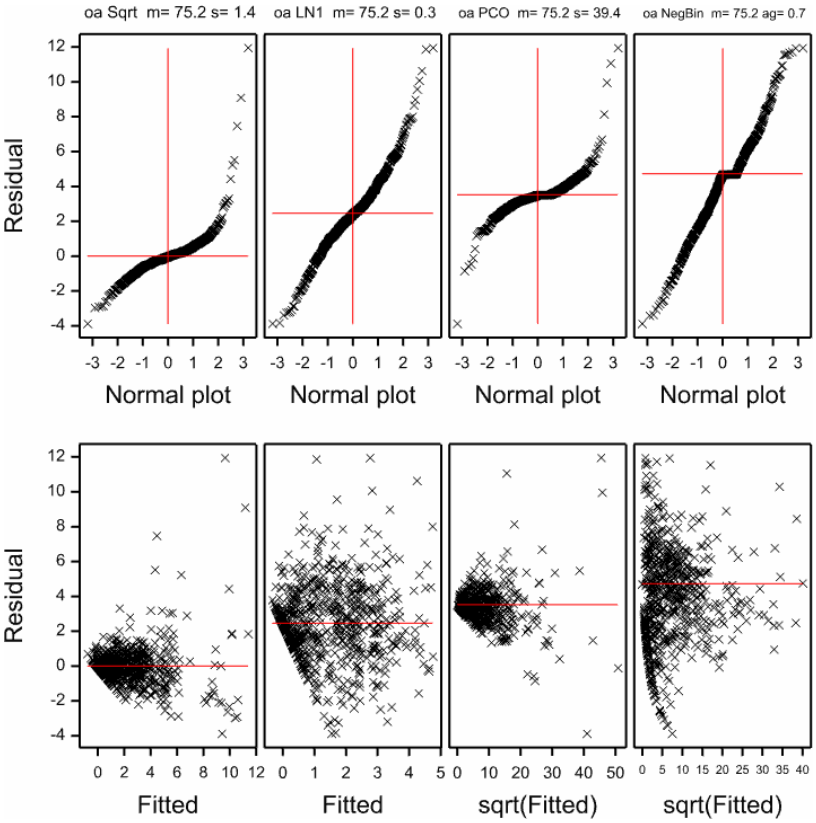
Some first comparisons are made using models

- 1a (square root transformed counts),
- 1bi (log(N+1) transformation),
- 2b (Poisson corrected for over-dispersion) and
- 2c (Negative Binomial).

The model terms fitted are: `plot+(year/week)*cultivar` in all cases. Normal plots and residual plots for the following three endpoints are shown:

- Oa which has high counts (mean count 75), see Figure 3-3
- Pxl which has moderate counts (mean count 10), see Figure 3-4
- Cce which has low counts (mean count 0.7), see Figure 3-5

Figure 3-3: Normal and residual plots for Oa. Comparison of four models.



Differences are largest for endpoint Oa, i.e. the high count case (Figure 3-3). The normal plots deviate most from a straight line for the square root transformation and the overdispersed Poisson model; moreover the residual plots show larger residuals for larger fitted values. So clearly a model in which the variance is proportional to the mean is not appropriate for these data. Normal and residual plots for the lognormal and the negative binomial model are satisfactory; the residual plot of the negative binomial model is to be preferred with a more homogeneous spread of residuals.

In the moderate (Figure 3-4) and low (Figure 3-5) count cases it is less obvious which model to choose. However, the large frequency of zeroes makes any model that needs an ad-hoc approach to replace these values, such as the log transformation, unattractive. We conclude that among the four models the negative binomial model has the best performance to be applicable in many cases, both with high and low counts. Other models, such as the zero-excess models still have to be investigated.

The final modelling choice will be made later in the project based on studies using the simulation model described in the next chapters.

Figure 3-4: Normal and residual plots for Pxl. Comparison of four models.

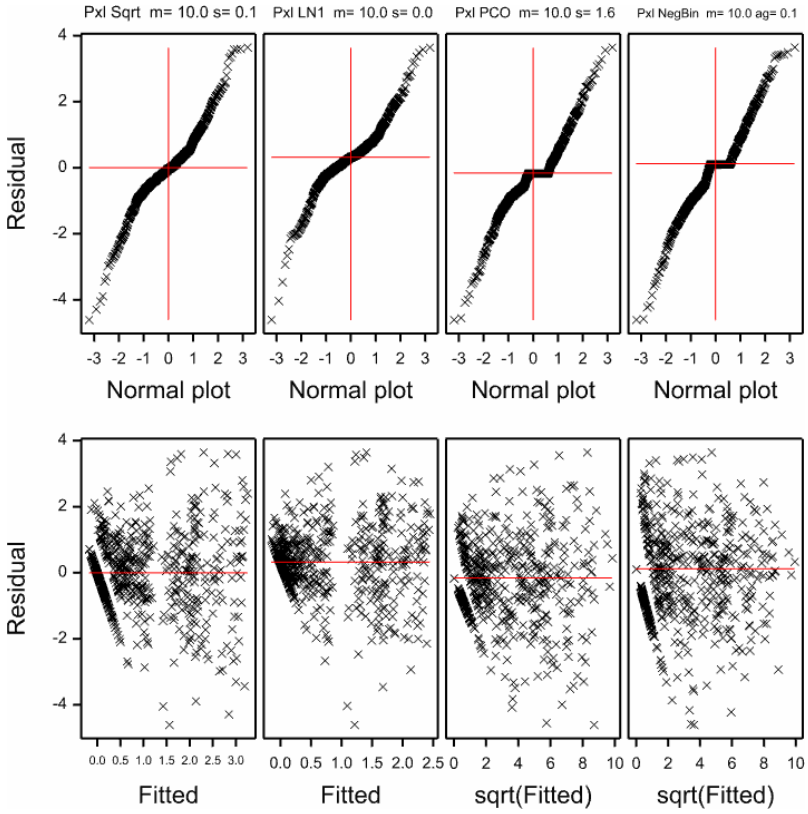
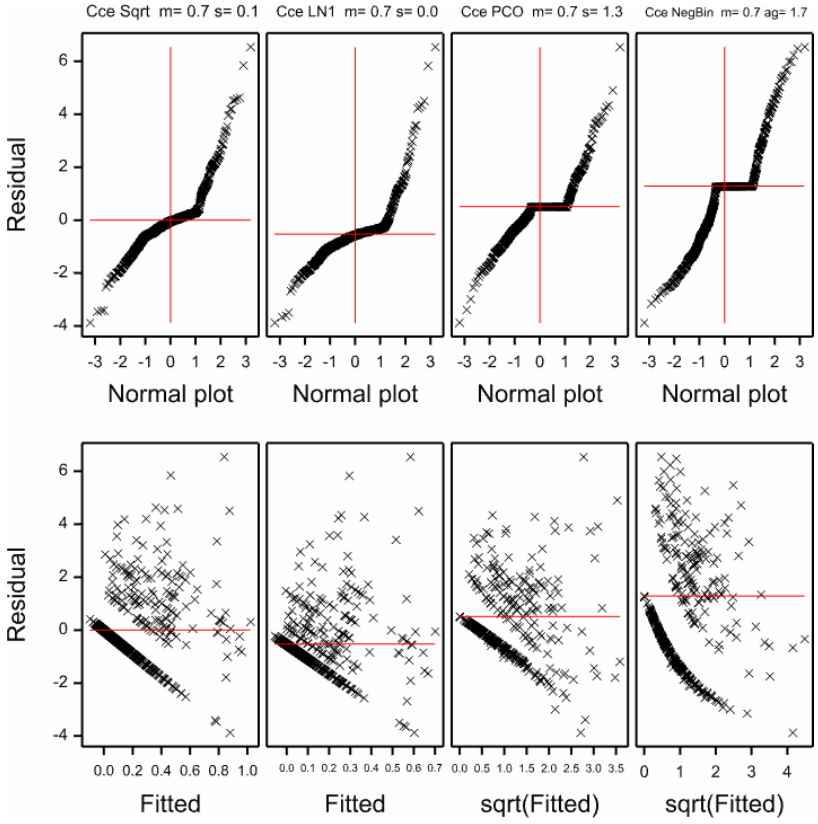


Figure 3-5: Normal and residual plots for Cce. Comparison of four models.



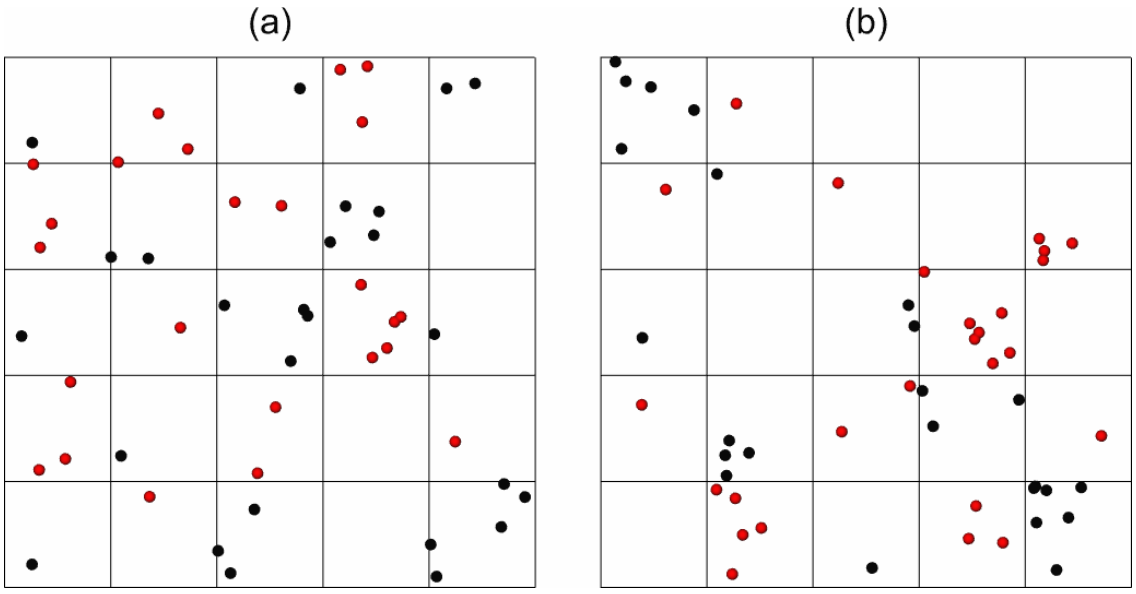
4 Statistical distributions for counts and presence/absence data

As was seen in previous chapters, typical environmental data are counts or presence/absence data. The basic distribution for counts without a formal upper limit, is the Poisson distribution. Presence/absence data arise when for instance the number of plants is counted on which an organism is present. Such data can only take the values 0 up to the total number of plants, and one might also think of such data as percentages. The basic distribution for presence/absence data is the binomial distribution. Clumping or mixing might give rise to over-dispersed distributions and some of these are considered here for both the Poisson and the binomial distribution. Finally the number of zero observations can be larger than predicted by the count distribution. This is termed excess-zeros and this class of distributions is described in the final paragraph of this chapter.

4.1 Poisson distribution

The basic distribution for counts is the Poisson distribution. The Poisson distribution arises when events occur independently of each other but at a fixed rate in time or space. The number of counts in a fixed time- or space-interval then follows a Poisson distribution. As a simple example consider N randomly drawn points in a unit square. Further suppose that the unit square is divided in K cells with equal area by a regular grid. The number of points in each cell then follows a Poisson distribution with mean N/K. An example of this is given in Figure 4-1(a) in which 50 points with random x- and y-coordinates are depicted. The number of points in each cell in this example equals, running from top left to bottom right, 1,3,1,3,2, 2,2,2,4,0, 1,1,4,5,1, 3,1,2,0,1, 1,1,3,0,5. The mean of the counts equals 2 and the variance equals 2.16. This is very close to the theoretical variance which equals the mean of the Poisson distribution.

Figure 4-1: (a) Example of a spatial Poisson process with 50 randomly drawn points in a unit square and 25 cells defined by a regular grid, and (b) example of a spatial process with 50 points which are clumped.

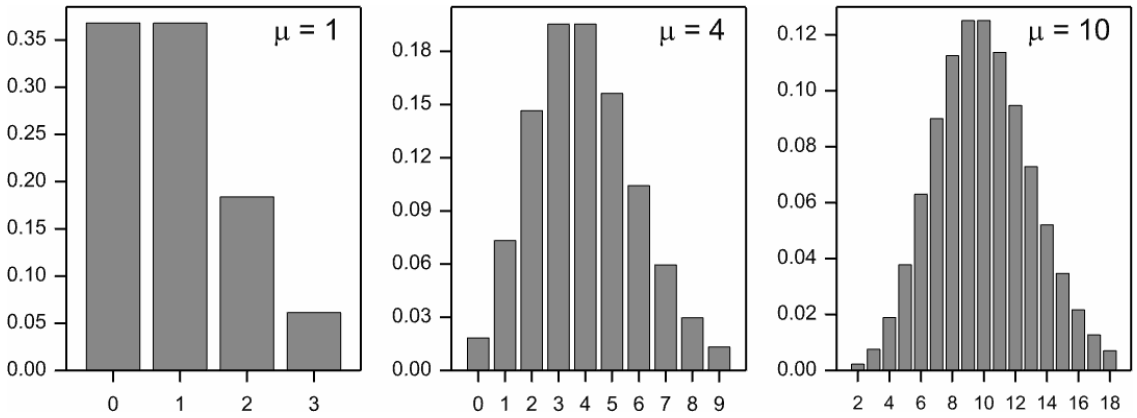


The probabilities of the Poisson distribution with positive mean μ are given by

$$P(X = x) = e^{-\mu}\frac{\mu^x}{x!} \qquad x \geq 0 \tag{1}$$

The mean and variance of the Poisson distribution both equal μ and the skewness of the distribution equals $1/\sqrt{\mu}$. Examples of three Poisson distributions are given in Figure 4-2.

Figure 4-2: Examples of three Poisson distributions with means 1, 4 and 10.



