Special Proceedings: ISBN#: 978-81-950383-1-2
24th Annual Conference, 23-27 February 2022; pp 83-90

One-inflated Intervened Poisson Distribution: Stochastic Representations and Estimation

V.S.Vaidyanathan and Jahnavi Merupula
Department of Statsitics
Pondicherry University, Puducherry, India.

Received: 01 July 2022; Revised: 16 July 2022; Accepted:20 July 2022

Abstract

Count data modelling using Poisson distribution has applications in medicine, biology, physical sciences etc,. For example, the number of people affected by a strain of virus, number of gamma ray emissions etc., can be modelled using Poisson distribution. Sometimes, it is necessary to alter the rate of occurrence of counts through intervention, like administering vaccines that alters the rate of people getting affected by a virus. Such an alteration of Poisson counts results in what is generally called in the literature as an intervened Poisson distribution whose support is the set of positive integers. There are situations in which these counts can occur with more frequency than what is expected from the underlying distribution. For example, the number of visits to a physician has more frequency of 1's. This can happen either due to people visiting for a general health checkup or treatment of any ailment. Inflated count data models are often used to model count data with excess counts. Popular inflated count data models include inflated Poisson and negative binomial distributions. In this paper, an intervened Poisson distribution with one inflated count is developed. Also, two stochastic representations of the model are discussed. The moment generating function of the model is derived, and parametric estimation using the frequentist approach is carried out. A real-life application of the model is also discussed.

Key words: EM algorithm; Intervened Poisson; Maximum likelihood estimation; Moment generating function; One inflation; Zero-truncated Poisson.

AMS Subject Classifications: 60E05, 62F10.

1. Introduction

Poisson distribution is one of the oldest distributions for modelling count data. Over the years, this distribution has evolved into various forms like truncated, intervened, inflated and generalized Poisson distribution. Applications of Poisson distribution can be seen in medical, epidemiological, environmental, physical sciences etc. For a detailed discussion on various Poisson models and their applications, one may refer to Johnson et al. (2005). Inflated count models are used when a particular count frequency is more prominent than expected from the model. The excess count frequencies are attributed to having come from other generating processes. Inflated Poisson models can be used to model count data with excess counts by considering them to be generated from a degenerate distribution. Lambert (1992) introduced

Corresponding Author: V.S.Vaidyanathan
Email: vaidya.stats@gmail.com

the zero-inflated Poisson distribution for modelling the number of defects of manufacturing equipment. Following Lambert (1992), many researchers have developed various inflated Poisson distributions. Godwin and Bohning (2017) have developed a one-inflated positive Poisson model to estimate the population size of an animal species. Melkersson and Olsson (1999) have proposed a zero-one-inflated Poisson model to analyze the number of visits to a dentist.

In certain situations, the count data generating process is altered due to an intervention mechanism. It is to be noted that the intervention mechanism is activated when at least one event has occurred. For example, the intervention mechanism can be administering a drug to control the spread of disease, adjusting the specifications of a manufacturing process to reduce the number of defects etc. To accommodate the effect of the intervention on the mean of the Poisson distribution, Shanmugam (1985) introduced an intervened Poisson distribution whose probability mass function (pmf) is as given in equation (1) using a zero-truncated Poisson distribution. The intervention parameter ρ alters the mean of the Poisson distribution after the intervention mechanism.

When the intervention mechanism decreases the mean of the underlying Poisson distribution, one can expect the frequency of the smaller counts to be high. As a consequence, there might be a surge in the one counts. Also, assuming some of these 1's to be arising from a degenerate distribution outside the intervened Poisson model, the overall counts can be modelled by a one-inflated intervened Poisson distribution (OIIPD). For example, in the context of controlling the spread of the SARS-CoV-2 virus through vaccination, it is observed that people can still be infected by the virus even after vaccination. Thus, if we consider the number of individuals infected exactly once, they belong either to the vaccinated group or the unvaccinated group. Thus, the number of 1's is from two generating processes.

