Computational Statistics and Data Analysis 55 (2011) 466–474

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

An extension of an over-dispersion test for count data

M. Fazil Baksh, Dankmar Böhning *, Rattana Lerdsuwansri

Section of Applied Statistics, School of Biological Sciences, University of Reading, Reading RG6 6BX, England, United Kingdom

ARTICLE INFO

Article history:
Received 10 December 2009
Received in revised form 14 May 2010
Accepted 14 May 2010
Available online 1 June 2010

Keywords:
Capture–recapture
Over-dispersion
Turing estimator
Zero-inflation
Zero-truncation

ABSTRACT

While over-dispersion in capture–recapture studies is well known to lead to poor estimation of population size, current diagnostic tools to detect the presence of heterogeneity have not been specifically developed for capture–recapture studies. To address this, a simple and efficient method of testing for over-dispersion in zero-truncated count data is developed and evaluated. The proposed method generalizes an over-dispersion test previously suggested for un-truncated count data and may also be used for testing residual over-dispersion in zero-inflation data. Simulations suggest that the asymptotic distribution of the test statistic is standard normal and that this approximation is also reasonable for small sample sizes. The method is also shown to be more efficient than an existing test for over-dispersion adapted for the capture–recapture setting. Studies with zero-truncated and zero-inflated count data are used to illustrate the test procedures.

1. Introduction

The Poisson distribution is commonly used in modelling zero-modified (i.e. zero-truncated or zero-inflated) count data. Zero-modified count data can be found in a range of disciplines including epidemiology, public health, biology, sociology, engineering and agriculture. For instance, count data such as length of hospital stay, number of car accidents, catch rates and wild fires within a particular time period are typically zero-modified. In particular, zero-truncated count data arise in capture–recapture studies concerned with estimating the size of populations that are hidden or difficult to measure, such as number of drug users within a region and elusive animal populations.

The conventional Poisson distribution $Po(\lambda)$ has mean λ equal to its variance. This is referred to as equi-dispersion. When the variance of the observed counts is greater than the mean, we are said to have over-dispersion. The presence of over-dispersed data in a study is often a result of sampling from different and unknown sub-populations and can lead to biased inference in many ways (Lindsay, 1995; Böhning, 2000). For example when over-dispersion is ignored it is well known (see for instance Aitkin et al., 1977) that estimates for the variance of parameter estimates might be too small. Furthermore, there is the additional complication that the population size estimate from capture–recapture studies can be severely negatively biased if population heterogeneity is ignored (Böhning et al., 2005).

Example 1. To illustrate the potential for biased inference we consider the capture–recapture study by van der Heijden et al. (2003b) on illegal gun ownership in the Netherlands. Data from this study for the 2-year period from 1998 and 1999 and for 5 regions of the Netherlands, obtained from police registers of violations against possession of firearms, is presented in Table 1.

There are $f_1 = 2561$ illegal gun owners who have been identified during the observational period *exactly once*, $f_2 = 72$ have been identified *exactly twice*, and exactly $f_3 = 5$ illegal gun owners have been identified three times; total size of the

* Corresponding author. Tel.: +44 118 378 6211; fax: +44 118 378 8032.
E-mail address: d.a.w.bohning@reading.ac.uk (D. Böhning).

0167-9473/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2010.05.015

Table 1
Zero-truncated count distribution on illegal gun owners for the period 1998–1999 for 5 regions of the Netherlands.

f_0	f_1	f_2	f_3	n
–	2561	72	5	2638

observed sample is $n = 2638$. Clearly, illegal gun owners who never got caught do *not* appear in the register and hence there are no zeros observed. Here, interest is in f_0, the number of hidden or unobserved gun owners. A simple estimate $\hat{N}$ of the population size $N = n + f_0$ can be obtained using the Horvitz–Thompson estimator $\hat{N} = n/(1 - p_0)$, where the probability of observing a zero count p_0 is to be estimated. Under the assumption of a Poisson distribution with mean λ, we get $p_0 = \exp(-\lambda) = \exp(-\lambda)\lambda/\lambda$ which leads to the estimate $\hat{p}_0 = f_1/S$ with $S = f_1 + 2f_2 + \cdots + mf_m$ (m being the largest observed count) and consequently to the Good–Turing estimate of N, $\hat{N} = n/(1 - f_1/S)$ (Good, 1953).

For this example the Good–Turing estimate is $\hat{N} = 45{,}128$. The Poisson assumption on which this estimate is built is known to be frequently violated in capture–recapture studies. This violation is often caused by the occurrence of heterogeneity implying that not one, but several Poisson parameters, are required in different parts of the population. Heterogeneity is closely connected to the occurrence of over-dispersion. We return to this example later.

