# Parameter Estimation and Application of Generalized Inflated Geometric Distribution 

Avishek Mallick*<br>Department of Mathematics, Marshall University,<br>Huntington, WV, USA<br>mallicka@marshall.edu<br>Ram Joshi<br>Department of Mathematics \& Statistics, Texas Tech University,<br>Lubbock, Texas, USA<br>ram.d.joshi@ttu.edu

## TReceived TOT-DQXDV 201] <br> Accepted 0 [\$ XJXV201]


#### Abstract

A count data that have excess number of zeros, ones, twos or threes are commonplace in experimental studies. But these inflated frequencies at particular counts may lead to overdispersion and thus may cause difficulty in data analysis. So to get appropriate results from them and to overcome the possible anomalies in parameter estimation, we may need to consider suitable inflated distribution. Generally, Inflated Poisson or Inflated Negative Binomial distribution are the most commonly used for modeling and analyzing such data. Geometric distribution is a special case of Negative Binomial distribution. This work deals with parameter estimation of a Geometric distribution inflated at certain counts, which we called Generalized Inflated Geometric (GIG) distribution. Parameter estimation is done using method of moments, empirical probability generating function based method and maximum likelihood estimation approach. The three types of estimators are then compared using simulation studies and finally a Swedish fertility dataset was modeled using a GIG distribution.


Keywords: Standardized bias; Standardized mean squared error; Akaike's Information Criterion (AIC); Bayesian Information Criterion (BIC); Monte Carlo study.

2000 Mathematics Subject Classification: 62F10, 62P10.

## 1. Introduction

A random variable $X$ that counts the number of trials to obtain the $r^{t h}$ success in a series of independent and identical Bernoulli trials, is said to have a Negative Binomial distribution whose probability mass function (pmf) is given by

$$
\begin{equation*}
P(X=k)=P(k \mid p)=\binom{k-1}{r-1} p^{r}(1-p)^{k-r} \tag{1.1}
\end{equation*}
$$

[^0]This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).
where $r=1,2,3, \ldots ; k=r, r+1, \ldots$ and $p>0$.
The above distribution is also the "Generalized Power Series distribution" as mentioned in Johnson et al. (2005) [9]. Some writers, for instance Patil et al. (1984) [12], called this the "PólyaEggenberger distribution", as it arises as a limiting form of Eggenberger and Pólya's (1923) [5] urn model distribution. A special case of Negative Binomial Distribution is the Geometric distribution which can be defined in two different ways

Firstly, the probability distribution for a Geometric random variable X (where X being the number of independent and identical trials to get the first success) is given by

$$
P(X=k \mid p)= \begin{cases}p(1-p)^{k-1} & \text { if } k=1,2, \ldots  \tag{1.2}\\ 0 & \text { otherwise }\end{cases}
$$

However, instead of counting the number of trials, if the random variable $X$ counts the number of failures before the first success, then it will result in the second type of Geometric distribution which again is a special case of Negative Binomial distribution when $r=1$ (first success) and its pmf is given by

$$
P(X=k)=P(k \mid p)= \begin{cases}p(1-p)^{k} & \text { if } k=0,1,2, \ldots  \tag{1.3}\\ 0 & \text { otherwise }\end{cases}
$$

The support set of this random variable is $\{0,1,2, \ldots\}$, which makes it different from the distribution in (1.2). The above model in (1.3), henceforth referred to as "Geometric(p)" has mean $\frac{1-p}{p}$ and variance $\frac{1-p}{p^{2}}$ and is the only distribution with non-negative integer support that can be characterized by the "memoryless property" or the "Markovian property". Many other characterizations of this distribution can be found in Feller $(1968,1969)$ [6] [7]. The distribution occurs in many applications and some of them are indicated in the references below:

- The famous problem of Banach's match boxes (Feller (1968) [6]);
- The runs of one plant species with respect to another in transects through plant populations (Pielou $(1962,1963)$ [13] [14]);
- A ticket control problem (Jagers (1973) [8]);
- A survillance system for congenital malformations (Chen (1978) [4]);
- The number of tosses of a fair coin before the first head (success) appears;
- The number of drills in an area before observing the first productive well by an oil prospector (Wackerly et al. (2008) [15]).

The Geometric model in (1.3) which is widely used for modeling count data may be inadequate for dealing with overdispersed count data. One such instance is the abundance of zero counts in the data, and (1.3) may be an inefficient model for such cases due to the presence of heterogeneity, which usually results in undesired over dispersion. Therefore, to overcome this situation, i.e., to explain or capture such heterogeneity, we consider a 'two-mass distribution' by giving mass $\pi_{1}$ to 0 counts, and mass $\left(1-\pi_{1}\right)$ to the other class which follows Geometric(p). The result of such a 'mixture distribution' is called the 'Zero-Inflated Geometric' (ZIG) distribution with the probability
mass function

$$
P\left(k \mid p, \pi_{1}\right)= \begin{cases}\pi_{1}+\left(1-\pi_{1}\right) p & \text { if } k=0  \tag{1.4}\\ \left(1-\pi_{1}\right) P(k \mid p) & \text { if } k=1,2, \ldots\end{cases}
$$

where, $p>0$, and $P(k \mid p)$ is given in (1.3). However the mixing parameter $\pi$ is chosen such that $P(k=0) \in(0,1)$ in (1.4), i.e., it ranges over the interval $-\frac{p}{1-p}<\pi_{1}<1$. This allows the distribution to be well defined for certain negative values of $\pi_{1}$, depending on $p$. Although the mixing interpretation is lost when $\pi_{1}<0$, these values have a natural interpretation in terms of zero-deflation, relative to a Geometric(p) model. Correspondingly, $\pi_{1}>0$ can be regarded as zero inflation as discussed in Johnson et al. (2005) [9].

