JASET KMUTNB
Research Article

Journal of Applied Science and Emerging Technology (JASET)
Published by Faculty of Applied Science, KMUTNB, Thailand
doi: 10.14416/JASET.KMUTNB.2023.03.003
Vol. 22, No. 3 [2023]: e251640

Confidence Intervals using Bootstrap Methods for the Parameter of the zero-truncated Poisson-Shanker Distribution and their Application

Panicha Kaskasamkul[1] and Wararit Panichkitkosolkul[2*]

[1]Department of Mathematics, Faculty of Science, Naresuan University, Thailand

[2]Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Thailand

ABSTRACT

Count data with non-zero values can be modeled using the zero-truncated Poisson-Shanker distribution in a variety of situations. However, the confidence interval estimation for the parameter has not yet been examined. In this paper, confidence interval estimation based on percentile bootstrap, simple bootstrap, biased-corrected and accelerated bootstrap, and bootstrap-t methods was examined in terms of coverage probability, lower and upper error rate, and average interval length through Monte Carlo simulation. The simulation results indicate that attaining the nominal confidence level using the bootstrap methods were not possible for small sample sizes, regardless of the other settings. Furthermore, in cases when the sample size was large, there was no significant difference observed in the performances of the all methods. In several scenarios, the bias-corrected and accelerated bootstrap method exhibited superior performance compared to the other methods. Last, we used the bootstrap methods to calculate the confidence interval for the parameter of the zero-truncated Poisson-Shanker distribution via three numerical examples, and the results of these examples matched those from the simulation study.

KEYWORDS: interval estimation, count data, Shanker distribution, bootstrap interval, simulation

*__Corresponding Author:__ __wararit@mathstat.sci.tu.ac.th__

Received: 11/02/2023; **Revised:** 17/08/2023; **Accepted:** 25/08/2023

1. INTRODUCTION

The Poisson distribution is a discrete probability distribution that quantifies the likelihood of a certain number of events taking place within specified time or spaces. (Kissell and Poserina, 2017; Siegel and Wanger, 2022). This distribution is used to model data like the number of orders a firm will receive tomorrow, the number of people who will apply for a job tomorrow, the number of defects on a finished product, the number of confirmed COVID-19 cases per day, the number of customers entering a post office on a given day, etc. (Siegel, 2016; Ross, 2006).

The probability mass function (pmf) of a Poisson distribution is defined as

$$p(x;\lambda)=\frac{e^{-\lambda}\lambda^{x}}{x!},\quad x=0,1,2,...\ ;\lambda>0 \qquad (1)$$

where e is a constant approximately equal to 2.71828 and λ is the mean number of events within a given interval of time or space. The Poisson distribution is frequently employed to model the number of occurrences under the assumption that events are independent and occur at a constant rate. In real situations, however, this assumption may not always hold true. The typical Poisson distribution may be inadequate if the underlying rate of occurrence is heterogeneous or variable. The mixed Poisson distribution lets the rate parameter be random, dealing with this constraint. Instead of having a constant rate parameter, the rate parameter is assumed to follow a certain distribution.

Shanker (2016) introduced the mixed Poisson distribution known as the Poisson- Shanker (PS) distribution and investigated its mathematical and statistical properties. Combining the Poisson distribution and the Shanker distribution (Shanker, 2015) yields this distribution. When the Poisson parameter λ follows a Shanker distribution, the Poisson distribution gives rise to the PS distribution

(Shanker, 2016). The pmf of the PS distribution is given by

$$p_0(x;\theta)=\frac{\theta^2}{\theta^2+1}\frac{x+(\theta^2+\theta+1)}{(\theta+1)^{x+2}},\ x=0,1,2,...\ ;\theta>0. \quad (2)$$

The mathematical and statistical properties of the PS distribution for modeling count data were established by Shanker (2016). When applied to several real-world data sets, the PS distribution proved to be more appropriate than the Poisson and Poisson-Lindley (Sankaran, 1970) distributions.

The Shanker distribution was introduced by Shanker (2015) as a lifetime continuous distribution with a probability density function (pdf) defined as

$$f(\lambda;\theta)=\frac{\theta^2}{\theta^2+1}(\theta+\lambda)e^{-\theta\lambda},\lambda>0,\theta>0. \quad (3)$$

It is a composite of an exponential distribution with a scale parameter θ, and a gamma distribution with a shape parameter 2 and a scale parameter θ, with respective proportions $\theta^2/(\theta^2+1)$ and $1/(\theta^2+1)$. Shanker (2015) showed that the pdf in Eq.(3) provides a better model than the exponential and Lindley (Lindley, 1958) distributions for modeling lifetime data.

The probability model of PS distribution can be utilized to analyze data that contains both zero and positive values with low occurrence probabilities within a predefined time or area range (Sangnawakij, 2021). However, probability models can become truncated when a range of possible values for the variables is either disregarded or impossible to observe. Zero truncation is commonly employed when analyzing count data without zeros. David and Johnson (1952) developed the zero-truncated (ZT) Poisson (ZTP) distribution, which has been applied to datasets such as the length of hospital stays, the number of published journal articles across various disciplines, the number of children ever born to mothers over 40 years old, and the number of passengers in cars (Hussain, 2020). A ZT distribution's pmf can be derived as

$$p(x;\theta)=\frac{p_0(x;\theta)}{1-p_0(0;\theta)},\ x=1,2,3,... \quad (4)$$

where $p_0(x;\theta)$ is the pmf of the un-truncated distribution. Shanker (2017) proposed the zero-truncated Poisson-Shanker (ZTPS) distribution and its properties, such as the moment, coefficient of variation, skewness, kurtosis and the dispersion index. The method of moments and the maximum likelihood have also been derived for estimating its parameter. Furthermore, when the ZTPS distribution was applied to real data set, it was more suitable than the ZTP and zero-truncated Poisson-Lindley (ZTPL) (Shanker et al., 2015) distributions.

