1st International Congress on Cancer Prevention
ICCP 2024
Zanjan, Iran
18-19 September 2024
University of Zanjan

# A novel multivariate Poisson distribution: forecasting the number of cancer deaths in Iran

**Maryam Sharafi[1], Zohre Shishebor[*2]**

[1] Department of Statistics, Shiraz University, Shiraz, Iran, msharafi@shirazu.ac.ir
[2] Department of Statistics, Shiraz University, Shiraz, Iran, shisheb@shirazu.ac.ir
* Corresponding Author

**Abstract**

Forecasting the number of cancer deaths in the future years will be the primary basis for policy-making in the field of cancer prevention. The forecasting can be made using statistical models. Accordingly, this article introduces a new multivariate Poisson distribution. This distribution has Poisson marginals and a joint probability mass function that is easier to understand than previously known distributions. The article also discusses some statistical properties of the distribution. A bivariate integer-valued autoregressive model (BINAR(1)) is constructed based on the bivariate version of the proposed distribution and the model is applied to forecast the number of deaths caused by cancer in Iran based on the number of deaths from 2011 to 2018.

**Keywords:** Bivariate integer-valued autoregressive time series, Cancer, Forecasting, Multivariate distribution, Multivariate Poisson distribution.

## 1- Introduction

A multivariate distribution is a probability distribution that involves multiple random variables simultaneously. The multivariate distributions are used in many statistical analyses for various applications: Market Basket Analysis for analyzing the relationships between items purchased together by customers [1], Genetics and Genomics for modeling the distribution of genetic variants across populations [2], in image analysis and computer

vision for modeling the distribution of pixel intensities or colors in images [3]. Recently, the applications of multivariate Poisson distributions have increased [4]. They are employed as models to describe count data with a positive correlation structure. However, the computational complexity in calculating the multivariate Poisson probability mass function prevents their widespread use in these counting models, for more detail see for example [5], and [6]. The importance of Poisson distribution and applications of multivariate discrete distributions motivated us to define a new multivariate distribution with marginal Poisson distribution. Although the new distribution is useful in many fields, we mention its application in counting time series and forecasting.

Time series models are widely used for analyzing time-dependent data. In particular, count time series, which are integer-valued, are a valuable type of time series for studying data obtained from random counting processes. These types of time series are applied in various fields, including economics, social sciences, behavioral health, electronic health records, and life insurance. Examples of data suitable for analysis using count time series include the number of children in a family, the frequency of doctor visits in a week, the number of days absent from work for employees, and the number of deaths due to a disease. On the other hand, several econometric methods are available for analyzing such data, including the Poisson, geometric, and negative binomial models. They can provide useful insights not obtained from the standard linear time series models. For the first time, the integer-valued autoregressive process of order one, denoted by INAR(1) in brief, was introduced by [7] based on the binomial thinning operator, which is introduced by [8]. Extensions of these INARs to the bivariate INARs have been accomplished significantly in the same trend. [9] and [10] introduced a bivariate integer-valued autoregressive process of order 1 (BINAR(1)) with bivariate Poisson and bivariate Poisson Weighted Exponential distributions.

The article is structured in the following manner: Section 2 includes the definition of the distribution, the study of its properties, and the definition of the BINAR(1) model with the bivariate version of the proposed multivariate Poisson distribution for innovation of the model. Section 3 deals with fitting the BINAR model to bivariate time series data about the number of male deaths and total deaths due to cancer in Iran from 2011 to 2018 to forecast the number of deaths in future years. The conclusion of the article is given in section 4.

## 2- Theory Part

In this section, we first define a new multivariate Poisson distribution and discuss its properties. Then, based on this, we proposed a bivariate integer-valued autoregressive model.

### 2-1- Definition and some properties

Several definitions of the multivariate Poisson distribution have already been established in the existing literature. However, this article introduces a new approach for defining a multivariate Poisson distribution. Our methodology considers an independent sequence $\{X_i\}_{i\geq 1}$ where each $X_i$ follows a Poisson distribution with parameter $\lambda_i \geq 0$. We define $Y_j$ as the sum of $X_i$ from i = 1,...,j, for j = 1,...,m, and represent the resulting vector as $\boldsymbol{W} = (Y_1, \dots, Y_m)$.

**Definition1**: The random vector $\boldsymbol{W} = (Y_1, \dots, Y_m)$ is considered to follow a multivariate Poisson distribution with parameter $\boldsymbol{\theta} = (\lambda_1, \dots, \lambda_m)$ with $\lambda_i \geq 0$; $i = 1, \dots, m$, if the joint probability mass function (pmf) is defined as:

$$f(y_1, \dots, y_m) = P(Y_1 = y_1, \dots, Y_m = y_m) = \frac{e^{-\sum_{i=1}^{m}\lambda_i}\prod_{i=1}^{m}\lambda_i^{(y_i - y_{i-1})}}{\prod_{i=1}^{m}(y_i - y_{i-1})!}, \qquad (1)$$

where $y_0 = 0$ and $y_i \geq y_{i-1}$ ; $y_i = 0,1, \dots,$ for $i = 1, \dots, m$.

