

An Alternative Distribution for Modelling Overdispersion Count Data: Poisson Shanker Distribution

A Meytrianti^{1*}, S Nurrohmah², M Novita³.

1,2,3 Department of Mathematics, Universitas Indonesia, Depok, 16424, Indonesia.

*Corresponding author: snurrohmah@sci.ui.ac.id

Abstract

Poisson distribution is a common distribution for modelling count data with assumption mean and variance has the same value (equidispersion). In fact, most of the count data have mean that is smaller than variance (overdispersion) and Poisson distribution cannot be used for modelling this kind of data. Thus, several alternative distributions have been introduced to solve this problem. One of them is Shanker distribution that only has one parameter. Since Shanker distribution is continuous distribution, it cannot be used for modelling count data. Therefore, a new distribution is offered that is Poisson-Shanker distribution. Poisson-Shanker distribution is obtained by mixing Poisson and Shanker distribution, with Shanker distribution as the mixing distribution. The result is a mixture distribution that has one parameter and can be used for modelling overdispersion count data. In this paper, we obtain that Poisson-Shanker distribution has several properties are unimodal, overdispersion, increasing hazard rate, and right skew. The first four raw moments and central moments have been obtained. Maximum likelihood is a method that is used to estimate the parameter, and the solution can be done using numerical iterations. A real data set is used to illustrate the proposed distribution. The characteristics of the Poisson-Shanker distribution parameter is also obtained by numerical simulation with several variations in parameter values and sample size. The result is MSE and bias of the

estimated parameter θ will increase when the parameter value rises for a value of n and will decrease when the value of n rises for a parameter value.

Keywords: maximum likelihood estimation, mixing, overdispersion, Poisson distribution, Shanker distribution

Introduction

Count data is the type of data in which the observations can take only the non-negative integers values. Count data are generally defined as numbers of events per interval. Count data are used to describe many phenomena such as the insurance claim numbers, number of yeast cells, number of chromosomes, etc (Panjer, 2006). To model the distribution of count data can be used counting distribution. Counting distributions are discrete distributions with probabilities are defined only at the points $0,1,2,3, \dots$ (Klugman et al., 2012)0. Poisson distribution is a common counting distribution for modelling count data with assumption mean and variance has the same value (equidispersion).

In fact, most of the count data have mean that is smaller than variance (overdispersion) and Poisson distribution cannot be used for modelling this kind of data. Thus, several alternative distributions have been introduced to solve this problem. One of them is Shanker distribution that only has one parameter, that has a distribution pattern similar to Poisson distribution. Based on Shanker (2015), Shanker distribution can be used in overcoming the assumptions of equidispersion, overdispersion, and underdispersion. Therefore, Shanker distribution can be used as an alternative to the Poisson distribution in overcoming the assumption of overdispersion and underdispersion in the data.

Since Shanker distribution is continuous distribution, the use of the Shanker distribution as an alternative to the Poisson distribution in modelling count data is less suitable. The other

way of handling overdispersion from count data is by mixed Poisson distribution. In Klugman et al. (2012), mixed Poisson distribution is a class of a mixture distribution, which is assumed by $X|\Lambda \sim \text{Poisson}(\Lambda)$ where the parameter Λ is a random variable with the certain distribution. It has been found that the general characteristics of the mixed Poisson distribution follow some characteristics of its mixing distribution. Depending on the choice of the mixing distribution, various mixed Poisson distributions have been constructed.

Furthermore, Shanker (2016) proposes a new distribution that is Poisson-Shanker distribution, an alternative distribution for overdispersion count data that only has one parameter. Poisson-Shanker distribution is obtained by mixing Poisson and Shanker distribution, with Shanker distribution as the mixing distribution. Shanker (2016) has discussed Poisson-Shanker distribution properties including its moments, skewness, kurtosis, hazard rate function, moment generating function, and estimation of the parameter using Maximum Likelihood Estimation and method of moments. Then, Shanker (2017) has discussed the applications of Poisson-Shanker distribution to model count data from biological sciences.

In this paper, we discuss a brief explanation about Poisson-Shanker distribution and its characteristics that have been discussed in Shanker (2016). Then, a simulation study is conducted to study the performance of the estimators. Furthermore, Poisson-Shanker distribution parameters will be estimated from the real case example using the Maximum Likelihood Estimation (MLE) method.

