Computational Statistics and Data Analysis 59 (2013) 41–51

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

On computing the distribution function for the Poisson binomial distribution

Yili Hong

Department of Statistics, Virginia Tech, Blacksburg, VA 24061, USA

ARTICLE INFO

Article history:
Received 29 September 2011
Received in revised form 10 August 2012
Accepted 4 October 2012
Available online 12 October 2012

Keywords:
Characteristic function
k-out-of-n system
Longevity risk
Normal approximation
Sum of independent random indicators
Warranty returns

ABSTRACT

The Poisson binomial distribution is the distribution of the sum of independent and non-identically distributed random indicators. Each indicator follows a Bernoulli distribution and the individual probabilities of success vary. When all success probabilities are equal, the Poisson binomial distribution is a binomial distribution. The Poisson binomial distribution has many applications in different areas such as reliability, actuarial science, survey sampling, econometrics, etc. The computing of the cumulative distribution function (cdf) of the Poisson binomial distribution, however, is not straightforward. Approximation methods such as the Poisson approximation and normal approximations have been used in literature. Recursive formulae also have been used to compute the cdf in some areas. In this paper, we present a simple derivation for an exact formula with a closed-form expression for the cdf of the Poisson binomial distribution. The derivation uses the discrete Fourier transform of the characteristic function of the distribution. We develop an algorithm that efficiently implements the exact formula. Numerical studies were conducted to study the accuracy of the developed algorithm and approximation methods. We also studied the computational efficiency of different methods. The paper is concluded with a discussion on the use of different methods in practice and some suggestions for practitioners.

1. Introduction

1.1. Motivation

The Poisson binomial distribution describes the distribution of the sum of independent and non-identically distributed random indicators. Each indicator is a Bernoulli random variable and the individual probabilities of success vary. A special case of the Poisson binomial distribution is the ordinary binomial distribution, when all success probabilities are equal. The Poisson binomial distribution has many applications in different areas such as reliability, actuarial science, survey sampling, econometrics, and so on. The following gives examples from different areas.

- In some reliability applications, it is often of interest to predict the total number of failures for a fleet of products in the field. Hong et al. (2009) considered the prediction for the total number of field failures for a fleet of high-voltage power transformers. Due to the staggered entry of units into service, individual units in the field have different failure probabilities at a specified future time. Thus the total number of field failures follows a Poisson binomial distribution.
- In actuarial science, the total payout of an insurance company is often related to the Poisson binomial distribution. For example, Pitacco (2007) considered a one-year insurance coverage only providing a death benefit for n insureds. Let C denote the payout due at each death. The individual payout is either 0 or C with probability $1-p_j$ and p_j, respectively, where the death probability p_j varies from individual to individual. Assuming that the individual lifetimes are independent, the total payout for those n insureds is C times the total number of deaths which follows the Poisson binomial distribution.

E-mail address: yilihong@vt.edu.

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.10.006

- In econometrics, it is sometimes of interest to predict the number of corporation defaults (e.g., Duffie et al., 2007). The default probabilities differ from corporation to corporation because each corporation has its own unique situation on assets, debts, stock returns and so on. The number of corporation defaults at a future time also follows a Poisson binomial distribution.
- In engineering, Fernández and Williams (2010) provided several interesting examples such as multi-sensor fusion and reliability of k-out-of-n systems, which are related to the Poisson binomial distribution.
- In survey sampling, Chen and Liu (1997) presented an example where the inclusion probabilities of sampling units are different. The total number of units in the sample follows a Poisson binomial distribution.
- The Poisson binomial distribution also has wide applications in areas such as data mining of uncertain databases (Tang and Peterson, 2011), bioinformatics (Niida et al., 2012), and wind energy (Bossavy et al., 2012).

While the Poisson binomial distribution has many applications in different disciplines, the computing of the cumulative distribution function (cdf) of the distribution is not straightforward. Because the individual probabilities of success vary, the naive way of computing the cdf by using enumeration is not practical, even when the number of indicators is small (i.e., around 30). Approximation methods such as the Poisson approximation and normal approximations have been used in literature. There are situations, however, in which approximation methods do not perform well. Thus it is desirable to have a method to compute the exact values of the cdf. It is also useful to know in which situation approximation methods work well. In applications such as predictions for the number of failures and corporation defaults, the number of indicators is usually large. Thus the efficiency of algorithms for computing the exact values of the cdf is also important. This motivates us to provide efficient methods to compute the exact values of the cdf of the Poisson binomial distribution.

1.2. Related literature and this work

The study on the Poisson binomial distribution has a long history. Le Cam (1960) provided an upper bound for the error of the Poisson approximation. Normal approximations are widely used in practice. Volkova (1996) gave a normal approximation with second order correction and provided an upper bound for the error of the approximation. Hong et al. (2009) and Hong and Meeker (2010) applied the approximation in Volkova (1996) to warranty prediction applications. Recursive formulae are available in literature to compute the exact values of the cdf of the Poisson binomial distribution. For example, Barlow and Heidtmann (1984) described a recursive formula for computing the cdf. Chen et al. (1994) provided another recursive formula. Details for these recursive formulae are described in Section 2.5. Fernández and Williams (2010) gave a closed-form expression for the cdf using the technique of polynomial interpolation and the discrete Fourier transform.

The contribution of this paper is summarized as follows.

- We propose a simple derivation for an exact formula for the cdf of the Poisson binomial distribution, which gives the same form as that in Fernández and Williams (2010).
- We develop an algorithm that efficiently implements the exact formula, which outperforms existing methods.
- Numerical studies were conducted to compare the accuracy of the algorithm and approximation methods. We also compared the computational efficiency of different methods.
- Based on the numerical studies, we provide a discussion on the advantages and disadvantages of different methods and some guidelines for practitioners.

The statistical software R (2012) is widely used. There was no package, however, for computing the Poisson binomial distribution function. We developed an R package that efficiently implements both exact methods and approximation methods. The package can be downloaded from the R website, see Section 5 for more details.

1.3. Overview

The rest of the paper is organized as follows. Section 2 describes several exact methods for computing the cdf and algorithms for their efficient implementations. Section 3 describes several approximation methods based on the Poisson and normal approximations. Section 4 conducts a comprehensive numerical study to assess the performance of various methods in terms of accuracy and efficiency. Section 5 discusses software implementation for both the exact and approximation methods. Section 6 provides some concluding remarks and suggestions for practitioners.

2. Exact methods

2.1. Notation

Let $I_j,\ j = 1, \ldots, n$ be a series of n independent and non-identically distributed random indicators. In particular,

$$I_j \sim \text{Bernoulli}(p_j), \quad j = 1, \ldots, n, \tag{1}$$

where $p_j = \Pr(I_j = 1)$ is the success probability of indicator I_j and not all p_j's are equal. The Poisson binomial random variable N is defined as the sum of independent and non-identically distributed random indicators (i.e., $N = \sum_{j=1}^{n} I_j$). Note that N can take values in $\{0, 1, \ldots, n\}$.