Extending Zelterman's Approach for Robust Estimation of Population Size to Zero-truncated Clustered Data

W. C. W. Navaratna[1], **Victor J. del Rio Vilas**[2], and **Dankmar Böhning**[*,3]

1 Department of Mathematics and Computer Science, Open University of Sri Lanka, Nawala, Nugegoda, Sri Lanka
2 Food and Farming group, Department for Environment, Food and Rural Affairs (Defra), Nobel House, 17 Smith Square, London, SW1P 3JR, UK
3 Quantitative Biology and Applied Statistics, School of Biological Sciences, Harry Pitt Building, Whiteknights, Reading, RG6 6FN, UK

Received 23 July 2007, revised 1 December 2008, accepted 21 January 2008

Summary

Estimation of population size with missing zero-class is an important problem that is encountered in epidemiological assessment studies. Fitting a Poisson model to the observed data by the method of maximum likelihood and estimation of the population size based on this fit is an approach that has been widely used for this purpose. In practice, however, the Poisson assumption is seldom satisfied. Zelterman (1988) has proposed a robust estimator for unclustered data that works well in a wide class of distributions applicable for count data. In the work presented here, we extend this estimator to clustered data. The estimator requires fitting a zero-truncated homogeneous Poisson model by maximum likelihood and thereby using a Horvitz–Thompson estimator of population size. This was found to work well, when the data follow the hypothesized homogeneous Poisson model. However, when the true distribution deviates from the hypothesized model, the population size was found to be underestimated. In the search of a more robust estimator, we focused on three models that use all clusters with exactly one case, those clusters with exactly two cases and those with exactly three cases to estimate the probability of the zero-class and thereby use data collected on all the clusters in the Horvitz–Thompson estimator of population size. Loss in efficiency associated with gain in robustness was examined based on a simulation study. As a trade-off between gain in robustness and loss in efficiency, the model that uses data collected on clusters with at most three cases to estimate the probability of the zero-class was found to be preferred in general. In applications, we recommend obtaining estimates from all three models and making a choice considering the estimates from the three models, robustness and the loss in efficiency.

Key words: Efficiency; Robustness; Truncated count distribution; Zelterman's estimator; Zero-class

1 Introduction

Consider a population made up of N_T clusters of which an unknown number of N clusters possess a certain characteristic of interest. Suppose an identification device is available for detecting the presence of the characteristic, but it is incomplete in the sense that it may fail to identify the presence of the characteristic in some clusters. Identified clusters are assumed to be correctly identified by the device giving no false positives. We are interested in estimating the total number of clusters, N, that possess the characteristic of interest based on an identified sample of n clusters, where cluster size and the number of units within the cluster possessing the characteristic of interest are observed.

[*] Corresponding author: e-mail: d.a.w.bohning@reading.ac.uk, Phone: +44(0)118 378 6211, Fax: +44(0) 118 975 3169

Table 1 Scrapie data for Great Britain 2004: Del Rio Vilas, V. J. et al. (2005).

Number of scrapie cases	0	1	2	3	4	5	6	7	9	12	18	32
Total number of holdings	–	72	28	14	5	5	4	1	2	2	1	1

Table 2 Cholera data of Mosely et al. (1972).

Cases per village	1	2	3	4	5–9	10 or more
Number of villages	21	20	8	7	11	8

As a typical example, we consider the estimation of the total number of scrapie-affected sheep holdings in Great Britain. Table 1 summarizes the field data for the year 2004. Here, sheep holdings define clusters and there are about $N_T = 70,000$ in Great Britain. The characteristic of interest is whether or not the holding has scrapie-affected animals. We are interested in estimating the size N of those holdings with scrapie. The cluster sizes are presented in Table 3. The identification device here is a passive surveillance stream called the Scrapie Notifications Database (SND) in which clinical cases are reported to the veterinary authorities for confirmatory diagnosis. Prevalence estimates based on the SND need to be adjusted for undercounting. Previously, the under-ascertainment adjusted prevalence has been estimated via anonymous postal surveys (PS) (Hoinville et al., 2000; Sivam et al., 2003). More recently, multiple-list capture-recapture (CRC) methods were applied to estimate the number of scrapie-affected holdings not detected by any of the surveillance streams in place (Del Rio Vilas et al., 2005). In the following we try to develop a methodology for providing an adjustment for disease undercount.

In the example we just described, the frequency of the zero-class is not observed. Mosley et al. (1972) reports another motivating application in epidemiological assessment, where data are available on the zero-class, but are suspected as incomplete due to imperfect surveillance. In their field study carried out in 132 villages in rural east Pakistan, they had recorded the population size and the number of clinically apparent cholera cases in each village. According to the field data (Table 2), 57 out of the 132 villages had no reported cases. However, they further comment that, due to the presence of unreported, potentially mild cases, the epidemiologic picture of cholera is incomplete when only clinically apparent and reported cases are considered. Thus, 57 only serves as an upper bound for the unknown actual number of villages with no cholera-affected individuals. Here the villages serve as clusters and the population size of the village corresponds to the size of the cluster. The purpose of this paper is to develop techniques that can be used in applications such as this for the estimation of the number of villages actually affected by cholera based on the observed cases in each village, taking the population size into account.

Table 3 Cluster size m_i, $i = 1, 2, \ldots, 135$.

