J. R. Statist. Soc. A (1999)
162, *Part* 2, *pp.*195–209

The zero-inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology

Dankmar Böhning, Ekkehart Dietz and Peter Schlattmann

Free University of Berlin, Germany

and Lisette Mendonça and Ursula Kirchner

Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

[Received January 1996. Final revision September 1998]

Summary. For frequency counts, the situation of extra zeros often arises in biomedical applications. This is demonstrated with count data from a dental epidemiological study in Belo Horizonte (the Belo Horizonte caries prevention study) which evaluated various programmes for reducing caries. Extra zeros, however, violate the variance–mean relationship of the Poisson error structure. This extra-Poisson variation can easily be explained by a special mixture model, the zero-inflated Poisson (ZIP) model. On the basis of the ZIP model, a graphical device is presented which not only summarizes the mixing distribution but also provides visual information about the overall mean. This device can be exploited to evaluate and compare various groups. Ways are discussed to include covariates and to develop an extension of the conventional Poisson regression. Finally, a method to evaluate intervention effects on the basis of the ZIP regression model is described and applied to the data of the Belo Horizonte caries prevention study.

Keywords: Caries prevention study; Decayed, missing and filled teeth index; Mixture model; Poisson model with zero inflation; Zero-inflated Poisson model graphics

1. Introduction

Data that consist of counts often occur in areas such as public health, epidemiology, sociology, psychology, engineering and agriculture. Typically, a Poisson model $\Pr(Y = y) = \mathrm{Po}(y, \mu)$ is assumed for modelling the distribution of the count observations Y or, at least, to approximate their distribution. However, the Poisson model often underestimates the observed dispersion. This phenomenon, also called *overdispersion* or *extra-Poisson variation*, occurs because a single Poisson parameter μ is often insufficient to describe the population (see for example McCullagh and Nelder (1989), Aitkin *et al.* (1989), Breslow (1984) and Campbell *et al.* (1991)). In fact, in many cases it can be suspected that *population heterogeneity* which has not been accounted for is causing this overdispersion. This population heterogeneity is unobserved, i.e. the population consists of several subpopulations, in this case of Poisson type, but the subpopulation membership is *not* observed in the sample. One approach to the problem is to assume that the heterogeneity involved in the data can be adequately described by some density $g(\mu)$ defined on the population of possible Poisson parameters μ. Since this

Address for correspondence: Dankmar Böhning, Department of Epidemiology, Institute for Social Medicine, Free University of Berlin, Fabeckstrasse 60–62, Haus 562, 14195 Berlin, Germany.
E-mail: boehning@zedat.fu-berlin.de

 0964–1998/99/162195

heterogeneity cannot be observed directly, it is also called *latent*. We can only observe counts coming from the *marginal* or *mixture* density

$$\int_0^\infty \mathrm{Po}(y, \mu)\, g(\mu)\, \mathrm{d}\mu.$$

Two approaches can be distinguished. The traditional approach is to follow a fully *parametric* model for the mixing density g. An example of this nature is the gamma distribution for g, for which the marginal density becomes the negative binomial distribution (Hogg and Tanis (1983), p. 380). The second approach, the *nonparametric* approach, does not specify any parametric density for $g(\mu)$. Here, the nonparametric maximum likelihood estimator of the mixing density g is always finite, giving weights p_j to the *latent* classes or subpopulations μ_j, $j = 1, \ldots, k$ (Simar, 1976; Böhning, 1982, 1995; Lindsay, 1983; Laird, 1978). This nonparametric approach is attractive, since it is not only easy to interpret but also requires *no* specification of the number of latent classes k.

In this paper we study the simplest mixing distribution, namely a two-mass distribution giving mass $1 - p$ to 0 and mass p to the second class with mean μ. This model is also called the *zero-inflated Poisson (ZIP)* model (Lambert, 1992; Johnson *et al.*, 1992; Yip, 1991; Fong and Yip, 1993). For medical applications of the ZIP model, see for example Demétrio and Ridout (1994) and Campbell *et al.* (1991). This model has been found to be useful in the study of the prevention of dental caries, described in the next section.

2. The decayed, missing and filled teeth index in dental epidemiology

In dental epidemiology, the decayed, missing and filled teeth (DMFT) index is an important and well-known indicator and overall measure of the dental status of a person. As an application, we consider here data from a prospective study of school-children from an urban area of Belo Horizonte (Brazil), the Belo Horizonte caries prevention (BELCAP) study. The data can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

The children were all 7 years of age at the beginning of the study, and schools with similar socioeconomic backgrounds were selected. See Mendonça and Böhning (1994) and Mendonça (1995). Fig. 1 shows the DMFT distribution at the beginning of the study. Only the eight deciduous molars were considered, which implies that the smallest possible value of the DMFT index is 0 and the largest is 8. There is a clear spike of extra zeros representing the caries-free children. This pattern is typical of DMFT data. Nevertheless, the line of argument followed in dental epidemiology is that the DMFT index is a count variable, and that Poisson distributions with log-linear modelling to include covariates would be an appropriate method of analysis. However, as can be seen in Fig. 1, the homogeneous Poisson distribution does not provide an adequate fit to these data.

The aim of the caries prevention study was to compare four methods to prevent dental caries. Interventions were carried out according to the following scheme: school 1, oral health education; school 2, all four methods together; school 3, the control group; school 4, enrichment of the school diet with rice bran; school 5, mouthwash with 0.2% sodium fluoride (NaF) solution; school 6, oral hygiene. The six treatments were randomized to the six schools, so that all children of a given school received the same treatment. 797 school-children were examined both before and after the trial, their dental status evaluated and the DMFT index computed.