A further generalization of (1.4) can be obtained by inflating/deflating the Geometric distribution at several specific values. To be precise, if the discrete random variable $X$ is thought to have inflated probabilities at the values $k_{1}, \ldots ., k_{m} \in\{0,1,2, \ldots$.$\} , then the following general probability$ mass function can be considered:

$$
P\left(k \mid p, \pi_{i}, 1 \leq i \leq m\right)= \begin{cases}\pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) P(k \mid p) & \text { if } k=k_{1}, k_{2}, \ldots, k_{m}  \tag{1.5}\\ \left(1-\sum_{i=1}^{m} \pi_{i}\right) P(k \mid p) & \text { if } k \neq k_{i} ; 1 \leq i \leq m\end{cases}
$$

where $k=0,1,2, \ldots ; p>0$ and $\pi_{i}$ 's are chosen in such a way that $P\left(k_{i}\right) \in(0,1)$ for all $i=1,2, \ldots, m$ in (1.5). For the remainder of this work, we will refer to (1.5) as the Generalized Inflated Geometric (GIG) distribution which is the main focus of this work.

Table 1. Observed number of children (= count) per woman.

| Count | Frequency | Proportion |
| :---: | :---: | :---: |
| 0 | 114 | .097 |
| 1 | 205 | .175 |
| 2 | 466 | .398 |
| 3 | 242 | .207 |
| 4 | 85 | .073 |
| 5 | 35 | .030 |
| 6 | 16 | .014 |
| 7 | 4 | .003 |
| 8 | 1 | .001 |
| 10 | 1 | .001 |
| 12 | 1 | .001 |
| Total | 1,170 | 1.000 |

We will consider some special cases of the (GIG) distribution such as Zero-One-Inflated Geometric (ZOIG) distribution in the case $k=2$ with $k_{1}=0$ and $k_{2}=1$ or Zero-One-Two Inflated Geometric (ZOTIG) models. Similar type of Generalized Inflated Poisson (GIP) models have been considered by Melkersson and Rooth (2000) [11] to study a women's fertility data of 1170 Swedish women of the age group 46-76 years (Table 1). This data set consists of the number of child(ren) per woman, who have crossed the childbearing age in the year 1991. They justified the Zero-Two Inflated Poisson distribution was the best to model it. However recently, Begum et al. (2014) [2]
studied the same data set and found that a Zero-Two-Three Inflated Poisson (ZTTIP) distribution was a better fit.

Instead of using an Inflated Poisson model, we will consider fitting appropriate Inflated Geometric models to the data in Table 1. Now, which model is the best fit whether a GIG model with focus on counts ( 0,1 ), i.e., ZOIG or a GIG model focusing on some other set $\left\{k_{1}, k_{2}, \ldots, k_{m}\right\}$ is appropriate for the above data will be eventually decided by different model selection criteria in section 5 . The rest of the paper is organized as follows. In the next section, we provide some properties of the GIG distribution. In section 3, we discuss different techniques of parameter estimation namely, the method of moments (MME), empirical probability generating function (epgf) based methods (EPGE) suggested by Kemp and Kemp (1988) [10], and the maximum likelihood estimation (MLE). In section 4, we compare the performances of MMEs, EPGEs and MLEs for different GIG model parameters using simulation studies. Finally conclusion is given in section 6.

## 2. Some Properties

In this section, we present some properties of the GIG distribution and that of some special cases of this distribution.
Proposition 1. The probability generating function (pgf) of a random variable $X$ following a GIG in (1.5) with parameters $\pi_{1}, \ldots, \pi_{m}$ and $p$ is given by

$$
\begin{equation*}
G_{X}(s)=\sum_{i=1}^{m} s^{k_{i}} \pi_{i}+\frac{\left(1-\sum_{i=1}^{m} \pi_{i}\right) p}{1-\{(1-p) s\}} \quad \text { if } s<\frac{1}{1-p} \tag{2.1}
\end{equation*}
$$

Proof.

$$
\begin{aligned}
G_{X}(s) & =\sum_{k=0}^{\infty} s^{k} P\left(k \mid p, \pi_{i}, 1 \leq i \leq m\right) \\
& =\sum_{i=1}^{m} s^{k_{i}} \pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) \sum_{k=0}^{\infty} s^{k} P(k \mid p) \\
& =\sum_{i=1}^{m} s^{k_{i}} \pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) \sum_{k=0}^{\infty} p\{(1-p) s\}^{k} \\
& =\sum_{i=1}^{m} s^{k_{i}} \pi_{i}+\frac{\left(1-\sum_{i=1}^{m} \pi_{i}\right) p}{1-\{(1-p) s\}}
\end{aligned}
$$