The rest of the paper is organized as follows. In Section 2, the pmf of the OIIPD is derived and its distributional properties are presented. Two stochastic representations (SR) of the proposed distribution are constructed in Section 3, and their equivalence is shown. In Section 4, the stochastic representations are used to derive the moment generating function and the moments of OIIPD. The estimation of parameters of the OIIPD is discussed in Section 5 through maximum likelihood (ML) estimation and EM algorithm. A Numerical illustration of the estimation procedure is presented in Section 6 using real-life data. The conclusion of the paper is given in Section 7.

2. Model Formulation and Properties

The pmf of zero-truncated Poisson distribution for positive integer-valued random variable T with mean λ is given by

$$P(T=t) = \frac{\lambda^t}{t!(e^\lambda - 1)}; \quad t = 1, 2, \ldots, \lambda > 0.$$

After some intervention mechanism, let us suppose that the mean changes from λ to $\rho\lambda$. The parameter ρ, $0 \leq \rho < \infty$ is called the intervention parameter. Let V denote a Poisson random variate with mean $\rho\lambda$. Define $X = T + V$. The pmf of X is then obtained using

convolution and is given by (Shanmugam (1985))

$$P(X=x) = \sum_{l=0}^{x-1} P(T=x-l)P(V=l|T=x-l)$$
$$= \frac{e^{-\rho\lambda}}{(e^{\lambda}-1)}\left[\frac{(\rho+1)^x-\rho^x}{x!}\right]\lambda^x; \quad x=1,2,\ldots \tag{1}$$

X defined above is said to have intervened Poisson distribution (IPD). The first two moments of X are respectively given by

$$E(X) = \lambda\left[\rho + \frac{e^{\lambda}}{(e^{\lambda}-1)}\right] \tag{2}$$

and

$$E(X^2) = \left[\frac{\lambda}{(e^{\lambda}-1)}((\rho+1)e^{\lambda}(1+\lambda(\rho+1)) - \rho(1+\rho\lambda)\right]. \tag{3}$$

To obtain the pmf of OIIPD, we proceed as follows. Let $\pi \in (0,1)$ denote the proportion of 1's obtained from outside the generating process. Thus, $(1-\pi)$ is the proportion of counts obtained from the IPD. The pmf of a random variable Y having OIIPD can thus be written as

$$P(Y=y) = \begin{cases} \pi + (1-\pi)\dfrac{e^{-\rho\lambda}\lambda}{(e^{\lambda}-1)} & , y=1 \\ (1-\pi)\dfrac{e^{-\rho\lambda}}{(e^{\lambda}-1)}\left[\dfrac{(\rho+1)^y-\rho^y}{y!}\right]\lambda^y & , y=2,3,\ldots \end{cases} \tag{4}$$

From equation (4), it is seen that OIIPD has three parameters, namely, $\lambda > 0$ that denotes the location parameter, $\rho \in [0,\infty)$ that denotes the intervention parameter and $\pi \in (0,1)$ that denotes the inflation parameter.

2.1. Distributional Properties

Using the pmf given in equation (4), the following properties of OIIPD are obtained. The moment generating function (mgf) and the probability generating function (pgf) of Y are respectively given by

$$M_Y(t) = \pi e^t + \frac{(1-\pi)}{(e^{\lambda}-1)}e^{\lambda\rho(e^t-1)}(e^{\lambda e^t}-1) \tag{5}$$

and

$$P_Y(s) = s\pi + (1-\pi)\frac{e^{s\lambda\rho(e^{\lambda s}-1)}}{(e^{\lambda}-1)e^{\rho\lambda}}.$$

Using equation (5), the mean and variance of Y are obtained, respectively as

$$E(Y) = \mu = \pi + (1-\pi)\lambda\left[\rho + \frac{e^{\lambda}}{(e^{\lambda}-1)}\right]$$

and

$$V(Y) = \mu(1-\pi)\left[1-\lambda\rho + \frac{\lambda}{(e^{\lambda}-1)}(\lambda e^{\lambda} - e^{\lambda} + \rho^2)\right].$$

The r^{th} factorial moment of OIIPD is obtained as

$$\mu_{[r]} = \pi I_r + \frac{(1-\pi)}{(e^\lambda - 1)}\lambda^r \left[(\rho+1)^r e^\lambda - \rho^r\right],$$

where $I_r = 1$ when $r = 1$ and $I_r = 0$ if $r > 1$.