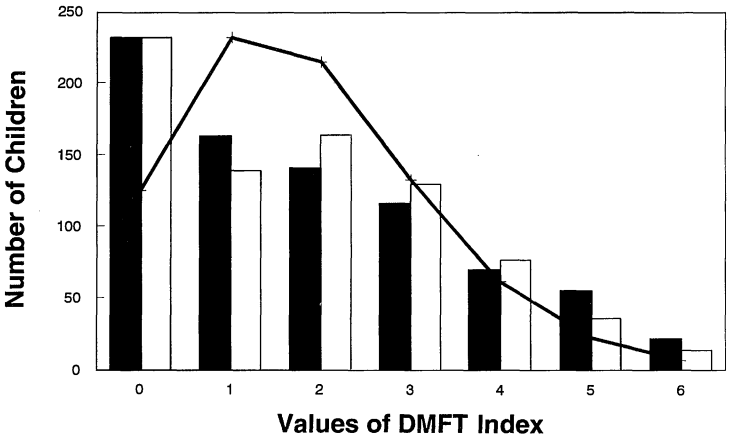

Fig. 1. DMFT distribution for 797 children at the beginning of the study: ■, observed; ——, homogeneous Poisson distribution; □, ZIP model

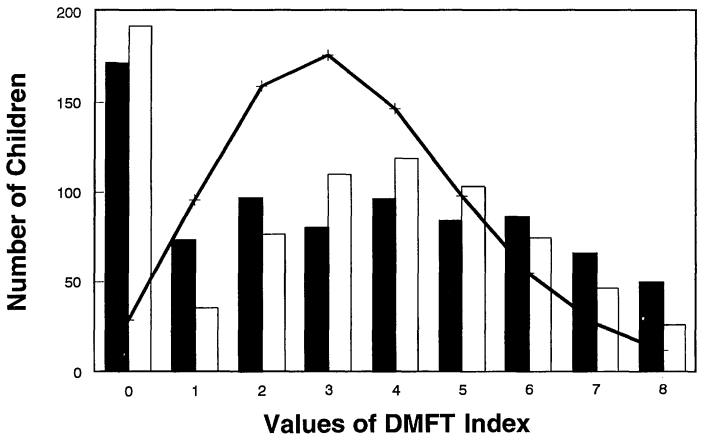

Fig. 2. DMFT distribution for 797 children at the end of the study: ■, observed; ——, homogeneous Poisson distribution; □, ZIP model

The distribution of the DMFT index at the end of the study is presented in Fig. 2. Let DMFT1 denote the dental status at the beginning of the study and DMFT2 the dental status 2 years later. It is evident from a comparison of Fig. 1 with Fig. 2 that the dental status has improved for many children. However, for 486 children the DMFT index after 2 years (DMFT2) is lower than the DMFT index at the beginning of the study (DMFT1). How can the DMFT value decrease with time? In the examinations the decay value in the DMFT index was measured on the conventional scale expressing not only decay but also lesions of the tooth graded above 0, where the grading is as follows: grade 0, healthy; grade 1, light, chalky spot; grade 2, thin brown–black line; grade 3, damage, not larger than 2 mm wide; grade 4, damage, wider than 2 mm. A DMFT index which also accounts for the grade of lesions is sometimes also denoted the $D_{1\text{-}4}MFT$ index (Pilz, 1985). It is possible—as in the BELCAP study—that a tooth with a low grade of lesion can recover in time. This might be completely different in studies on older populations where only negative development can be expected.

For every child in the BELCAP study, every tooth missing, or having a filling or decayed (or even a lesion graded 1–4) was counted as contributing 1 unit to the DMFT index. For example, a child with a DMFT index of 3 had three teeth which were either missing, filled or decayed.

3. The zero-inflated Poisson model

A simple and frequently applied statistical model for a count distribution is the Poisson model. In this model Y is assumed to follow a Poisson density

$$\Pr(Y = y) = \exp(-\mu)\mu^y/y! = \mathrm{Po}(y, \mu). \tag{3.1}$$

As we have seen in Section 2, model (3.1) does not fit the DMFT data, because of the large frequency of extra zeros in the distribution. A simple way to model this *zero inflation* is to include a proportion $1-p$ of extra zeros and a proportion $p\exp(-\mu)$ from the Poisson distribution (Johnson *et al.* (1992), p. 314, and Lambert (1992)). We can write this ZIP density f as

$$f(x; p, \mu) = \begin{cases} 1 - p + p\exp(-\mu), & \text{if } y = 0, \\ p\,\mathrm{Po}(y, \mu), & \text{if } y > 0, \end{cases} \tag{3.2}$$

or

$$f(y; p, \mu) = (1-p)\,\mathrm{Po}(y, 0) + p\,\mathrm{Po}(y, \mu). \tag{3.3}$$

Representation (3.3) indicates that the ZIP model is a special mixture model having two classes, where the first class has a *fixed* value at 0. In the case of the DMFT index, this class could include those children with *no* caries risk at all.

For the ZIP model with zero inflation we find that

$$\begin{aligned} \mathrm{var}(Y) &= E(Y) + E(Y)\{\mu - E(Y)\}, \\ E(Y) &= p\mu. \end{aligned} \tag{3.4}$$

For the DMFT2 data, we find overdispersion $s^2 - \bar{y} = 1.05$. The maximum likelihood estimators for the ZIP model (for details see Appendix B) are $\hat{p} = 0.78$ and $\hat{\mu} = 2.37$ in this case, leading to a fitted overdispersion (under equations (3.4)) of $\widehat{E(Y)}\{\hat{\mu} - \widehat{E(Y)}\} = 0.95$, corresponding to an explained overdispersion of

$$\widehat{E(Y)}\{\hat{\mu} - \widehat{E(Y)}\}/(s^2 - \bar{y}) = 0.90.$$

Thus 90% of the overdispersion would be *explained* by the ZIP model. Indeed, we could consider a formal way of testing the hypothesis H_0: $E(S^2) = E(Y)\{1 + \mu - E(Y)\}$, i.e. whether the ZIP model satisfactorily explains the overdispersion. One way to test this hypothesis formally would be to compare the likelihood under this null hypothesis with the likelihood of the nonparametric maximum likelihood estimator. In Appendix A moment and maximum likelihood estimators are discussed for the ZIP model. In what follows, let $\hat{\mu}$ and $\hat{p}$ denote the maximum likelihood estimates of μ and p respectively.

4. A graphical representation of the zero-inflated Poisson model

In this section we provide a graphical display which presents summary information on the

various parts of the ZIP model. The display contains a rectangle with a base-line of length μ, so that the two end points of this base-line represent the two component means of the mixture model. The height of the rectangle is p, thus showing the *distribution of the mixing distribution*. Now, the area of the rectangle is $p\mu = E(Y)$. For sample replacement this equation becomes $\bar{Y} = \hat{\mu}\hat{p}$, because of the estimating equations given in Appendix A. The graphical display is helpful in comparing the various parameters between groups. For example, we could define group A to be *strongly* better than group B if and only if $p^A \leqslant p^B$ *and* $\mu^A \leqslant \mu^B$ (with at least one inequality being strict). Graphically this means that the rectangle of group B contains the rectangle of group A. Group A could be defined to be *weakly* better than group B if $E(Y^A) < E(Y^B)$, the latter being the more traditional criterion of comparison. Again, graphically this means that the area of the rectangle associated with group B is larger than the rectangle associated with group A. Note that 'strongly better' implies 'weakly better'.

