Accident Analysis and Prevention 43 (2011) 1738–1742

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros

Dominique Lord[a,*], Srinivas Reddy Geedipally[b,1]

[a] Zachry Department of Civil Engineering, Texas A&M University, 3136 TAMU, College Station, TX 77843-3136, USA
[b] Engineering Research Associate, Texas Transportation Institute, Texas A&M University, 3135 TAMU, College Station, TX 77843-3135, USA

ARTICLE INFO

Article history:
Received 17 January 2011
Received in revised form 1 April 2011
Accepted 2 April 2011

Key words:
Count data
Safety
Negative binomial distribution
Poisson distribution
Negative binomial–Lindley distribution

ABSTRACT

The modeling of crash count data is a very important topic in highway safety. As documented in the literature, given the characteristics associated with crash data, transportation safety analysts have proposed a significant number of analysis tools, statistical methods and models for analyzing such data. Among the data issues, we find the one related to crash data which have a large amount of zeros and a long or heavy tail. It has been found that using this kind of dataset could lead to erroneous results or conclusions if the wrong statistical tools or methods are used. Thus, the purpose of this paper is to introduce a new distribution, known as the negative binomial–Lindley (NB-L), which has very recently been introduced for analyzing data characterized by a large number of zeros. The NB–L offers the advantage of being able to handle this kind of datasets, while still maintaining similar characteristics as the traditional negative binomial (NB). In other words, the NB–L is a two-parameter distribution and the long-term mean is never equal to zero. To examine this distribution, simulated and observed data were used. The results show that the NB–L can provide a better statistical fit than the traditional NB for datasets that contain a large amount of zeros.

1. Introduction

The modeling of crash count data is a very important topic in highway safety. As documented in Lord and Mannering (2010), given the characteristics associated with crash data, transportation safety analysts have proposed a significant number of analysis tools and models for analyzing such data. Among the data issues documented in the paper, we find the one related to crash data which have a large amount of zeros and a long or heavy tail. For such datasets, the number of sites where no crash is observed is so large that traditional statistical distributions or models, such as the Poisson and Poisson-gamma or negative binomial (NB) distributions, cannot be used efficiently. The Poisson distribution tends to underestimate the number of zeros given the mean of the data, while the NB may over-estimate zeros, but under-estimate observations with a count. This is obviously dependent upon the characteristics of the tail.

The large amount of zeros observed in crash data have initially been attributed to observations or sites that can be categorized under two states: a safe state, where no crash can occur, and a non-safe state (Miaou, 1994; Shankar et al., 1997, 2003). A portion of the zero counts come from the safe state, while the rest of the zero counts come from a Poisson or NB distribution. The observations classified under the first state are considered as 'added' zeros. The zero-inflated model (both used for the Poisson and NB) has been consequently proposed to analyze this kind of dataset, usually because they provide better statistical fit (Shankar et al., 1997, 2003; Kumara and Chin, 2003). Some researchers (Warton, 2005; Lord et al., 2005, 2007) however have raised important methodological issues about the use of such models, including the fact that the safe state has a distribution with a long-term mean equal to zero, which is theoretically impossible. Lord et al. (2005) noted that the large amount of zeros can be attributed to the following factors: (1) sites with a combination of low exposure, high heterogeneity, and sites categorized as high risk; (2) analyses conducted with small time or spatial scales; (3) data with a relatively high percentage of missing or mis-reported crashes; and (4) crash models with omitted important variables. More recently, Mayshkina and Mannering (2009) have proposed a zero-state Markov switching model, which overcomes some of the criticisms discussed above, for analyzing longitudinal datasets characterized by a large number of zeros.

In cases in which the characteristics of the dataset cannot or is very difficult to be changed (as it will be discussed further below), the large number of zeros could still create a lot of difficulties for properly analyzing such dataset. This could obviously lead to erroneous results or conclusions if the wrong statistical tools or

* Corresponding author. Tel.: +1 979 458 3949; fax: +1 979 845 6481.
E-mail addresses: d-lord@tamu.edu (D. Lord), srinivas-g@ttimail.tamu.edu (S.R. Geedipally).
[1] Tel.: +1 979 862 1651; fax: +1 979 845 6006.

0001-4575/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2011.04.004

methods are used. Thus, the purpose of this paper is to introduce a new distribution that has very recently been introduced for analyzing data characterized by a large number of zeros. This mixed distribution is known as the NB–Lindley (NB–L) distribution (Zamani and Ismail, 2010), which as the name implies, is a mixture of the NB and the Lindley distributions (Lindley, 1958; Ghitany et al., 2008). This two-parameter distribution has interesting and sound theoretical properties in which the distribution is characterized by a single long-term mean that is never equal to zero and a single variance function, similar to the traditional NB distribution. The properties of the NB–Lindley distribution are examined using simulated and observed data and a discussion is presented about the potential use of the NB-L distribution for traffic safety analyses. It important to point out that all documented distributions, such as the Poisson-gamma, Poisson-lognormal, Poisson-Pascal or the NB–L in highway safety research are in fact used as an approximation to describe the crash process. This process is known as the Poisson trials with unequal probability of events (See Lord et al., 2005, for additional details).