2	6	9	10	20	26	31	33	35	35	42	45	48	53	58
58	60	61	61	62	62	63	68	73	76	76	78	78	83	86
90	93	95	98	98	110	112	117	117	122	128	130	131	135	144
146	151	155	155	157	159	184	204	231	241	263	291	296	304	307
310	322	375	494	503	520	673	680	680	780	1136	1577	12	20	29
39	43	46	50	73	85	101	104	120	123	129	141	158	161	168
190	200	210	232	293	333	421	432	530	748	4	22	27	36	57
58	66	90	100	125	131	195	232	873	23	58	70	106	150	14
60	97	98	769	65	128	184	334	1330	95	294	46	121	93	190

The context focused on in this paper falls into the general frame-work of estimating the missing zero-class with truncated data. Many authors have addressed this problem (McKendrick, 1926; Chao, 1987; Zelterman, 1988) for the case of non-clustered data. In fact, the estimation problem addressed in the historic paper by McKendrick was of a clustered type but the clustering was ignored. There are many similar applications, where the data are clustered with cluster size possibly having an impact on the number of infected cases belonging to the cluster. In the case of scrapie, for example, there is support for the hypothesis (McLean et al., 1999) that the larger the holding, the greater the likelihood of having scrapie. Longini and Koopman (1982) have focused on infectious epidemiology data arising in household infections and have devised a model for the distribution of the total number of infected cases in households from a homogeneous community taking the cluster sizes into account. Their model describes the infection pattern of the disease as to infections from the community and infections within the household through the infected members but the computation of estimates become complex as the cluster size increases. This limits the application of their technique to problems such as the estimation of the total number of infected animal holdings. In this paper, we extend the Zelterman's estimator of population size (Zelterman, 1988), which was developed as a robust estimator for unclustered data, to the case of clustered data with possibly different cluster sizes. This allows us to deal with more general applications that arise widely in epidemiological assessment studies.

The paper is organized as follows. In Section 2, we introduce the truncated count distribution for clustered data. In Section 2, we present the models and discuss the estimation of parameters in the truncated count distribution. Estimators for the population size and computation of standard errors are described in Section 3. In Section 4, we present the results of a simulation study that examines the robustness and the efficiency. Some applications of the proposed techniques are illustrated in Section 5.

2 Truncated Count Distribution Modeling for Clustered Data

For ease of reference, here onwards we refer to the units possessing the characteristic of interest as cases. Let m_i denote the size of the i-th cluster and Y_i denote the total number of cases of the i-th cluster. Y_i is considered as a zero-truncated count variable, since clusters are typically identified if there are positive cases in the clusters (In some situations a frequency of clusters with a zero-count of cases is observed, but is ignored since it must be assumed to be contaminated with an unknown number of clusters with positive cases). Let N denote the number of clusters with cases. Suppose we have observations collected on n clusters each of which has cases. As a working model for the untruncated case, assume that for given m_i, $Y_i \sim \text{Pois}(\lambda m_i)$ so that $E(Y_i \mid m_i) = \lambda m_i$ and $P(Y_i = j) = \mathrm{e}^{-\lambda m_i}(\lambda m_i)^j / j!$, where λ is a parameter that represents the expected number of cases in a cluster of unit size.

As proposed in Zelterman (1988), we regard the *observed* data as coming from a zero-truncated count distribution with associated zero-truncated count variable Y_i^+. As a *working model* we consider the *zero-truncated Poisson*

$$P(Y_i^+ = j) = \frac{\mathrm{e}^{-\lambda m_i}}{1 - \mathrm{e}^{-\lambda m_i}} \frac{(\lambda m_i)^j}{j!}, \quad \text{for} \quad i = 1, 2, \ldots, N \quad \text{and} \quad j = 1, 2, \ldots \tag{1}$$

and

$$P(Y_i = 0 \mid \lambda, m_i) = \mathrm{e}^{-\lambda m_i}, \quad \text{for} \quad i = 1, 2, \ldots, N.$$

It is understood that this distribution might not fit well to all parts of the observed data. However, the intention is to develop an estimator of λ which is *robust* to certain misspecifications of the Poisson

model. In particular, we will suggest two distributional alternative models which only require the count distribution to behave *like* a Poisson for counts 1 and 2 (model $\mathcal{S}$) or counts 1, 2 and 3 (model $\mathcal{M}$).

Note that (1) is a specific form of the Poisson distribution in that it accounts for the different cluster sizes where the Y_i^+ arise from. Furthermore, this implies that $E(Y_i^+) = \lambda m_i/(1 - \exp(-\lambda m_i))$, implying that the expected count is proportional to the cluster size. This means, for example in the case of the scrapie surveillance, that the number of scrapie cases is proportional to the holding size, an assumption under debate in scrapie epidemiology. Less critical is the Poisson assumption relative to a binomial distribution in this example, since the number of cases is small relative to the size of the holdings.

We first estimate λ and use that to estimate the frequency of the zero-class of the truncated distribution. In contrast to the disease-free interpretation of the zero-class, $P(Y_i = 0 \mid \lambda, m_i)$ is now the proportion of disease affected clusters with no detected cases.

The first model we consider is the Poisson model where we assume that the observed counts follow the truncated Poisson model presented in the previous section. This model makes no allowance for model misspecification in situations where the unknown true distribution deviates from the specified Poisson model. In the case of non-clustered data, or equivalently, when all m_i's are equal to 1, observations are identically distributed and this model is referred to as the homogeneous Poisson model. In our case, with possibly different cluster sizes, the observed counts are not identically distributed. However, the model predicts no heterogeneity across clusters of the same size. Therefore, here onwards we refer to this model as the homogeneous Poisson model in our context.

Many authors have described applications (see Böhning et al., 1999; Lowe, 1999; Griffiths, 1973) where the number of zeros are over and above what is predicted by a homogeneous Poisson model. Negative binomial distribution, beta binomial distribution and zero inflated Poisson models are found to provide satisfactory fits in such cases. This motivated us to look for procedures that make allowance for model misspecification. In this regard, we consider using the homogeneous Poisson Model as a working model and estimating the population size using techniques that are robust to departures from the working model.

Zelterman (1988) proposes a series of estimators, which allow exclusion of parts of the data that intuitively have less bearing on the zero-class. In the application to capture-recapture experiments, he argues that individuals never seen are more similar to those rarely seen than to those captured many times. Extending his idea to our case, we consider the estimators of the population size when λ in the working model is estimated using only those clusters with few cases. The heuristic justification for using only those clusters with few cases to estimate λ is that such clusters are likely to have more bearing on the hidden diseased fraction compared to those that have more detected cases. Technically, this makes more allowance for model misspecification.

