Varieties of over- and underdispersion models for binary data

J.K. Lindsey

Biostatistics, Limburgs Universitair Centrum, Diepenbeek

Email: jlindsey@luc.ac.be

and P.M.E. Altham

Statistical Laboratory, University of Cambridge, CB2 1SB, UK

Email: p.m.e.altham@statslab.cam.ac.uk

Summary. Overdispersion is commonly treated as a nuisance factor in the analysis of binomial-type data. With the aid of an example, we consider various ways in which departures from the binomial distribution can arise. We fit four different generalizations of the binomial distribution, as well as a finite mixture model, to the data set and study why not all of these distributions provide reasonable fits. We conclude that the reasons that certain distributions may not be appropriate include the presence of underdispersion, an excess of extreme events, and nonhomogeneity of the reaction of subjects to some treatment.

Keywords: AIC; beta-binomial distribution; direct likelihood inference; double binomial distribution; mixture; 'multiplicative' binomial distribution; normal-binomial distribution; overdispersion; toxicology.

1 Introduction

Overdispersion in binary data is widely thought to be a nuisance, something to be adequately allowed for in making inferences about some linear model. Corrections range from simply modifying the standard errors by means of a heterogeneity factor to fitting a mixture distribution such as the beta-binomial. Rarely does thought appear to be invested in considering why overdispersion arises, other than stating that some important covariates must be missing.

In certain cases, the dispersion in the data can be modelled directly. For example, dispersion may itself vary systematically with the available covariates. Lindsey (1999) gives examples of experiments concerning fish eggs hatching where overdispersion only occurs in biologically extreme conditions, requiring response surface models both for the probability of an event and for the correlation among events.

Another possibility is that the form of the distribution differs significantly from a binomial distribution. Various generalizations of this distribution are now available that allow for overdispersion, and sometimes for underdispersion. Some of these will be discussed in the next section; they can yield quite different forms when fitted to various data sets.

From our investigations, we conclude that a generally applicable solution to over- or underdispersion is not available. Simple corrections to standard errors are usually inadequate. The widely used beta-binomial mixture does not always perform well. On the other hand, in the presence of overdispersion, the normal-binomial mixture often does do well. If underdispersion is present, one may choose between the double binomial and the multiplicative binomial distributions.

2 Models for overdispersion

The most commonly used model for overdispersion in binomial data is the beta-binomial (Skellam, 1948):

$$f(y;\pi,\psi) \quad = \quad \binom{n}{y} \frac{B(\pi e^{\psi}+y,(1-\pi)e^{\psi}+n-y)}{B(\pi e^{\psi},(1-\pi)e^{\psi})}$$

where typically n is the total number of siblings of a family, y the number of these giving a positive response, and $B(\cdot)$ is the beta function. Thus y may be written as say $X_1+\cdots+X_n$, where X_i is the response (either 0 or 1) of the ith sibling. This parametrization is convenient for comparison with the two to follow. The correlation between X_i and X_j, for any distinct i,j is $\rho = 1/(\exp(\psi)+1)$. For finite ψ, ρ is strictly positive, indicating overdispersion relative to the binomial distribution. This distribution can be derived from the binomial distribution by a mixture argument: the Bernoulli probability can be thought to vary in the population according to a beta distribution, and the marginal distribution taken. Although the binomial distribution is a member of the exponential family, the beta-binomial is not.

A closely related approach uses the normal distribution as the mixing distribution for the logit of the Bernoulli probability, in a way analogous to the model introduced by Hinde (1982) for overdispersed Poisson data:

$$\int_{-\infty}^{\infty} f(y;\pi(\lambda))\phi(\lambda;\mu,\psi^2)d\lambda$$

where $f(y;\pi)$ is the binomial frequency function, with π as the corresponding probability, and with $\log(\pi/(1-\pi)) = \lambda$, and $\phi(\cdot)$ the normal density with μ the mean logit and ψ the standard deviation. Here, however, numerical integration must be used in fitting the model. We shall use Gauss-Hermite integration.

Two members of the exponential family have also been proposed to handle overdispersion. Unfortunately, both have intractable normalizing constants and thus have not yet been widely used. However, these constants can now easily be calculated with fast computers by summing over possible values of the response variable.

