Journal of Current Science and Technology, April - June 2025 Vol. 15 No. 2, Article 102

ISSN 2630-0656 (Online)

Cite this article: Sangnawakij, P. (2025). Parametric and nonparametric estimation of population mean in poisson-xgamma distribution with applications to count data. *Journal of Current Science and Technology*, *15*(2), Article 102. https://doi.org/10.59796/jcst.V15N2.2025.102

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Parametric and Nonparametric Estimation of Population Mean in Poisson-Xgamma Distribution with Applications to Count Data

Patarawan Sangnawakij*

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 12120, Thailand

*Corresponding author; E-mail: patarawan@mathstat.sci.tu.ac.th

Received 24 October 2024; Revised 8 December 2024; Accepted 9 December 2024: Published online 25 March 2025

Abstract

This study proposes new estimators and confidence intervals for the population mean of the Poisson-Xgamma distribution, which are useful for overdispersed count data analysis. We prove that the proposed estimators using maximum likelihood and method of moments estimation are consistent and establish the variance of the estimators. Moreover, the confidence intervals are constructed based on large-sample theory and bootstrap method. The former method utilizes the properties of the maximum likelihood and moment estimators, the likelihood ratio, and the asymptotic normality property of the log-transformed maximum likelihood estimator. Percentile bootstrap and bias-corrected and accelerated confidence intervals are considered. The performance of the estimators is investigated through simulations in terms of bias, mean squared error, coverage probability, and length of interval. According to the simulations, the log-transformed maximum likelihood estimation-based confidence interval for the mean provides excellent and better coverage rates than the other competitive methods. Furthermore, two real data sets are used to demonstrate our estimators and perform a comparison that supports the findings obtained from the simulation study.

Keywords: *asymptotic theory; bootstrap; compounding discrete distribution; count data; overdispersion*

1. Introduction

Statistical analysis of count data plays a crucial role in many areas of social sciences, biometrics, medical sciences, and healthcare service applications. More specifically, in the context of count data, they consist of discrete, non-negative counts and typically a right-skewed distribution. These data indicate the number of events within a fixed interval, time period, or space (Hilbe, 2017; Tang et al., 2023). Under this assumption, the Poisson distribution is a commonly used probability model because of its simple form and tractable properties. Suppose that X is a Poisson random variable. The probability mass function (pmf) of X is given by

$$P(X=x;\lambda)=\frac{e^{-\lambda}\lambda^{x}}{x!} \tag{1}$$

where $x=0,1,...$ are the observed values of X and $\lambda>0$ is the average number of occurrences in the interval time. The expected value and variance of X are $E(X)=\lambda$ and $Var(X)=\lambda$, respectively. The Poisson model assumes equidispersion because the mean and variance are equal (Oseni et al., 2023; Weems et al., 2023). However, this situation is rarely reflected in practical and real data. Hence, the Poisson distribution cannot accurately represent data that is either overdispersed, where the variance exceeds the mean, or underdispersed, where the variance falls below the mean (Gschlößl, & Czado, 2008; Huang, 2023). Nevertheless, overdispersion is a common situation in count data. Therefore, various mixed-Poisson distributions have been introduced in the statistical literature to deal with this limitation. They included Poisson-gamma, Poisson-Lindley, Poisson-Ishita, Poisson-Shanker, Poisson-Amarendra, Conway-Maxwell- Poisson, uniform-Poisson, and Poisson Quasi-XLindley (Alghamdi et al., 2024; Conway, & Maxwell, 1962; Déniz, 2013; Hilbe, 2011; Imoto et al., 2017; Sankaran, 1970; Shanker, 2016; Shanker et al., 2017; Shukla, & Shanker, 2019). These are some examples of the references.

Parameter estimation is a statistical inference used to estimate the value of an unknown parameter in a probability model. It consists of two types: point estimation and interval estimation. Point estimation uses a single value to estimate an unknown parameter based on the sample data, while interval estimation calculates a range of plausible values using confidence intervals with a given probability (Casella, & Berger, 2002; Sangnawakij, & Anlamlert, 2023). So, interval estimation can be used to reveal the precision of parameter estimate. In this paper, we are interested in estimating the parameter of the Poisson-Xgamma (PX) distribution, which was introduced by Altun et al. (2022). The PX model compounds the Poisson and Xgamma distributions (Sen et al., 2016) to allow the modeling of the overdispersed count data. Moreover, it is a flexible and sufficient distribution for modeling the overdispersed time series of counts. The pmf of the PX variable with unknown parameter θ is given as

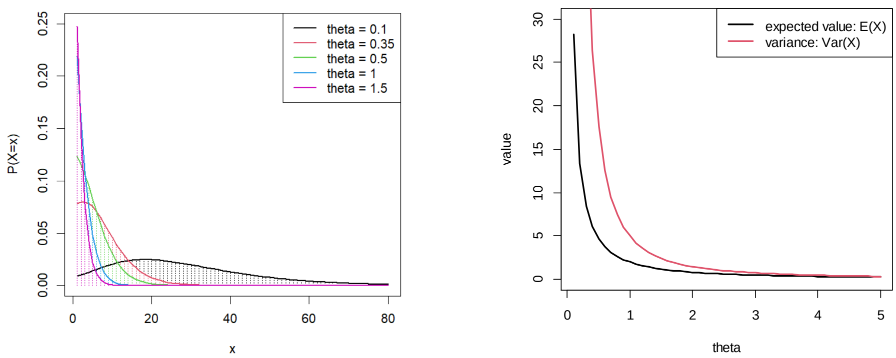
$$P(X=x;\theta)=\frac{\theta^2\left[2(1+\theta)^2+\theta(x+2)(x+1)\right]}{2(1+\theta)^{x+4}}, \qquad (2)$$

where $x=0,1,\ldots$ and $\theta>0$. The shapes of PX distribution varying by θ are displayed in Figure 1(a). When θ is increased, the PX distribution has a long tail towards the right side, say a positive-skewed distribution. Meanwhile, the distribution tends to symmetry when θ is greater. Hence, the PX distribution can occur in right-skewed and symmetric data sets. Under model (2), the mean and variance of X are given by

$$E(X)=\frac{\theta+3}{\theta(1+\theta)}, \text{ and}$$
$$Var(X)=\frac{\theta^3+5\theta^2+11\theta+3}{\theta^2(1+\theta)^2},$$

respectively. The possible values of mean and variance are positive real values. The following evidence supports the overdispersion of the PX distribution. We generate the PX data to investigate the behaviour of $E(X)$ and $Var(X)$. Then, we plot the graphs between these values for several possible values of θ. They are displayed in Figure 1(b). It is clear that the mean and variance decrease when θ is increased. The variance is greater than the mean for all possible values of θ. If θ is small, the difference between the mean and variance has an impact. Hence, the PX distribution has utility for the overdispersed data.

In statistical inference, the method of moment and maximum likelihood estimators for θ were introduced in Altun et al., (2022). They evaluated the performance of these estimators in terms of bias and mean squared error using simulations. The estimators for estimating θ from those two methods performed satisfactorily, as they provided a small bias and mean squared error. However, θ represents the shape of the distribution only. It is rarely used as a stand-alone parameter for describing the interesting characteristics of a population in applied research, like the mean, total, and variance.



(a) (b)

Figure 1 (a) density plot of Poisson-Xgamma distributions and (b) plot of expected value and variance with varying parameter values of θ

The key points of this current work are highlighted. The measure of central tendency is essential to determining the midpoint of the quantitative data, including discrete and continuous variables. Unfortunately, no research has introduced an estimator and a confidence interval for the population mean of the PX distribution. In this paper, we propose two simple estimators to estimate the population mean. These are constructed based on the moment method and maximum likelihood estimation. Moreover, the mean estimate from nonparametric bootstrap method is introduced. The statistical properties of the mean estimators are then studied comprehensively. For interval estimation, the large-sample approximation, asymptotic normal approximation, and bootstrap methods are applied to construct the confidence intervals for the population mean. We conduct a simulation study to evaluate the performance of the proposed estimators. Moreover, our methods are applied to two overdispersed real data sets on the number of chromosome aberrations in human genetic disease and the number of victims of unrest events in the southern region of Thailand.

The remaining parts of the paper are organized as follows. Section 2 proposes the new estimator and its statistical properties. We also derive the confidence intervals for the population mean. In Section 3, the performance of the proposed methods is investigated via the simulation study under several situations. In Section 4, the two real data sets are analyzed to demonstrate the usefulness of the proposed estimators and confidence intervals. Section 5 provides some concluding remarks.

2. Statistical Methodology

The Poisson distribution is a basic probability model in count data analysis. If X follows a Poisson random variable with parameter λ, namely Pois(λ), the pmf of X is shown in (1). The dispersion index of X is Var(X)/E(X)= 1. Hence, the Poisson model assumes equidispersion. Here, we are interested in the Poisson-Xgamma (PX) distribution. A basic idea for constructing the model is briefly introduced at the beginning of this section. Then, the proposed estimators for the mean in the PX distribution are intensively discussed.

