Austrian Journal of Statistics
2025, Volume 54, 85–100.
http://www.ajs.or.at/
doi:10.17713/ajs.v54i2.1906

Goodness-of-Fit Tests for COM-Poisson Distribution Using Stein's Characterization

Traison T
Department of Statistics
Pondicherry University
Puducherry, India

V. S. Vaidyanathan
Department of Statistics
Pondicherry University
Puducherry, India

Abstract

A crucial part of data analysis is to assess the adequacy of the fit of the probability models used, and such assessment is generally made using the goodness-of-fit tests. In this context, we propose goodness-of-fit tests for the two-parameter generalization of the Poisson distribution known as the Conway-Maxwell Poisson (COM-Poisson) distribution. The ability of COM-Poisson distribution to handle a wide range of dispersion makes it a suitable candidate for modelling discrete data. The usual goodness-of-fit tests like chi-square, Cramér–von Mises and Anderson-Darling tests, when applied to COM-Poisson distribution, are computationally complex due to the presence of the normalizing constant in the probability mass function. In this article, we overcome this complexity by representing the probability mass function of the COM-Poisson distribution using Stein's characterization and propose modified test statistics. The performance of the modified test statistics is assessed in terms of the empirical level and percentage of rejection through simulation study. Applicability of the modified tests is demonstrated through real data sets.

Keywords: Anderson-Darling test, bootstrap test, Cramér–von Mises test, Conway Maxwell Poisson distribution, Stein's characterization.

1. Introduction

Count data model involves modelling non-negative integers that correspond to the outcome of a random experiment. Some common examples of count data include web page hits, disease incidence, number of goals scored in a football match, etc. Count data can be either equi-, under- or over-dispersed. When the counts are equi-dispersed (variance equals mean), Poisson distribution is often used to model them. Under-dispersed (variance smaller than mean) and over-dispersed (variance larger than mean) count data are often modelled using weighted Poisson, Good and negative Binomial distributions. In the literature, variants of Poisson distribution are available to model non-equi-dispersed count data. See, for example, the generalized Poisson distribution (Consul and Jain (1973)) and the Poisson-Tweedie distribution (Jorgensen (1997)). However, due to restrictions in the parameter space, these distributions are less mathematically appealing. An elegant form of a two-parameter Poisson

distribution was introduced by Conway and Maxwell (1962) in the context of modelling state-dependent queues. This distribution (henceforth referred to as COM-Poisson distribution) can model equi-, under- and over-dispersed count data. The probability mass function (pmf) of a discrete random variable X having COM-Poisson distribution is given by

$$p(X = x; \lambda, \nu) = \frac{\lambda^x}{(x!)^\nu} \frac{1}{Z(\lambda, \nu)}, \quad x = 0, 1, 2, \ldots, \quad \lambda > 0, \quad \nu \geq 0, \tag{1}$$

where $Z(\lambda, \nu) = \sum_{j=0}^{\infty} \frac{\lambda^j}{(j!)^\nu}$ is the normalizing constant. Here, λ and ν, respectively, denote the location and dispersion parameters. When $\nu = 1$, the COM-Poisson distribution is equi-dispersed (Poisson distribution). The distribution is under-dispersed when $\nu > 1$ and over-dispersed when $\nu < 1$. The moments of the COM-Poisson distribution do not have simple expressions because the normalizing constant $Z(\lambda, \nu)$ in the pmf contains an infinite series. However, Shmueli, Minka, Kadane, Borle, and Boatwright (2005) have given an approximate expression for the mean and variance as $E(X) \approx \lambda^{\frac{1}{v}} - \frac{v-1}{2v}$ and $V(X) \approx \frac{\lambda^{\frac{1}{v}}}{v}$ when $\nu \leq 1$ or $\lambda > 10^\nu$. The flexibility of the distribution to model both under- and over-dispersed count data has appealed to theoretical and applied researchers. In the past two decades, many research articles on the properties, characterizations, extensions and applications of COM-Poisson distribution have appeared in the literature. Notable among them include Sellers, Borle, and Shmueli (2012), Daly and Gaunt (2016), Melo and Alencar (2022), Chanialidis, Evers, Neocleous, and Nobile (2018), Huang and Kim (2021), Benson and Friel (2021), Ong, Gupta, Ma, and Sim (2021) and Piancastelli, Friel, Barreto-Souza, and Ombao (2023).

However, the problem of testing the goodness-of-fit of the observed data for the COM Poisson distribution has not yet been addressed. Implementation of goodness-of-fit tests like Kolmogorov–Smirnov, Anderson-Darling and Cramér–von Mises that involve the cumulative distribution function will be computationally challenging due to the presence of the infinite series in the normalizing constant. The computational complexity can be overcome by employing Stein's characterization for a discrete probability distribution. The main focus of this paper is to apply Stein's characterization on COM-Poisson distribution and thereby propose modified Anderson-Darling and Cramér–von Mises test statistics for testing the goodness-of-fit of a COM-Poisson distribution. Also, a test statistic is proposed based on the distance between the empirical and estimated probabilities of the COM-Poisson distribution. Recently, Yang, Liu, Rao, and Neville (2018), Aleksandrov, Weiß, and Jentsch (2022a) and Aleksandrov, Weiß, Jentsch, and Faymonville (2022b) have used Stein's identity to test the goodness-of-fit for discrete probability distributions. However, these test procedures involve weight functions and are based on the asymptotic normality of the underlying test statistic. Here, we propose a different approach that considers the estimated probabilities to construct the test statistic.

The paper is organized as follows: In Section 2, we define the mean reparametrized COM-Poisson distribution and represent its pmf using Stein's characterization. In Section 3, the modified Anderson-Darling and Cramér–von Mises test statistics are proposed based on the pmf obtained by Stein's characterization. Also, a test statistic based on probability distance is introduced. The performance of the test statistics is assessed based on the empirical level and percentage of rejection through a detailed simulation study in Section 4. Section 5 discusses the application of the goodness-of-fit tests to real data. Section 6 contains concluding remarks.