Materials

Shanker Distribution

Shanker distribution was first introduced by Shanker (2015). Shanker distribution is a two-component mixture of an exponential (θ) and gamma distribution ($2, \theta$) with mixing

propotion $\frac{\theta^2}{\theta^2+1}$ and $\frac{1}{\theta^2+1}$ respectively. The probability density function (pdf) and cumulative distribution of Shanker distribution given by:

Shanker (2015) have detailed study on modelling lifetime data using one parameter and

$$f(x) = \begin{cases} \frac{\theta^2}{1 + \theta^2} (x + \theta) e^{-\theta x}, & x > 0, \theta > 0 \\ 0, & \text{otherwise} \end{cases}$$

$$F(x) = \begin{cases} 1 - \frac{\theta x + \theta^2 + 1}{\theta^2 + 1} e^{-\theta x}, & x > 0, \theta > 0 \\ 0, & \text{otherwise} \end{cases}$$

gives better fitting model than Exponential and Lindley distribution.

Maximum Likelihood Estimation (MLE)

Let X_1, X_2, \dots, X_n be random sample size n from a certain distribution with pdf of X is $f(x; \theta)$ that depends on $\theta \in \Omega$, where Ω is a space of parameters. Then likelihood function can be obtained as joint pdf of X_1, X_2, \dots, X_n , denoted by $L(\theta; x_1, x_2, \dots, x_n)$ or $L(\theta)$, as follows:

$$L(\theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) = \prod_{i=1}^n f(x_i; \theta); \theta \in \Omega$$

Let $u(X_1, X_2, \dots, X_n)$ be a function from X_1, X_2, \dots, X_n , x_1, x_2, \dots, x_n are the observed experimental values of X_1, X_2, \dots, X_n so that if θ replaced by $u(x_1, x_2, \dots, x_n)$, the likelihood function $L(\theta; x_1, x_2, \dots, x_n)$ reach maximum value. So, $u(x_1, x_2, \dots, x_n)$ be the maximum likelihood estimator (MLE) for θ which denoted by $\hat{\theta}$. To find maximum likelihood estimator, $L(\theta)$ can be modified to log-likelihood function (Hogg et al., 2018).

Methods

Mixing

In Klugman et al. (2012), mixing is a method to establish new mixture distribution where one of the parameters of the main distribution is a random variable that has another distribution, where another distribution acts as a mixing distribution.

Let X is a random variable that has a certain distribution which is the main distribution with pdf $f_{X|\Lambda}(x|\lambda)$, where random variable X depends on the parameter Λ . Let λ is the value of a random variable with a pdf $f_{\Lambda}(\lambda)$. Then, the unconditional pdf of X is given by:

$$f_X(x) = \begin{cases} \sum_{\lambda} f_{X|\Lambda}(x|\lambda) f_{\Lambda}(\lambda), & \text{if } \Lambda \text{ is a discrete random variable} \\ \int_{\lambda} f_{X|\Lambda}(x|\lambda) f_{\Lambda}(\lambda), & \text{if } \Lambda \text{ is a continuous random variable} \end{cases} \quad (3)$$

where the sums and integral is taken over all values of λ with positive probability. The resulting distribution is a mixture distribution.

Poisson-Shanker Distribution (PSD)

Let $X|\Lambda$ be a random variable following a Poisson distribution with parameter Λ , denoted by $X|\Lambda \sim \text{Poisson}(\Lambda)$ and probability density function (pdf) is

$$f_{X|\Lambda}(x|\lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}, & x = 0, 1, 2, 3, \dots \\ 0, & \text{elsewhere.} \end{cases}$$

If λ is the value of a random variable Λ following a Shanker distribution with parameter θ , denoted by $\Lambda \sim \text{Shanker}(\theta)$ with pdf of Λ is

$$f_{\Lambda}(\lambda) = \frac{\theta^2}{\theta^2 + 1} (\theta + \lambda) e^{-\theta\lambda}; \quad \lambda > 0, \theta > 0$$