The infinite sum converges, if $(1-p) s<1 \Leftrightarrow s<\frac{1}{1-p}$.
Proposition 2. The $r^{\text {th }}$ raw moment of a random variable $X$ following a GIG in (1.5) with parameters $\pi_{1}, \ldots, \pi_{m}$ and $p$ is given by

$$
\begin{equation*}
E\left[X^{r}\right]=\sum_{i=1}^{m} k_{i}^{r} \pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) \mu_{r}^{\prime}(p) \tag{2.2}
\end{equation*}
$$

where $\mu_{r}^{\prime}(p)$ is the the $r^{\text {th }}$ raw moment of Geometric $(\mathrm{p})$ distribution and can be calculated easily by differentiating its moment generating function (mgf) given by $\frac{p}{1-(1-p) e^{t}}$, i.e., $\mu_{r}^{\prime}(p)=$
$\left.\frac{d^{r}}{d t^{r}}\left(\frac{p}{1-(1-p) e^{t}}\right)\right|_{t=0}$.
Proof.

$$
\begin{aligned}
E\left[X^{r}\right] & =\sum_{k=0}^{\infty} k^{r} P\left(k \mid p, \pi_{i}, 1 \leq i \leq m\right) \\
& =\sum_{i=1}^{m} k_{i}^{r} \pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) \sum_{k=0}^{\infty} k^{r} P(k \mid p) \\
& =\sum_{i=1}^{m} k_{i}^{r} \pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) \mu_{r}^{\prime}(p)
\end{aligned}
$$

For the ZIG distribution, i.e., when $m=1$ and $k_{1}=0$, using (2.1) we get the pgf as

$$
G_{X}(s)=\pi_{1}+\frac{\left(1-\pi_{1}\right) p}{1-\{(1-p) s\}}
$$

Using (2.2) the population mean and population variance in this special case can be obtained as

$$
\begin{equation*}
E(X)=\left(1-\pi_{1}\right) \frac{(1-p)}{p} \text { and } \operatorname{Var}(X)=\left(1-\pi_{1}\right)\left\{1+\pi_{1}(1-p)\right\} \frac{(1-p)}{p^{2}} \tag{2.3}
\end{equation*}
$$

Note that $\partial E(X) / \partial \pi_{1}<0$ and $\partial E(X) / \partial p<0$, and therefore the mean decreases with both parameters $\pi_{1}$ and $p$. Table (2) shows the variance-to-mean ratio, i.e., $\operatorname{Var}(X) / E(X)$ for different values of the parameters $\pi_{1}$ and $p$. It can be observed that this ratio is increasing (decreasing) with the values of $\pi_{1}(p)$. Also it is always larger than 1 , and thus ZIG distribution seems to be overdispersed.

Table 2. $\operatorname{Var}(X) / E(X)$ for ZIG distribution

| $\pi_{1}$ |  |  | $p$ |  |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 |
| 0.1 | 10.90 | 5.40 | 3.57 | 2.65 | 2.10 | 1.73 | 1.47 | 1.28 | 1.12 |
| 0.2 | 11.80 | 5.80 | 3.80 | 2.80 | 2.20 | 1.80 | 1.51 | 1.30 | 1.13 |
| 0.3 | 12.70 | 6.20 | 4.03 | 2.95 | 2.30 | 1.87 | 1.56 | 1.33 | 1.14 |
| 0.4 | 13.60 | 6.60 | 4.27 | 3.10 | 2.40 | 1.93 | 1.60 | 1.35 | 1.16 |
| 0.5 | 14.50 | 7.00 | 4.50 | 3.25 | 2.50 | 2.00 | 1.64 | 1.38 | 1.17 |
| 0.6 | 15.40 | 7.40 | 4.73 | 3.40 | 2.60 | 2.07 | 1.69 | 1.40 | 1.18 |
| 0.7 | 16.30 | 7.80 | 4.97 | 3.55 | 2.70 | 2.13 | 1.73 | 1.43 | 1.19 |
| 0.8 | 17.20 | 8.20 | 5.20 | 3.70 | 2.80 | 2.20 | 1.77 | 1.45 | 1.20 |
| 0.9 | 18.10 | 8.60 | 5.43 | 3.85 | 2.90 | 2.27 | 1.81 | 1.48 | 1.21 |

The cumulative distribution function (cdf) of the ZIG distribution is given by

$$
F\left(k \mid p, \pi_{1}\right)= \begin{cases}0 & \text { if } k<0  \tag{2.4}\\ \pi_{1}+\left(1-\pi_{1}\right)\left\{1-(1-p)^{k+1}\right\} & \text { if } k=0,1,2, \ldots\end{cases}
$$

Thus, the survival function for the ZIG distribution is given by

$$
S(k)=P(X>k)= \begin{cases}1 & \text { if } k<0  \tag{2.5}\\ \left(1-\pi_{1}\right)(1-p)^{k+1} & \text { if } k=0,1,2, \ldots\end{cases}
$$

The failure rate (hazard function) for ZIG, using (1.4) and (2.5), is given by

$$
\begin{aligned}
r(k) & =P(X=k \mid X \geq k) \\
& =\frac{P(X=k)}{P(X>k-1)}=\frac{P(k)}{S(k-1)} \\
& = \begin{cases}\pi_{1}+\left(1-\pi_{1}\right) p & \text { if } k=0 \\
p & \text { if } k=1,2, \ldots\end{cases}
\end{aligned}
$$

which is a constant function of $X$, except at 0 .
Using (1.5), the probability mass function of 'Zero-One-Inflated Geometric' (ZOIG) distribution, i.e., when $m=2$ and $k_{1}=0$ and $k_{2}=1$, can be obtained as

$$
P\left(k \mid p, \pi_{1}, \pi_{2}\right)= \begin{cases}\pi_{1}+\left(1-\pi_{1}-\pi_{2}\right) p & \text { if } k=0  \tag{2.6}\\ \pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) p(1-p) & \text { if } k=1 \\ \left(1-\pi_{1}-\pi_{2}\right) P(k \mid p) & \text { if } k=2,3, \ldots\end{cases}
$$

