The Journal of Applied Science **Vol. 20 No. 2: 13-22 [2021]**
วารสารวิทยาศาสตร์ประยุกต์ doi: 10.14416/j.appsci.2021.02.002
ISSN 1513-7805 Printed in Thailand ISSN 2586-9663 (Online)

Research Article

Confidence interval for the parameter of the zero-truncated Poisson distribution

Patarawan Sangnawakij [1]*

[1]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, 12120 Thailand
***E-mail**: patarawan@mathstat.sci.tu.ac.th
Received: 27/12/2020; Revised: 27/03/2021; Accepted: 17/04/2021

Abstract

This paper introduces a confidence interval for the parameter in a zero-truncated Poisson distribution. We adjust the profile likelihood method to construct this confidence interval by using a function of parameter as a nuisance. The performance of the proposed estimator is investigated through simulations, and compared with the conventional Wald confidence interval. From the results, the proposed estimator provides a good performance in terms of coverage probability in all cases in the study. It also has the short interval length. The practicality of our approach is confirmed by application to two real datasets, on a cholera-epidemic and on mortality rates of infants on an estate.

Keywords: coverage probability, interval estimation, non-zero count, Wald method

Introduction

Parameter estimation is an important method in statistical inference. It is widely used in applications and research areas that rely on continuous or count data. Many approaches have been proposed for estimating the parameter of interest with good accuracy. In particular, interval estimation has been developed for the parameters, reliability functions, and applied in many areas, such as medical science, social science, and engineering. This method is used to calculate an interval, or range, of plausible values of an unknown parameter (Casella & Berger, 2002). It can also describe the probability level at which the confidence interval will contain the true value, in contrast with point estimation which provides an approximate value only.

In this paper, we focus on the count outcome variable as a Poisson distribution. Let X be a Poisson variable with parameter mean $\lambda > 0$. It is denoted as $X \sim P(\lambda)$. The probability density function (PDF) of X is given by

$$P(X = x) = \frac{\exp(-\lambda)\lambda^{x}}{x!},$$

where the observed value x = 0, 1, 2, ... This probability model is usually used in analysis of data containing zero and positive events that have low probabilities of occurrence within some definite time or area range. However, observed data can be truncated. Only positive values of the Poisson variable are available, or no zero counts are observed at all. In such a case, the *zero-truncated Poisson* (ZTP) distribution is therefore more appropriate than the Poisson model (Dietz & Böhning, 2000). The ZTP model is often used in socio-economical applications, including research on alcohol and illicit drug use in the social sciences, and biological sciences. The general PDF of the ZTP variable is given by

$$\frac{P(X=x,\lambda)}{P(X>0,\lambda)}=\frac{P(X=x,\lambda)}{1-P(X=0,\lambda)},$$

where x = 1, 2, 3, ... (Tang et al., 2012). Figure 1 depicts the ZTP distribution for different values of λ. Papers related to parameter estimation in the Poisson distribution with missing zero have been discussed by Dahiya & Gross (1973), Johnson et al. (2005), and Nasiri (2011; 2015). For interval estimation, Daidoji & Iwasaki (2012) introduced a confidence interval for λ in a ZTP distribution. They derived the likelihood function and estimated the variance of the maximum likelihood (ML) estimator for building the confidence interval using the Wald method. Simulations were used to confirm the performance of the confidence interval. It was found that the coverage probabilities of the confidence interval proposed by Daidoji & Iwasaki (2012) were lower than the target probability in many cases, especially when the sample sizes were less than 50. Many techniques have been developed to estimate the functions of parameter in the distribution related to the Poisson model. However, most of them were considered in the zero-inflated Poisson (ZIP) distribution, for example, Taylor et al. (2001), Zhou & Tu (2000), Chen et al. (2010), and Paneru et al. (2018).