3. Stochastic Representations

In this section, two SRs for the pmf given in equation (4) are presented, and their equivalence is discussed. Zhang et al. (2016) contain SRs for zero-one inflated Poisson distribution. The same methodology is adopted in the sequel.

3.1. First SR

Let Z denote a Bernoulli random variable having outcomes Z_1, Z_2. Suppose the probability of Z_1 happening is ϕ_1 and the probability of Z_2 happening is ϕ_2 i.e., $P(Z_1 = 1) = \phi_1$, $P(Z_2 = 1) = \phi_2$, $\phi_1 + \phi_2 = 1$. Let $X \sim IPD(\lambda, \rho)$ with pmf as defined in equation (1) and let $Y \sim OIIPD(\phi_1, \lambda, \rho)$. The first SR of Y is given by

$$Y = Z_1 + Z_2 X. \tag{6}$$

Note that Y takes the value one when $Z_1 = 1$ or $\{Z_2 = 1 \text{ and } X = 1\}$. Also Y takes value other than one when $\{Z_2 = 1 \text{ and } X = y\}$. Assuming X and Z are independent, the pmf of Y is obtained as

$$P(Y=y) = \begin{cases} \phi_1 + \phi_2 \dfrac{e^{-\rho\lambda}\lambda}{(e^\lambda - 1)} & , y = 1 \\ \phi_2 \dfrac{e^{-\rho\lambda}}{(e^\lambda - 1)} \left[\dfrac{(\rho+1)^y - \rho^y}{y!}\right] \lambda^y & , y = 2, 3, \ldots \end{cases} \tag{7}$$

Note that the pmf in equation (7) obtained through the first SR is the same as the pmf given in equation (4). The advantage of using the SR in equation (6) is that the moments of Y can be obtained easily as discussed in the next section.

3.2. Second SR

Let Z and η be two Bernoulli random variables such that $P(Z = 1) = 1 - \phi$ and $P(\eta = 1) = p$. Let $X \sim IPD(\lambda, \rho)$. Also Z, η and X are assumed to be independent. The second SR of Y is given by

$$Y = (1 - Z)\eta + ZX. \tag{8}$$

Note that Y takes the value one when $\{Z = 0, \eta = 1\}$ or $\{Z = 1, X = 1\}$. Also, Y takes value other than one when $\{Z = 1, X = y\}$.

$$P(Y=y) = \begin{cases} \phi p + (1-\phi) \dfrac{e^{-\rho\lambda}\lambda}{(e^\lambda - 1)} & , y = 1 \\ (1-\phi) \dfrac{e^{-\rho\lambda}}{(e^\lambda - 1)} \left[\dfrac{(\rho+1)^y - \rho^y}{y!}\right] \lambda^y & , y = 2, 3, \ldots \end{cases} \tag{9}$$

It can be observed from the right-hand side of equations (7) and (9) that,

$$\begin{cases} \phi p = \phi_1 \\ (1-\phi) = \phi_2 \end{cases} \iff \begin{cases} \phi = \phi_1 \\ p = 1. \end{cases}$$

Hence the equivalence of the two SRs.