It is clear that the Poisson distribution becomes more symmetric as the mean increases. For large μ the distribution is well approximated by the normal distribution. The variance stabilizing transformation is $\sqrt{X}$ in the sense that for large μ the mean of the transformed stochastic variable is $\mathbb{E}(\sqrt{X}) \approx \sqrt{\mu}$ and the variance $\mathbb{V}(\sqrt{X}) \approx 1/4$, see McCullagh and Nelder (1989). The square root transformation is sometimes used for statistical inference based on the normal distribution rather than on the Poisson distribution.

4.2 Over-dispersion relative to the Poisson distribution

The Poisson distribution assumes a fixed rate of events in time or space. However frequently this rate might vary in different time- or space-intervals. In the spatial context this gives rise to what is called clumping, i.e. points tend to clump together. An example of this is given in Figure 4-1(b). The number of points in each cell now equals, again running from top left to bottom right, 5,1,0,0,0, 1,1,1,0,4, 1,0,2,7,0, 1,4,2,3,1, 0,5,1,3,7. The mean of the counts equals 2 and the variance equals 4.75 which is much larger than the mean. A common way to model this is to assume inter-subject variability, also called mixing. It is then assumed that a count X follows a Poisson distribution with mean Z, where Z itself is a random variable with mean μ and variance σ^2. The marginal mean and variance of the distribution of X is then given by

$$\mathbb{E}(X) = \mu \quad \text{and} \quad \mathbb{V}(X) = \mu + \sigma^2 \tag{2}$$

Since σ^2 is a positive variance parameter this results in a distribution with a variance larger than the mean and this is termed over-dispersion. Also, Feller (1943) has shown that the probability of zero in a mixed Poisson distribution is greater than the probability of zero in an

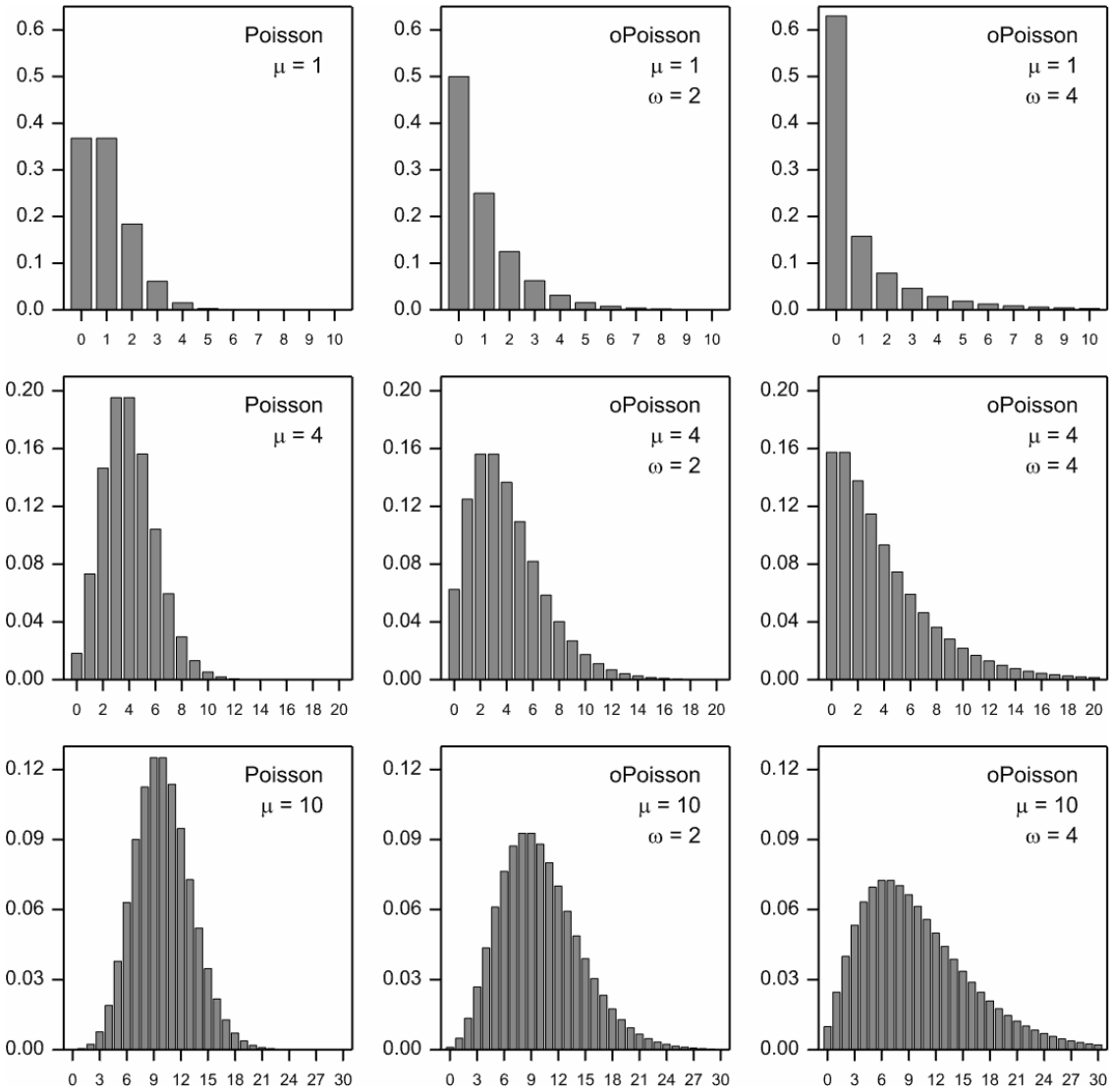
ordinary Poisson distribution with the same mean. There are various ways to specify the mixing distribution of Z and the most common ones are given below.

4.2.1 Over-dispersed Poisson distribution

Suppose that the variance of Z is proportional to the mean such that $\sigma^2 = (\omega - 1)\mu$ where $\omega > 1$. The marginal variance of X is given by equation (2) which results in $\mathbb{V}(X) = \omega\mu$. So in this case the variance of X itself is also proportional to the mean. Further assuming that Z follows a gamma distribution with mean μ and variance $(\omega - 1)\mu$ leads to a special form of the negative binomial distribution, with $\phi = (\omega - 1)^{-1}$ for ease of notation:

$$P(X = x) = \frac{\Gamma(x + \phi\mu)}{x!\ \Gamma(\phi\mu)} \frac{\phi^{\phi\mu}}{(1 + \phi)^{x+\phi\mu}} \qquad x \geq 0 \tag{3}$$

Figure 4-3: Comparison of the Poisson and the over-dispersed Poisson distribution.



This distribution will be named over-dispersed Poisson in order to distinguish it from the more common form of the negative binomial distribution which is given in the next section. The skewness of this distribution equals $(2\omega - 1)/\sqrt{\omega\mu}$, which shows that the over-dispersed

Poisson distribution becomes symmetric less quickly than the Poisson distribution as μ increases. Cumulative probabilities can be calculated directly by means of the regularized incomplete beta function, i.e. $P(X \leq x) = I_{\omega^{-1}}(\mu/(\omega - 1), x + 1)$. Figure 4-3 shows some examples of the over-dispersed Poisson distribution.

For modest amounts of over-dispersion, the difference between maximum likelihood estimates based on (3) and based on the Poisson likelihood (1) may be neglected (McCullagh and Nelder, 1989). Also, using the Poisson likelihood, the dispersion parameter can be estimated by the Pearson Chi-squared statistic or the residual deviance after a Poisson fit. The standard errors of the Poisson maximum likelihood estimates can then be easily adjusted by multiplication with the squared root of the dispersion parameter. This is the so-called quasi likelihood approach (McCullagh and Nelder, 1989). This approach is quite popular, and it is therefore that the over-dispersed Poisson distribution is not frequently used. It is however a convenient vehicle to simulate over-dispersed counts.

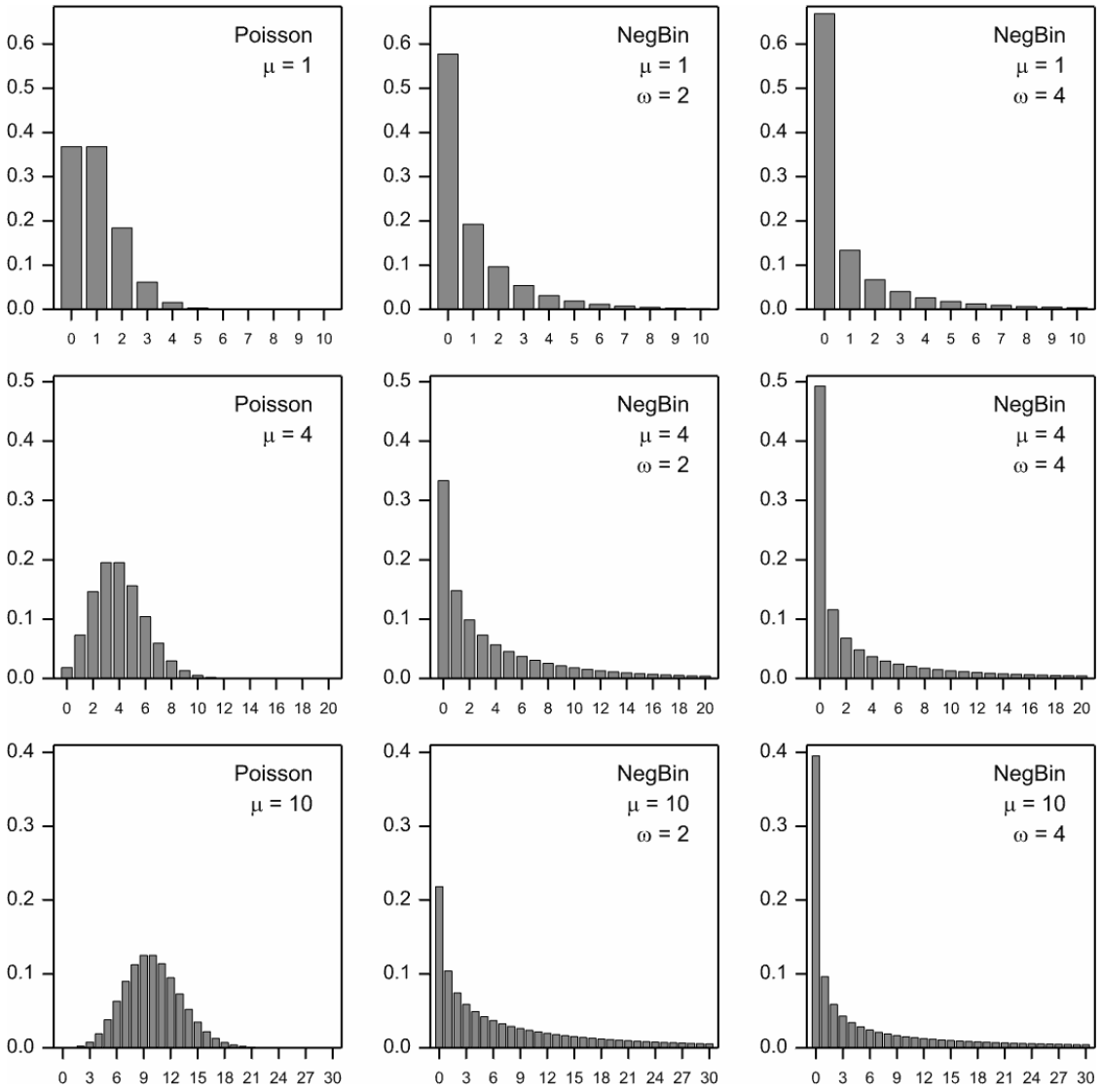
4.2.2 Negative binomial distribution

The negative binomial distribution arises when the mixing distribution Z follows a gamma distribution with mean μ and variance $\omega\mu^2$. The marginal mean of X is then again μ and the variance equals $\mu + \omega\mu^2$. The probability distribution is given by what is generally termed the negative binomial distribution

$$P(X = x) = \frac{\Gamma(x + \omega^{-1})}{x!\ \Gamma(\omega^{-1})} \left(\frac{\omega\mu}{1 + \omega\mu}\right)^x (1 + \omega\mu)^{-\omega^{-1}} \qquad x \geq 0 \tag{4}$$

The skewness of the negative binomial distribution equals $(2\omega\mu + 1)/\sqrt{(1 + \omega\mu)\mu}$. Cumulative probabilities can be calculated directly by means of the regularized incomplete beta function, i.e. $P(Y \leq y) = I_{(1+\omega\mu)^{-1}}(\omega^{-1}, y + 1)$. Some examples of the negative binomial distribution are given in Figure 4-4 along with a Poisson distribution with the same mean. This shows that the negative binomial distribution with large dispersion parameter ω has a large zero probability and a rather flat tail.

Figure 4-4: Comparison of the Poisson and the negative binomial distribution.