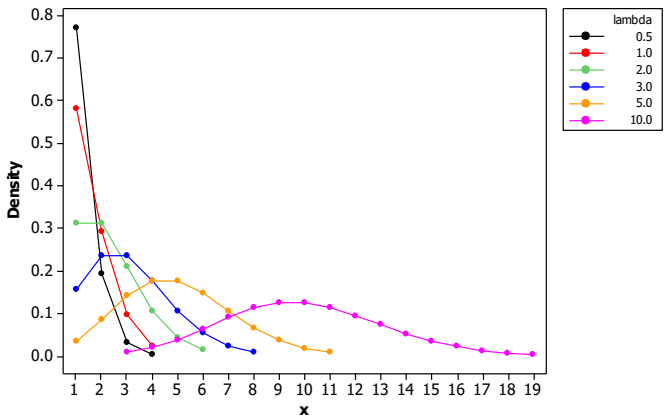

Figure 1 Zero-truncated Poisson probability density for different values of λ

We know that confidence interval which has a high coverage probability will cover the true parameter value better than that has a low coverage probability. However, as noted in Daidoji & Iwasaki (2012) little papers on interval estimation for the ZTP distribution have been shown. The confidence interval introduced in that paper is also unsatisfactory in terms of coverage probability. So, we see that this is an important problem and needed to address. The *profile likelihood* is an alternative approach for dealing with the nuisance parameters in a distribution. It can be used to derive the variance of the ML estimator (Young & Smith, 2005; Böhning et al., 2008). For the ZTP distribution, we know that λ is the only one parameter in the model. The profile method is then reasonably adjusted in this case. The idea for constructing the confidence interval in this paper is that we assume a function of the population mean, in terms of exponential, of the ZTP distribution to be a nuisance parameter, and eliminate this function using the profile method. Then, the parameter of interest is estimated. The variance of the estimator obtained from this method is used to build the new confidence interval for λ in the ZTP distribution, which may improve the coverage probability of the confidence interval. Our approach will show here that eliminating a complex function of parameter by using a simple form can be used and will provide a good estimator.

The rest of this paper is organized as follows. In Section 2, the definition of the ZTP distribution and the confidence interval of Daidoji & Iwasaki (2012) for λ are explained. We also derive the likelihood using the adjusted profile function method and introduce the novel confidence interval in Section 2. In Section 3, we investigate the performance of the proposed confidence interval using simulations in various situations, and compare it with that of the existing estimator. Two real data examples are used to illustrate our method and presented in Section 4. Finally, Section 5 presents our conclusions.

Methods

Let $Y = (Y_1, Y_2, ..., Y_n)$ be a random sample of size n from a zero-truncated Poisson (ZTP) distribution. The conditional PDF of Y is given by

$$P(Y = y_i \mid y_i > 0, \lambda) = \frac{\exp(-\lambda)\lambda^{y_i} / y_i!}{1 - \exp(-\lambda)},$$

for i = 1, 2, ..., n. The observed value y_i = 1, 2, 3, ... and the mean parameter of the un-truncated Poisson distribution $\lambda > 0$. The mean and variance of Y are given by

$$E(Y) = \frac{\lambda}{1 - \exp(-\lambda)} \qquad (1)$$

and

$$Var(Y) = \frac{\lambda}{1 - \exp(-\lambda)}\left(1 - \frac{\lambda}{\exp(\lambda) - 1}\right),$$

respectively (Winkelmann, 2008). The point estimator of λ is obtained by maximizing the log-likelihood function $\log L(\lambda, y_i)$ or the logarithm of joint PDF of $Y_1, Y_2, ..., Y_n$. Thus, the ML estimator for λ of the ZTP model is derived by the following processes:

$$\frac{\partial}{\partial\lambda}\log L(\lambda, y_i) = \frac{\partial}{\partial\lambda}\left(\log\prod_{i=1}^{n} P(Y = y_i \mid y_i > 0, \lambda)\right)$$

or

$$\frac{\partial}{\partial\lambda}\log L(\lambda, y_i) = \frac{\partial}{\partial\lambda}\left(\sum_{i=1}^{n} y_i \log\lambda - n\lambda - \sum_{i=1}^{n}\log(y_i!) - n\log(1 - \exp(-\lambda))\right).$$

Solving the equation $\frac{\partial}{\partial\lambda}\log L(\lambda, y_i) = 0$ for λ, we have

$$\frac{\lambda}{1 - \exp(-\lambda)} = \bar{Y}, \qquad (2)$$

where $\bar{Y} = \sum_{i=1}^{n} Y_i / n$ denotes the sample mean. Since the ML estimator for λ does not provide the closed-form solution, the estimated parameter is then approximated by the iterative approach, using the expression:

$$\hat{\lambda}^{(t+1)} = \bar{Y}(1 - \exp(-\hat{\lambda}^{(t)})). \qquad (3)$$

In calculation, the suggested initial value is corresponded to the sample mean of variable Y. The procedure will be iterated until the value of $\hat{\lambda}$ in the $(t+1)$ th and the value of $\hat{\lambda}$ in the $t-$ th converge. In other word, the difference of these values must be small and close to zero. Note that since $\hat{\lambda}$ is ML estimator, its function has invariance property (Tan & Drossos, 1975).