To our knowledge, no research has been conducted to estimate the confidence interval for the ZTPS distribution parameter θ. Bootstrap methods provide a way to estimate confidence intervals by quantifying uncertainties in statistical inferences based on a sample of data. The concept is to run a simulation study based on the actual data for estimating the likely extent of sampling error (Wood, 2004). Therefore, the objective of the current study is to assess the efficiencies of four bootstrap methods, namely the percentile bootstrap (PB), the simple bootstrap (SB), the bias- corrected and accelerated (BCa) bootstrap, and bootstrap-t (B-t) to estimate the confidence interval for the parameter (θ) of the ZTPS distribution. Due to a theoretical comparison is not possible, a simulation study was conducted to compare their performance and used the results to determine the best performing method based on the coverage probability, lower and upper error rate and the average length.

2. THEORETICAL BACKGROUND

Compounding probability distributions is a robust and innovative technique for obtaining new probability distributions that better fit datasets not adequately modeled by common parametric distributions. Shanker (2017) proposed a novel compounding distribution by combining the Poisson distribution with the Shanker distribution, aiming to find a more flexible model for analyzing statistical data. The pmf of the PS distribution is given in Eq.(2).

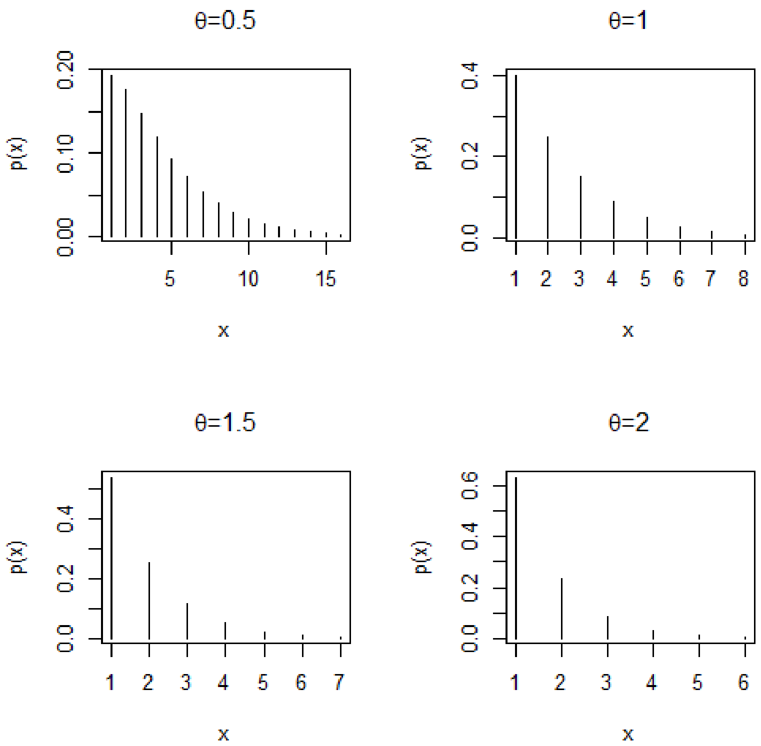

Figure 1 The plots of the pmf of the ZTPS distribution with =0.5, 1, 1.5, and 2.

Table 1 Empirical coverage probability and average length of the 95% bootstrap confidence intervals for the parameter of the ZTPS distribution

n	θ	Coverage probability				Average length			
		PB	SB	BCa	B-t	PB	SB	BCa	B-t
10	0.25	0.900	0.885	0.912	0.900	0.2795	0.2801	0.2552	0.2328
	0.5	0.882	0.879	0.881	0.867	0.7057	0.7049	0.6318	0.5155
	1	0.879	0.858	0.883	0.786	2.2385	2.2524	2.0565	1.2764
	1.5	0.920	0.846	0.928	0.750	3.9184	3.9368	3.9010	2.9430
	2	0.961	0.833	0.953	0.716	5.1095	5.1023	5.0118	4.7787
30	0.25	0.923	0.921	0.921	0.920	0.1453	0.1450	0.1400	0.1379
	0.5	0.939*	0.918	0.944*	0.933	0.3184	0.3183	0.3056	0.2979
	1	0.912	0.898	0.923	0.907	0.8177	0.8211	0.7607	0.7073
	1.5	0.926	0.903	0.933	0.909	1.6302	1.6218	1.4655	1.2649
	2	0.897	0.899	0.903	0.919	3.2230	3.2036	2.7655	2.2254
50	0.25	0.944*	0.929	0.940*	0.932	0.1112	0.1112	0.1089	0.1081
	0.5	0.938	0.933	0.940*	0.932	0.2438	0.2437	0.2377	0.2344
	1	0.928	0.918	0.931	0.923	0.5870	0.5875	0.5643	0.5429
	1.5	0.934	0.907	0.941*	0.922	1.1121	1.1118	1.0438	0.9744
	2	0.941*	0.908	0.934	0.924	1.8518	1.8520	1.7044	1.5345
100	0.25	0.948*	0.942*	0.943*	0.948*	0.0777	0.0776	0.0769	0.0767
	0.5	0.939*	0.940*	0.939*	0.939*	0.1699	0.1696	0.1677	0.1667
	1	0.936	0.943*	0.941*	0.941*	0.3998	0.3999	0.3922	0.3855
	1.5	0.960*	0.949*	0.958*	0.952*	0.7267	0.7271	0.7061	0.6844
	2	0.938	0.921	0.936	0.929	1.1647	1.1686	1.1229	1.0735
500	0.25	0.941*	0.947*	0.934	0.946*	0.0346	0.0345	0.0344	0.0344
	0.5	0.944*	0.943*	0.948*	0.945*	0.0747	0.0746	0.0745	0.0743
	1	0.960*	0.950*	0.955*	0.955*	0.1740	0.1742	0.1736	0.1733
	1.5	0.948*	0.959*	0.948*	0.953*	0.3109	0.3114	0.3089	0.3087
	2	0.949*	0.938	0.948*	0.949*	0.4705	0.4709	0.4659	0.4638