In almost all the data sets we examined, less than 20% of the clusters had more than three cases. Therefore, using clusters with at most three cases is almost similar to using data collected on all the clusters to estimate λ. This led us to consider three models for estimating λ; using all the clusters with cases, using only those with at most three cases and using only those with at most two cases. Once we find an estimate for λ, we use observations on all the clusters with cases to estimate the frequency of the zero-class or the number of disease affected clusters with no apparent cases.

The model that uses all the clusters with cases to estimate λ makes no allowance for model misspecification. Here onwards we refer to this model as Model $\mathcal{N}$. The second model that uses only those clusters with at most three cases to estimate λ makes moderate allowance for model misspecification and hence will be referred to as Model $\mathcal{M}$. The third model that uses only those clusters with at most two cases to estimate λ makes strong allowance for model misspecification and will be referred to as Model $\mathcal{S}$. The estimation of λ under each of these models is described next.

2.1 Model $\mathcal{N}$ (Poisson model with no allowance for misspecification)

Here we use all the clusters with cases to estimate λ. Consider a random sample $y_1, \ldots, y_n$ from the truncated Poisson distribution described by Eq. (1). The likelihood function is

$$L = \prod_{i=1}^{n} \frac{e^{-\lambda m_i}(\lambda m_i)^{y_i}}{(y_i!)(1 - e^{-\lambda m_i})} = (e^{-\lambda \sum m_i}) \prod_{i=1}^{n} \frac{(\lambda m_i)^{y_i}}{y_i!(1 - e^{-\lambda m_i})} .$$

Apart from a constant, the log-likelihood is given by

$$l = -\lambda(\sum m_i) + \sum_{i=1}^{n} y_i \log(\lambda m_i) - \sum_{i=1}^{n} \log(1 - e^{-\lambda m_i}) - \sum \log(y_i!)$$

with score function

$$\frac{\partial l}{\partial \lambda} = -(\sum m_i) + \sum \frac{y_i}{\lambda} - \sum \frac{(e^{-\lambda m_i})(m_i)}{1 - e^{-\lambda m_i}} \tag{2}$$

and the score equation leads to

$$\lambda = \frac{\sum y_i}{\sum m_i + \sum m_i e^{-\lambda m_i}/(1 - e^{-\lambda m_i})} = \frac{\sum y_i}{\sum m_i/(1 - e^{-\lambda m_i})} = \Psi(\lambda) \text{ (say)} .$$

The derivative of $\Psi(\lambda)$ with respect to λ is strictly positive. Hence, the iterative scheme

$$\lambda^{j+1} = \frac{\sum_i y_i}{\sum_i m_i/\{1 - \exp(-\lambda^{(j)} m_i)\}}$$

is guaranteed to converge. Alternatively, λ can be estimated using Newton Raphson iterative scheme. Considering the case of equal cluster sizes, $(\sum_i y_i)/(nm)$ can be used as an initial estimate for λ, where m is the average cluster size.

2.2 Model $\mathcal{S}$ (Poisson model with strong allowance for misspecification)

Here we estimate λ using only those clusters with at most two cases. Let k' denote the total number of clusters in the sample with only one or two cases. Based on the zero-truncated model described in Eq. (1), $P(Y_i = 1) = (e^{-\lambda m_i} \lambda m_i)/(1 - e^{-\lambda m_i})$ and $P(Y_i = 2) = \{e^{-\lambda m_i}(\lambda m_i)^2/2\}/(1 - e^{-\lambda m_i})$.

Since we are only considering clusters with one or two cases, the likelihood can simply be written as $L = \prod_{i=1}^{k'} p_1^{(i)^{1-\delta_i}} p_2^{(i)^{\delta_i}}$ where δ_i is the indicator variable defined as

$$\delta_i = \begin{cases} 1, \text{ if the } i\text{th cluster has two cases} \\ 0, \text{ if the } i\text{th cluster has only one case} \end{cases}$$

and $p_1^{(i)} = \frac{1}{1+\lambda m_i/2}$ and $p_2^{(i)} = \frac{\lambda m_i/2}{1+\lambda m_i/2}$ are the probabilities of observing exactly one case and exactly two cases respectively, based only on those clusters with one or two cases. The model underlying this procedure is simply

$$P(Y_i^+ = j) = (1 + \lambda m_i/2)^{j-1}/(1 + \lambda m_i/2) , \quad \text{for} \quad i = 1, 2, \ldots, N \quad \text{and} \quad j = 1, 2 . \tag{3}$$

The likelihood function can therefore be written as

$$L = \prod_{i=1}^{k'} \left(\frac{\lambda m_i/2}{1 + \lambda m_i/2}\right)^{\delta_i} \left(\frac{1}{1 + \lambda m_i/2}\right)^{1-\delta_i} = \prod_{i=1}^{k'} (\lambda m_i/2)^{\delta_i} \left(\frac{1}{1 + \lambda m_i/2}\right) .$$

Apart from a constant, the log-likelihood is given by

$$l = \sum_{i=1}^{k'} \delta_i \log(\lambda) - \sum_{i=1}^{k'} \log(1 + \lambda m_i/2) = f_2 \log \lambda - \sum_{i=1}^{k'} \log(1 + \lambda m_i/2),$$

where f_2 denotes the total number of clusters with exactly two cases. The score function $\frac{\partial l}{\partial \lambda}$ is given by

$$\frac{\partial l}{\partial \lambda} = \frac{f_2}{\lambda} - \sum_{i=1}^{k'} \frac{m_i/2}{1 + \lambda m_i/2}. \tag{4}$$

and the score equation leads to $\lambda = f_2/(\sum_{i=1}^{k'} \frac{m_i}{2+\lambda m_i})$.

As in the previous case, it is easy to find that the sequence $\lambda^{(j+1)} = \Phi_1(\lambda^{(j)})$ converges. Considering the case of equal cluster sizes, $(2f_2)/(f_1 m)$ can be used as an initial estimate for λ. Here, m is the average cluster size and f_1 and f_2 are the numbers of clusters with exactly one and exactly two cases respectively.