Fig. 3 presents the DMFT2 index for the six schools in terms of this graphical device, where the estimates are adjusted for potential confounders as will be described later in Sections 6 and 7. As can be seen from Fig. 3, school 2 (all four intervention methods together) is strongly (in both μ and p) better than all the other schools whereas school 4 (diet) is only *weakly* better than school 3 (control).

Another favourable property of this graphical device is that we can see to what extent the variation in the DMFT2 index between schools is due to the second component mean μ or due to the second component weight p or both. It is quite striking, as seen in Fig. 3, that the variation between schools is due more to the mean $\hat{\mu}$ than to the component weight $\hat{p}$.

5. Inference for the population mean

Although interest lies in the components of the ZIP model, in practice the overall mean $E(Y)$ is still required as a summary. In this section we investigate the question of constructing a confidence interval for the population mean $E(Y)$. The large sample approximate method of constructing a $100(1-\alpha)\%$ confidence interval would be $\bar{Y} \pm z\sqrt{\text{var}(Y)}/\sqrt{n}$, with $z = z_{1-\alpha/2}$ being the $(1-\alpha/2)$-quantile of the standard normal distribution. Replacing the unknown

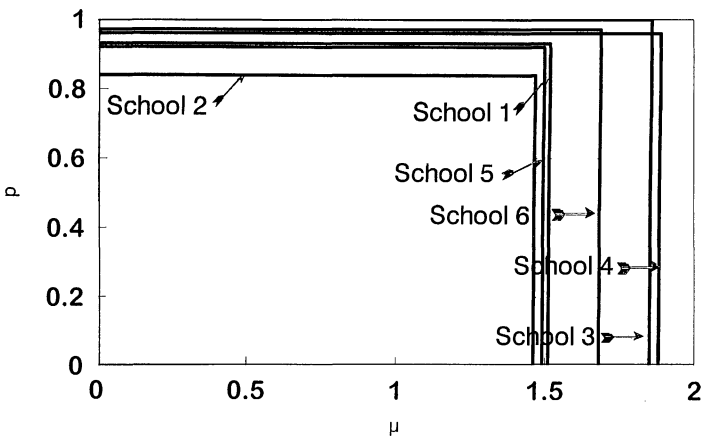

Fig. 3. Six schools in the BELCAP study in a graphical comparison of DMFT2 values: the area of a rectangle corresponds to the mean of DMFT2

var(Y) by its unbiased estimator S^2 gives $\bar{Y} \pm zS/\sqrt{n}$. Call this method 1. In contrast, equation (3.4) gives $\text{var}(Y) = E(Y) + E(Y)\{\mu - E(Y)\}$. We consider the statement

$$L\{E(Y)\} \leqslant \bar{Y} \leqslant U\{E(Y)\} \tag{5.1}$$

with

$$L\{E(Y)\} = E(Y) - z\sqrt{[E(Y)\{1 + \mu - E(Y)\}]}/\sqrt{n}$$

and

$$U\{E(Y)\} = E(Y) + z\sqrt{[E(Y)\{1 + \mu - E(Y)\}]}/\sqrt{n}.$$

It is easily seen by the monotonicity of the functions L and U that inequality (5.1) is equivalent to

$$U^{-1}(\bar{Y}) \leqslant E(Y) \leqslant L^{-1}(\bar{Y}). \tag{5.2}$$

$U^{-1}(\bar{Y})$ and $L^{-1}(\bar{Y})$ can be found as the *two* solutions of the quadratic equation in x

$$(1 + z^2/n)x^2 - (2\bar{Y} + z^2\mu/n + z^2/n)x + \bar{Y}^2 = 0. \tag{5.3}$$

Call this method 2. If $p = 1$, i.e. if $E(Y) = \mu$, then $\text{var}(Y) = E(Y)$ and inequality (5.2) leads to the classical approximating confidence interval for the expected value of a Poisson count (Breslow and Day, 1985)

$$\bar{Y} + \frac{z^2}{2n} \pm \frac{z}{\sqrt{n}}\sqrt{\left(\bar{Y} + \frac{z^2}{4n}\right)}.$$

The expressions $L\{E(Y)\}$ and $U\{E(Y)\}$ still contain the unknown parameter μ. It can be replaced by its maximum likelihood or moment estimator. In our case we used the maximum likelihood estimator of μ. Methods 1 and 2 were compared in terms of coverage probability and expected length in a small simulation experiment. Two combinations were studied, one with a sample size of 100, $p = 0.5$ and $\mu = 0.1, 0.2, \ldots, 2.9, 3.0$, which corresponds closely to the situation found in the BELCAP study. The other combination is $n = 10$, $p = 0.1$ and $\mu = 0.05, 0.10, 0.15, \ldots, 0.95, 1.0$, which models a situation of a small sample size with a large (90%) zero inflation.

Fig. 4 shows the coverage probabilities of both methods for the two combinations. Clearly, method 2 gives results that are closer to the nominal level for all values of μ. In the small sample case with high zero inflation, method 1 is a complete failure, whereas method 2 is still behaving reasonably well. For all parameter combinations, the replication size was set to 10000. Fig. 5 presents the means with 95% confidence interval constructed by method 2 for the DMFT index for the six schools in the BELCAP study. Significant DMFT mean differences between schools as well as significant differences of DMFT1 mean and DMFT2 mean within schools can be observed.