1. Confidence interval of Daidoji and Iwasaki (2012)

Basically, the $(1-\alpha)$ 100% confidence interval for λ is constructed based on the Wald method. The general form is given by

$$\hat{\lambda} \pm Z_{\alpha/2}\sqrt{Var(\hat{\lambda})},$$

where $\hat{\lambda}$ is the ML estimator for λ, $Z_{\alpha/2}$ is the $(\alpha/2)$ th quantile of the standard normal distribution, and $Var(\hat{\lambda})$ is the estimated variance of $\hat{\lambda}$. Based on a property of the ML estimator, $\hat{\lambda}$ approximately converges to a normal distribution with mean λ and variance $1/I(\lambda)$, where $I(\lambda)$ is the expected Fisher information (Casella & Berger, 2002). From the ZTP distribution, $I(\lambda)$ is given as

$$I(\lambda) = -E\left(\frac{\partial^2}{\partial\lambda^2}\log L(\lambda, y_i)\right) = \frac{n(1-(\lambda+1)\exp(-\lambda))}{\lambda(1-\exp(-\lambda))^2}.$$

Using the estimated variance of $\hat{\lambda}$ from the inverse of $I(\hat{\lambda})$, Daidoji & Iwasaki (2012) introduced the Wald-type confidence interval for λ, which is given as follows:

$$CI_{DI} = \hat{\lambda} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\lambda}(1-\exp(-\hat{\lambda}))^2}{n(1-(\hat{\lambda}+1)\exp(-\hat{\lambda}))}}. \qquad (4)$$

As can be seen from their paper, CI_{DI} provides coverage probabilities lower than the nominal level in many cases in simulations. The new confidence interval is therefore considered to deal with this problem.

2. Proposed confidence interval

The interesting point of our method is started from the two equivalent equations related to the mean: the population mean of the ZTP distribution given in (1) and the expression corresponding to the mean given in (2). From these two formulas, we have

$$E(Y) = \frac{\lambda}{1-\exp(-\lambda)} = \bar{Y}$$

which follows that

$$1-\exp(-\lambda) = \frac{\lambda}{\bar{Y}}. \qquad (5)$$

Here, the function $1-\exp(-\lambda)$ in the log-likelihood, $\log L(\lambda, y_i)$ as noted above equation (2), is assumed as the nuisance parameter. It will be eliminated by substituting $\lambda/\bar{Y}$ as used in the profile likelihood method. Then, we achieve the log-likelihood:

$$\log L_p(\lambda, y_i) = \sum_{i=1}^{n} y_i \log\lambda - n\lambda - \sum_{i=1}^{n}\log(y_i!) - n\log\lambda + n\log\bar{y}.$$

This function is used to derive the expected Fisher information, which is given by

$$I_p(\lambda) = -E\left(\frac{\partial^2}{\partial\lambda^2}\log L_p(\lambda, y_i)\right) = \frac{n(\lambda-1+\exp(-\lambda))}{\lambda^2(1-\exp(-\lambda))}.$$

Note that this function is entirely different from $I(\lambda)$ presented in the previous method derived by Daidoji & Iwasaki (2012). Using the inverse of $I_p(\hat{\lambda})$ as the estimated variance of the estimator, the novel confidence interval for λ is then given by

 doi: 10.14416/j.appsci.2021.02.002

$$CI_{PR} = \hat{\lambda} \pm Z_{\alpha/2}\sqrt{\frac{\hat{\lambda}^2(1-\exp(-\hat{\lambda}))}{n(\hat{\lambda}-1+\exp(-\hat{\lambda}))}}, \qquad (6)$$

where $\hat{\lambda}$ is the ML estimator obtained by an iterative method shown in (3). Again, we note that the confidence intervals given in (6) and (4) have entirely different formula. The performance of these two confidence intervals is investigated using simulations.