* represents the empirical coverage probability is greater than or equal to the nominal confidence level.

ZTPS distribution (Shanker, 2017) with parameter θ, it is denoted as $X \sim \text{ZTPS}(\theta)$. Using Eqs. (2) and (4), the pmf of the ZTPS distribution can be obtained as

$$p(x;\theta)=\frac{\theta^2}{\theta^3+\theta^2+2\theta+1}\frac{x+\left(\theta^2+\theta+1\right)}{(\theta+1)^x},\ x=1,2,3,...;\theta>0.$$

Figure 1 shows the graphs of the pmf of the ZTPS distribution with some specified parameter values θ. We noticed that the parameter values θ affect the form of Figure 1. When the parameter value is larger, the $p(1)$ is larger and $p(x)\to 0$ for the large value of x. The expected value and variance of X are as follows:

$$E(X)=\mu=\frac{(\theta+1)(\theta^3+\theta^2+2\theta+2)}{\theta\left(\theta^3+\theta^2+2\theta+1\right)},$$

and

$$Var(X)=\sigma^2=\frac{(\theta+1)(\theta^6+2\theta^5+6\theta^4+9\theta^3+10\theta^2+8\theta+2)}{\theta^2\left(\theta^3+\theta^2+2\theta+1\right)^2}.$$

The point estimator of θ is obtained by maximizing the log-likelihood function $\log L(x_i;\theta)$ or the logarithm of joint pmf of $X_1,...,X_n$. Therefore, the maximum likelihood (ML) estimator for θ of the ZTPS distribution is derived by the following processes:

$$\frac{\partial}{\partial\theta}\log L(x_i;\theta)=\frac{\partial}{\partial\theta}\left[n\log\left(\frac{\theta^2}{\theta^3+\theta^2+2\theta+1}\right)-\sum_{i=1}^{n}x_i\log(\theta+1)+\sum_{i=1}^{n}\log\left[x_i+\left(\theta^2+\theta+1\right)\right]\right]$$
$$=\frac{2n}{\theta}-\frac{n\left(3\theta^2+2\theta+2\right)}{\left(\theta^3+\theta^2+2\theta+1\right)}-\frac{n\bar{x}}{\theta+1}+\sum_{i=1}^{n}\frac{2\theta+1}{x_i+\left(\theta^2+\theta+1\right)}.$$

Solving the equation $\frac{\partial}{\partial\theta}\log L(x_i;\theta)=0$ for θ, we have the non-linear equation

$$\frac{2n}{\theta}-\frac{n\left(3\theta^2+2\theta+2\right)}{\left(\theta^3+\theta^2+2\theta+1\right)}-\frac{n\bar{x}}{\theta+1}+\sum_{i=1}^{n}\frac{2\theta+1}{x_i+\left(\theta^2+\theta+1\right)}=0,$$

where $\bar{x}=\sum_{i=1}^{n}x_i/n$ denotes the sample mean. Since the ML estimator for θ does not provide the closed-form solution, the non-linear equation can be solved by numerical iteration methods such as Newton-Raphson method, bisection method and Ragula-Falsi method. In this research, we use maxLik package (Henningsen and Toomet, 2011) with Newton-Raphson method for ML estimation in the statistical software R. The R code for estimating the parameter of the ZTPS distribution is shown in Appendix.

3. BOOTSTRAP METHODS

Confidence intervals are obtained from a parametric estimator of the standard errors of a quantity of interest ϕ. Then, the $(1-\alpha)100\%$ confidence interval for ϕ is obtained by adding or subtracting the standard error multiplied by a critical value (for example, $\hat{\phi}\pm z_{1-(\alpha/2)}SE(\hat{\phi})$). This calculation assumes that the distribution of the estimator of ϕ is approximately normal (Flowers-Cano et al., 2018). However, there are several situations in which the assumption of normality is violated. In these cases, or when the standard error is very difficult to estimate, one alternative is to use techniques based on the bootstrap method (Van den Boogaard and Hall, 2004). The computer-intensive bootstrap methods described in this study provide an alternative for constructing approximate confidence intervals without assuming the underlying distribution (Meeker, 2007). Furthermore, it is essential to note that the score function of ZTPS is complicated, and the maximum likelihood has no closed form. Therefore, likelihood-based, score, and Wald-type confidence intervals have no closed forms. In such cases, finding these confidence intervals can be challenging; alternative methods, such as numerical techniques or resampling methods like the bootstrap, can be utilized.