6. Including covariates

Frequently, a variety of further variables are considered in a study, either as explanatory factors or as confounders such as sex, age and exposure covariates. The data could be arranged to form strata so that stratum-specific μs and ps could be estimated as described in Section 4. However, the stratified approach has its limitations when the number of covariates

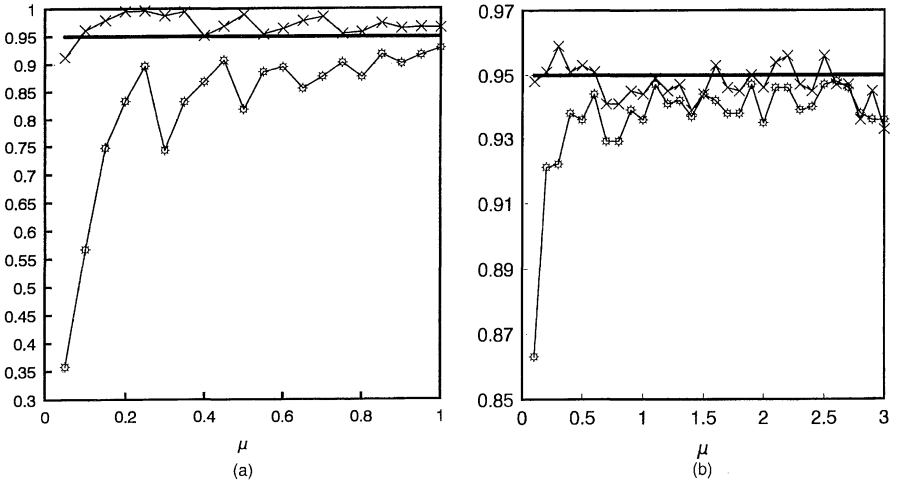

Fig. 4. Coverage probabilities of method 1 (☼) and method 2 (×) and the nominal level (——): (a) $n = 10$, $p = 0.1$; (b) $n = 100$, $p = 0.5$

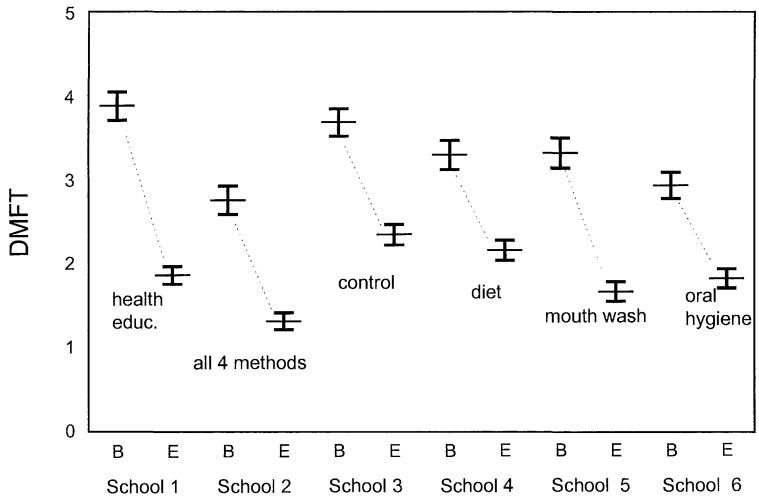

Fig. 5. Mean of the DMFT values at the beginning (B) and end (E) of the study, with 95% confidence intervals (based on method 2) for the six schools (the treatments applied to the schools are named above the school names)

is increasing. Alternatively, we might try to combine the specific error structure of the ZIP model with the framework of generalized linear models, in particular with Poisson regression. The conventional log-linear model

$$E(Y) = \exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}),$$

where $\mathbf{x}$ is the vector of covariates, α is an unknown intercept parameter and $\boldsymbol{\beta}$ is an unknown vector of regression coefficients, can easily be generalized to the ZIP regression giving

$$Y|\mathbf{x} \sim (1-p)\,\mathrm{Po}(y,\,0) + p\,\mathrm{Po}(y,\,\mu) = (1-p)\,\mathrm{Po}(y,\,0) + p\,\mathrm{Po}\{y,\ \exp(\alpha + \beta^{\mathrm{T}}\mathbf{x})\} \tag{6.1}$$

conditional on the values of the covariates. A further generalization, which is taken into consideration here, is also to allow the mixture weight p to depend on covariates. We generalize equation (6.1) to

$$\begin{aligned} Y|\mathbf{x} \sim (1-p)\,\mathrm{Po}(y,\,0) + p\,\mathrm{Po}(y,\,\mu) &= \{1 - p(\alpha' + \beta'^{\mathrm{T}}\mathbf{x}')\}\,\mathrm{Po}(y,\,0) \\ &\quad + p(\alpha' + \beta'^{\mathrm{T}}\mathbf{x}')\,\mathrm{Po}\{y,\ \exp(\alpha + \beta^{\mathrm{T}}\mathbf{x})\} \end{aligned} \tag{6.2}$$

where $\mathbf{x}'$ is a (potentially different) vector of covariates influencing the zero inflation and α' and β' are additional unknown parameters of the model. Various choices of the function $p(\cdot)$ are possible, e.g. $p(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$ as suggested by Lambert (1992). However, choosing a suitable function $p(\cdot)$ might be an additional problem. This can be avoided if a few strata, J say, can be created by means of the covariate $\mathbf{x}'$ such that homogeneous zero inflation can be assumed within each stratum. Then equation (6.2) simplifies to

$$Y|\mathbf{x}, j \sim (1-p_j)\,\mathrm{Po}(y,\,0) + p_j\,\mathrm{Po}\{y,\ \exp(\alpha + \beta^{\mathrm{T}}\mathbf{x})\} \tag{6.3}$$

for each stratum $j = 1, \ldots, J$.

We have a model which can be placed in a class of generalized linear models for heterogeneity as recently considered by Dietz (1992). In Dietz (1992) and Dietz and Böhning (1995), methods for finding the maximum likelihood estimates via the EM algorithm were discussed. A variation and adaptation to the case considered here is provided in Appendix B.

For the data of the BELCAP study introduced in Section 2, the five indicator variables for the intervention schools were used as components of $\mathbf{x}$, each representing a specific prevention strategy. These are the covariates of main interest in this study. As additional potential confounders, the following variables were considered: SEX (a binary covariate with girls as the base-line) and COLOUR (ethnic group, a covariate with three categories: *dark* (base-line), *white* (2) and *black* (3)).

Also the mixture weight is potentially allowed to depend on the school indicators. In this case we must consider $J = 6$ strata when using the stratification approach.

7. Modelling the intervention effect

It is quite striking, as seen in Fig. 5, that a statistically significant improvement in each school, even in the control school (school 3) exists. To decide whether or not an effect of intervention in an intervention school actually exists, we can use model (6.1) by considering DMFT2 as the response variable and log(DMFT1 + 0.5) as an additional covariate in the linear predictor. In the model, this variable fulfils two functions. One is to serve as an offset, so that it becomes possible to analyse to which covariates the change in the DMFT index (from the beginning to the end of the study) can be ascribed. It is easy to see that in this case the linear predictor, obtained by deleting the offset, is a linear predictor of $\log\{E(\mathrm{DMFT2})\} - \log\{E(\mathrm{DMFT1})\}$, which is a sensible measurement of the change in the DMFT index. Notice that $E\{\log(Y + 0.5)\} \approx \log\{E(Y)\}$ for Y a Poisson variable with mean μ and $1 < \mu < 10$, so the values of log(DMFT1 + 0.5) can be considered as estimates of $\log\{E(\mathrm{DMFT1})\}$.