2.3 Model $\mathcal{M}$ (Poisson model with moderate allowance for misspecification)

Here we estimate λ using only those clusters with at most three cases. Let k denote the total number of clusters in the sample with only one, two or three cases. Based on the zero-truncated model described in (1) we have $P(Y_i = 1) = \{e^{-\lambda m_i}(\lambda m_i)\}/\{1 - e^{-\lambda m_i}\}$, $P(Y_i = 2) = \{e^{-\lambda m_i}(\lambda m_i)^2/2\}/\{1 - e^{-\lambda m_i}\}$ and $P(Y_i = 3) = \{e^{-\lambda m_i}(\lambda m_i)^3/6\}/\{1 - e^{-\lambda m_i}\}$, for $i = 1, 2, \ldots, k$, so that the additional truncation of all counts *larger than* 3 leads to the model

$$P(Y_i^+ = j) = \{(\lambda m_i)^{j-1}/j!\}/\{1 + \lambda m_i/2 + (\lambda m_i)^2/6\}, \tag{5}$$

for $i = 1, 2, \ldots, N$ and $j = 1, 2, 3$. The likelihood based only on observations collected on clusters with one, two or three observations is according to (5)

$$L = \prod_{i=1}^{k} \frac{(\lambda m_i)^{y_i - 1}}{y_i!(1 + \lambda m_i/2 + (\lambda m_i)^2/6)}.$$

The log-likelihood function l is given by

$$l = \sum_{i=1}^{k} (y_i - 1) \log(\lambda m_i) - \sum_{i=1}^{k} \log(y_i!) - \sum_{i=1}^{k} \log\left(1 + \frac{\lambda m_i}{2} + \frac{(\lambda m_i)^2}{6}\right), \quad \text{for} \quad y_i = 1, 2, 3.$$

The score function

$$\frac{\partial l}{\partial \lambda} = \frac{1}{\lambda}\left(\sum_{i=1}^{k} y_i - k\right) - \sum_{i=1}^{k} \frac{(m_i/6)(3 + 2\lambda m_i)}{(1 + \lambda m_i/2 + (\lambda m_i)^2/6)}. \tag{6}$$

provides the score equation

$$\lambda = \left(\sum_{i=1}^{k} y_i - k\right) \Big/ \left(\sum_{i=1}^{k} \frac{m_i(3 + 2\lambda m_i)}{6 + 3\lambda m_i + (\lambda m_i)^2}\right) \tag{7}$$

and in turn leads to the iterative scheme

$$\lambda^{(j+1)} = \left(\sum_{i=1}^{k} y_i - k\right) \Big/ \left(\sum_{i=1}^{k} \frac{m_i(3 + 2\lambda^{(j)} m_i)}{6 + 3\lambda^{(j)} m_i + (\lambda^{(j)} m_i)^2}\right)$$

which is guaranteed to converge (see Appendix). In parallel with the initial estimate for λ used in the case of Model $\mathcal{N}$, $(\sum_{i=1}^{k} y_i)/(km)$ can be used as an initial estimate.

3 Estimation of Population Size

We recall that our emphasis is on estimating the population size, N. The population includes clusters that are already observed to have cases as well as those that have cases but are not detected by the identification device. Once an estimate $\hat{N}$ is found for N, the frequency of the zero-class, or the number of clusters that have cases but are not detected by the identification device, can simply be estimated as $\hat{N} - n$, where n is the total number of clusters that are already detected to have cases. Estimation of the frequency of the zero-class will therefore be omitted from the discussion that follows.

We now construct the Horvitz–Thompson estimator of the population size for our situation. Let I_i be an indicator variable defined as

$$I_i = \begin{cases} 1, \text{if the } i\text{th cluster is detected to have cases} \\ 0, \text{otherwise}\,. \end{cases}$$

As we already noted in Section 2, a cluster will be included in the sample if it is detected to have cases. Based on the zero-truncated Poisson model, the probability that the i-th cluster belongs to the sample is $(1 - \mathrm{e}^{-\lambda m_i})$.

The original Horvitz-Thompson estimator (Horvitz and Thompson, 1952) for our case given by $\hat{N} = \sum_{i=1}^{N} I_i/(1 - \mathrm{e}^{-\lambda m_i})$ requires the weight or the probability that each unit is included in the sample to be known. This means λ will have to be known, and if this is the case, $E(\hat{N}) = \sum_{i=1}^{N} E(I_i)/(1 - \mathrm{e}^{-\lambda m_i}) = N$ and hence the estimator will be unbiased.

When λ is unknown, which is always the case in our applications, one of the three models presented in Section 2 can be used to obtain an estimate $\hat{\lambda}$ for λ. Since all the clusters belonging to the sample are detected to have cases, our estimate for N, based on the sample of n observed clusters is

$$\hat{N} = \sum_{i=1}^{n} \frac{1}{1 - \mathrm{e}^{-\hat{\lambda} m_i}}\,. \tag{8}$$

In passing we note that the estimates for N from all three models use the observations collected on all the clusters with cases, even though in Model $\mathcal{M}$ and Model $\mathcal{S}$ only those clusters with more than three and more than two cases are used for fitting the truncated Poisson model.

To assess random error it is important to have an estimate of the standard error of $\hat{N}$ which we consider in the following. Using conditional expectation and closely following the work of van der Heijden et al. (2003), we compute the standard error of $\hat{N}$ as (using the notation $\sigma^2(X) = \mathrm{Var}\,(X)$ for random variable X):

$$\sigma^2(\hat{N}) = E[\sigma^2(\hat{N} \mid I_1, I_2, \ldots, I_N)) + \sigma^2(E(\hat{N} \mid I_1, \ldots, I_N))$$

which can be estimated by means of the multivariate δ-method[1] to provide

$$\widehat{\sigma^2(\hat{N})} = \sum_{i=1}^{n} \frac{(m_i\, \mathrm{e}^{-\hat{\lambda} m_i})^2}{(1 - \mathrm{e}^{-\hat{\lambda} m_i})^4} \frac{1}{\left(-\frac{\partial^2 l}{\partial \lambda^2}\right)_{\lambda = \hat{\lambda}}} + \sum_{i=1}^{n} \mathrm{e}^{-\hat{\lambda} m_i} \left(\frac{1}{1 - \mathrm{e}^{-\hat{\lambda} m_i}}\right)^2, \tag{9}$$

