Sveriges lantbruksuniversitet
Swedish University of Agricultural Sciences
Biostokastikum

Overdispersion

Workshop in generalized linear models
Uppsala, June 11-12, 2014

Johannes Forkman, Field Research Unit, SLU

”Overdispersion is not uncommon in practice. In fact, some would maintain that overdispersion is the norm in practice and nominal dispersion the exception”

McCullagh and Nelder (1989)

Outline

What is overdispersion and how do we detect it?

An overview of methods for overdispersed data

- Generalized linear mixed models
- Generalized estimating equations
- Adjustment using an overdispersion factor
- Negative binomial distribution
- Mixture distributions for zero-inflated data

Overdispersion

In Poisson and binomially distributed data, the variance is a known function of the mean:

Proportions: $V(y_i) = \mu_i(1 - \mu_i)/n_i$

Counts: $V(y_i) = \mu_i$

In practice, the variance is often much larger. This is called *overdispersion*.

Measures of goodness of fit

Deviance (D): Twice the difference between the log likelihood of a model with a perfect fit and the log likelihood of the fitted model

Pearson's chi-sqare:

$$\sum_{i=1}^{N} \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}$$

that is, the sum of all squared Pearson residuals

How to detect overdispersion

No overdispersion	Overdispersion
Deviance ≈ df	Deviance ≫ df
Pearson χ^2 ≈ df	Pearson χ^2 ≫ df

df = residual degrees of freedom

Example Leaves

20 plants

2 treatments: Active (10 plants)
Control (10 plants)

Approx. 20 leaves per plant

Plant	Treatment	Infested	N		Plant	Treatment	Infested	N
1	Active	6	20		11	Control	5	20
2	Active	3	20		12	Control	9	19
3	Active	7	20		13	Control	14	20
4	Active	1	20		14	Control	3	20
5	Active	0	18		15	Control	20	20
6	Active	0	20		16	Control	8	20
7	Active	4	18		17	Control	7	15
8	Active	9	20		18	Control	7	20
9	Active	10	20		19	Control	8	20
10	Active	2	20		20	Control	5	20

The data was analyzed using a generalized linear model with a **binomial distribution** and a **logit link**.

The probability that a leaf was infested was estimated to 0.21 and 0.44, for Active and Control, respectively. This difference was significant (**P < 0.0001**).

The data was analyzed using a gene
linear model with a **binomial distrib**
a **logit link**.

This analysis was incorrect!

The probability that a leaf was infested was estimated to 0.21 and 0.44, for Active and Control, respectively. This difference was significant (**P < 0.0001**).

Why incorrect? Well, because…

Residual degrees of freedom: 18

Deviance: 94.53 (P < 0.0001)
Pearson chi-square: 79.57 (P < 0.0001)

The observations are clearly overdispersed!

How could the data be overdispersed?

Each binomial observation is a cluster of approx. 20 Bernoulli (Yes/No) observations.

Bernoulli observations from the same plant might be correlated.

This correlation is the source of overdispersion

Note that we get exactly the same results if we analyze the following summary table:

Treatment	Infested	N
Active	42	196
Control	86	194

These analyses do not consider variation between plants (i.e. correlation within plants).

Generalized Linear Mixed Models (GLMM)

GLMM

We have fitted the model

$$\text{logit}(\mu) = \beta_0 + \beta_i \qquad i = 1,2$$

But since the data is overdispersed and clustered, a generalized linear mixed model (GLMM):

$$\text{logit}(\mu) = \beta_0 + \beta_i + b_{ij} \qquad \begin{matrix} i = 1,2 \\ j = 1, 2, \dots, 10 \end{matrix}$$

$$b_i \sim N\left(0, \sigma_b^2\right)$$

is more appropriate

In SAS, the glimmix procedure gives

Type III Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Treatment	1	14.35	5.60	**0.0325**

In R, the glmer function can be used, and the likelihood ratio test is

	Df	AIC	deviance	Chisq	Chi DF	Pr(>Chisq)
Model.Null	2	**65.864**	61.864			
Model.Mixed	3	**62.157**	56.157	5.7071	1	**0.0169***

Are there no problems?

- Complicated GLMM models are hard to fit
- For GLMM, statistical inference is an issue
- We might be interested in the effects on the population, rather than in the effects on the individual subjects

Generalized Estimating Equations (GEE)

GEE – What's specified?