However, it could be argued that the potential for improvement is larger in those schools with a higher DMFT value at the beginning of the study. Since the children were not randomly assigned to the schools (treatments), a considerable variation between schools is

Table 1. Effect estimates with standard errors for Poisson and ZIP regression on the DMFT2 index for the covariates SCHOOL, SEX, COLOUR and log(DMFT1)†

Estimate		*Standard error*		*Z-value*		*Parameter*
Log-linear model: $Po\{y, \exp(\alpha + \beta^T \mathbf{x})\}$						
−0.211	(0.759)	0.093	(0.072)	−2.275	(10.568)	Intercept
0.013	(0.131)	0.053	(0.053)	0.245	(2.484)	SEX
0.046	(0.099)	0.058	(0.058)	0.800	(1.718)	COLOUR(2)
−0.050	(−0.138)	0.087	(0.087)	−0.571	(−1.591)	COLOUR(3)
−0.279	(−0.233)	0.087	(0.087)	−3.201	(−2.663)	SCHOOL1
−0.401	(−0.588)	0.096	(0.096)	−4.166	(−6.139)	SCHOOL2
−0.018	(−0.089)	0.082	(0.082)	−0.218	(−1.094)	SCHOOL4
−0.289	(−0.349)	0.085	(0.084)	−3.425	(−4.140)	SCHOOL5
−0.116	(−0.299)	0.089	(0.089)	−1.299	(−3.362)	SCHOOL6
0.800		0.042		19.254		log(DMFT1)
ZIP model: $(1-p)\,Po(y, 0) + p\,Po\{y, \exp(\alpha + \beta^T \mathbf{x})\}$; $\hat{p} = 0.9502$						
−0.181	(0.943)	0.097	(0.075)	−1.862	(12.495)	Intercept
−0.003	(0.098)	0.055	(0.055)	−0.055	(1.765)	SEX
0.058	(0.083)	0.060	(0.061)	0.971	(1.384)	COLOUR(2)
−0.043	(−0.117)	0.092	(0.091)	−0.466	(−1.287)	COLOUR(3)
−0.238	(−0.222)	0.091	(0.092)	−2.603	(−2.417)	SCHOOL1
−0.350	(−0.470)	0.101	(0.101)	−3.467	(−4.671)	SCHOOL2
−0.001	(−0.064)	0.085	(0.086)	−0.016	(−0.746)	SCHOOL4
−0.262	(−0.223)	0.089	(0.089)	−2.961	(−2.512)	SCHOOL5
−0.105	(−0.226)	0.094	(0.093)	−1.123	(−2.407)	SCHOOL6
0.799		0.044		18.261		log(DMFT1)

†The Z-value is the estimate divided by its standard error; numbers in parentheses refer to the corresponding model *without* log(DMFT1).

Table 2. Log-likelihood for the three distributional models and the various covariates†

Model	*Log-likelihoods for models with the following parameters:*	
	SEX, COLOUR, SCHOOL	*With log(DMFT1)*
Log-linear	−1473.20	−1252.37
ZIP	−1410.27	−1246.89
Mixture	−1406.30	−1246.89
ZIP, p dependent on school	−1402.27	−1242.68

†The response variable is DMFT2.

possible. Indeed, Fig. 5 shows that school 1 has not only the largest improvement but also the largest DMFT mean at the start of the study. Thus, the second function of this covariate is to control for this kind of effect modification. Because, in Poisson regression or, more generally, in the regression part of the ZIP model, the logarithm of the expected value of the response (here DMFT2) is modelled as a linear combination of the covariates, it is appropriate to include DMFT1 on the log-scale as well. Note that the purpose of including the DMFT1 index as a covariate can be seen in an attempt to explain the change in the DMFT index through the various intervention measures (and other covariates).

In Table 1 the results are given not only for the conventional Poisson regression and its zero-inflated generalization but also for the more general mixture model which allows mixing on the intercept. It is clear from Table 2 that the major gain in the increase in likelihood

is from the non-inflated to the ZIP model. Note that this model has only one additional parameter, whereas the general mixture model and the ZIP model with school-dependent weights need two and six parameters respectively.

Table 1 provides effect estimates with associated standard errors and Z-values (the estimate of an effect divided by its standard error). As can be seen, neither SEX nor COLOUR plays a relevant role in any of the models. Note that school 3 serves as the control (*base-line category*): school 1 (oral health education), school 2 (all treatments) and school 5 (mouthwash) show significant *negative* effects (changes to a lower DMFT2 value), whereas schools 4 and 6 are non-significant.

The school effects express the adjusted differences in the DMFT2 value to the control school at the end of the study. Clearly, with this kind of modelling it is very important to include the DMFT1 value as the base-line. If base-line values were not considered, the treatment associated with school 6 would be falsely assumed to be effective.

Table 2 gives a justification for the ZIP model. When the log(DMFT1)-value is included as a base-line value, there is no difference between the ZIP model and the next higher mixture model with two free components. This again underlines the importance of the ZIP distribution as a more general error distribution for a count variable in a generalized linear model. The last row in Table 2 considers the ZIP regression model (6.3) where for each school a different proportion of zero inflation is allowed. Though there is an improvement in the log-likelihood, it is not significant on a χ^2-scale with 6 degrees of freedom (see also Fig. 3 for comparison).

8. Discussion

Using the example of the data of the BELCAP study, it has been demonstrated that the ZIP model is useful in describing the DMFT index which is used to evaluate the effects of the various prevention programmes of the BELCAP study.

8.1. Belo Horizonte caries prevention study

When the preventive effects of the five programmes were evaluated, from the *univariate* analysis it was seen that the strongest results were achieved in school 1 (oral health education), school 5 (mouthwash with 0.2% NaF solution) and school 2 (all four programmes together). The programmes performed in school 4 (enrichment of the school diet with rice bran) and school 6 (oral hygiene) did not show preventive effects.