4. Moment Generating Function based on SRs

Consider the second SR of OIIPD given in equation (8). The mgf of Y is given by

$$\begin{aligned} M_Y(t) &= E(\exp(tY)) \\ &= E\left\{\exp[t(1-Z)\eta + tZX]\right\} \\ &= E_Z\left[E_Y\left\{\exp[t(1-Z)\eta + tZX]|Z\right\}\right] \\ &= E_Z\left[E_Y\left(e^{t(1-Z)\eta}e^{tZX}|Z\right)\right] \\ &= E_Z\left[M_\eta(t(1-Z))M_X(tZ)\right] \\ &= E_Z\left[\{(1-p)+pe^{t(1-Z)}\}\frac{e^{\rho\lambda(e^{tZ}-1)}(e^{\lambda e^{tZ}}-1)}{(e^\lambda-1)}\right] \\ &= \phi[(1-p)+pe^t] + (1-\phi)\frac{e^{\rho\lambda(e^t-1)}(e^{\lambda e^t}-1)}{(e^\lambda-1)}. \end{aligned} \tag{10}$$

Using the equivalence of the two SRs, substituting $p = 1$ and taking $\phi = \phi_1$ in equation (10), the mgf of Y based on the first SR can be obtained as below.

$$M_Y(t) = \phi_1 e^t + \phi_2 \frac{e^{\rho\lambda(e^t-1)}(e^{\lambda e^t}-1)}{(e^\lambda-1)}.$$

From equation (6), using the binomial expansion, we get

$$E(Y^r) = \phi_1 + \phi_2 E(X^r), r = 1, 2, \ldots \tag{11}$$

Using the equations (2), (3) and (11), the first two moments of Y are respectively obtained as

$$E(Y) = \phi_1 + \phi_2\lambda\left[\rho + \frac{e^\lambda}{(e^\lambda-1)}\right]$$

and

$$E(Y^2) = \phi_1 + \phi_2\left[\frac{\lambda}{(e^\lambda-1)}((\rho+1)e^\lambda(1+\lambda(\rho+1)) - \rho(1+\rho\lambda)\right].$$

Thus,

$$V(Y) = E(Y)\phi_2\left[1 - \lambda\rho + \frac{\lambda}{(e^\lambda-1)}(\lambda e^\lambda - e^\lambda + \rho^2)\right].$$

5. Parametric Estimation

5.1. Method of Maximum Likelihood

Let $\vec{y} = (y_1, y_2, \dots y_n)$ be a sample of n iid observations from $OIIPD(\pi, \lambda, \rho)$. Let m denote the number of 1's in the sample and $(n-m)$ denote the number of observations taking values other than one. The likelihood function of (π, λ, ρ) corresponding to the pmf given in equation (4) is

$$L(\pi, \lambda, \rho|\vec{y}) = \left[\pi + (1-\pi)\frac{\lambda e^{-\rho\lambda}}{(e^{\lambda}-1)}\right]^{m} \times (1-\pi)^{n-m} \prod_{i=1}^{n-m} \frac{e^{-\rho\lambda}}{(e^{\lambda}-1)} \left[\frac{(\rho+1)^{y_i} - \rho^{y_i}}{y_i!}\right] \lambda^{y_i}.$$

The corresponding log-likelihood function is

$$\begin{aligned} l(\pi, \lambda, \rho|\vec{y}) = m \ln\left[\pi + (1-\pi)\frac{\lambda e^{-\rho\lambda}}{(e^{\lambda}-1)}\right] + (n-m)\left[\ln(1-\pi) - \rho\lambda - \ln(e^{\lambda}-1)\right] \\ + \ln(\lambda)\sum_{i=1}^{n-m} y_i + \sum_{i=1}^{n-m} \ln\left((\rho+1)^{y_i} - \rho^{y_i}\right) - \sum_{i=1}^{n-m} \ln(y_i!). \end{aligned}$$

The score functions of the parameters (π, λ, ρ) are respectively obtained as below.