Then, the unconditional pdf of X

$$\begin{aligned} p_x &= \int_0^{\infty} f_{X|\Lambda}(x|\lambda) f_{\Lambda}(\lambda) d\lambda \\ &= \int_0^{\infty} \frac{e^{-\lambda} \lambda^x}{x!} \cdot \frac{\theta^2}{\theta^2 + 1} (\theta + \lambda) e^{-\theta\lambda} d\lambda = \dots \\ p_x &= \Pr(X = x) = \frac{\theta^2}{\theta^2 + 1} \cdot \frac{x + \theta^2 + \theta + 1}{(\theta + 1)^{x+2}}; \quad x = 0, 1, 2, \dots; \theta > 0 \end{aligned} \quad (1)$$

is pmf of random variable X following a Poisson-Shanker distribution with parameter θ , denoted by $X \sim \text{PSD}(\theta)$. Distribution function (cdf) and survival function of Poisson-Shanker distribution is

$$F(x) = 1 - \frac{\theta^3 + \theta^2 + (2+x)\theta + 1}{(\theta^2 + 1)(\theta + 1)^{x+2}} e^{-\theta x}; \quad n \leq x < n + 1 \text{ where } n = 0, 1, 2, \dots, \theta$$

$$> 0$$

$$S(x) = \frac{\theta^3 + \theta^2 + (2+x)\theta + 1}{(\theta^2 + 1)(\theta + 1)^{x+2}} e^{-\theta x}; \quad n \leq x < n + 1 \text{ where } n = 0, 1, 2, \dots, \theta$$

$$> 0$$

The graphs of the pmf, cdf and survival function of Poisson-Shanker distribution for different values of the parameter are shown in Figure 1, Figure 2, and Figure 3.

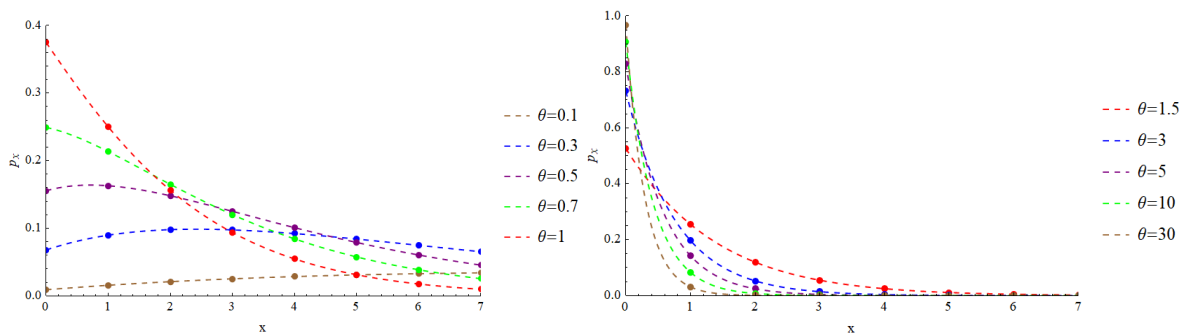


Figure 1. Graphs of pmf of the PSD for different values of the parameter θ

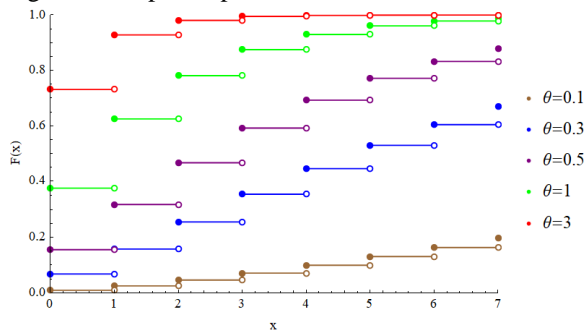


Figure 2. Graphs of cdf of the PSD for different values of the parameter θ

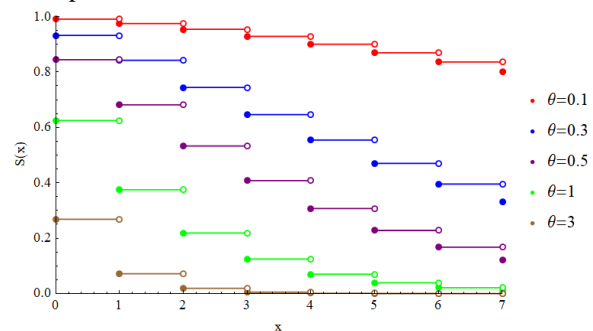


Figure 3. Graphs of the survival function of the PSD for different values of the parameter θ