- Link function : e.g. log or logit
- The variance function: $\phi V(\mu)$, where ϕ is a dispersion factor
- Correlation pattern: e.g.
  exchangeable (i.e. compound symmetry),
  autoregressive,
  unstructured

GEE – What's not specified?

- The exact distribution (i.e. the exact likelihood)
- Random effects

This enables

✓ Simple estimating equations (makes it easier to fit the model)

✓ A robust estimator of the standard errors: the so called empirical sandwich estimator

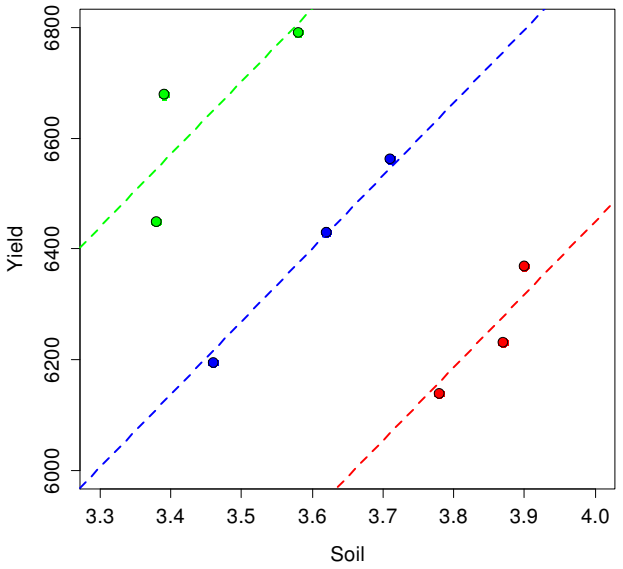
✓ Population-averaged statistical inference

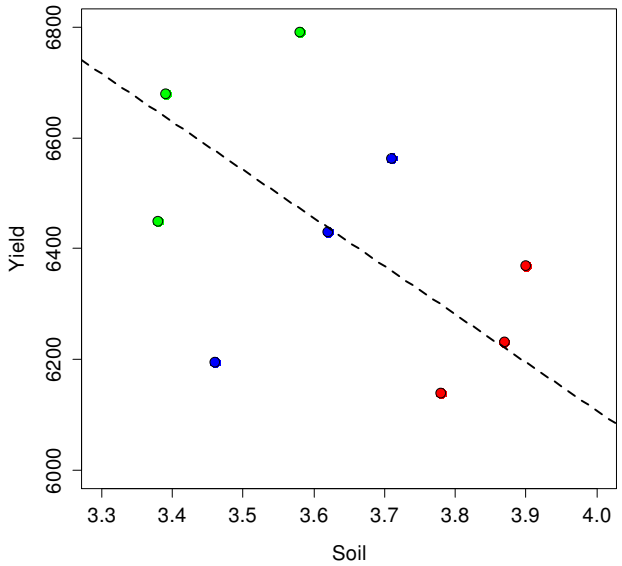
In SAS, using the genmod procedure:

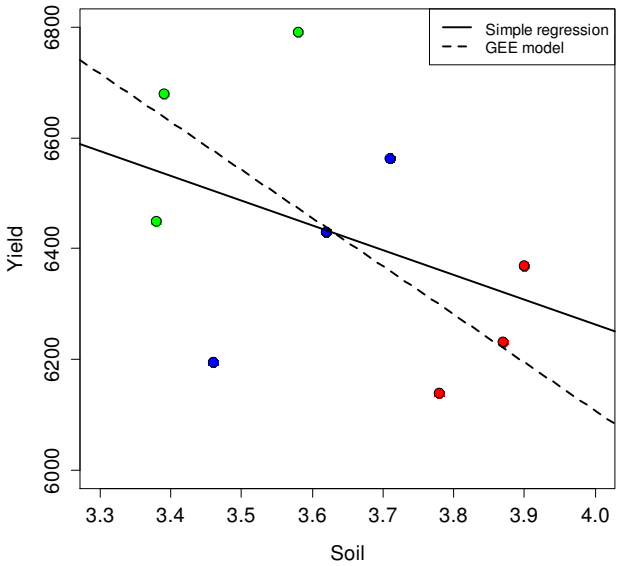
Wald: P = 0.017
Score: P = 0.032

In R, using the ggeglm function:

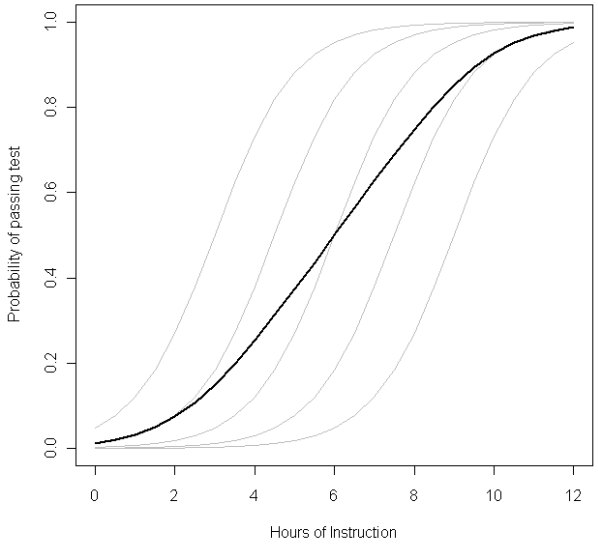
Wald: P = 0.017







GEE is a model for the population average



Black: GEE (population-averaged)

Gray: GLMM (subject-specific)

Source: http://stats.stackexchange.com/questions/32419/difference-between-generalized-linear-models-generalized-linear-mixed-models-i

Adjustment using an overdispersion factor

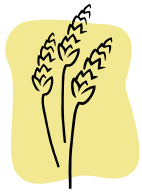
Block	Mix	Count
1	1	24
1	2	12
1	3	8
1	4	13
2	1	9
2	2	9
2	3	9
2	4	18
3	1	12
3	2	8
3	3	44
3	4	0
4	1	8
4	2	12
4	3	25
4	4	0

Example Seeding mixes

Four blocks

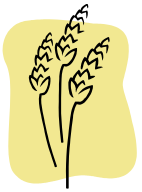
Four seeding mixes

Observed number of plants of a specific species



Residual degrees of freedom: 9

Deviance: 90.23 (P < 0.0001)
Pearson's chi-square: 79.55 (P < 0.0001)



The data is clearly overdispersed.

In this example, we have no clusters!

Adjustment using overdispersion factor

Simply assume that the variance is $\phi V(\mu)$, where ϕ is a dispersion factor

Response	Variance
Proportions	$\phi\mu(1-\mu)/n$
Counts	$\phi\mu$

Estimate ϕ as Pearson's chi-square / df:

$$\hat{\phi} = 79.55/9 = 8.84$$

Within the glm function, specify

```
family = quasipoisson
```

or

```
family = quasibinomial
```

Within the model statement of the genmod procedure, give the option

```
dist = poisson pscale
```

or

```
dist = binomial pscale
```

Analysis of deviance

Without overdispersion factor:

$$\chi^2 = D(Reduced) - D(Complete)$$

With overdispersion factor:

$$F = \frac{D(Reduced) - D(Complete)}{[df(Complete) - df(Reduced)]\hat{\phi}}$$

Results without dispersion factor

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Block	3	4.98	0.1735
Mix	3	30.95	**<.0001**

	Means on the log scale				**Back-transformed means**		
Mix	Mean	SEM	95% conf.		Mean	95% conf.	
1	2.57	0.138	2.3	2.8	13.1	10.0	17.2
2	2.32	0.157	2.0	2.6	10.1	7.5	13.8
3	3.06	0.108	2.8	3.3	21.2	17.2	26.3
4	2.04	0.180	1.7	2.4	7.7	5.4	10.9

Results with dispersion factor

LR Statistics For Type 3 Analysis				
Source	Num DF	Den DF	F Value	Pr > F
Block	3	9	0.19	0.9021
Mix	3	9	1.17	**0.3749**

Means on the log scale **Back-transformed means**

Mix	Mean	SEM	95% conf.		Mean	95% conf.	
1	2.57	0.410	1.77	3.38	13.1	5.9	29.2
2	2.32	0.465	1.40	3.23	10.1	4.1	25.2
3	3.06	0.322	2.43	3.69	21.2	11.3	40.0
4	2.04	0.535	0.99	3.08	7.7	2.7	21.9