4.2.3 Poisson-Lognormal distribution

In Poisson regression it is common to introduce random effects e on the scale of the linear predictor, i.e. to write $\log(\mu) = \alpha + e$, in which e follows a normal distribution with mean 0 and some variance. This is equivalent to assuming that Z follows a lognormal distribution with mean, say λ, and variance say σ^2. For obvious reasons this distribution can be termed Poisson-Lognormal. The mean and variance of the marginal distribution are given below. Note that the mean is not equal to $\exp(\lambda)$.

$$\mathbb{E}(X) = \exp\left(\lambda + \tfrac{1}{2}\sigma^2\right) \quad \text{and} \quad \mathbb{V}(X) = \mathbb{E}(X) + (\exp(\sigma^2) - 1)\mathbb{E}^2(X) \tag{5}$$

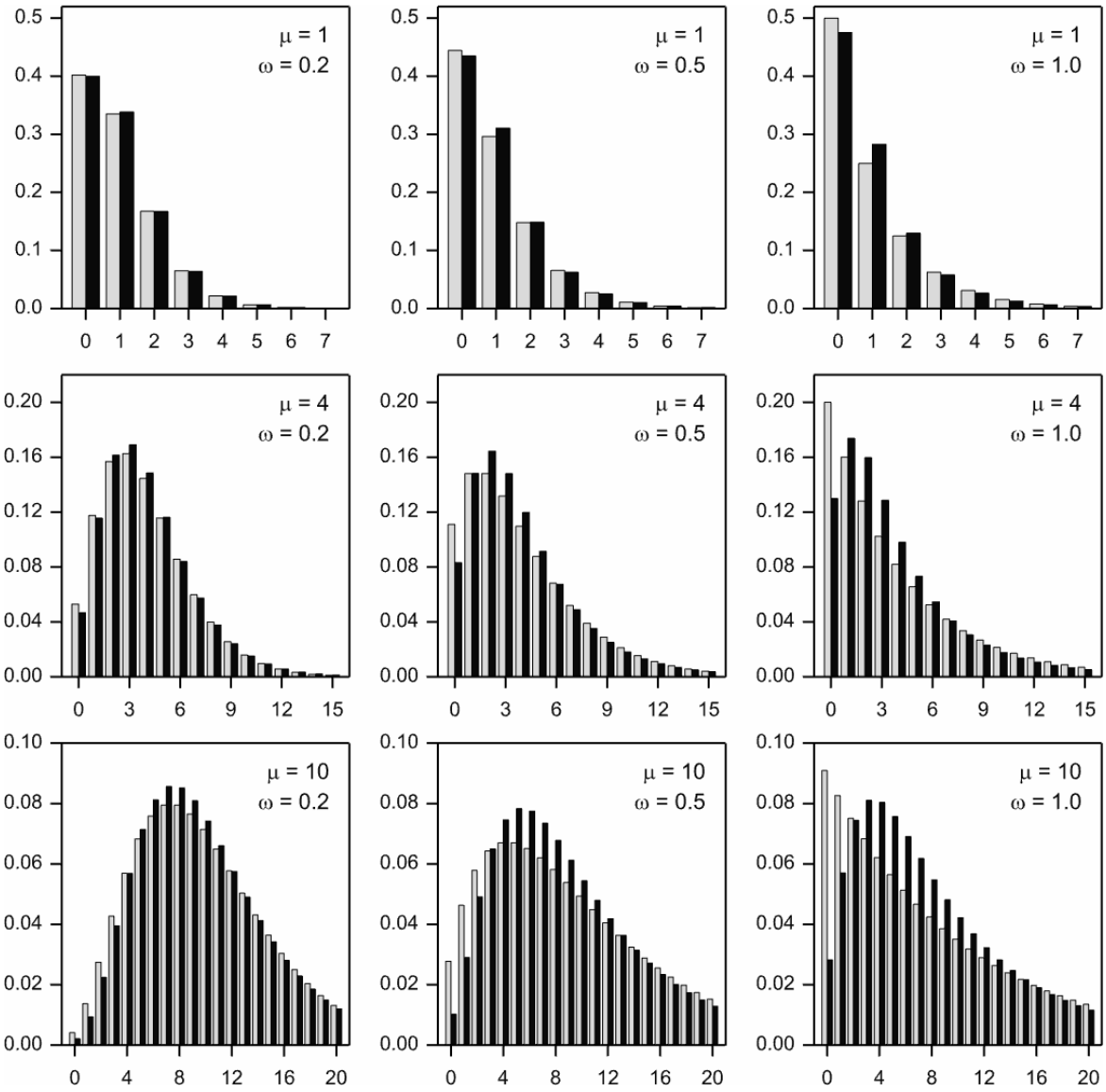
This is thus the same variance function as the negative binomial distribution. Indeed writing $\sigma^2 = \log(\omega + 1)$ and $\lambda = \log(\mu) - \frac{1}{2}\log(\omega + 1)$, the mean and variance of the negative binomial distribution are obtained, i.e. $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \mu + \omega\mu^2$. Probabilities can be obtained by integrating out the random effect. There is no analytic solution to the integral, but a very good numerical approximation can be obtained by what is called Gauss-Hermite

integration. This approximates the integral by a weighted sum of a limited number of function evaluations. In this case, with η_j the so-called Gauss-Hermite nodes and w_j the accompanying weights the Poisson-Lognormal probabilities can be approximated in the following way

$$P(X = x) = \sum_j \frac{w_j}{\sqrt{\pi}} P\left[Y = x \mid Y \sim \text{Poisson}\left(\exp\left(\sqrt{2}\sigma\,\eta_j + \lambda\right)\right)\right] \tag{6}$$

Although the mean and variance of the negative binomial and Poisson-Lognormal are equivalent, the distributions can be quite different for large μ and ω as is shown in Figure 4-5. Note that for small μ there is hardly a difference between the Poisson-Lognormal and the negative binomial distribution, although the difference increases for larger ω with a larger zero probability for the negative binomial distribution.

Figure 4-5: Negative binomial distribution (grey) and Poisson-lognormal distributions (black) with the same mean μ and variance $\mu + \omega\mu^2$.



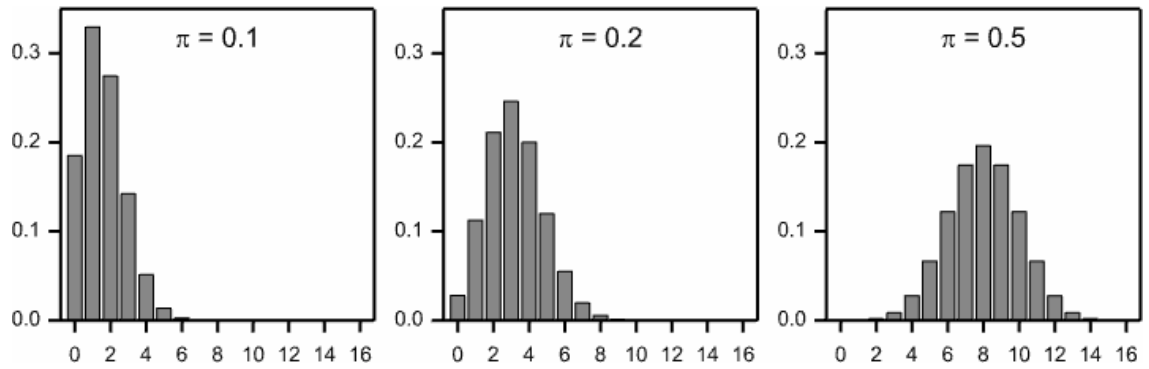
4.3 Binomial distribution

The most familiar example of a binomial distribution is provided by counting the number of heads in successive tosses of a coin. Assuming that the, say *n*, tosses are independent and that the probability of a head equals π, the total number of heads follows a binomial distribution with probability π and so-called binomial denominator *n*. In the AMIGA context this distribution may arise when, rather than counting organisms, the presence or absence of organisms is recorded. The response might then be the number of plants on which a specific organism is present for each experimental unit. The probabilities of the binomial distribution are given by

$$P(X = x) = \binom{n}{x} \pi^x \, (1 - \pi)^{n-x} \qquad x \geq 0 \tag{7}$$

The mean of the binomial distribution is given by $n\pi$ and the variance equals $n\pi(1 - \pi)$. The skewness equals $(1 - 2\pi)/\sqrt{n\pi(1 - \pi)}$ which shows that the distribution is symmetric for $\pi = 0.5$. Examples of the binomial distribution are given in Figure 4-6.

Figure 4-6: Examples of three binomial distributions with *n*=16 and π = 0.1, 0.2 and 0.5.



4.4 Over-dispersion relative to the Binomial distribution

Over-dispersed binomial distributions are obtained by assuming that the number of successes *X* follows a binomial distribution with binomial denominator *n* and probability of success Z where Z itself follows some statistical distribution with mean π and some variance σ^2. It follows that the marginal mean of X itself equals $n\pi$ and the variance equals $n\pi(1 - \pi) + n(n - 1)\sigma^2$ which is larger than the variance of the binomial distribution. Note that, since Z is a probability, its distribution must be defined on the interval (0,1). The most popular choice for *Z* is the beta distribution which results in the so-called beta-binomial distribution. An alternative is to assume that the logit transform of *Z* follows a normal distribution. Details of both distributions are given below.

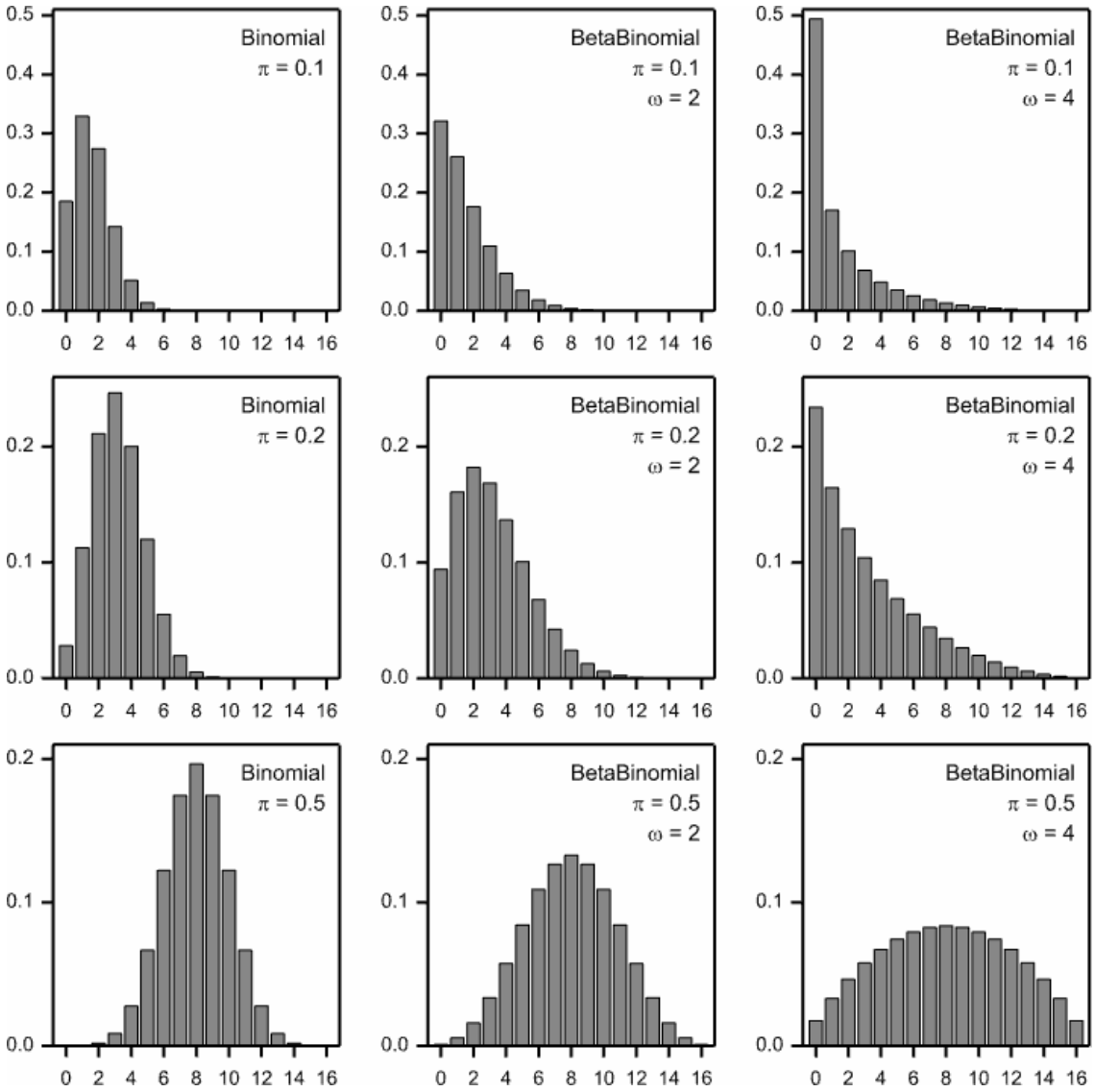
4.4.1 Beta-Binomial distribution

The beta-binomial distribution arises when it is assumed that the probability of success of a binomial distribution itself follows a Beta(α, β) distribution. The beta distribution is defined on the interval (0,1). A convenient re-parameterization is given by $\pi = \alpha/(\alpha + \beta)$ and $\varphi = 1/(\alpha + \beta + 1)$. The mean of the beta-binomial distribution is then given by $n\pi$ and the variance is given by $n\pi(1 - \pi)[1 + (n - 1)\varphi]$. When the number of binomial trials is equal across experimental units, the term between squared brackets is constant and we can write

$\omega = [1 + (n - 1)\varphi]$. It follows that the variance is proportional to the binomial variance, i.e. the variance equals $\omega\, n\pi(1 - \pi)$ in which ω is an over-dispersion parameter. Since $0 < \varphi < 1$, the over-dispersion parameter ω must be in the interval $(1, n)$. In this case data can be easily analysed by the quasi likelihood approach, see the over-dispersed Poisson distribution.

Some examples of the beta-binomial distribution are given in Figure 4-7 along with a binomial distribution with the same mean. This shows that for large values of ω the range of possible outcomes is extended. However for very large values of ω the distribution becomes bath-tub like with large probabilities for outcomes 0 and n and small probabilities for intermediate values.

Figure 4-7: Comparison of the binomial and the beta-binomial distribution for $n = 16$.



4.4.2 Binomial-Logitnormal distribution

In logistic regression, i.e. regression with the binomial distribution, it is common to introduce random effects e on the scale of the linear predictor, i.e. to write $\text{logit}(\pi) = \alpha + e$ in which e follows a normal distribution with mean 0 and some variance σ^2. This is equivalent to assuming that Z follows a logit-normal distribution. For obvious reasons this distribution can

be termed binomial-logitnormal. Unfortunately the mean and variance of the logit-normal distribution cannot be written in analytical form, and this is thus also the case for the binomial-logitnormal distribution itself. Note however that the mean is not given by $n\,\text{logit}^{-1}(\alpha)$. Probabilities can be obtained by integrating out the random effect. There is no analytic solution to the integral, but a very good numerical approximation can be obtained again by Gauss-Hermite integration:

$$P(X = x) = \sum_j \frac{w_j}{\sqrt{\pi}}\, P\left[Y = x \mid Y \sim \text{Binomial}\left(n, \text{logit}^{-1}\left(\sqrt{2}\sigma\,\eta_j + \alpha\right)\right)\right] \qquad (8)$$

in which η_j are the so-called Gauss-Hermite nodes and w_j the accompanying weights.