2. Mean reparametrized COM-Poisson distribution

Reparametrizing ν and λ as e^ϕ and $\left(\mu + \frac{e^\phi - 1}{2e^\phi}\right)^{e^\phi}$ respectively in the pmf given in equation (1), Ribeiro Jr, Zeviani, Bonat, Demétrio, and Hinde (2020) introduced a mean reparametrized form of the COM-Poisson distribution. The pmf of a random variable X having a mean reparametrized COM-Poisson distribution (henceforth referred to as mCMP distribution) is

given by

$$p(X=x;\mu,\phi)=\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{xe^{\phi}}\frac{(x!)^{-e^{\phi}}}{Z(\mu,\phi)},\quad x=0,1,2,\ldots,\quad \begin{matrix}0<\mu<+\infty,\\ -\infty<\phi<+\infty,\end{matrix} \tag{2}$$

where $Z(\mu,\phi)=\sum_{j=0}^{\infty}\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{je^{\phi}}}{(j!)^{e^{\phi}}}$ is the normalizing constant. We denote X having the pmf in equation (2) as $X\sim mCMP(\mu,\phi)$. Here, μ and ϕ, respectively, denote the location and dispersion parameters. When $\phi=0$, the mCMP distribution is equi-dispersed (Poisson distribution). The distribution is under-dispersed when $\phi>0$ and over-dispersed when $\phi<0$. The mean and variance of X are, respectively, $E(X)\approx\mu$ and $V(X)\approx\mu e^{-\phi}$. The advantage of the reparametrized form of the distribution is that the parameter μ approximates the mean of the distribution.

2.1. Representation of mCMP distribution using Stein's characterization

Betsch, Ebner, and Nestmann (2022) have introduced a characterization for discrete probability distributions using the discrete density approach identity given in Ley and Swan (2013). Let X be a discrete random variable with pmf $p(x)$ that has support in the set of non-negative integers (N_0). Using Stein's characterization on $p(x)$, from Betsch *et al.* (2022), the pmf is represented as $\rho_X(k)$ and is given by

$$\rho_X(k)=E\left[-\frac{\Delta^{+}p(X)}{p(X)}I(X\geq k)\right],k\in N_0 \tag{3}$$

where $\Delta^{+}p(x)=p(x+1)-p(x)$ is the forward difference operator and $I(.)$ is the indicator function provided $p(x)$ satisfies the following conditions.

- C1 Support of $p(x)$ is $\{L,L+1,...,R\}$, where $L,R\in N_0, L<R$.
- C2 $\lim\limits_{k\to\infty}\left|\frac{\Delta^{+}p(k)(1-P(k))}{p(k)p(k+1)}\right|<\infty$, where $P(k)=\sum\limits_{l=0}^{k}p(l)$ is the cumulative distribution function corresponding to $p(x)$.
- C3 $E|X|<\infty$.

It is to be noted that $\rho_X(k)=p(k)$ has to hold if $p(x)$ is indeed the pmf of X. The primary advantage of using Stein's characterization is that the normalizing constant, if any, in the original pmf of the distribution vanishes due to the ratio $\frac{\Delta^{+}p(x)}{p(x)}$. Hence, from an inferential point of view, the pmf given in equation (3) is flexible for mathematical treatments. Since the mCMP distribution has a normalizing constant that involves an infinite sum, we employ Stein's characterization to represent its pmf using equation (3). This is done in the following theorem.

Theorem 1. *Let $X\sim mCMP(\mu,\phi)$. Using Stein's characterization, the pmf of X is represented as*

$$\rho_X(k)=E\left[\left(1-\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{X+1}\right)^{e^{\phi}}\right)I(X\geq k)\right],k\in N_0. \tag{4}$$

Proof. For the mCMP distribution with the pmf given in equation (2), condition C1 holds. Since $E(X)\approx\mu$ ($<\infty$), condition C3 also holds. To check whether the condition C2 holds,

note that

$$\frac{\Delta^{+}p\left(k\right)}{p\left(k\right)}=\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+1}\right)^{e^{\phi}}-1,\quad k\in N_{0}.$$

Therefore,

$$\begin{aligned}\left|\frac{\Delta^{+}p\left(k\right)\cdot\left(1-P\left(k\right)\right)}{p\left(k\right)p\left(k+1\right)}\right| &= \left|\frac{\Delta^{+}p\left(k\right)}{p\left(k\right)}\right|\left[\frac{\left(1-P\left(k\right)\right)}{p\left(k+1\right)}\right] \quad \left(\text{since } \left[\tfrac{\left(1-P\left(k\right)\right)}{p\left(k+1\right)}\right] \text{ is always positive}\right)\\ &= \left|\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+1}\right)^{e^{\phi}}-1\right|\sum_{i=k+1}^{\infty}\left[\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{(i-k-1)}(k+1)!}{(i)!}\right]^{e^{\phi}}\\ &= \left|\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+1}\right)^{e^{\phi}}-1\right|\sum_{l=0}^{\infty}\left[\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{l}(k+1)!}{(l+k+1)!}\right]^{e^{\phi}}\\ &= \left|\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+1}\right)^{e^{\phi}}-1\right|\left\{1+\left(\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)}{(k+2)}\right)^{e^{\phi}}+\left(\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{2}}{(k+2)(k+3)}\right)^{e^{\phi}}+...\right\}\end{aligned}$$

Since for $k\in N_0$, $\frac{1}{k+2},\frac{1}{(k+2)(k+3)},...$ is a decreasing series, we have,

$$\left\{1+\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{e^{\phi}}+\left(\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{2}}{(k+2)(k+3)}\right)^{e^{\phi}}+\dots\right\}\leq\left\{1+\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{e^{\phi}}+\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{2e^{\phi}}+\dots\right\}$$

Therefore,

$$\begin{aligned}\sum_{l=0}^{\infty}\left(\frac{\left(\mu+\frac{e^{\phi}-1}{2e^{\phi}}\right)^{l}(k+1)!}{(l+k+1)!}\right)^{e^{\phi}} &\leq \sum_{l=0}^{\infty}\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{e^{\phi}l}\\ &= \frac{1}{1-\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{e^{\phi}}}.\end{aligned}$$