Whereas the question of over-dispersion and general goodness-of-fit is well discussed in various textbooks including Cameron and Trivedi (1998, chap. 5), Winkelmann (2003, chap. 3) and Collett (2003, chap. 6), model evaluation and goodness-of-fit testing is less discussed for zero-truncated modelling. However, there is the grounding work by Rao and Chakravarthi (1956) and, more recently, the assessment and review paper by Best et al. (2007) who compare a number of tests for goodness-of-fit. Rao and Chakravarthi (1956) dispersion test statistic, in the spirit of exploratory data analysis, is

$$D = \frac{(S^{(2)} - S^2/n)(1 - \mathrm{e}^{-\hat{\lambda}})^2}{\hat{\lambda}[1 - (1 + \hat{\lambda})\mathrm{e}^{-\hat{\lambda}}]}, \tag{1}$$

where $S^{(2)}$ is the sum of squares of the observed counts and $\hat{\lambda}$ is the maximum likelihood estimate for the parameter λ of the zero-truncated Poisson distribution. In the comparison (Best et al., 2007) of the dispersion test based on $U = (D - n)/\sqrt{2n}$ with four other tests, it was shown that U is most efficient for the various alternatives considered.

In this paper we suggest a simple test statistic for examining the presence of over-dispersion in zero-modified count data. This statistic will help practitioners develop trust in their inference, such as when estimating population size from capture–recapture data under the assumption of homogeneity. It will also identify when a different procedure, such as one capable of coping with heterogeneity, is more appropriate. In Section 2 we introduce the generalization of the over-dispersion statistic suggested in Böhning (1994) for zero-truncated count data, including a correction to improve the normal approximation and examine the sampling distribution of the test statistic. Also, type I error and efficiency of the proposed test is compared with type I error and efficiency of the over-dispersion test using U. Finally, we illustrate an application of the over-dispersion test to zero-inflated data and introduce a slightly different version of the test, suitable for sparse count data.

2. The over-dispersion test

2.1. The test statistic $\tilde{T}$

Let $X_1, \ldots, X_N$ be a sample of size N of counts from an unknown distribution with mean λ, and suppose it is of interest to test whether the sample is over-dispersed. When N is fixed and known, the test statistic

$$T = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X})^2 - \bar{X}}{\sqrt{\frac{2}{N-1}}\bar{X}}, \tag{2}$$

where $\bar{X} = \sum X_i/N$, was suggested (Böhning, 1994) for testing the null hypothesis $H_0 : X \sim Po(\lambda)$ against the alternative $H_1 : var(X) > \lambda$. This statistic was proposed as the correction to the one of Tiago de Oliveira (1965) and is based on the fact that, under the null, the expected value of the over-dispersion estimate $\sum_{i=1}^{N}(X_i - \bar{X})^2/(N - 1) - \bar{X}$ is equal to zero with variance equal to $2\lambda^2/(N-1)$. These properties will be used later to develop similar results for our proposed over-dispersion statistic.

In studies with zero-truncated count data, such as in capture–recapture studies, the only observed counts are those for which the random variable X is non-zero. Let N denote the population size and, without loss of generality, denote the observed sample of non-zero counts from a capture–recapture study as $X_1, \ldots, X_n$ and let $X_{n+1}, \ldots, X_N$ be the remaining unobserved zero counts. Thus the sample is now truncated at known n while N is unknown, but assumed fixed.

Unlike the Poisson random variable, the mean of a zero-truncated Poisson random variable X_+ is not equal to the variance. Rather, the mean $E(X_+) = \lambda/(1 - \exp(-\lambda))$ is related to the variance $var(X_+)$ by $var(X_+) = E(X_+)\{1 - E(X_+)\exp(-\lambda)\}$,

(see also Cameron and Trivedi, 1998, p. 119). Thus the zero-truncated, homogeneous Poisson is itself an under-dispersion model and a test procedure for over-dispersion may be based on this mean–variance relationship. Alternatively, since all sample elements $X_{n+1}, \ldots, X_N$ are known to be zero counts, all that is required is an estimate $\hat{N}$ of N to impute the residual sample elements $X_{n+1}, \ldots, X_{\hat{N}}$. We then use the available over-dispersion statistic T of Eq. (2) for the observed sample $X_1, \ldots, X_n$ jointly with the imputed sample $X_{n+1}, \ldots, X_{\hat{N}}$ to construct the over-dispersion test. The details follow.

The proposed test for over-dispersion first extends the statistic T in Eq. (2) by replacing $N-1$ by N. This gives,

$$T \approx \frac{\sum_{i=1}^{N}(X_i-\bar{X})^2 - \sum_{i=1}^{N} X_i}{\sqrt{\frac{2}{N}}\sum_{i=1}^{N} X_i} = \frac{S^{(2)} - S\bar{X} - S}{\sqrt{2\bar{X}S}}, \tag{3}$$

where as before $S^{(2)} = \sum_{i=1}^{N} X_i^2$, $S = \sum_{i=1}^{N} X_i$ and $\bar{X} = S/N$. This approximation is valid when N is large. It is also important to note that $S^{(2)}$ and S are known irrespective of whether the population size N is known or not while $\bar{X}$ will only be known if N is known. Thus, as can be seen from Eq. (3), in order to calculate the test statistic all that is now required is an estimate for the mean λ, or equivalently, an estimate for N.