The Xgamma distribution is derived by Sen et al. (2016). It is a probability model for a continuous variable, generated by a special finite mixture of exponential and gamma distributions. Let X be a random variable from an Xgamma distribution with parameter θ. The probability density function can be expressed as

$$f(x;\theta)=\frac{\theta^2}{1+\theta}\left(1+\frac{\theta x^2}{2}\right)e^{-\theta x}, \quad (3)$$

where $x>0$ and $\theta>0$. Also, we can rewrite it as

$$f(x;\theta)=\frac{\theta}{1+\theta}\theta e^{-\theta x}+\left(1-\frac{\theta}{1+\theta}\right)\frac{\theta^3x^2e^{-\theta x}}{2} \quad (4)$$

this is the two-component mixture of the exponential distribution with rate parameter θ, namely $\text{Exp}(\theta)$, and the gamma distribution with shape parameter 3 and scale parameter $1/\theta$, denoted as $\text{Gam}(3,1/\theta)$. Consequently, their weights are $\theta/(1+\theta)$ and $1/(1+\theta)$, respectively. To construct the probability model of the PX distribution, λ is assumed to be a random variable following the Xgamma distribution with an unknown parameter θ. Therefore, the unconditional probability distribution of the Poisson variable given λ is defined by

$$P(X=x;\theta)=\int_0^\infty P(X=x|\lambda)f(\lambda;\theta)d\lambda=\int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!}\left(\frac{\theta^2}{1+\theta}\right)\left(1+\frac{\theta\lambda^2}{2}\right)e^{-\theta\lambda}d\lambda$$

$$P(X=x;\theta)=\frac{\theta^2\left[2(1+\theta)^2+\theta(x+2)(x+1)\right]}{2(1+\theta)^{x+4}},$$

where $x=0, 1, \ldots$. The above equation is the pmf of the PX distribution as given in (2). The mean of PX variable can be determined from the factorial moment. It is established as

$$E(X)=\frac{\theta+3}{\theta(1+\theta)}=\phi \quad (5)$$

this is the parameter of interest to be estimated in this paper. In the following subsections, the new point estimators for ϕ are derived, and their properties are discussed. Moreover, the bootstrap estimates and confidence intervals will be introduced.

2.1 Point Estimation for the Population Mean

Maximum likelihood estimation is a statistical method usually used to estimate the unknown parameter in a probability distribution. The main concept of this method is to find the values of the parameter that maximize the likelihood of observed data.

Here, we assume tha $X=(X_1,X_2,...,X_n)t$ is a random sample of size n from a PX distribution. It is denoted as X~PX(θ). The likelihood function of θ given $X_i=x_i$ is given by

$$L(\theta)=\frac{\theta^{2n}}{2^n(1+\theta)^{\sum_{i=1}^n x_i+4n}}\prod_{i=1}^n\left[2(1+\theta)^2+\theta(x_i+2)(x_i+1)\right] \quad (6)$$

then, we maximize the log-likelihood function, specifically $\log L(\theta)$, with respect to θ and set it equal to zero. This gives the normal equation for estimating θ:

$$\theta=\frac{2n}{\frac{1}{\theta+1}(\sum_{i=1}^{n}x_i+4n)-\sum_{i=1}^{n}\frac{4(1+\theta)+(x_i+2)(x_i+1)}{2(1+\theta)^2+\theta(x_i+2)(x_i+1)}} \tag{7}$$

since there is no closed-form solution for the ML estimator for θ, a numerical method is applied. For example, the fixed point method based on equation (7) can be used to derive the ML estimate. Moreover, the *optim* function in the *R* package (R Core Team, 2024) can be applied. We denote the ML estimate for θ as $\hat{\theta}_{ML}$. Substituting $\hat{\theta}_{ML}$ into its parameter of (5), the ML estimator of ϕ is therefore given by

$$\hat{\phi}_{ML}=\frac{\hat{\theta}_{ML}+3}{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})} \tag{8}$$

next, we consider the mean and variance of $\hat{\theta}_{ML}$ (see Theorem 1). They are derived using the delta method, which is based on the Taylor series expansion (Oehlert, 1992). The benefit of this method is that it can be used to find the asymptotic mean and variance of a complex function, such as a ratio of variables, and construct the pivot function, which is useful for statistical inference.

Theorem 1. Let $X=(X_1,X_2,...,X_n)$ be a random sample of size n from a PX distribution. The ML estimator of ϕ is $\hat{\phi}_{ML}$. The expected value and variance of $\hat{\phi}_{ML}$ are approximated by $E(\hat{\phi}_{ML})=\phi$ and

$$Var(\hat{\phi}_{ML})=\left[\frac{\theta(1+\theta)-(\theta+3)(1+2\theta)}{\theta^2(1+\theta)^2}\right]^2\frac{1}{I(\theta)},$$

respectively, where $I(\Theta)$ is the observed Fisher information of θ.

Proof. According to the property of the ML estimator for large n, $\hat{\theta}_{ML}$ has an asymptotic normal distribution with mean $E(\hat{\theta}_{ML})=\theta$ and variance $Var(\hat{\theta}_{ML})=1/I(\theta)$. The observed Fisher information $I(\Theta)$ is obtained from the negative of the second derivative of $\log L(\theta)$. So, we have

$$I(\theta)=-\frac{\partial^2}{\partial\theta^2}\log L(\theta)$$
$$I(\theta)=\frac{2n}{\theta^2}-\frac{\sum_{i=1}^{n}x_i+4n}{(1+\theta)^2}-\frac{\sum_{i=1}^{n}\left[4[2(1+\theta)^2+\theta(x_i+2)(x_i+1)]-[4(1+\theta)+(x_i+2)(x_i+1)]^2\right]}{\left[2(1+\theta)^2+\theta(x_i+2)(x_i+1)\right]^2}.$$

consequently, $Var(\hat{\theta}_{ML})=1/I(\theta)$, for $n\to\infty$. Suppose that Y is a generic random variable, $E(Y)=\theta$ and $Var(Y)$ exists. Let $g(Y)$ be a function of Y. By the delta method, the variance of $g(Y)$ is given as

$$Var(g(Y))\approx\left[\frac{\partial}{\partial Y}g(Y)\right]^2_{E(Y)=\theta}Var(Y).$$

in our case, $\hat{\theta}_{ML}$ is the function of $\hat{\theta}_{ML}$, and its mean and variance exist. So, we can find that

$$Var(\hat{\phi}_{ML})=Var\left(\frac{\hat{\theta}_{ML}+3}{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})}\right)$$
$$Var(\hat{\phi}_{ML})\approx\left[\frac{\partial}{\partial\hat{\theta}_{ML}}\left(\frac{\hat{\theta}_{ML}+3}{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})}\right)\right]^2_{E(\hat{\theta}_{ML})=\theta}$$
$$Var(\hat{\theta}_{ML})=\left[\frac{\theta(1+\theta)-(\theta+3)(1+2\theta)}{\theta^2(1+\theta)^2}\right]^2\frac{1}{I(\theta)}. \tag{9}$$

note that $Var(\hat{\theta}_{ML})$ contains the unknown parameter θ. It can be estimated by substituting $\hat{\theta}_{ML}$ into θ of the above equation. We denote the estimated value of $Var(\hat{\theta}_{ML})$ as $var(\hat{\theta}_{ML})$.

In the following, the method of moment (MM) is used to derive the estimator for ϕ. This approach uses the information from the r-th moment about the origin of a random variable to estimate the parameter of interest in the probability model. Under this approach, the r-th population moment is assumed to be equal to the corresponding sample moment. The equation can be written as

$$E(X^r)=\frac{1}{n}\sum_{i=1}^{n}X_i^r.$$

then, this equation is used to solve for the moment estimator. Consider the PX distribution, if r = 1, the first population moment is the expected value of X, i.e., $E(X)=\phi$. The first sample moment is the sample mean, i.e., $\bar{X}=\sum_{i=1}^{n}X_i/n$. By equating the population and sample moments, it follows that $\phi=\bar{X}$ and the MM estimator for ϕ is denoted by

$$\hat{\phi}_{MM}=\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i. \tag{10}$$

we can see that $\hat{\phi}_{MM}$ has an explicit closed-form solution. Moreover, it needs only a single stage to estimate ϕ. This differs from ML estimation, which requires estimating θ first to obtain the estimate of ϕ. The mean and variance of $\hat{\phi}_{MM}$ are derived and given in Theorem 2.

Theorem 2. Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample of size n from a PX distribution. The MM estimator of ϕ is $\hat{\phi}_{MM}$. The expected value and variance of $\hat{\phi}_{MM}$ are $E(\hat{\phi}_{MM})=\phi$ and

$$Var(\hat{\phi}_{MM})=\frac{1}{n}\left[\frac{\theta^3+5\theta^2+11\theta+3}{\theta^2(1+\theta)^2}\right],$$

respectively.

Proof. Recall that $\phi=\frac{\theta+3}{\theta(1+\theta)}$. We can derive that $E(\hat{\phi}_{MM})=E(\bar{X})=\phi$. Therefore, $\hat{\phi}_{MM}$ is the unbiased estimator of ϕ. The variance of $\hat{\phi}_{MM}$ is directly derived by

$$\mathrm{Var}(\hat{\phi}_{MM})=\mathrm{Var}(\bar{X})=\frac{1}{n}\mathrm{Var}(X)$$

$$\mathrm{Var}(\hat{\phi}_{MM})=\frac{1}{n}\left[\frac{\theta^3+5\theta^2+11\theta+3}{\theta^2(1+\theta)^2}\right]$$

since $\lim_{n\to\infty}E(\hat{\phi}_{MM})=\phi$ and $\lim_{n\to\infty}\mathrm{Var}(\hat{\phi}_{MM})=0$, then $\hat{\phi}_{MM}$ is the consistent estimator.

