Modeling Rare Events using a Zero-Inflated Poisson (ZIP) Distribution: Some New Results on Point Estimation

Suntaree Unhapipat[1,2,3,*], Nabendu Pal[2] and Montip Tiensuwan[1,3]

[1]Department of Mathematics, Faculty of Science, Mahidol University, Thailand
[2]Department of Mathematics, University of Louisiana at Lafayette, Louisiana, USA
[3]Center of Excellence in Mathematics, Commission on Higher Education, Bangkok

This paper takes a fresh look on point estimation of model parameters under a Zero-Inflated Poisson (ZIP) distribution. The reason is that some finer details of point estimation, if overlooked, may lead to wrong estimates as was done by the earlier researchers. In this paper we have achieved the following new results: (a) A new set of corrected method of moments estimators has been proposed; (b) We have shown how the standard technique of differentiating the log-likelihood function to find the maximum likelihood estimators may lead to wrong estimates, as well as how to avoid this problem; and (c) A new adjusted maximum likelihood estimation technique has been proposed which not only produces meaningful estimates always, but also appears to work better compared to all other estimation techniques in terms of standardized mean squared error (SMSE) when ZIP is used to model rare events. Finally, datasets on rare events have been used to demonstrate the estimation techniques, and how the ZIP distribution can be used to model such datasets.

Keywords: Maximum likelihood estimation, method of moments estimation, standardized mean squared error, standardized bias, goodness of fit test.

1. Introduction

The Poisson distribution is widely used to model count data that involves number of events (or elements) per unit time (or space). There are several applications in the literature that used Poisson distribution to model the real-life datasets, e.g., the number of blemishes per sheet of white bond paper (Doane and

Seward (2010)), the number of network failures per day (Levine *et al*. (2011)) and the number of arrivals at a car wash in one hour (Anderson *et al*. (2012)). However, Poisson distribution is not very helpful to model rare events, that means when the dataset contains many zeros. The Poisson parameter, which represents both mean and variance of the distribution, does not provide enough flexibility to model the data if the sample mean differs greatly from the sample variance. In such a situation Zero-Inflated Poisson (ZIP) distribution is found to be extremely useful. A ZIP distribution with parameters π and λ, denoted by ZIP(π,λ), has the following probability mass function (*pmf*), with $0 \le \pi \le 1$ and $\lambda \ge 0$, as

$$P(X=k)=\begin{cases}\pi+(1-\pi)e^{-\lambda}, \ if\ k=0,\\ (1-\pi)e^{-\lambda}\dfrac{\lambda^k}{k!}, \ if\ k\neq 0.\end{cases} \tag{1.1}$$

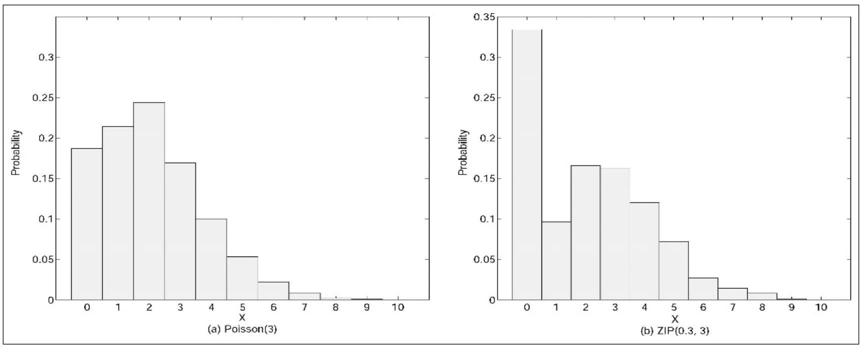


Figure 1.1. Plots of Poisson(λ) and ZIP(π,λ) with π = 0.3 and λ = 3

The above Figure 1.1 shows the similarities and dissimilarities between the usual Poisson(λ) and the ZIP(π, λ) distributions. The ZIP distribution assigns an extra probability π (called the inflation probability) to the value {0}, on top of the regular Poisson probability at this value due to its over-occurrences, and then multiplies the Poisson probabilities over the rest of the region {1, 2, 3, …} by a factor of ($1-\pi$) as a compensation. This is illustrated further with the following real-life example which will be discussed further in Section 5.

Example 1.1. The Bureau of Vector-Borne Diseases, Department of Disease Control, Ministry of Public Health (Royal Government of Thailand), 2015, provides the number of yearly new Elephantiasis patients reported in Thailand (by Provinces). Elephantiasis is a symptom of a variety of diseases, where parts of a person's body (especially the limbs) swell to massive proportions (http://en.wikipedia.org/wiki/Elephantiasis). In this paper, we use the dataset from Phuket province which is one of famous provinces for tourists. Table 1.1 provides the frequency distribution of Elephantiasis patients from 1992 - 2013.

Table 1.1. Number of yearly new elephantiasis patients of Phuket (1992 - 2013)

New Elephantiasis Patients	Count										
1992 – 2002	1	1	3	3	2	0	0	0	0	1	1
2003 – 2013	1	0	0	0	0	0	0	0	0	0	0

Table 1.2. Frequency distribution of elephantiasis patients of Phuket (1992 - 2013)

Number of new Elephantiasis patients	0	1	≥ 2	Total
Frequency	14	5	3	22

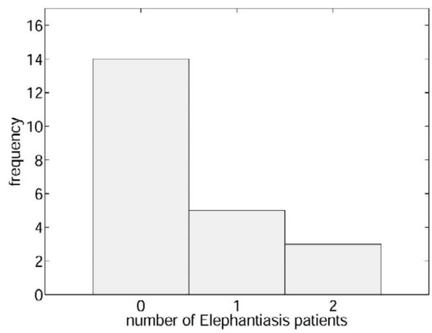


Figure 1.2. Histogram of elephantiasis patients dataset in Table 1.2

For the above data set on the yearly number of Elephantiasis patients in Phuket province (see the frequency histogram in Figure 1.2), if one wants to fit a regular Poisson(λ) distribution, then the corresponding goodness-of-fit test p-value comes out to be 0.0574 which accepts the Poisson model. It will be shown in Section 5 that a suitable ZIP model gives a far better fit to the above data set. For other applications of ZIP one can see Böhning *et al.* (1999) to model tooth decay data from the Brazilian school children; Xie (2001) to model the number of read-write errors discovered in a computer hard disk used in a manufacturing process; and Davidson (2012) to model recurrent colorectal adenomas. Lord *et al.* (2005) investigated several distributions for modeling the crash data, i.e., Poisson, Poisson-gamma, ZIP and zero inflated negative binomial (ZINB), and multinomial probability models; moreover, the authors provided some guidance for choosing the appropriate model for the crash data. In 2006, Ghosh *et al.* proposed the application of ZIP to model the number of defects that resulted from an experiment for improving printed circuit board (PCB) manufacturing quality at Nortel, RTP, North Carolina. Beckett *et al.* (2014) showed how ZIP provides a far superior approach (compared to Poisson(λ)) to model annual frequency of major earthquakes in the US, the number of US wildfires, the number of Atlantic hurricanes having landfall in the US, the number of lighting fatalities in the State of Louisiana (USA), and the number of tornado occurrences in Lafayette Parish (Louisiana, USA).

The main objective of this paper is to provide superior point estimators of π and λ. In the next section (Section 2) we revisit the two popular estimation techniques, i.e., method of moments estimation (MME) and the maximum likelihood estimation (MLE). Beckett *et al.* (2014) pointed out that the MME can take values outside the parameter space, and hence it needs to be corrected by truncating the MME suitably (henceforth called 'corrected MME' or CMME). In this paper we have proposed a new corrected MME which often works better than the previous CMME. Further, certain computational difficulties in obtaining the MLE was overlooked in earlier studies which may lead to wrong estimates, and this has been pointed out in Section 2 also. In Section 3 we propose a new method of estimating the parameters using the partial likelihood functions. This approach, called adjusted MLE (or AMLE), is much simpler and does not face the computational difficulties as seen for regular MLE. Further, our comprehensive simulation study presented in Section 4 shows that the AMLE can be superior to all other estimators while dealing with rare events. Finally, in Section 5 we demonstrate all the four estimation methods with some real-life datasets from Thailand and how the ZIP distribution can be useful to model rare events.