From (2.6), the cdf of ZOIG can be easily obtained as

$$
F\left(k \mid p, \pi_{1}, \pi_{2}\right)= \begin{cases}0 & \text { if } k<0  \tag{2.7}\\ \pi_{1}+\left(1-\pi_{1}-\pi_{2}\right) p & \text { if } k=0 \\ \pi_{1}+\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right)\left\{1-(1-p)^{k+1}\right\} & \text { if } k=1,2, \ldots\end{cases}
$$

Thus, the survival function of ZOIG distribution is given by

$$
S(k)= \begin{cases}1 & \text { if } k<0  \tag{2.8}\\ \left(1-\pi_{1}\right)(1-p)+\pi_{2} p & \text { if } k=0 \\ \left(1-\pi_{1}-\pi_{2}\right)(1-p)^{k+1} & \text { if } k=1,2, \ldots\end{cases}
$$

Therefore, the hazard function becomes

$$
r(k)= \begin{cases}\pi_{1}+\left(1-\pi_{1}-\pi_{2}\right) p & \text { if } k=0  \tag{2.9}\\ \frac{\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) p(1-p)}{\left(1-\pi_{1}\right)(1-p)+\pi_{2} p} & \text { if } k=1 \\ p & \text { if } k=2,3, \ldots\end{cases}
$$

which is a constant function of $X$, except at 0 and 1 .
Also using (2.1), we obtain the pgf of ZOIG distribution as

$$
G_{X}(s)=\pi_{1}+s \pi_{2}+\frac{\left(1-\pi_{1}-\pi_{2}\right) p}{1-\{(1-p) s\}}
$$

## 3. Estimation of Model Parameters

In this section, we estimate the parameters by three well known methods of parameter estimation namely the method of moment estimations, methods based on empirical probability generating function and the method of maximum likelihood estimations.

### 3.1. Method of Moments Estimation (MME)

The easiest way to obtain estimators of the parameters is through the method of moments estimation (MME). Given a random sample $X_{1}, \ldots, X_{n}$, i.e. independent and identically distributed (iid) observations from the GIG distribution, we equate the sample moments with the corresponding population moments (2.2) to get a system of $(\mathrm{m}+1)$ equations involving the $(\mathrm{m}+1)$ model parameters $p, \pi_{1}, \ldots, \pi_{m}$ of the form

$$
\begin{equation*}
m_{r}^{\prime}=\sum_{i=1}^{m} k_{i}^{r} \pi_{i}+\left(1-\sum_{i=1}^{m} \pi_{i}\right) \mu_{r}^{\prime}(p) \tag{3.1}
\end{equation*}
$$

where $r=1,2, \ldots,(m+1)$. Note that $m_{r}^{\prime}=\sum_{j=1}^{n} X_{j}^{r} / n$ is the $r^{t h}$ raw sample moment. The values of $\pi_{i}, i=1,2, \ldots, m$, and $p$ obtained by solving the system of equations (3.1) are denoted by $\hat{\pi}_{i(M M)}$ and $\hat{p}_{M M}$ respectively. The subscript "(MM)" indicates the MME approach. Note that the parameter $p$ is non-negative and hence the estimate must obey this restriction. But there is no such guarantee, as such we propose the corrected MME's to ensure non-negativity of this moment estimator as

$$
\begin{equation*}
\hat{p}_{M M}^{c}=\hat{p}_{M M} \text { truncated at } 0 \text { and } 1 \text { and } \hat{\pi}_{i(M M)}^{c}=\hat{\pi}_{i(M M)} \tag{3.2}
\end{equation*}
$$

where $\hat{\pi}_{i(M M)}^{c}$ is the solution of $\pi_{i}$ in (3.1) obtained after substituting $\hat{p}_{M M}$.
Consider the case of ZIG distribution when $m=1$ i.e., $k_{1}=0$, resulting into only two parameters to estimate, i.e. $\pi_{1}$ and $p$. In order to obtain the method of moments estimators, we can equate the mean and variance given in (2.3) with sample mean $(\bar{X})$ and sample variance $\left(s^{2}\right)$ respectively. This is an alternative approach to estimate the parameters instead of dealing with the sample raw moments. Hence we obtain,

$$
\begin{array}{r}
\left(1-\pi_{1}\right) \frac{(1-p)}{p}=\bar{X} \\
\left(1-\pi_{1}\right)\left\{1+\pi_{1}(1-p)\right\} \frac{(1-p)}{p^{2}}=s^{2} \tag{3.3}
\end{array}
$$

Solving the above equations simultaneously for $p$ and $\pi_{1}$, we get, $\hat{p}_{M M}^{c}=\frac{2 \bar{X}}{\left(s^{2}+\bar{X}+\bar{X}^{2}\right)}$ and $\hat{\pi}_{1(M M)}^{c}=\frac{s^{2}-\bar{X}-\bar{X}^{2}}{s^{2}-\bar{X}+\bar{X}^{2}}$.