Simulation study

The simulations were performed to explore the performance of the methods for estimating the confidence interval for λ of the ZTP distribution. The study was designed to cover cases with different sample sizes, as n = 10, 20, 50, and 100, reflecting small to large samples. The true parameter (λ) was given by 0.5, 1, 2, and 3. The confidence level ($1-\alpha$) was set at 0.95. Following the simulation method in Daidoji & Iwasaki (2012), we generated the data from standard Poisson random numbers by discarding the zero values using R programming (R Core Team, 2019). We then computed the maximum likelihood estimator for λ from $\lambda^{(t+1)} = \bar{y}(1-\exp(-\lambda^{(t)}))$. The sequence started at $\lambda^{(0)} = \bar{y}$, and stopped when $|\lambda^{(t+1)} - \lambda^{(t)}| < 0.00001$. Each combination of situation was repeated 10,000 times. The performance of the confidence interval was calculated by

$$ACP = \frac{n(L \le \lambda \le U)}{10,000}$$

for the average coverage probability and

$$AL = \frac{\sum_{h=1}^{10,000}(U_h - L_h)}{10,000}$$

for the average length, where $n(L \le \lambda \le U)$ is the number of simulation runs for λ that lies between the lower limit L and upper limit U. A confidence interval which has a coverage probability greater than or close to the nominal coverage level means that it contains the true value with a given probability. In other word, it can precisely estimate the parameter of interest. The confidence interval that satisfies the criterion is the best in comparison. If the confidence intervals perform well in terms of coverage probability and have the same average probability value, they will be used to compare the average length. The confidence interval which has a short length interval denotes that the estimate is close to the parameter value, which is needed in interval estimation.

The performance of the confidence intervals considered in this paper is summarized in Table 1. The coverage probabilities of the proposed confidence interval, namely CI_{PR}, were grater than or close to the nominal coverage probability at 0.95 in many cases in the study. They are increased, when λ or n increased. Obviously, CI_{PR} performed better than the compaired estimators in terms of coverage probability. The confidence interval of Daidoji & Iwasaki (2012), namely CI_{DI}, had the coverage probability much lower than 0.95 when $n < 50$. The behavior of CI_{DI} in the current simulation study was similar to that presented in Daidoji & Iwasaki (2012). Next, we considered the performance of the confidence intervals in terms of average length. CI_{PR} had the short expected length, which was acceptable. The expected lengths of CI_{PR} and CI_{DI} were slightly different. However, we noted that CI_{PR} actually covered the true parameter λ in computation. These results are also shown graphically in Figure 2.

Table 1 Coverage probability and expected length of the 95% confidence intervals for λ in the zero-truncated Poisson distribution

		Coverage probability		Expected length	
n	λ	CI_{DI}	CI_{PR}	CI_{DI}	CI_{PR}
10	0.5	0.9209	0.9409	1.0575	1.1013
	1	0.9000	0.9466	1.4725	1.5700
	2	0.9210	0.9556	1.9440	2.1340
	3	0.9392	0.9678	2.2723	2.5188
20	0.5	0.9310	0.9421	0.7863	0.8170
	1	0.9207	0.9472	1.0603	1.1301
	2	0.9356	0.9606	1.3843	1.5214
	3	0.9501	0.9703	1.6080	1.7841
50	0.5	0.9460	0.9485	0.5075	0.5267
	1	0.9485	0.9547	0.6779	0.7227
	2	0.9449	0.9679	0.8781	0.9656
	3	0.9457	0.9671	1.0190	1.1312
100	0.5	0.9337	0.9496	0.3600	0.3735
	1	0.9456	0.9544	0.4811	0.5129
	2	0.9479	0.9689	0.6213	0.6833
	3	0.9479	0.9691	0.7206	0.8001