2. The Two Common Estimation Methods with Corrections

Throughout this paper we assume that $X_1, X_2, ..., X_n$ are *iid* ZIP(π, λ) unless mentioned otherwise. Let $\bar{X}$ and s^2 be the sample mean and the sample variance respectively. The mean and variance of ZIP(π, λ) are

$$E(X) = \lambda(1-\pi) \text{ and } V(X) = \lambda(1-\pi)(1+\pi\lambda) \tag{2.1}$$

2.1 Method of Moments Estimation (MME)

By equating the population moments with their sample counterparts we obtain the MME of (π, λ) as (with subscript 'MME')

$$\hat{\lambda}_{MME} = \bar{X} + \frac{s^2}{\bar{X}} - 1, \text{ and} \tag{2.2}$$

$$\hat{\pi}_{MME} = \left\{\frac{s^2}{\bar{X}} - 1\right\} / \hat{\lambda}_{MME} = D / \left(D + \bar{X}^2\right), \tag{2.3}$$

where $D = \left(s^2 - \bar{X}\right)$. Define the events E_1 and E_2 as

$$E_1 = \{\bar{X} < \left(s^2 + \bar{X}^2\right)\} \quad \text{and} \quad E_2 = \{\bar{X} < s^2\}. \tag{2.4}$$

Note that $E_2 \subset E_1$.

Beckett *et al.* (2014) proposed the following corrected MME (i.e., CMME), denoted with the subscript 'MME1', as follows:

$$\hat{\lambda}_{MME1} = \begin{cases} \bar{X}, & on\, E_2^c \\ \hat{\lambda}_{MME}, & on\, E_2 \end{cases} \quad (2.5)$$

and

$$\hat{\pi}_{MME1} = \begin{cases} 0, & on\, E_2^c \\ \hat{\pi}_{MME}, & on\, E_2. \end{cases} \quad (2.6)$$

The justification for CMME was that on E_2^c the statistic D takes a negative value. Hence, $\hat{\pi}_{MME}$ takes a negative value, and therefore it needs to be truncated at zero. Also, when $\hat{\pi}$ takes the value zero, the ZIP reduces to a regular Poisson(λ), and therefore, $\hat{\lambda}$ should be $\bar{X}$.

Now we are going to propose a new type of correction on the MME as follows:

(i) λ_{MME} should take the value which is greater than 0, i.e., $s^2 + \bar{X}^2 > \bar{X}$. Hence,

$$\hat{\lambda}_{MME} > 0 \text{ on } E_1. \quad (2.7)$$

(ii) The space of π is (0, 1), i.e., the estimator of π is expected to be between 0 and 1. It can be considered in two following cases: **(a)** $\hat{\pi}_{MME} > 0$ if (a.1) $s^2 - \bar{X} > 0$ and $s^2 - \bar{X} + \bar{X}^2 > 0$ or (a.2) $s^2 - \bar{X} < 0$ and $s^2 - \bar{X} + \bar{X}^2 < 0$. Note that (a.1) holds if $s^2 > \bar{X}$ which is E_2. On the other hand, (a.2) holds if $s^2 < \bar{X}$ and $s^2 + \bar{X}^2 < \bar{X}$ which is equivalent to $s^2 + \bar{X}^2 < \bar{X}$, i.e., E_1^c holds. So, $\hat{\pi}_{MME} > 0$ on $E_1^c \cup E_2$. **(b)** $\hat{\pi}_{MME} < 1$, i.e., $D/(D + \bar{X}^2) < 1$. This means either (b.1) $D < (D + \bar{X}^2)$ if $D + \bar{X}^2 > 0$ or (b.2) $D > (D + \bar{X}^2)$. Note that (b.2) can never happen, and (b.1) is equivalent to E_1. Hence, $\hat{\pi}_{MME} \in (0,1)$ provides $(E_1^c \cup E_2) \cap E_1$ which is equivalent to $(E_1 \cap E_2) = E_2$. So,

$$0 < \hat{\pi}_{MME} < 1 \text{ on } E_2. \quad (2.8)$$

Therefore, a new correction on $\hat{\pi}_{MME}$ and $\hat{\lambda}_{MME}$ is done by truncating each parameter estimator at the respective parameter space boundary as given below. The resultant estimators are identified with the subscript 'MME2'.

$$\hat{\lambda}_{MME2} = \begin{cases} \hat{\lambda}_{MME}, & on\, E_1 \\ 0, & on\, E_1^c \end{cases} \tag{2.9}$$

and

$$\hat{\pi}_{MME2} = \begin{cases} \hat{\pi}_{MME}, & on\, E_2 \\ 0, & on\, E_1 \cap E_2^c \\ 1, & \text{on } E_1^c. \end{cases} \tag{2.10}$$

Tables A.1 - A.2 (in the Appendix) show the approximated $P(E_1)$ and $P(E_2)$ for selected values of n and (π,λ). These approximated probabilities have been obtained by replicating data from ZIP(π,λ) with 10^5 replications. Also, $P\left(E_1 \cap E_2^c\right) = P\left(E_1\right) - P\left(E_2\right)$.

2.2 Maximum likelihood estimation (MLE)

Define the random variable Y as

$$Y = \sum_{i=1}^{n} I\left(X_i = 0\right) = \text{number of zero observations.} \tag{2.11}$$

Then the likelihood function of the given data is

$$L\left(\pi, \lambda | X_1, X_2, \ldots, X_n\right) = \left\{\pi + \left(1-\pi\right)e^{-\lambda}\right\}^{Y} \prod_{X_i \neq 0} \left\{\left(1-\pi\right)e^{-\lambda}\lambda^{X_i} / \left(X_i!\right)\right\}, \tag{2.12}$$

i.e., the log-likelihood function, called L_*, is

$$L_* = Yln\left\{\pi + \left(1-\pi\right)e^{-\lambda}\right\} + \left(n-Y\right)\ln\left(1-\pi\right) - \left(n-Y\right)\lambda + \left(n\bar{X}\right)ln\lambda - \sum_{i=1}^{n}\ln\left(X_i!\right). \tag{2.13}$$

Beckett *et al.* (2014) suggested setting partial derivative of L_* w.r.t. π and λ equal to zero, to obtain the MLE of (π,λ) denoted by $\hat{\pi}_{MLE}$ and $\hat{\lambda}_{MLE}$ by solving

$$\frac{n\bar{X}}{\lambda} = \frac{Y\left(1-\pi\right)e^{-\lambda}}{\pi + \left(1-\pi\right)e^{-\lambda}} + \left(n-Y\right), \quad \text{and} \tag{2.14}$$

$$\left(n-Y\right) = \frac{Y\left(1-\pi\right)e^{-\lambda}}{\pi + \left(1-\pi\right)e^{-\lambda}}. \tag{2.15}$$

It is easily seen by substitution that solving (2.14) and (2.15) is equivalent to solving

$$\pi = 1 - \left(\bar{X} / \lambda \right), \text{ and} \tag{2.16}$$

$$g(\lambda) := \frac{\lambda e^{\lambda}}{e^{\lambda} - 1} = \frac{n\bar{X}}{n - Y}. \tag{2.17}$$

First, one needs to solve (2.17) to get $\hat{\lambda}_{MLE}$ which is then used in (2.16) to get $\hat{\pi}_{MLE}$. However, solving the above system of equations can be problematic for certain datasets as shown in the following two remarks.

Remark 2.1 Consider a dataset where all observations are zero, i.e., $Y = n$ and $\bar{X} = 0$. This can happen when π is close to 1, and/or λ is close to 0. In this case, the right hand side (RHS) of (2.17) is undefined (because we have 0/0), rendering the equation unsolvable. Even if we use $Y = n$ and $\bar{X} = 0$ in (2.14) – (2.15), we obtain $\hat{\pi}_{MLE} = 1$ and $\hat{\lambda}_{MLE}$ = any real value. This implies that the ZIP distribution is degenerating at $\{0\}$, and it really does not matter what value of $\hat{\lambda}_{MLE}$ is obtained. From an estimation point of view this perhaps is acceptable, but it makes studying the sampling distribution of $\hat{\lambda}_{MLE}$ ambiguous. As a result, bias and variance of $\hat{\lambda}_{MLE}$ are going to be nonunique.