Thus,

$$\left|\frac{\Delta^{+}p\left(k\right)\cdot\left(1-P\left(k\right)\right)}{p\left(k\right)p\left(k+1\right)}\right|\leq\left|\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+1}\right)^{e^{\phi}}-1\right|\left[\frac{1}{1-\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{e^{\phi}}}\right].$$

$$\Rightarrow\lim_{k\to\infty}\left|\frac{\Delta^{+}p\left(k\right)\left(1-P\left(k\right)\right)}{p\left(k\right)p\left(k+1\right)}\right|\leq\lim_{k\to\infty}\left|\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+1}\right)^{e^{\phi}}-1\right|\left[\frac{1}{1-\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{k+2}\right)^{e^{\phi}}}\right]$$

$$=1$$

provided $\mu+\frac{e^{\phi}-1}{2e^{\phi}}>0$. Hence, condition C2 is satisfied. Since the conditions C1, C2 and C3 holds good for mCMP distribution, using equation (3), the pmf of mCMP distribution is represented as

$$\rho_{X}\left(k\right)=E\left[\left(1-\left(\frac{\mu+\frac{e^{\phi}-1}{2e^{\phi}}}{X+1}\right)^{e^{\phi}}\right)I\left(X\geq k\right)\right],k\in N_{0}. \tag{5}$$

□

The condition $\mu + \frac{e^{\phi}-1}{2e^{\phi}} > 0$ is essentially a restriction in the pmf of the mCMP distribution as mentioned in Ribeiro Jr *et al.* (2020). As pointed out by the authors, the parametric space under this restriction relates to small under-dispersed counts that are unlikely of interest in practice.

Remark. *Using equation (5), an estimator of $\rho_X(k)$ given the random sample $x_1, x_2, \ldots, x_n$ of size n from mCMP distribution is defined as*

$$\hat{\rho}_X(k) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \left(\frac{\mu + \frac{e^{\phi}-1}{2e^{\phi}}}{x_i + 1}\right)^{e^{\phi}}\right) I(x_i \geq k),\ k \in N_0 \tag{6}$$

By law of large numbers, $\hat{\rho}_X(k) \xrightarrow{p} \rho_X(k)$.

3. Goodness of fit tests

Goodness-of-fit test is a statistical procedure to decide whether a given random sample of observations is generated from some specified probability distribution. In the present context, the problem of interest is to test whether the random sample is generated from a mCMP distribution. The testing problem is stated in terms of hypotheses H_0 and H_1 as

$$H_0 : \text{Sample data is from mCMP}(\mu, \phi)$$

against

$$H_1 : \text{Sample data is not from mCMP}(\mu, \phi).$$

Let $Y_x,\ x = 0, 1, 2, ..., r$ denote the frequency of x occurring in the random sample of size n, where r is the largest observation in the sample. Thus, $n = \sum_{i=0}^{r} Y_i$. Let $p_n(x) = \frac{Y_x}{n}$ denote the empirical probability and let

$$F_n(x) = \sum_{i=0}^{x} p_n(i), \quad x = 0, 1, 2, ..., r \tag{7}$$

denote the empirical distribution function (EDF). From equation (2), the cumulative distribution function (CDF) of a mCMP distribution is given by

$$F(x; \mu, \phi) = \sum_{i=0}^{x}\left(\mu + \frac{e^{\phi}-1}{2e^{\phi}}\right)^{ie^{\phi}} \frac{(i!)^{-e^{\phi}}}{Z(\mu, \phi)}, \quad x = 0, 1, 2, ... \tag{8}$$

To test the hypotheses, one may use the traditional chi-square (χ^2) test or tests based on EDF, like the Cramér–von Mises (CV) and Anderson-Darling (AD) tests. The test statistic of the above tests are respectively defined as

$$\chi^2 = \sum_{x=0}^{r} \frac{(O_x - E_x)^2}{E_x}, \tag{9}$$

$$CV = n\sum_{x=0}^{r} [F_n(x) - F(x; \mu, \phi)]^2 p(x; \mu, \phi) \tag{10}$$

and

$$AD = n\sum_{x=0}^{r} \frac{[F_n(x) - F(x; \mu, \phi)]^2 p(x; \mu, \phi)}{F(x; \mu, \phi)[1 - F(x; \mu, \phi)]}, \tag{11}$$

where O_x and E_x denote respectively the observed and expected frequency. From Henze and Klar (1995), it is clear that the chi-squared statistic inflates the probability of Type I error when the sample size and observed frequencies are small. Also, the presence of the normalizing constant in the CDF given in equation (8) makes the computation of the CV and AD test statistics difficult. In the sequel, we define the modified CV and AD test statistics using the estimated CDF of mCMP distribution obtained through Stein's characterization. In addition, we propose a test statistic based on the distance between the empirical probability and the pmf of mCMP distribution obtained through Stein's characterization.

3.1. Modified Cramér–von Mises and Anderson-Darling statistic

The proposed CD and AD test statistics are obtained by replacing the CDF of mCMP distribution with the CDF of Stein's pmf. The modified CV and AD test statistics, denoted respectively by CV_M and AD_M, are defined below.

$$CV_M = n\sum_{x=0}^{r}\left[F_n\left(x\right) - F_{\hat{\rho}_x}\left(x;\mu,\phi\right)\right]^2\hat{\rho}_X(x) \tag{12}$$

and

$$AD_M = n\sum_{x=0}^{r}\frac{\left[F_n\left(x\right) - F_{\hat{\rho}_X(x)}\left(x;\mu,\phi\right)\right]^2\hat{\rho}_X(x)}{F_{\hat{\rho}_X(x)}\left(x;\mu,\phi\right)\left[1 - F_{\hat{\rho}_X(x)}\left(x;\mu,\phi\right)\right]}, \tag{13}$$

where $F_{\hat{\rho}_x}\left(x;\mu,\phi\right)$ is CDF of mCMP distribution based on Stein's pmf in equation (6) and is given by

$$F_{\hat{\rho}_X}\left(x;\mu,\phi\right) = \sum_{j=0}^{x}\hat{\rho}_X(j). \tag{14}$$

The modified CV and AD statistics in equations (12) and (13) do not involve the normalizing constant, thereby making the computation easy.