2.2 Interval Estimation for the Population Mean

The six confidence intervals for ϕ are introduced. These are constructed using the large-sample approximation with the properties of ML and MM estimators derived in the previous section, the likelihood ratio method, the log-transformation approach, and the two bootstraps based on percentile and bias-corrected and accelerated bootstrap approaches. The processes for constructing the confidence intervals are given as follows.

2.2.1 Asymptotic Confidence Interval using the ML Estimator

The first confidence interval relies on the asymptotic normality property of the ML estimator. In Theorem 1, we show the variance of $\hat{\phi}_{ML}$ with the approximated variance

$$\mathrm{var}(\hat{\phi}_{ML})=\left[\frac{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})-(\hat{\theta}_{ML}+3)(1+2\hat{\theta}_{ML})}{\hat{\theta}_{ML}^2(1+\hat{\theta}_{ML})^2}\right]^2\frac{1}{I(\hat{\theta}_{ML})},$$

where $I(\hat{\theta}_{ML})$ is the estimator of $I(\theta)$. Using the Wald-type method, $(1-\alpha)$a 100% confidence interval for ϕ is established as

$$CI_{pr1}:\hat{\phi}_{ML}\pm z_{1-\alpha/2}\sqrt{\mathrm{var}(\hat{\phi}_{ML})},$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ 100% percentile of a standard normal distribution, or N(0,1), and $\alpha\in(0,1)$ is a significance level. The Wald-type method is simple, but it is typically used as a starting point for constructing the confidence interval in theory. Also, the Wald-type confidence interval is widely applied in applications, particularly when the sample size (n) is large.

2.2.2 Asymptotic Confidence Interval using the MM Estimator

The second confidence interval is constructed using a large- sample approximation based on the properties of the MM estimator. As shown in Theorem 2, $\hat{\phi}_{MM}$ has the explicit expected value as $E(\hat{\phi}_{MM})=\phi$. The estimated variance of $\hat{\phi}_{MM}$ is derived as

$$\mathrm{var}(\hat{\phi}_{MM})=\frac{1}{n}\left[\frac{\hat{\theta}_{MM}^3+5\hat{\theta}_{MM}^2+11\hat{\theta}_{MM}+3}{\hat{\theta}_{MM}^2(1+\hat{\theta}_{MM})^2}\right],$$

where $\hat{\phi}_{MM}$ is the MM estimator for θ obtained based on the normal equation:

$$\frac{\theta+3}{\theta(1+\theta)}=\bar{X}$$

this equation is obtained from the relationship between the population and sample moments. It can also be written in quadratic form as

$$\bar{X}\theta^2+(\bar{X}-1)\theta-3=0$$

solving this equation for $\theta>0$, we have

$$\hat{\theta}_{MM}=\frac{-(\bar{X}-1)+\sqrt{(\bar{X}-1)^2+12\bar{X}}}{2\bar{X}}$$

under the central limit theorem, the sampling distribution of the sample mean is normally distributed. The pivotal quantity of ϕ given by

$$Z_1=\frac{\hat{\phi}_{MM}-\phi}{\sqrt{\mathrm{var}(\hat{\phi}_{MM})}}=\frac{\bar{X}-\phi}{\sqrt{\mathrm{var}(\bar{X})}}$$

has an N(0,1), for $n\to\infty$. Using the probability statement $1-\alpha=P(-z_{1-\alpha/2}\le Z_1\le z_{1-\alpha/2})$ and solving for ϕ, a $(1-\alpha)$ 100% confidence interval for ϕ is established as

$$CI_{pr2}:\hat{\phi}_{MM}\pm z_{1-\alpha/2}\sqrt{\mathrm{var}(\hat{\phi}_{MM})}$$

2.2.3 Asymptotic LR Confidence Interval

In this section, we introduce the likelihood ratio (LR) confidence interval. Generally, the LR statistic (Hudson, 1971) is the ratio of the likelihoods of two models, where the parameter estimates are obtained from the possible parameter space Ω and the restricted subspace ω, or parameter space under the null hypothesis (H_0). Under model (2), the LR statistic for parameter θ is given by

$$\Lambda=\frac{\sup_{\theta\in\omega}L(\theta)}{\sup_{\theta\in\Omega}L(\theta)}=\frac{L(\tilde{\theta})}{L(\hat{\theta}_{ML})}, \quad (11)$$

where $\Omega=\{\theta:\theta>0\}$ is the parameter space of θ, $\omega=\{\theta:\theta=\theta_0\}$ is the parameter space under $H_0:\theta=\theta_0$, and θ_0 is a specific value of θ. According to the likelihood function for θ given in (6), we have that

$$L(\hat{\theta}_{ML})=\frac{\hat{\theta}_{ML}^{2n}}{2^n(1+\hat{\theta}_{ML})^{\sum_{i=1}^{n}x_i+4n}}\prod_{i=1}^{n}\left[2(1+\hat{\theta}_{ML})^2+\hat{\theta}_{ML}(x_i+2)(x_i+1)\right]$$

is the likelihood under Ω and

$$L(\tilde{\theta})=\frac{\theta_0^{2n}}{2^n(1+\theta_0)^{\sum_{i=1}^{n}x_i+4n}}\prod_{i=1}^{n}\left[2(1+\theta_0)^2+\theta_0(x_i+2)(x_i+1)\right]$$

is the likelihood under H_0. Therefore, the approximate LR statistic is of the form

$$\Lambda_a=-2\log\left(\frac{L(\tilde{\theta})}{L(\hat{\theta}_{ML})}\right)=-2\left[\log L(\tilde{\theta})-\log L(\hat{\theta}_{ML})\right]. \qquad (12)$$

Λ_a is distributed as an asymptotically chi-squared random variable with one degree of freedom (df) for the large sample size. The LR confidence interval for θ consists of all parameter values that would not be rejected at the α significance level (Doganaksoy, 2021; Sangnawakij, 2024). Consequently, the asymptotic confidence interval for θ contains the lower and upper limits that satisfy

$$\log L(\hat{\theta}_{ML})-\kappa=\frac{1}{2}\chi^2_{1-\alpha,df=1}, \qquad (13)$$

where κ is a constant once data are observed and $\chi^2_{1-\alpha,df=1}$ is the (1-α)100% percentile of a chi-square distribution with df = 1.

In practice, numerical computing with software will be used to find the bounds of a two-sided confidence interval for θ. For instance, the *gosolnp* function within the *R* software (Ghalanos, & Theussl, 2015) can be applied. The lower and upper limits that satisfy (13) are then obtained. They are denoted as L_θ and U_θ. To estimate the population mean of the PX distribution, a (1-α) 100% asymptotic LR confidence interval for ϕ can be computed by

$$CI_{pr3}:\left(\frac{L_\theta+3}{U_\theta(1+U_\theta)},\frac{U_\theta+3}{L_\theta(1+L_\theta)}\right)$$

2.2.4 Asymptotic Confidence Interval using Log-transformation of Estimator

According to model (2), the support of θ is a positive real value; consequently, $\phi>0$. However, the lower bound of confidence intervals given in previous methods may sometimes provide a negative value (Gangopadhyay et al., 2024). To deal with this problem, it is often set to zero, which is the default lower bound of the parameter. A question arises whether setting a lower limit equal to zero is appropriate. This is because the estimation is unrealistic. Therefore, in this work, we address this problem by using the log-transformed estimation (Meeker et al., 2022). Next, the asymptotic confidence interval using the log-transformation of the MM estimator is constructed. This uses the mean and variance of $\log\hat{\theta}_{ML}$ and $\log\hat{\phi}_{ML}$ given in Theorem 3.