We propose the Good–Turing estimator $\hat{N} = n/(1-\frac{f_1}{S})$, where f_1 is the frequency of observed counts 1, as an estimate for N. Using this, we get $\tilde{\lambda} = \bar{X} = (S-f_1)/n$ and the proposed test statistic becomes

$$\hat{T} = \frac{S^{(2)} - S\left(\frac{S-f_1}{n}+1\right)}{\sqrt{2S\frac{S-f_1}{n}}}. \tag{4}$$

Note that use of the Good–Turing estimator $\hat{N}$ of N is justified under the Poisson assumption and, as we show below, the over-dispersion estimate based on $\hat{N}$ is unbiased. Furthermore, the distribution of $\hat{T}$ is approximately standard normal, with convergence for large N and λ.

Let $\bar{X} = \sum_{i=1}^{\hat{N}} X_i/\hat{N}$ and $V^2 = \sum_{i=1}^{\hat{N}}(X_i-\bar{X})^2/(\hat{N}-1)$ be the estimated mean and variance using the Good–Turing estimator $\hat{N}$ of N. From the fact that under the Poisson assumption, $E\left(\sum_{i=1}^{N}(X_i-\bar{X})^2/(N-1)-\bar{X}\right) = 0$ (see Eq. (2)) we get $E(V^2-\bar{X}|\hat{N}) = 0$ for all possible values of $\hat{N}$ and hence $E(V^2-\bar{X}) = 0$. Thus the over-dispersion estimate based on an estimated population size is also unbiased. We can also apply this same conditional expectation argument to the variance. Using $var\{\sum_{i=1}^{N}(X_i-\bar{X})^2/(N-1)-\bar{X}\} = 2\lambda^2/(N-1)$ (Böhning, 1994) we get $E[(V^2-\bar{X})^2|\hat{N}] = 2\lambda^2/(\hat{N}-1)$ for any given $\hat{N}$ and hence $var(V^2-\bar{X}) \approx 2\lambda^2/(N-1)$ if N is large.

We next use simulated data to investigate the moments of our over-dispersion test statistic $\hat{T}$ in Eq. (4). Figs. 1 and 2 show the mean and variance respectively of $\hat{T}$ from 1000 replicates, for values of the Poisson mean λ in the range 0.5 to 5 and selected values of N. Also shown in the variance plot is the curve $1-\exp(-\lambda)$. While the mean is consistently close to zero, the variance seems to be reasonably close to one when $\lambda \geq 3$ but is less than one for smaller values of λ. In order to improve the standard normal approximation to the distribution of T, based on these findings we introduce the correction factor $1-\exp(-\lambda)$, where the parameter λ is again estimated under the null hypothesis of equi-dispersion. This gives the over-dispersion test statistic

$$\tilde{T} = \frac{S^{(2)} - S(\tilde{\lambda}+1)}{\sqrt{2S\tilde{\lambda}(1-e^{-\tilde{\lambda}})}}, \tag{5}$$

where $\tilde{\lambda} = (S-f_1)/n$.

2.2. Sampling distribution of $\tilde{T}$

For values of the mean $\lambda = 0.5, 1, 2, 5$ and true population size $N = 50, 100, 1000$ and 10,000, a total of 2000 replicate values each of the statistic $\tilde{T}$ in Eq. (5) and the statistic $U = (D-n)/\sqrt{2n}$ (see Eq. (1)) were generated and compared with the corresponding quantiles of the standard normal distribution. The resulting Q–Q plots in Fig. 3 suggest that the sampling distribution of $\tilde{T}$ converges to the standard normal distribution as either N or λ increases. Similar findings (not shown) were obtained for the sampling distribution of U.

The type I error of the over-dispersion tests based on $\tilde{T}$ and U was next evaluated by simulations. For each of a set of values for λ and N we conducted 10,000 simulations under the null hypothesis of no over-dispersion; the achieved type I errors for the test based on $\tilde{T}$ are shown in Table 2. As can be seen from this table, all analyses generate approximately the correct type I error rate; similar results were found for U.

In the final simulation assessment, 10,000 replicate sets of over-dispersed data were generated using a two component mixture of Poisson distributions and the number of times the null hypothesis was rejected using $\tilde{T}$ and U respectively, were

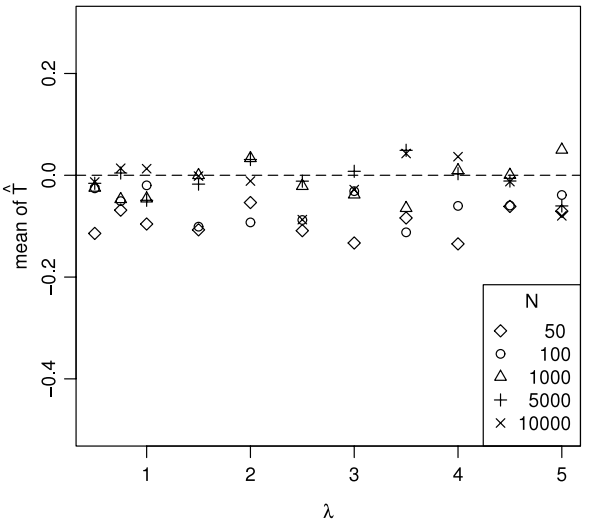

Fig. 1. Simulated values of mean of the over-dispersion statistic $\hat{T}$ for a range of values of λ and true population size N.