Altham (1978) introduced two generalizations of the binomial distribution. That which she called the 'multiplicative' generalization is a member of the exponential family. It can be written

$$f(y;\pi,\psi) = c(\pi,\psi)\binom{n}{y}\pi^y(1-\pi)^{n-y}\mathrm{e}^{\psi y(n-y)}$$

where $c(\pi,\psi)$ is the intractable normalizing constant. The distribution will be overdispersed for $\psi < 0$, with $\psi = 0$ yielding the usual binomial distribution. This parameter, as -2ψ, is the log conditional cross-ratio for the responses of any pair of siblings given all of the others, for example the cross-ratio derived from the 2×2 contingency table for $P(X_1 = x_1, X_2 = x_2 | X_3 = x_3, ..., X_n = x_n)$ for $x_1, x_2 = 0, 1$, in the notation introduced above.

This distribution has y and $y(n-y)$ as the joint sufficient statistics for the probability and the dispersion parameter. This yields a special characteristic of this distribution in that the dispersion parameter disappears from the model, except in the normalization constant, when $y = 0$ or $y = n$. It also means that, when the distribution is fairly asymmetric, it can have a small secondary mode in the longer tail.

Efron (1986) proposed a family, which he called double exponential, that is also of exponential family type. For overdispersed binomial data, the double binomial distribution in this family may be appropriate:

$$f(y;\pi,\psi) = c(\pi,\psi)\binom{n}{y}\frac{n^{n\psi}\pi^{y(\psi+1)}(1-\pi)^{(n-y)(\psi+1)}}{y^{y\psi}(n-y)^{(n-y)\psi}}$$

where $c(\pi,\psi)$ is again an intractable normalizing constant. Again, the distribution will be overdispersed for $-1 < \psi < 0$. In this model, $1/(\psi+1)$ has an approximate interpretation as the variance inflation factor. Efron showed that for large n the normalizing constant c is approximately 1. However, Aitkin (1995) demonstrated that this approximation is inadequate for 'moderate' n. For the rat litters dataset used below the estimated parameter values (Table 3) with $n = 10$ give c(0.9,−0.217)=0.875 and c(0.9,−0.951)=0.426, both clearly rather different from 1.

One important advantage of the double-binomial and the multiplicative binomial models is that they allow for underdispersion as well as overdispersion.

Lindsey and Altham (1998) fit these three models to frequency data on the proportions of the two sexes among births of children in nineteenth century Saxony. In contrast to the direct procedure used here, they fitted generalized models with a Poisson distribution and log link to frequency (histogram) data (Lindsey and Mersch, 1992). This method is restricted to studies where a large number of occurrences of each possible outcome is observed. Hence, it cannot be applied to ordinary contingency tables where overdispersion may be present.

3 Examples

Except for the 'pure' binomial model, these models are non-nested. They will be compared using a direct likelihood approach whereby the negative log likelihood is penalized by adding to it the

Table 1: Numbers of offsprings of pregnant rats surviving at four and 21 days, by litter. (Williams, 1975)

Control		Treated	
4 days	21 days	4 days	21 days
13	13	12	12
12	12	11	11
9	9	12	12
9	9	9	9
8	8	11	10
8	8	10	9
13	12	10	9
12	11	9	8
10	9	9	8
10	9	5	4
9	8	9	7
13	11	7	4
5	4	10	5
7	5	6	3
10	7	10	3
10	7	7	0

number of parameters estimated (a form of the Akaike (1973) information criterion: AIC). Smaller values indicate relatively better models. (See Bai *et al* (1992) for a similar use of AIC's in modelling contingency tables.) These relative values of the AIC's are of course only indicative of the fits of the various models. Because of sampling errors, we cannot attach much importance to the fact that one AIC is, say, 58.7, whilst another is 57.6 although the latter model could contain one completely redundant parameter and still have equivalent fit to the former one.

The deviances quoted below are the standard *two times* the difference in negative log likelihood with respect to the saturated model.

3.1 Rat litters

Consider an experiment in which 16 female rats received a control diet during pregnancy while 16 others received a chemically treated diet. The numbers (n) of offspring alive in the litters at four days were recorded and they were followed to ascertain how many (y) were still alive at 21 days (reproduced in Table 1). The data have previously been analyzed by Williams (1975) and Ochi and Prentice (1984).