Figure 4-8: Comparison of the beta-binomial (grey) and binomial-logitnormal (black) distributions for $n = 16$ with the same mean $n\pi$ and variance $n\omega\pi(1-\pi)$.

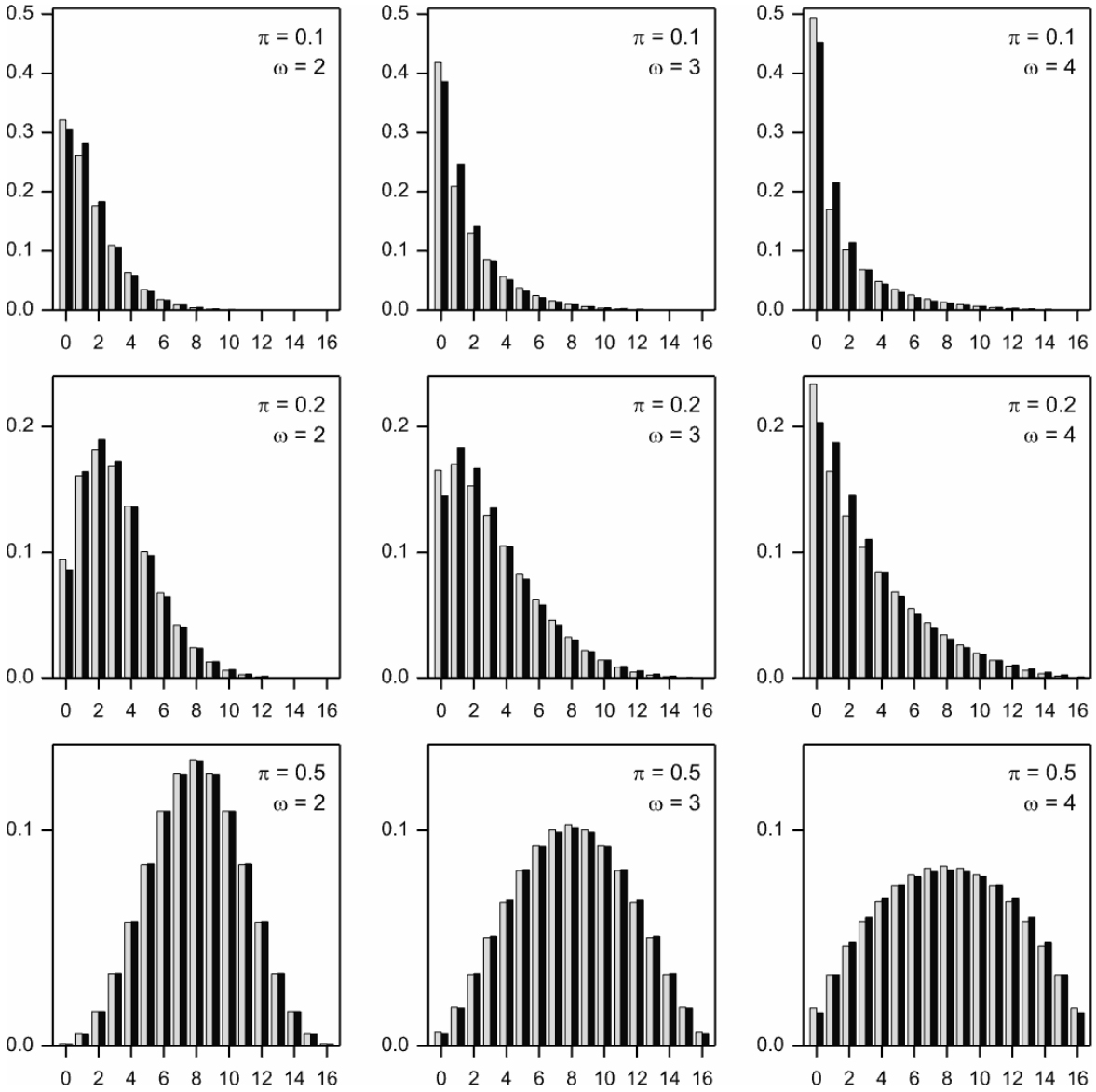


Figure 4-8 compares the binomial-logitnormal distribution with the beta-binomial distribution for parameter values which yield the same mean and variance. This shows that, for π = 0.1

and π = 0.2, the beta-binomial distribution has a somewhat larger zero probability and a smaller probability for values close to zero. For the symmetric case π = 0.5 there is hardly any difference between the two distributions.

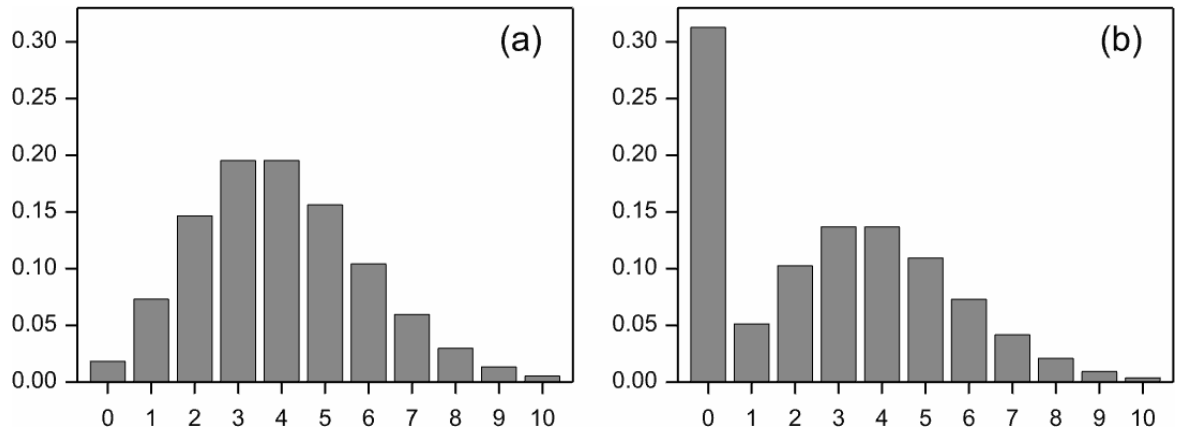
4.5 Excess-zeros distribution

Although the over-dispersed count distributions have a larger zero probability than the corresponding Poisson or binomial distribution, the number of zero observations can still be larger than predicted by the count distribution. This is termed excess-zeros or zero-inflated counts. Examples of situations with excess-zeros are given by Cunningham and Lindenmayer (2005), Sileshi (2008) and Lewin *et al* (2010). Excess zeros can be of interest because zero counts frequently have special status. For example, in counting non-target organisms on plants, a plant may have none of them either because the plant is of no interest whatsoever to the organism, or simply because it so happens by chance that there are no organism on the plant. This is the distinction between structural zeros, which are inevitable, and sampling zeros which occur by chance. A common model employs this distinction by assuming that a proportion δ of the plants have a structural zero and the remaining proportion $(1-\delta)$ of plants follows one of the count distributions given above. The zero-inflated distribution for the resulting count Y is then given by

$$P(Y=y) = \begin{cases} \delta + (1-\delta)P_c(X=0) & y=0 \\ (1-\delta)P_c(X=y) & y>0 \end{cases} \tag{9}$$

in which $P_c(X=x)$ is the distribution of the counts. Note that the probability of observing a zero is given by the probability δ of obtaining a structural zero plus the probability of obtaining a zero by chance. Although in this definition it is possible that $\delta < 0$, we will further assume that $0 \leq \delta < 1$.

Figure 4-9: (a) Poisson distribution with mean 4 and (b) zero-inflated Poisson distribution with the same mean and a zero-inflated probability equal to 0.3.



An example of a Poisson distribution and its zero-inflated counterpart is given in Figure 4-9. The relative probabilities for positive counts of the zero-inflated distribution are equal to the relative probabilities of the ordinary distribution. Figure 4-9b clearly shows the need for an excess-zero distribution because there is a large spike at zero. However, having a lot of zeroes

in itself does not necessarily mean that a zero-inflated model is needed. Examples of this are given in Figure 4-4 with $\omega = 4$.

The mean of a zero-inflated Poisson distribution equals $\mu(1 - \delta)$ and its variance equals $\mu(1 - \delta)(1 + \delta\mu)$. The variance of the zero-inflated Poisson distribution can thus not be written as a function only of its mean and the same holds for the other zero-inflated count distributions. Regression models based on the zero-inflated Poisson distributions were introduced by Lambert (1992) who considered simultaneous modelling of μ and π which are related to possibly different sets of covariates. Greene (1994) brought regression modelling to the zero-inflated negative binomial distribution. Hall (2000) and Vieira et al (2000) seem to be the first papers which employ a zero-inflated binomial model. Finally, Cheung (2006) uses a zero-inflated beta-binomial model to analyse cognitive function test scores of Indonesian children.

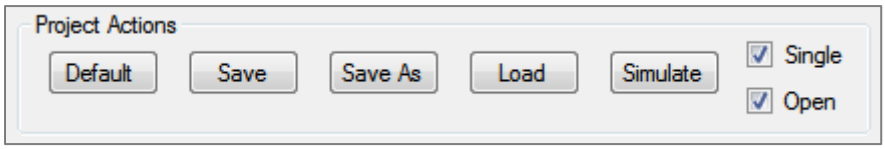
5 General principles of the simulation tool

5.1 AMIGA simulation projects

The AMIGA simulation tool consists of

- A Windows form interface written in C# which can be used to specify all the simulation settings.
- A simulation program written in the R package (R Core Team, 2012) which performs the simulation itself. This requires installation of the R package. Moreover the R packages MASS and XML are needed to perform a simulation.

The Windows form interface stores all simulation settings in a so-called project. Project actions are given at the bottom of the Windows interface:



The *Default* button opens a default project with the most simple simulation settings. This is a good starting point. The *Save* button can be used to save the current project in a XML file. The name of this XML file equals the name of the project (given in the title bar) appended with the file extension ".ami". The project can be saved in a different file by using the *Save As* button which also enables the user to specify a different directory. Previously saved projects can be loaded by means of the *Load* button. Finally the *Simulate* button can be used to first save the current settings to its project file and then performing the requested simulation. The *Single* checkbox can be employed to request that all the simulated experiments should be written to (a) a single file when checked, or (b) to separate files when unchecked. In the former case the simulated data will be written to a single CSV file which has the same name as the project and which will reside in the same directory. In the latter case the separate CSV files will be named "Simulation-#.CSV" where # is a serial number, and the files will be saved in a subdirectory below the directory where the project file is stored; this subdirectory has the same name as the project. The *Open* checkbox can be used to open the single or first CSV file. Note that pushing the *Simulation* button again will automatically delete all previous simulation files, irrespective of the setting of the *Single* checkbox. Finally, running the Windows form program will open the last saved or loaded project.

5.2 General settings of the simulation tool

The most simple use of the simulation tool is for a single trial with a single measurement for each experimental unit. The general settings of the simulation tool for this situation are given below, assuming that a count is recorded rather than presence/absence data. Section 6.1 gives a formal description of this situation. In the context of the AMIGA project an experiment will minimally consist of a genetically modified plant variety (GMO) and a conventional non-GMO comparator which are both replicated a number of times. When there is no blocking, no excess zeroes and only a single measurement, simulating such an experiment only requires specification of the mean μ of the count distribution for both the GMO and the comparator. Adding blocking to this requires specification of a random blocking effect. It is natural and common to introduce blocking effects on the natural log scale, i.e.

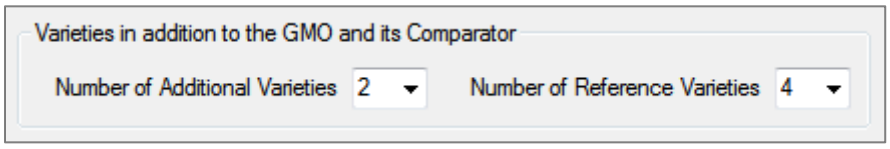
$$\log(\mu) = \text{(variety effect)} + \text{(blocking effect)}$$

since this ensures that the mean μ is always positive. Note that this also requires that the variety effect of the GMO and the comparator are both specified on the natural log scale. Likewise, using the logit transformation for specification of the excess zeroes probability δ ensures that this probability is always in the interval (0,1):

$$\text{logit}(\delta) = \text{(variety effect)} + \text{(blocking effect)}$$

So in the simulation model all effects will be introduced on the log scale for counts and on the logit scale for probabilities.

In addition to the GMO and the comparator other varieties might be introduced in the experiment. These might be other comparators or other GMOs, or alternatively the GMO and/or comparator itself which are treated with for instance herbicides. Although the latter are not varieties but rather treatments, these will also be termed additional varieties in the simulation tool. A special case is an experiment in which the GMO and its comparator are to be compared with a group of reference varieties which have a history of safe use (Van der Voet et al, 2011). In this case the individual reference varieties themselves are not of interest, but they are used to derive baselines or equivalence limits. The reference varieties in the experiment might thus be considered as representing a population of reference varieties. It is then natural to assume that the (logarithm of the) mean of each reference variety is drawn from a statistical distribution. For convenience a normal distribution is used for this. The difference between additional and reference varieties is that for each additional variety a variety effect must be specified, whereas for reference varieties only a common variety effect and an associated variance must be specified. The number of additional and reference varieties can be specified at the top of the Windows interface:



Means on Log-scale

Variety	Level
GMO	2
Comparator	2
Variety 1	2
Variety 2	2
REF Mean	2
REF Variance	0

These settings results in a table of log-means given at the right and requires a user to specify a mean of the count distribution for the GMO, the comparator, two additional varieties, and a common mean and variance for four reference varieties. In each simulated trial the mean of the GMO, comparator and additional varieties will be set accordingly, while the mean of the four reference varieties will vary from trial to trial. When employing a REF Variance equal to zero as in the example above, all the reference varieties will have the same mean as specified by the REF Mean value.