Remark 2.2 Consider another situation where $0 < Y < n$, but the (n-Y) nonzero observations are all equal to 1, i.e., $n\bar{X} = \sum_{i=1}^{n} X_i = (n - Y)$. In this case solving (2.17) amounts to solving

$$h(\lambda) := \lambda e^{\lambda} - e^{\lambda} + 1 = 0. \tag{2.18}$$

The function $h(\lambda)$ is nondecreasing since $h'(\lambda) = \lambda e^{\lambda} \geq 0$ and $h(0) = 0$. This means that $h(\lambda)$ starts at 0 and keeps increasing. Thus the only solution of (2.18) is $\hat{\lambda}_{MLE} = 0$. This creates two difficulties: (a) If $\hat{\lambda}_{MLE} = 0$, then the estimated ZIP distribution is again degenerating at $\{0\}$ which seems impractical since some observations (i.e., (n-Y) observations) are strictly greater than 0. (b) If $\hat{\lambda}_{MLE} = 0$, then (2.16) yields $\hat{\pi}_{MLE} = -\infty$, which is again absurd, and hence $\hat{\pi}_{MLE}$ needs to be truncated at 0. But if we plug in $\lambda = \hat{\lambda} = 0$ in (2.14) or (2.15), then any value of $\pi = \hat{\pi} \in [0,1]$ will work. This makes studying sampling distribution of $\hat{\lambda}_{MLE}$ again ambiguous.

The difficulties with the computation of MLE described in the above two remarks was overlooked in Beckett *et al*. (2014) and this requires a fresh attention as we have discussed next.

Remark 2.3 (a) Recall the Remark 2.2 discussed earlier with all nonzero X_i's being equal to 1. Though the MLE form (2.16) – (2.17) comes out to be $(\hat{\pi}_{MLE}, \hat{\lambda}_{MLE})$ = (0, 0) (after suitable boundary value adjustment), a direct grid search method reveals that the actual maxima is somewhat different.

(b) A grid search method divides the parameter space (all possible values of (π,λ)) $\Theta = [0,1]\otimes[0,\infty]$ into a grid with a sufficiently large upper bound for λ. In other words, first we restrict (π,λ) to $\Theta_0 = [0,1]\otimes[0, M]$, for a sufficiently large M, and then study the value of L_* by varying π and λ with small increments over E_0. This is quite time consuming with π varying with an increment of $\delta_1 = 10^{-3}$ and λ varying with an increment of $\delta_2 = 10^{-2}$ unless M is not too large. This method coupled with M=15 gives a value of $(\hat{\pi}_{MLE}, \hat{\lambda}_{MLE})$ which is quite different from (2.16) – (2.17) in Remark 2.2. The grid search method, though accurate, is hard to employ in simulation studies where number of replications is large, say 10^4 or higher and M is very large. Thus, the grid search method can be used when λ is expected to be small (for rare events), i.e., M is not too large.

(c) The properties of the estimators, including MLE, can be studied through simulation, especially when the theoretical sampling distributions of the estimators are intractable. In the light of our findings of the difficulties with (2.16) – (2.17) it seems appropriate to study the standardized mean squared error (SMSE) and standardized bias (SB) of all estimators afresh. The equations (2.16) – (2.17) are helpful in finding the MLE in some cases, but not always. In the following subsection we propose a better way to compute the MLE depending on the observations, and this is somewhat new from Beckett *et al.* (2014).

2.3 A better approach to compute the MLE

The MLE of (π,λ) can be computed as follows.

Case – (i): All observations are zero, i.e., $Y = n$,, $\overline{X} = 0$.

From (2.13) it is seen that

$$L_* = n\left[\ln\{\pi+(1-\pi)e^{-\lambda}\}-1\right]. \tag{2.19}$$

Since $(\partial L_*/\partial\pi) \geq 0$ and $(\partial L_*/\partial\lambda) \leq 0$, L_* is maximized at $\pi = 1$ and $\lambda = 0$. Thus

$$\hat{\pi}_{MLE} = 1 \text{ and } \hat{\lambda}_{MLE} = 0. \tag{2.20}$$

Case – (ii): All nonzero observations are equal to 1 and $Y < n$.

Here a grid search method needs to be applied with $M = 15$, δ_1 = increment of $\pi = 10^{-3}$ and δ_2 = increment of $\lambda = 10^{-2}$. The selection of $M = 15$ has been discussed in Remark 2.4 later.

As a demonstration of how different the grid search maxima can be from the one in (2.14) - (2.15), the following table (Table 2.1) provides the MLE from two methods.

Table 2.1. Computation of the MLE of λ, i.e., $\hat{\lambda}$, by (a) using grid search and (b) solving equations (2.14) – (2.15) with $n = 10$ under Case - (ii)

$\hat{\lambda}$	Y	0	1	2	3	4	5	6	7	8	9
using	$\bar{X}$	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
(a)		1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
(b)		0	0	0	0	0	0	0	0	0	0

The following Figure 2.1 shows the contour plot of L_* against π and λ with n=10 and Y=5 ($\bar{X}$ =0.5). Note that the grid search implies $\hat{\lambda}_{MLE}$ = 0.5, whereas (2.14) - (2.15) gives $\hat{\lambda}_{MLE}$ = 0.

Case – (iii): All other datasets not falling under Case - (i) or Case - (ii).

Under this case, RHS of (2.17) is strictly larger than 1 (since $n\bar{X} = \sum X_i$ = sum of nonzero observations > number of nonzero observations = $n-Y$). Here, a nonnegative solution of (2.17) exists, which in turn produces a nonnegative solution of (2.16). But $\hat{\pi}_{MLE}$ needs to be truncated at 0 if needed.

Remark 2.4. How likely are we to encounter these three cases for our dataset? It is easily seen that

$$\text{P(Case – (i))} = P(X_i = 0, 1 \le i \le n) = \{\pi + (1-\pi)\, e^{-\lambda}\}^n;$$

$$\text{P(Case – (ii))} = \text{P(some } X_i\text{'s are zero, and all others are 1)}$$

$$= \sum_{k=1}^{n} P\,(k \text{ of } X_i\text{'s are 1 and remaining of } X_i\text{'s are 0})$$

$$= \sum_{k=1}^{n} \binom{n}{k} \left\{(1-\pi)e^{-\lambda}\lambda\right\}^k \left\{\pi + (1-\pi)e^{-\lambda}\right\}^{n-k};$$

$$\text{P(Case – (iii))} = 1 - [\text{P(Case – (i))} + \text{P(Case – (ii))}].$$

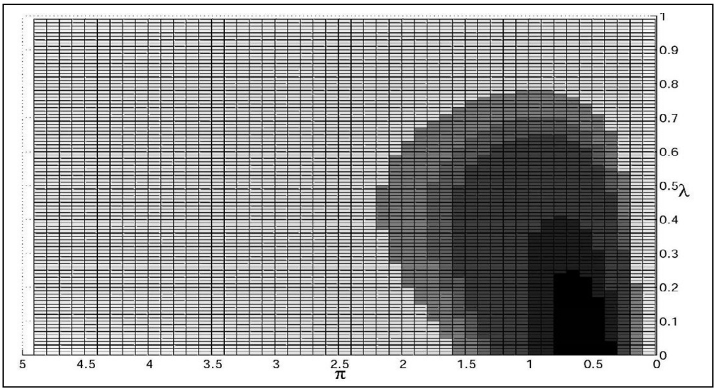


Figure 2.1. The contour plot of $L*$ in Case - (ii) for n = 10 and Y = 5.

Tables A.3 – A.4 (see Appendix) show the above probabilities for selected values of n and (π, λ). Note that for λ greater than or equal to 10, Case – (ii) is not likely to be encountered. Therefore, the grid search method, which is needed for Case – (ii) only, is employed with $M = 15$, a sufficiently large upper bound for λ.

3. The 'Adjusted MLE' (AMLE) Approach

Here we are going to propose a completely new approach to estimate the ZIP parameters. This calls for splitting the likelihood function in two parts, estimate the parameters based on each part, and then combining the two sets of estimates using a suitable convex combination.