The binomial model, allowing for a difference between food treatments, has an AIC of 68.2. The deviance is 87.2 with 30 degrees of freedom, indicating overdispersion. All four overdispersion distri-

Table 2: AICs for various overdispersion models for the rat litter data with either the probability and/or the dispersion parameter differing between treatment groups. (For the technical reason that software is not available, the normal-binomial distribution with the same probability and different dispersion could not be fitted.)

Distribution	Both same	Different probability	Different dispersion	Both different
Binomial	71.5	68.2	—	—
Beta-binomial	58.7	58.7	58.7	56.9
Normal-binomial	57.8	57.4	—	56.5
Double binomial	57.6	57.1	55.3	56.3
Multiplicative binomial	54.3	54.2	55.3	53.4

Table 3: Parameter estimates for the best model from each distribution for the rat litter data. (The standard error for the normal-binomial logit difference is not available because the model was fitted separately to the two treatment groups, not simultaneously as for the other distributions.)

	Control		Treated		Logit difference	
	$\hat{\pi}$	$\hat{\psi}$	$\hat{\pi}$	$\hat{\psi}$	Estimate	s.e.
Binomial	0.899	—	0.776	—	−0.943	0.330
Beta-binomial	0.898	3.880	0.741	0.750	−1.126	0.463
Normal-binomial	0.906	0.484	0.844	1.757	−0.575	—
Double binomial	0.901	−0.217	0.901	−0.951	0.071	1.768
Multiplicative binomial	0.815	−0.096	0.564	−0.270	−1.228	0.738

butions provide improved fits, as shown in Table 2. The multiplicative binomial, with different values for both the probability and the dispersion, gives the best fit. Notice how the double binomial does not indicate a difference in probability of survival.

The parameter estimates for the best model from each distribution are presented in Table 3. Obviously, the estimates of the dispersion parameter are not comparable among the distributions. The estimates for the probability (π) given by the binomial distribution, the observed proportions in the raw data, are 0.899 and 0.776. Those for the beta-binomial are also close to these. However, this is not true of the better fitting models.

This can be clarified by examining a plot of the various models, given in Figure 1. In fact, the three better fitting models are bimodal for the treated group, indicating a group with low survival probability. A look at the data reveals that this is an accurate reflection of what is occurring. Several of the treated litters have low probability of survival. Thus, one single probability is not an adequate