Theorem 3. Let $\hat{\theta}_{ML}$ and $\hat{\phi}_{ML}$ be the ML estimators of θ and ϕ of the PX distribution, respectively. The mean and variance of $\log\hat{\theta}_{ML}$ are $\log\hat{\theta}_{ML}=\log\theta$ and $Var(\log\hat{\theta}_{ML})=1/\theta^2 I(\theta)$. Hence, $E(\log\hat{\phi}_{ML})=\log\phi$ $E(\log\hat{\phi}_{ML})=\log\phi$ and

$$Var(\log\hat{\phi}_{ML})=\left[\frac{\theta(1+\theta)}{\theta+3}\right]^2\left[\frac{\theta(1+\theta)-(\theta+3)(1+2\theta)}{\theta^2(1+\theta)^2}\right]^2\frac{1}{I(\theta)}$$

Proof. Using the delta method and assuming that n is large enough, the mean of $\log\hat{\theta}_{ML}$ is estimated by $E(\log\hat{\theta}_{ML})=\log\theta$ and the mean of $\log\hat{\theta}_{ML}$ is $E(\log\hat{\phi}_{ML})=\log\phi$. The variance of $\log\hat{\theta}_{ML}$ can be approximated by

$$Var(\log\hat{\theta}_{ML})\approx\left[\frac{\partial}{\partial\hat{\theta}_{ML}}\left(\log\hat{\theta}_{ML}\right)\right]^2_{E(\hat{\theta}_{ML})=\theta}Var(\hat{\theta}_{ML})=\frac{1}{\theta^2 I(\theta)},$$

where $I(\theta)$ is the observed Fisher information of θ and $Var(\log\hat{\theta}_{ML})$ is the estimated variance given in Theorem 1. For the large sample size, it follows that $\log\hat{\theta}_{ML}\sim N\left(\log\theta,1/\theta^2 I(\theta)\right)$. Then, we show the variance of $\log\hat{\theta}_{ML}$ Again, it is based on the delta method. Since $Var(\hat{\phi}_{ML})$ is proved in Theorem 1, we yield

$$Var(\log\hat{\phi}_{ML})=Var\left(\log\frac{\hat{\theta}_{ML}+3}{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})}\right)\approx\left[\frac{\partial}{\partial\hat{\theta}_{ML}}\left(\log\frac{\hat{\theta}_{ML}+3}{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})}\right)\right]^2_{E(\hat{\theta}_{ML})=\theta}Var(\hat{\theta}_{ML})$$

$$Var(\hat{\phi}_{ML})=\left[\frac{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})}{\hat{\theta}_{ML}+3}\right]^2\left[\frac{\hat{\theta}_{ML}(1+\hat{\theta}_{ML})-(\hat{\theta}_{ML}+3)[\hat{\theta}_{ML}+(1+\hat{\theta}_{ML})]}{\hat{\theta}^2_{ML}(1+\hat{\theta}_{ML})^2}\right]^2_{E(\hat{\theta}_{ML})=\theta}\frac{1}{I(\theta)}$$

$$Var(\hat{\phi}_{ML})=\left[\frac{\theta(1+\theta)}{\theta+3}\right]^2\left[\frac{\theta(1+\theta)-(\theta+3)(1+2\theta)}{\theta^2(1+\theta)^2}\right]^2\frac{1}{I(\theta)}=\left[\frac{\theta(1+\theta)}{\theta+3}\right]^2Var(\hat{\phi}_{ML}) \qquad (14)$$

the estimated variance for $Var(\log\hat{\phi}_{ML})$ is obtained by substituting $\hat{\phi}_{ML}$ into θ of equation (14). The proof ends here.

The information given in Theorem 3 is used to find the pivotal quantity for $\log\phi$. It is given by the following function:

$$Z_2=\frac{\log\hat{\phi}_{ML}-\log\phi}{\sqrt{var(\log\hat{\phi}_{ML})}}$$

and follows an $N(0,1)$. Therefore, a (1 – α) 100% confidence interval for ϕ is derived as

$$CI_{pr4}:\exp\left\{\widehat{\log\phi}_{ML}\pm z_{1-\alpha/2}\sqrt{var(\widehat{\log\phi}_{ML})}\right\},$$

where $\mathrm{Var}(\log\hat{\phi}_{ML})$ is the estimated variance for $\mathrm{Var}(\log\hat{\phi}_{ML})$. Note that the basic exponential function is always positive. This method can address the problem of the negative lower bound of the confidence interval for the parameter truncated at zero.

2.2.5 Percentile Bootstrap Confidence Interval

Bootstrapping is a version of the statistical techniques used to estimate statistics about a population by resampling a dataset with replacement (Efron, & Tibshirani, 1993). It is a computer-intensive method for determining the sampling distribution of any statistics derived from a random sample. Therefore, it proves beneficial for statistical inferences involving complex statistics or situations where adequate statistical theory is unavailable. As a consequence, the approximated variance of the estimate and confidence limits for the parameter of interest can be obtained, regardless of the underlying distribution of the data (Mooney, & Duval, 1993; Wilcox, 2012).

The percentile bootstrap confidence interval is defined as the interval between (α/2) 100% and (1-α/2) 100% percentiles of the distribution for the B estimates of the interested parameter obtained from the sampling. In this paper, the mean parameter ϕ is needed to estimate. Since the bootstrap involves drawing a series of random samples from the original sample with replacement. Repeating the process, a large number of times is required to decrease the sampling error. We provide the algorithm of the percentile bootstrap confidence interval for ϕ. The steps are as follows:

1) Draw a bootstrap sample of size n with replacement from the original data $x_1, x_2,\ldots,x_n$ and denote the nonparametric bootstrap sample as $x_1^*,x_2^*,\ldots,x_n^*$.
2) Estimate a bootstrap statistic for ϕ using ML estimation, denoted by $\hat{\phi}_{ML}^*$.
3) Repeat steps 1 and 2 B times to obtain the B bootstrap statistics $\hat{\phi}_{ML}^{*(1)},\hat{\phi}_{ML}^{*(2)},\ldots,\hat{\phi}_{ML}^{*(B)}$.
4) Calculate the median of $\hat{\phi}_{ML}^{*(1)},\hat{\phi}_{ML}^{*(2)},\ldots,\hat{\phi}_{ML}^{*(B)}$ to get the bootstrap median $\hat{\phi}_{BT}$.
5) Calculate the lower and upper limits from (α/2)100% and (1-α/2)100% percentiles of $\hat{\phi}_{ML}^{*(1)},\hat{\phi}_{ML}^{*(2)},\ldots,\hat{\phi}_{ML}^{*(B)}$ to obtain the percentile bootstrap confidence interval for ϕ, namely $CI_{pr5}=\left(L_{\hat{\phi}_{ML}^{*(1)},\hat{\phi}_{ML}^{*(2)},\ldots,\hat{\phi}_{ML}^{*(B)}}(\alpha/2),U_{\hat{\phi}_{ML}^{*(1)},\hat{\phi}_{ML}^{*(2)},\ldots,\hat{\phi}_{ML}^{*(B)}}(1-\alpha/2)\right)$,

where $L(\cdot)$ and $U(\cdot)$ denote the lower and upper bounds of the confidence interval, and α is a significance level.

In the application and simulation, the bootstrap replication $B = 1000$ is used. Note that the bootstrap statistic introduced is based on the ML estimates. In fact, we can apply other methods, such as the MM estimates. However, we primarily take into account the invariant property of the ML estimator, leading us to select it here.

2.2.6 BCa Confidence Interval

An alternative bootstrap method is considered. We refer to the bias-corrected and accelerated (BCa) percentile interval (Efron, 1987), which is a distribution-free method used to construct a bootstrap confidence interval. It is similar to percentile bootstrapping; however, the BCa method introduces corrections for bias and skewness in the distribution of bootstrap estimates. In estimation, the two parameters, the bias-correction parameter (z_0) and the acceleration parameter (a) will be estimated. z_0 represents the proportion of bootstrap estimates that are less than the observed statistics, and a denotes the adjusted value for the skewness of the bootstrap distribution. Then, these values are used to adjust the endpoints of the percentile confidence interval. See more details of the theoretical concept and applications of the BCa method in Böhning et al., (2022), Chernick, & LaBudde (2011), and Efron, & Narasimhan (2020).

To estimate ϕ of the PX distribution, we define a (1-α)100% two-sided BCa percentile confidence interval as

$$CI_{pr6}=\left(\hat{\phi}^*(\alpha_1),\hat{\phi}^*(\alpha_2)\right),$$

where $\alpha_1=\Phi\left(\hat{z}_0+\frac{\hat{z}_0+z_\alpha}{1-\hat{a}(\hat{z}_0+z_\alpha)}\right)$ and $\alpha_2=\Phi\left(\hat{z}_0+\frac{\hat{z}_0+z_{1-\alpha}}{1-\hat{a}(\hat{z}_0+z_{1-\alpha})}\right)$ are the BCa level endpoints obtained from Monte Carlo simulations, where $\hat{z}_0$ is the estimator of z_0, $\hat{a}$ is the estimator of a calculated using a jackknife approach, $\Phi(\cdot)$is the cumulative distribution function (cdf) of an N(0,1), and z_α and $z_{1-\alpha}$ are the $\alpha\times100\%$ and (1-α) 100% percentiles of an N(0,1), respectively (Efron, & Narasimhan, 2020). The bias corrector is given by

$$\hat{z}_0=\Phi^{-1}\left(\frac{\#\left(\hat{\phi}_{ML}^{*(b)}<\hat{\phi}_{ML}\right)}{B}\right),$$

where $\Phi^{-1}(\cdot)$is the inverse cdf of an N(0,1), and $\#\left(\hat{\phi}_{ML}^{*(b)}<\hat{\phi}_{ML}\right)$ is the number of replications that give $\hat{\phi}_{ML}^{*(b)}$, or the mean estimate in the bootstrap sample b, less than $\hat{\phi}_{ML}$, or the mean estimate from the sample data, for b=1,2,...,B. Moreover, the acceleration factor is expressed as

$$\hat{a}=\frac{\sum_{i=1}^{n}\left(\hat{\phi}_{ML(\cdot)}-\hat{\phi}_{ML_i}\right)^3}{6\left[\sum_{i=1}^{n}(\hat{\phi}_{ML(\cdot)}-\hat{\phi}_{ML_i})^2\right]^{3/2}},$$

where $\hat{\phi}_{ML(\cdot)}$ is the average of jackknife estimates from n observations and $\hat{\phi}_{ML_i}$ is the estimate of ϕ having x_i removed from the dataset. We give the procedure for estimating the confidence limits of ϕ in the following steps:

1) Compute the ML estimate for ϕ from the n original observations $x_1,x_2,...,x_n$ using $\hat{\phi}_{ML}$.
2) Draw a bootstrap sample of size n with replacement from the original data to get the nonparametric bootstrap sample $x_1^*,x_2^*,...,x_n^*$.
3) Estimate a bootstrap statistic for ϕ using ML estimation, denoted as $\hat{\phi}_{ML}^*$.
4) Repeat steps 2 and 3 B times to obtain the B bootstrap statistics $\hat{\phi}_{ML}^{*(1)},\hat{\phi}_{ML}^{*(2)},...,\hat{\phi}_{ML}^{*(B)}$.
5) Calculate bias correction estimate $\hat{z}_0$ and acceleration value $\hat{a}$.
6) Compute the adjusted-percentile values α_1 and α_2

Find the lower and upper limits for ϕ from the α_1 and α_2 quantiles of the bootstrap distribution for $\hat{\phi}_{ML}^{*(b)}$, for b=1,2,…,B, to obtain a BCa bootstrap CI_{pr6}.