where $\frac{\partial^2 l}{\partial \lambda^2}$ in the three models are given by:

$$\frac{\partial^2 l}{\partial \lambda^2} = \begin{cases} -(\sum y_i)/\lambda^2 + \sum_{i=1}^{n} (m_i^2\, \mathrm{e}^{-\lambda m_i})/(1 - \mathrm{e}^{-\lambda m_i})^2\,, & (\text{Model } \mathcal{N}) \\ -(\sum_{i=1}^{k} y_i - k)/\lambda^2 - \sum_{i=1}^{k} (m_i/6)^2\, (3 - 6\lambda m_i - 2\lambda^2 m_i^2)/\left(1 + \frac{\lambda m_i}{2} + \frac{\lambda^2 m_i^2}{6}\right)^2, & (\text{Model } \mathcal{M}) \\ -f_2/\lambda^2 + \sum_{i=1}^{k'} (m_i/2)^2/(1 + \lambda m_i/2)^2\,, & (\text{Model } \mathcal{S})\,. \end{cases}$$

[1] Details are available in an extended report.

4 Robustness of Estimators: A Simulation Study

Here we describe a simulation study that illustrate the performance of the estimators when the model is correctly specified and the robustness of the estimators. We examined several alternative distributions that are widely used in the literature for count data and will be described in the sequel.

4.1 Design of the study

We fixed the population size N and the cluster sizes $m_i, i = 1, 2, \ldots, N$. The number of cases in each of these clusters were generated from the distributions:

- Poisson distribution with mean λm_i for fixed λ. In this case, $P(Y_i = j) = \mathrm{e}^{-\lambda m_i}(\lambda m_i)^j/j!$. (homogeneous Poisson model)
- Mixture of two Poissons with mixing proportions p and $(1-p)$ from Poisson distributions with means $\lambda_1 m_i$ and $\lambda_2 m_i$ respectively. The chosen values of λ_1 and λ_2 are presented with the results. We examined a series of p values: 0.90, 0.80, 0.70, 0.60 and 0.50.

In this case, $P(Y_i = j \mid \Theta = \theta) = \theta\, \mathrm{e}^{-\lambda_1 m_i}(\lambda_1 m_i)^j/j! + (1-\theta)\, \mathrm{e}^{-\lambda_2 m_i}(\lambda_2 m_i)^j/j!$, where Θ is a Bernoulli random variable taking values 1 and 0 with probabilities p and $(1-p)$ respectively.

- Negative binomial distribution which can be regarded as a heterogeneous mixture of Poisson distributions with mean parameter distributed as Gamma with parameters α and β. The values of α and β used are presented with the results. In this case, $P(Y_i = j \mid \Theta = \theta) = \mathrm{e}^{-\theta}\theta^j/j!$ where $\Theta \sim$ Gamma (α, β) so that $P(\Theta = \theta) = \frac{1}{\beta^\alpha \Gamma(\alpha)}\, \theta^{\alpha-1}\, \mathrm{e}^{-\theta/\beta}$.

At each simulation, once we generated the data, we ignored the clusters with no cases and regarded the rest of the clusters as the sample. Then, we used Eq. (8) to estimate the total number of clusters, N, from each sample under each of the three models presented earlier. The results presented in Table 4 to Table 5 are based on 10 000 simulations, where we reported the average of the 10 000 values of $\hat{N}$ computed from each simulated sample, mean $(\hat{N})$ and in parentheses, the standard deviations of the $\hat{N}$ values, SD $(\hat{N})$,

In the following we present some further details of the study. The intention here was to mimic the setting of the surveillance data on scrapie.

- The total population size N was fixed at 135.
- The cluster sizes m_i $(i = 1, 2, \ldots, 135)$ used are presented in Table 3 and are the sizes of the sheep holdings in the study referred in Section 1.
- The cluster sizes were arranged in the ascending order of number of cases detected from the cluster so that the interested readers can derive the complete data set together with the information presented in Table 1.

We examined the performance of the estimators for a series of λ values around the estimates obtained for the real data set. This is for the purpose of examining the behaviour of the estimators as λ is reduced resulting in an increase in the proportion of affected clusters with no cases.

4.2. Comparison of the estimators when the model is correctly specified

Firstly, we looked at the behavior of the estimators when the model is correctly specified. To save space we just report the major results.[2] We find that when the data follow a homogeneous Poisson distribution, thus with model correctly specified, Model $\mathcal{N}$ that makes no allowance for model misspecification performs quite well. Model $\mathcal{M}$ that makes moderate allowance for model misspecification also performs reasonably well except when λ is quite small resulting in a high proportion of disease affected clusters with no detectable cases. In this case, Model $\mathcal{M}$ tends to slightly overestimate the

[2] More details are available in a report which can received from the authors.

population size. Model $\mathcal{S}$ that makes strong allowance for model misspecification always overestimates the population size and furthermore, the deviation from the true value increase as the proportion of disease affected clusters with no cases increase. The simulation study also shows that in comparison with the estimator from Model $\mathcal{N}$, the estimator from Model $\mathcal{M}$ is less efficient. The estimator from Model $\mathcal{S}$ is the least efficient out of the three estimators.

4.3 Comparison of the estimators when the true distribution is a mixture of two Poissons

Table 4 presents the results when the data follow a distribution which is a mixture of two Poissons with means $\lambda_1 m_i$ and $\lambda_2 m_i$ where a proportion of p $(0 < p < 1)$ was taken from the distribution with mean $\lambda_1 m_i$. It is worth noting that since clusters are of possibly different sizes, we do not have an identically distributed sample. The mean, $E(X)$, and variance $V(X)$, of the Poisson mixture are $p\lambda_1 m_i + (1-p)\,\lambda_2 m_i$ and $m_i^2(\lambda_1 - \lambda_2)^2\, p(1-p) + p\lambda_1 m_i + (1-p)\,\lambda_2 m_i$ respectively. It is easy to see that this variance is always greater than that of the homogeneous Poisson model.