Without loss of generality, assume that the first m observations, i.e., X_1, X_2, ..., X_m, are nonzero. The partial likelihood function of these nonzero observations is

$$L^{(m)} = L\left(\pi, \lambda | X_1, X_2, \ldots, X_m\right) = \left(1-\pi\right)^m e^{-m\lambda} \lambda^{\sum_{i=1}^{m} X_i} / \prod_{i=1}^{m} X_i ! \tag{3.1}$$

Define $L_*^{(m)} = lnL^{(m)}$. Note that $(\partial L_*^{(m)}/\partial \pi) = 0$ and $(\partial L_*^{(m)}/\partial \lambda) = 0$ the MLE of π and λ based on the partial dataset as

$$\hat{\pi}^{(m)} = 0 \quad \text{and} \quad \hat{\lambda}^{(m)} = \sum_{i=1}^{m} X_i / m = \bar{X}^{(m)} \quad \text{(say)}. \tag{3.2}$$

For the other zero observations, the remaining partial likelihood function, denoted by $L^{(n-m)}$, is

$$L^{(n-m)} = L(\pi, \lambda \,|\, X_{m+1}, \ldots, X_n = 0) = \{\pi + (1 - \pi)\, e^{-\lambda}\}^{(n-m)}. \tag{3.3}$$

It is easy to see that $L_*^{(n\text{-}m)} = lnL^{(n-m)}$ is maximized at

$$\hat{\pi}^{(n-m)} = 1 \text{ and } \hat{\lambda}^{(n-m)} = 0. \quad (3.4)$$

(In terms of our earlier notation, $Y = (n\text{-}m)$.)

A new type of estimator of (π, λ) is obtained by a convex combination as follows

$$(\hat{\pi}(\epsilon),\ \hat{\lambda}(\epsilon)) = \epsilon(\hat{\pi}^{(n\text{-}m)},\ \hat{\lambda}^{(n\text{-}m)}) + (1-\epsilon)(\hat{\pi}^{(n\text{-}m)},\ \hat{\lambda}^{(n\text{-}m)}) = (\epsilon,\ (1-\epsilon)\overline{X}^{(m)}), \quad (3.5)$$

where the convex coefficient $\epsilon \in [0,1]$ will be chosen suitably. The expression (3.5) is now plugged in the combined likelihood function $L = L^{(m)} \cdot L^{(n\text{-}m)}$, i.e., $L_* = L_*^{(m)} + L_*^{(n\text{-}m)}$, which is a function of ϵ only. Therefore,

$$L_* = (n-m)\ ln\{\epsilon + (1-\epsilon)\ e^{-(1-\epsilon)\overline{X}(m)}\} + m \cdot ln(1-\epsilon) \quad (3.6)$$
$$-\ m\overline{X}^{(m)}(1-\epsilon) + m\overline{X}^{(m)}\ ln\,(1-\epsilon) + C,$$

where the last term C depends only on X_i's. The value of ϵ, say $\hat{\epsilon}$, is the one which maximizes L_* w.r.t. ϵ. The resultant estimator in (3.5), with $\epsilon = \hat{\epsilon}$, is called the adjusted MLE (or AMLE).

Under *Case* – (*ii*), all nonzero observations are equal to 1, therefore $\overline{X}^{(m)} = 1$. Then the following table (Table 3.1) provides values of $\hat{\epsilon}$, $\hat{\pi}(\hat{\epsilon})$, and $\hat{\lambda}(\hat{\epsilon})$ for $n = 10$. Since $\hat{\pi}(\hat{\epsilon}) = \hat{\epsilon}$ and $\hat{\lambda}(\hat{\epsilon}) = (1-\hat{\epsilon})\overline{X}^{(m)}$, the results of Table 3.1 appear to be more reasonable since each parameter estimate is strictly within the respective parameter space.

Table 3.1. Values of $\hat{\epsilon}$, $\hat{\pi}(\hat{\epsilon})$, and $\hat{\lambda}(\hat{\epsilon})$ with $n = 10$ under Case – (ii)

m	1	2	3	4	5	6	7	8	9	10
$\hat{\epsilon}$	0.6723	0.5293	0.4152	0.3137	0.2189	0.1256	0.0305	0.0000	0.0000	0.0000
$\hat{\pi}(\hat{\epsilon})$	0.6723	0.5293	0.4152	0.3137	0.2189	0.1256	0.0305	0.0000	0.0000	0.0000
$\hat{\lambda}(\hat{\epsilon})$	0.3277	0.4707	0.5848	0.6863	0.7811	0.8744	0.9695	1.0000	1.0000	1.0000

4. Comparison of All Four Estimation Methods

We are going to compare all the four estimations techniques (i.e., MME1, MME2, MLE and AMLE) in terms of standardized mean squared error (SMSE) and standardized bias (SB). Note that SMSE and SB are more informative than the usual mean squared error (MSE) and the regular bias since the error in estimation is measured with respect to the actual parameter value.

Let be a general notation of an estimator of π which can be $\hat{\pi}_{MME1}$, $\hat{\pi}_{MME2}$, $\hat{\pi}_{MLE}$ and $\hat{\pi}_{AMLE}$. The SMSE and SB of the estimated parameter $\hat{\pi}$ are defined as

$$\text{SMSE}(\hat{\pi}) = (1/\pi^2)\,\text{MSE}(\hat{\pi}) = (1/\pi^2)\,E(\hat{\pi}-\pi)^2 \quad \text{and} \tag{4.1}$$

$$\text{SB}(\hat{\pi}) = (1/\pi)\,E(\hat{\pi}-\pi). \tag{4.2}$$

The SMSE and SB of ˆ can be obtained by using $\hat{\lambda}$ instead of $\hat{\pi}$ and λ in place of π in equations (4.1) and (4.2), respectively.

In this paper, the SMSE and SB of the estimators are approximated by a simulation which is conducted through the following steps:

For each fixed value of sample size n and various combinations of (π,λ),

(i) draw random samples $X_1, X_{2,\ldots,}X_n$, from a ZIP (π,λ);

(ii) based on the random sample in (i), compute $(\hat{\pi}_{MME1}, \hat{\lambda}_{MME1})$, $(\hat{\pi}_{MME2}, \hat{\lambda}_{MME2})$, $(\hat{\pi}_{MLE}, \hat{\lambda}_{MLE})$ and $(\hat{\pi}_{AMLE}, \hat{\lambda}_{AMLE})$;

(iii) repeat step (i) and (ii) for $Q = 10^4$times, thus we have $\hat{\pi}^{(q)}$ and $\hat{\lambda}^{(q)}$ for $1 \le q \le Q$. Then approximate the SMSE of $\hat{\pi}$ as

$$\text{SMSE}(\hat{\pi}) = (1/\text{Q})\sum_{q=1}^{Q}\left[\left(\hat{\pi}^{(q)}-\pi\right)^2\right]/\pi^2. \tag{4.3}$$

The SB($\hat{\pi}$) is approximated in a similar way. The following Figures 4.1 – 4.4 show the plots of SMSE($\hat{\pi}$) and SMSE($\hat{\lambda}$). Similarly, Figures 4.5 – 4.8 represent the plots of SB($\hat{\pi}$) and SB($\hat{\lambda}$). Tables 4.1 and 4.2 provide the simulated MSE($\hat{\pi}$) and MSE($\hat{\lambda}$) of four estimators.

Table 4.1. The simulated MSE of four estimators for π = 0.2.

MSE(·)	Estimator	λ				
		1.0	**2.0**	**3.0**	**5.0**	**10.0**
$\hat{\pi}$	MME1	0.1636	0.0897	0.0624	0.0449	0.0364
	MME2	0.1636	0.0897	0.0624	0.0449	0.0364
	MLE	0.0700	0.1600	0.0600	0.0400	0.0400
	AMLE	0.0600	0.0200	0.0400	0.0400	0.0600
$\hat{\lambda}$	MME1	0.1159	0.0938	0.0831	0.0588	0.0310
	MME2	0.1454	0.0986	0.0849	0.0588	0.0310
	MLE	0.0729	0.1648	0.0663	0.0388	0.0255
	AMLE	0.0541	0.0402	0.0711	0.0980	0.0835

Table 4.2. The simulated MSE of four estimators for $\lambda = 3.0$.

MSE(•)	Estimator	π			
		0.2	0.4	0.6	0.8
$\hat{\pi}$	MME1	0.0437	0.0272	0.0176	0.0125
	MME2	0.0437	0.0272	0.0176	0.0094
	MLE	0.0400	0.0400	0.0000	0.0000
	AMLE	0.0400	0.0800	0.0600	0.0800
$\hat{\lambda}$	MME1	0.0435	0.0591	0.0906	0.2157
	MME2	0.0435	0.0591	0.0906	0.2157
	MLE	0.0288	0.0390	0.0588	0.1554
	AMLE	0.0741	0.1932	0.4272	0.9453

Remark 4.1 The results in Tables 4.1 – 4.2 as well as those in Figures 4.1 – 4.8 help us get an idea about the standard error (SE) of various estimators. For example, while estimating π by $\hat{\pi}$ (which can be any one of those four estimators mentioned earlier), MSE($\hat{\pi}$) = π^2 SMSE($\hat{\pi}$) = Var($\hat{\pi}$) + (π SB($\hat{\pi}$))^2, i.e., Var($\hat{\pi}$) = MSE($\hat{\pi}$) – (π SB($\hat{\pi}$))^2. Therefore, SE($\hat{\pi}$) = estimated $\sqrt{\mathrm{Var}(\hat{\pi})}$. Since the actual parameters π and λ are unknown, the above estimation in SE is carried out by using $\pi = \hat{\pi}$ and $\lambda = \hat{\lambda}$.