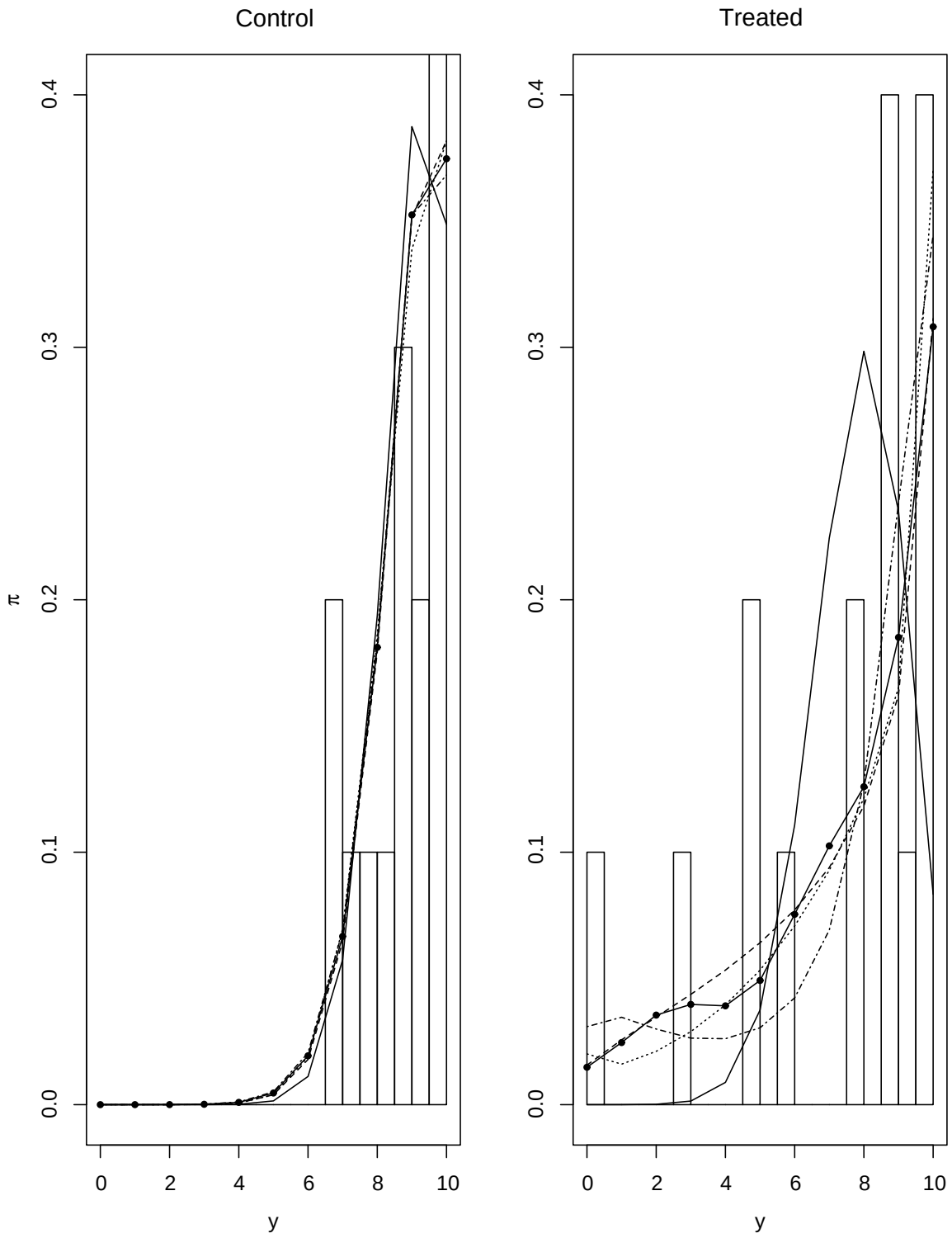

Figure 1: Fitted values for the five models applied to the rat data for a litter of size $n = 10$ in each of the two treatment groups. The vertical scale has been exaggerated to show the second mode. Binomial: solid; beta-binomial: dashed; double binomial: dotted; multiplicative binomial: dash-dotted; normal-binomial: solid with dots. The heights of the bars represent the empirical proportions of survivors with all litters standardized to size 10.

measure of survival for the treated group. As well, not only does the probability of survival vary between the two groups but the variability in survival is also different.

The second small mode for these skewed distributions could, for example, occur if the treatment has little effect for most of that group but negative effect for a minority. Thus, another useful approach to analyzing these data might be to apply a two-component finite mixture model (Brooks *et al.*, 1997). When this is done with a mixture of two binomial distributions for the treatment group (and one binomial distribution for the control), the AIC is 65.3, fitting more poorly than any of the overdispersion models considered. It would also be possible to fit a finite mixture of any one of the three overdispersion distributions considered above, but such a model would be difficult to interpret biologically.

3.2 Other examples

We have considered a number of other examples. We describe briefly two of them, omitting the computational details here. These were

1. Anderson's (1988) data on grasshopper chromosomes, for which the normal-binomial provides the best fit and

2. the data set given by Hand *et al.* (1994, p. 138) on sizes of Duroc-Jersey pig litters and the sex of the young. These data exhibit clear under-dispersion relative to the binomial. We found that for this data set the multiplicative and double binomial distributions fit almost equally well. Both are much better than the binomial, which in turn is better than either of the beta-binomial or normal-binomial.

4 Discussion

Departures from a binomial distribution can occur in a variety of ways. In checking for such departures, the following points should be considered:

- Over- or underdispersion may be the result of a single error in the table, arising, for example, either when conducting a study or in recording the results.

- An excess number of events may occur for the two extreme possibilities ($y = 0$ and/or $y = n$) as compared to the binomial distribution. This can be modelled by using a finite mixture with extra probability masses at these two extremes. Lindsey and Altham (1998) used such a model for the sex ratio, but it did not account for all of the overdispersion. Such an approach might also be appropriate for the rat data where many litters had all young alive after 21 days, as can be seen in Table 1.

- Treatment, or some covariate, may not have a uniform effect on all subjects. This may cause a second small mode to appear for certain distributions as with the rat data; this can be seen in

Figure 3 for the treated group. In such a situation, a scientifically more useful procedure might be to fit a finite mixture of two binomial distributions (Brooks *et al.*, 1997); this did not turn out to produce a good model for the rat data.