Hazard Rate of Poisson-Shanker Distribution

The hazard rate of Poisson-Shanker distribution is

$$h(x) = \frac{\theta^2(\theta^2 + \theta + x + 1)}{(\theta + 1)(1 + \theta(\theta^2 + \theta + x + 1))}; \quad x = 0, 1, 2, \dots; \quad \theta > 0$$

The graphs of the hazard rate of Poisson-Shanker distribution for different values of the parameter are shown in Figure 4.

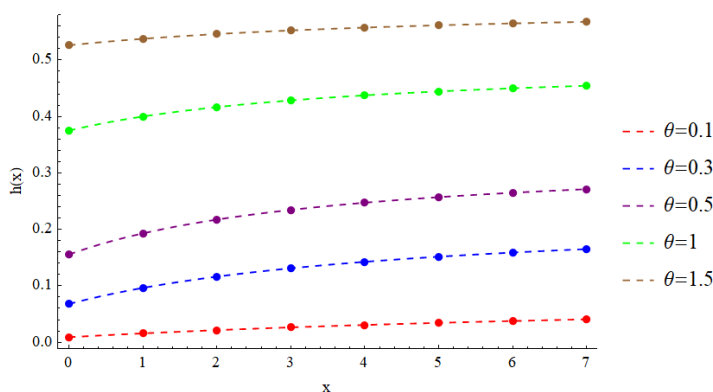


Figure 4. Graphs of hazard rate of Poisson-Shanker distribution for different values of the parameter θ

Based on Figure 4 above, the Poisson-Shanker distribution has an increasing hazard rate.

Moment Generating Function (mgf) of Poisson-Shanker Distribution

The moment generating function (mgf) of Poisson-Shanker distribution is

$$M_X(t) = \frac{\theta^2}{(\theta^2 + 1)(\theta + 1)} \left[\frac{e^t}{(\theta + 1 - e^t)^2} + \frac{\theta^2 + \theta + 1}{\theta + 1 - e^t} \right]; t \in \mathbb{R}; \theta > 0 \quad (5)$$

Moments of Poisson-Shanker Distribution

The r th factorial moment of Poisson-Shanker distribution is:

$$\begin{aligned} \mu'_{[r]} &= E[X(X-1)(X-2) \dots (X-r+1)] = \frac{r! (\theta^2 + r + 1)}{\theta^r (\theta^2 + 1)}; \theta > 0, r \\ &= 1, 2, 3, \dots \end{aligned} \quad (6)$$

In particular, we have the first four factorial moments of the Poisson-Shanker distribution:

$$\mu'_{[1]} = \frac{\theta^2 + 2}{\theta(\theta^2 + 1)}; \mu'_{[2]} = \frac{2\theta^2 + 6}{\theta^2(\theta^2 + 1)}; \mu'_{[3]} = \frac{6\theta^2 + 24}{\theta^3(\theta^2 + 1)}; \mu'_{[4]} = \frac{24\theta^2 + 120}{\theta^4(\theta^2 + 1)}$$

Using the relationship between factorial moments and moments about the origin, then

the four moments about the origin of the Poisson-Shanker distribution are:

$$\begin{aligned} \mu'_1 &= E[X] = \frac{\theta^2 + 2}{\theta(\theta^2 + 1)} & \mu'_3 &= E[X^3] = \frac{\theta^4 + 6\theta^3 + 8\theta^2 + 18\theta + 24}{\theta^3(\theta^2 + 1)} \\ \mu'_2 &= E[X^2] & \mu'_4 &= E[X^4] \\ &= \frac{\theta^3 + 2\theta^2 + 2\theta + 6}{\theta^2(\theta^2 + 1)} & &= \frac{\theta^5 + 14\theta^4 + 38\theta^3 + 66\theta^2 + 144\theta + 120}{\theta^4(\theta^2 + 1)} \end{aligned}$$