In practice, statistical software is commonly used to obtain bootstrap samples and statistics. Statistical software packages, including *R*, *S-Plus*, and *SPSS* are all capable of doing these calculations to compute parameter estimates. In this paper, we use the *bca* function in the *coxed* package (Kropko, & Harden, 2020) of *R* for estimating the BCa bootstrap interval.

3. Simulation Study

The aim of this simulation study is to evaluate the performance of proposed estimators and confidence intervals for the mean parameter of the PX distribution. We use the simulation procedure described below. The sample sizes are set at $n=$ 5, 10, 30, 50, 100, and 500 to represent small to large sample sizes. The data are sampled from a PX distribution with parameters θ = 0.35, 0.5, 1, and 1.5. So, the population means are ϕ =1.2, 2, 4.67, and 7.09. Note that the values of θ are varied to provide different data distributions. The shapes of PX distributions with various parameters are also shown in Figure 1. To generate sample data from the PX distribution, a mixture model, we follow these steps:

1) Set the initial value for θ.
2) Sample a variable u_i from a uniform distribution between 0 and 1, where i = 1,2,…,n.
3) If $u_i<\theta/(1+\theta)$, we sample $\lambda_i\sim Exp(\theta)$, otherwise $\lambda_i\sim Gam(3,1/\theta)$, see also equation (4).
4) Generate the data $x_i\sim Pois(\lambda_i)$ to get a PX sample of size n.

we then estimate ϕ using $\hat{\phi}_{ML}$, $\hat{\phi}_{MM}$, and $\hat{\phi}_{BT}$, and six confidence intervals, CI_{pr1} - CI_{pr6}. Here, the confidence level is given as 1 - α = 0.95. Each scenario is repeated $H=5000$ times using the *R* statistical software (R Core Team, 2024). On average, we compute the bias and mean squared error (MSE) of the estimator using the following quantities:

$$Bias(\hat{\phi})=\frac{1}{H}\sum_{h=1}^{H}\hat{\phi}_h-\phi \text{ and } MSE(\hat{\phi})=\frac{1}{H}\sum_{h=1}^{H}(\hat{\phi}_h-\phi)^2,$$

where $\hat{\phi}_h$ is the estimate for parameter ϕ in the h-th replication. The average coverage probability and expected length of the confidence interval for ϕ are approximated by

$$CP=\frac{\#(L_h\le\phi\le U_h)}{H} \text{ and } EL=\frac{1}{H}\sum_{h=1}^{H}(U_h-L_h),$$

respectively, where $\#(L_h\le\phi\le U_h)$ is the number that ϕ lies within the lower and upper limits of the confidence interval. An estimator with low variation across simulation replications and a confidence interval that has a close-to-nominal coverage probability of 0.95 with a short interval length is preferred.

The performances of the proposed estimators and confidence intervals for ϕ are presented in Tables 1 and 2, respectively. They are also displayed in Figure 2. We summarize the major findings of the simulation studies as follows:

1) The results indicate that the biases of $\hat{\phi}_{ML}$ and $\hat{\phi}_{MM}$ are similar. They approach zero when the sample size n is increased. The efficiency of the estimators does not depend on the true parameter ϕ.
2) Biases of $\hat{\phi}_{BT}$ are large if $\phi\le 2$ and n<50. However, when $\phi\le 2$, $\hat{\phi}_{BT}$ is superior to $\hat{\phi}_{ML}$ and $\hat{\phi}_{MM}$ in all cases. Note that a large value of ϕ leads to a small value of θ, reflecting less skewed data.
3) The three estimators exhibit small MSEs. Moreover, their MSEs tend to decrease if n is increased. When n < 50, the MSEs of $\hat{\phi}_{ML}$ are slightly smaller than those of the comparators.

4) CI_{pr1} and CI_{pr2} (the confidence intervals using ML and MM estimators, respectively) have similar coverage probabilities. Their coverage rates are much lower than the nominal coverage level of 0.95 when n<30, otherwise, they have better performance.
5) The coverage probabilities of CI_{pr3} (the likelihood ratio confidence interval) are much greater than 0.95 for all situations in the study. It follows that the expected lengths of CI_{pr3} are greater than the other confidence intervals.
6) Coverage probabilities of CI_{pr4} (the confidence interval using the log-ML estimator) maintain the nominal coverage rate in general. The performance of this confidence interval does not depend on n and ϕ. Moreover, CI_{pr4} provides an acceptable, small, expected interval length.
7) The percentile bootstrap (CI_{pr5}) and BCa bootstrap (CI_{pr6}) confidence intervals have the coverage probabilities lower than 0.95. They cannot cover the true parameter in many cases, especially for small sample sizes. The results also show that their expected lengths are too small. However, the BCa method improves the percentile interval in terms of coverage probability, comparing with CI_{pr5}.

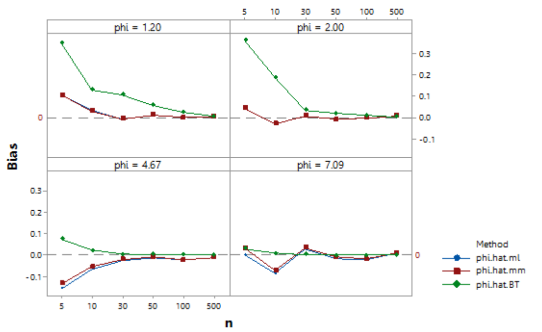
In summary, the bias and MSE of the estimators proposed in this paper tend to zero for a large sample size. These verify that they satisfy the property of a consistent estimator. The performance of and is similar in terms of bias, but has a better MSE. is then recommended for use when However, outperforms the MM and ML estimators in terms of bias when. It is suggested to estimate the population mean in such a case. For interval estimation, is the best confidence interval compared to the other proposed confidence intervals in terms of coverage probability and interval width. It can be used to estimate the mean parameter of the PX distribution in a wide range of situations, including small, moderate, and large sample sizes.

Table 1 Estimated bias and mean squared error (MSE) of the three estimators for the population mean ϕ in the PX distribution from simulations

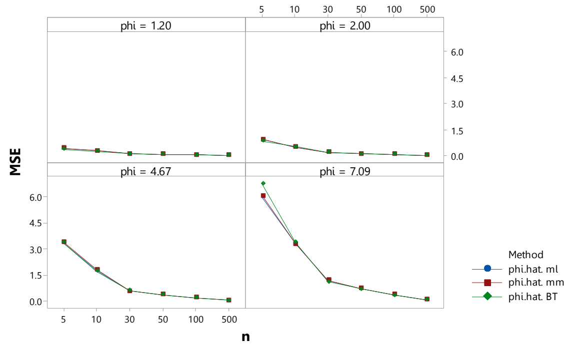
ϕ	n	Bias			MSE		
		$\hat{\phi}_{ML}$	$\hat{\phi}_{MM}$	$\hat{\phi}_{BT}$	$\hat{\phi}_{ML}$	$\hat{\phi}_{MM}$	$\hat{\phi}_{BT}$
1.20	5	0.1043	0.1040	0.3462	0.3947	0.4003	0.3451
	10	0.0315	0.0300	0.1277	0.2561	0.2569	0.2280
	30	-0.0072	-0.0076	0.1058	0.0754	0.0755	0.0845
	50	0.0120	0.0115	0.0573	0.0450	0.0450	0.0479
	100	0.0011	0.0009	0.0261	0.0241	0.0241	0.0238
	500	0.0031	0.0030	0.0048	0.0046	0.0046	0.0048
2.00	5	0.0413	0.0440	0.3597	0.9138	0.9277	0.8301
	10	-0.0287	-0.0283	0.1863	0.4696	0.4731	0.4935
	30	0.0063	0.0067	0.0345	0.1623	0.1629	0.1662
	50	-0.0083	-0.0083	0.0182	0.1017	0.1018	0.1001
	100	-0.0020	-0.0020	0.0086	0.0518	0.0518	0.0495
	500	0.0079	0.0079	0.0016	0.0103	0.0103	0.0099
4.67	5	-0.1551	-0.1332	0.0743	3.3410	3.3932	3.3763
	10	-0.0652	-0.0549	0.0222	1.7835	1.7932	1.7249
	30	-0.0240	-0.0190	0.0055	0.5653	0.5671	0.5772
	50	-0.0112	-0.0091	0.0031	0.3493	0.3518	0.3445
	100	-0.0230	-0.0221	0.0015	0.1746	0.1744	0.1780
	500	-0.0120	-0.0119	0.0003	0.0343	0.0344	0.0353
7.09	5	-0.0006	0.0315	0.0295	5.9799	6.0751	6.7542
	10	-0.0865	-0.0709	0.0091	3.2744	3.2874	3.3573
	30	0.0269	0.0348	0.0022	1.1695	1.1784	1.1069
	50	-0.0152	-0.0091	0.0013	0.7034	0.7092	0.6732
	100	-0.0222	-0.0173	0.0006	0.3536	0.3539	0.3357
	500	0.0091	0.0097	0.0001	0.0647	0.0646	0.0669