In this paper, we focus on the four bootstrap methods for estimating confidence interval for the parameter of the ZTPS distribution. In practice, the popular bootstrap methods are the percentile bootstrap, the simple bootstrap, the bias-corrected and accelerated bootstrap and bootstrap-t methods. See the details of some bootstrap methods in DiCiccio and Efron (1996) and Manoharan et al. (2017).

3.1 Percentile bootstrap (PB) method

The PB two-sided confidence interval is the range between the $(\alpha/2)\times 100$ and $(1-(\alpha/2))\times 100$ percentiles of the resampled distribution of θ estimations or the distribution of $\hat{\theta}^*$, where θ

Table 2 Lower and upper error rates of the 95% bootstrap confidence intervals for the parameter of the ZTPS distribution

n	θ	Lower error rates				Upper error rates			
		PB	SB	BCa	B-t	PB	SB	BCa	B-t
10	0.25	0.077	0.005	0.057	0.044	0.023	0.110	0.031	0.056
	0.5	0.101	0.002	0.085	0.079	0.017	0.119	0.034	0.054
	1	0.100	0.000	0.089	0.159	0.021	0.142	0.029	0.055
	1.5	0.060	0.000	0.040	0.180	0.020	0.154	0.032	0.070
	2	0.002	0.000	0.001	0.207	0.038	0.167	0.045	0.077
30	0.25	0.056	0.008	0.045	0.042	0.021	0.071	0.034	0.038
	0.5	0.038	0.004	0.023	0.027	0.023	0.078	0.033	0.040
	1	0.067	0.000	0.047	0.051	0.021	0.102	0.030	0.042
	1.5	0.057	0.000	0.042	0.054	0.017	0.097	0.025	0.037
	2	0.089	0.000	0.069	0.042	0.014	0.101	0.028	0.039
50	0.25	0.035	0.007	0.032	0.024	0.021	0.064	0.028	0.044
	0.5	0.041	0.002	0.030	0.031	0.021	0.065	0.030	0.037
	1	0.048	0.000	0.035	0.034	0.024	0.082	0.034	0.043
	1.5	0.049	0.000	0.033	0.044	0.017	0.093	0.026	0.034
	2	0.044	0.000	0.038	0.041	0.015	0.092	0.028	0.035
100	0.25	0.031	0.013	0.032	0.025	0.021	0.045	0.025	0.027
	0.5	0.039	0.010	0.030	0.029	0.022	0.050	0.031	0.032
	1	0.049	0.003	0.035	0.036	0.015	0.054	0.024	0.023
	1.5	0.030	0.000	0.022	0.026	0.010	0.051	0.020	0.022
	2	0.044	0.000	0.034	0.039	0.018	0.079	0.030	0.032
500	0.25	0.040	0.024	0.042	0.035	0.019	0.029	0.024	0.019
	0.5	0.030	0.019	0.027	0.028	0.026	0.038	0.025	0.027
	1	0.023	0.015	0.020	0.022	0.017	0.035	0.025	0.023
	1.5	0.042	0.017	0.040	0.036	0.010	0.024	0.012	0.011
	2	0.018	0.003	0.015	0.015	0.033	0.059	0.037	0.036

Table 3 The descriptive measures of the number of unrest events

Min	Mean	Median	Var	Skewness	Kurtosis	Q1	Q3	Max
1.000	6.714	6.000	17.915	0.546	2.183	3.750	9.500	15.000

Table 4 The the log-likelihood (Log L), AIC and BIC for the number of unrest events

Distribution	Log L	AIC	BIC
ZTPS	-77.328	**156.657**	**157.989**
ZTPL	-77.552	157.104	158.436
Poisson	-86.529	175.057	176.390

Table 5 The number of unrest events in the southern border area of Thailand

Number of unrest events	1-3	4-5	6-7	≥ 8
Observed frequency	7	6	7	8
Expected frequency	8.7702	5.3505	4.2138	9.6656

Table 6 The 95% confidence intervals and their corresponding lengths for the parameter in the unrest events example, using all intervals

Methods	Confidence intervals	Lengths
PB	(0.2463, 0.4023)	0.1560
SB	(0.2102, 0.3651)	0.1549
BCa	(0.2430, 0.3946)	0.1516
B-t	(0.2375, 0.3818)	0.1443

Table 7 The descriptive measures of the number of fertile mothers

Min	Mean	Median	Var	Skewness	Kurtosis	Q1	Q3	Max
1.000	1.607	1.000	1.091	1.941	6.446	1.000	2.000	6.000

significance (e.g., α = 0.05 for 95% confidence intervals) (Efron, 1982). Obtaining a PB confidence interval for θ is as follows:

1) With a replacement, B random bootstrap samples of the underlying distribution are created, where B is the number of bootstrap replications,

2) From each bootstrap sample, a parameter estimate $\hat{\theta}^*$ is determined,

3) All parameter estimates from the B bootstrap are ranked from lowest to highest, and

4) The $(1-\alpha)100\%$ PB two-sided confidence interval for θ is created as follows:

$$CI_{PB} = \left[\hat{\theta}^*_{(r)}, \hat{\theta}^*_{(s)}\right], \quad (5)$$

where $\hat{\theta}^*_{(r)}$ is the r^{th} quantile of a set of quantiles ordered from lowest to highest, $\hat{\theta}^*_{(s)}$ is the s^{th} quantile of the same set, $r = \left[(\alpha/2)B\right]$, $s = \left[(1-(\alpha/2))B\right]$, and $1-\alpha$ is the confidence level. This study utilized B = 1,000 and α = 0.05; the quantile corresponding to the lower limit of the confidence interval was $\hat{\theta}^*_{(r)} = \hat{\theta}^*_{(25)}$, and the quantile corresponding to the upper limit was $\hat{\theta}^*_{(s)} = \hat{\theta}^*_{(975)}$.