So far no caries epidemiological study has been carried out in the Belo Horizonte metropolitan area. This implies that no comparison of the results found in the BELCAP study with other data from local surveys is possible. Comparing the values of the DMFT index from 7- and 10-year-old children in the BELCAP study with international data, it becomes evident that the values in Belo Horizonte are very high. Another important effect that was detected in the BELCAP study was the reduction in initial dental caries on the enamel surface of teeth, which was present even in the control group. There are two possible explanations for this. One possibility is a trend in dental caries that has affected all the schools in the BELCAP study in a similar way. However, it could be that during the study, especially while the intervention phase was in progress, information from one school to another could have been passed over (the *spillover effect*). Frequently meetings were held between the co-ordinators of the BELCAP study and the heads of the schools, to discuss matters concerning the execution of the programmes. So, in this case a spillover effect cannot be completely excluded.

However, this also raises the question why in this study only the schools were randomized and *not* the individual children. Individual randomization would have created a large potential for biases including bias due to non-blinding of treatments and the spillover effect mentioned. However, in epidemiology non-randomized studies are not untypical and the only option is to adjust for imbalance with statistical tools as has been done in the BELCAP study. Also, in the BELCAP study the schools were randomized to the treatments. Thus, this kind of study falls under the category of *community randomized trials*, in which a treatment is applied to more than one person as part of the experimental unit (Piantadosi (1997), p. 68). These trials are gaining popularity, recently, for example, in the evaluation of health education programmes (see also Piantadosi (1997) and Gail *et al.* (1996)). However, it must be kept in mind that the conclusions drawn from this type of non-individual randomized interventional study are dependent on the statistical model used. Consequently, the results must be interpreted with caution.

It remains unclear why the various schools had such different initial values at the beginning of the study. We might imagine different socioeconomic backgrounds in these schools. However, the study was performed with urban school-children belonging to six distinct local government schools, all in the same socioeconomic unit region. The clinical assessment was carried out in the yard of each school in daylight by a research team consisting of two assistant professors from the dental school, two assistants (students from the dental school) and two associated research fellows (dentists). The examination took on average 10 minutes to be performed. The oral examination was carried out with the child lying on a school table. For this procedure a plain dental mirror number 4 and an explorer number 5 were used. Attention was paid to the accuracy of data entry and checking was carried out.

The sample of the BELCAP study originated from a low socioeconomic class and shows a high prevalence of dental caries. In Brazil refined sugar as well as products containing fermentable carbohydrates are very cheap and are used by low income families. This would explain partially why such a high prevalence of caries is found in this population. The results of the BELCAP study show also that the programme of oral health education proved to be the best method to prevent initial dental caries in primary molars. In the literature regarding different methods for the prevention of caries until now no study could be found which has used and tested exclusively the pedagogical aspects of this programme. The material published on this research area has been used for the implementation of oral health programmes along with the pedagogical instruments and fluoride and tooth brushing. Therefore a comparison is not possible. To understand in its complexity the mechanism of action of the programme of oral health education further studies should be designed. This would allow the validation of the results presented here. In this case the following points may be considered in further studies:

(a) whether a reduction in sugar consumed outside the school has taken place;
(b) whether a change in the frequency of oral hygiene outside the school has taken place;
(c) whether the parents have increased their knowledge of and changed their opinion about sugar consumption and oral hygiene.

The mouthwash programme with 0.2% NaF solution in comparison with the control group also showed a significant reduction in initial dental caries lesions in primary molars. This agrees closely with the hypotheses of Disney *et al.* (1990), who suggested that only the use of popular fluoride products such as dentifrice, gel and varnish can lead to a significant prevention of dental caries. The reduction achieved by following this programme as described by Ripa *et al.* (1983) could be partly confirmed in the BELCAP study.

8.2. Statistical modelling

It has been seen that ZIP models are very special *mixture* or *latent class* models which can be used in a variety of applications in which an extra proportion of zeros occur. It has been demonstrated that there are ways which provide maximum likelihood estimators reliably. Since ZIP models are special (Poisson) mixtures, software for mixture modelling such as C.A.MAN (Böhning *et al.*, 1992; Böhning, 1995) might be used for fitting ZIP models. The ZIP model might be tested by means of the likelihood ratio test. However, the classical asymptotic result of a χ^2-distribution with 1 degree of freedom is *not* valid here since the boundary condition is violated. Instead, the limiting distribution follows a two-point mixture of χ^2-distributions with equal weights: $\frac{1}{2}\chi_0^2 + \frac{1}{2}\chi_1^2$ (Self and Liang, 1987; Feng and McCulloch, 1992). An alternative way to test for general overdispersion is a test based on comparing the empirical variance with the empirical mean (Böhning, 1994). In addition, ways to include covariates have been discussed. The approach followed in Section 6 assumes that the amount of zero inflation does not depend on the values of the covariates. It appears appropriate to validate this assumption. If there is a variation in p, a more complete modelling can be adopted (Lambert, 1992) and possibilities have been outlined. Suggestions have been made to the authors on various occasions (including during the reviewing process) to consider nonparametric methods as an alternative analysis instrument. However, if the amount of zero inflation is above 50%, then the median is necessarily 0. Thus a procedure based on ranks will not detect any changes in the second component, in this case. As pointed out by one reviewer, alternatively a zero-inflated binomial model, e.g. $(1-p)\,\mathrm{bin}(y, 0, N) + p\,\mathrm{bin}(y, \mu, N)$, could have been considered. For the DMFT index in the BELCAP study, N is the number of teeth that could be become decayed, missing or filled, e.g. $N = 8$. This is a valid alternative. In fact, we can consider the ZIP as a (good) approximation to the zero-inflated binary model. Rather similar results with respect to the Z-values of the school effects have been obtained.

Acknowledgements

The research of Dr Böhning and Dr Dietz is currently supported by the German Research Foundation. Lisette Mendonça's research is carried out under a scholarship from the German Academic Exchange Service. Dr Böhning would like to thank the Department of Statistics, PennState University, in particular Jim Rosenberger, Marllyn Boswell, Cliff Clogg (who unexpectedly died in 1995) and Bruce Lindsay for making a research visit possible, so that this work could be finalized. Thanks go also to Gerhard Arminger for various helpful discussions. The authors are grateful to the Associate Editor and two referees for their helpful comments.