The distribution of the endpoint can be specified in the *Distribution of Counts* groupbox.

The distribution can be either Poisson, OverdispersedPoisson, NegativeBinomial or PoissonLogNormal for counts, or for presence/absence data Binomial, BetaBinomial or BinomialLogitNormal. The latter three distributions requires specification of the binomial denominator *n*. The over-dispersed distributions involve an over-dispersion parameter which is linked to the variance of the distribution as given in Table 5-1 (also see Chapter 0). For the BinomialLogitNormal distribution the over-dispersion parameter is the variance of the extra random effect on the logit scale (see paragraph 4.4.2).

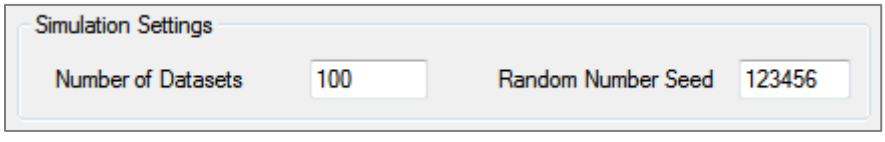
Table 5-1: Mean and variance of the various distributions and definition of the over-dispersion parameter ω.

Distribution	Mean	Variance
Poisson	μ	μ
Over-dispersedPoisson	μ	$\omega\,\mu$
NegativeBinomial	μ	$\mu + \omega\,\mu^2$
PoissonLogNormal	μ	$\mu + \omega\,\mu^2$
Binomial	$n\,\pi$	$n\,\pi\,(1-\pi)$
BetaBinomial	$n\,\pi$	$\omega\,n\,\pi\,(1-\pi)$
BionomialLogitNormal	Not available analytically	

Checking the *Excess Zeroes* box will result in excess zero count distributions which requires specification of excess zero probabilities δ in the same way as the log-means.

The design of an experiment can be either completely randomized or randomized blocks. In either case the number of replications or blocks must be specified. Checking the *Randomized Block* radio button displays a table in which the variance of the block effects must be specified for both the counts and the excess zeroes when required.

The number of datasets to simulate and the seed for the random number generation can be specified in the *Simulation Settings* groupbox. A random number seed of zero uses the current computer time to form a seed. The default random number generator of the R package is used, which is Mersenne-Twister.



6 Simulation model for single trials

6.1 Single trial with one measurement per experimental unit

The mean μ_{ij} of the observed count of variety *i* in replicate or block *j*, or of the success probability π_{ij} for presence/absence data, is given by:

$$\log(\mu_{ij}) = a_j + \alpha_i$$
$$\text{logit}(\pi_{ij}) = a_j + \alpha_i \quad (10)$$

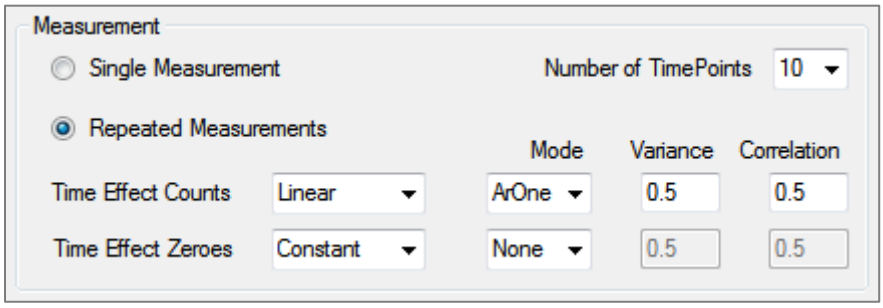
where $a_j \sim \text{N}(0, \sigma_a^2)$ are independent random block effects, and α_i is a fixed variety effect. The variety effects of reference varieties are generated by means of $\alpha_i \sim N(\alpha, \sigma_\alpha^2)$, and this is done separately for each simulated trial. The log link ensures that the mean μ_{ij} is larger than zero, and the logit link ensures that π_{ij} is in the interval (0,1). For a completely randomized experiment the random block effects a_j are all set to zero, which is equivalent to setting $\sigma_a^2 = 0$. Data are subsequently simulated by employing the selected count distribution.
When excess zeroes are requested a similar model is used for the excess zero probability:

$$\text{logit}(\delta_{ij}) = b_j + \beta_i \quad (11)$$

where $b_j \sim \text{N}(0, \sigma_b^2)$ are independent random block effects, and β_i is a fixed variety effect. The logit link ensures that the excess-zero probability δ_{ij} is in the interval (0,1). The variety effects of reference varieties are again generated by means of $\beta_i \sim N(\beta, \sigma_\beta^2)$, and this is done separately for each simulated trial. It is assumed that the random block effects a_j and b_j are independent.

6.2 Single trial with repeated measurements

Non target organism at the same experimental units are frequently sampled at different points in time. Selection of the *Repeated Measurements* radio button will simulate such data:



The number of time points *T* can be specified; it is assumed that time points are equidistant, i.e. 1, 2, …, *T*. The time effect can be either constant, linear or quadratic in time, all on the transformed scale, and this can be set separately for the mean of the counts or for the success probability and also for the excess zero probability. Moreover the repeated observations on the same experimental unit can be independent or can be correlated. This is the purpose of the *Mode* setting: the setting None represents independence, the setting ArOne represents an autoregressive model of order one in time, and the setting Equal results in equal correlations between the repeated measures. Formally the mean of variety *i* in block *j* at time point *t* is given by

$$\log(\mu_{ijt}) = a_j + f_i(t) + v_{ijt} \quad (12)$$

where $f_i(t)$ is a polynomial up to order 2 in time *t* for treatment *i*, which replaces the fixed variety effect α_i in (10). The extra random effect v_{ijt} in (12) specifies the correlation between repeated measures; see below.

A constant time effect is given by $f_i(t) = \alpha_i$, a linear time effect by $f_i(t) = \beta_{i0} + \beta_{i1}t$, and a quadratic time effect by $f_i(t) = \beta_{i0} + \beta_{i1}t + \beta_{i2}t^2$. An alternative parameterization for the second order polynomial with more meaningful parameters is given by

$$f_i(t) = \beta_{i,max} - \left(t - \beta_{i,opt}\right)^2 / \left(2\beta_{i,tol}\right) \quad (13)$$

where the maximum $\beta_{i,max}$ is attained for the optimal time point $\beta_{i,opt}$ and the parameter $\beta_{i,tol}$ represents the width of the quadratic curve, also called the tolerance. This latter parameterization is used by the simulation tool. For a positive tolerance the parabola has a maximum, while for a negative tolerance it has a minimum.

The vector of random effects $\boldsymbol{v}_{ij} = \left(v_{ij1}, \ldots, v_{ijT}\right)$ is assumed to follow a multivariate normal distribution, i.e. $\boldsymbol{v}_{ij} \sim \mathrm{MN}(0, \sigma_v^2 V)$ where V is a T x T symmetric correlation matrix. The value of σ_v^2 can be specified by the *Variance* textbox in the Windows form and the *Mode* and *Correlation* settings specifies the matrix V as follows:

1. *Mode* = None: no extra variability as given by $\sigma_v^2 = 0$.
2. *Mode* = Equal: equal correlation across time by setting $V_{kk} = 1$ and $V_{kl} = \rho$ for $k \neq l$
3. *Mode* = ArOne: autoregressive correlation across time by setting $V_{kl} = \rho^{|k-l|}$.

It is assumed that the block effect a_j and the time effect v_{ijt} are independent.

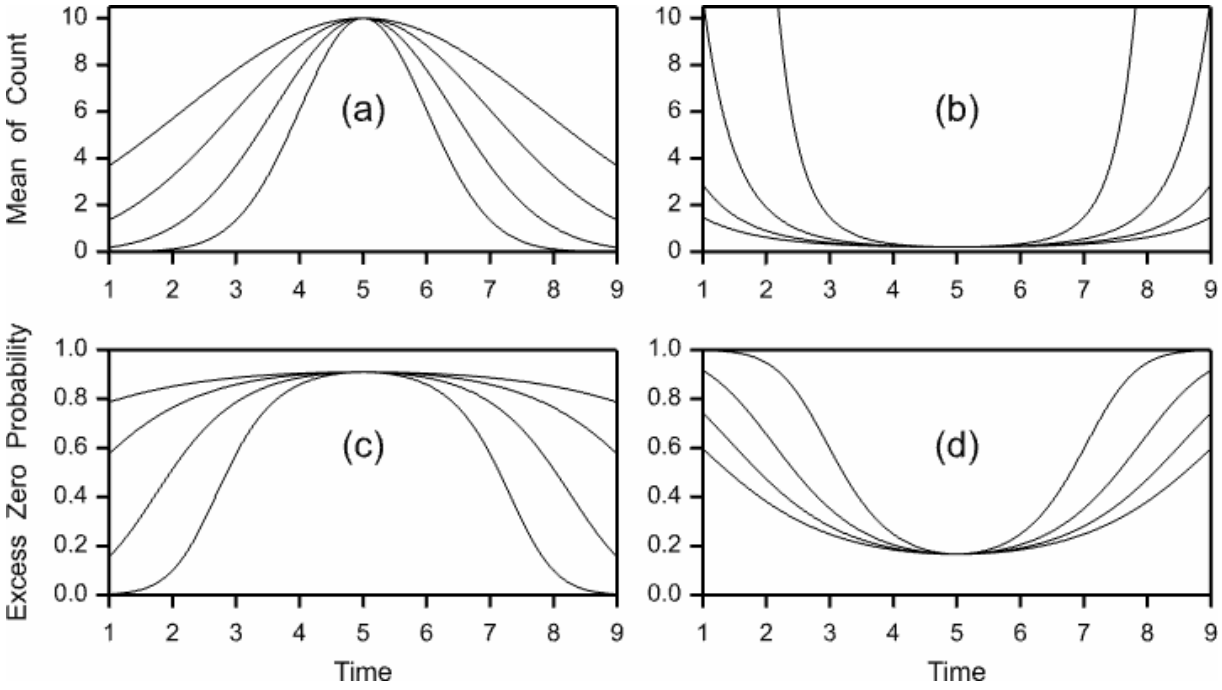
The same repeated measure model is used for the excess zero probability:

$$\mathrm{logit}(\delta_{ijt}) = b_j + f_i(t) + w_{ijt} \quad (14)$$

Again the time effect $f_i(t)$ is a polynomial in time up to order 2 and the vector $\boldsymbol{w}_{ij}$ follows a multivariate normal distribution. It is assumed that b_j and w_{ijt} are independent, and also independent of a_j and v_{ijt}.

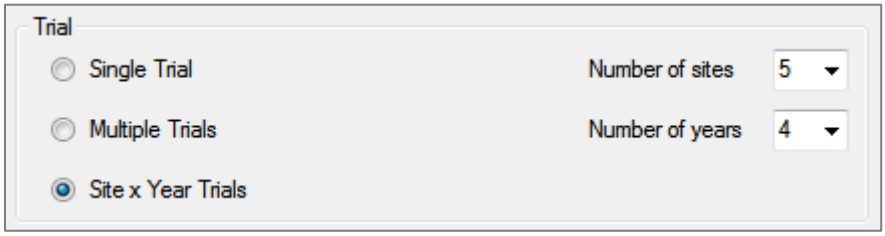
Examples of the second order polynomial on the original count and excess zero probability scale for positive and negative values of the tolerance are given in Figure 6-1. Note that the β parameters which describe the time effect must be specified for each variety in the experiment. For the reference varieties each β parameter is drawn from a normal distribution with specified mean and variance. Also note that it is not possible to have a linear effect in time for one variety and a quadratic time effect for another variety.

Figure 6-1: Examples of second order polynomials on the original count scale (a) and (b), and on the original excess probability scale (c) and (d). The parameters used are for (a) and (c): β_{opt}=5, β_{max}=log(10), and β_{tol}=1, 2, 4, 8; for (b) and (d): β_{opt}=5, β_{max}=log(0.2), and β_{tol}=-1, -2, -3, -4.



7 Simulation model for multiple trials

A single simulated dataset can also consists of multiple trials which can be used to study the design and analysis of multiple field trials across environments. Two variants of multiple trials are implemented: one in which there is no further structure across trials and one in which trials follow a Site x Year structure. In the latter case it is assumed that the same number of experiments are conducted (one in each year) at a limited number of sites. This can be set in the *Trial* groupbox. Note that when *Multiple Trials* is selected, the number of trials, rather than the number of sites and years, must be specified.