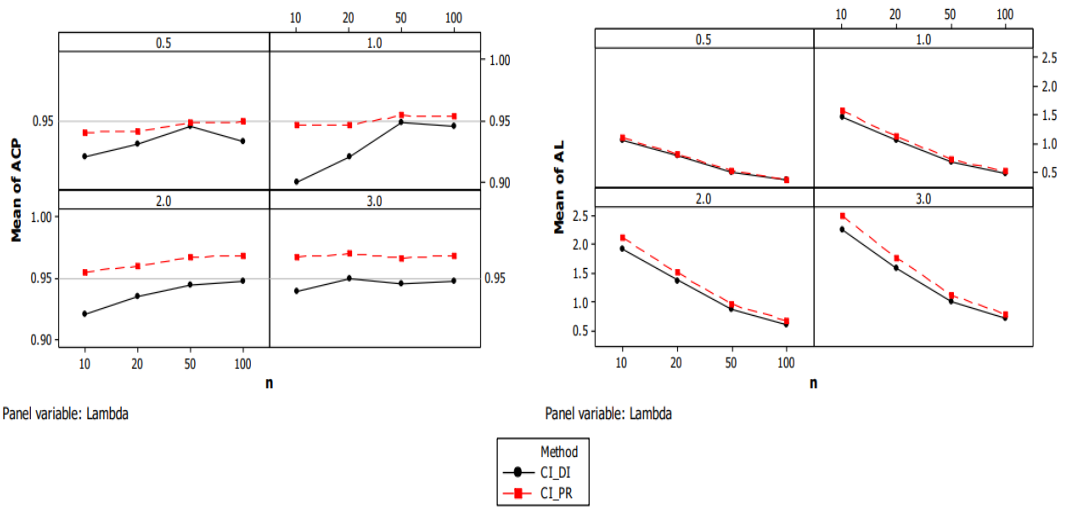

Figure 2 Plots of coverage probability (left) and average length (right) of the 95% confidence intervals on the settings: n = 10, 20, 50, 100 and λ = 0.5, 1, 2, 3.

Overall, the coverage probability of the proposed confidence interval outperformed that of the compared confidence interval in all cases. The length of interval was also small on average. The confidence interval proposed in this paper is suggested to estimate the Poisson parameter in the ZTP distribution.

Numerical illustration

There are two real data examples in this section.

1. Cholera data

We used the data on a cholera epidemic in an Indian village obtained from Böhning & Schön (2005). The dataset included the number of households (observed frequency: O_i) with exact numbers of cholera i cases: O_1 = 32, O_2 = 16, O_3 = 6, and O_4 = 1, so that the sample size n = 55. Böhning & Schön (2005) pointed out that, although the original data presented in McKendrick (1926) reported the frequency of houses with no cases of cholera, households with zero cases were ignored because they were not relevant to determination of the number of affected houses. Only the associated $i-$th household that was clearly affected by cholera, or any case count which was greater than zero, was applied. The histogram of this dataset is shown in Figure 3.

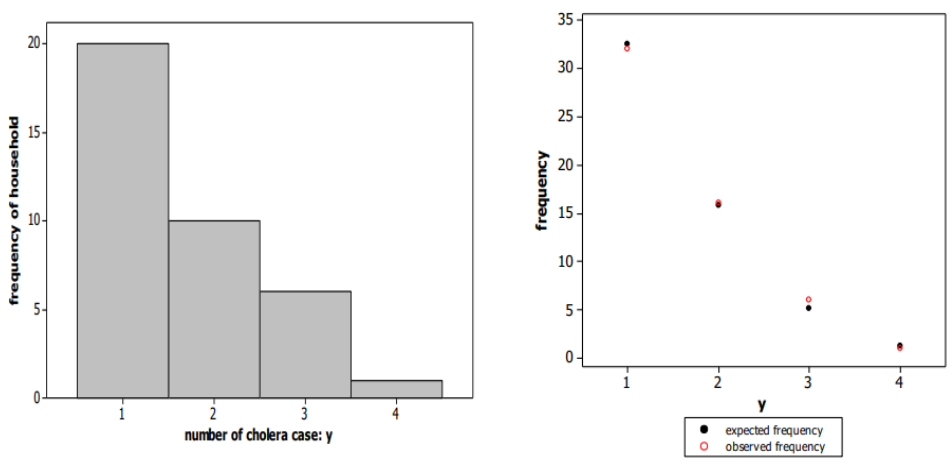

Figure 3 Histogram (left) and plot of observed and expected frequencies (right) under the zero-truncated Poisson distribution for cholera data (AIC = 111.56)

Before applying our method, the distribution of cholera data was checked. We tested the following hypotheses:

H_0 : The data follow the ZTP distribution

H_1 : The data do not follow the ZTP distribution.