We denote as $\boldsymbol{W} \sim NMP(\lambda_1, \dots, \lambda_m)$. The statistical properties of the distribution are outlined in the following propositions. Note that the proofs of these propositions are not included here due to space constraints.

**Proposition 1**: If $\boldsymbol{W} \sim NMP(\lambda_1, \dots, \lambda_m)$, then

$$M_{\boldsymbol{W}}(t_1, \dots, t_m) = E\left(e^{\sum_{i=1}^{m} t_i Y_i}\right) = e^{\lambda_1 e^{\sum_{i=1}^{m} t_i} + \lambda_2 e^{\sum_{i=2}^{m} t_i} + \dots + \lambda_m e^{t_m} - \sum_{i=1}^{m}\lambda_i}. \qquad (2)$$

**Proposition 2**: If $\boldsymbol{W} \sim NMP(\lambda_1, \dots, \lambda_m)$, then:

I. $Y_j \sim P(\sum_{k=1}^{j}\lambda_k)$.

II. $(Y_i, Y_j) \sim NMP(\sum_{k=1}^{i}\lambda_k, \sum_{k=i+1}^{j}\lambda_k)$ ; $j \geq i$ .

III. $\mathrm{Cov}(Y_i, Y_j) = \sum_{k=1}^{i}\lambda_k$; $j \geq i$.

IV. $\mathrm{Corr}(Y_i, Y_j) = \sqrt{\frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{j}\lambda_k}}$ ; $j \geq i$, note that the correlation is positive.

**Proposition 3**: Suppose $\boldsymbol{W} \sim NMP(\lambda_1, \dots, \lambda_m)$. Then

I. $Y_i | Y_j = y_j \sim B\left(y_j, \frac{\sum_{k=1}^{i}\lambda_k}{\sum_{k=1}^{j}\lambda_k}\right)$ ; $j \geq i$.

II. $Y_j|Y_i = y_i \overset{D}{=} Z + y_i \quad ; \quad Z \sim P\left(\sum_{k=i+1}^{j} \lambda_k\right) \quad ; j \geq i.$

III. $E\left(Y_j \middle| Y_i = y_i\right) = y_i + \sum_{k=i+1}^{j} \lambda_k$, which is a linear function of $y_i$.

IV. $Var\left(Y_j \middle| Y_i = y_i\right) = \sum_{k=i+1}^{j} \lambda_k$.

**2-2 - The BINAR(1) model**

Suppose that $\boldsymbol{Y}_t = \left(Y_{t,1}, Y_{t,2}\right);\ t = 1,2,\ldots$ is a bivariate integer-valued time series (BINAR) as follows

$$Y_{t,1} = \alpha_1 o Y_{t-1,1} + \epsilon_{t,1} \qquad (3)$$
$$Y_{t,2} = \alpha_2 o Y_{t-1,2} + \epsilon_{t,2}\ ,$$

where $\alpha_i \in (0,1)$ for $i = 1,2,$ and "$o$" is a binomial thinning operator that has a definition $\alpha o Y = \sum_{i=1}^{Y} W_i$ , $\{W_i\}$ is a sequence of i.i.d random variables with Bernoulli($\alpha$). We assumed that the innovation vector $\boldsymbol{\epsilon}_t = \left(\epsilon_{t,1}, \epsilon_{t,2}\right)$ follows NMP($\lambda_1, \lambda_2$) and $\epsilon_{t,j}$ is independent of $Y_{s,j}$; j=1, 2, for each fixed t and $s < t$. Also, innovations are independent of the counting series in the binomial thinning operator.

The conditional joint pmf of the process is given by

$$p(\boldsymbol{y}_t|\boldsymbol{y}_{t-1}) = \sum_{k=0}^{u}\sum_{s=0}^{v} p_1(k)\, p_2(s) P\left(\epsilon_{t,1} = y_{t,1} - k, \epsilon_{t,2} = y_{t,2} - s\right), \qquad (4)$$

where $u = \min(y_{t,1}, y_{t-1,1})$, $v = \min(y_{t,2}, y_{t-1,2})$ , $p_i(.)$ is the density function of the binomial distribution with parameters $y_{t-1,i}$ and $\alpha_i$, Bin($y_{t-1,i}, \alpha_i$); i=1,2 and $P\left(\epsilon_{t,1} = \epsilon_1, \epsilon_{t,2} = \epsilon_2\right)$ is the pmf of NMP($\lambda_1, \lambda_2$) in relation (1) at point $(y_{t,1} - k, y_{t,2} - s)$.