Now let us consider another special case of GIG, the Zero-One Inflated Geometric (ZOIG) distribution, i.e., $m=2$ and $k_{1}=0$ and $k_{2}=1$. It has three parameters $\pi_{1}, \pi_{2}$ and $p$ and to estimate them we need to equate the first three raw sample moments with the corresponding population moments and we thus obtain the following system of equations:

$$
\begin{align*}
& m_{1}^{\prime}=\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) \frac{(1-p)}{p} \\
& m_{2}^{\prime}=\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) \frac{\left(p^{2}-3 p+2\right)}{p^{2}} \\
& m_{3}^{\prime}=\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) \frac{(1-p)\left(p^{2}-6 p+6\right)}{p^{3}} \tag{3.4}
\end{align*}
$$

Similarly, another special case of GIG is the Zero-One-Two Inflated Geometric (ZOTIG) distribution. Here we have four parameters $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ to estimate. This is done by solving the
following system of four equations obtained by equating the first four raw sample moments with their corresponding population moments to have,

$$
\begin{align*}
& m_{1}^{\prime}=\pi_{2}+2 \pi_{3}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) \frac{(1-p)}{p} \\
& m_{2}^{\prime}=\pi_{2}+4 \pi_{3}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) \frac{(1-p)(2-p)}{p^{2}} \\
& m_{3}^{\prime}=\pi_{2}+8 \pi_{3}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) \frac{(1-p)\left(p^{2}-6 p+6\right)}{p^{3}} \\
& m_{4}^{\prime}=\pi_{2}+16 \pi_{3}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) \frac{(1-p)(2-p)\left(p^{2}-12 p+12\right)}{p^{4}} \tag{3.5}
\end{align*}
$$

Algebraic solutions to these systems of equations in (3.4) and (3.5) are obtained by using Mathematica. We note that these solutions may not fall in the feasible regions of the parameter space, so we put appropriate restrictions to these solutions as discussed for the ZIG distribution to obtain the corrected MMEs.

### 3.2. Methods based on Probability Generating Function (EPGE)

Given a random sample $X_{1}, X_{2}, \ldots, X_{n}$ of size $n$ from a distribution with $\operatorname{pgf} G(s)$, the empirical probability generating function (epgf) is defined as

$$
\begin{equation*}
G_{n}(s)=\frac{1}{n} \sum_{i=1}^{n} s^{X_{i}} \tag{3.6}
\end{equation*}
$$

Number of parameter estimation methods have been proposed based on the the epgf, due to the fact that the epgf $G_{n}(s)$ converges pgf $G(s)$ in various modes as $n \rightarrow \infty$. The simplest of these method for estimating $\theta=\left(\theta_{1}, \theta_{2}, \ldots, \theta_{m}\right)$ is by equating the epgf to the pgf at $m$ different numerical values of $s$, and then solving the system of $m$ equations of the form

$$
\begin{equation*}
G_{n}\left(s_{i}\right)=G\left(s_{i}\right), i=1,2, \ldots m \tag{3.7}
\end{equation*}
$$

where, $-1 \leq s_{i} \leq 1$. Another system of estimating equations can be obtained and solved by equating the derivatives,

$$
\begin{equation*}
G_{n}^{\prime}\left(s_{i}\right)=G^{\prime}\left(s_{i}\right), i=1,2, \ldots m \tag{3.8}
\end{equation*}
$$

or by considering the derivatives of logs,

$$
\begin{equation*}
G_{n}^{\prime}\left(s_{i}\right) G\left(s_{i}\right)=G^{\prime}\left(s_{i}\right) G_{n}\left(s_{i}\right), i=1,2, \ldots m \tag{3.9}
\end{equation*}
$$

A combination of these three types of estimating equations can be used as well. Kemp and Kemp (1988) [10] showed that some well known estimation procedures, such as, the method of moments discussed earlier are special cases of these methods. For the case of ZIG, i.e. when $m=2$, we used one such well-known procedure, the method of mean-and-zero-frequency, where the sample mean is equated to the population mean and the relative frequency of $X=0$ in the sample $\left(f_{0}\right)$ is equated
to $P(X=0)$. Hence we obtain,

$$
\begin{align*}
& \bar{X}=\left(1-\pi_{1}\right) \frac{(1-p)}{p} \\
& f_{0}=\pi_{1}+\left(1-\pi_{1}\right) p . \tag{3.10}
\end{align*}
$$

Solving the above set of equations for $p$ and $\pi_{1}$, we get, $\hat{p}_{P G}=\frac{1-f_{0}}{\bar{X}}$ and $\hat{\pi}_{1(P G)}=\frac{\bar{X} f_{0}+f_{0}-1}{\bar{X}+f_{0}-1}$. The subscript "(PG)" indicates the EPGE approach. As in the case of the MME approach, the parameter estimates obtained using EPGE approach may lie outside the parameter space. In that case, like the MME approach we need to correct them by appropriately truncating them. We can extend this approach to any GIG distribution by equating the sample mean with the population mean and probability of the inflated points $P\left(X=k_{i}\right)$ with the relative frequencies of those in the sample $\left(f_{k_{i}}\right)$. For example, in case of ZOIG distribution, we obtain the following set of equations:

$$
\begin{align*}
& \bar{X}=\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) \frac{(1-p)}{p} \\
& f_{0}=\pi_{1}+\left(1-\pi_{1}-\pi_{2}\right) p \\
& f_{1}=\pi_{2}+\left(1-\pi_{1}-\pi_{2}\right) p(1-p) \tag{3.11}
\end{align*}
$$

and for ZOTIG distribution, we have

$$
\begin{align*}
& \bar{X}=\pi_{2}+2 \pi_{3}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) \frac{(1-p)}{p} \\
& f_{0}=\pi_{1}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) p \\
& f_{1}=\pi_{2}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) p(1-p) \\
& f_{2}=\pi_{3}+\left(1-\pi_{1}-\pi_{2}-\pi_{3}\right) p(1-p)^{2} \tag{3.12}
\end{align*}
$$

Again, algebraic solutions to these systems of equations in (3.11) and (3.12) are obtained by using Mathematica.