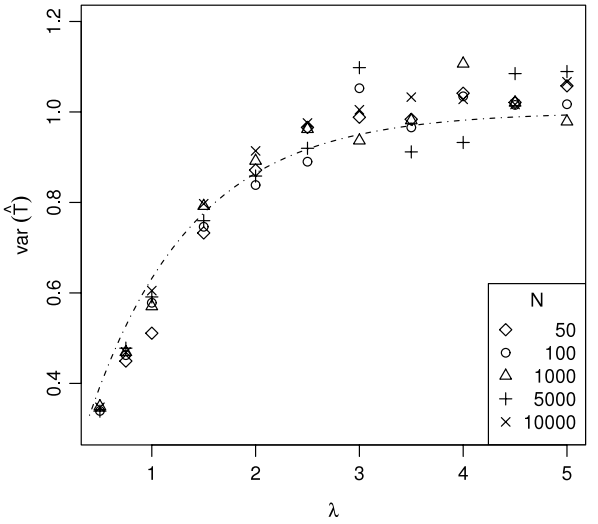

Fig. 2. Simulated values of $var(\hat{T})$ for a range of values of λ and true population size N. The broken line is the curve $1 - \exp(-\lambda)$.

noted. The results for selected ratios of true variance σ^2 to mean μ of the mixture distributions and varying study sizes are plotted in Fig. 4. Values of μ were 1, 1.3 and 1.9. As can be seen, the over-dispersion test based on $\tilde{T}$ is consistently more efficient than the test using U. An added benefit is that $\tilde{T}$ is easily calculated and avoids the potential computational burden of obtaining the maximum likelihood estimate of the mean λ.

Example 1 (*Continued*)**.** For the illegal gun-owner data of van der Heijden et al. (2003b) introduced earlier, we get $\tilde{T} = 2.30$ with a p-value of 0.0108, providing evidence for some amount of over-dispersion. The test based on U leads to the same conclusion (p-value $= 0.0019$). It has been shown that estimators developed under the assumption of homogeneity experience under-estimation bias if in fact heterogeneity is present (van der Heijden et al., 2003a; Böhning and Schön, 2005). Thus the Good–Turing estimate of 45,128 for the population size is likely to be low. Indeed, the lower bound estimator of Chao (1987), which was developed under the assumption of heterogeneity, is $n + f_1^2/2f_2 = 48{,}185$ which is a bit larger than the Good–Turing estimate. Here we see that a small amount of over-dispersion correlates with a small amount of under-estimation in the population size estimator.

Example 2. This example is from a study on dystrophin density in human muscle (Matthews and Appleton, 1993). Dystrophin, a gene product of possible importance in muscular dystrophies, may be located within muscle fibres using an electron microscope. Units (epitops) of dystrophin cannot be detected until they have been labelled by a suitable electron-dense substance; gold-conjugated antibodies which adhere to the dystrophin were used. Not all units can be labelled and

