Statistics and Probability Letters 87 (2014) 158–166

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

A COM–Poisson type generalization of the binomial distribution and its properties and applications

Patrick Borges [a,*], Josemar Rodrigues [b], Narayanaswamy Balakrishnan [c,d], Jorge Bazán [b]

[a] Department of Statistics, Federal University of Espírito Santo, Vitória, Brazil
[b] Department of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
[c] Department of Mathematics and Statistics, McMaster University, Hamilton, Ontario, Canada
[d] Department of Statistics, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Article history:
Received 20 January 2014
Accepted 20 January 2014
Available online 29 January 2014

Keywords:
COM–Poisson–binomial distribution
Dependent Bernoulli variables
Correlation coefficient
Exponential family
Weighted Poisson distributions

ABSTRACT

Shmueli et al. (2005) introduced the COM–Poisson–binomial distribution, but they did not study the mathematical properties of this family of distributions. In this paper, we discuss some properties and an asymptotic approximation of it by the COM–Poisson distribution. Moreover, three datasets are also considered.

1. Introduction

Usually the binomial and Poisson distributions are used to analyze discrete data. However, it seems wise to consider flexible alternative models to take into account the overdispersion or underdispersion (see Hinde and Demetrio (1998)). For this reason, the binomial and Poisson distributions have been generalized in several ways to handle the problem of dispersion inherent in the analysis of discrete data that may arise in the presence of aggregation of the individuals. For instance:

(i) in plant selection study, the association among two plants arises when they compete for the quantity of nutrients;
(ii) in biological study (see Yakovlev and Tsodikov (1996) and Borges et al. (2012)), it is usually assumed that cells in a tissue are independent. However, the biological independence assumption may not be true when the dynamics of the cell population of a normal tissue is considered. It is therefore desirable to construct new models with strong biological interpretation of the dependence incorporated in the carcinogenesis process.

The binomial distribution has been generalized in various ways. Rudolfer (1990), Madsen (1993) and Luceño and Ceballos (1995) have summarized most of these generalizations. Among these extensions, there are the multiplicative and additive generalized binomial distributions which were derived by Altham (1978). The probability mass function (pmf) of the multiplicative binomial distribution is a multiplication of its pmf by a factor. It makes the variance greater or less than the

* Correspondence to: Departamento de Estatística, Universidade Federal do Espírito Santo - UFES. Av. Fernando Ferrari 514, Goiabeiras, CEP 29075-910, Vitória ES, Brazil.
E-mail addresses: patrick@cce.ufes.br, patrickborges@yahoo.com.br (P. Borges).

0167-7152/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spl.2014.01.019

corresponding binomial variance depending on the values of the factor. On the other hand, the additive binomial distribution is a mixture of three conventional binomial models. Altham (1978) developed the correlated binomial model by correcting the binomial model via the method of Bahadur (1961) to encompass dependent Bernoulli variables. A three-parameter binomial distribution was derived by Paul (1985, 1987), which is a generalization of the binomial, beta-binomial and the correlated binomial distribution proposed by Kupper and Haseman (1978). Ng (1989) developed the modified binomial distributions. In this approach, the binomial distribution is changed and the resulting distribution becomes more spread out (indicating positive correlation among the Bernoulli variables), or more peaked (indicating negative correlation among the Bernoulli variables) than the binomial distribution. A four-parameter binomial distribution was derived by Fu and Sproule (1995). This new distribution assumes values between α and β for $\alpha < \beta$, rather than the usual values 0 or 1. Lindsey (1995) and Luceño and Ceballos (1995) proposed a generalized binomial distribution which is discussed in detail in Diniz et al. (2010). Chang and Zelterman (2002) generalized the binomial distribution by considering Bernoulli variables with probability of success depending on the previous one. Tsai et al. (2003) presented a model that studies the overall error rate when testing multiple hypotheses. This model involves the distribution of the sum of dependent Bernoulli trials, and it is approximated by means of a beta-binomial structure. Instead of using the beta-binomial model, Gupta and Tao (2010) derived the exact distribution of the sum of dependent Bernoulli variables and not identically distributed. Another extension of the binomial distribution is the COM–Poisson–binomial distribution (CMPB, for short) introduced in Shmueli et al. (2005), but they did not study the mathematical properties of this family; they are studied in detail in this paper. A recent application of CMPB distribution can be found in Kadane and Naeshagen (2013).