Remark 4.2 Since our objective here has been to use ZIP to model rare events, the value of λ has been taken deliberately not too large in our simulation study.

(a) For estimating π, the following patterns have been observed. (i) When λ is very small, the SMSE plot of AMLE seems to have the best performance with respect to π. But as λ increases, MME2 and MLE appear to provide a better performance over a larger range of π values. (ii) On the other hand, when the SMSE is seen as a function of λ (and for fixed π), the AMLE provides the best performance for small π and small sample sizes. As the sample size increases and/or the value of π increases, the performance of AMLE gets overshadowed by MME2 and MLE.

(b) For estimating λ, the following patterns have been observed. (i) When λ is very small, the SMSE as a function of π, shows that the best performance is offered by AMLE. But as λ gets larger, the performance of MME2 and MLE get superior. (ii) On the other hand, when the SMSE is viewed as a function of λ, for fixed π, we note that the AMLE is good only when both n and π are very small. For large n and/or larger π, a better performance is provided by MME2 and MLE.

(c) In all practical applications we recommend that all the four types of estimators be tried to see which one provides the best fit for the data as shown in Section 5.

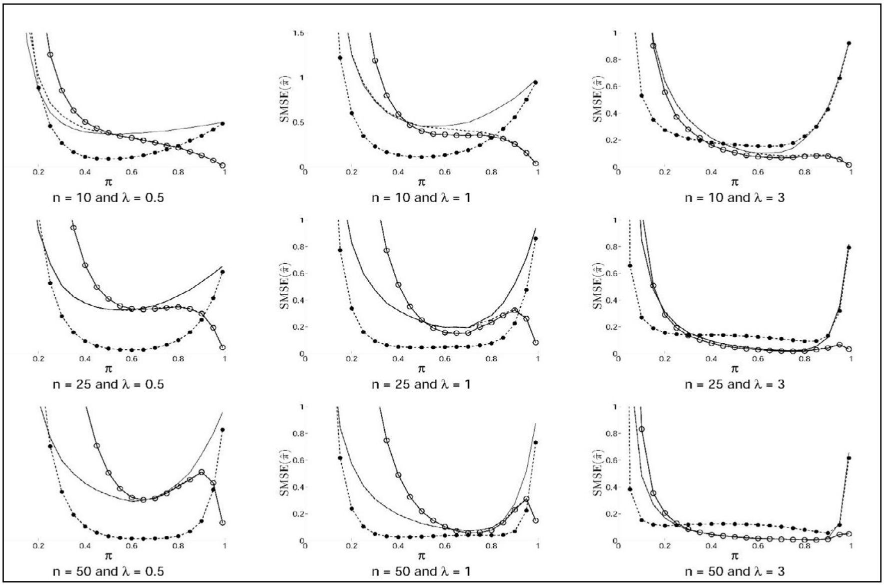


Figure 4.1. The simulated SMSE($\hat{\pi}$) of four estimators are plotted against π for selected values of λ and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

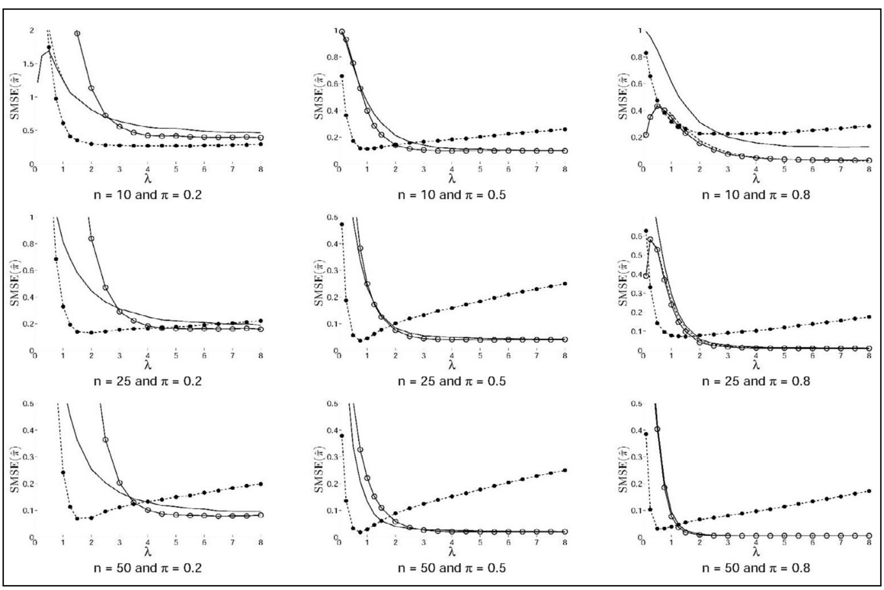


Figure 4.2. The simulated SMSE($\hat{\pi}$) of four estimators are plotted against λ for selected values of π and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

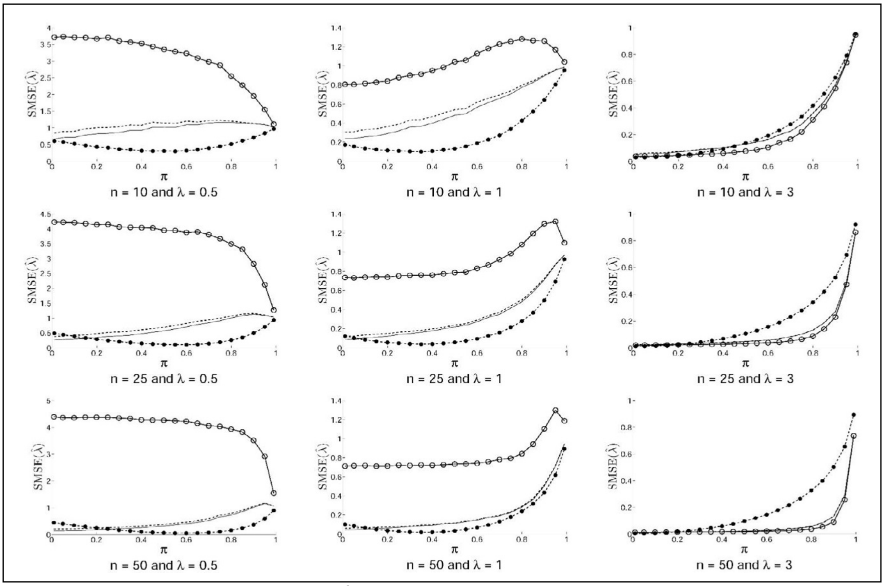


Figure 4.3. The simulated SMSE($\hat{\lambda}$) of four estimators are plotted against π for selected values of λ and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

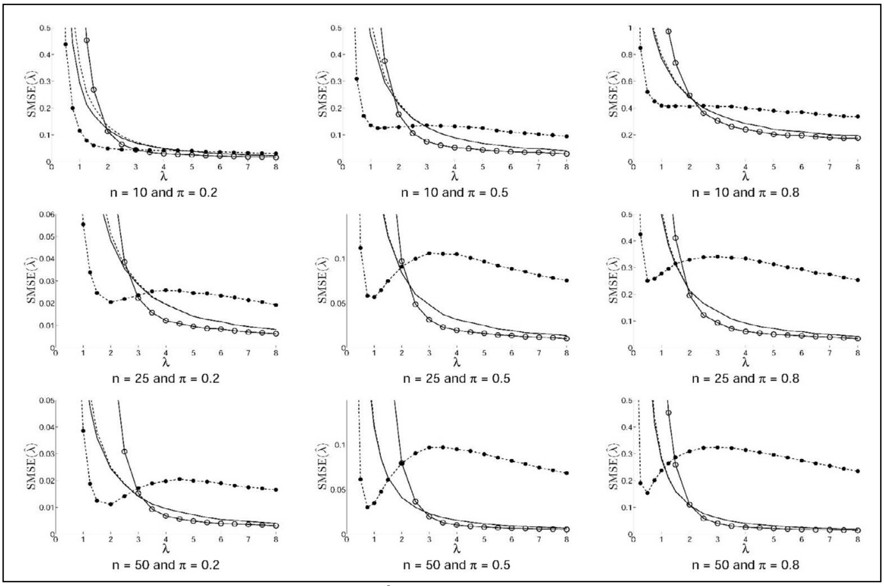


Figure 4.4. The simulated SMSE($\hat{\lambda}$) of four estimators are plotted against λ for selected values of π and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