3.2 Simple bootstrap (SB) method

The SB method is as straightforward to implement as the PB method and is sometimes referred to as the simple bootstrap method. Consider the quantity of interest to be θ and the estimator of θ to be $\hat{\theta}$. The straightforward bootstrap method implies that the distributions of $\hat{\theta}-\theta$ and $\hat{\theta}^*-\hat{\theta}$ are roughly equivalent (Meeker et al., 2017). The $(1-\alpha)100\%$ SB two-sided confidence interval for θ is

$$CI_{SB} = \left[2\hat{\theta} - \hat{\theta}^*_{(s)}, 2\hat{\theta} - \hat{\theta}^*_{(r)}\right], \quad (6)$$

where the quantiles $\hat{\theta}^*_{(r)}$ and $\hat{\theta}^*_{(s)}$ represent the same percentile of the empirical distribution of bootstrap estimates $\hat{\theta}^*$ that are utilized in Eq. (5) to calculate the PB confidence interval.

3.3 Bias-corrected and accelerated (BCa) bootstrap method

The calculation of BCa bootstrap confidence intervals commonly involves the utilization of influence statistics derived from jackknife simulations. However, incorporating jackknife simulation alongside ordinary bootstrapping is computationally expensive for the intended purposes. The BCa bootstrap confidence interval uses a bias- correction factor and an acceleration factor to correct for the bias and skewness of the bootstrap parameter estimates, mitigating the over- coverage problems seen with the PB confidence interval (Efron, 1987; Efron and Tibshirani, 1993). Davison and Hinkley (1997) and Chernick and LaBudde (2011) described the mathematical particulars of the BCa adjustment. The bias-correction factor $\hat{z}_0$ calculated as the proportion of bootstrap estimates that are smaller than the original parameter estimate $\hat{\theta}$,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^* \le \hat{\theta}\}}{B}\right),$$

where Φ^{-1} is the inverse function of a standard normal cumulative distribution function (e. g. , $\Phi^{-1}(0.975) \approx 1.96$). The acceleration factor $\hat{a}$ is estimated through jackknife resampling, also known as 'leave- one- out' resampling. This technique involves generating n replicates of the original sample, where n represents the number of observations in the sample. The initial jackknife replicate is derived by excluding the first case $(i=1)$ from the original sample. Subsequently, the second replicate is obtained by excluding the second case $(i=2)$, and this process continues until n samples of size $n-1$ are generated. The value of $\hat{\theta}_{(-i)}$ is obtained for each of the jackknife resamples. The mean of these estimates is denoted by $\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(-i)} / n$. The acceleration factor $\hat{a}$ is then determined as follows:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)}\right)^2\right\}^{3/2}}.$$

With the values of $\hat{z}_0$ and $\hat{a}$, the values α_1 and α_2 are calculated,

Table 8 The the log-likelihood (Log L), AIC and BIC for the number of fertile mothers

Distribution	Log L	AIC	BIC
ZTPS	-143.976	**289.951**	**292.856**
ZTPL	-144.091	290.182	293.087
Poisson	-191.055	384.110	387.015

Table 9 The number of fertile mothers who have experienced at least one child death

Number of child deaths	1	2	3	≥ 4
Observed frequency	89	25	11	10
Expected frequency	83.4756	32.3839	12.2451	6.8953

Table 10 The 95% confidence intervals and their corresponding lengths for the parameter in the demographic example, using all intervals

Methods	Confidence intervals	Lengths
PB	(1.5810, 2.5710)	0.9900
SB	(1.3739, 2.3301)	0.9562
BCa	(1.5567, 2.4995)	0.9428
B-t	(1.5527, 2.4537)	0.9010

Table 11 The descriptive measures of the number of flower heads as per the number of fly eggs

Min	Mean	Median	Var	Skewness	Kurtosis	Q1	Q3	Max
1.000	3.034	3.000	3.344	0.812	3.135	1.750	4.000	9.000

Table 12 The the log-likelihood (Log L), AIC and BIC for the number of flower heads as per the number of fly eggs

Distribution	Log L	AIC	BIC
ZTPS	-167.154	**336.308**	**338.785**
ZTPL	-167.381	336.763	339.240
Poisson	-171.281	344.561	347.039

Table 13 The number of flower heads as per the number of fly eggs

Number of fly eggs	1	2	3	4	5	6	≥ 7
Observed frequency	22	18	18	11	9	6	4
Expected frequency	24.9287	19.7204	14.6526	10.2922	6.9078	4.4711	7.0272

Table 14 The 95% confidence intervals and corresponding lengths using all methods for the parameter in the flower heads example