Appendix A: Estimating the model parameters

A.1. Moment estimation

From equation (3.4) we have the moment equations $E(Y) = \bar{Y}$ and $S^2 = E(Y)\{1 + \mu - E(Y)\}$ which are readily solved by

$$\hat{\mu}_{\mathrm{MO}} = S^2/\bar{Y} - 1 + \bar{Y}$$

and

$$\hat{p}_{\mathrm{MO}} = \bar{Y}/\hat{\mu}_{\mathrm{MO}}.$$

A.2. Maximum likelihood estimation

Let n_i be the number of is in the sample; in particular, n_0 is the number of zeros in the sample. Then the log-likelihood function is

$$L(p, \mu) = n_0 \log\{1 - p + p\exp(-\mu)\} + \sum_{y=1}^{m} n_y \log\{p\,\mathrm{Po}(y, \mu)\} \tag{A.1}$$

and the *score vector* is

$$\left(n_0 \frac{\exp(-\mu)}{1 - p + p\exp(-\mu)} + \frac{n - n_0}{p},\ n_0 \frac{-p\exp(-\mu)}{1 - p + p\exp(-\mu)} - (n - n_0) + n\bar{y}\mu \right)^{\mathrm{T}},$$

leading to the *score equations*

$$p = \frac{1 - n_0/n}{1 - \exp(-\mu)},$$

$$\mu = \bar{y}/p,$$

which can be written in one equation as

$$\mu = \bar{y}\left\{ \frac{1 - n_0/n}{1 - \exp(-\mu)} \right\}^{-1} =: G(\mu).$$

Because

$$\frac{\mathrm{d}}{\mathrm{d}\mu} G(\mu) = \frac{\bar{y}}{1 - n_0/n} \exp(-\mu) > 0,$$

$\mu_{j+1} = G(\mu_j)$ converges for any initial value μ_0 to the maximum likelihood estimate $\hat{\mu}_{\mathrm{MLE}}$ satisfying the fixed point equation $\mu = G(\mu)$. As the initial value for iteration, $\mu_0 = \hat{\mu}_{\mathrm{MO}}$ could be chosen. The convergence of this algorithm is usually linear and acceleration procedures exist (Böhning, 1993). A problem with this algorithm is that it does not take the parameter restriction $0 < p < 1$ into consideration, and this procedure can lead to an estimated p-value that is larger than 1. In that case, we must set $p = 1$ and use the usual mean as an estimate of μ. Alternatively, we can use a simplified version of the EM algorithm described in Appendix B. This EM algorithm always leads to an estimate which fulfils the parameter restriction on p.

The *sample information matrix*

$$I = I(p, \mu) = -\frac{\partial^2}{\partial p \partial \mu} L(p, \mu)$$

can be found to be

$$\begin{pmatrix} n_0\{\exp(-\mu) - 1\}^2/C & n_0\exp(-\mu)/C \\ n_0\exp(-\mu)/C & n\bar{y}/\mu^2 - n_0 p(1 - p)\exp(-\mu)/C \end{pmatrix} \tag{A.2}$$

with $C = C(p, \mu) = \{1 - p + p\exp(-\mu)\}^2$. From the inverse of matrix (A.2) asymptotic standard errors can be constructed in the conventional way.

Appendix B: Maximum likelihood estimation for the zero-inflated Poisson model with covariates

Let $(y_i, \mathbf{x}_i)$ be a random sample of size n with counts y_i and a vector $\mathbf{x}_i$ of covariates for $i = 1, \ldots, n$. Then, the log-likelihood function for model (6.1) is

$$L(p, \alpha, \boldsymbol{\beta}) = \sum_{i=1}^{n} \log[(1 - p)\,\mathrm{Po}(y_i, 0) + p\,\mathrm{Po}\{y_i, \exp(\alpha + \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}_i)\}].$$

The 'complete-data likelihood' is given by

$$\prod_{i|y_i=0} (1-p)^{1-z_i}[p\,\mathrm{Po}\{y_i,\ \exp(\alpha+\beta^{\mathrm{T}}\mathbf{x}_i)\}]^{z_i} \prod_{i|y_i>0} [p\,\mathrm{Po}\{y_i,\ \exp(\alpha+\beta^{\mathrm{T}}\mathbf{x}_i)\}]^{z_i}$$

where z_i is an *unobserved* binary variable indicating whether sample unit $(y_i, \mathbf{x}_i)$ comes from latent class *zero* ($z_i = 0$) or *non-zero* ($z_i = 1$). Note that $\mathrm{Po}(y, 0) = 0$ if $y > 0$ and z is known to be equal to 1 in this case. The complete-data log-likelihood is given by

$$L^{\mathrm{C}}(p, \alpha, \beta) = \sum_{i=1}^{n} z_i \log(p) + (1-z_i)\log(1-p) + z_i \log[\mathrm{Po}\{y_i,\ \exp(\alpha+\beta^{\mathrm{T}}\mathbf{x}_i)\}].$$

The EM algorithm proceeds in the usual way to calculate the maximum likelihood estimates of p, α and β as follows. Let $p^{(m)}$, $\alpha^{(m)}$ and $\beta^{(m)}$ be the iterates achieved at the mth M-step. Then, in the $(m+1)$th E-step, $E\{L^{\mathrm{C}}(p^{(m)}, \alpha^{(m)}, \beta^{(m)})\}$ must be computed. This leads to the computation of the expected values of z_i:

$$E(z_i|p^{(m)}, \alpha^{(m)}, \beta^{(m)}, y_i, \mathbf{x}_i) = w_i = p^{(m)} \frac{\mathrm{Po}\{y_i,\ \exp(\alpha^{(m)}+\beta^{(m)\mathrm{T}}\mathbf{x}_i)\}}{p^{(m)}\,\mathrm{Po}\{y_i,\ \exp(\alpha^{(m)}+\beta^{(m)\mathrm{T}}\mathbf{x}_i)\} + (1-p^{(m)})\,\mathrm{Po}(y_i, 0)}.$$

Note that $w_i = 1$, if $y_i \geqslant 0$. In the M-step we must find those values of p, α and β which maximize

$$E(L^{\mathrm{C}}) = \sum_{i=1}^{n} w_i \log(p) + (1-w_i)\log(1-p) + w_i \log[\mathrm{Po}\{y_i,\ \exp(\alpha+\beta^{\mathrm{T}}\mathbf{x}_i)\}]. \tag{B.1}$$

Here, the nice feature is that the first term of equation (B.1) involves only p, whereas the second term is independent of p. Maximizing the first term we find that $p^{(m+1)} = (w_1 + \ldots + w_n)/n = \bar{w}$. The maximizing of α and β cannot be given in closed form. However, we observe that they can easily be found by a software package which contains Poisson regression and allows the specification of case weights (the w_i in this case).

