Zero-Inflated Poisson Models and C.A.MAN: 
A Tutorial Collection of Evidence

DANKMAR BÖHNING

Department of Epidemiology
Institute for Social Medicine and Medical Psychology
Free University of Berlin
Germany

Summary

Analysis of count data is required in many areas of biometric interest. Often the simple Poisson distribution is not appropriate, since an extra-number of zero counts occur in the count data. Some current approaches for the problem at hand are reviewed. It will be argued that these situations can often be easily modeled using the zero-inflated Poisson distribution. A variety of applications are considered in which this occurs. Possibilities are outlined on how the validity of the zero-inflated Poisson can be validated including a comparison with the nonparametric Poisson mixture maximum likelihood estimator.

Key words: Zero-inflated poisson; C.A.MAN; Count data analysis.

1. Introduction

In a recent issue, BROEK (1995) analyzed data on 98 HIV-infected men attending the Department of Internal Medicine at the Utrecht University Hospital. The data consisted of the number of times these men had an urinary tract infection as seen in Figure 1. Broek suggested modeling these data with a zero-inflated Poisson (ZIP) model, namely \( f(y; \lambda, p) = p + (1 - p) \text{Po}(y, \lambda), \) if \( y = 0, \) and \( f(y; \lambda, p) = (1 - p) \text{Po}(y, \lambda), \) if \( y > 0; \) Po denotes the Poisson density.

The excellent fit of the model is also visualized in Figure 1. The \( \chi^2 \)-value is \( \chi^2 = 1.3723 \) corresponding to a p-value of 0.2414. The \( \chi^2 \)-value for the single Poisson is \( \chi^2 = 16.135 \) with corresponding p-value 0.0003, indicating a strong lack-of-fit for the single Poisson distribution.\(^1\) Let \( y_1, \ldots, y_n \) denote the random sample and let \( x_1, \ldots, x_m \) the m distinct elements of the sample with frequencies \( n(x_1), \ldots, n(x_m). \) Also, let \( f(y) \) denote the discrete density under consideration.

\(^1\) Inspection of the Pearson residuals shows that most of the contribution to the \( \chi^2 \)-value comes from cell 1 and 2 (and not from the 0-cell!). This is due to the fact that the heavy weight on the 0-cell forces the mean strongly to the left, thus leading to a reasonable fit in the 0-cell.
Then, the Pearson $\chi^2$ is given (and used here throughout the paper) as $\chi^2 = \sum_i (n(x_i) - nf(x_i))^2 / nf(x_i)$, where the summation is from 1 to $m$. Degrees of freedom are computed conventionally as $(m - 1 - \text{number of parameters estimated})$.

Zero-inflation occurs rather frequently when dealing with count data. Before we give an illustration and collection of ZIP-examples of data sets in section 3, we present some broader introduction into count data modeling in section 2.

2. ZIP Modeling and C.A.MAN

One could think of the zero-inflated Poisson model as the special mixture model

$$p \Po(y, 0) + (1 - p) \Po(y, \lambda), \quad y = 0, 1, 2, \ldots$$

with a fixed first component mean 0. Note that the first component $\Po(y, 0)$ is a one-point distribution with all it’s mass at 0. In fact, this model could be considered as a special case of a wide class of mixture models

$$p_1 \Po(y, \lambda_1) + \ldots + p_k \Po(y, \lambda_k), \quad y = 0, 1, 2, \ldots$$

with $p_j \geq 0$ and $p_1 + \ldots + p_k = 1$. The number of components is considered to be an unknown parameter. The discrete distribution $P$ giving mass $p_1$ to $\lambda_1, \ldots, p_k$ to $\lambda_k$ is called the nonparametric mixing distribution and it’s maximum likelihood estimator (including an estimate of the number of components $k$) can be computed easily with the statistical package C.A.MAN\(^2\) for mixture analysis (BÖHNING, SCHLATTMANN, and LINDSAY, 1992; BÖHNING, DIETZ, and SCHLATTMANN, 1998). For the HIV-data the estimate of $P$ is a distribution with $k = 2$ components, with first mean $\lambda_1 = 0$ receiving weight $\hat{p}_1 = 0.7116$ and second mean $\lambda_2 = 0.9198$ receiving weight $\hat{p}_2 = 0.2884$. The log-likelihood is $-61.0214$. The log-likelihood