### 3.3. Maximum Likelihood Estimation (MLE)

Now we discuss the approach of estimating our parameters by the method of Maximum Likelihood Estimation(MLE). Based on the random sample $\mathbf{X}=\left(X_{1}, X_{2}, \ldots, X_{n}\right)$, we define the likelihood function $L=L\left(p, \pi_{i}, 1 \leq i \leq m \mid \mathbf{X}\right)$ as follows. Let $Y_{i}=$ the number of observations at $k_{i}$ with inflated probability, i.e., if $I$ is an indicator function, then $Y_{i}=\sum_{j=1}^{n} I\left(X_{j}=k_{i}\right), 1 \leq i \leq m$, which means $Y_{i}$ is the total number of observed counts at $k_{i}$. Also, let $Y .=\sum_{i=1}^{m} Y_{i}=$ total number of observations with inflated observations, $n=$ total number of observations and, $(n-Y$.) is the total number of non-inflated observations. Then,

$$
\begin{align*}
L & =\prod_{i=1}^{m}\left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}^{Y_{i}} \prod_{X_{j} \neq k_{i}}\left\{\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(X_{j} \mid p\right)\right\}  \tag{3.13}\\
& =\prod_{i=1}^{m}\left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}^{Y_{i}}\left(1-\sum_{l=1}^{m} \pi_{l}\right)^{(n-Y .)} \prod_{X_{j} \neq k_{i}} P\left(X_{j} \mid p\right)
\end{align*}
$$

The $\log$ likelihood function $l=\ln L$ is given by

$$
l=\sum_{i=1}^{m} Y_{i} \ln \left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}+(n-Y .) \ln \left(1-\sum_{l=1}^{m} \pi_{l}\right)+\sum_{X_{j} \neq k_{i}} \ln P\left(X_{j} \mid p\right)
$$

But we have,

$$
\sum_{X_{j} \neq k_{i}} \ln P\left(X_{j} \mid p\right)=(n-Y .) \ln p+\ln (1-p)\left(\sum_{j=1}^{n} X_{j}-\sum_{l=1}^{m} k_{l} Y_{l}\right)
$$

hence the log likelihood function becomes

$$
\begin{align*}
l & =\sum_{i=1}^{m} Y_{i} \ln \left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}+(n-Y .) \ln \left(1-\sum_{l=1}^{m} \pi_{l}\right)  \tag{3.14}\\
& +(n-Y .) \ln p+\ln (1-p)\left(\sum_{j=1}^{n} X_{j}-\sum_{l=1}^{m} k_{l} Y_{l}\right)
\end{align*}
$$

Now to obtain the MLEs, we maximize $l$ in (3.14) with respect to the parameters $\pi_{i}, 1 \leq i \leq m$, and $p$ over the appropriate parameter space. Differentiating $l$ partially w.r.t the parameters and setting them equal to zero yields the following system of likelihood equations or score equations.

$$
\begin{align*}
& \frac{\partial l}{\partial \pi_{i}}= \frac{Y_{i}}{\left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}}-\sum_{i=1}^{m} \frac{Y_{i} P\left(k_{i} \mid p\right)}{\left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}} \\
&-\frac{(n-Y .)}{\left(1-\sum_{l=1}^{m} \pi_{l}\right)}=0, \quad \forall i=1, \ldots, m  \tag{3.15}\\
& \frac{\partial l}{\partial p}= \sum_{i=1}^{m} \frac{Y_{i}\left(1-\sum_{l=1}^{m} \pi_{l}\right) P^{(p)}\left(k_{i} \mid p\right)}{\left\{\pi_{i}+\left(1-\sum_{l=1}^{m} \pi_{l}\right) P\left(k_{i} \mid p\right)\right\}}+\frac{(n-Y .)}{p}-\frac{\left(n \bar{X}-\sum_{l=1}^{m} k_{l} Y_{l}\right)}{(1-p)}=0
\end{align*}
$$

where, $P^{(p)}\left(k_{i} \mid p\right)=(\partial / \partial p) P\left(k_{i} \mid p\right)$.
Usually, of the three methods discussed, MLE is known to work best but in absence of closed form expression we have to resort to numerical optimization techniques, which makes it the most difficult one to obtain. Using the other two methods (MME and EPGE), estimators are much easier to obtain and do not require extensive computation, but they may lie outside the feasibility region of the parameter space. So we need to appropriately correct them. Therefore, we conduct simulation studies in the next section which can provide some guidance about their performances.