To make the comparison easier, we selected λ_1 to be $\frac{\lambda}{2p}$ and λ_2 to be $\frac{\lambda}{2(1-p)}$ so that the resulting mixture will have the same mean as the homogeneous Poisson model, which is λm_i. The extent of heterogeneity compared to the homogeneous Poisson model is $m_i^2(\lambda_1 - \lambda_2)^2\, p(1-p)$. For the choice of λ_1 and λ_2, the heterogeneity reduces to $\frac{m_i^2\lambda^2(1-2p)^2}{4p(1-p)}$. It is easy to see that when p is 0.5, the components of the mixture are identically distributed and there is no heterogeneity. As p is increased, the heterogeneity increases. The results for a series of p values are presented in Table 4. Here we notice that Model $\mathcal{N}$ that makes no allowance for misspecification underestimates the population size in all the cases where heterogeneity is present. Model $\mathcal{M}$ that makes moderate allowance for model misspecification performs quite well. As heterogeneity increases, the relative efficiency compared to Model $\mathcal{N}$ increases. Model $\mathcal{S}$ that makes strong allowance for misspecification overestimates the population size. Furthermore, we find that the estimator from Model S is the least efficient. It appears that Model $\mathcal{M}$ represents a good compromise between bias and variance as expressed in the root mean squared error (RMSE) provided in Table 4. We examined several other λ values as well and the results were similar.

Table 4 Mixture of Poissons: p Pois $(\lambda_1 m_i) + (1-p)$ Pois $(\lambda_2 m_i)$; $\lambda_1 = \frac{\lambda}{2p}, \lambda_2 = \frac{\lambda}{2(1-p)}$; True $N = 135$.

parameters	Clusters with no cases (%)	mean $(\hat{N})$, (SD $(\hat{N})$, RMSE $(\hat{N})$)			Relative Efficiency	
		Model $\mathcal{N}$	Model $\mathcal{M}$	Model $\mathcal{S}$	Model $\mathcal{M}$	Model $\mathcal{S}$
$\lambda = 0.01$						
$p = 0.99$	53.02	104.35 (31.41, 43.89)	143.21 (34.74, 35.70)	148.34 (49.07, 50.85)	1.51	0.74
$p = 0.9$	46.69	107.37 (18.32, 33.15)	146.87 (31.87, 34.01)	154.88 (46.31, 50.40)	0.95	0.43
$p = 0.8$	41.71	119.94 (16.54, 22.37)	138.14 (25.05, 25.25)	145.45 (37.44, 38.87)	0.79	0.33
$p = 0.7$	38.43	128.93 (16.61, 17.68)	136.51 (22.18, 22.23)	141.63 (32.30, 33.00)	0.63	0.29
$p = 0.6$	36.64	133.93 (16.39, 16.42)	136.64 (21.06, 21.12)	140.40 (30.00, 30.48)	0.60	0.29
$p = 0.5$	36.07	135.35 (15.79, 15.79)	136.81 (20.49, 20.57)	140.01 (28.99, 29.42)	0.59	0.29

Table 5 Negative Binomial: Pois $(\lambda), \lambda \sim$ Gamma (α_i, β); True $N = 135$.

parameters	Clusters with no cases (%)	mean $(\hat{N})$, (SD $(\hat{N})$, RMSE $(\hat{N})$)			Relative Efficiency	
		Model $\mathcal{N}$	Model $\mathcal{M}$	Model $\mathcal{S}$	Model $\mathcal{M}$	Model $\mathcal{S}$
$\bar{\lambda} = 0.01$						
$\alpha_i = m_i/80$,	43.92	109.90	115.54	122.96	1.14	0.82
$\beta = 80\bar{\lambda}$		(13.31, 28.41)	(18.15, 26.61)	(28.92, 31.33)		
$\alpha_i = m_i/40$,	40.50	120.53	123.77	129.48	0.84	0.49
$\beta = 40\bar{\lambda}$		(14.54, 20.51)	(19.36, 22.38)	(28.74, 29.27)		
$\alpha_i = m_i/20$,	38.45	127.31	129.65	134.10	0.68	0.35
$\beta = 20\bar{\lambda}$		(15.27, 17.10)	(20.11, 20.81)	(28.82, 28.83)		
$\alpha_i = m_i/10$,	37.34	130.94	132.56	136.11	0.63	0.31
$\beta = 10\bar{\lambda}$		(15.55, 16.07)	(20.10, 20.25)	(28.90, 28.92)		
$\alpha_i = m_i/5$,	36.66	133.13	134.52	137.90	0.60	0.30
$\beta = 5\bar{\lambda}$		(15.58, 15.69)	(20.21, 20.22)	(28.49, 28.64)		
$\alpha_i = m_i/3$,	36.50	133.03	135.40	138.39	0.61	0.30
$\beta = 3\bar{\lambda}$		(15.91, 16.03)	(20.60, 20.60)	(28.85, 29.05)		
$\alpha_i = m_i$,	36.20	134.60	136.03	138.92	0.60	0.30
$\beta = \bar{\lambda}$		(15.86, 15.87)	(20.49, 20.52)	(28.87, 29.13)		
$\alpha_i = 2m_i$,	36.15	135.21	136.95	139.58	0.58	0.29
$\beta = \bar{\lambda}/2$		(15.96, 15.96)	(20.78, 20.87)	(29.28, 29.64)		

4.4 Comparison of the estimators when the true distribution is negative binomial

Table 5 presents the results when the data follow a negative binomial distribution. Here we have regarded the negative binomial distribution as a heterogeneous mixture of Poissons with a gamma mixing distribution with parameters α_i and β. The mean and the variance of the resulting negative binomial data are $\alpha_i\beta$ and $\alpha_i\beta(1+\beta)$ respectively. Since $\beta > 0$, this variance is always more than that of the homogeneous Poisson model. The extent of heterogeneity is reflected from the component $\alpha_i\beta^2$.