The CMPB distribution arises as the conditional distribution of a COM–Poisson variable (Conway and Maxwell, 1962) given a sum of two COM–Poisson variables with the same dispersion parameter. It generalizes the binomial distribution and can be interpreted as the sum of dependent Bernoulli variables with a specific joint distribution (see Remark 1). The dispersion parameter governs the correlation among the Bernoulli variables, with the overdispersion and underdispersion being relative to the binomial distribution. The CMPB distribution is appealing from a theoretical point of view since it belongs to the exponential and weighted Poisson families (Castillo and Pérez-Casany, 1998, 2005), and the sufficient statistic is defined by the mean and the log-geometric mean of the data. We refer to Barndorff-Nielsen (1978) for a general theory of exponential families.

The rest of this paper is organized as follows. In Section 2, we present the CMPB distribution along with its mathematical properties. Section 3 describes the maximum likelihood method for estimating the parameters. In Section 4, we apply the CMPB distribution to three real datasets and show that this model provides an excellent fit to these datasets. Finally, some concluding remarks are made in Section 5.

2. The CMPB distribution and its properties

The probability mass function (pmf) of the CMPB distribution (Shmueli et al., 2005) is given by

$$\mathbb{P}[X = x|m, p, \nu] = \frac{\binom{m}{x}^{\nu} p^x (1-p)^{m-x}}{\sum_{k=0}^{m} \binom{m}{k}^{\nu} p^k (1-p)^{m-k}}, \quad x = 0, 1, \ldots, m, \tag{1}$$

for $m \in \mathbb{Z}^+$ (set of known non-negative integers), $p \in (0, 1)$ and $\nu \in \Re$. For $\nu = 1$, we have the usual binomial distribution. The values of $\nu > 1$ correspond to underdispersion (the mean is greater than the variance) while values of $\nu < 1$ represent overdispersion (the variance exceeds the mean) with respect to the binomial distribution. For $\nu \to \infty$, the pmf is concentrated at the point mp and for $\nu \to -\infty$ it is concentrated at 0 or m. Fig. 2.1 presents the pmf of the CMPB distribution for $m = 6$ and different choices of p and ν.

Remark 1. The CMPB distribution can be interpreted as a sum of equicorrelated Bernoulli variables Z_i ($i = 1, \ldots, m$) with joint distribution (see Shmueli et al. (2005))

$$\mathbb{P}[Z_1 = z_1, \ldots, Z_m = z_m] = \frac{\binom{m}{x}^{\nu-1} p^x (1-p)^{m-x}}{\sum_{z_1=0}^{1} \cdots \sum_{z_m=0}^{1} \binom{m}{x}^{\nu-1} p^x (1-p)^{m-x}}, \qquad \boldsymbol{z} = (z_1, \ldots, z_m) \in \{0, 1\}^m, \tag{2}$$

where $x = \sum_{i=1}^{m} z_i$. The measure of linear association between Bernoulli variables, i.e., the correlation between Z_i and Z_j, say $\rho = \mathbb{C}orr(Z_i, Z_j)$, is given by

$$\rho = \frac{\mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j]}{\sqrt{\mathbb{V}[Z_i]\mathbb{V}[Z_j]}} = \frac{p(1-p)(1-4^{\nu-1})}{(p + (1-p)2^{\nu-1})(1-p(1-2^{\nu-1}))}, \quad i \neq j, i, j = 1, \ldots, m, \tag{3}$$