3.2. Test statistic based on probability distance

Under H_0, one would expect the difference between the empirical probabilities and probabilities of the mCMP distribution to be small. Hence, for testing H_0 against H_1, a test statistic can be defined by taking the average of the squared distances between these probabilities. A similar test statistic has been used to test the goodness-of-fit of Poisson distribution (see Betsch *et al.* (2022)). In the present context, we use Stein's pmf for mCMP distribution given in equation (6) to define the test statistic. The test statistic (PD) based on the probability distance is defined as

$$PD = \frac{1}{n}\sum_{j=0}^{r}\left(p_n(j) - \rho_X\left(j\right)\right)^2. \tag{15}$$

In the sequel, we derive the asymptotic distribution of PD.

Let x_1, x_2, x_3,..., x_n be the observed sample data of size n from an $mCMP(\mu,\phi)$. we have $p_n(j) = \frac{Y_j}{n}$, $j = 0,1,2,3,..,r$. Under H_0, $Y_j \sim$ Binomial$(n,\rho_X(j))$, with $E(Y_j) = n\rho_X(j)$ and $V(Y_j) = n\rho_X(j)(1-\rho_X(j))$.

Since $Y_j = np_n(j)$, $E(p_n(j)) = \rho_X(j)$ and $V(p_n(j)) = \dfrac{\rho_X(j)(1-\rho_X(j))}{n}$. Thus

$$\begin{aligned}
E(PD) =& \frac{1}{n}\sum_{j=0}^{r} E(p_n(j) - \rho_X(j))^2 \\
=& \frac{1}{n}\sum_{j=0}^{r} E(p_n(j) - E(p_n(j)))^2 \\
=& \frac{1}{n}\sum_{j=0}^{r} V(p_n(j)) \\
=& \frac{1}{n^2}\sum_{j=0}^{r} \rho_X(j)(1-\rho_X(j)).
\end{aligned}$$

Also,

$$\begin{aligned}
V(PD) =& \frac{1}{n^2}\sum_{j=0}^{r} V(p_n(j) - \rho_X(j))^2 \\
=& \frac{1}{n^2}\sum_{j=0}^{r} V(p_n(j) - E(p_n(j)))^2 \\
=& \frac{1}{n^2}\sum_{j=0}^{r} \left\{ \frac{1}{n^4} E(Y_j - E(Y_j))^4 - \left(E(p_n(j) - E(p_n(j)))^2\right)^2 \right\} \\
=& \frac{1}{n^2}\sum_{j=0}^{r} \left\{ \frac{1}{n^4} E(Y_j - E(Y_j))^4 - (V(p_n(j)))^2 \right\} \\
=& \frac{1}{n^2}\sum_{j=0}^{r} \frac{1}{n^4} n\rho_X(j)(1-\rho_X(j))\left[1 + 3(n-2)\rho_X(j)(1-\rho_X(j))\right] \\
& - \frac{1}{n^2}\sum_{j=0}^{r} \left(\frac{1}{n^2}\sum_{j=0}^{r} \rho_X(j)(1-\rho_X(j)) \right)^2 \\
=& \frac{1}{n^5}\sum_{j=0}^{r} \rho_X(j)(1-\rho_X(j))\left[1 + 3(n-2)\rho_X(j)(1-\rho_X(j))\right] \\
& - \frac{r+1}{n^6} \left((\sum_{j=0}^{r} \rho_X(j)(1-\rho_X(j))) \right)^2 .
\end{aligned}$$

Thus, for large n, PD has a normal distribution with mean $E(PD)$ and variance $V(PD)$.

4. Simulation study

A simulation study is conducted to assess the performance of the CV_M, AD_M and PD goodness-of-fit test statistics and compare them with the CV, AD and χ^2 test statistics. The performance is assessed based on the empirical level and percentage of rejection. Since the exact distribution for all the test statistics under H_0 (except the chi-squared statistic) is difficult to obtain, we propose a bootstrap test procedure similar to Gürtler and Henze (2000) to compute the critical value and thereby arrive at a decision. The steps involved in the bootstrap test procedure is given below. Let Z^* denote a goodness-of-fit test statistic.

- **Step 1** Generate a random sample of size n from mCMP(μ, ϕ) and obtain the estimates $\hat{\mu}$ and $\hat{\phi}$. Compute the test statistics Z^* for testing H_0 against H_1.
- **Step 2** Generate B bootstrap samples each of size n from mCMP$(\hat{\mu}\ ,\ \hat{\phi})$. For each of the B bootstrap samples, find the estimates $\hat{\mu}_i, \hat{\phi}_i$ and compute $Z_i^*, i = 1, 2, 3, ..., B$.

- **Step 3** Arrange the Z_i^*'s in ascending order and compute the critical value c_n defined as
$$c_n = Z^*_{(a)} + (1-\alpha)(Z^*_{(a+1)} - Z^*_{(a)}),$$
where $a = [(1-\alpha)B]$, $\alpha \in (0,1)$and [.] denote the floor function. Here $Z^*_{(a+1)}$ and $Z^*_{(a)}$ denote the value of Z_i^* at the $(a+1)^{th}$ and a^{th} position in the ordered arrangement.
- **Step 4** Reject H_0 if $Z^* > c_n$.

The estimates $\hat{\mu}$ and $\hat{\phi}$ are obtained by minimizing $S(\mu,\phi) = \sum_{j=0}^{r} (p_n(j) - \hat{\rho}_X(j))^2$ with respect to μ and ϕ. Since $\hat{\rho}_X(x)$ does not involve the normalizing constant, minimizing $S(\mu,\phi)$ is computationally easy. One may also use maximum likelihood estimation to obtain the estimates. However, the complexity involved in the likelihood equations due to the presence of the normalizing constant makes the estimation process difficult (see Benson and Friel (2021) and Bedbur, Kamps, and Imm (2023)). We carry out the simulation study by fixing $\mu \in \{1,2,3,4,5\}$ and $\phi \in \{-0.5,-0.25,0,0.25,0.5\}$. These choices of ϕ include the equi-, under- and over-dispersed scenarios. For each combination of μ and ϕ, samples of size $n = 25$ and 50 are generated from mCMP distribution using the **COMPossionReg** package in R (Kimberly, Thomas, and Andrew (2019)). Based on each of these samples, 10,000 runs of the bootstrap test procedure is implemented taking $B \in \{50,100\}$ and fixing $\alpha = 0.05$. The empirical level of the test is computed as the proportion of rejecting H_0 out of 10,000 runs. In a similar manner, the percentage of rejection is computed under H_1. To accommodate the equi-, under- and over-dispersion scenarios, the following distributions, namely, binomial (B), negative binomial (NB), discrete uniform (DU), generalized Poisson (GP), a mixture of Poisson (Poi), a mixture of generalized Poisson and a mixture of mCMP, are considered under H_1.