As before, the parameters α_i and β of the gamma mixing distribution were chosen in such a way that the resulting negative binomial distribution will have the same mean as the homogeneous Poisson model, which is, $\bar{\lambda}m_i$. While keeping the mean fixed, the heterogeneity was reduced by reducing the scale parameter β.

Here again we find that Model $\mathcal{N}$ that makes no allowance for model misspecification underestimates the population size, unless the heterogeneity is quite small. Model $\mathcal{M}$ that makes moderate allowance for misspecification performs better than Model $\mathcal{N}$ but as heterogeneity is increased further Model $\mathcal{M}$ also underestimates. The relative efficiency of Model $\mathcal{M}$ increases as heterogeneity increases. Model $\mathcal{S}$ that makes strong allowance for misspecification is more robust in comparison to Model $\mathcal{N}$ and Model $\mathcal{M}$ but is least efficient. As a trade-off between the gain in robustness and loss in efficiency, Model $\mathcal{M}$ is superior as can again seen in the RMSE in Table 5.

4.5 Conclusions

In summary, in all the cases we examined in the simulation study, we find that Model $\mathcal{N}$ that makes no allowance for model misspecification underestimates the population size even when the variance of the true distribution is slightly in excess to that of the homogeneous Poisson model. Model $\mathcal{M}$ performs well with moderate heterogeneity. Model $\mathcal{S}$ overestimates the population size, if the frequency of cases follows a homogeneous Poisson model. However, in the presence of heterogeneity, Model $\mathcal{S}$

Table 6 Estimated number of scrapie-affected holdings in Great Britain

parameter	Estimate (SE)		
	Model $\mathcal{N}$	Model $\mathcal{M}$	Model $\mathcal{S}$
λ	0.010	0.007	0.005
	(0.001)	(0.001)	(0.001)
N	351.76	498.56	584.87
	(61.37)	(97.26)	(119.67)

is more robust but is relatively less efficient to Model $\mathcal{N}$ and Model $\mathcal{M}$. If the heterogeneity is quite large, the estimate from Model $\mathcal{S}$ is preferred as a conservative estimate for the population size. Considering the robustness and the relative efficiency, we recommend Model $\mathcal{M}$, in general, unless there is strong evidence to support that the distribution is a homogeneous Poisson.

5 An Application to Scrapie Surveillance in Great Britain

This is the data set referred in Section 1 on scrapie-affected sheep holdings in Great Britain. The data presented in Table 1 contained the frequencies of counts of scrapie-cases detected in the 135 holdings and Table 3 presented the cluster sizes. Here again, data on the zero-class is not observed. The population size, N, here refers to the total number of scrapie-affected sheep holdings and the estimates obtained using the techniques proposed in this paper are presented in Table 6.

As in the previous example, estimates for the population size from all three models exceed the total number of holdings detected with scrapie. Thus, there is evidence of the presence of scrapie-affected holdings with no reported cases.

A 95% confidence interval for N, based on Model $\mathcal{N}$ using standard normal quantiles is (231.47, 472.05). The estimate for Model $\mathcal{M}$ falls outside this interval. Thus, the estimates for the population size from Model $\mathcal{N}$ and Model $\mathcal{M}$ are quite different. A 95% confidence interval for N based on Model $\mathcal{M}$ is (307.93, 689.19). The estimate from Model $\mathcal{S}$ falls into this interval. Recall that Model $\mathcal{S}$ is more robust but is less efficient in comparison with Model $\mathcal{M}$. Therefore, we prefer the estimate from Model $\mathcal{M}$ and estimate the total number of animal holdings that are infected to be around 499.

6 Discussion

In many epidemiological assessment studies, a fixed number of households, animal holdings or some such clusters are examined for the presence of the disease. Usually, the data take the form of the number of units examined in each cluster and the number of cases found. A case generally refers to a unit that has shown symptoms of the disease. However, it is more realistic to assume that such data do not give a complete picture of the spread of the epidemic (Mosley et al., 1972; Longini and Koopman, 1982). It is required to estimate the total population size that include clusters that are detected with cases as well as those that are affected by the disease but remain undetected by the identification mechanism. Many authors have discussed this problem for non-clustered data (McKendrick, 1926; Zelterman, 1988). In this paper, we considered robust estimation of the population size from zero-truncated clustered data. The work presented here is an extension of Zelterman's approach for robust estimation in the non-clustered case. We have considered extension of Zelterman's approach in two directions. On the one hand, we have extended the robust estimation proposed for the unclustered situation to deal with clustered data. On the other hand, our approach allows use of several frequencies ($f_1, f_2, f_3, \ldots$ etc. where f_j denote the number of clusters with exactly j cases) as opposed to using only two frequencies (eg. f_1 and f_2) in the Zelterman's approach.

The robustness of the proposed estimators and loss in efficiency were examined using a simulation study. This revealed that, when the true distribution is a mixture of two Poissons or a negative-binomial, thus having more heterogeneity compared to the Poisson model, the usual procedure of fitting a truncated homogeneous Poisson model by maximum likelihood and thereby using a Horvitz–Thompson estimator of population size underestimates the population size and is least robust. This agrees with the observations made by Böhning und Schön (2005) for unclustered data. Extending the approach proposed by Zelterman (1988) for the unclustered data and fitting the truncated homogeneous Poisson model considering only those clusters with at most two cases was found to be most robust but least efficient. As a trade-off between gain in robustness and loss in efficiency, Model $\mathcal{M}$ which uses only those clusters with at most three cases was preferred, in general. In applications, we recommend to fit all three models and choose an estimate comparing the values of the estimates, robustness and efficiency.

Model (1) implies that expected counts are proportional to cluster sizes. Although this assumption seems meaningful it might not be appropriate in other situations and needs to be assessed with suitable methodology. Our approach here outlinesways to proceed *given* this assumption of proportionality holds. In addition, for models S and $\mathcal{M}$ the assumption of proportionality is less crucial.