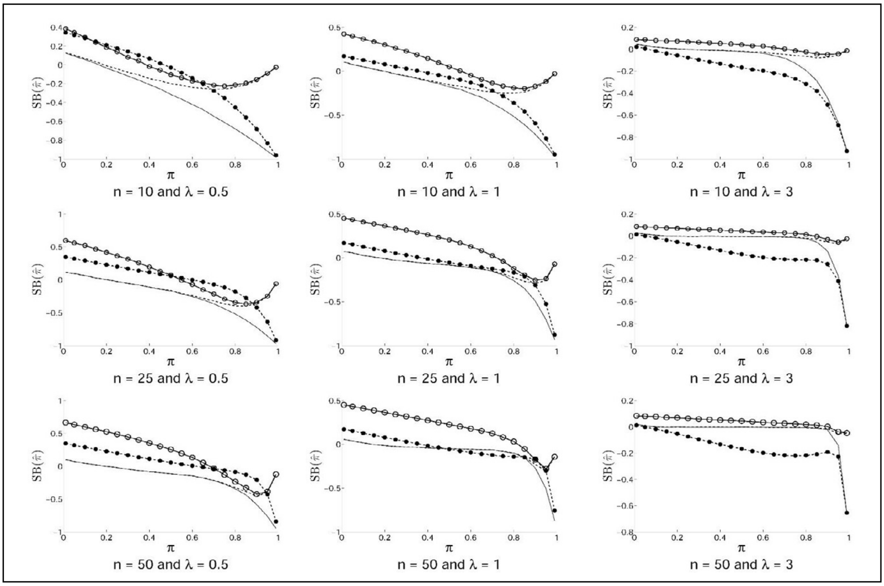


Figure 4.5. The simulated SB($\hat{\pi}$) of four estimators are plotted against π for selected values of λ and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

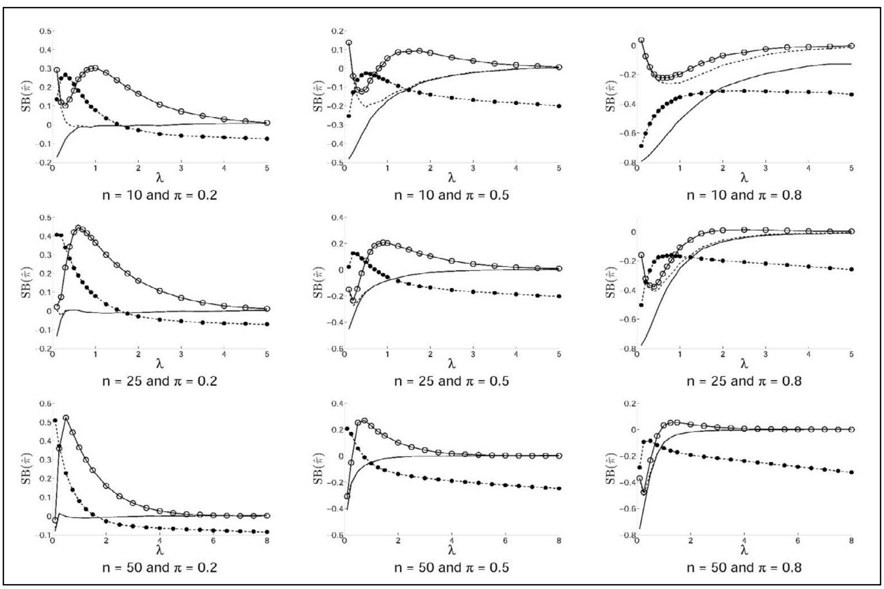


Figure 4.6. The simulated SB($\hat{\pi}$) of four estimators are plotted against λ for selected values of π and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

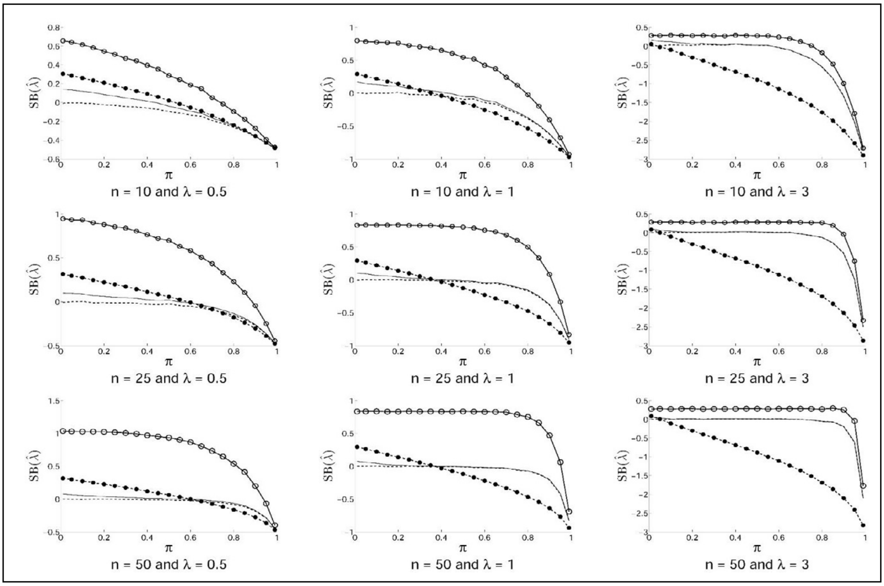


Figure 4.7. The simulated SB($\hat{\lambda}$) of four estimators are plotted against π for selected values of λ and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

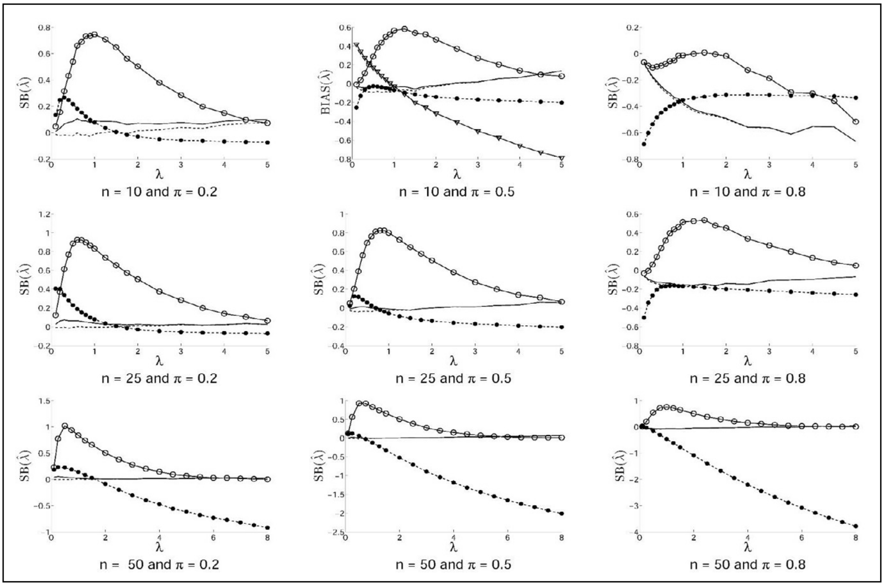


Figure 4.8. The simulated SB($\hat{\lambda}$) of four estimators are plotted against λ for selected values of π and n = 10, 25, 50. The solid line = MME1, dash line = MME2, solid line with empty bullets = MLE and dash line with solid bullets = AMLE.

5. Illustrative Examples

In this section all of four estimation methods, i.e., MME1, MME2, MLE, AMLE, are applied to estimate the parameters of ZIP(π,λ) to fit real-life datasets on rare events. The first example (Example 5.1) is the application of ZIP(π,λ) to fit a public health data in Thailand. In Example 5.2, the ZIP distribution is used to model the number of major Earthquakes in Thailand with magnitude greater than 6.0 Richter scale.