4.1. Results and discussions

Table 1 and 3, respectively, display the empirical level of the test obtained under CV, AD, χ^2, CV_M, AD_M and PD for n=25 and 50. The percentage of rejection for various distributions considered under H_1 are displayed respectively in Table 2 and 4 for $n = 25$ and 50. It is seen from Table 1 and 3 that, by and large, all the test statistics are conservative when the samples are equi- or under-dispersed. However, for over-dispersed samples, the test statistics are mildly less conservative. From the percentage of rejection displayed in Table 2 and 4, it is seen that the modified test statistics CV_M and AD_M have a higher percentage of rejection compared to other test statistics for uniform and mixture distributions under H_1. Also, all the test statistics have a low percentage of rejection under binomial and negative binomial distribution under H_1. This is due to the fact that mCMP distribution can accommodate under- and over-dispersed data. Besides, the proposed test statistic PD based on probability distance performs comparatively better in terms of empirical level and percentage of rejection than the traditional test statistics CV, AD and χ^2. Hence, from the simulation results, it is inferred that the modified Cramér–von Mises and Anderson-Darling tests based on Stein's characterization are conservative and have a higher percentage of rejection.

Table 1: Empirical level for n= 25

	μ	CV	AD	χ^2	CV_M	AD_M	PD
ϕ =-0.5	1	0.047	0.048	0.085	0.070	0.075	0.066
	2	0.047	0.035	0.045	0.047	0.044	0.046
	3	0.037	0.034	0.040	0.044	0.044	0.053
	4	0.059	0.055	0.031	0.047	0.047	0.055
	5	0.047	0.053	0.034	0.051	0.055	0.056
ϕ =-0.25	1	0.043	0.039	0.071	0.061	0.063	0.055
	2	0.035	0.026	0.043	0.047	0.041	0.044
	3	0.045	0.048	0.033	0.050	0.049	0.052
	4	0.059	0.053	0.039	0.050	0.043	0.048
	5	0.042	0.042	0.028	0.045	0.055	0.054
ϕ =0	1	0.038	0.051	0.066	0.058	0.056	0.041
	2	0.055	0.052	0.034	0.064	0.056	0.050
	3	0.035	0.036	0.027	0.040	0.041	0.039
	4	0.055	0.049	0.029	0.048	0.051	0.039
	5	0.052	0.042	0.034	0.041	0.038	0.052
ϕ =0.25	1	0.025	0.031	0.043	0.034	0.033	0.036
	2	0.059	0.057	0.039	0.059	0.058	0.042
	3	0.045	0.028	0.037	0.045	0.044	0.045
	4	0.041	0.038	0.029	0.047	0.047	0.049
	5	0.050	0.041	0.026	0.054	0.058	0.045
ϕ =0.5	1	0.034	0.037	0.042	0.041	0.045	0.036
	2	0.072	0.055	0.037	0.067	0.062	0.055
	3	0.044	0.025	0.030	0.051	0.046	0.039
	4	0.040	0.029	0.034	0.051	0.044	0.038
	5	0.037	0.030	0.028	0.049	0.048	0.029

Table 2: Percentage of rejection for n=25

Distribution under H_1	CV	AD	χ^2	CV_M	AD_M	PD
B(5,0.25)	4.0	3.1	1.6	4.2	4.2	3.3
B(5,0.5)	6.1	6.4	7.7	5.7	5.0	6.2
B(10,0.25)	13.3	8.8	6.4	13.2	11.8	5.6
NB(5,0.25)	4.0	4.0	7.3	6.6	6.1	6.0
NB(5,0.5)	6.1	6.4	7.7	5.7	5.0	6.2
NB(10,0.75)	4.6	3.9	4.7	4.7	4.7	5.4
GP(2,0.25)	6.0	6.2	9.1	7.9	7.7	7.3
GP(10,0.1)	5.5	5.8	5.0	5.7	7.2	5.5
GP(10,0.25)	4.2	3.4	4.4	5.1	4.1	5.1
GP(10,0.5)	4.5	3.4	7.0	7.5	6.6	6.1
DU(0,4)	21.2	19.5	0.1	30.7	39.8	19.4
DU(0,8)	20.9	32.2	13.9	15.8	34.9	18.9
0.25Poi(5)+0.75Poi(20)	30.7	36.2	28.0	60.9	68.6	22.3
0.5Poi(2)+0.5Poi(10)	7.3	8.0	0.5	17.2	23.1	10.0
0.5CMP(1,0.5)+0.5CMP(5,0.5)	100.0	100.0	94.6	97.9	100.0	90.0

Table 3: Empirical level for n= 50

	μ	CV	AD	χ^2	CV_M	AD_M	PD
ϕ =-0.5	1	0.056	0.044	0.079	0.068	0.062	0.059
	2	0.045	0.034	0.055	0.044	0.050	0.043
	3	0.051	0.042	0.036	0.058	0.057	0.066
	4	0.057	0.062	0.038	0.054	0.049	0.050
	5	0.039	0.035	0.033	0.048	0.044	0.046
ϕ =-0.25	1	0.064	0.050	0.059	0.067	0.063	0.063
	2	0.051	0.036	0.046	0.047	0.051	0.049
	3	0.067	0.058	0.056	0.076	0.079	0.060
	4	0.043	0.040	0.041	0.043	0.042	0.049
	5	0.051	0.049	0.039	0.051	0.049	0.047
ϕ =0	1	0.041	0.036	0.062	0.045	0.041	0.041
	2	0.051	0.043	0.046	0.060	0.063	0.051
	3	0.044	0.043	0.045	0.051	0.051	0.049
	4	0.037	0.038	0.037	0.051	0.045	0.045
	5	0.048	0.048	0.052	0.051	0.044	0.042
ϕ =0.25	1	0.051	0.047	0.049	0.053	0.053	0.046
	2	0.062	0.059	0.038	0.060	0.057	0.065
	3	0.043	0.032	0.035	0.044	0.042	0.055
	4	0.047	0.039	0.047	0.053	0.043	0.045
	5	0.051	0.052	0.045	0.055	0.062	0.048
ϕ =0.5	1	0.055	0.058	0.066	0.058	0.057	0.055
	2	0.064	0.056	0.051	0.054	0.039	0.039
	3	0.032	0.028	0.040	0.036	0.037	0.036
	4	0.036	0.032	0.054	0.047	0.045	0.048
	5	0.043	0.042	0.050	0.047	0.046	0.049