\(^2\) C.A.MAN is the abbreviation for Computer Assisted Analysis of Mixtures.
for the single Poisson is $-67.14243$, leading to a value of $2 \ln (\xi) = 2(-61.0214 + 67.14243) = 12.24206$ in the log-likelihood ratio. According to Self and Liang (1987) the asymptotic distribution of $2 \ln (\xi)$ is a mixture of $\chi^2$-distributions: $\frac{1}{2} \chi_0^2 + \frac{1}{2} \chi_1^2$, leading to the simple formula for the $\text{P-value}$: $\frac{1}{2} [1 - F_\chi(12.24206)] = 0.00023$; where $F_\chi$ is the distribution function of the $\chi^2$ with 1 df. In other words, the conventionally determined $\text{P-value}$ (using a $\chi^2$ with 1df) needs to multiplied only by the factor $\frac{1}{2}$. See also Feng and McCulloch (1992) for further discussion on this point.

In more generality, one might be interested in finding the mixing distribution $P$ which maximizes the log-likelihood $L(P) = \sum_i \ln \text{Po}(y_i; P)$ in the class of all discrete probability distributions, where $\text{Po}(y, P) = p_1 \text{Po}(y, \lambda_1) + \ldots + p_k \text{Po}(y, \lambda_k)$. Note that the number of components $k$ is allowed to vary. It is straightforward to show that $L$ is a concave functional in the set of all discrete probability distributions. Also, the number of components $k$ is always bounded above (for details see Lindsay (1995) or Böhning (1998)). This concavity property allows the construction of a powerful theory, from which we want to highlight only two aspects.

For one, it can be easily verified that this $\hat{P}$ is the nonparametric mixture maximum likelihood estimator (NPMLLE) by using the general mixture maximum likelihood theorem (Lindsay, 1995) saying that $\hat{P}$ is NPMLLE if and only if the gradient function $D_P(\lambda) = \frac{1}{n} \sum \text{Po}(y_i, \lambda)/\text{Po}(y_i, \hat{P}) \leq 1$ for all $\lambda = 0$. Similar theorems have been derived before in the area of optimal design (Atkinson and Donev (1992); Fedorov (1972); see also Böhning (1995) for detailed outline of this link). Here $\text{Po}(y, \hat{P}) = \hat{p}_1 \text{Po}(y, \hat{\lambda}_1) + \ldots + \hat{p}_k \text{Po}(y, \hat{\lambda}_k)$, the marginal likelihood for observation $y$. It might be easily numerically verified for the HIV-data from section 1 that this condition is fulfilled for $\hat{P}$, giving weight $\hat{p}_1 = 0.7116$ to 0 and giving weight $\hat{p}_2 = 0.2884$ to the second mean $\hat{\lambda}_2 = 0.9198$.

For two, though the general mixture maximum likelihood theorem is very useful in determining if a given $\hat{P}$ is the NPMLLE, it is less useful in constructing the NPMLLE. To accomplish this task algorithms have to be applied. Meanwhile, a variety of globally converging algorithms have been made available and implemented in C.A.MAN including the vertex-direction-method (Wu, 1978), the vertex-exchange method (Böhning, 1985) and a version of the EM-algorithm (Laird, 1978) to mention a few. For a detailed review of algorithmic aspects the interested reader is referred to Böhning (1995, 1998).

We have seen that the C.A.MAN-approach is nonparametric in terms of the mixing distribution. Thus, there is no other mixture model of Poisson distributions leading to a likelihood higher than the one of the zero-inflated Poisson model. If the NPMLLE turns out to be a ZIP-model, it can therefore be validly concluded that it is the best (in the likelihood sense) mixture model. Thus, C.A.MAN might be used as a diagnostic device for ZIP-Models.