Table 2 Estimated coverage probability and expected length of the six 95% confidence intervals for ϕ of the PX distribution from simulations

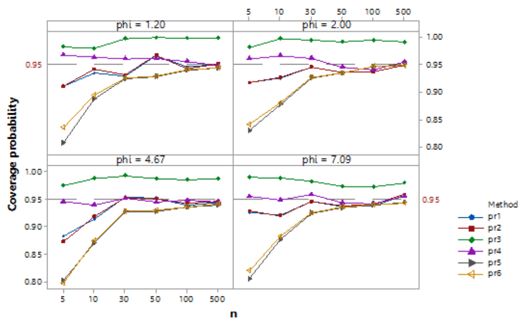
ϕ	n	CI_{pr1}	CI_{pr2}	CI_{pr3}	CI_{pr4}	CI_{pr5}	CI_{pr6}
				Coverage probability			
1.20	5	0.9100	0.9100	0.9822	0.9671	0.8075	0.8350
	10	0.9342	0.9412	0.9794	0.9632	0.8877	0.8940
	30	0.9294	0.9310	0.9972	0.9602	0.9240	0.9260
	50	0.9656	0.9670	0.9986	0.9608	0.9278	0.9286
	100	0.9458	0.9428	0.9974	0.9558	0.9394	0.9406
	500	0.9492	0.9512	0.9980	0.9482	0.9436	0.9440
2.00	5	0.9171	0.9172	0.9810	0.9600	0.8300	0.8413
	10	0.9252	0.9266	0.9968	0.9662	0.8774	0.8802
	30	0.9452	0.9452	0.9940	0.9612	0.9258	0.9270
	50	0.9358	0.9358	0.9914	0.9454	0.9344	0.9342
	100	0.9366	0.9366	0.9940	0.9400	0.9460	0.9464
	500	0.9490	0.9490	0.9904	0.9546	0.9496	0.9482
4.67	5	0.8834	0.8732	0.9758	0.9464	0.8021	0.7979
	10	0.9142	0.9194	0.9890	0.9404	0.8708	0.8746
	30	0.9550	0.9538	0.9936	0.9524	0.9286	0.9296
	50	0.9532	0.9518	0.9880	0.9456	0.9286	0.9306
	100	0.9404	0.9436	0.9866	0.9490	0.9366	0.9366
	500	0.9452	0.9464	0.9882	0.9460	0.9404	0.9418
7.09	5	0.9270	0.9294	0.9908	0.9562	0.8060	0.8203
	10	0.9216	0.9206	0.9892	0.9498	0.8772	0.8832
	30	0.9468	0.9466	0.9834	0.9594	0.9254	0.9260
	50	0.9372	0.9384	0.9746	0.9444	0.9350	0.9366
	100	0.9390	0.9416	0.9730	0.9424	0.9404	0.9410
	500	0.9578	0.9596	0.9808	0.9562	0.9450	0.9440
				Expected length			
1.20	5	2.8287	2.8399	13.3804	3.4694	1.8285	1.9159
	10	1.9372	1.9384	1.4497	2.1544	1.6712	1.7077
	30	1.1051	1.1052	1.8344	1.1457	1.0725	1.0813
	50	0.8670	0.8672	1.3866	0.8858	0.8369	0.8420
	100	0.6100	0.6101	0.9539	0.6167	0.5981	0.6001
	500	0.2735	0.2735	0.4192	0.2740	0.2707	0.2710
2.00	5	3.8942	3.9152	14.7286	4.5532	2.8658	2.9149
	10	2.7134	2.7176	9.7741	2.9399	2.3863	2.4206
	30	1.5979	1.5991	2.3577	1.6409	1.5249	1.5361
	50	1.2329	1.2335	1.7831	1.2528	1.2068	1.2130
	100	0.8749	0.8750	1.2446	0.8819	0.8599	0.8619
	500	0.3930	0.3930	0.5516	0.3936	0.3893	0.3897
4.67	5	7.0496	7.1099	19.1812	7.8134	5.5849	5.6089
	10	5.0978	5.1200	7.8397	5.3667	4.5530	4.6008
	30	2.9772	2.9845	3.8084	3.0288	2.8754	2.8911
	50	2.3138	2.3170	2.9227	2.3377	2.2530	2.2611
	100	1.6340	1.6357	2.0460	1.6425	1.6152	1.6182
	500	0.7325	0.7330	0.9103	0.7333	0.7291	0.7295
7.09	5	10.0561	10.1384	13.1738	10.9365	7.8698	7.8903
	10	7.0728	7.1065	9.0846	7.3802	6.3559	6.4132
	30	4.1484	4.1610	5.0208	4.2075	3.9750	3.9927
	50	3.2002	3.2088	3.8378	3.2276	3.1328	3.1426
	100	2.2621	2.2677	2.6934	2.2718	2.2344	2.2384
	500	1.0159	1.0175	1.2015	1.0168	1.0072	1.0076



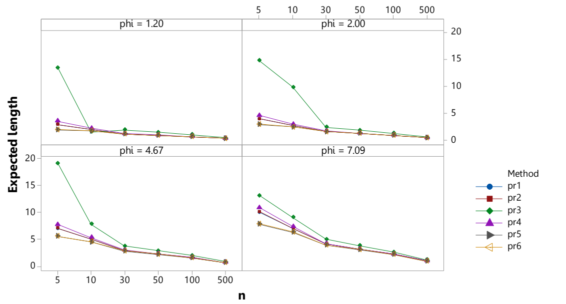
(a) Bias



(b) Mean squared error



(c) Coverage probability



(d) Expected length

Figure 2 (a) average bias; (b) mean squared error of the mean estimators; (c) average coverage probability of the confidence intervals; (d) length of interval under simulated data with true population mean ϕ and sample size n

4. Numerical Illustration

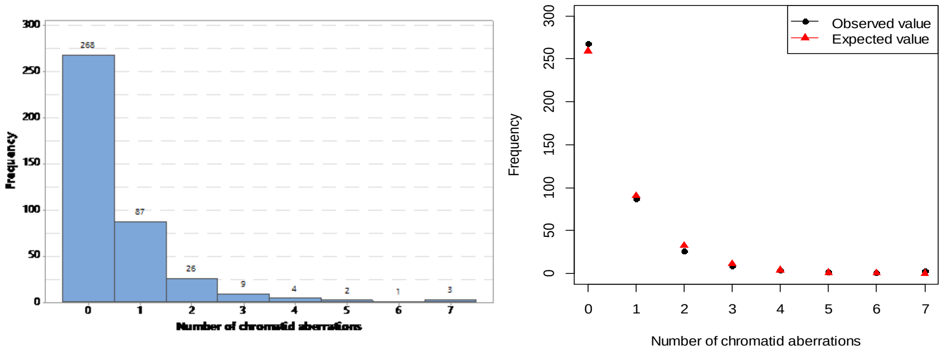
This section uses two real data sets to demonstrate the newly proposed estimators for the population mean of the PX distribution. The chromosome aberrations in genetic applications and the number of victims of unrest events in the southern region of Thailand are among the data considered.