The second, third, and fourth central moment of the Poisson-Shanker distribution are:

$$\mu_2 = \frac{\theta^5 + \theta^4 + 3\theta^3 + 4\theta^2 + 2\theta + 2}{\theta^2(\theta^2 + 1)^2}$$

$$\mu_3 = \frac{\theta^8 + 3\theta^7 + 6\theta^6 + 15\theta^5 + 17\theta^4 + 18\theta^3 + 14\theta^2 + 6\theta + 4}{\theta^3(\theta^2 + 1)^3}$$

$$\mu_4 = \frac{\theta^{11} + 10\theta^{10} + 23\theta^9 + 69\theta^8 + 135\theta^7 + 188\theta^6 + 247\theta^5 + 224\theta^4 + 182\theta^3 + 122\theta^2 + 48\theta + 24}{\theta^4(\theta^2 + 1)^4}$$

Overdispersion

In particular, the mean and variance of Poisson-Shanker distribution is

$$E[X] = \mu = \frac{\theta^2 + 2}{\theta(\theta^2 + 1)} = \frac{\theta^5 + 3\theta^3 + 2\theta}{\theta^2(\theta^2 + 1)^2}$$

$$\text{Var}(X) = \sigma^2 = \frac{\theta^5 + \theta^4 + 3\theta^3 + 4\theta^2 + 2\theta + 2}{\theta^2(\theta^2 + 1)^2}$$

From the mean and variance can be written by

$$E[X] \leq \text{Var}(X)$$

$$\Leftrightarrow \frac{\theta^5 + 3\theta^3 + 2\theta}{\theta^2(\theta^2 + 1)^2} \leq \frac{\theta^5 + \theta^4 + 3\theta^3 + 4\theta^2 + 2\theta + 2}{\theta^2(\theta^2 + 1)^2}$$

Thus, the Poisson-Shanker distribution has a mean value smaller than the variance for each parameter $\theta > 0$, or can be said to be overdispersion.

Skewness and Kurtosis

The skewness (γ_1) and kurtosis (γ_2) of Poisson-Shanker distribution is

$$\gamma_1 = \frac{\theta^8 + 3\theta^7 + 6\theta^6 + 15\theta^5 + 17\theta^4 + 18\theta^3 + 14\theta^2 + 6\theta + 4}{(\theta^5 + \theta^4 + 3\theta^3 + 4\theta^2 + 2\theta + 2)^{\frac{3}{2}}} \quad (7)$$

$$\gamma_2 = \frac{\theta^{11} + 10\theta^{10} + 23\theta^9 + 69\theta^8 + 135\theta^7 + 188\theta^6 + 247\theta^5 + 224\theta^4 + 182\theta^3 + 122\theta^2 + 48\theta + 24}{(\theta^5 + \theta^4 + 3\theta^3 + 4\theta^2 + 2\theta + 2)^2} \quad (8)$$

The graphs of the skewness and kurtosis of Poisson-Shanker distribution for different values of the parameter are shown in Figure 5 and Figure 6.

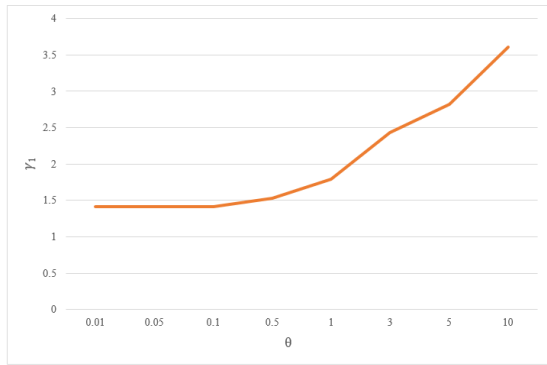


Figure 5. Graphs of skewness of the PSD for different values of the parameter θ

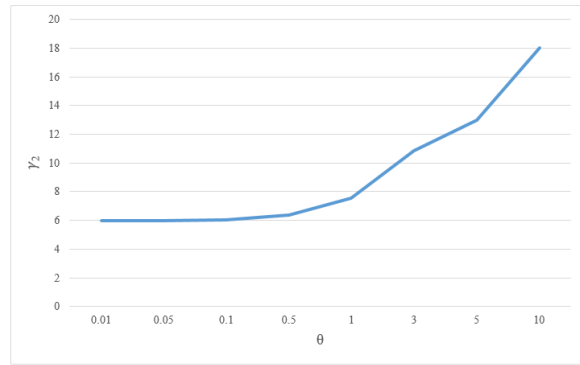


Figure 6. Graphs of kurtosis of the PSD for different values of the parameter θ

Based on Figure 5, the skewness of Poisson-Shanker distribution has a value greater than zero for each parameter θ , so it can be said that the distribution of Poisson-Shanker distribution is right-skew. Furthermore, based on Figure 6, the kurtosis of Poisson-Shanker distribution has a value greater than three for each parameter θ , so it can be said that the distribution of Poisson-Shanker distribution is leptokurtic.