Nevertheless, alternative approaches are popular and well-known in the literature. Westermeier and Michaelis (1995) point out that the current alternatives
to the simple Poisson distribution are the families of *compound* or *generalized* Poisson distributions. The compounding process is identically to mixing homogeneous Poisson distributions, though typically one thinks of mixing the Poisson parameter $\lambda$ with some continuous, often parametric density $p(\lambda)$: $\int_0^\infty \text{Po}(y, \lambda) p(\lambda) \, d\lambda$. A prominent example is the mixing with a Gamma-distribution, since the marginal distribution $\text{Po}(y, p)$ is the negative binomial distribution. A result now increasingly used by epidemiologists (see for example Martuzzi and Hills, 1995) to adjust for unobserved population heterogeneity. All parametric mixture distributions, however, are somewhat inferior to the nonparametric approach, since it is guaranteed that their likelihood is at most as large as the nonparametric mixture likelihood. Even if the continuous, parametric mixture density were the true density (which will seldom be possible to be verified), the NPMLE will give a reasonable approximation, since it is a consistent estimator of the mixing distribution (Kiefer and Wolfowitz, 1956). In addition, it is argued sometimes that the quantity $\lambda$ is desired to be of continuous character; even if this were the case one can still use the discrete NPMLE as reasonable approximation, just as one is taking the empirical (discrete) distribution function as an estimate of a continuous, theoretical distribution function.

Another class are the generalized Poisson distributions. Following again Westermeyer and Michaelis (1995) they arise as the sum of $N$ independent and identically distributed variables $Y_i, i = 1, \ldots, N$ where $N$ is distributed according to a Poisson distribution. A prominent example is the Neyman-type-A-distribution which arises if all $Y_i, i = 1, \ldots, N$ are Poisson distributions, in addition. Lautenschläger (1989) studies these distributions in detail. It should be pointed out that all of these distributions share the characteristic that their variance is at least as large as their expected value, a property called *overdispersion*. In turn, as a diagnostic device *overdispersion tests* use the fact that under the simple Poisson distribution, mean and variance coincide. A prominent test statistic is $T = (S^2 - \bar{Y}) / \{(2\bar{Y} / (n - 1))\}^{1/2}$, which is under the hypothesis of a homogeneous Poisson asymptotically standard normal (Boëhning, 1994). If, however, such a test is leading to a significant result, constructive alternatives are required and the ZIP-model is one of those, having often a plausible interpretation on its side.

In the next section we provide a collection of potential applications of the ZIP-model from various sciences.

3. A Variety of Biometric Applications with the Occurrence of Zero-Inflation

In this section we collected a number of examples from which we think that the ZIP-model is providing a reasonable alternative distributional model.

*Traffic Accident Research.* In traffic accident research a variable of interest is often the number of accidents per driver. A possible motivation can be seen in the
possibility of finding risk factors involved in the accidents. Kuan et al. (1991), as one example consider this variable at data coming from the California Department of Motor Vehicles master driver license file. The authors observed that the simple Poisson model is not fitting well. Different modelling approaches were investigated. Here, we consider a data set of Thyrion (1961) which occasionally appears in the literature.

**Accident research.** A second data set consisted of the number of accidents of 647 female workers in an ammunition factory (following Greenwood and Yule, 1920).

For the ZI-Poisson we find a $\chi^2$-value is $\chi^2_{(3)} = 7.838$ corresponding to a p-value of 0.0495. The $\chi^2$-value for the single Poisson is $\chi^2_{(4)} = 115.35$ with corresponding p-value $<0.00001$, indicating a strong lack-of-fit for the single Poisson distribution.

**Crime and Deviating Behavior.** Dieckmann (1981) provided a data set from crime sociology consisting of a sample of people with deviating behavior. $Y$ is here the number of criminal acts.

Dieckmann (1981) pointed out that the simple Poisson model with $\chi^2_{(4)} = 47135$ does not fit the data well. If the ZIP-model is fitted a drastic improvement in the goodness-of-fit with $\chi^2_{(3)} = 62.596$ can be found. However, the
fit of the ZIP-model is not too good in the upper classes (number of criminal acts 3, 4, and 5).

**Rehabilitation Data.** The following data (see Figure 5) refer to Stallmann (1994) and were collected by the state insurance company responsible for regulation of costs concerning rehabilitation measures. $Y$ is the number of episodes of working inability three years before a rehabilitation measure was performed.