Using the chi-square goodness of fit test (Cochran, 1952),

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i is the observed value and E_i is the expected value related to the ZTP probability. Under the null hypothesis, χ^2 has a chi-square distribution with $k-2$ degrees of freedom. Note that, in calculation the expected frequency E_i which was less than 5 was pooled. Then, the observed test statistic was given by 6.63 with a p-value of 0.01. It can be seen that the p-value was borderline at a significance level at 0.01. We next considered the plot of observed and estimated frequencies to support the homogeneity of the distribution. It is shown in Figure 3 (right). Clearly, the cholera data followed a ZTP distribution with Akaike information criterion (AIC) of 111.56. The ML estimator for λ was 0.97. The 95% confidence intervals obtained from the proposed and existing methods were calculated. CI_{PR} was (0.63, 1.31) and CI_{DI} was (0.65, 1.29) with lengths of interval of 0.68 and 0.64, respectively. Based on the invariant property of ML estimation and our method, the 95% confidence interval for the mean of cholera cases in an Indian village was given as (1.35, 1.80), or 1.56 on average.

2. Infant mortality data

We used the data on the number of mothers on an estate who had at least one live birth and one neonatal death. They were obtained from Shanker et al. (2015). The original data reported the number of mothers from neonatal deaths i cases: O_1 = 71, O_2 = 32, O_3 = 7, O_4 = 5, and O_5 = 3, to the total n = 118. Only observable counts were reported. The frequency distribution of the observed used data is shown in Figure 4 (left). The chi-square statistic was used to test the distribution of a ZTP. This was given as 0.68 with a p-value of 0.41. Therefore, the infant mortality data considered here significantly followed a ZTP distribution with AIC = 259.46. This dataset was suitable for our purposes. The ML estimate for λ obtained by the iterative method was estimated as 1.06. The 95% confidence interval from CI_{PR} was given by (0.81, 1.29) with interval length 0.48. The CI_{DI} was (0.83, 1.28) with interval length of 0.45. The mean of number of mothers from infant deaths was 1.62 cases with 95% confidence interval of (1.46, 1.78).

From this example, if the data were assumed to be the Poisson distribution as often used with $\hat{\lambda}$ = 1.62 and AIC = 330, interval estimation for the mean was given as (1.33, 1.91). It can be seen that the interval length of this method, where it did not come from the reasonable probability model, was greater than that of the proposed method. We just point out that the use of appropriate statistical tool for the available data will lead to the right solution.

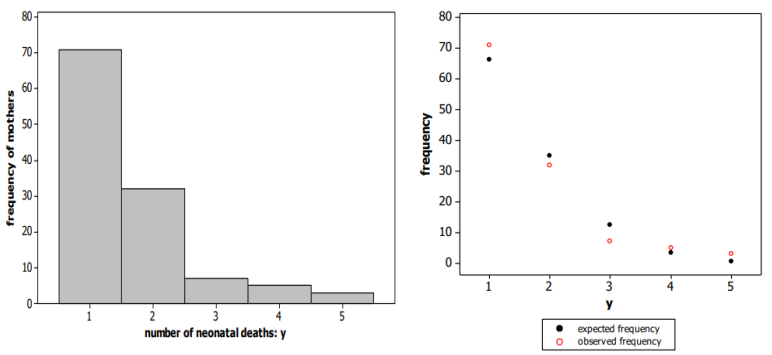

Figure 4 Histogram (left) and plot of observed and expected frequencies (right) under the zero-truncated Poisson distribution for infant mortality data (AIC = 259.46)

Based on these two examples, we conclude that the lengths of the confidence intervals are small, with CI_{DI} having a slightly smaller length of interval than CI_{PR}. The findings considered in this section therefore support the simulation results.