SEM are $\sqrt{\hat{\phi}} = 2.97$ times larger than before

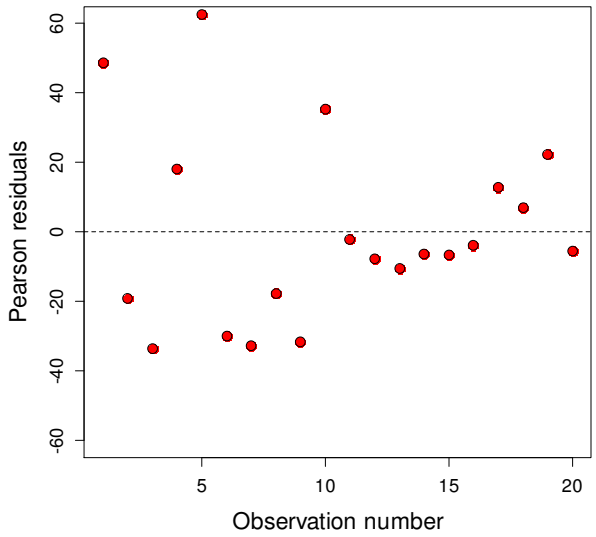
The negative binomial distribution

Example Sheep milk

Somatic cell count in sheep milk using mechanical or manual milking

Method	Count	Method	Count
Mechanical	2966	Manual	186
Mechanical	569	Manual	107
Mechanical	59	Manual	65
Mechanical	1887	Manual	126
Mechanical	3452	Manual	123
Mechanical	189	Manual	164
Mechanical	93	Manual	408
Mechanical	618	Manual	324
Mechanical	130	Manual	548
Mechanical	2493	Manual	139

Pearson residuals



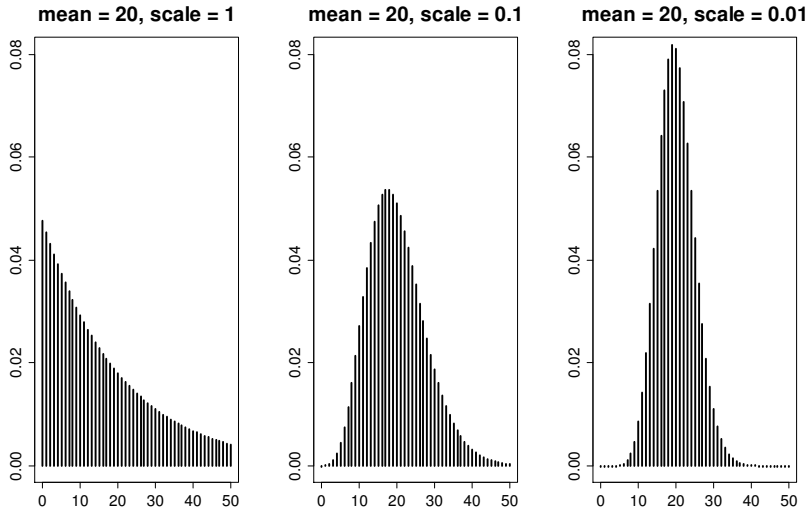
Poisson distribution

Residual degrees of freedom: 18

Deviance: 14203 (P < 0.0001)
Pearson chi-square: 13643 (P < 0.0001)
The observations are clearly overdispersed.

Using the previous method, standard errors would be multiplied by $\sqrt{\hat{\phi}} = \sqrt{13643/18} = 27.5$