4.1 Chromosome Aberrations

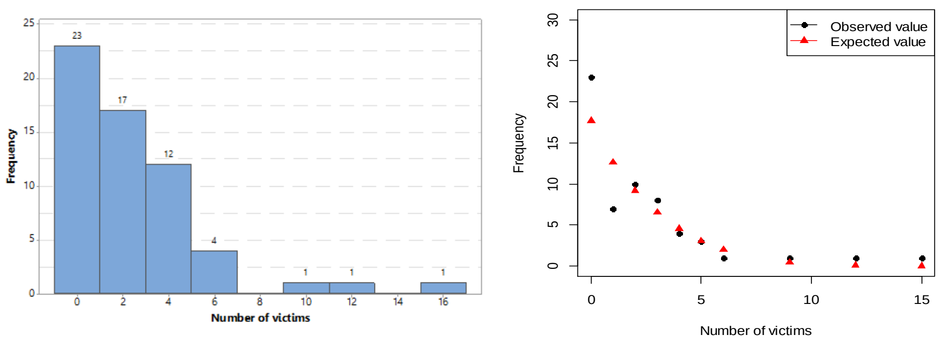
The first data set is related to the number of chromatid aberrations (0.2 g chinon 1, 24 hours) in human leukocytes. It was obtained from Altun et al., (2022), discussed previously by Shanker, & Fesshaye (2015). The data consist of 400 observations and are displayed in Figure 3(a), left panel. We can see that they have a right-skewed distribution; most data fall to the right; and there is zero inflation. Altun et al., (2022) used the Akaike information criterion (AIC) and Bayesian information criteria (BIC) to select the suitable model. They investigated that the data fitted to the PX distribution better than Poisson, Poisson-Lindley, generalized Poisson-Lindley, and negative binomial distributions, as the PX distribution had the smallest AIC and BIC compared to the other models. Herein, we show the observed and expected frequencies for this data set in Figure 3(a), right panel. The chi-square goodness of fit (GoF) test statistic for the PX distribution was 4.86 (df = 3) with a p-value of 0.18. As a result, the chromosome aberration data are suitable for the PX distribution at the 0.05 level of significance, as well as for use in this paper.

4.2 Unrest Events in the Southern Area of Thailand

For the second dataset, we analyze the number of victims affected by unrest events that occurred in the southern Thailand in 2023. These data were collected and reported by the Southern Border Area News Summarizes team, which is part of the Office of Academic Resources, Prince of Songkla University, Pattani Campus, Thailand (see https://summarise.wbns.oas.psu.ac.th/home/). Specifically, the data were collected from five provinces: Satun, Songkhla, Pattani, Yala, and Narathiwat. They are shown on the map given in Figure 4. In this paper, we refer to the victim as a person directly harmed by a crime in the study area, including physical injury or death. However, we do not include the number of insurgents who suffered injuries or lost their lives. Figure 3(b) shows the number of victims in the study area from January to December 2023, as well as the fitted frequencies. It can be seen that there were 59 days of unrest. The maximum observed frequency was 15 people per day. According to the data, there are 23 days without injury or death, seven days with one victim, and so on. Using the GoF test, the number of victims followed a PX distribution at a significance level of 0.05. The GoF statistic and corresponding p-value were 4.86 (df = 4) and 0.30, respectively.



a) Chromatid aberrations data



(b) Unrest event in the southern border area of Thailand in 2023

Figure 3 Observed frequencies with fitted frequencies under the PX distribution for (a) chromatid aberrations data and (b) unrest event in the southern border area of Thailand in 2023

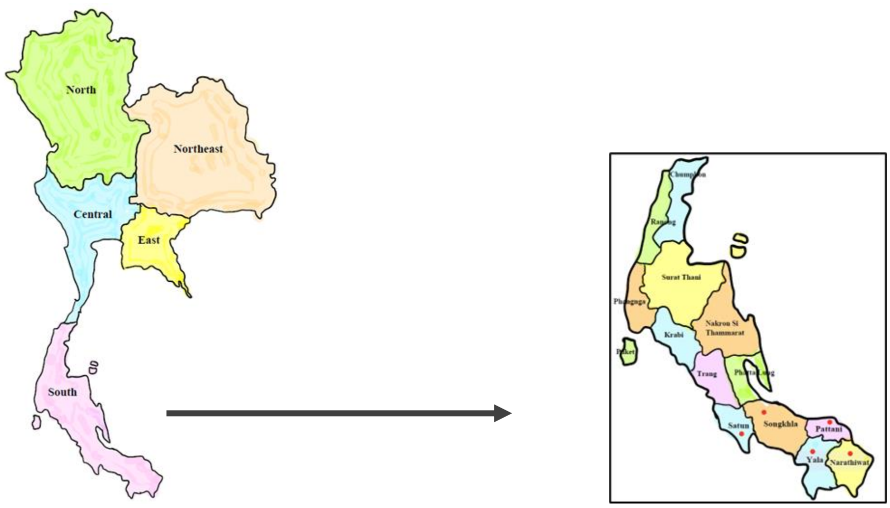


Figure 4 A map of southern Thailand: Highlight the five provinces (Satun, Songkhla, Pattani, Yala, and Narathiwat) using red dots

4.3 Parameter Estimation using Real Data

The estimators proposed in the methodology section are applied to estimate the mean number of chromatid aberrations and the average number of victims during unrest events in Thailand in 2023. The estimated means computed from the three methods using the first data set are

$$\hat{\phi}_{ML}= 0.5447,\ \hat{\phi}_{MM}=0.5475, \text{ and } \hat{\phi}_{BT}= 0.5424$$

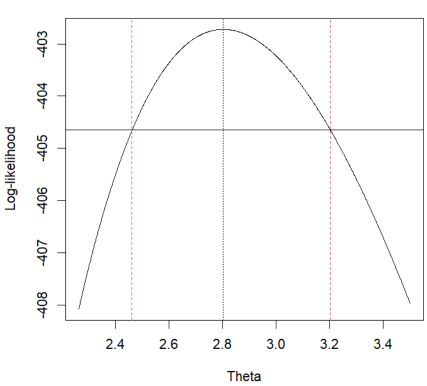
for the second data set, they are

$$\hat{\phi}_{ML}=2.1026,\ \hat{\phi}_{MM}=2.1017, \text{ and } \hat{\phi}_{BT}=2.0852$$

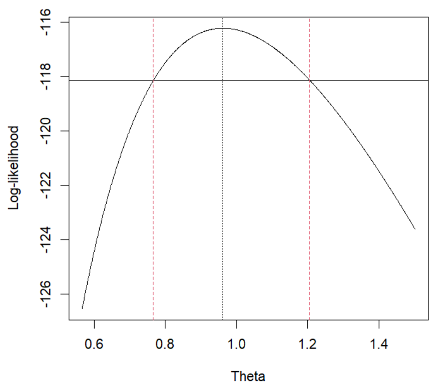
based on the ML estimate of the PX distribution, we calculate $\widehat{E}(X)=\hat{\phi}_{ML}$ and $\widehat{V}ar(X)=\frac{\hat{\theta}^3_{ML}+5\hat{\theta}^2_{ML}+11\hat{\theta}_{ML}+3}{\hat{\theta}^2_{ML}(1+\hat{\theta}_{ML})^2}$ hence, the dispersion index is approximated by $\widehat{V}ar(X)/\widehat{E}(X)$. For example, the dispersion for data in Thailand is 5.0115, meaning that the variance of the data is around five times greater than the mean. We conclude that the data exhibit overdispersion. The 95% confidence intervals from the six methods are given in Table 3. The results clearly show that the asymptotic likelihood ratio confidence interval, or CI_{pr3}, has the largest interval width. The results from two real data examples match those from the simulation study. Although CI_{pr3} appears to have a large interval length, it works well in computation as the confidence interval converges and can be solved for the lower and upper limits (see Figure 5).

Table 3 The dispersion estimate, estimated mean, 95% confidence interval for ϕ, and length of interval using the two data examples

Dataset (Sample size)	Dispersion Index	Estimate for ϕ	95% Confidence interval for ϕ	Interval length
Chromosome aberrations	5.8481	$\hat{\phi}_{ML}$ = 0.5447	CI_{pr1} = (0.4550, 0.6344)	0.1794
($n=400$)		$\hat{\phi}_{MM}$ = 0.5475	CI_{pr2} = (0.4575, 0.6375)	0.1800
		$\hat{\phi}_{BT}$ = 0.5424	CI_{pr3} = (0.4061, 0.7279)	0.3218
			CI_{pr4} = (0.4620, 0.6422)	0.1802
			CI_{pr5} = (0.4454, 0.6569)	0.2114
			CI_{pr6} = (0.4497, 0.6662)	0.2165
Victims in Thailand	5.0115	$\hat{\phi}_{ML}$ = 2.1026	CI_{pr1} = (1.5083, 2.6969)	1.1886
($n=59$)		$\hat{\phi}_{MM}$ = 2.1017	CI_{pr2} = (1.5103, 2.6930)	1.1827
		$\hat{\phi}_{BT}$ = 2.0852	CI_{pr3} = (1.4170, 3.1111)	1.6941
			CI_{pr4} = (1.5849, 2.7894)	1.2045
			CI_{pr4} = (1.4362, 2.8523)	1.4161
			CI_{pr6} = (1.4746, 2.9373)	1.4626



(a)



(b)

Figure 5 Log-likelihood plot of 95% likelihood ratio confidence limits for parameter θ of the PX distribution using (a) chromatid aberrations data ($\hat{\theta}_{ML}$ = 2.80) and (b) unrest event in the southern border area of Thailand ($\hat{\theta}_{ML}$ = 0.96). Note that the red-dash lines indicate the lower and upper limits

5. Conclusions

This paper introduces estimators and confidence intervals for the population mean of the PX distribution, which are then applied to count data sets. For point estimation, the maximum likelihood, moment, and bootstrap methods are used. The large-sample theory and bootstrap approach are applied to derive the two-sided confidence intervals. The performance of these estimators is evaluated through simulation studies in various situations. The results show that the maximum likelihood and moment estimators have excellent performance in terms of bias if the data from the PX distribution has a long right tail. Otherwise, we suggest using the bootstrap for point estimation. To quantify uncertainty in an estimate of the parameter, the confidence interval is used in statistical inference. Nonparametric bootstrap methods need computer software to calculate the confidence limits. Our simulations show that these methods are limited to estimate the mean for $n < 500$. In such a case, the percentile bootstrap and BCa confidence intervals are not suggested to estimate the population mean of the PX distribution. The highlight of this work is the confidence interval based on log-transformed maximum likelihood estimation. It can address the issue of a negative lower bound, as the mean parameter of the PX distribution is always a positive real value. Furthermore, the simulation results indicate that this confidence interval performs well and maintains coverage probability across all sample sizes in the study. The confidence interval using log-transformed maximum likelihood estimation is recommended for use in applications. The asymptotic confidence intervals based on the properties of maximum likelihood and moment estimators can be used as alternative methods for sample sizes larger than or equal to 50.