For each estimation technique we also use the goodness of fit test (GFT) to see how well the estimated parameters provide the model fitting to the dataset. Given the observations $X_1, X_2,..., X_n$, we note that observed frequencies (O_i' s) for the cells 0, 1, 2, ..., ($k-1$), where k is the first cell with zero frequency. In other words, O_i = empirical frequency of the cell-i, $0 \leq i \leq (k\text{-}1)$, such that $O_k = 0$, or $O_{(k\text{-}1)}$ = (number of X_i's $\geq (k - 1)$). The expected frequency of cell-i (i.e., E_i) is found based on the fitted model ZIP($\hat{\pi},\hat{\lambda}$), where ($\hat{\pi},\hat{\lambda}$) is found by one of the four techniques discussed earlier. The GFT statistic (Δ) is $\Delta = \sum_{i=1}^{k} \left(O_i - E_i\right)^2 / E_i$

In this section we use the GFT statistic to test the null hypothesis H_0: ZIP(π, λ) fits the given data, against the alternative H_A: ZIP(π, λ) does not fit the given data. The p-values are calculated accordingly using the approximate Chi-square distribution with ($k-3$) df.

Example 5.1 Revisit Example 1.1 The empirical and all fitted ZIP($\hat{\pi},\hat{\lambda}$) relative frequency histograms are plotted in Figure 5.1. Table 5.1 provides the estimated parameters of each estimation methods, the value of Δ and the GFT p-value. The bar diagrams in Figure 5.1 show that the AMLE method gives the best fit for this dataset. This is also seen in Table 5.1, where ZIP distribution with all estimation methods are providing good fit of the number of Elephantiasis patients. However, AMLE is providing the highest p-value.

Remark 5.1 It is interesting to note that though all the four estimation techniques make the ZIP distribution suitable at 5% level to fit the dataset in Example 5.1, the goodness of fit statistics and the corresponding p-values are quite different. The AMLE approach appears to provide the best fit of ZIP distribution for the given data followed by both MME1 and MME2 (though they are identical for this dataset which satisfies the set E_2 in (2.4)). The MLE approach appears to provide the worst fit for ZIP. Also note that using the AMLE approach ZIP(π,λ) fits the Elephantiasis data far better than the regular Poisson(λ) model as discussed in Section 1.

Table 5.1. The estimated parameters of four estimation methods, GFT values and *p*-values for fitting data in Table 1.2 with ZIP

Method	$(\hat{\pi}, \hat{\lambda})$	Observed Δ	p-value	Conclusion
MME1	(0.4851, 1.1477)	0.2922	0.8641	ZIP is suitable
MME2	(0.4851, 1.1477)	0.2922	0.8641	ZIP is suitable
MLE	(0.6847, 1.8744)	5.0060	0.0818	ZIP is suitable
AMLE	(0.4014, 0.9727)	0.0438	0.9783	ZIP is suitable

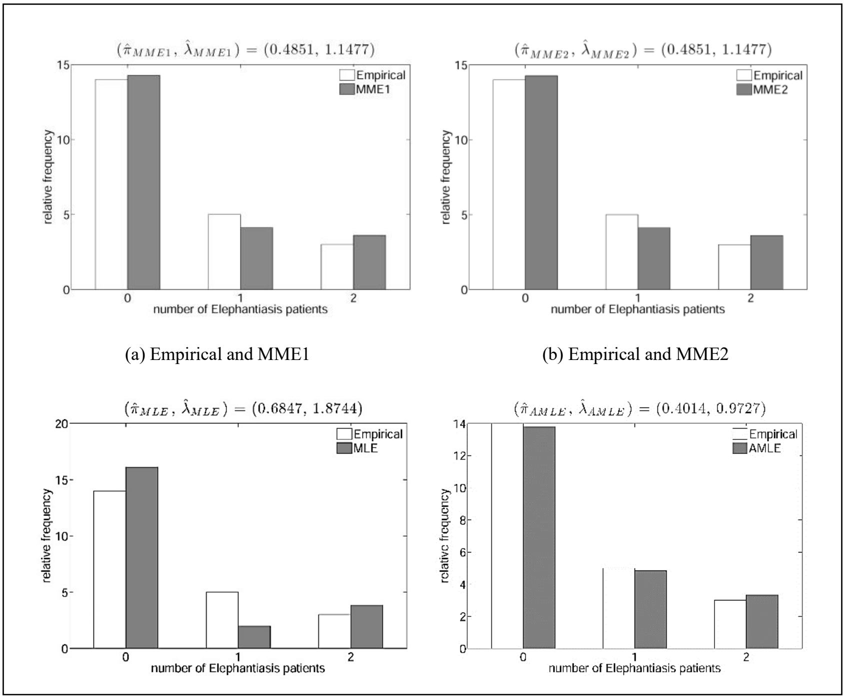


Figure 5.1. The empirical relative frequency of the number of elephantiasis patients of Phuket province (Thailand) is plotted with four fitted ZIP (*π*,*λ*) models.

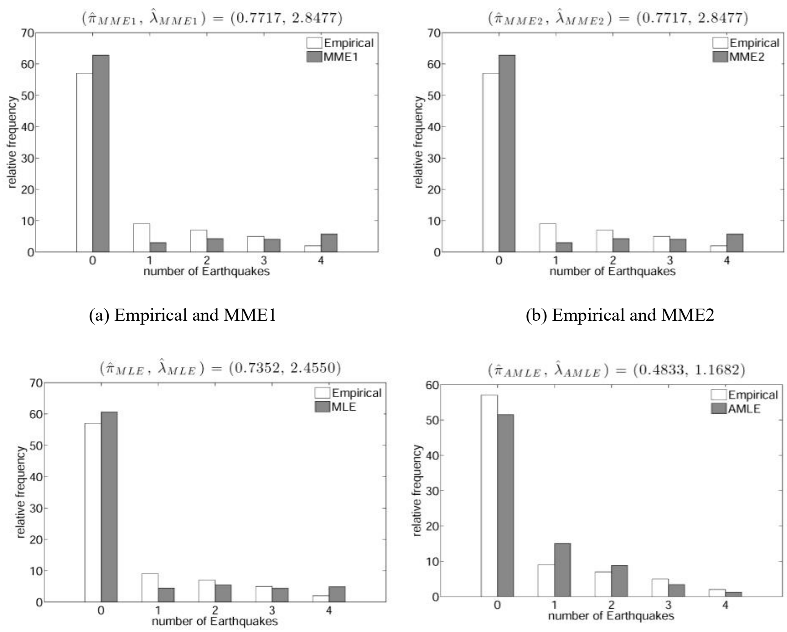
Example 5.2 The number of earthquakes per year witnessed by Thailand is reported by the Seismological Bureau, Thai Meteorological Department (Royal Thai Government), 2015. The variable of interest is the number of major earthquakes (those with magnitude at least 6.0 on the Richter scale) from 1933 to 2012. The observed yearly frequencies of major Thailand earthquakes are given in Table 5.2. Figure 5.2 shows the bar diagrams of the sample and four fitted ZIP($\hat{\pi}, \hat{\lambda}$) models. Table 5.3 provides the estimated parameters from four estimation techniques, the GFT statistic value of each technique and the corresponding p-value.

Table 5.2. Frequency distribution of major earthquakes in Thailand from 1933 to 2012

Number of major Earthquakes per year	**0**	**1**	**2**	**3**	**≥ 4**	**Total**
Frequency	57	9	7	5	2	80

Table 5.3. The Estimated parameters of four estimation methods, GFT values and *p*-values for fitting data in Table 5.2 with ZIP.

Method	$(\hat{\pi}, \hat{\lambda})$	Observed Δ	p-value	Conclusion
MME1	(0.7717, 2.8477)	16.8406	0.0021	ZIP is not suitable
MME2	(0.7717, 2.8477)	16.8406	0.0021	ZIP is not suitable
MLE	(0.7352, 2.4550)	7.0480	0.1334	ZIP is suitable
AMLE	(0.4833, 1.1682)	4.4842	0.3444	ZIP is suitable



(a) Empirical and MME1 (b) Empirical and MME2

(c) Empirical and MLE (d) Empirical and AMLE

Figure 5.2. The empirical relative frequency of the number of major earthquakes in Thailand is plotted with four fitted ZIP(π, λ) models.

Remark 5.2 For the earthquake dataset, it is interesting to see that while MME1 and MME2 (the same estimation for this dataset which satisfies E_2 in (2.4)) appear to reject the ZIP model at 5% level, the MLE and AMLE techniques tend to retain the ZIP model. Also, AMLE appears to provide the best fit of ZIP to the earthquake data. At the same time, if one tries to fit a regular Poisson(λ) model, then the p-value would be 2×10^{-7} making it unsuitable to fit the earthquake data.