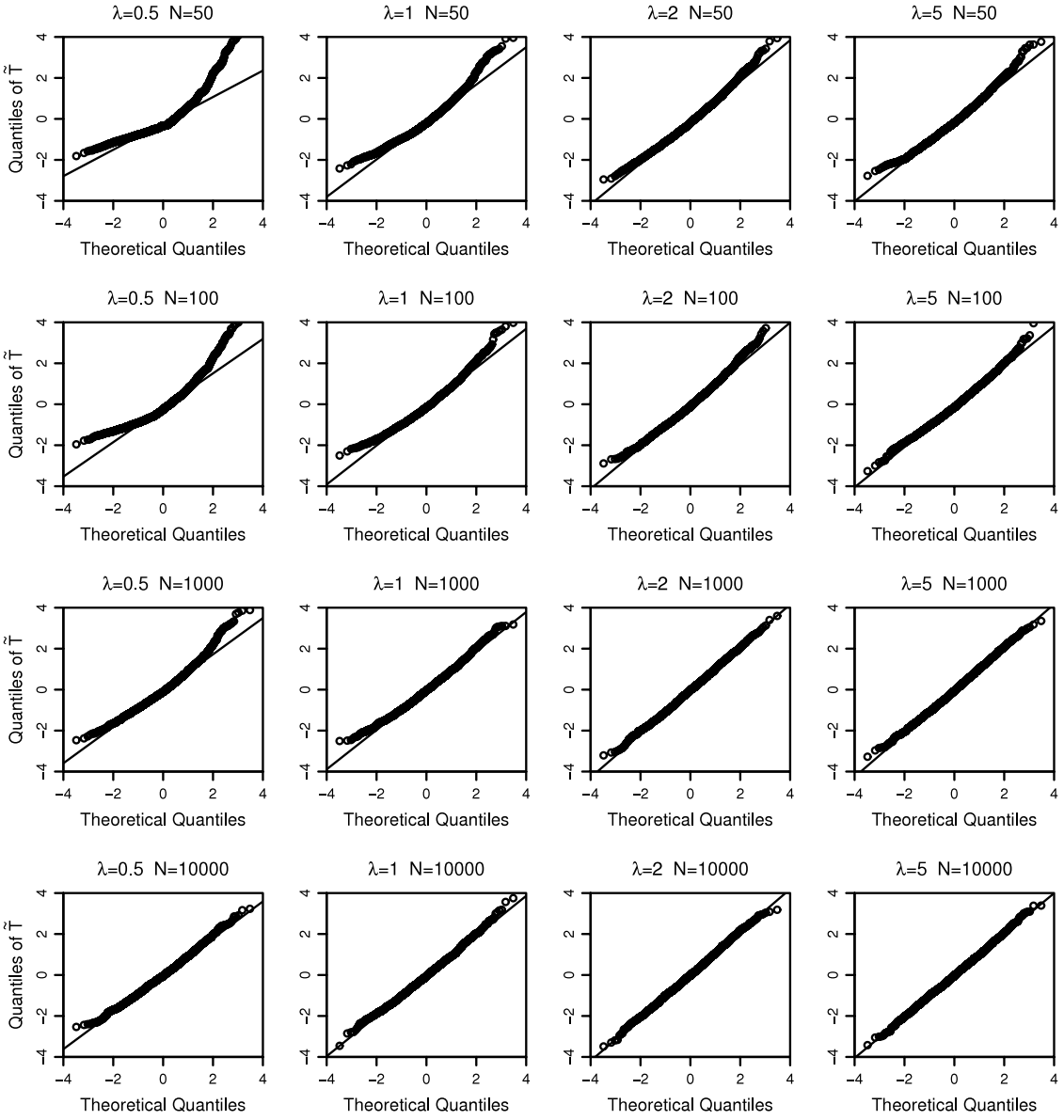

Fig. 3. Q–Q plots of simulated values of $\tilde{T}$ against standard normal quantiles.

more than one antibody molecule may attach to a dystrophin unit. To achieve an unbiased estimate of the dystrophin density, it is important to account for all labelled and unlabelled units. Shown in Table 3 is the observed count of the number of antibody molecules on each dystrophin unit within the muscle fibres of biopsy specimens taken from normal patients; f_0 represents the number of units that are unlabelled and hence not observed.

Using the data in Table 3 we get $\tilde{\lambda} = 0.9596$ and $\tilde{T} = 1.90$ (p-value $= 0.0284$), providing evidence for over-dispersion. The same conclusion is obtained using U (p-value $= 0.0192$). In such situations, we can consider use of the following modification of the Turing estimate. The rationale for this procedure is the assumption of an underlying Poisson model which has been contaminated by a component with a potential large mean value. For this dataset, there is evidence (for example from the analysis of the residuals) that the contamination is caused by the observed count of units labelled with five antibody molecules f_5. For the uncontaminated part f_1, f_2, f_3, f_4, we proceed now as follows. The original Turing estimate of p_0 is $f_1/S = (f_1/N)/\bar{X}$. Replacing $\bar{X}$ by a more robust, but consistent estimate of λ such as $\hat{\lambda}^* = (2f_2 + 3f_3 + 4f_4)/(f_1 + f_2 + f_3)$ leads to the more robust Turing estimate $\hat{N} = n/(1 - \frac{f_1/N}{\hat{\lambda}^*})$ which can be written in closed form as $\hat{N} = n + f_1/\hat{\lambda}^*$. In this example we get $\hat{\lambda}^* = 0.8947$ with associated population size estimate $\hat{N} = 334$ and the value of the over-dispersion test statistic is now $\tilde{T} = 0.007$ (p-value $= 0.4971$). This is further evidence that the observed f_5 count is a source of heterogeneity. Note that this robustified estimate not only reduces the under-estimation bias involved in the original Turing estimate, but

Table 2
Empirical type I error based on 10,000 replications of $\tilde{T}$ and U for selected values of λ and population size N.

N		$\tilde{T}$				U			
	$\lambda \rightarrow$	0.50	1	2	5	0.50	1	2	5
		$\alpha = 0.10$							
50		0.072	0.079	0.092	0.097	0.071	0.084	0.088	0.090
100		0.081	0.086	0.095	0.095	0.088	0.089	0.095	0.088
1,000		0.085	0.090	0.098	0.098	0.113	0.113	0.095	0.095
5,000		0.087	0.083	0.100	0.104	0.118	0.107	0.099	0.094
10,000		0.085	0.092	0.107	0.107	0.114	0.116	0.100	0.096
		$\alpha = 0.05$							
50		0.040	0.045	0.055	0.059	0.042	0.050	0.053	0.052
100		0.051	0.050	0.052	0.053	0.056	0.056	0.054	0.048
1,000		0.047	0.048	0.052	0.051	0.065	0.068	0.050	0.048
5,000		0.040	0.043	0.050	0.052	0.067	0.059	0.050	0.048
10,000		0.043	0.046	0.054	0.054	0.064	0.063	0.050	0.048
		$\alpha = 0.01$							
50		0.022	0.015	0.016	0.017	0.022	0.018	0.016	0.015
100		0.018	0.018	0.017	0.016	0.025	0.021	0.017	0.012
1,000		0.011	0.011	0.013	0.012	0.022	0.019	0.013	0.009
5,000		0.007	0.009	0.011	0.010	0.020	0.015	0.011	0.010
10,000		0.007	0.009	0.013	0.012	0.018	0.015	0.010	0.010