Count data can sometimes contain many zeros, as demonstrated in the first application of this paper. When there are too many zeros, a zero-inflated distribution (Perumean-Chaney et al., 2013) may be more appropriate. On the other hand, sometimes zero cannot occur. As in capture-recapture applications, some units are detected but some remain hidden, so a zero-event count is not reported (Böhning et al., 2018; Sangnawakij, & Böhning, 2024). A zero-truncated probability model could be used instead of the untruncated distribution. Future research will focus on estimating the mean parameter for the PX distribution with zero inflation or truncation.

6. Acknowledgements

The author gratefully thanks to the Editor-in-chief and referees for the valuable comments and recommendations which help to improve this paper. This study was supported by Thammasat University Research Fund, Contract No TUFT 84/2567.

7. References

Alghamdi, F. M., Ahsan-ul-Haq, M., Hussain, M. N. S., Hussam, E., Almetwally, E. M., Aljohani, H. M., ... & Yusuf, M. (2024). Discrete Poisson Quasi-XLindley distribution with mathematical properties, regression model, and data analysis. *Journal of Radiation Research and Applied Sciences*, *17*(2), Article 100874. https://doi.org/10.1016/j.jrras.2024.100874

Altun, E., Cordeiro, G. M., & Ristic, M. M. (2022). An one-parameter compounding discrete distribution. *Journal of Applied Statistics*, *49*(8), 1935-1956. https://doi.org/10.1080/02664763.2021.1884846

Böhning, D., Bunge, J., & Heijden, P. G. (Eds.). (2018). *Capture-recapture methods for the social and medical sciences*. Boca Raton: CRC Press.

Böhning, D., Sangnawakij, P., & Holling, H. (2022). Confidence interval estimation for the Mantel–Haenszel estimator of the risk ratio and risk difference in rare event meta-analysis with emphasis on the bootstrap. *Journal of Statistical Computation and Simulation*, *92*(6), 1267-1291. https://doi.org/10.1080/00949655.2021.1991347

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

Chernick, M. R., & LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R*. John Wiley & Sons.

Conway, R. W., & Maxwell, W. L. (1962). A queueing model with state dependent service rates. *Journal of Industrial Engineering*, *12*(2), 132-136.

Doganaksoy, N. (2021). A simplified formulation of likelihood ratio confidence intervals using a novel property. *Technometrics*, *63*(1), 127-135. https://doi.org/10.1080/00401706.2020.1750488

Déniz, E. G. (2013). A new discrete distribution: Properties and applications in medical care. *Journal of Applied Statistics*, *40*(12), 2760-2770. https://doi.org/10.1080/02664763.2013.827161

Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, *82*(397), 171-185. https://doi.org/10.2307/2289144

Efron, B., & Narasimhan, B. (2020). The automatic construction of bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, *29*(3), 608-619. https://doi.org/10.1080/10618600.2020.1714633

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman and Hall/CRC.

Gangopadhyay, A. K., Mondal, R., Lodhi, C., & Maiti, K. (2024). Bayesian inference on parameters and reliability characteristics for inverse Xgamma distribution under adaptive-general progressive Type-II censoring. *Journal of Radiation Research and Applied Sciences*, *17*(2), 1-17. https://doi.org/10.1016/j.jrras.2024.100890

Ghalanos, A., & Theussl, S. (2015). *General non-linear optimization*. Retrieved July 2024, from https://cran.r-project.org/web/packages/Rsolnp/Rsolnp.pdf

Gschlößl, S., & Czado, C. (2008). Modelling count data with overdispersion and spatial effects. *Statistical Papers*, *49*, 531-552. https://doi.org/10.1007/s00362-006-0031-6

Hilbe, J. M. (2011). *Negative binomial regression*. Cambridge University Press.

Hilbe, J. M. (2017). The statistical analysis of count data. *Culture and Education*, *29*(3), 409-460. https://doi.org/10.1080/11356405.2017.1368162

Huang, A. (2023). On arbitrarily underdispersed discrete distributions. *American Statistical Association*, *77*(1), 29-34. https://doi.org/10.1080/00031305.2022.2106305

Hudson, D. J. (1971). Interval estimation from the likelihood function. *Journal of the Royal Statistical Society Series B. 33*(2), 256-262. https://www.jstor.org/stable/2985006

Imoto, T., Ng, C. M., Ong, S. H., & Chakraborty, S. (2017). A modified Conway-Maxwell-Poisson type binomial distribution and its applications. *Communications in Statistics - Theory and Methods*, *46*(24), 12210-12225. https://doi.org/10.1080/03610926.2017.1291974

Kropko, J., & Harden, J. J. (2020). Beyond the hazard ratio: generating expected durations from the cox proportional hazards model. *British Journal of Political Science*, *50*(1), 303-320. https://doi.org/10.1017/S000712341700045X

Meeker, W. Q., Escobar, L. A., & Pascual, F. G. (2021). *Statistical methods for reliability data*. John Wiley & Sons.

Mooney, C. Z., & Duval, R. D. (1993). *Bootstrapping: A nonparametric approach to statistical inference*. Sage Publications.

Oehlert, G. W. (1992). A note on the delta method. *The American Statistician*, *46*(1), 27-29. https://doi.org/10.2307/2684406

Oseni, B. M., Makinde, O. S., & Adepetun, A. O. (2024). Dispersion insensitive truncated models for inflated count data with upper bound. *Communications in Statistics-Simulation and Computation*, *53*(12), 5884-5898. https://doi.org/10.1080/03610918.2023.2202843

Perumean-Chaney, S. E., Morgan, C., McDowall, D., & Aban, I. (2013). Zero-inflated and overdispersed: what's one to do?. *Journal of Statistical Computation and Simulation*, *83*(9), 1671-1683. https://doi.org/10.1080/00949655.2012.668550

R Core Team. (2024). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing. https://www.r-project.org/

Sangnawakij, P. (2024). A Simple Profile Likelihood-based Confidence Interval for the Risk Ratio in Rare Events Meta-analysis. *Thailand Statistician*, *22*(2), 312-327.

Sangnawakij, P., & Anlamlert, W. (2023). Alternative confidence interval for variability parameters in the normal distribution with applications to stock exchange index data set. *Austrian Journal of Statistics*, *52*(5), 63-81. https://doi.org/10.17713/ajs.v52i5.1534

Sangnawakij, P., & Böhning, D. (2024). On repeated diagnostic testing in screening for a medical condition: How often should the diagnostic test be repeated?. *Biometrical Journal*, *66*(3), 1-16. https://pubmed.ncbi.nlm.nih.gov/38637326/

Sankaran, M. (1970). The discrete Poisson-Lindley distribution. *Biometrics*, *26*, 145-149. https://doi.org/10.2307/2529053

Sen, S., Maiti, S. S., & Chandra, N. (2016). The xgamma distribution: statistical properties and application. *Journal of Modern Applied Statistical Methods*, *15*(1), Article 38.

Shanker, R. (2016). The discrete Poisson-Amarendra distribution. *International Journal of*

Statistical Distributions and Applications, 2(2), 14-21. https://doi.org/10.11648/j.ijsd.20160202.11

Shanker, R., & Fesshaye, H. (2015). On Poisson-Lindley distribution and its applications to biological sciences. *Biometrics & Biostatistics International Journal*, *2*(4), 103-107. https://doi.org/10.15406/bbij.2015.02.00036

Shanker, R., Fesshaye, H., Shanker, R., Leonida, T. A., & Sium, S. (2017). On discrete Poisson-Shanker distribution and its applications. *Biometrics & Biostatistics International Journal, 5*(1), Article 00121. https://doi.org/10.15406/bbij.2017.05.00121

Shukla, K. K., & Shanker, R. (2019). A discrete Poisson-Ishita distribution and its applications. *International Journal of Statistics and Economics*, *20*(2), 109-122.

Tang, W., He, H., & Tu, X. M. (2023). *Applied categorical and count data analysis*. Chapman and Hall/CRC.

Weems, K. S., Sellers, K. F., & Li, T. (2023). A flexible bivariate distribution for count data expressing data dispersion. *Communications in Statistics - Theory and Methods*, *52*(13), 4692-4718. https://doi.org/10.1080/03610926.2021.1999474

Wilcox, R. R. (2011). *Introduction to robust estimation and hypothesis testing*. Academic press.