$$\frac{\partial l}{\partial \pi} = \frac{m\left(1 - \frac{\lambda \mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)}\right)}{\pi + \frac{\lambda \mathrm{e}^{-\rho\lambda}(1-\pi)}{(\mathrm{e}^{\lambda}-1)}} - \frac{n-m}{1-\pi}, \tag{12}$$

$$\frac{\partial l}{\partial \lambda} = \frac{m\left(-\frac{(1-\pi)\lambda \mathrm{e}^{\lambda(1-\rho)}}{\left(\mathrm{e}^{\lambda}-1\right)^2} - \frac{(1-\pi)\rho\lambda \mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)} + \frac{(1-\pi)\mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)}\right)}{\pi + \frac{(1-\pi)\lambda \mathrm{e}^{-\rho\lambda}}{(\mathrm{e}^{\lambda}-1)}} + (n-m)\left(-\frac{\mathrm{e}^{\lambda}}{(\mathrm{e}^{\lambda}-1)} - \rho\right) + \frac{1}{\lambda}\sum_{i=1}^{n-m} y_i, \tag{13}$$

$$\frac{\partial l}{\partial \rho} = \frac{m(1-\pi)\lambda^2 \mathrm{e}^{-\lambda\rho}}{(\mathrm{e}^{\lambda}-1)\left(\pi + \frac{(1-\pi)\lambda \mathrm{e}^{-\lambda\rho}}{(\mathrm{e}^{\lambda}-1)}\right)} + \sum_{i=1}^{n-m}\left[\frac{y_i(\rho+1)^{y_i-1} - y_i\rho^{y_i-1}}{(\rho+1)^{y_i} - \rho^{y_i}}\right] - (n-m)\lambda. \tag{14}$$

Equating the score functions in equations (12), (13) and (14) to zero and solving them simultaneously, the ML estimates of the parameters (π, λ, ρ) are obtained provided the Hessian matrix evaluated at the ML estimates is negative definite. Since the score functions are nonlinear in the parameters, one has to use numerical methods to obtain the ML estimates. To ease out the computation, in the sequel, the parameters are estimated using the EM algorithm by treating the 1's coming from the degenerate distribution as latent.

5.2. ML Estimation via EM Algorithm

Let us assume the 1's from OIIPD are from two distributions, namely, the degenerate distribution and the IPD. Let U be the latent variable that denotes the number of 1's from the degenerate distribution. Suppose there are a total of m 1's observed, then $(m-U)$ 1's are from the IPD. Thus, the distribution of U given Y is Binomial(m, p) , where

$$p = \frac{\pi}{\pi + (1-\pi)\frac{\lambda e^{-\rho\lambda}}{(e^{\lambda}-1)}}.$$

The likelihood based on the complete sample $Y_{comp} = (Y, U)$ is proportional to

$$L(\pi, \lambda, \rho | Y_{comp}) \propto \pi^u \left[(1-\pi)\frac{\lambda e^{-\rho\lambda}}{(e^\lambda - 1)}\right]^{m-u} \times (1-\pi)^{n-m} \left(\frac{e^{-\rho\lambda}}{(e^\lambda - 1)}\right)^{n-m} \lambda^N \prod_{i=1}^{n-m} \left[(\rho+1)^{y_i} - \rho^{y_i}\right],$$

taking $N = \sum_{i=1}^{n-m} y_i$. The corresponding log-likelihood function is thus proportional to

$$\begin{aligned} l(\pi, \lambda, \rho | Y_{comp}) \propto\ & u \ln(\pi) + (n-u)\ln(1-\pi) + (n-u)(-\rho\lambda) - (n-u)\ln(e^\lambda - 1) \\ & + (N + m - u)\ln(\lambda) + \sum_{i=1}^{n-m} \ln([(\rho+1)^{y_i} - \rho^{y_i}]). \end{aligned} \tag{15}$$

In the E-step of the EM algorithm, the latent U is estimated as

$$\hat{u} = \frac{m\pi}{\pi + (1-\pi)\dfrac{\lambda e^{-\rho\lambda}}{(e^\lambda - 1)}}. \tag{16}$$

The ML estimates of the parameters, namely, $\hat{\pi}, \hat{\lambda}$ and $\hat{\rho}$ are obtained using the complete log-likelihood given in equation (15) through the M-step of the EM algorithm by solving the following simultaneous equations.