Parameter Estimation

Let X_1, X_2, \dots, X_n be a random sample of size n from the Poisson-Shanker (θ) distribution with its pdf

$$f(x_i; \theta) = \frac{\theta^2}{\theta^2 + 1} \cdot \frac{x + \theta^2 + \theta + 1}{(\theta + 1)^{x+2}} ; \quad x_i = 0, 1, 2, \dots ; \quad \theta > 0$$

The likelihood function of the Poisson-Shanker distribution is given by

$$\begin{aligned} L(\theta; x) &= f(x_1, x_2, \dots, x_n; \theta) \\ &= f(x_1; \theta)f(x_2; \theta) \dots f(x_n; \theta) \\ &\quad \vdots \\ L(\theta; x) &= \frac{\theta^{2n}}{(\theta^2 + 1)^n} \cdot \frac{\prod_{i=1}^n (x_i + \theta^2 + \theta + 1)}{(\theta + 1)^{\sum_{i=1}^n x_i + 2n}} \end{aligned}$$

The log-likelihood function is thus obtained as

$$\begin{aligned} \log L(\theta; x) &= 2n \log \theta - n \log(\theta^2 + 1) + \sum_{i=1}^n \log(x_i + \theta^2 + \theta + 1) \\ &\quad - \left(\sum_{i=1}^n x_i + 2n \right) \log(\theta + 1) \end{aligned}$$

The first derivative of the log-likelihood function is given by

$$\frac{d \log L(\theta; x)}{d\theta} = \frac{2n}{\theta(\theta^2 + 1)} + \sum_{i=1}^n \frac{2\theta + 1}{x_i + \theta^2 + \theta + 1} - \frac{n(\bar{x} + 2)}{\theta + 1} = 0$$

where \bar{x} is the sample mean.

The maximum likelihood estimate (MLE), $\hat{\theta}$ of θ of the PSD is the solution of the equation $\frac{d \log L(\theta; x)}{d\theta} = 0$ and is given by the solution of the following non-linear equation.

Results and Discussion

Simulation Study

In this section, we perform a simulation study to compare the performance of each estimate of the parameter θ numerically in terms of the mean-squared error (MSE) and the bias. A simulation study was carried out $N = 1000$ times, with different values of parameter θ ($\theta = 0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1, 1.5, 3, 5$) and sample size ($n = 50, 100, 150, 200, 250, 300, 350, 400, 450, 500$). The following measures were computed:

- The average bias of the simulated estimates $\hat{\theta}_i, i = 1, 2, \dots, N$.

$$\text{Average Bias} = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta)$$

- Average MSE (Mean Square Error) of the simulated estimates $\hat{\theta}_i, i = 1, 2, \dots, N$.

$$\text{Average MSE} = \frac{1}{N} \sum_{i=1}^N (\hat{\theta}_i - \theta)^2$$

To generate random data $X_i, i = 1, 2, \dots, n$ from the Poisson-Shanker distribution with parameter θ , we have the following steps:

1. Determine the value of parameter θ and the sample size n .
2. Generate the random data λ_i follows a Shanker distribution with parameter θ .
3. Generate random data X_i follows a Poisson distribution with parameter λ_i .
4. Estimate the parameter of Poisson-Shanker distribution using maximum likelihood method.
5. Calculating the value of bias and MSE between $\hat{\theta}$ and θ .
6. Repeat step 1-4 N times
7. Then calculate the average bias and MSE

8. Repeat step 1-6 for several variations the pairs of values of n and θ .

The simulation study obtained by using software RStudio 1.0.153. The results are summarized in Figure 7, Figure 8, Figure 9, and Figure 10 below.