Methods	Confidence intervals	Lengths
PB	(0.6342, 0.8684)	0.2342
SB	(0.6045, 0.8312)	0.2267
BCa	(0.6323, 0.8597)	0.2274
B-t	(0.6253, 0.8525)	0.2272

$$\alpha_1 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1-\hat{a}\left(\hat{z}_0 + z_{\alpha/2}\right)}\right\} \quad \text{and}$$

$$\alpha_2 = \Phi\left\{\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1-\hat{a}\left(\hat{z}_0 + z_{1-\alpha/2}\right)}\right\},$$

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution (e.g. $z_{0.05/2} \approx -1.96$). Then, the $(1-\alpha)100\%$ BCa bootstrap two-sided confidence interval for θ is as follows

$$CI_{BCa} = \left[\hat{\theta}^*_{(j)}, \hat{\theta}^*_{(k)}\right], \tag{7}$$

where $j = [\alpha_1 B]$ and $k = [\alpha_2 B]$. When the value $\hat{z}_0$ is set to 0 and $\hat{a}$ is also set to 0, it can be observed that the BCa confidence interval is equal to the PB confidence interval.

3.4 Bootstrap-t (B-t) method

Assuming the quantity of interest is θ, one can compute the estimate $\hat{\theta}$ and its corresponding estimate of the standard error $s.e.(\hat{\theta})$ from the given data. The bootstrap estimates $\hat{\theta}^*_j$ and their estimated standard errors $s.e.(\hat{\theta}^*_j)$ are then computed for each bootstrap sample, $j = 1,2,...,B.$ The bootstrap-t (studentized) statistics are derived from these values;

$$R_j^* = \frac{\hat{\theta}_j^* - \hat{\theta}}{s.e.(\hat{\theta}_j^*)}, \quad j = 1,2,...,B.$$

The $(1-\alpha)100\%$ bootstrap-t two-sided confidence interval for θ is

$$CI_{B-t} = \left[\hat{\theta} - t_r^* \times s.e.(\hat{\theta}), \hat{\theta} - t_s^* \times s.e.(\hat{\theta})\right], \quad (8)$$

where $r = 1-(\alpha/2)$ and $s = \alpha/2$, and t_q^* denotes the q quantile of the distribution of R_j^*.

4. SIMULATION STUDY AND RESULTS

The confidence interval for the parameter of a ZTPS distribution estimated via various bootstrap methods was considered in this study. Due to the absence of a theoretical comparison, a Monte Carlo simulation study was conducted using R version 4.2.2 (Ihaka and Gentleman, 1996). The random samples from the ZTPS distribution were generated using the accept-reject method (Robert and Casella, 2004). The simulation covered different sample sizes (n = 10, 30, 50, 100, and 500) to observe the effects of small and large variances. To observe the effect of small and large variances, the true parameter (θ) was set as 0.25, 0.5, 1, 1.5, and 2 (the variance of the random variables decreases as the value of θ increases). For each simulation, 1,000 bootstrap samples of size n were generated from the original sample, and the simulation was repeated 1,000 times. To evaluate the performance of the bootstrap methods, the empirical coverage probabilities, lower and upper error rates, and average lengths were compared. A bootstrap method with a coverage probability close to or greater than the nominal confidence level $(1-\alpha)$ of 0.95 indicates that it contains the true value and provides a precise estimate of the confidence interval for the parameter. In this study, we can conclude that a coverage probability is greater than or equal to the nominal confidence level when the empirical coverage probability is greater than or equal to 0.939 by using the one-proportion z-test with a significance level of 0.95. Moreover, the shortest average length can be used to more accurately estimate the bootstrap confidence interval for the parameter.

The empirical coverage probability and average length from the simulation study are presented in Table 1. For n=10, 30 and 50, the empirical coverage probability of the four approaches had a tendency to go below 0.95 and did not attain the expected nominal confidence level. However, in these scenarios, the BCa bootstrap method provided higher empirical coverage probabilities than the other methods. For $n \geq 100$, all four methods performed comparably well in terms of empirical coverage probability, and they provided empirical coverage probabilities greater than 0.95 in almost all of situations. In this case ($n \geq 100$), the BCa bootstrap method and B-t method produced shorter lengths on average than the PB and SB methods. Therefore, when the size of the sample was enlarged, the empirical coverage probabilities of the bootstrap methods exhibited a tendency to rise and converge towards 0.95. Furthermore, the average lengths of the bootstrap methods exhibited a reduction when the value of θ declined, reflecting the relationship between the variance and θ. As expected, the average lengths of the four approaches dropped with an increase in the sample size. While the B-t method demonstrated the lowest average length for small sample sets, it exhibited an empirical coverage probability that was notably lower than the nominal confidence level. In brief, the BCa bootstrap method demonstrated persistent superior performance in terms of empirical coverage probability across many scenarios.

Table 2 shows the lower and upper error rates of the bootstrap confidence intervals. The lower error rates of the PB and BCa bootstrap methods exceeded the upper error rates across all scenarios. On the other hand, in every situation, the lower error rates of the SB method are lower than the upper error rates. Furthermore, there is no substantial difference observed between the lower and upper error rates in the BCa bootstrap method.

5. NUMERICAL EXAMPLES

We used three real-world examples to demonstrate the applicability of the bootstrap methods for estimating the confidence interval for the parameter of the ZTPS distribution.