## 4. Simulation Study

We have considered the following three cases for our simulation study:
(i) $m=1, k_{1}=0$ (Zero Inflated Geometric (ZIG) distribution)
(ii) $m=2, k_{1}=0, k_{2}=1$ (Zero-One Inflated Geometric (ZOIG) distribution)
(iii) $m=3, k_{1}=0, k_{2}=1, k_{3}=2$ (Zero-One-Two Inflated Geometric (ZOTIG) distribution)

For each model mentioned above, we generate random data $X_{1}, \ldots, X_{n}$ from the distribution (with given parameter values) $N=10,000$ times. Let us denote a parameter (either $\pi_{i}$ or $p$ ) by the generic notation $\theta$. The parameter $\theta$ is estimated by three possible estimators $\hat{\theta}_{M M}^{(c)}$ (the corrected MME), $\hat{\theta}_{P G}^{(c)}$ (the corrected EPGE) and $\hat{\theta}_{M L}$ (the MLE). At the $l$ th replication, $1 \leq l \leq N$, the estimates of $\theta$
are $\hat{\theta}_{M M}^{(c)(l)}, \hat{\theta}_{P G}^{(c)(l)}$ and $\hat{\theta}_{M L}^{(l)}$ respectively. Then the standardized bias (called 'SBias') and standardized mean squared error (called 'SMSE') are defined and approximated as below

$$
\begin{array}{r}
\operatorname{SBias}(\hat{\theta})=E(\hat{\theta}-\theta) / \theta \approx\left\{\sum_{l=1}^{N}\left(\hat{\theta}^{(l)}-\theta\right) / \theta\right\} / N \\
\operatorname{SMSE}(\hat{\theta})=E(\hat{\theta}-\theta)^{2} / \theta^{2} \approx\left\{\sum_{l=1}^{N}\left(\hat{\theta}^{(l)}-\theta\right)^{2} / \theta^{2}\right\} / N \tag{4.1}
\end{array}
$$

Note that $\hat{\theta}$ will be replaced by $\hat{\theta}_{M M}^{(c)}, \hat{\theta}_{P G}^{(c)}$ and $\hat{\theta}_{M L}$ in our simulation study. Further observe that we are using SBias and SMSE instead of the actual Bias and MSE, because the standardized versions provide more information. An error of magnitude 0.01 in estimating a parameter with true value 1.00 is more severe than a situation where the parameter's true value is 10.0 . This fact is revealed through SBias and/or SMSE more than the actual bias and/or MSE.

### 4.1. The ZIG Distribution



Fig. 1. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}$ and $p$ (from ZIG distribution) plotted against $\pi_{1}$ for $p=0.2$ and $n=25$. The solid blue line represents the SBias or SMSE of the CMME. The dashed red line represents the SBias or SMSE of the EPGE and the longdashed black line is for the MLE. (a) Comparison of SBias of $\pi_{1}$ estimators. (b) Comparison of SBias of $p$ estimators. (c) Comparison of SMSE of $\pi_{1}$ estimators. (d) Comparison of SMSE of $p$ estimators.

In our simulation study for the Zero Inflated Geometric (ZIG) distribution, we fix $p=0.2$ and vary $\pi_{1}$ from 0.1 to 0.8 with an increment of 0.1 for $n=25,100$. The constrained optimization algorithm "L-BFGS-B" (Byrd et al. (1995)) [3] is implemented in R programming language to obtain the maximum likelihood estimators (MLEs) of the parameters $p$ and $\pi_{1}$, and the corrected MMEs and EPGEs are obtained by solving a system of equations and imposing appropriate restrictions on the parameters. In order to compare the performances of the MLEs with that of the CMMEs and EPGEs, we plot the standardized biases (SBias) and standardized MSE (SMSE) of these estimators obtained over the allowable range of $\pi_{1}$. The SBias and SMSE plots are presented in Figure 1 for sample size 25 . Due to space constraint, the plot for sample size 100 is presented in Figure 8 in the Appendix. Also we have presented in the Appendix a plot when $p=0.7$ and $n=100$ (Figure 9).

For the ZIG distribution with $n=25$, from Figure 1(a) we see that MLE and EPGE outperforms CMME for all values of $\pi_{1}$ with respect to SBias except at $\pi_{1}=0.1$. All the estimators beyond 0.2 are negatively biased. In Figure 1(b), we again see that MLE and EPGE uniformly outperforms the CMME and are increasing with values of $\pi_{1}$. In Figure 1(c), MLE and EPGE consistently outperforms CMME at all points w.r.t SMSE. For all the estimators, the SMSE starts off at their highest values and then decreases rapidly until it reaches nearly zero. The SMSE of MLE and EPGE consistently outperforms that of CMME in Figure 1(d). They both start off at their lowest values, and increase as values of $\pi_{1}$ get higher. For $n=100$, from Figure 8 we see a similar pattern. However, as expected the values of sBias and SMSE are considerably smaller for all estimators.

When $p=0.7$, from Figure 9 we observe that MLE outperforms both CMME and EPGE. Again all estimators of $\pi_{1}$ seems to be negatively biased and MLE and EPGE of $p$ is almost unbiased up to $\pi_{1}=0.7$. We also observe that SMSE of MLE for both the parameters $\pi_{1}$ and $p$ is smallest among the three estimators.