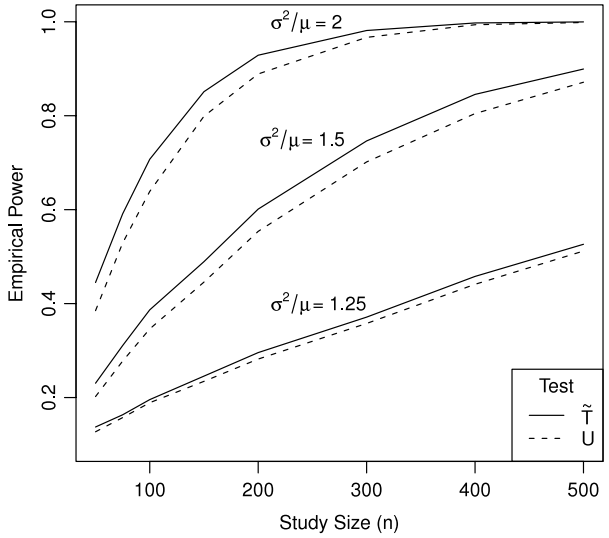

Fig. 4. Empirical power based on 10,000 replications for selected levels of over-dispersion and varying study size.

Table 3
Distribution of antibody counts attached to dystrophin units.

f_0	f_1	f_2	f_3	f_4	f_5	n
–	122	50	18	4	4	198

is also larger than the estimates obtained using any of the three estimators considered in Matthews and Appleton (1993), which were developed under homogeneity.

Example 3. This example, concerning estimation of the grizzly bear population in Yellowstone, illustrates the over-dispersion test procedure for small populations. The data, shown in Table 4 is taken from Chao and Huggins (2005) and gives the sighting frequencies of female grizzly bears with cubs-of-the-year for three different observational periods. Also shown in this table are the over-dispersion tests based on $\tilde{T}$ and on U. As in the preceding examples, the two tests lead to similar conclusions.

It is also possible to apply the proposed over-dispersion test to zero-inflated data. Zero-inflation and over-dispersion modelling have recently received a lot of attention (Deng and Paul, 2005; Hall and Berenhaut, 2002; Jansakul and Hinde, 2009; Moghimbeigi et al., 2008, 2009; Xie et al., 2009; Xiang et al., 2007). Zero-inflation leads to over-dispersion. However,

Table 4
Female grizzly bears in the yellowstone ecosystem.

Year	Counts							Over-dispersion test			
	f_1	f_2	f_3	f_4	f_5	f_6	f_7	$\tilde{T}$	p-value	U	p-value
1996	15	10	2	1	0	0	0	−0.529	0.702	−0.522	0.699
1997	13	7	4	1	3	0	1	2.543	0.005	2.447	0.007
1998	11	13	5	1	1	0	2	1.682	0.046	2.067	0.019

Table 5
DMFT distribution at the end of the BELCAP study.

School	**All**	Counts									Over-dispersion test			
		0	1	2	3	4	5	6	7	8	$\tilde{T}$	p-value	U	p-value
1	**124**	16	6	19	10	22	14	16	12	9	−0.187	0.574	−0.211	0.583
2	**127**	40	8	18	16	11	8	13	7	6	1.155	0.124	1.119	0.131
3	**136**	18	13	16	16	19	20	14	10	10	1.049	0.147	0.718	0.236
4	**132**	31	16	9	14	15	14	12	10	11	2.426	0.008	1.781	0.037
5	**155**	39	17	11	11	15	22	17	13	10	1.712	0.043	1.035	0.150
6	**123**	28	13	23	13	13	5	13	13	2	2.234	0.013	2.064	0.019
All	797	172	73	96	80	95	83	85	65	48	3.561	0.000	2.752	0.003

it is often important to determine whether there is any residual over-dispersion in the presence of zero-inflation, i.e. over-dispersion not caused by the amount of zero counts in the data, or whether there is over-dispersion given there is no zero-inflation. This question is of interest because, as in the example given below, the presence of over-dispersion which is not explained by zero-inflation or incorporated covariates will require further investigation for other associated covariates. Conversely, if there is no evidence for residual over-dispersion, the investigator can be assured that there is then no need for further covariate investigation as there is no residual variances that could be explained. We propose a test of over-dispersion for zero-inflated data by applying the over-dispersion statistic $\tilde{T}$ to zero-inflated samples with all zero counts removed. We also point out that this area has not only received interest from a theoretical side, but is also important for several application fields (Carrivick et al., 2003; Xie et al., 2001).