5.1 The Unrest Events Example

The number of unrest events occurring in the southern border area of Thailand from July 2020 to October 2022 collected by the Southern Border Area News Summarises (http://summarise.wbns.oas.psu.ac.th) was used for this example (the total sample size is 28). The number of unrest events per month during this time period in the five southern provinces of Pattani, Yala, Narathiwat, Songkhla, and Satun is as follows: 1, 1, 1, 2, 3, 3, 3, 4, 4, 5, 5, 5, 5, 6, 6, 6, 7, 7, 7, 7, 9, 11, 12, 12, 13, 14, 14, 15. Table 3 provides some descriptive measures of the number of unrest events.

The fitting performance of the ZTPS distribution was compared to the Poisson and the ZTPL (Shanker et al., 2015) distributions. The parameters of all distributions were estimated using the ML approach. For distribution comparison, we have considered the log-likelihood (Log L), Akaike's information criterion (AIC) (Akaike, 1973), and Bayesian information criterion (BIC) (Schwarz, 1978) presented in Table 4. The AIC and BIC values illustrate that the ZTPS distribution provides an adequate fit to the number of unrest events as compared with other distributions.

For the Chi-squared goodness-of-fit test (Turhan, 2020) for the ZTPS distribution in Table 5, the Chi-squared statistic was 2.5655 and the p-value was 0.2772. Thus, a ZTPS distribution with ML estimator $\hat{\theta}$ = 0.3058 and $\hat{\mu}$ = 6.7331 is suitable for this dataset. Table 6 reported the 95% confidence intervals for the parameter of the ZTPS distribution. The results correspond with the simulation results for n = 30 and θ = 0.5 because the average lengths of the BCa bootstrap method and B-t method were shorter than those of the PB and SB methods.

5.2 Demographic Example

The 135 demographic data points on the number of fertile mothers who have experienced at least one child death (Shanker et al., 2015) are as follows: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 5, 5, 5, 6. Some descriptive measures of the number of fertile mothers are reported in Table 7.

A comparison was conducted between the fitting performance of the ZTPS distribution and the Poisson and ZTPL (Shanker et al., 2015) distributions. Table 8 shows the log-likelihood (Log L), Akaike′s information criterion (AIC) (Akaike, 1973), and Bayesian information criterion (BIC) (Schwarz, 1978) that we used to compare the distributions. The adequacy of the ZTPS distribution in fitting the number of fertile mothers is supported by the AIC and BIC values, which indicate its comparative performance against other distributions.

The Chi-squared goodness-of-fit test (Turhan, 2020) was conducted to assess the fit of the ZTPS distribution shown in Table 9. The obtained Chi-squared statistic was 3.1070, and the corresponding p-value was 0.2115. Thus, a ZTPS distribution with ML estimator $\hat{\theta}$ = 1.9538 and $\hat{\mu}$ = 1.6052 is suitable for this dataset. The 95% confidence intervals for the parameter of the ZTPS distribution are reported in Table 10. The results correspond with the simulation results for n = 100 and θ = 2 because the average lengths of the BCa bootstrap method and B-t method were shorter than those of the PB and SB methods.

5.3 Flower Heads Example

The third dataset consists of 88 observations, which is the number of flower heads as per the number of fly eggs reported by Finney and Varley (1955). The data are as follows: 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 5, 5, 6, 6, 6, 6, 6, 6, 7, 7, 7, 9. Some descriptive measures of the number of flower heads as per the number of fly eggs are reported in Table 11.

The ability of the ZTPS distribution to fit data was compared to the Poisson and ZTPL distributions (Shanker et al., 2015). In order to compare distributions, we have taken into account the log-likelihood (Log L), Akaike's information criterion (AIC) (Akaike, 1973), and Bayesian information criterion (BIC) (Schwarz, 1978) as provided in Table 12. The AIC and BIC values indicate that the ZTPS distribution fits the number of flower heads as per the number of fly eggs better than other distributions.

For the Chi-squared goodness-of-fit test (Turhan, 2020) with the ZTPS distribution in Table 13, the Chi-squared statistic was 3.7681 and p-value was 0.5832. Thus, a ZTPS distribution with ML estimator $\hat{\theta}$ = 0.7334 and $\hat{\mu}$ = 3.0588 is suitable for this dataset. The 95% confidence intervals for the parameter of the ZTPS distribution are reported in Table 14. The results correspond with the simulation results for n = 100 and θ = 1 because the average lengths of all bootstrap methods are not significantly different.

5. CONCLUSIONS AND DISCUSSION

We present four bootstrap methods, PB, SB, BCa, and B-t, to estimate the confidence interval for the ZTPS distribution parameter. The coverage probabilities of all bootstrap confidence intervals were markedly lower than 0.95 when the sample size was 10, 30, or 50. However, when the sample size was sufficiently large ($n \geq 100$), the coverage probabilities and average lengths derived from four bootstrap methods did not differ significantly. Based on our findings, the BCa bootstrap method consistently outperformed the others in terms of the empirical coverage probability and average length for almost all situations, as confirmed by both the simulation study and the analysis of real datasets.