$$\hat{\pi} = \frac{\hat{u}}{n}, \tag{17}$$

$$\hat{\lambda} = \frac{1}{(n - \hat{u})} \sum_{i=1}^{n-m} \frac{[y_i(\hat{\rho}+1)^{y_i - 1} - y_i \hat{\rho}^{y_i - 1}]}{[(\hat{\rho}+1)^{y_i} - \hat{\rho}^{y_i}]}, \tag{18}$$

and

$$\hat{\rho} = \frac{(N - m - \hat{u})}{(n - \hat{u})\hat{\lambda}} - \frac{e^{\hat{\lambda}}}{e^{\hat{\lambda}} - 1}. \tag{19}$$

The E and the M steps in the equations (16) to (19) are repeated till the estimates converge. To start the iterative procedure, initial values of the parameters, say $\pi^{(0)}, \lambda^{(0)}$ and $\rho^{(0)}$ need to be specified. The advantage of using the EM algorithm is that the estimators have closed-form expressions, unlike the ML method, making the computations easier.

6. Numerical Illustration

The application of the proposed OIIPD to a real-life dataset is illustrated in this section. We consider the data on an epidemic of cholera in a village in India used in Shanmugam (1985) to fit a intervened Poisson distribution. The data relate to the spread of cholera in an Indian village and was earlier reported in McKendrick (1926). The data was observed when preventive treatment to contain the spread of cholera had been initiated. The data excluding the households not affected by cholera is tabulated below.

x	1	2	3	4+	Total
f_x	32	16	6	1	55

Here, x denotes the number of cholera cases, and f_x denotes the number of households with x cases. The primary reason for many households having cholera cases was attributed to one particular infected well which was used by a large section of the people in the village. However, other wells near its vicinity can also be the source of infection. Since the frequency of the number of households having one cholera case is large, OIIPD model is used to fit the data. The EM-algorithm steps given in the previous section are implemented to estimate the parameters (π, λ, ρ) by fixing their initial values as $(0.01, 0.5, 0.5)$ respectively. The initial values of the parameters λ and ρ were fixed near to their moment estimates obtained thorough the intervened Poisson model. The difference between the proportion of the observed and the expected 1's (rounded to two decimals) based on intervened Poisson model is taken as the initial value of π. The final estimates of the parameters (π, λ, ρ) are obtained as $(0.0099, 0.7050, 0.2492)$. From the estimate of π, it is clear that the proportion of 1's emerging from outside the IPD is small. This means that the primary source of the spread of cholera among the people in the village is the particular infected well. Also, a small value of the estimate of the intervention parameter ρ suggests that the preventive mechanism had a considerable effect in bringing down the number of cholera cases per household.

7. Concluding Remarks

The OIIPD introduced in this paper not only accounts for the excess 1's but also provides information on the effectiveness of the intervention mechanism. The two equivalent stochastic representations of the model given in this work provide an efficient way to derive the moment generating function and moments of OIIPD. EM algorithm approach is used to estimate the model parameters, circumventing the need to solve simultaneously the nonlinear equations given by the ML method. The proposed distribution can be used to model count data process altered by an intervention mechanism resulting in 1's with high frequency.

References

Godwin, R. T. and Böhning, D. (2017). Estimation of the population size by using the one-inflated positive Poisson model. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **66(2)**, 425-448.

Johnson, N. L., Kotz, S. and Kemp, A. W. (2005). *Univariate Discrete Distributions*. John Wiley & Sons.

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, **34(1)**, 1-14.

Melkersson, M. and Olsson, C. (1999). Is visiting the dentist a good habit?: Analyzing count data with excess zeros and excess ones. *Umeå Economic Studies*, **492**, 1–18.

McKendrick, A.G. (1926). Application of Mathematics to medical problems, *Proceedings of the Edinburgh Mathematical Society*, **44**, 98-130.

Shanmugam, R. (1985). An intervened Poisson distribution and its medical application. *Biometrics*, 1025-1029.

Zhang, C., Tian, G. L. and Ng, K. W. (2016). Properties of the zero-and-one inflated Poisson distribution and likelihood-based inference methods. *Statistics and its Interface*, **9(1)**, 11-32.