ACKNOWLEDGEMENTS

The research of the first author (Suntaree Unhapipat) is supported by the Development and Promotion of Science and Technology Talent (DPST) Project of the Royal Thai Government. She is indebted to the Royal Thai Government for generous financial support to undertake her research work at the University of Louisiana at Lafayette, USA. The authors are thankful to an anonymous referee for many helpful and constructive comments which have led to a better presentation of materials in this paper.

REFERENCES

ANDERSON, D. R., SWEENEY, D. J., and WILLIAMS, T. A., 2012, *Essentials of Modern Business Statistics with Microsoft Excel*. Mason, OH: South-Western, Cengage Learning.

BECKETT S., JEE J., NCUBE T., POMPILUS S., WASHINGTON Q., SINGH A. and PAL N., 2014, Zero-inflated Poisson (ZIP) distribution: parameter estimation and applications to model data from natural calamities. *INVOLVE - A Journal of Mathematics* 7(6): 751-767.

BÖHNING D., DIETZ E., SCHLATTMANN P., MENDONCA L. and KIRCHNER U., 1999, The zero inflated Poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. 162(2): 195-209.

DAVIDSON C. L., 2012, A model selection paradigm for modeling recurrent adenoma data in polyp prevention trials, Master's thesis, University of Arizona, retrieved from http://arizona.openrepository.com/arizona/handle/10150/228465.

DOANE D., and SEWARD L., 2010, *Applied Statistics in Business and Economics*, 3rd Edition, Mcgraw-Hill.

GHOSH S. K., MUKHOPADHYAY P. and LU J.C. 2006. Bayesian analysis of zero inflated regression models, *Journal of Statistical Planning and Inference*, 136: 1360 1375.

LEVINE, D. M., STEPHAN, D.F., KREHBIEL, T.C., and BERENSON, M.L., 2011, *Essentials of Modern Business Statistics with Microsoft Excel*. Mason, OH: South-Western, Cengage Learning.

LORD D., WASHINGTON S. P. and IVAN J. N., 2005, Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory, *Accident Analysis and Prevention*, 37: 35 46.

THE BUREAU OF VECTOR-BORNE DISEASES, DEPARTMENT OF DISEASE CONTROL, MINISTRY OF PUBLIC HEALTH (ROYAL GOVERNMENT OF THAILAND), 2015, retrieved from http://www.thaivbd.org/n/.

THE SEISMOLOGICAL BUREAU THAI METEOROLOGICAL DEPARTMENT (ROYAL THAI GOVERNMENT), 2015, retrieved from http://www.seismology.tmd.go.th/home.html.

XIE. M., HE. B., and GOH T.N., 2001, Zero-inflated Poisson model in statistical process control. *Computational Statistics and Data Analysis* 38: 191 - 201.

Appendix

Table A.1. The approximated probability of E_1 and E_2 for $n = 10$.

P(·)	π	λ						
		0.5000	1.0000	2.0000	5.0000	10.0000	15.0000	20.0000
	0.1000	0.9863	0.9995	1.0000	1.0000	1.0000	1.0000	1.0000
	0.3000	0.9599	0.9969	1.0000	1.0000	1.0000	1.0000	1.0000
E_1	0.5000	0.8905	0.9775	0.9958	0.9987	0.9986	0.9900	0.9987
	0.7000	0.7217	0.8767	0.9478	0.9696	0.9722	0.9705	0.9724
	0.9000	0.3325	0.4709	0.5993	0.6450	0.6430	0.6428	0.6493
	0.1000	0.3930	0.5076	0.5913	0.7642	0.7992	0.8100	0.8100
	0.3000	0.4773	0.6335	0.8417	0.9692	0.9872	0.9849	0.9832
E_2	0.5000	0.5413	0.6969	0.9295	0.9979	0.9980	0.9983	0.9981
	0.7000	0.5361	0.7057	0.8994	0.9692	0.9722	0.9705	0.9724
	0.9000	0.2972	0.4333	0.5819	0.6446	0.6430	0.6428	0.6493

Table A.2. The approximated probability of E_1 and E_2 for $n = 25$.

P(·)	π	λ						
		0.5000	1.0000	2.0000	5.0000	10.0000	15.0000	20.0000
	0.1000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.3000	0.9998	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
E_1	0.5000	0.9960	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
	0.7000	0.9553	0.9956	0.9994	1.0000	1.0000	0.9999	0.9998
	0.9000	0.6386	0.8040	0.8990	0.9278	0.9301	0.9272	0.9261
	0.1000	0.5010	0.5697	0.7210	0.8927	0.9494	0.9592	0.9623
	0.3000	0.5960	0.7903	0.9608	0.9985	0.9998	1.0000	0.9998
E_2	0.5000	0.6393	0.8883	0.9960	1.0000	1.0000	1.0000	1.0000
	0.7000	0.5864	0.8706	0.9931	0.9999	1.0000	0.9999	0.9998
	0.9000	0.4434	0.6540	0.8510	0.9266	0.9301	0.9272	0.9261

Table A.3. The probability of three cases of MLE computation for $n = 10$.

		0.5000	1.0000	2.0000	5.0000	10.0000	15.0000	20.0000
Case – (i)	0.1000	0.0126	0.0002	0.0000	0.0000	0.0000	0.0000	0.0000
	0.3000	0.0399	0.0029	0.0001	0.0000	0.0000	0.0000	0.0000
	0.5000	0.1118	0.0224	0.0035	0.0010	0.0010	0.0010	0.0010
	0.7000	0.2848	0.1221	0.0496	0.0291	0.0283	0.0282	0.0282
	0.9000	0.6694	0.5205	0.4048	0.3513	0.3487	0.3487	0.3487
Case – (ii)	0.1000	0.4162	0.0659	0.0005	0.0000	0.0000	0.0000	0.0000
	0.3000	0.4810	0.1264	0.0045	0.0000	0.0000	0.0000	0.0000
	0.5000	0.5185	0.2200	0.0260	0.0004	0.0000	0.0000	0.0000
	0.7000	0.4753	0.3157	0.0909	0.0045	0.0001	0.0000	0.0000
	0.9000	0.2440	0.2446	0.1373	0.0134	0.0002	0.0000	0.0000
Case – (iii)	0.1000	0.5712	0.9338	0.9995	1.0000	1.0000	1.0000	1.0000
	0.3000	0.4791	0.8707	0.9954	1.0000	1.0000	1.0000	1.0000
	0.5000	0.3697	0.7576	0.9705	0.9985	0.9990	0.9990	0.9990
	0.7000	0.2399	0.5622	0.8595	0.9665	0.9717	0.9718	0.9718
	0.9000	0.0866	0.2349	0.4579	0.6353	0.6511	0.6513	0.6513

Table A.4. The probability of three cases of MLE computation for $n = 25$.

		0.5000	1.0000	2.0000	5.0000	10.0000	15.0000	20.0000
Case – (i)	0.1000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.3000	0.0003	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	0.5000	0.0042	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000
	0.7000	0.0433	0.0052	0.0005	0.0001	0.0001	0.0001	0.0001
	0.9000	0.3666	0.1954	0.1043	0.0731	0.0718	0.0718	0.0718
Case – (ii)	0.1000	0.1204	0.0011	0.0000	0.0000	0.0000	0.0000	0.0000
	0.3000	0.1955	0.0060	0.0000	0.0000	0.0000	0.0000	0.0000
	0.5000	0.3113	0.0289	0.0001	0.0000	0.0000	0.0000	0.0000
	0.7000	0.4604	0.1216	0.0069	0.0001	0.0000	0.0000	0.0000
	0.9000	0.4307	0.3165	0.1121	0.0072	0.0001	0.0000	0.0000
Case – (iii)	0.1000	0.8796	0.9989	1.0000	1.0000	1.0000	1.0000	1.0000
	0.3000	0.8042	0.9940	1.0000	1.0000	1.0000	1.0000	1.0000
	0.5000	0.6846	0.9711	0.9999	1.0000	1.0000	1.0000	1.0000
	0.7000	0.4963	0.8732	0.9926	0.9998	0.9999	0.9999	0.9999
	0.9000	0.2027	0.4880	0.7837	0.9197	0.9281	0.9282	0.9282