In the simulation study, we have sampled (with replacement) cluster sizes according to the observed cluster sizes in our main example from scrapie surveillance. This procedure is based upon the assumption that clusters with undetected cases follow the same distribution in their cluster sizes. This might be violated in practice, for example, holdings with undetected cases might be smaller in size. Unfortunately, this assumption cannot be supported unless a validation sample is available.

Appendix

Consider the likelihood function based only on observations collected on clusters with one, two or three observations.

$$L = \prod_{i=1}^{k} \frac{\mathrm{e}^{-\lambda m_i}\,(\lambda m_i)^{y_i}}{y_i!(1-\mathrm{e}^{-\lambda m_i})\,(\mathrm{e}^{-\lambda m_i}\,\lambda m_i\{1+\lambda m_i/2+(\lambda m_i)^2/6\})/(1-\mathrm{e}^{-\lambda m_i})}\,, \quad \text{for} \quad y_i = 1,2,3\,.$$

The maximum likelihood estimate satisfies $\lambda = \left(\sum_{i=1}^{k} y_i - k\right) \Big/ \left(\sum_{i=1}^{k} \frac{m_i(3+2\lambda m_i)}{(6+3\lambda m_i+(\lambda m_i)^2}\right) = \Phi(\lambda)$, (say). The derivative of $\Phi(\lambda)$ with respect to λ is given by

$$\Phi'(\lambda) = \frac{\sum_{i=1}^{k} y_i - k}{\left(\sum_{i=1}^{k} \frac{m_i(2\lambda m_i)}{(6+3\lambda m_i+(\lambda m_i)^2}\right)^2} \sum_{i=1}^{k} \frac{2m_i^2(\lambda^2 m_i^2 + 3\lambda m_i - 3/2)}{l(6+3\lambda m_i+(\lambda m_i)^2)^2}\,.$$

Notice that $(\sum y_i) - k$ is always positive. Also, notice that $(\lambda^2 m_i^2 + 3\lambda m_i - 3/2)$ is positive if $\lambda m_i > \sqrt{15}/2 - 3/2 = 0.4365$, or equivalently if $m_i > 0.4365/\lambda$. In the situation of the scrapie data and model $\mathcal{M}$ we have $\hat{\lambda} = 0.007$ and most holdings have larger sizes then $0.4365/0.007 = 62.36$. We find here that $\Phi'(\lambda)$ is strictly positive at $\lambda = 0.007$ and hence the sequence $\lambda^{(j+1)} = \Phi(\lambda^{(j)})$ is guaranteed to converge if it is started in a vicinity close enough to $\hat{\lambda} = 0.007$.

Acknowledgements *The authors wish to thankfully acknowledge that this research was carried out while the first author was on a Visiting Fellowship from the Commonwealth Fellowship Commission, UK. The authors also wish to thank the Section of Quantitative Biology and Applied Statistics, School of Biological Sciences of the University of Reading, UK for making the computing facilities available for the research.*

Conflict of Interests Statement
The authors have declared no conflict of interest.

References

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., and Kirchner, U. (1999). The Zero-Inflated Poisson Model and the Decayed, Missing and Filled Teeth Index in Dental Epidemiology. *Journal of the Royal Statistical Society: Series A* **162**, 195–209.

Böhning, D. and Schön, D. (2005). Nonparametric Maximum Likelihood Estimation of Population Size based on the Counting Distribution. *Journal of the Royal Statistical Society: Series C* **54**(4), 721–737.

Chao, A. (1987) Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics* **43**, 783–791.

Del Rio Vilas, V. J., Sayers, R., Sivam, K., Pfeiffer, D. U., Guitian, J., and Wilesmith, J. W. (2005). A case study of capture-recapture methodology using scrapie surveillance data in Great Britain. *Preventive Veterinary Medicine* **67**, 303-317.

Griffiths, D. A. (1973). Maximum Likelihood Estimation for the Beta Binomial Distribution and an Application to the Household Distribution of the Total Number of Cases of a Disease. *Biometrics* **29**, 637–648.

Hoinville, L. J., Hoek, A., Gravenor, M. B., and Mclean, A. R. (2000). Descriptive epidemiology of scrapie in Great Britain: results of a postal survey. *Veterinary Record* **146**, 455-461.

Horvitz, D. G. and Thompson, D. J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. *Journal of the American Statistical Association* **47**, 663–685.

Longini, I. M. and Koopman, J. S. (1982). Household and Community Transmission Parameters from Final Distributions of Infections in Households. *Biometrics* **38**, 115 –126.

Lowe, S. A. (1999). The Beta-Binomial Mixture Model for Word Frequencies in Documents with Application to Information Retrieval. *EUROSPEECH'99*, 2443–2446.

McLean, A. R., Hoek, A., Hoinville, L. J., and Gravenor, M. B. (1999). Scrapie Transmission in Britain: A Recipe for a Mathematical Model. *Proceedings of the Royal Society: Biological Sciences* **266**, 2531–2538.

McKendrick, A. G. (1926). Applications of Mathematics to Medical Problems. *Proceedings of the Edinburgh Mathematical Society* **44**, 98–130.

Mosley, W. H., Bart, K. J., and Sommer, A. (1972). An Epidemiological Assessment of Cholera Control Programs in Rural East Pakistan. *International Journal of Epidemiology* **1**, 5–11.

Sivam, K., Baylis, M., Gravenor, M. B., Gubbins, S., and Wilesmith, J. W. (2003). Occurrence of scrapie in GB: results of a postal survey in 2002. *Veterinary Record* **153**, 782–783.

van der Heijden, P., Bustami, R., Cruyff, M. J., Engbersan, G., and van Houwelingen, H. C. (2003). Point and Interval Estimation of Population Size Using the Truncated Poisson Regression Model. *Statistical Modelling* **3**, 305–322.

Zelterman, D. (1988). Robust Estimation in Truncated Discrete Distributions with Application to Capture Recapture Experiments, *Journal of Statistical Planning and Inference* **18**, 225–237.