The negative binomial distribution



Negative binomial distribution

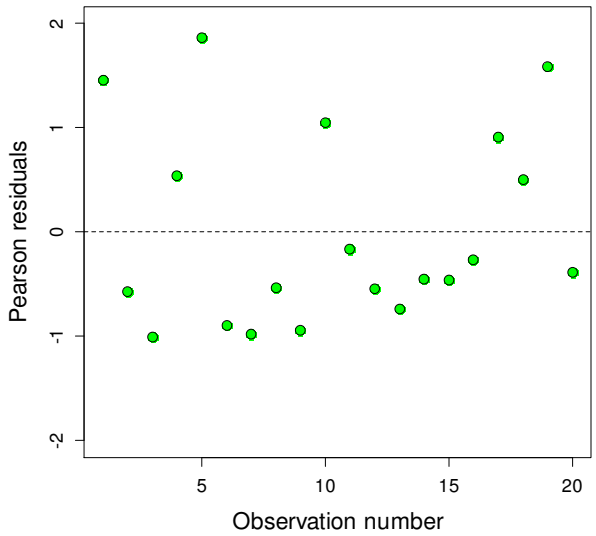
glm.nb in R, and genmod in SAS

Residual degrees of freedom: 18

Deviance: 22.77 (P = 0.200)
Pearson chi-square: 16.27 (P = 0.574)
The data is not overdispersed

Problem solved!

Pearson residuals



Negative binomial distribution

SAS

Differences of Method Least Squares Means					
Method	_Method	Estimate	Standard Error	z Value	Pr > |z|
Manual	Mechanical	**-1.7384**	0.4256	-4.08	**<.0001**

R

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	5.389	0.301	17.89	< 2e-16 ***
MethodMechanical	**1.738**	0.426	4.08	**4.4e-05** ***

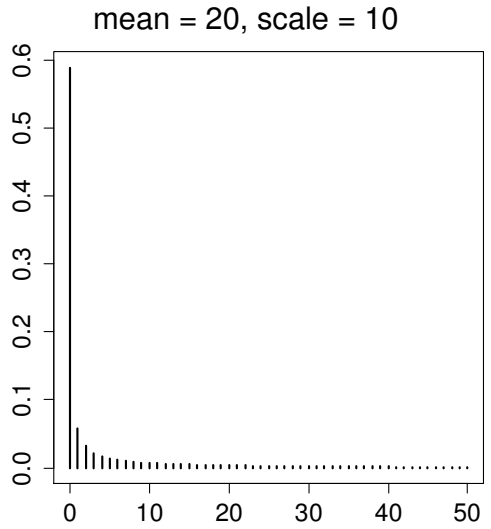
Zero-inflated data

0 0 0 0 0 0 4 0 0 16 0 0 5 0 0 0 10 0 0 0 0 12 0 0 0 0

Zero-inflated data

- There are more zeros than expected according to the Poisson or negative binomial distribution
- A special case of overdispersion
- Common in ecology when the numbers of various species are counted

A negative binomial distribution

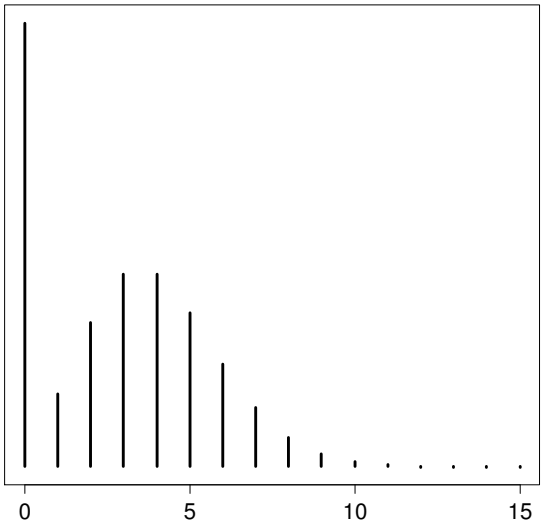


Zero-inflated data

This is not a Poisson or a negative binomial distribution

But it might be a mixture of a Poisson and a binomial distribution

Or a mixture of a negative binomial and a binomial distribution



Zero-inflated Poisson distribution

A mixture of a Poisson distribution and a binomial distribution

$$\Pr(\text{Extra zero}) = \pi$$

$$\Pr(\text{Observation zero})$$

$$= \pi + (1 - \pi)\Pr(\text{Poisson distribution gives a zero})$$

How to fit zero-inflated models

Zero-inflated Poisson and zero-inflated negative binomial models can be fitted

- using the zeroinfl function of the pscl package
- using the genmod procedure, through dist = zip and dist = zinb, respectively

See Exercise 5

Summary

Overdispersion is the rule rather than the exception. When not accounted for, the statistical inference is not valid.

The following methods were presented:

- Generalized linear mixed models
- Generalized estimating equations
- Adjustment using an overdispersion factor
- Negative binomial distribution
- Mixture distributions for zero-inflated data