The paper is divided into five sections. Section 2 describes the characteristics of the NB–Lindley distribution. Section 3 presents the comparison analysis between the Poisson, NB and NB–L using simulated and observed data. Section 4 provides additional information for future work. Section 5 summarizes the study results.

2. Characteristics of the negative binomial–Lindley distribution

As discussed above, the NB–L distribution is a mixture of negative binomial and Lindley distributions. This mixed distribution has a thick tail and can be used when the data contains large number of zeros.

The negative binomial distribution is a mixture of Poisson and gamma distribution. The probability mass function (pmf) of the NB distribution can be given as:

$$P(Y=y;\phi,p)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\times y!}(1-p)^{\phi}(p)^{y};\quad \phi>0,\quad 0<p<1 \tag{1}$$

The parameter 'p' is defined as the probability of success in each trial and is given as:

$$p=\frac{\mu}{\mu+\phi} \tag{2}$$

where, $\mu = E(Y)$ = mean; and, ϕ = inverse dispersion parameter.

Then, it can be shown that the variance is (Casella and Berger, 1990):

$$\mathrm{Var}(Y)=\phi\frac{p}{(1-p)^2}=\frac{1}{\phi}\mu^2+\mu \tag{3}$$

Using Eqs. (2) and (3), the pmf of the NB distribution can be re-parameterized this way:

$$P(Y=y;\mu,\phi)=\frac{\Gamma(\phi+y)}{\Gamma(\phi)\Gamma(y+1)}\left(\frac{\phi}{\mu+\phi}\right)^{\phi}\left(\frac{\mu}{\mu+\phi}\right)^{y} \tag{4}$$

The pmf in Eq. (4) is the one normally used for crash count data.

The Lindley distribution is a mixture of exponential and gamma distribution (Lindley, 1958; Ghitany et al., 2008; Zamani and Ismail, 2010). The pmf of the Lindley distribution can be defined as follows:

$$f(X=x;\theta)=\frac{\theta^2}{\theta+1}(1+x)e^{-\theta x};\quad \theta>0,\quad x>0 \tag{5}$$

A random variable Z is assumed to follow a NB–L (r,θ) distribution when the following conditions satisfy:

$$Z\sim \mathrm{NB}\left(r, P=1-e^{-\lambda}\right) \text{ and } \lambda\sim \text{Lindley}\left(\theta\right)$$

The pmf of the NB–L distribution is given as (Zamani and Ismail, 2010):

$$P(Z=z;r,\theta)=\frac{\Gamma(r+z)}{\Gamma(r)\times z!}\frac{\theta^2}{\theta+1}\sum_{j=0}^{z}\frac{\Gamma(z+1)}{\Gamma(j+1)\times\Gamma(z+j+1)}(-1)^{j}\times\frac{\theta+r+j-1}{(\theta+r+j)^2} \tag{6}$$

The parameter 'r' is the shape parameter of NB–L distribution, similar to the inverse dispersion parameter 'ϕ' of the NB distribution. The parameter 'θ', in combination with shape parameter 'r' dictates the mean and variance of the NB–L distribution.

The first moment (i.e., the mean) of the NB–L (r,θ) is given as:

$$E(Z)=r\left[\frac{\theta^3}{(\theta+1)(\theta-1)^2}-1\right] \tag{7}$$

It should be noted that $E[Z] = E[Y] = \mu$

The second moment of the NB–L (r,θ) is given as:

$$E(Z^2)=(r+r^2)\left[\frac{\theta^2(\theta-1)}{(\theta+1)(\theta-2)^2}\right]-(r+2r^2)\left[\frac{\theta^3}{(\theta+1)(\theta-1)^2}\right]+r^2 \tag{8}$$

The variance of the NB–L (r,θ) is calculated as:

$$\mathrm{Var}(Z)=E(Z^2)-(E(Z))^2 \tag{9}$$

As described above, the NB–L distribution is a two-parameter distribution, which implies that the mean is never equal to zero. The NB–L is in fact an extension of the NB distribution.

To estimate the parameters r and θ, Eqs. (7) and (8) need to be solved iteratively and both parameters should be greater than 0 ($r>0$ and $\theta>0$). The parameters can also be estimated by solving the likelihood function, but the function is difficult to manipulate, since the partial derivatives contain multiple solutions. Additional work is therefore needed for finding the optimal solution among all the possible ones. More detailed information can be found in Zamani and Ismail (2010).

3. Application of the negative binomial–Lindley distribution

This section presents the comparison analysis results between the Poisson, NB and NB–L distributions using simulated and observed data.

3.1. Simulated data

The simulation protocol is the same one used by Lord et al. (2005), which consisted in simulating a Poisson distribution and add observations with the value zero to 'simulate' a two-state process, one of which is characterized by a long-term equal to zero. For this example, count data with 100 observations were simulated using a Poisson distribution with a mean equal to 0.50. Then, 100, 150 and 200 observations with the value zero were added to the data. The simulated data are summarized in Fig. 1. The original simulated data produced 57 observations with the value zero, 34 with the value 1, 8 with the value 2, and 1 with a value above 3 (5 to be exact).

The Poisson, NB and NB–L distributions were fitted based on the simulated data. Using the mean and variance of the data, the parameters were estimated with Eqs. (2) and (3) for the NB distribution and Eqs. (7)–(9) for the NB–L distribution. After the parameters