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 $\lambda$ 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 $\lambda$, 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 $\lambda$ 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 $\lambda$, 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 $var(X_+) = E(X_+)\{1 - E(X_+)\exp(-\lambda)\}$,