The simple Poisson model with $\chi^2 = 51.97$ does not fit the data well. If the ZIP-model is fitted a clear improvement in the goodness-of-fit with $\chi^2 = 4.96$ (P-value 0.2914) can be found.

**Psychological Scale Data.** Often in psychiatric diagnosis a scale is used consisting out of number of binary items is used. A typical example is the obsessive compulsive neurosis syndrome (Wilson et al. 1991, p. 2134) in geriatric research. In Figure 6, a sample of 2645 elderly patients is shown in their distribution of positive item responses. The simple Poisson has a $\chi^2$-value of 8,822,035,456 indicating a very bad fit, whereas the ZIP-model gives a $\chi^2$-value of 434.36. It should be noted that the lack-of-fit in the ZIP-model is due to the sparsity of the data in the last cell.
Standardized Mortality Ratio in Geographic Epidemiology. In a recent publication, Martuzzi and Hills (1995) discussed questions of heterogeneity in the true standardized mortality ratios of perinatal mortality and suggested to model this heterogeneity by using using a Gamma-distribution, while the distribution of the observed perinatal death is assumed to be a Poisson distribution. The problem of heterogeneity (spatial variation) in geographic epidemiology is of prime interest, since it might indicate that certain areas are at higher risk than others, if heterogeneity is present. This is the target of many disease atlases (usually mortality) which are now available for many diseases (cancer, infectious diseases) and for many countries. Martuzzi and Hills (1995) considered a data set on the geographic distribution of perinatal mortality in the North West Thames Health Region, England, in the period 1986–1990 on the basis of 515 small areas as units for the statistical analysis. The distribution of the SMR is shown in Figure 7. For mortality data of SMR-type the Poisson distribution is frequently used. Let $E_i$ be the expected number of death for the $i$th area calculated from an external reference population and $\lambda$ the population SMR. Often it is assumed that the observed number of death follow the well-known Poisson distribution, that is $\Pr [O_i = o_i] = \exp(-\lambda E_i)(\lambda E_i)^{o_i}/o_i!^3$. Note that the expected number of deaths in the $i$th county are incorporated into the Poisson parameter. Consequently the ZIP-model takes the form

$$f(o_i, E_i, \lambda, p) = (1 - p) \text{Po}(o_i, 0) + p \text{Po}(o_i, \lambda E_i),$$

$$o_i = 0, 1, \ldots, i = 1, \ldots, n.$$  

The simple Poisson model with $\chi^2_{(11)} = 285.84$ does not fit the data well. Again, if the ZIP-model is fitted a clear improvement in the goodness-of-fit with $\chi^2_{(10)} = 13.89$ (P-value 0.3816) can be found.

3 This form of Poisson density can also be used for binomial rate data in which $o_i = y_i$ is the number of events and $E_i = n_i$ is the number at risk (assuming small event probability).
The point of many zeros in large cancer data sets such as those coming from cancer registries has also been noted by Westermeier and Michaelis (1995) as they write: The large proportion of municipalities without any observed malignancy is one of the major characteristics of the data (of the German Children’s Cancer Registry, insertion by D.B.), i.e. 61% for all malignancies, 79% for leukemia’s and 76% for leukemia’s and lymphomas (p. 8). It can therefore be argued that the ZIP-model might be a useful alternative to the simple Poisson for many cancer registry data.

4. Discussion

Is every count distribution with a lot of zero counts a ZIP model? Although there might be a high percentage of zero counts in the sample, this is not necessarily an indication for a ZIP model. Consider the following sample data of size 100: \( n_0 = 65, n_1 = 33, n_2 = 1, n_3 = 1 \). They have been generated from a homogeneous Poisson with \( \mu = 0.5 \). This points out that a more complete statistical analysis of the question is required and ways to do so have been outlined above.