The simple zero-inflated Poisson model is given as $(1-p)+p\exp(-\lambda)$ if $X=0$ and $pPo(x|\lambda)$ if $X=x$ is larger than 0. It was shown (Dietz and Böhning, 2000) that inference for λ in the zero-inflation model can be fully based on the zero-truncation model (maximum likelihood estimates are identical) without any loss of efficiency. Recall that, in obtaining the over-dispersion test statistic $\tilde{T}$, we denoted the observed sample of non-zero counts by $X_1,\ldots,X_n$ and the remaining unobserved zero counts by $X_{n+1},\ldots,X_N$. For the test with zero-inflated count data, we still have $X_1,\ldots,X_n$ as the observed sample of non-zero counts but we now treat $X_{n+1},\ldots,X_N$ as the unknown zero counts attributable to the "Poisson component" of the zero-inflated distribution. In this case, the estimate for N should be less than the total number of individuals sampled.

Example 4. To illustrate our method we use data from the Belo Horizonte Caries Prevention (BELCAP) study of school children from Belo Horizonte, Brazil (Böhning et al., 1999). Our data, which is zero-inflated, consist of counts of the decayed, missing and filled teeth for each child taken at the end of the study. This count, known as the DMFT index in dental epidemiology, is commonly used to quantify dental status of an individual. The frequency distribution of the DMFT index for the six schools studied is presented in Table 5. Also given are results of over-dispersion tests for individual schools and for the pooled data from all schools (final four columns of table). The test, based on $\tilde{T}$, of the overall data from the six schools show evidence for over-dispersion (p-value $=0.0002$) not due to zero-inflation. This is indicative of an important covariate. Indeed, considering individual schools, the tests based on $\tilde{T}$ shows evidence of over-dispersion for schools 4 and 6, and to a lesser extent, school 5 and confirms that the covariate school is important—the schools had received different caries preventative treatments in this community randomized trial. The tests based on U will lead to the same conclusions, apart from school 5 where it shows no evidence of over-dispersion.

The test statistic (5) is built upon $\tilde{\lambda}=(S-f_1)/n$ which requires a good estimate of f_1. Such an estimate is possible in many data sets, particularly from capture–recapture studies, as f_1 is large relative to $f_2,\ldots,f_m$. However, it is not always the case that f_1 is large relative to the other counts. In such situations it is advisable to consider maximum likelihood estimation. So, for example, we would estimate λ using the value $\hat{\lambda}$ maximizing the zero-truncated Poisson likelihood

$$\prod_{j=1}^{m}\left(\frac{p_j}{1-p_0}\right)^{f_j}=\prod_{j=1}^{m}\left(\frac{1}{1-\exp(-\lambda)}\exp(-\lambda)\lambda^j/j!\right)^{f_j},$$

in λ. Alternatively, the maximum likelihood estimate $\hat{\lambda}$ can be found by solving the equation

$$\lambda=\frac{S}{n}(1-\exp(-\lambda)),$$

in λ iteratively by cycling between the steps (i) $\lambda^{(j)} = S/N^{(j)}$ and (ii) $N^{(j+1)} = n/\{1 - \exp(-\lambda^{(j)})\}$. This iterative solution is a special case of the EM algorithm (Böhning et al., 2005). The modified over-dispersion statistic is then obtained by simply replacing $\tilde{\lambda}$ in Eq. (5) with $\hat{\lambda}$ leading to

$$\hat{T} = \frac{S^{(2)} - S(\hat{\lambda} + 1)}{\sqrt{2S\hat{\lambda}[1 - e^{-\hat{\lambda}}]}}.$$

3. Discussion

As over-dispersion is a consequence of latent heterogeneity, a simple diagnostic test for over-dispersion can rule out the need for subsequent investigation of latent heterogeneity (Böhning, 1994), as well as the need for more complex modelling. If over-dispersion can be ruled out, the use of an estimator under homogeneity such as the Good–Turing (for the truncated situation) or the maximum likelihood estimator under homogeneity can be well justified.

Latent heterogeneity is an important issue in capture–recapture modelling since the presence of latent heterogeneity invalidates inference on the population size parameter (Böhning et al., 2005). However, a simple and general test for over-dispersion for truncated count data is currently not available in the literature. Clearly tests are available when the alternative is specific, such as a Poisson–Gamma mixture; here likelihood ratio and score tests have been developed (Deng and Paul, 2005; Jansakul and Hinde, 2002, 2009; Ridout et al., 2001; van den Broek, 1995). Although these tests might be more powerful than the proposed test, they might be testing the wrong alternative since it is seldom clear which alternative is valid, as is the case in capture–recapture approaches using truncated count modelling, for example. Hence it is beneficial to have a test procedure readily available which is valid for a wider family of alternatives. In our case the alternative is a general, arbitrary mixing distribution on the Poisson parameter. We have shown that the proposed test, under these circumstances, is consistently more powerful than similar procedures recently advocated by Best et al. (2007). In future work we will extend the approach to incorporate available covariate information.

Acknowledgements

We are grateful to the editor and two anonymous reviewers for their helpful comments and suggestions. Funding for RL was provided by the Royal Thai Government.