The bootstrap method resamples using replacement from observed data to estimate the sampling distribution of a statistic. However, there are certain limitations to using the bootstrap method, especially when dealing with small sample sizes. With a small sample size, we have limited information about the underlying population. This means that resampling from the small dataset may not adequately capture the variability present in the population, leading to biased or unreliable estimates. Additionally, the coverage probability of bootstrap confidence intervals might deviate from the desired level. Bootstrap confidence intervals may end up too narrow or too wide, leading to incorrect inferences about the parameter. On the other hand, the bootstrap method becomes more effective at approximating the sampling distribution of a statistic as the sample size increases. This is because with a large sample, the bootstrap resamples from the original data, capturing the underlying distribution of the population better.

Future research could focus on the other interval estimations to compare with the proposed bootstrap confidence intervals. The constructions of the confidence intervals for the functions of the parameter such as the population mean and the dispersion index are interest. Moreover, there is no research on the hypothesis testing for the parameter of the ZTPS distribution. The bootstrap confidence intervals studied in this paper can be applied for other distribution. These issues can be studied in the future.

ACKNOWLEDGEMENTS

The authors would like to express their gratitude for the reviewers' dedication and commitment in enhancing the overall quality of this manuscript.

REFERENCES

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & B. F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Academiai Kiado: Budapest.

Chernick, M. R., & LaBudde, R. A. (2011). *An Introduction to Bootstrap Methods with Applications to R*. New York: John Wiley & Sons.

David, F. N., & Johnson, N. L. (1952). The truncated Poisson. *Biometrics*, *8*(4), 275-285.

Davison, A. C., & Hinkley, D.V. (1997). *Bootstrap Methods and Their Application*. Cambridge: Cambridge University Press.

DiCiccio, T. J., & Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, *11*(3), 189-212.

Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. *CBMS-NSF Regional Conference Series in Applied Mathematics*, Monograph 38. SIAM.

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171-185.

Efron, B., & Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman and Hall.

Finney, D. J., & Varley, G. C. (1955). An example of the truncated Poisson distribution. *Biometrics*, *11*(3), 387-394.

Flowers-Cano, R. S., Ortiz-Gómez, R., León-Jiménez, J. E., López Rivera, R., & Perera Cruz, L. A. (2018). Comparison of bootstrap confidence intervals using Monte Carlo simulations. *Water*, *10*(2), 166. https://doi.org/10.3390/w10020166

Ghitany, M. E., Al-Mutairi, D. K., & Nadarajah, S. (2008). Zero-truncated Poisson-Lindley distribution and its application. *Mathematics and Computers in Simulation*, *79*(3), 279-287.

Henningsen, A., & Toomet, O. (2011). A package for maximum likelihood estimation in R. *Computational Statistics*, *26*(3), 443-458.

Hussain, T. (2020). A zero truncated discrete distribution: Theory and applications to count data. *Pakistan Journal of Statistics and Operation Research*, *16*(1), 167-190.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, *5*(3), 299-314.

Kissell, R., & Poserina, J. (2017). *Optimal sports math, statistics, and fantasy*. London: Academic Press.

Lindley, D. V. (1958). Fiducial distributions and Bayes' theorem. *Journal of the Royal Statistical Society: Series B*, *20*(1), 102-107.

Meeker, W. Q., Hahn, G. J., & Escobar, L. A. (2017). *Statistical intervals: A guide for practitioners and researchers*. New Jersey: John Wiley and Sons.

Manoharan, T., Arasan, J. Midi, H., & Adam, M. B. (2017). Bootstrap intervals in the presence of left-truncation, censoring and covariates with a parametric distribution. *Sains Malaysiana*, *46*(12), 2529-2539.

Robert, C. P., & Casella, G. (2004). *Monte Carlo Statistical Methods*. New York: Springer.

Ross, S. (2006). *A First Course in Probability*. New Jersey: Pearson Education.

Sankaran, M. (1970). The discrete Poisson-Lindley distribution. *Biometrics*, *26*(1), 145-149.

Sangnawakij, P. (2021). Confidence interval for the parameter of the zero-truncated Poisson distribution. *The Journal of Applied Science*, *20*(2), 13-22.

Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461-464.

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). A zero-truncated Poisson-Shanker distribution and its applications. *International Journal of Statistics and Applications*, *7*(3), 159-169.

Shanker, R., Fesshaye, H., Sujatha, S., & Yemane, A. (2015). On zero-truncation of Poisson and Poisson-Lindley distributions and their applications. *Biometrics and Biostatistics International Journal*, *2*(6), 168-181.

Siegel, A. F., & Wanger, M. R. (2022). *Practical Business Statistics*. San Diego: Academic Press.

Turhan, N. S. (2020). Karl Pearson's chi-square tests. *Educational Research Review*, *15*(9), 575-580.

Van den Boogaard, H. F. P., & Hall, M. J. (2004). The construction of confidence intervals for frequency analysis using resampling techniques: a supplementary note. *Hydrology and Earth System Sciences*, *8*(6), 1174-1178.

Wood, M. (2004). Statistical inference using bootstrap confidence intervals. *Significance*, *1*(4), 180-182.

APPENDIX

```
library(maxLik)
theta.hat <- function(x) {
options(warn=-1)
logLikFun <- function(param) {
th <- param
n <- length(x)
term1 <- n*log((th^2)/(th^3+th^2+2*th+1))
term2 <- sum(x)*log(th+1)
term3 <- sum(log(x+(th^2+th+1)))
return(term1-term2+term3)
}
mle <- maxLik(logLik=logLikFun,start=0.01)
return(c(mle$estimate,vcov(mle)))
}
```