When *Multiple Trials* is selected, there is one extra level of variation. In addition to random block effects, there might be random trial effects such that trials will vary in their level of response without affecting differences between varieties. Also when Site x Year Trials is selected there might be random Site and random Year effects as well as a random Site.Year interaction effect. These random effects are then added to the other effects on the transformed (log or logit) scale. The variances of these additional random effects can be specified in the *Variance parameters* groupbox:

Variances on Transformed scale

Response	Block	Trial
Counts	0	0

or

Variances on Transformed scale

Response	Block	Site	Year	SiteYear
Counts	0	0	0	0

In addition to additive multiple trial effects on the transformed scale, variety effects might be different from trial to trial, from site by site or from year to year. This might be termed genotype by environment interaction. This is implemented by additional random effects which operate on the variety effects. When simulating Multiple Trials for instance, in each separate trial the variety effect is drawn from a normal distribution with some mean and variance. So with the values given on the right, the GMO effects across trials have a mean of 2 and a variance of 0.5 and similarly for the comparator effect. For the reference varieties there are two stages of simulation. In the first stage a reference mean, say *M*, for the trial is drawn from a normal distribution with mean 1 and variance 0.3. In the second stage the effects of the

Means on Log-scale

Variety	Level
GMO	2
Comparator	1.5
REF Mean	1
REF Variance	0.5

Variance Trials

Level
0.5
0.4
0.3

Means on Log-scale

Variety	Level
GMO	2
Comparator	1.5
REF Mean	1
REF Variance	0.5

Variance Sites

Level
0.5
0.4
0.3

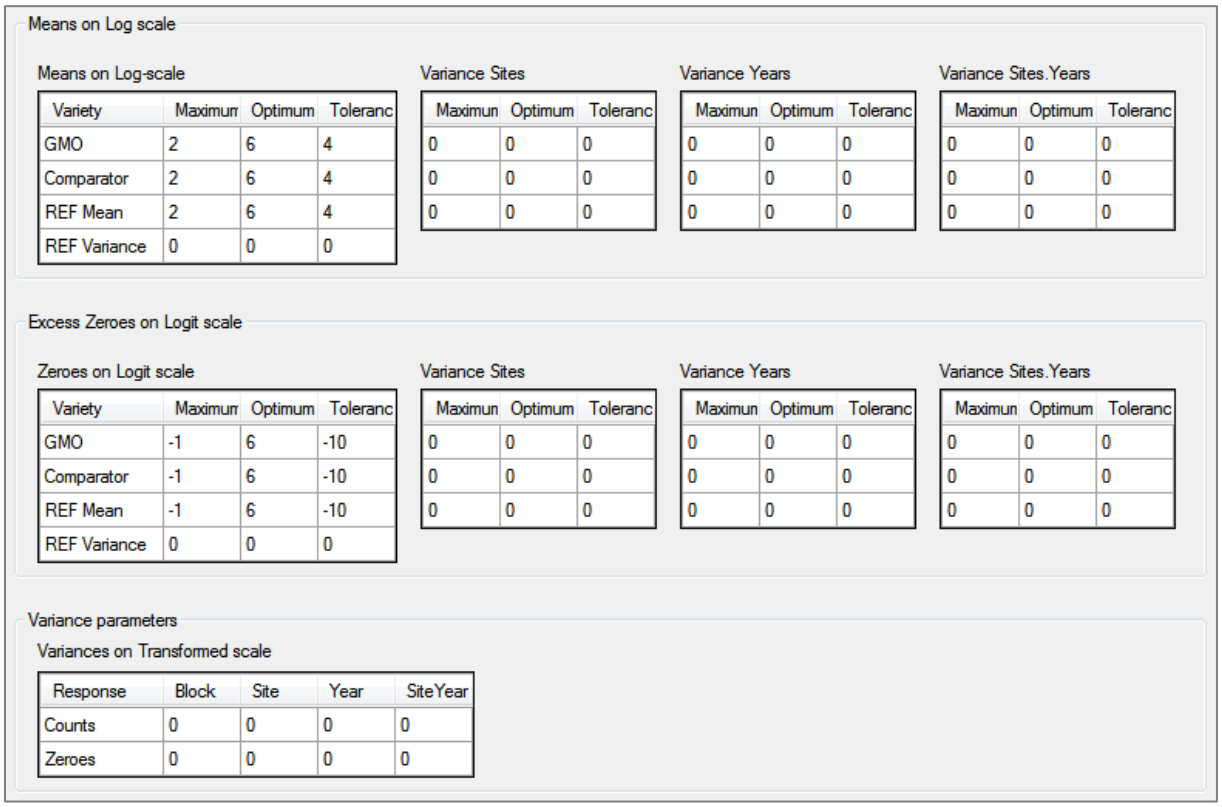
Variance Years

Level
0.1
0.2
0.3

Variance Sites.Years

Level
0
0
0

reference varieties in that specific trial are simulated from a normal distribution with mean M and variance 0.5. Similarly for Site x Year Trials three variance components can be distinguished for every variety effect as depicted on the right. The examples given here relate to a single measurement experiment without excess zeroes in which only the mean of the variety effect must be specified. In case of repeated measurements with say a quadratic time effect, all the time effect β parameters have their associated variance components and many parameters need to be specified. Moreover the same statistical model can be specified for the excess zero probability. So for a Site x Year Trial with some reference varieties and repeated counts with a quadratic time effect along with excess zeroes with a quadratic time effect the following values must be specified.



Variety	Maximum	Optimum	Toleranc
GMO	2	6	4
Comparator	2	6	4
REF Mean	2	6	4
REF Variance	0	0	0

Maximun	Optimum	Toleranc
0	0	0
0	0	0
0	0	0

Maximun	Optimum	Toleranc
0	0	0
0	0	0
0	0	0

Maximun	Optimum	Toleranc
0	0	0
0	0	0
0	0	0

Variety	Maximum	Optimum	Toleranc
GMO	-1	6	-10
Comparator	-1	6	-10
REF Mean	-1	6	-10
REF Variance	0	0	0

Maximun	Optimum	Toleranc
0	0	0
0	0	0
0	0	0

Maximun	Optimum	Toleranc
0	0	0
0	0	0
0	0	0

Maximun	Optimum	Toleranc
0	0	0
0	0	0
0	0	0

Response	Block	Site	Year	SiteYear
Counts	0	0	0	0
Zeroes	0	0	0	0

8 Examples of a simple simulation

The simulation tool was used to perform two preliminary simulation studies to assess properties of statistical difference and equivalence testing in simple experiments. To this end single trials with single measurements were simulated without excess zeroes. A completely randomized experiment was employed and in addition to the GMO and its comparator one additional variety was included. This setup, without any random effects, only requires specification of three variety means on the log scale and specification of the count distribution. Since the power of a difference test or the properties of an equivalence test depends heavily on the level of replication, different levels of replication were included.

8.1 Power of difference test for the negative binomial distribution

In the first simulation study the power of a likelihood ratio test for the difference between the GMO and the comparator was studied. Data were simulated according to the negative binomial distribution without excess zeroes. The mean of the comparator and the additional variety were assumed to be equal to say μ. Denote the mean of the GMO with λ and the number of replications with *N*. Data were simulated for all 864 combinations of the following values:

- μ 1, 2, 5, 10, 20, 40
- $\lambda = \delta\mu$ with δ = 1, 1.2, 1.4, 1.6, 1.8 , 2.0
- N 4, 6, 8, 10, 15, 20, 30, 40
- ω 0.25, 0.5, 1

For each combination 1000 datasets were simulated. The negative binomial distribution was fitted to each dataset, first under the restriction that the mean of the GMO equals the mean of the comparator and secondly without this restriction. A likelihood ratio test statistic is then given by twice the difference between the log-likelihoods of the two models. The large sample distribution of this test statistic is χ_1^2 and this distribution was used to calculate P values. The (simulated) power of the difference test is then given by the fraction of the 1000 datasets for which the null hypothesis of no difference is rejected. A significance level of $\alpha = 0.05$ was used. The simulated negative binomial data were also analysed with the over-dispersed Poisson and with the lognormal distribution. For the over-dispersed Poisson distribution the quasi likelihood approach was used to fit the model and the likelihood ratio test was replaced by a scaled deviance test whenever the residual deviance of the full model (with separate parameters for the three varieties) was larger than one. An analysis employing the lognormal distribution simply involves taking the logarithm of the data and then analysing according to the normal distribution. Whenever a dataset has a zero observation, 0.5 was added to all observations before taking logs. The resulting power curves for the negative binomial model, the Poisson model and the lognormal model are given in Appendix 1 to Appendix 3. Appendix 4 contains a direct comparison between the power for the different models for *N*=10 and *N*=40. As expected the power is larger when there is less over-dispersion (smaller values of ω) and when the mean of the distribution is large. The number of replications required to detect a quotient $\delta = 2$ between the mean of the GMO and the comparator with probability 0.80 using a two-sided test at $\alpha = 0.05$ is given in Table 8-1 for

the various values of μ and ω. These values are interpolated from the values of *N* which are used in the simulation.

Table 8-1: Number of replications needed to obtain a significant result with probability 80% using a two-sided test at $\alpha = 0.05$ when the quotient of the mean of the GMO and the comparator equals $\delta = 2$ for data which have a negative binomial distribution with mean μ for the comparator and dispersion parameter ω. The number of replications is given for the negative binomial, over-dispersed Poisson and lognormal model.

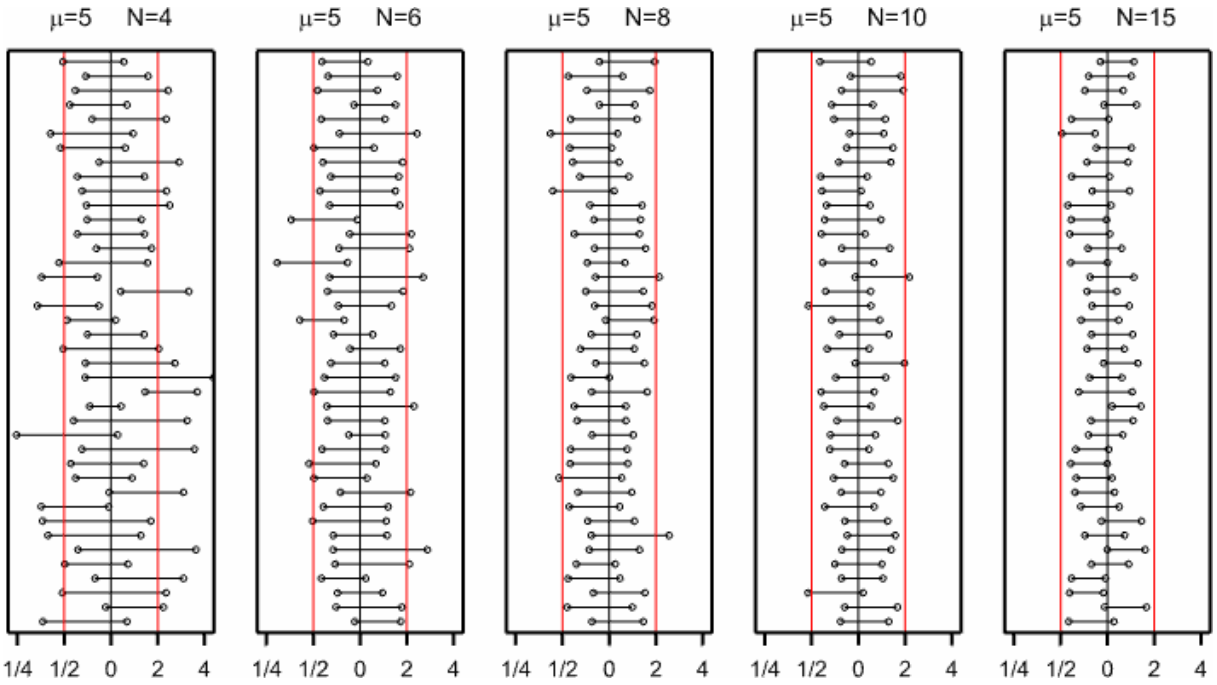
NegBinomial	$\mu = 1$	$\mu = 2$	$\mu = 5$	$\mu = 10$	$\mu = 20$	$\mu = 40$
$\omega = 0.25$	29	21	13	10	9	9
$\omega = 0.50$	> 40	27	21	19	17	16
$\omega = 1.00$	> 40	> 40	37	35	33	32
overPoisson	$\mu = 1$	$\mu = 2$	$\mu = 5$	$\mu = 10$	$\mu = 20$	$\mu = 40$
$\omega = 0.25$	32	22	13	10	10	9
$\omega = 0.50$	> 40	27	21	18	16	16
$\omega = 1.00$	> 40	39	32	32	28	27
logNormal	$\mu = 1$	$\mu = 2$	$\mu = 5$	$\mu = 10$	$\mu = 20$	$\mu = 40$
$\omega = 0.25$	36	26	17	13	12	11
$\omega = 0.50$	> 40	37	30	27	24	23
$\omega = 1.00$	> 40	> 40	> 40	> 40	> 40	> 40

Surprisingly, an analysis under the wrong over-dispersed Poisson model requires slightly less replication for $\omega = 1$ as compared to an analysis with the correct negative binomial distribution.

8.2 Properties of equivalence test for the Poisson distribution

In the second simulation study the properties of the TOST approach to equivalence testing was assessed for count data which were simulated according to the Poisson distribution. The TOST, or two one-sided tests, approach uses a two-sided confidence interval for the difference between the GMO and the comparator (Schuirmann, 1987). When the confidence interval completely lies in the interval determined by fixed lower and upper equivalence limits, then the null hypothesis of non-equivalence is be rejected in favour of equivalence. The same simulation setting as in section 8.1 was used, however since the Poisson distribution was used to simulate data there is no over-dispersion. Hypothetical equivalence limits of ½ and 2 were used to perform equivalence testing. A 95% likelihood ratio confidence interval for the ratio of the GMO mean and the comparator mean was calculated for each simulated dataset. The number of times this interval lies within the equivalence interval (½, 2) can then be counted. As an example the confidence interval for 40 simulated datasets is given in Figure 8-1 with $\mu = 5$ for both the GMO and the comparator, and for various values of the number of replications *N*. In this case the GMO and comparator have equal means and are thus theoretically equivalent. However for small numbers of replications the confidence intervals frequently crosses the equivalence limits implying that the null hypothesis of non-equivalence is not always rejected.

Figure 8-1: 95% likelihood ratio confidence intervals for the ratio of the mean of the GMO and the mean of the comparator when the underlying mean of both is $\mu = 5$, and various number of replication *N*. The red lines denote the artificial equivalence limits set at factor 2 Limits of Concern.



Appendix 5 displays the power of the likelihood ratio difference tests as a function of the number of replication while Appendix 6 gives the probability to reject the null hypothesis of non-equivalence.

9 References

Al-Deeb MA, Wilde GE & Zhu KY (2001). Effect of Insecticides Used in Corn, Sorghum, and Alfalfa on the Predator *Orius insidiosus* (Hemiptera: Anthocoridae). *Journal of Economic Entomology*, 94(6): 1353-1360.

Al-Deeb MA & Wilde GE (2003). Effect of Bt Corn Expressing the Cry3Bb1 Toxin for Corn Rootworm Control on Aboveground Nontarget Arthropods. *Environmental Entomology*, 32(5): 1164-1170.