For $i = 1,2$ the conditional mean and conditional variance of the component of the process are calculated as

$$E\left(Y_{t,i}\middle|Y_{t-1,i}\right) = \alpha_i Y_{t-1,i} + \sum_{k=1}^{i} \lambda_k, \qquad (5)$$
$$Var\left(Y_{t,i}\middle|Y_{t-1,i}\right) = \alpha_i(1-\alpha_i) Y_{t-1,i} + \sum_{k=1}^{i} \lambda_k. \qquad (6)$$

One of the benefits of time series is forecasting. For this purpose, we should first estimate the parameters of the model. We estimate the parameters $(\alpha_1, \alpha_2, \lambda_1, \lambda_2)$ by the conditional least squares (CLS) method and substitute in relation (5). In this way, forecasting can be made by

$$\hat{E}\left(Y_{t,i}\middle|Y_{t-1,i}\right) = \hat{\alpha}_i Y_{t-1,i} + \sum_{k=1}^{i} \hat{\lambda}_k \ ;\ i = 1,2,\ \text{t=2,3,…..} \qquad (7)$$

## 3-Forecasting the number of deaths due to cancer

Forecasting cancer deaths is a complex task that requires analyzing historical data on cancer incidence, mortality rates, population demographics, and trends in cancer diagnosis and treatment. Statistical models, including time series analysis and regression models [4], are used to predict future trends. Age, gender, ethnicity, socioeconomic status, lifestyle choices, and environmental exposures are considered to improve accuracy. Understanding the underlying factors contributing to cancer mortality helps researchers and healthcare professionals develop strategies for prevention, early detection, and improved treatment, to reduce deaths from this devastating disease.

In this section, we consider a bivariate time series that includes the number of male deaths ($Y_{t,1}$) and total deaths ($Y_{t,2}$) due to cancer in Iran from 2011 to 2018, [11]. The descriptive statistics of data are given in Tab.1.

**Table 1: The descriptive statistics of data**

| Data | Mean | Standard Deviation | Median |
|---|---|---|---|
| No. Male Deaths | 20327 | 2671.423 | 20724 |
| No. Total Deaths | 34788 | 4722.582 | 35349 |

The correlation coefficient between $Y_{t,1}$ and $Y_{t,2}$ is 0.9988, indicating a strong positive correlation. The p-value of the correlation significance test is $3.811\times10^{-9}$, which is highly significant. The Kolmogorov-Smirnov test was used to assess the goodness of fit of the Poisson distribution for $Y_{t,1}$ and $Y_{t,2}$. The p-values obtained were 0.02259, which are greater than 0.01, suggesting that they follow a Poisson distribution. Without losing the generality of the problem, we divide the data into 1000 and then round them and consider the BINAR(1) model for the data. The estimate of the parameters is obtained as $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\lambda}_1, \hat{\lambda}_2)$ = (0.9311,0.9585,2.5,0.7541), and by relation (7), we forecast the average number of male deaths and total deaths of 2019 as 24848 and 42555. It should be noted that the calculations of this section were done using the statistical language package R version 4.3.3.

## 4- Conclusions

In this work, we introduced a new multivariate Poisson distribution and studied some properties. Based on the bivariate version of the new distribution, we proposed a bivariate integer-valued time series model and, used the model to forecast the number of male deaths and total deaths due to cancer in Iran.

## References


[1] G.J. Russel, A. Petersen, "Analysis of cross-category dependence in market basket selection", Journal of Retailing, 76(2000) 367-392

[2] P. Behrouzi, "Extensions of graphical models with applications in genetics and genomics", University of Groningen, (2018)

[3] I.L. Dryden, K. V. Mardia, " Multivariate Shape Analysis", Sankhyā, 55(1993) 460-480

[4] P.-F. Su, Y.-L. Mau, Y. Guo, C.-I. Li, Q. Liu, J. D. Boice, Y. Shyr, "Bivariate Poisson models with varying offsets: an application to the paired mitochondrial DNA dataset", Statistical Applications in Genetics and Molecular Biology, 16 (2017) 47-58

[5] F.Novoa-Mu˜noz, M.D. Jim´enez-Gamero, " A goodness-of-fit test for the multivariate Poisson distribution", SORT, 40(2014) 113–138

[6] B. Çekyay, J. Frenk, S. Javadi," On Computing the Multivariate Poisson Probability Distribution", Methodol Comput Appl Probab, 25(2023) 1-22

[7] M.A. Al-Osh, A.A. Alzaid, " First-order integer-valued autoregressive (INAR(1)) process", Journal of Time Series Analysis, 8 (1987) 261–75

[8] F. M. Steutel, K. Van Harn, " Discrete analogs of self-decomposability and stability", The Annals of Probability, 7 (1979) 893–899

[9] x. Pedeli, D. Karlis, " A bivariate INAR(1) process with application", Statistical Modelling,11 (2011) 325–349

[10] Z. Sajjadnia, M. Sharafi, N. Mamode Khan, A.D. Soobhug," A new bivariate INAR(1) model with paired Poisson-weighted exponential distributed innovations", Commun. Stat. Simul. Comput. ,(2023) 1–19

[11] M. Torkashvand Moradabadi, W. Soroush, Z. Torkashvand, " Mortality rate and years of life lost due to cancer in Iran from 2011 to 2018", Payesh, 20 (2021) 333-345