Table 4: Percentage of rejection for n=50

Distribution under H_1	CV	AD	χ^2	CV_M	AD_M	PD
B(5,0.25)	6.0	2.2	1.2	4.2	3.8	5.0
B(5,0.5)	6.5	4.8	6.0	5.8	4.4	4.3
B(10,0.25)	5.0	5.1	9.0	7.6	6.0	5.6
NB(5,0.25)	6.1	5.2	12.2	11.4	9.3	6.4
NB(5,0.5)	7.9	6.5	12	7.9	7.7	5.6
NB(10,0.75)	4.6	4.7	7.7	6.2	6.4	6.6
GP(2,0.25)	4.6	5.6	13.8	8.7	7.6	5.3
GP(10,0.1)	5.0	5.0	5.9	6.1	6.1	3.8
GP(10,0.25)	6.1	6.2	6.7	8.8	8.4	7.7
GP(10,0.5)	5.3	4.2	13.8	15.5	11.0	6.9
DU(0,4)	47.1	51.4	4.6	59.3	72.9	42.2
DU(0,8)	34.9	45.7	7.8	48.2	74.0	34.2
0.25Poi(5)+0.75Poi(20)	57.0	57.0	30.0	84.0	96.0	55.0
0.5Poi(2)+0.5Poi(10)	14.0	16.8	0.1	36.9	51.0	22.2
0.5CMP(1,0.5)+0.5CMP(5,0.5)	100.0	100.0	100.0	100.0	100.0	100.0

5. Real data example

This section demonstrates the applicability and performance of the goodness-of-fit test statistics given in Section 3 on three real data sets. The estimation of the parameters is carried out by minimizing $S(\mu, \phi)$ given in Section 4. The p-value of the test statistics is computed using bootstrap methodology with 100 bootstrap runs. For details on obtaining p-value using Bootstrap methodology, see Davison and Hinkley (1997).

1. Home Injury data: Adelstein (1952) considered a data set on the frequency of home injuries of 122 male shunters over an 11-year period from 1937-42 in the context of fitting Poisson distribution. The data set is displayed in Table 5. Here, we investigate the goodness-of-fit of mCMP distribution to this data. The estimates of the parameters μ and ϕ of mCMP distribution are respectively obtained as $\hat{\mu} = 1.081$ and $\hat{\phi} = -0.713$, implying the data is over-dispersed. The bar plot of the expected frequencies obtained using the pmf of the mCMP distribution, its Stein form (given in equation (5)), and the observed frequencies are displayed in Figure 1. From Figure 1, it is clear that the expected frequencies under both the pmf's are similar and are close to the observed frequencies. The p-value of the test statistics are given in Table 6. It is evident from Table 6 that all the p-values exceed the 5% level of significance. Thus, it can be concluded that mCMP distribution is a good fit for the data.

Table 5: Home Injury data

Number of Injuries	0	1	2	3	4
Observed frequencies	73	36	10	2	1

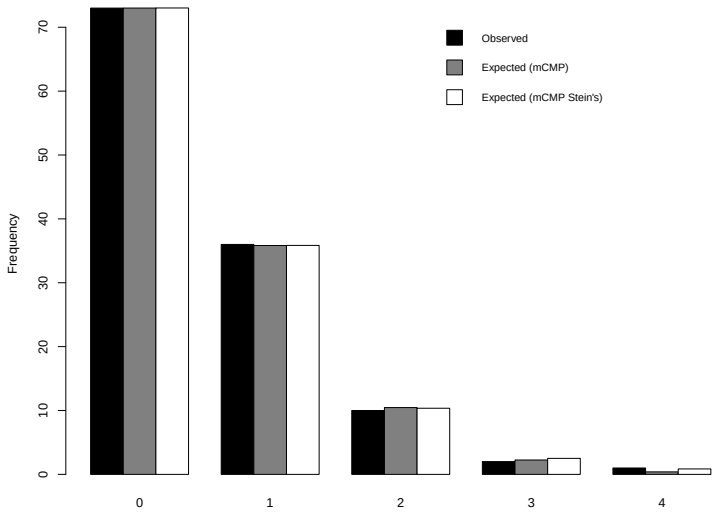

Figure 1: Bar plot of observed and expected frequencies of Home Injuries data

Table 6: p-value of the test statistics for Home Injuries data

	CV	AD	χ^2	CV_M	AD_M	PD
p-value	0.84	0.85	0.61	0.75	0.86	0.87

2. Seed data: Sarma, Rao, and Rao (1990) used the seed data set given in Table 7 to fit a family of bimodal distributions. Through the χ^2 test, they found that bimodal

distributions provide a good fit to the data. We examine the goodness-of-fit of mCMP distribution to this data. The estimates of the parameters μ and ϕ of mCMP distribution are found to be $\hat{\mu} = 6.413$ and $\hat{\phi} = 0.815$, indicating the data is under-dispersed. The bar plot of the expected frequencies obtained using the pmf of the mCMP distribution, its Stein form, and the observed frequencies are displayed in Figure 2. From Figure 2, it is evident that the data is bimodal. Also, the expected frequencies do not match with the observed frequencies. The p-value of the test statistics associated with testing the goodness-of-fit of mCMP distribution to the data is given in Table 8. It is seen from Table 8 that all the test statistics except the χ^2 have p-value less than 5%. Therefore, it is reasonable to conclude that mCMP does not fit the data. This conclusion aligns with the fact that mCMP distribution is unimodal and hence cannot fit bimodal data.