- Although all of the models used for departures from the binomial distribution can handle overdispersion, of those considered here, only the multiplicative and double binomial can adequately allow for underdispersion. The latter may arise in the context of repulsion as with plants growing in a plot, or other forms of negative dependence, as when subjects are competing for a finite supply of resources. An example of the latter might occur in the survival of the strongest in a litter, although this does not appear to be relevant for the examples used above.

It is particularly interesting that the normal-binomial mixture can apparently be bimodal because Holgate (1970) demonstrated that the normal-Poisson mixture is unimodal. (We used 14-point Gauss–Hermite integration which should provide a good numerical approximation.)

Interpretation of the biological reasons for over- or underdispersion is always difficult for secondary analysis of data. It can only properly be done through close interaction with the scientists involved in the study.

Diagnostics are reasonably well known for generalized linear models, as well as for some of their extensions such as the beta-binomial and negative binomial distributions (Pregibon, 1981; Williams, 1987). However, their application is not as clear for other dispersion models less closely related to generalized linear models. We are currently investigating this problem.

Acknowledgments We thank Robert Gentleman and Ross Ihaka for developing the R software in which two functions were written to fit the models to the data in the examples. These are available from the first author or on CRAN (`http://www.ci.tuwien.ac.at/R/`) in the R libraries, `gnlm` and `repeated`.

References

Aitkin, M. (1995) Model choices in single samples from the exponential and double exponential families using the posterior Bayes factor. *Statist. Comput.* **5**, 113–120.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. in Petrov, B.N. and Csàki, F., eds *Second International Symposium on Inference Theory*, Budapest: Akadémiai Kiadó, pp. 267–281.

Altham, P.M.E. (1978) Two generalizations of the binomial distribution. *Appl. Statist.* **27**, 162–167.

Anderson, D.A. (1988) Some models for overdispersed binomial data. *Aust. J. Statist.* **30**, 125–148.

Bai, Z.D., Krishnaiah, P.R., Sambamoorthi, N., and Zhao, L.C. (1992) Model selection for log-linear model. *Sankhya* **B54**, 200-219.

Brooks, S.P., Morgan, B.J.T., Ridout, M.S., and Pack, S.E. (1997) Finite mixture models for proportions. *Biometrics* **53**, 1097–1115.

Efron, B. (1986) Double exponential families and their use in generalized linear regression. *J. Am. Statist. Ass.* **81**, 709–721.

Hand, D.J., Daly, F., Lunn, A.D., McConway, K.J., and Ostrowski, E. (1994) *A Handbook of Small Data Sets.* London: Chapman and Hall.

Hinde, J. (1982) Compound Poisson regression models. In Gilchrist, R. (1982, ed.) *GLIM82.* Berlin: Springer-Verlag, 109–121.

Holgate, P. (1970) The modality of some compound Poisson distributions. *Biometrika* **59**, 666–667.

Ihaka, R. and Gentleman, R. (1996) R: a language for data analysis and graphics. *J. Comput. Graph. Statist.*, **5**, 299–314.

Lindsey, J.K. (1999) Response surfaces for overdispersion in the study of the conditions for fish eggs hatching. *Biometrics* **55**, (in press).

Lindsey, J.K. and Altham, P.M.E. (1998) Analysis of the human sex ratio using overdispersion models. *App. Statist.* **47**, 149–157.

Lindsey, J.K. and Mersch, G. (1992) Fitting and comparing probability distributions with log linear models. *Comput. Statist. Data Anal.* **13**, 373–384.

Ochi, Y. and Prentice, R.L. (1984) Likelihood inference in a correlated probit regression model. *Biometrika* **71**, 531–543.

Pregibon, D. (1981) Logistic regression diagnostics. *Ann. Statist.* **9**, 705–724.

Prentice, R.L. (1986) Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement error. *J. Am. Statist. Ass.* **81**, 321–327.

Skellam, J.G. (1948) A probability distribution derived from the binomial distribution by regarding the probability of success as variable between sets of trials. *J. R. Statist. Soc.* **B10**, 257–261.

Williams, D.A. (1975) The analysis of binary responses from toxicological experiments involving reproduction and teratogenicity. *Biometrics* **31**, 949–952.

Williams, D.A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. *App. Statist.* **36**, 181–191.