References

Aitkin, M., Anderson, D., Francis, B., Hinde, J., 1977. Mathematical Statistics: Basic Ideas and Selected Topics. Holden-Day, Inc., CA.
Best, D.J., Rayner, J.C.W., Thas, O., 2007. Goodness of fit for the zero-truncated Poisson distribution. Journal of Statistical Computation and Simulation 77, 585–591.
Böhning, D., 1994. A note on a test for Poisson overdispersion. Biometrika 81, 418–419.
Böhning, D., 2000. Computer-assisted Analysis of Mixtures and Applications. Meta-analysis, Disease Mapping and Others. Chapman and Hall/CRC, Boca Raton.
Böhning, D., Dietz, E., Kuhnert, R., Schön, D., 2005. Mixture models for capture–recapture count data. Statistical Methods & Applications 14, 29–43.
Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., Kirchner, U., 1999. The zero-inflated Poisson model and the DMFT index in dental epidemiology. Journal of the Royal Statistical Society, Series A 162, 195–209.
Böhning, D., Schön, D., 2005. Nonparametric maximum likelihood estimation of the population size based upon the counting distribution. Journal of the Royal Statistical Society, Series C 54, 721–737.
Cameron, A., Trivedi, P.K., 1998. Regression Analysis of Count Data, 1st ed. Cambridge University Press, Cambridge.
Carrivick, P., Lee, A., Yau, K., 2003. Zero-inflated Poisson modeling to evaluate occupational safety interventions. Safety Science 41, 53–63.
Chao, A., 1987. Estimating the population size for capture–recapture data with unequal catchability. Biometrics 43, 783–791.
Chao, A., Huggins, R.M., 2005. Modern closed-population capture–recapture models. In: Amstrup, S.C., McDonald, T.L., Manly, B. F.J. (Eds.), Handbook of Capture–recapture Analysis. Princeton University Press, Princeton, pp. 58–87.
Collett, D., 2003. Modelling Binary Data, 2nd ed. Chapman and Hall/CRC, London.
Deng, D., Paul, S., 2005. Score tests for zero-inflation and over-dispersion in generalized linear models. Statistica Sinica 15, 257–276.
Dietz, E., Böhning, D., 2000. On estimation of the Poisson parameter in zero-modified Poisson models. Computational Statistics and Data Analysis 34, 441–459.
Good, I., 1953. On the population frequencies of species and the estimation of population parameters. Biometrika 40, 237–264.
Hall, D., Berenhaut, K., 2002. Score tests for heterogeneity and overdispersion in zero-inflated Poisson and binomial regression models. The Canadian Journal of Statistics 30, 415–430.
Jansakul, N., Hinde, J., 2002. Score tests for zero-inflated Poisson models. Computational Statistics and Data Analysis 40, 75–96.
Jansakul, N., Hinde, J., 2009. Score tests for extra-zero models in zero-inflated negative binomial models. Communications in Statistics. Simulation and Computation 38, 92–108.
Lindsay, B., 1995. Mixture Models: Theory, Geometry, and Applications. In: NFS-CBMS Regional Conference Series in Probability and Statistics, Institute of Statistical Mathematics, Hayward.
Matthews, J.N.S., Appleton, D.R., 1993. An application of the truncated Poisson distribution to immunogold assay. Biometrics 49, 617–621.
Moghimbeigi, A., Eshraghian, M., Mohammad, K., McArdle, B., 2008. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. Journal of Applied Statistics 35, 1193–1202.
Moghimbeigi, A., Eshraghian, M., Mohammad, K., McArdle, B., 2009. A score test for zero-inflation in multilevel count data. Computational Statistics and Data Analysis 53, 1239–1248.
Rao, C.R., Chakravarthi, I.M., 1956. Some small sample tests of significance for a Poisson distribution. Biometrics 12, 264–282.
Ridout, M.S., Demetrio, C.G.B., Hinde, J., 2001. A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. Biometrics 57, 219–223.

Tiago de Oliveira, J., 1965. Some elementary tests of mixtures of discrete distributions. In: Patil, G.P. (Ed.), Classical and Contagious Discrete Distributions, 2nd ed. Pergamon, New York, pp. 183–187.
van den Broek, J., 1995. A score test for zero inflation in a Poisson distribuiton. Biometrics 51, 738–743.
van der Heijden, P.G.M., Bustami, R., Cruyff, M., Engbersen, G., van Houwelingen, H., 2003a. Point and interval estimation of the population size using the truncated Poisson regression model. Statistical Modelling: An International Journal 3, 305–322.
van der Heijden, P.G.M., Cruyff, M., van Houwelingen, H., 2003b. Estimating the size of a criminal population from police records using the truncated Poisson regression model. Statistica Neerlandica 57, 1–16.
Winkelmann, R., 2003. Econometric Analysis of Count Data, 4th ed. Springer, Heidelberg.
Xiang, L., Lee, A., Yau, K., McLachlan, G., 2007. A score test for overdispersion in zero-inflated Poisson mixed regression model. Statistics in Medicine 26, 1608–1622.
Xie, F., Wei, B., Lin, J., 2009. Score tests for zero-inflated generalized Poisson mixed regression model. Computational Statistics and Data Analysis 53, 3478–3489.
Xie, M., He, B., Goh, T., 2001. Zero-inflated Poisson model in statistical process control. Computational Statistics and Data Analysis 38, 191–201.