Conclusion

Profile likelihood method has been generally used to construct the confidence interval. This method keeps the parameter of interest fixed and maximizes the nuisance parameter (for elimination the nuisance parameter). From the ZTP model, we point out that the population mean of this distribution is equivalent to the sample mean, leading to $\bar{Y} = \lambda / (1 - \exp(-\lambda))$. For this, the function in the denominator of the previous equation is eliminated to keep *only* the parameter of interest (λ). The likelihood function under this method is then used to derive the variance of the ML estimator, and is applied to estimate the confidence interval. We note here that this adjusts the idea of the profile likelihood method.

The performance of the proposed confidence interval was conducted by simulations. The results confirmed its good performance in terms of coverage probability and expected length. The coverage probabilities of the proposed confidence interval were satisfied the nominal coverage level, while the expected lengths were small. The confidence interval for λ proposed in this paper outperformed the confidence interval constructed based on the traditional method using the likelihood function. This shows that our proposed method is accuracy and precision to estimate the true parameter. Moreover, it is easy to compute our confidence interval using a basic programming language. In practical terms, the novel confidence interval is therefore recommended for estimating the parameter in the zero-truncated Poisson distribution.

Acknowledgement

The author would like to thank Editor and the reviewers for the valuable comments and suggestions to improve this paper.

References

Böhning, D., Kuhnert, R., & Rattanasiri, S. (2008). *Meta-Analysis of Binary Data using Profile Likelihood*. Boca Raton, USA: Chapman & Hall/CRC.

Böhning, D., & Schön, D. (2005). Nonparametric maximum likelihood estimation of population size based on the counting distribution. *The Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *54*(4), 721-737.

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Duxbury Press.

Chen, H., Chen, J., & Chen, S. Y. (2010). Confidence intervals for the mean of a population containing many zero values under unequal-probability sampling. *Canadian Journal of Statistics*, *38*(4), 582-597.

Cochran, W. G. (1952). The χ^2 test of goodness of fit. *The Annals of Mathematical Statistics*, *23*(3), 315-345.

Dahiya, R. C., & Gross, A. J. (1973). Estimating the zero class from a truncated poisson sample. *The Journal of the American Statistical Association*, *68*(343), 731-733.

Daidoji, K., & Iwasaki, M. (2012). On interval estimation of the poisson parameter in a zero-truncated poisson distribution. *Journal of the Japanese Society of Computational Statistics*, *25*(1), 1-12.

Dietz, E., & Böhning, D. (2000). On estimation of the poisson parameter in zeromodified poisson models. *Computational Statistics and Data Analysis*, *34*(4), 441-459.

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

McKendrick, A. G. (1926). Application of mathematics to medical problems. *Edinburgh Mathematical Society*, *44*(1), 98-130.

Nasiri, P. (2011). Estimation parameter of zero truncated mixture poisson models. *International Journal of Mathematical Analysis*, *5*(9), 465-470.

Nasiri, P. (2015). Estimating the parameters of modified poisson distribution at zero. *Journal of Applied Sciences*, *15*(4), 719-722.

Paneru, K., Padget, R. N., & Chen, H. (2018). Estimation of zero-inflated population mean: a bootstrapping approach. *The Journal of Modern Applied Statistical Methods*, *17*(1), 1-14.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Shanker, R., Fesshaye, H., Selvaraj, S., & Yemane, A. (2015). On zero-truncation of poisson and poisson-lindley distributions and their applications. *Biometrics & Biostatistics International Journal*, *2*(6), 1-14.

Tan, P., & Drossos, D. (1975). Invariance properties of maximum likelihood estimators. *Mathematics Magazine*, *48*(1), 37-41.

Tang, W., He, H., & Tu, X. M. (2012). *Applied Categorical and Count Data Analysis*. CRC Press.

Taylor, D. J., Kupper, L. L., Rappaport, S. M., & Lyles, R. H. (2001). A mixture model for occupational exposure mean testing with a limit of detection. *Biometrics*, *57*(3), 681-688.

Winkelmann, R. (2008). *Econometric Analysis of Count Data*. Springer-Verlag Berlin Heidelberg.

Young, G. A., & Smith, R. L. (2005). *Essentials of Statistical Inference*. Cambridge University Press.

Zhou, X. H., & Tu, W. (2000). Confidence intervals for the mean of diagnostic test charge data containing zeros. *Biometrics*, *56*(4), 1118-1125.