If also p is allowed to depend on covariates (by a logistic link, for example), the respective additional model parameters can also be computed by a GLIM fit where the w_i are considered as response variables. To be more specific, in the M-step for the parameters α' and β' in model (6.2) the following term must be maximized:

$$\sum_{i=1}^{n} w_i \log\{p(\alpha'+\beta'^{\mathrm{T}}\mathbf{x}_i')\} + (1-w_i)\log\{1-p(\alpha'+\beta'^{\mathrm{T}}\mathbf{x}_i')\}. \tag{B.2}$$

To solve the M-step in the simplified model (6.3) we must find those values for the p_js which maximize

$$\sum_{i=1}^{n} \{w_i \log(p_j) + (1-w_i)\log(1-p_j)\}\delta_{ij},$$

where $\delta_{ij} = 1$ if the ith data point comes from the jth stratum and $\delta_{ij} = 0$ otherwise.

These values are easily derived as $p_j^{(m+1)} = (w_1\delta_{1j} + \ldots + w_n\delta_{nj})/n_j$ where n_j denotes the size of the jth stratum. A GLIM macro to do these computations can easily be constructed or obtained from the authors.

References

Aitkin, M., Anderson, D., Francis, B. and Hinde, J. (1989) *Statistical Modelling in GLIM*. Oxford: Clarendon.

Böhning, D. (1982) Convergence of Simar's algorithm for finding the maximum likelihood estimate of a compound Poisson process. *Ann. Statist.*, **10**, 1006–1008.

——— (1993) Acceleration techniques in fixed-point methods for finding percentage points. *Statist. Comput.*, **3**, 1–5.

——— (1994) A note on test for Poisson overdispersion. *Biometrika*, **81**, 418–419.

——— (1995) A review of reliable maximum likelihood algorithms for the semi-parametric mixture maximum likelihood estimator. *J. Statist. Planng Inf.*, **47**, 5–28.

Böhning, D., Schlattmann, P. and Lindsay, B. G. (1992) Computer assisted analysis of mixtures (C.A.MAN): statistical algorithms. *Biometrics*, **48**, 283–303.

Breslow, N. E. (1984) Extra-Poisson variation in log-linear models. *Appl. Statist.*, **33**, 38–44.

Breslow, N. E. and Day, N. E. (1985) The standardized mortality ratio. In *Biostatistics: Statistics in Biomedical, Public Health and Environmental Sciences* (ed. P. K. Sen), pp. 55–74. Amsterdam: North-Holland.

van den Broek, J. (1995) A score test for zero inflation in a Poisson distribution. *Biometrics*, **51**, 731–743.

Campbell, M. J., Machin, D. and D'Arcangues, C. (1991) Coping with extra-Poisson variability in the analysis of factors influencing vaginal ring expulsions. *Statist. Med.*, **10**, 241–251.

Demétrio, C. G. B. and Ridout, M. S. (1994) Letter to the Editor: Coping with extra-Poisson variability in the analysis of factors influencing vaginal ring expulsions. *Statist. Med.*, **13**, 873–876.

Dietz, E. (1992) Estimation of heterogeneity—a GLM-approach. *Lect. Notes Statist.*, **78**, 66–72.

Dietz, E. and Böhning, D. (1995) Statistical inference based on a general model of unobserved heterogeneity. *Lect. Notes Statist.*, **104**.

Disney, J. A., Bohannan, H. M., Klein, S. P. and Bell, R. M. (1990) A case study in contesting the conventional wisdom: school-based fluoride mouthrinse programmes in the USA. *Commty Dent. Oral Epidem.*, **18**, 46–56.

Feng, Z. and McCulloch, C. (1992) Statistical inference using maximum likelihood estimation and the generalized likelihood ratio when the true parameter is on the boundary of the parameter space. *Statist. Probab. Lett.*, **13**, 325–332.

Fong, D. Y. T. and Yip, P. (1993) An EM algorithm for a mixture model of count data. *Statist. Probab. Lett.*, **17**, 53–60.

Gail, M. H., Mark, S. D., Carroll, R. J., Green, S. B. and Pee, D. (1996) On design considerations and randomization based inference for community intervention trials. *Statist. Med.*, **15**, 1069–1092.

Hogg, R. V. and Tanis, E. A. (1988) *Probability and Statistical Inference*, 3rd edn. New York: Macmillan.

Johnson, N., Kotz, S. and Kemp, A. W. (1992) *Univariate Discrete Distributions*, 2nd edn. New York: Wiley.

Laird, N. M. (1978) Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Statist. Ass.*, **73**, 805–811.

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

Lindsay, B. G. (1983) The geometry of mixture likelihoods: a general theory. *Ann. Statist.*, **11**, 86–94.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Mendonça, L. (1995) Longitudinalstudie zu kariespräventiven Methoden, durchgeführt bei 7- bis 10-jährigen urbanen Kindern in Belo Horizonte (Brasilien). *Dissertation*. Free University of Berlin, Berlin.

Mendonça, L. and Böhning, D. (1994) Die Auswirkung von Gesundheitsunterricht und Mundspülung mit Na-Fluorid auf die Prävention von Zahnkaries: eine Kohortenstudie mit urbanen Kindern in Brasilien. *39th A. Conf. Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie, Dresden, September 18th–25th.*

Piantadosi, S. (1997) *Clinical Trials: a Methodologic Perspective*. New York: Wiley.

Pilz, M. E. W. (1985) *Praxis der Zahnerhaltung und Oralen Prävention*. Munich: Hanser.

Ripa, L. W., Leske, G. S., Sposato, A. L. and Rebich, T. (1983) Supervised weekly rinsing with a 0.2% neutral NaF solution results after 5 years. *Commty Dent. Oral Epidem.*, **11**, 1.

Schlattmann, P. and Böhning, D. (1993) Mixture models and disease mapping. *Statist. Med.*, **12**, 1943–1950.

Self, S. and Liang, K. Y. (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Statist. Ass.*, **82**, 605–610.

Simar, L. (1976) Maximum likelihood estimation of a compound Poisson process. *Ann. Statist.*, **4**, 1200–1209.

Yip, P. (1991) Conditional inference on a mixture model for the analysis of count data. *Communs Statist. Theory Meth.*, **20**, 2045–2057.