Table 7: Seed data

Length of the seed (mm)	6	7	8	9	10	11	12	13	14
Observed frequencies	1	7	22	16	7	16	23	7	1

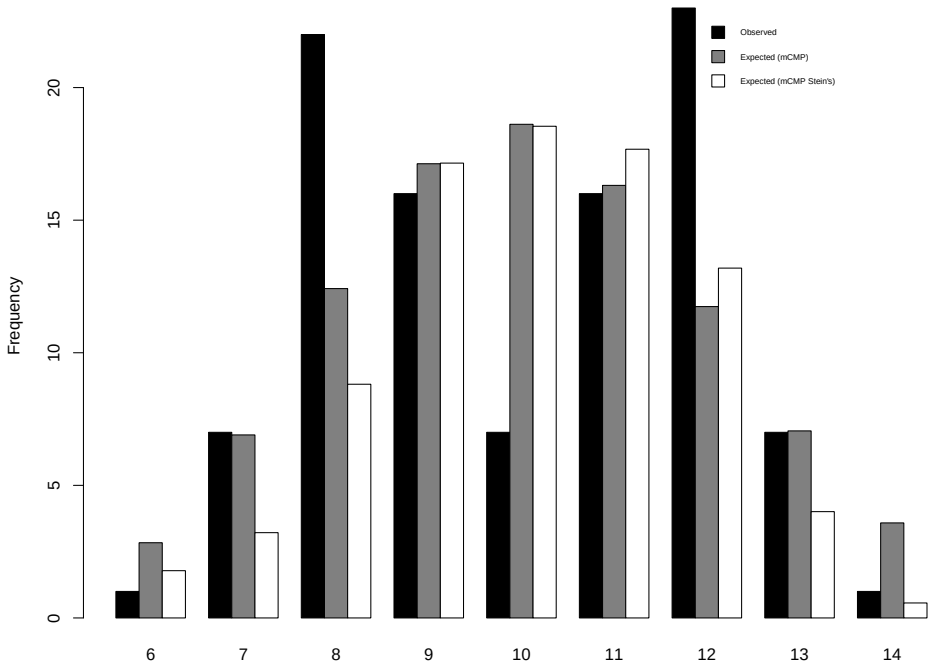

Figure 2: Bar plot of observed and expected frequencies of Seed data

Table 8: p-value of the test statistics for Seed data

p-value	CV	AD	χ^2	CV_M	AD_M	PD
	0.03	0.02	0.09	0.00	0.00	0.00

3. Coal factory strikes data: The data set displayed in Table 9 gives the number of strikes in four-week periods at a Coal factory in the United Kingdom during the years 1948-1959. Consul (1988) used the chi-square test on this data to fit a generalized Poisson distribution and found that it provides a poor fit. We inspect the goodness-of-fit of mCMP distribution to this data. The estimates of the parameters μ and ϕ of mCMP distribution are obtained as $\hat{\mu} = 1.0143$ and $\hat{\phi} = 0.708$ respectively, indicating the data

is under-dispersed. The bar plot of the expected frequencies obtained using the pmf of the mCMP distribution, its Stein form, and the observed frequencies are displayed in Figure 3. From Figure 3, it is clear that both the expected frequencies are similar and are close to the observed frequencies. The p-value of the test statistics is given in Table 10. From Table 10, it is evident that the p-value associated with all the test statistics exceeds the 5% significance level. Therefore, the mCMP distribution provides a good fit for the data.

Table 9: Coal factory strikes data

Number of strikes	0	1	2	3	4
Observed frequencies	46	76	24	9	1

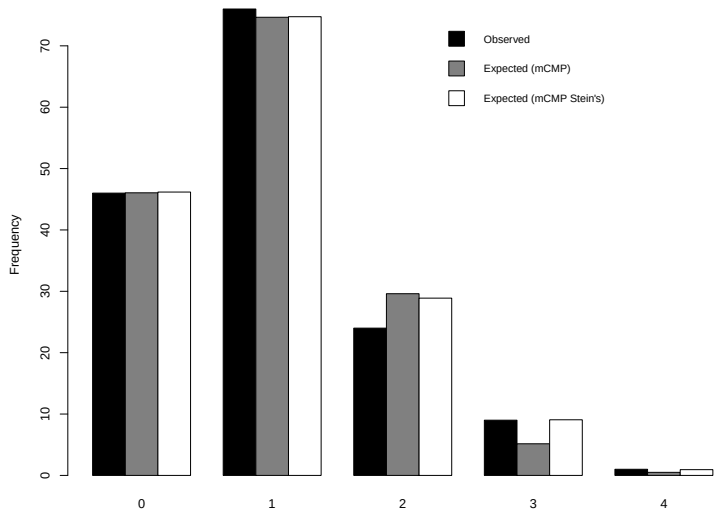

Figure 3: Bar plot of observed and expected frequencies of Coal factory strikes data

Table 10: p-value of the test statistics for Coal factory strikes data

p-value	CV	AD	χ^2	CV_M	AD_M	PD
	0.09	0.12	0.17	0.14	0.15	0.09

6. Conclusion

The test statistics proposed in this article demonstrate the applicability of goodness-of-fit tests for the COM-Poisson distribution. The test statistics are obtained using a modified form of the COM-Poisson probability mass function through Stein's characterization. The primary advantage of the probability mass function obtained through Stein's characterization is that it is free from the normalizing constant. Since COM-Poisson distribution can model equi-, under- and over-dispersed count data, and it includes Poisson, geometric and Bernoulli distributions as special cases, the proposed test statistics can test the goodness-of-fit of such distributions. This is evident from the low percentage of rejection of the tests as reported in the simulation study for these distributions under the alternate hypothesis. The results from the simulation study indicate that the chi-square test performs poorly in assessing the goodness-of-fit. The percentage of rejection of the chi-square test when the data is not from

the COM-Poisson distribution is low compared to other tests. Also, the chi-square test does not reject the hypothesis of a good fit when the data is from mixture distributions. This is clear from the real data illustration wherein the chi-square test fits the bimodal seed data set. Based on the simulation results, it is observed that the Cramér–von Mises and Anderson-Darling tests perform better than the chi-square test. However, these tests have a lesser percentage of rejection when compared to the modified Cramér–von Mises, modified Anderson-Darling and probability distance tests. For a practitioner, we recommend the modified Cramér–von Mises, modified Anderson-Darling and probability distance tests because they yield a high percentage of rejection when the data is not from COM-Poisson distribution. Also, the empirical levels of these tests are close to the chosen significance level.