Bourguet D, Chaufaux J, Micoud A, Delos M, Naibo B, Bombarde F, Marque G, Eychenne N & Pagliari C (2002). *Ostrinia nubilalis* parasitism and the field abundance of non-target insects in transgenic *Bacillus thuringiensis* corn (Zea mays). *Environmental Biosafety Research*, 1(1): 49‑60.

Buckelew LD, Pedigo LP, Meroc HM, Owen MDK & Tylkad GL (2000). Effects of Weed Management Systems on Canopy Insects in Herbicide-Resistant Soybeans. *Journal of Economic Entomology*, 93(5): 1437-1443.

Candolfi M, Brown K, Grimm C, Reber B & Schmidli H (2004). A faunistic approach to assess potential side-effects of genetically modified Bt-Corn on non-target arthropods under field conditions. *Biocontrol Science and Technology*, 14(2): 129-170

Cheung YB (2006). Growth and cognitive function of Indonesian children: Zero-inflated proportion models. *Statistics in Medicine*, 25(17): 3011-3022.

Cunningham RB & Lindenmayer DB (2005). Modeling count data of rare species: some statistical issues. *Ecology*, 86(5): 1135-1142.

de la Poza M, Pons X, Farinós GP, López C, Ortego F, Eizaguirre M, Castañera P & Albajes R (2005). Impact of farm-scale Bt-maize on abundance of predatory arthropods in Spain. *Crop Protection*, 24(7): 677-684.

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

EFSA (2010a). EFSA Panel on Genetically Modified Organisms (GMO). Statistical considerations for the safety evaluation of GMOs. *EFSA Journal*, 8(1), 1250. [59 pp.], doi:10.2903/j.efsa.2010.1250

EFSA (2010b). EFSA Panel on Genetically Modified Organisms (GMO). Guidance on the environmental risk assessment of genetically modified plants. *EFSA Journal*, 8(11): 1879. [111 pp.], doi:10.2903/j.efsa.2010.1879.

Feller W (1943). On a general class of “contagious” distributions. *Annals of Mathematical Statistics*, 14(4): 389-400.

Greene W (1994). Accounting for excess zeros and sample selection in Poisson and negative binomial regression models. Working Paper EC-94-10, Department of Economics, New York University.

Hall DB (2000). Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, 56(4): 1030-1039.

Jasinski JR, Eisley JB, Young CE, Kovach J & Wilson H (2003). Select nontarget arthropod abundance in transgenic and nontransgenic field crops in Ohio. *Environmental Entomology*, 32(2): 407-413.

Johnson MT (1997). Interaction of Resistant Plants and Wasp Parasitoids of Tobacco Budworm (Lepidoptera: Noctuidae). *Environmental Entomology*, 26(2): 207-214.

Johnson MT & Gould F (1992). Interaction of Genetically Engineered Host Plant Resistance and Natural Enemies of *Heliothis virescens* (Lepidoptera: Noctuidae) in Tobacco. *Environmental Entomology*, 21(3): 587-597

Kos M (2012). Multitrophic effects of plant resistance: from basic ecology to application in transgenic crops. PhD Thesis, Wageningen University.

Lambert D (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1): 1-14.

Lewin W-C, Freyhof J, Huckstorf V, Mehner T & Wolter C (2010). When no catches matter: Coping with zeros in environmental assessments. *Ecological Indicators*, 10(3): 572-583.

Manachini B, & Lozzia GC (2002). First investigations into the effects of Bt corn crop on Nematofauna. *Bollettino di Zoologia Agraria e Bachicoltura Serie II*, 34: 85‑96.

Manachini B, Landi S, Fiore MC, Festa M & Arpaia S (2004). First investigations on the effects of Bt-transgenic *Brassica napus* L. on the trophic structure of the nematofauna. *IOBC/WPRS Bullettin*, 27: 103‑108.

Marvier M, McCreedy C, Regetz J & Kareiva P (2007). A meta-analysis of effects of Bt cotton and maize on nontarget invertebrates. *Science*, 316: 1475-1477.

Mascarenhas VJ & Luttrell RG (1997). Combined Effect of Sublethal Exposure to Cotton Expressing the Endotoxin Protein of *Bacillus thuringiensis* and Natural Enemies on Survival of Bollworm (Lepidoptera: Noctuidae) Larvae. *Environmental Entomology*, 26(4): 939-945

McCullagh P & Nelder JA (1989). *Generalized Linear Models*. Chapman and Hall. London.

Men XY, Ge F, Liu XH & Yardim EN (2003). Diversity of Arthropod Communities in Transgenic Bt Cotton and Nontransgenic Cotton Agroecosystems. *Environmental Entomology*, 32(2): 270-275.

Musser FR & Shelton AM (2003). Bt Sweet Corn and Selective Insecticides: Impacts on Pests and Predators. *Journal of Economic Entomology*, 96(1): 71-80.

Naranjo SE (2005a). Long-term assessment of the effects of transgenic Bt cotton on the abundance of nontarget arthropod natural enemies. *Environmental Entomology*, 34(5): 1193-1210.

Orr DB & Landis DA (1997). Oviposition of European Corn Borer (Lepidoptera: Pyralidae) and Impact of Natural Enemy Populations in Transgenic Versus Isogenic Corn. *Journal of Economic Entomology*, 90(4): 905-909.

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.

Pilcher CP, Obrycki JJ, Rice ME & Lewis LC (1997). Preimaginal development, survival and field abundance of insect predators on transgenic *Bacillus thuringiensis* corn. *Environmental Entomology*, 26(2): 446-454.

R Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, www.R-project.org.

Reed GL, Jensen AS, Riebe J, Head G & Duan JJ (2001). Transgenic Bt potato and conventional insecticides for Colorado potato beetle management: comparative efficacy and non-target impacts. *Entomologia Experimentalis et Applicata*, 100(1): 89–100

Riddick EW, Dively G & Barbosa P (1998). Effect of a seed-mix deployment of Cry3A-transgenic and nontransgenic potato on the abundance of *Lebia grandis* (Coleoptera: Carabidae) and *Coleomegilla maculata* (Coleoptera: Coccinellidae). *Annals of the Entomological Society of America*, 91(5): 647-653.

Schuirmann DJ (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6): 657-680.

Sileshi G (2008). The excess-zero problem in soil animal count data and the choice of appropriate models for statistical inference. *Pedobiologica*, 52(1): 1-17.

Van der Voet H, Perry JN, Amzal B & Paoletti C (2011). A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. *BMC Biotechnology*, 11:15. [20 pp.].

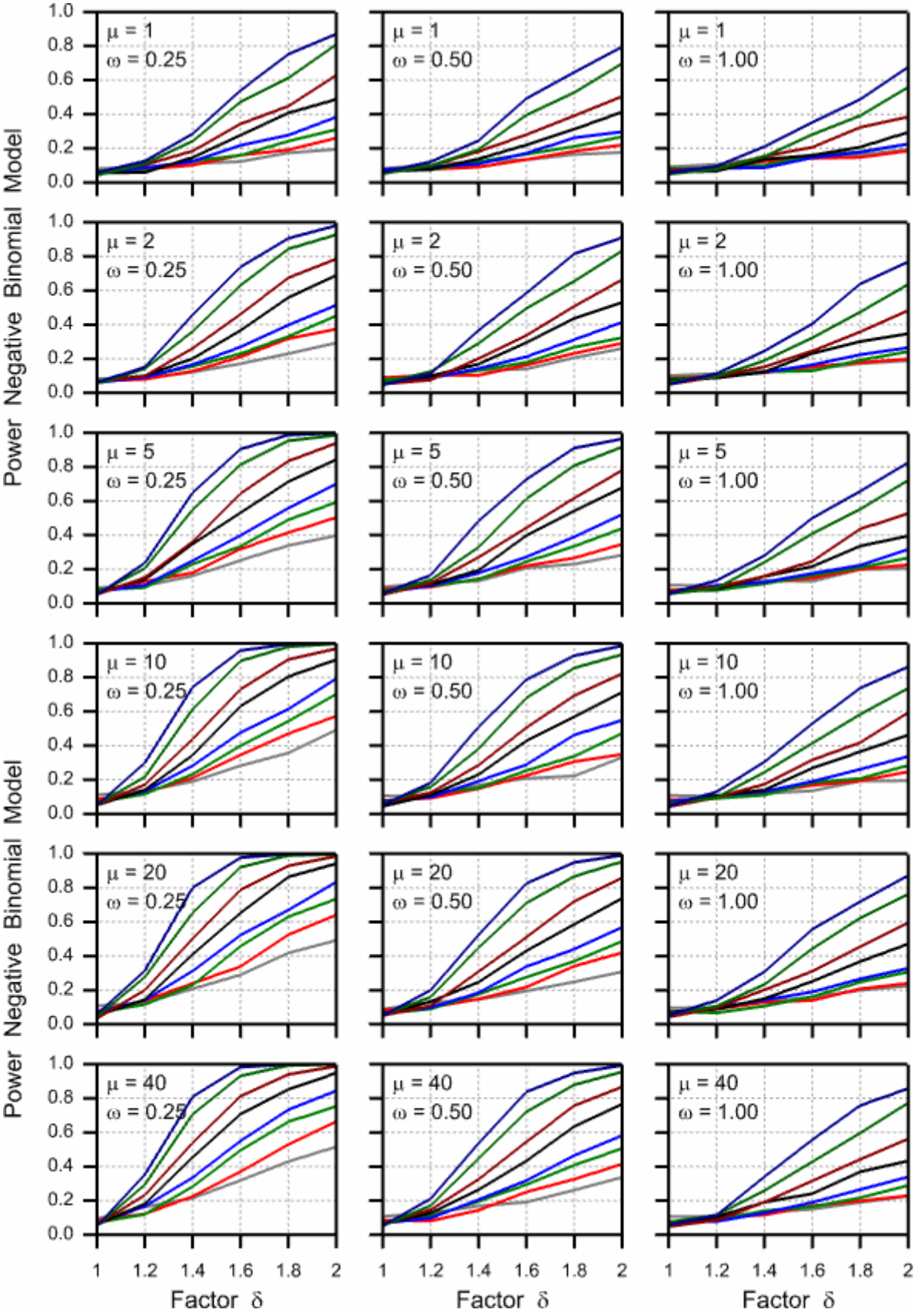
Vieira AMC, Hinde JP & Demetrio CGB (2000). Zero-inflated proportion data models applied to a biological control assay, *Journal of Applied Statistics*, 27(3): 373-389.

Wade French B, Chandler LD, Ellsbury MM, Fuller BW & West M (2004). Ground Beetle (Coleoptera: Carabidae) Assemblages in a Transgenic Corn–Soybean Cropping System. *Environmental Entomology*, 33(3): 554-563.

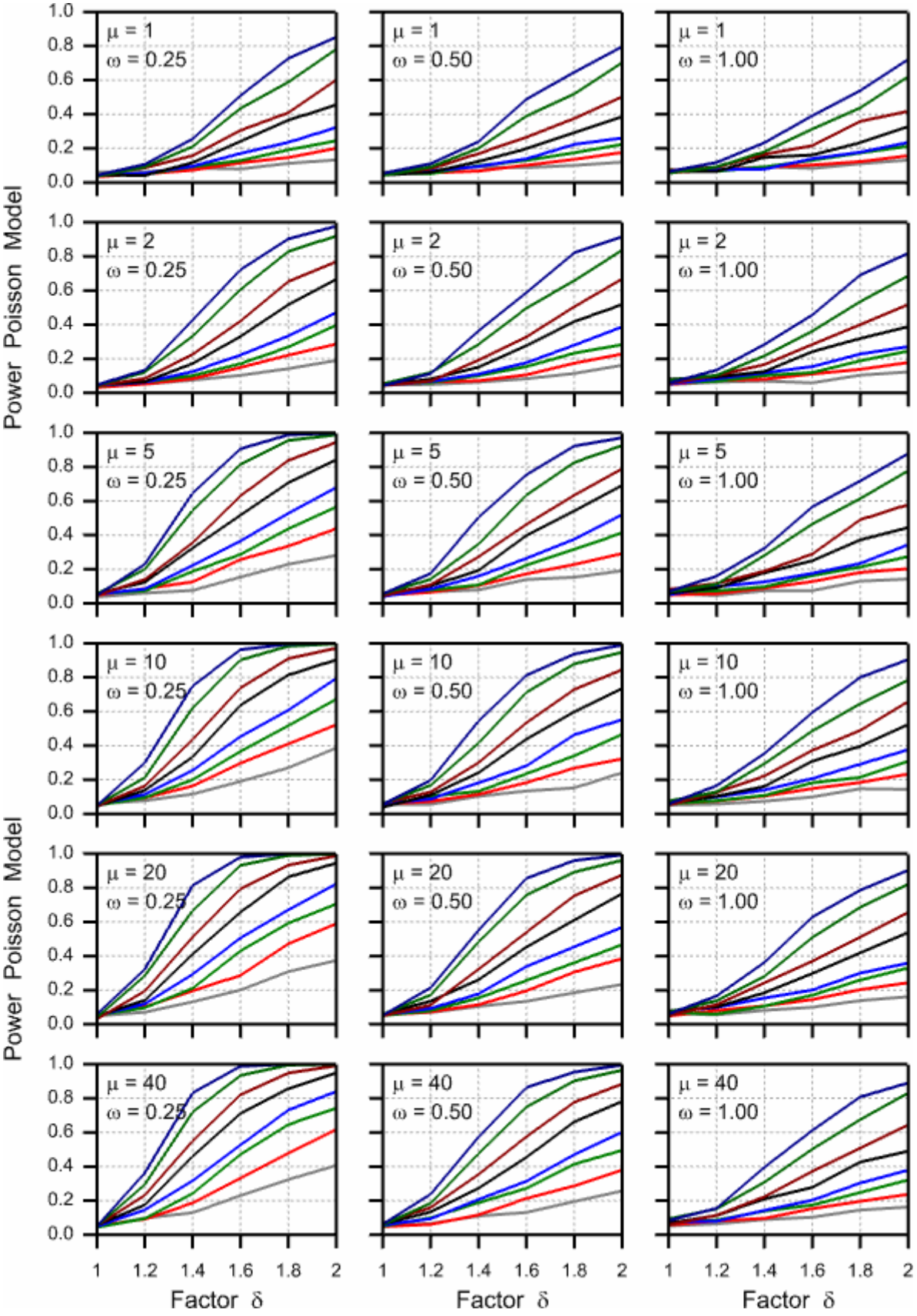
Wei-Di L, Wu KM, Chen XX, Feng HQ, Xu G & Guo YY (2004). Effects of transgenic cotton carrying *Cry1A+CpTI* and *Cry1Ac* genes on diversity of arthropod communities in cotton fields in North China. *Chinese Journal of Agricultural Biotechnology*, 1(1): 17-21.