How does the ZIP-model relate to other (parametric) approaches? The zero-inflated Poisson models a very special form of heterogeneity. The general nonparametric form of the mixture model can be thought of as a model for unobserved heterogeneity (consisting of various subpopulations), which, at least in principle, would vanish if subpopulation membership is observed. The ZIP-model, however, remains a necessity for many of the presented data sets even if a variety of covariates is added. Thus, the ZIP-model is more like a distributional characteristic of the count variable under consideration than is standing for a certain unobserved covariate. In some cases, it seems appropriate to allow a more general definition of the ZIP-model. Besides the zero-component, there might be further components...
necessary to achieve a good fitting model. We call \( f(y, P) \) a *generalized ZIP-model* if the density can be written

\[
f(y, P) = p_0 \text{Po}(y, 0) + p_1 \text{Po}(y, \lambda_1) + \ldots + p_k \text{Po}(y, \lambda_k),
\]

\( y = 0, 1, 2, \ldots \)

with \( p_j \geq 0 \) and \( p_0 + \ldots + p_k = 1 \). Note that the generalized ZIP-model has exactly \( 2k \) free parameters \( P = \begin{pmatrix} 0 & \lambda_1 & \ldots & \lambda_k \\ p_0 & p_1 & \ldots & p_k \end{pmatrix} \), whereas the conventional mixture model has \( 2k + 1 \) free parameters (one could even say that the generalized ZIP-model is filling the gap between a mixture model with \( k \) and \( k + 1 \) components).

To demonstrate the point we consider again the Dieckmann-data on criminal acts. It turns out that the NPMLE is a generalized ZIP model with \( k^* = 2 \) components:

\[
p_0^* = 0.5609, \quad p_1^* = 0.4248, \quad \lambda_1 = 0.124, \quad p_2^* = 0.0143, \quad \lambda_2 = 1.744.
\]

The corresponding density \( f(y, \hat{P}) \) is shown in Figure 4.

**How can the optimality of the ZIP model be evaluated?** Clearly, if the NPMLE is itself the ZIP model (a 2-mass point solution with first component mean 0), there is no discussion about which model to take. However, there might be other situations in which, for example in addition to the ZIP model, a third component with (potentially) tiny mass occurs. Here, one could consider the likelihood ratio \( 2\{L(\hat{P}) - L(\hat{\lambda}, \hat{p})\} \) as a goodness-of-fit criterion, although care must be taken again, since the asymptotic distribution does not follow the conventional \( \chi^2 \)-distribution results; a parametric bootstrap might be an exit to find the critical values of the test statistic as discussed in BÖHNING et al. (1994).

**Can the ZIP-model be generalized to other component densities?** The answer is yes. For example, one can consider the zero-inflated Binomial (ZIB)-model:

\[
\text{Bin}(y, N, \lambda, p) = (1 - p) \text{Bin}(y, N, 0) + p \text{Bin}(y, N, \lambda),
\]

where \( \text{Bin}(y, N, \pi) = \binom{N}{y} \pi^y (1 - \pi)^{N - y} \) the Binomial distribution with parameters \( \pi \) and \( N \). Note that \( \text{Bin}(y, N, 0) \) is the “special” Binomial which is 0 for all \( y \), unless \( y = 0 \) where it takes the value 1. A verification of the ZIB-model could be again worked out along the lines of the ZIP-model by using the NPMLE for the mixing distribution of a Binomial distribution.

This ZIB-model might be more appropriate for rate distributions where one is observing \( y \) out of \( N \) possible ones. A typical example would be the DMFT-index used in dental epidemiology. It counts the number of teeth in every person having a decay, a filling or are missing. Obviously, there is a maximum count in the DMFT-index for each person defined by the ever available teeth. Specifically, for child-populations it is documented that there is a large amount of extra-zeros (caries-free children) in these populations (see for this point MENDONÇA (1994) or PILZ (1985)). Here, often the dental epidemiologist is concentrating on the available 8 molars, leading to ZIB-model with \( N = 8 \) for each child.
Though it is possible to think of generalizations of the ZIP-model to other univariate discrete distributions, it is less clear how to proceed in the case of bivariate or multivariate discrete distributional forms (Krummenauer, 1998).

Acknowledgements

The author is grateful to the editor and the two referees for their comments and help in improving the paper. This work was done with support of the German Research Foundation.

References


DANKMAR BÖHNING
Department of Epidemiology
Institute for Social Medicine and Medical Psychology
Free University of Berlin
Fabeckstr. 60–62, Haus 562
14195 Berlin
Germany
E-Mail: boehning@zedat.fu-berlin.de

Received, December 1997
Revised, July 1998
Accepted, August 1998