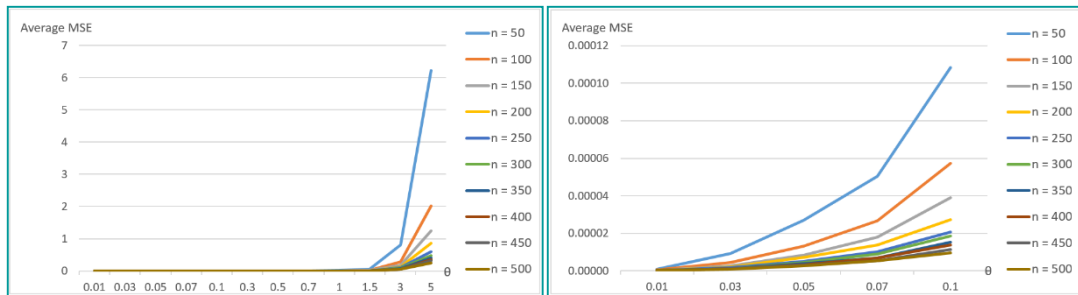


Figure 7. Graphs of MSE of the estimated parameter θ for some value of n .

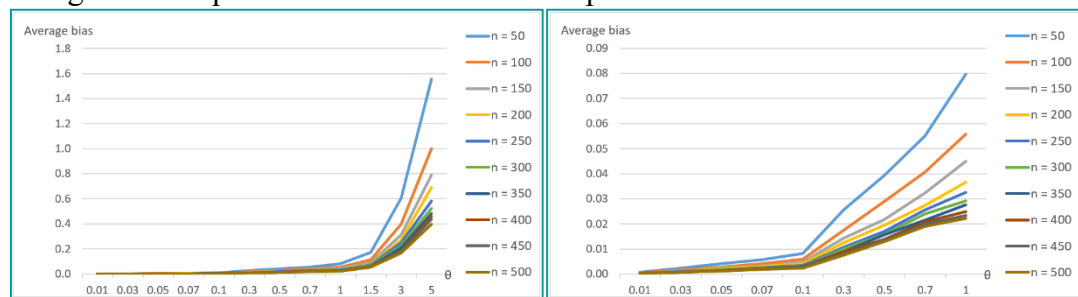


Figure 8. Graphs of bias of the estimated parameter θ for some value of n

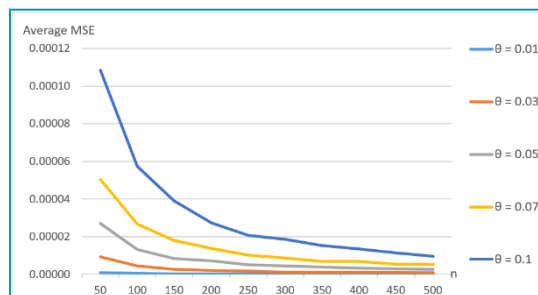
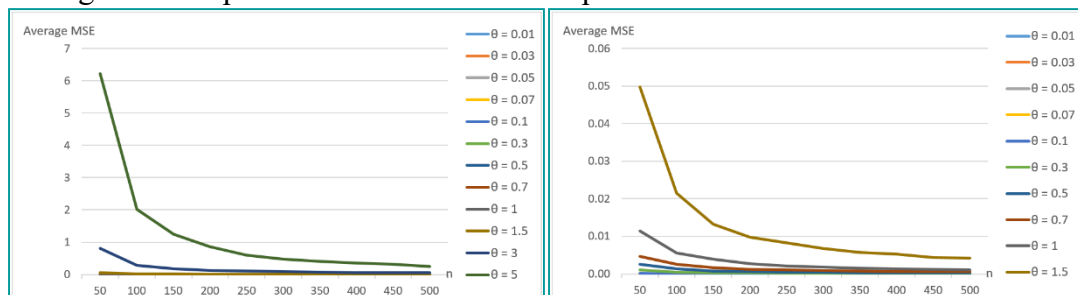