Wu KM & Guo YY (2003). Influences of *Bacillus thuringiensis* Berliner Cotton Planting on Population Dynamics of the Cotton Aphid, *Aphis gossypii* Glover, in Northern China. *Environmental Entomology*, 32(2): 312-318.

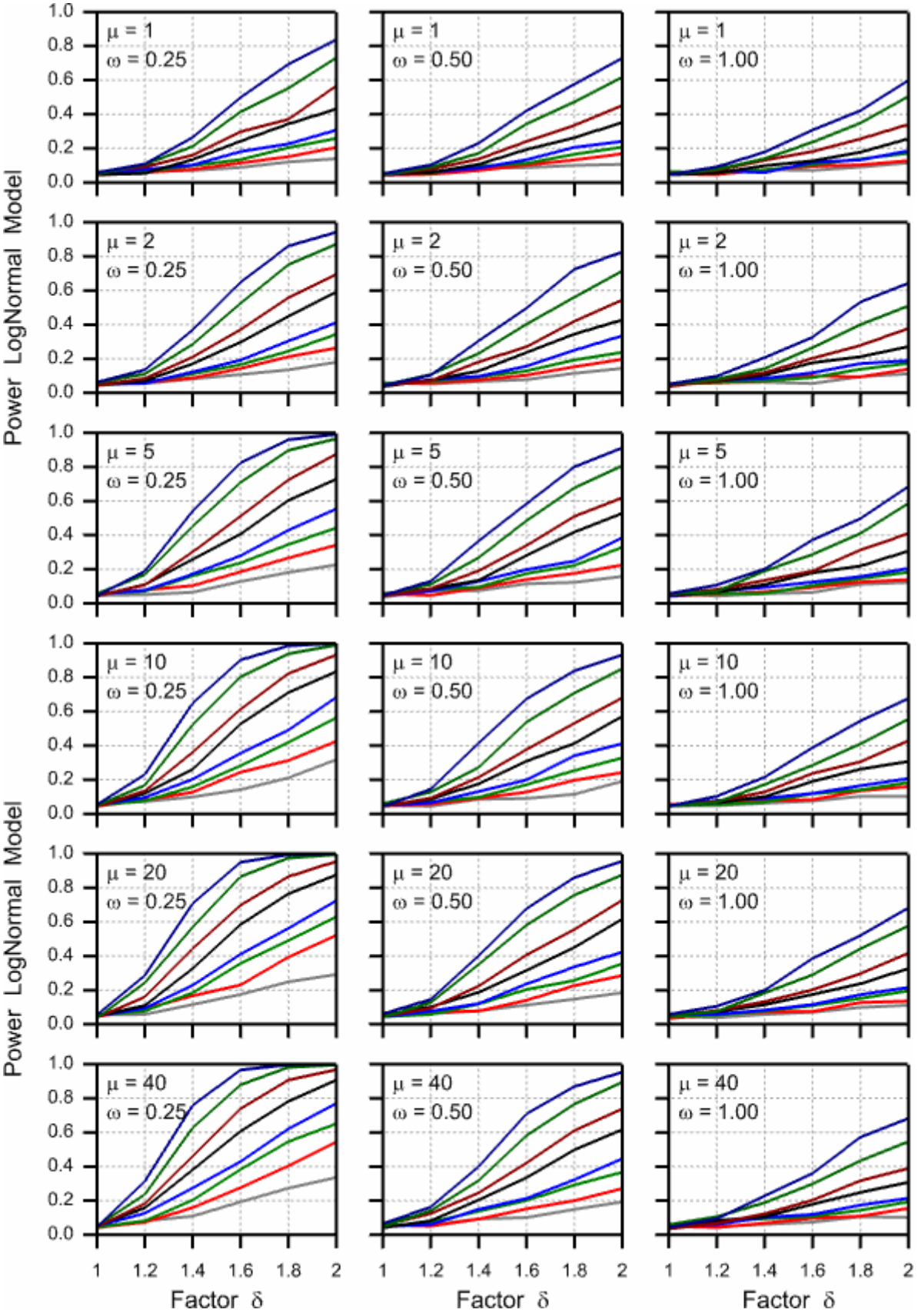
Appendix 1: Power of a difference test for the negative binomial distribution when the simulated data are analysed with the negative binomial distribution for replication levels *N*=4, 6, 8, 10, 15, 20, 30, 40 (bottom to top curves).



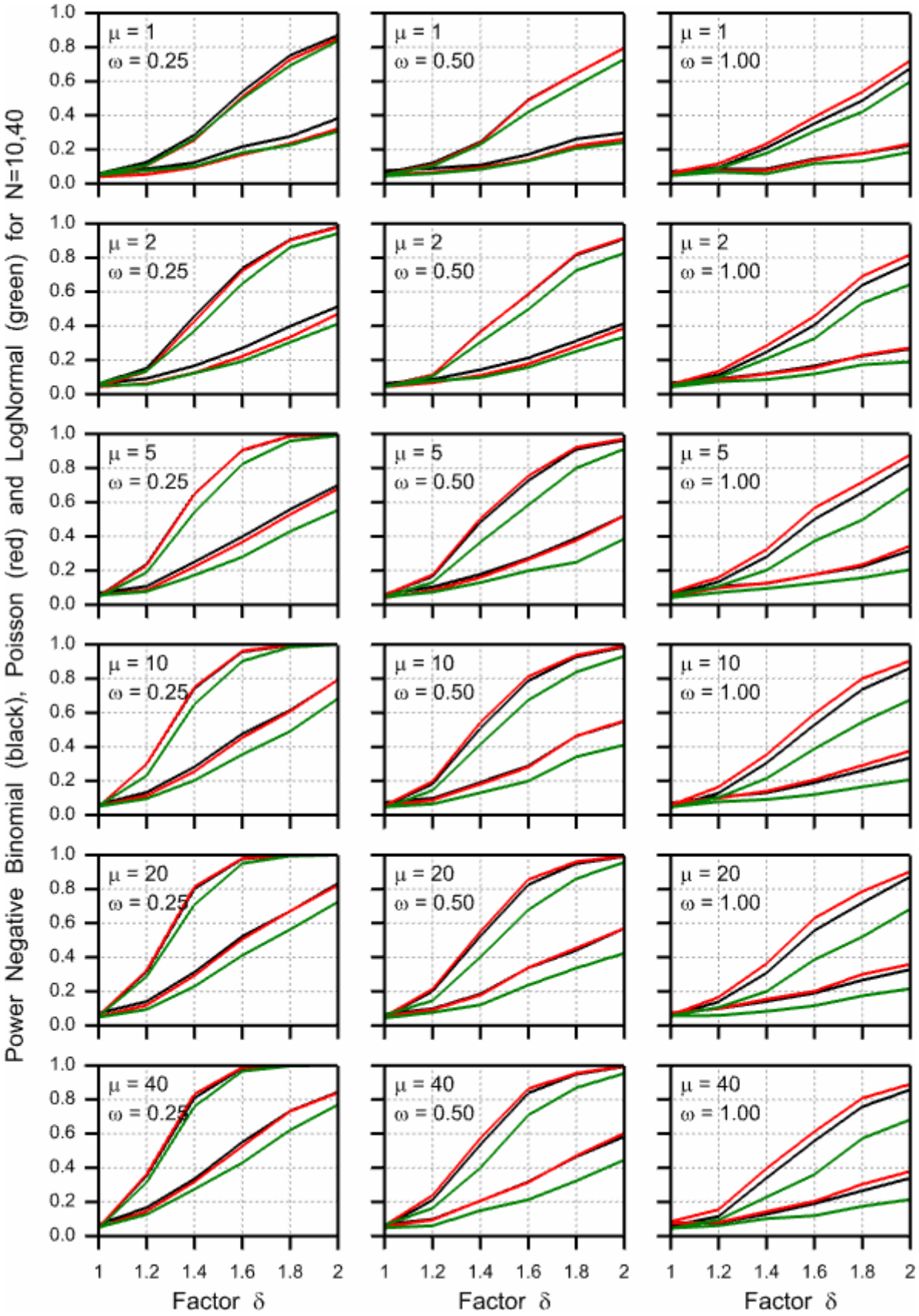
Appendix 2: Power of a difference test for the negative binomial distribution when the simulated data are analysed with the Poisson distribution for replication levels *N*=4, 6, 8, 10, 15, 20, 30, 40 (bottom to top curves).



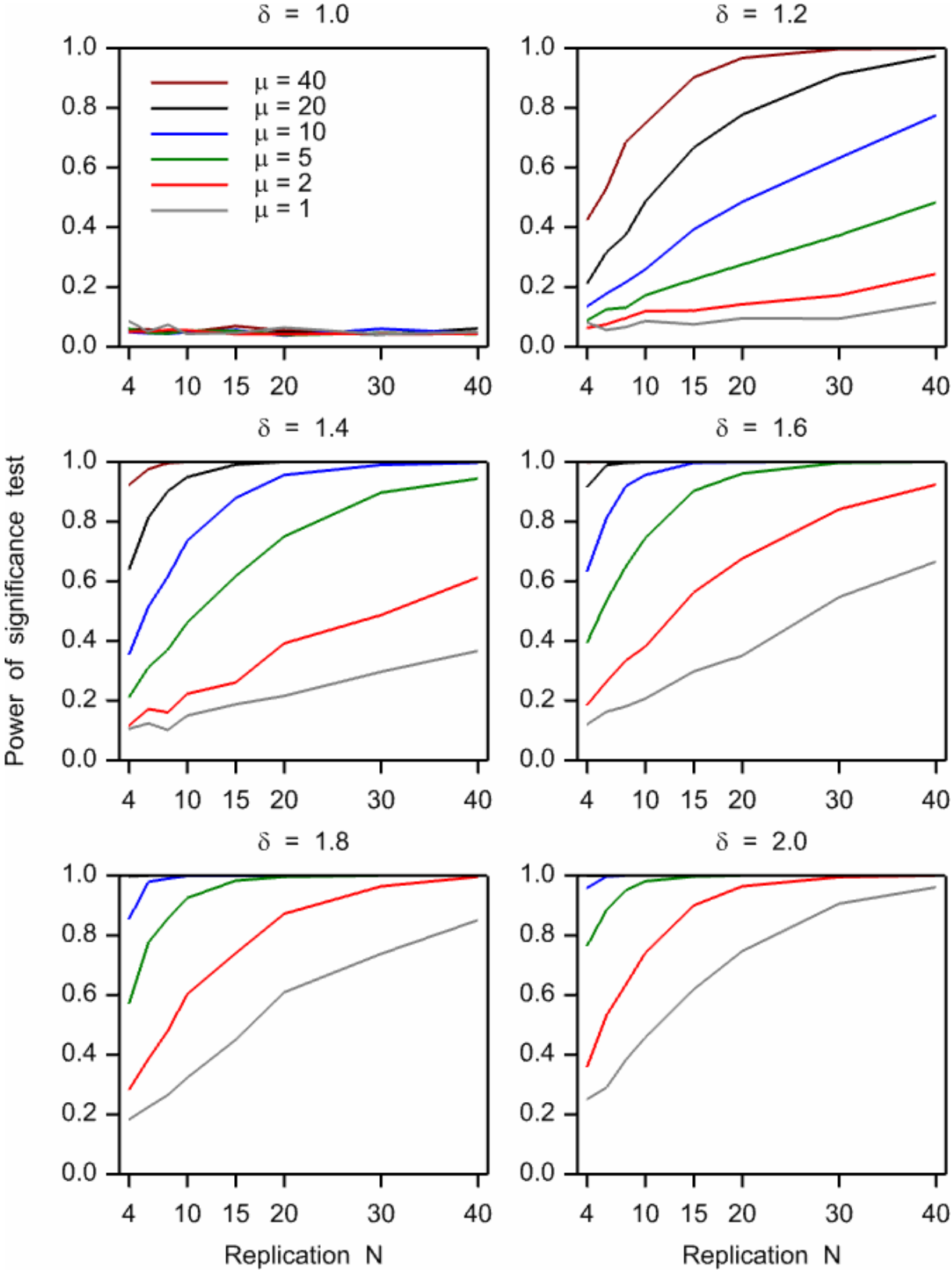
Appendix 3: Power of a difference test for the negative binomial distribution when the simulated data are analysed with the LogNormal distribution for replication levels *N*=4, 6, 8, 10, 15, 20, 30, 40 (bottom to top curves).



Appendix 4: Power of a difference test for the negative binomial distribution when the simulated data are analysed with the Negative Binomial (black), Poisson (red) or LogNormal (green) distribution for replication levels *N*=10 (bottom curves) and 40 (top curves).



Appendix 5: Power of a difference test for the Poisson distribution when the simulated data are also analysed with the Poisson distribution for means μ for the comparator and means $\delta\mu$ for the GMO for replication levels N=4, 6, 8, 10, 15, 20, 30, 40.



Appendix 6: Probability of rejecting the null hypothesis of non-equivalence with respect to a factor 2 Limit of Concern for the Poisson distribution when the simulated data are also analysed with the Poisson distribution for means μ for the comparator and means $\delta\mu$ for the GMO for replication levels N=4, 6, 8, 10, 15, 20, 30, 40.