References

Adelstein AM (1952). "Accident Proneness: A Criticism of the Concept Based upon an Analysis of Shunters' Accidents." *Journal of the Royal Statistical Society. Series A (General)*, **115**(3), 354–410. doi:10.2307/2980739.

Aleksandrov B, Weiß CH, Jentsch C (2022a). "Goodness-of-fit Tests for Poisson Count Time Series Based on the Stein–Chen Identity." *Statistica Neerlandica*, **76**(1), 35–64. doi:10.1111/stan.12252.

Aleksandrov B, Weiß CH, Jentsch C, Faymonville M (2022b). "Novel Goodness-of-fit Tests for Binomial Count Time Series." *Statistics*, **56**(5), 957–990. doi:10.1080/02331888.2022.2134384.

Bedbur S, Kamps U, Imm A (2023). "On the Existence of Maximum Likelihood Estimates for the Parameters of the Conway-Maxwell-Poisson Distribution." *ALEA: Latin American Journal of Probability and Mathematical Statistics*, **20**, 561–575. doi:10.30757/ALEA.v20-20.

Benson A, Friel N (2021). "Bayesian Inference, Model Selection and Likelihood Estimation Using Fast Rejection Sampling: The Conway-Maxwell-Poisson Distribution." *Bayesian Analysis*, **16**(3), 905–931. doi:10.1214/20-BA1230.

Betsch S, Ebner B, Nestmann F (2022). "Characterizations of Non-normalized Discrete Probability Distributions and Their Application in Statistics." *Electronic Journal of Statistics*, **16**(1), 1303–1329. doi:10.1214/22-EJS1983.

Chanialidis C, Evers L, Neocleous T, Nobile A (2018). "Efficient Bayesian Inference for COM-Poisson Regression Models." *Statistics and Computing*, **28**, 595–608. doi:10.1007/s11222-017-9750-x.

Consul PC (1988). *Generalized Poisson Distributions*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis.

Consul PC, Jain GC (1973). "A Generalization of the Poisson Distribution." *Technometrics*, **15**(4), 791–799. doi:10.2307/1267389.

Conway RW, Maxwell WL (1962). "A Queuing Model with State Dependent Service Rates." *Journal of Industrial Engineering*, **12**(2), 132–136.

Daly F, Gaunt RE (2016). "The Conway-Maxwell-Poisson Distribution: Distributional Theory and Approximation." *ALEA: Latin American Journal of Probability and Mathematical Statistics*, **13**, 635–658. doi:10.30757/ALEA.v13-25.

Davison AC, Hinkley DV (1997). *Bootstrap Methods and Their Application*. Cambridge University press.

Gürtler N, Henze N (2000). "Recent and Classical Goodness-of-fit Tests for the Poisson Distribution." *Journal of Statistical Planning and Inference*, **90**(2), 207–225. doi:10.1016/S0378-3758(00)00114-2.

Henze N, Klar B (1995). "Bootstrap Based Goodness of Fit Tests for the Generalized Poisson Model." *Communications in Statistics-Theory and Methods*, **24**(7), 1875–1896. doi:10.1080/03610929508831592.

Huang A, Kim ASI (2021). "Bayesian Conway–Maxwell–Poisson Regression Models for Overdispersed and Underdispersed Counts." *Communications in Statistics-Theory and Methods*, **50**(13), 3094–3105. doi:10.1080/03610926.2019.1682162.

Jorgensen B (1997). *The Theory of Dispersion Models*. CRC Press.

Kimberly S, Thomas L, Andrew R (2019). "COMPoissonReg: Conway-Maxwell Poisson (COM-Poisson) Regression." R package version 0.7.0, URL https://CRAN.R-project.org/package=COMPoissonReg.

Ley C, Swan Y (2013). "Local Pinsker Inequalities via Stein's Discrete Density Approach." *IEEE Transactions on Information Theory*, **59**(9), 5584–5591. doi:10.1109/TIT.2013.2265392.

Melo MdS, Alencar AP (2022). "Conway–Maxwell–Poisson Seasonal Autoregressive Moving Average Model." *Journal of Statistical Computation and Simulation*, **92**(2), 283–299. doi:10.1080/00949655.2021.1955887.

Ong SH, Gupta RC, Ma T, Sim SZ (2021). "Bivariate Conway–Maxwell Poisson Distributions with Given Marginals and Correlation." *Journal of Statistical Theory and Practice*, **15**, 1–19. doi:10.1007/s42519-020-00141-4.

Piancastelli LSC, Friel N, Barreto-Souza W, Ombao H (2023). "Multivariate Conway-Maxwell-Poisson Distribution: Sarmanov Method and Doubly Intractable Bayesian Inference." *Journal of Computational and Graphical Statistics*, **32**(2), 483–500. doi:10.1080/10618600.2022.2116443.

Ribeiro Jr EE, Zeviani WM, Bonat WH, Demétrio CGB, Hinde J (2020). "Reparametrization of COM–Poisson Regression Models with Applications in the Analysis of Experimental Data." *Statistical Modelling*, **20**(5), 443–466. doi:10.1177/1471082X19838651.

Sarma PVS, Rao KSS, Rao RP (1990). "On a Family of Bimodal Distributions." *Sankhyā: The Indian Journal of Statistics, Series B (1960-2002)*, **52**(3), 287–292.

Sellers KF, Borle S, Shmueli G (2012). "The COM-Poisson Model for Count Data: A Survey of Methods and Applications." *Applied Stochastic Models in Business and Industry*, **28**(2), 104–116. doi:10.1002/asmb.918.

Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P (2005). "A Useful Distribution for Fitting Discrete Data: Revival of the Conway–Maxwell–Poisson Distribution." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **54**(1), 127–142. doi:10.1111/j.1467-9876.2005.00474.x.

Yang J, Liu Q, Rao V, Neville J (2018). "Goodness-of-fit Testing for Discrete Distributions via Stein Discrepancy." In *International Conference on Machine Learning*, pp. 5561–5570. PMLR.

Affiliation:

V. S. Vaidyanathan
Department of Statistics
Pondicherry University
Puducherry, India-605014
E-mail: vaidya.stats@pondiuni.ac.in

Traison T
Department of Statistics
Pondicherry University
Puducherry, India-605014
E-mail: traison.thomas@pondiuni.ac.in

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 54
2025

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2024-03-13
Accepted: 2024-07-30