Figure 9. Graphs of MSE of the estimated parameter θ for some value of θ

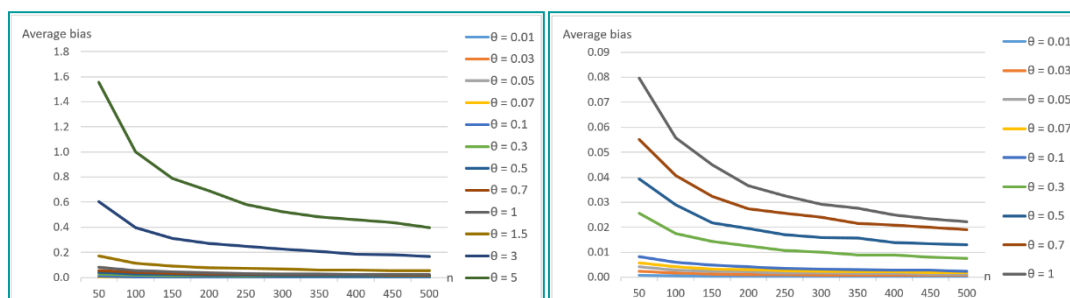


Figure 10. Graphs of bias of the estimated parameter θ for some value of θ

Based on Figure 7 and Figure 8, the larger value of the parameter we use, the larger MSE and bias value we obtain. Otherwise, based on Figure 9 and Figure 10, the larger sample size we use, the smaller MSE and bias value we obtain.

Applications

In this section, we use a real data set to model claim frequency with the Poisson-Shanker distribution. The data set that we use in this paper is claim frequencies for automobile portfolio of a Turkish insurance company occurred between 2012 and 2014 (Sarul & Sahin, 2015). The data contains information from 10.814 policyholders and it is a company-specific data.

Claim frequency	0	1	2	3	4	5
Observed values	8544	1796	370	81	22	1

We fitted both the Poisson and Poisson-Shanker distributions to this data set. The method of maximum likelihood was used to estimates $\hat{\theta}$.

Table 1. Number of trapped response *snowshoe hares*

Claim frequency	Observed values	Expected values	
		Poisson	Poisson-Shanker
0	8544	8543.470932	8291.74264
1	1796	1794.777952	2202.130518
2	370	376.251502	292.4235368
3	81	78.72592	25.88741832
≥ 4	23	16.448094	1.718809602
ML estimate		$\hat{\lambda} = 0.265582$	$\hat{\theta} = 3.988029$
χ^2		483.9765	2.78029
d.f.		3	3
p-value		1.4161E-104	0.426753

It is obvious from the table above that Poisson-Shanker distribution gives a better fit in modelling data than Poisson distribution.

Conclusion

Poisson-Shanker distribution is obtained by mixing the Poisson and Shanker distribution. The properties of the Poisson-Shanker distribution which includes the pmf, cdf, survival function, increasing hazard rate, factorial moments, mean, variance, overdispersion, skewness, and kurtosis have obtained. Parameters estimation are also obtained using the maximum likelihood method and the usefulness of the Poisson-Shanker distribution is illustrated by a real data set. The characteristics of the Poisson-Shanker distribution parameter is also obtained by numerical simulation with several variations in parameter values and sample size. The result is average MSE and average bias of the estimated parameter θ will increase when the parameter value rises for a value of n and will decrease when the value of n rises for a parameter value.

References

- Hogg, R. V., McKean, Craig, A. T. (2018). *Introduction to mathematical statistics*. New Jersey: Prentice-Hall.
- Klugman, S. A., Panjer, H. H. & Wilmot, G. E. (2012). *Loss Models: From Data to Decisions (4th ed.)*. New Jersey: Wiley.
- Panjer, H. H. (2006). *Mixed Poisson Distributions (3rd ed.)*. New Jersey: Wiley.
- Sarul, L. S. & Sahin, S. (2015). An Application of Claim Frequency Data Using Zero Inflated and Hurdle Models in General Insurance. *The Journal of Bussiness, Economics, & Finance*, 4(4), 732-743.
- Shanker, R. (2015). Shanker Distribution and Its Applications. *International Journal of Statistics and Applications*, 5(6), 338-348.
- Shanker, R. (2016). The Discrete Poisson-Shanker Distribution. *Jacobs Journal of Biostatistics*, 1(1), 1-7.
- Shanker, R. (2017). On Discrete Poisson-Shanker Distribution and Its Applications. *Biometrics & Biostatistics International Journal*, 5(1).