### 4.2. The ZOIG Distribution

In the case of the Zero-One Inflated Geometric (ZOIG) distribution we have three parameters to consider, namely $\pi_{1}, \pi_{2}$ and $p$. For fixed $p=0.3$ we vary $\pi_{1}$ and $\pi_{2}$ one at a time for sample sizes 25 and 100 . Figure 2 presents the comparisons for the three different type of estimators for three parameters $\pi_{1}, \pi_{2}$ and $p$ in terms of standardized bias and standardized MSE for $n=25$, varying $\pi_{1}$ from 0.1 to 0.5 and keeping $\pi_{2}$ and $p$ fixed at 0.15 and 0.3 respectively. Figure 10 in the Appendix presents the same for sample size 100. Also in the Appendix a plot when $p$ is fixed at 0.7 and $n=100$ is given in Figure 11. In Figure 2(a), both MLE and EPGE outperforms CMME at all points with respect to SBias. SBias of MLE and EPGE starts above zero and quickly becomes negative, whereas that of CMME is throughout negative. In Figure 2(b), we see that the MLE is almost unbiased for values of $\pi_{1}$ up to 0.4 , and SBias of both CMME and EPGE are always negative. Again in Figure 2(c), we see that MLE outperforms both CMME and EPGE. In Figure 2(d), SMSE of all the estimators are decreasing, but SMSE of MLE stays below that of other two estimators. In Figure 2(e), SMSE of MLE stays constant at 0.55 and in 2(f), SMSE of MLE stays below that of the other estimators.

For sample size $n=100$ from Figure 10, we observe that both MLE and EPGE for all the parameters are almost unbiased and performs better than CMME. However, SMSE wise MLE is marginally better. We also observe that with the increase in sample size the SBias and SMSE of all the estimators have gone down.


Fig. 2. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}$ and $p$ (from ZOIG distribution) by varying $\pi_{1}$ for fixed $\pi_{2}=0.15, p=0.3$ and $n=25$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(c) Comparisons of SBiases of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively. (d)-(f) Comparisons of SMSEs of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively.

When $p=0.7$ and $n=100$, from Figure 11 we observe that SBias of $\pi_{1}$ is decreasing for all the estimators. However for $\pi_{2}$, SBias of MLE starts of with very small positive value and then becomes negative, whereas both EPGE and CMME are negatively biased and MLE is almost unbiased for parameter $p$. Also SMSE wise, MLE seems to be the best estimator.

In our second scenario which is presented in Figure 3 for sample size 25, we vary $\pi_{2}$ keeping $\pi_{1}$ and $p$ fixed at 0.15 and 0.3 respectively. Figure 12 in the Appendix contains the same plots for sample size 100. Also in the Appendix we present comparison plots of SBias and SMSE in Figure 13 for sample size 100 , varying $\pi_{2}$ keeping $\pi_{1}$ and $p$ fixed at 0.15 and 0.7 respectively. In Figure 3(a), we see that all three estimators are negatively biased, but MLE is always performing better. However in Figure 3(b), MLE starts of with a positive bias and then becomes almost unbiased for $\pi_{2}$. SBias of EPGE is negative throughout but stays very close 0 . SBias of CMME stays throughout between -0.6 and -0.8. In Figure 3(c), we see that MLE outperforms both CMME and EPGE throughout with respect to SBias. From Figures 3(d), 3(e) and 3(f), it is clear that MLE outperforms both CMME
and EPGE with respect to SMSE for all permissible values of $\pi_{2}$. Thus we observe that the MLEs of the all three parameters perform better in terms of the both SBias and SMSE.


Fig. 3. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}$ and $p$ (from ZOIG distribution) by varying $\pi_{2}$ for fixed $\pi_{1}=0.15, p=0.3$ and $n=25$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(c) Comparisons of SBiases of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively. (d)-(f) Comparisons of SMSEs of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively.

We observe very similar behavior from the plots for $n=100$, presented in Figure 12. MLE seems to be performing best and SBias and SMSEs are going down as expected with an increase in sample size except for SMSE of $\pi_{2}$ (Figure 12(e)).

Again from Figure 13, when $p=0.7$ and $n=100$, we observe that MLE is performing better than EPGE and CMME w.r.t both SBias and SMSE.

### 4.3. The ZOTIG Distribution

For the Zero-One-Two Inflated Geometric (ZOTIG) distribution we have four parameters to consider, namely $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$. For fixed $p=0.3$ we vary $\pi_{1}, \pi_{2}$, and $\pi_{3}$ one at a time from 0.1


Fig. 4. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ (from ZOTIG distribution) by varying $\pi_{1}$ for fixed $\pi_{2}=\pi_{3}=0.2$ and $p=0.3$ and $n=25$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(d) Comparisons of SBiases of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively. (e)-(h) Comparisons of SMSEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively.
to 0.5 for sample size $n=25$. Thus we have 12 estimators $\hat{\pi}_{1(M M)}^{(c)}, \hat{\pi}_{2(M M)}^{(c)}, \hat{\pi}_{3(M M)}^{(c)}, \hat{p}_{(M M)}^{(c)}, \hat{\pi}_{1(P G)}$, $\hat{\pi}_{2(P G)}, \hat{\pi}_{3(P G)}, \hat{p}_{(P G)}, \hat{\pi}_{1(M L)}, \hat{\pi}_{2(M L)}, \hat{\pi}_{3(M L)}$ and $\hat{p}_{(M L)}$. The comparisons of these estimators in terms of standardized bias and standardize MSE for sample size $n=25$ are presented in Figures 4-6 and for sample size 100 are presented in the Appendix (Figures 14-16).

In the first scenario of ZOTIG distribution, which is presented in Figure 4, we vary $\pi_{1}$ keeping $\pi_{2}, \pi_{3}$ and $p$ fixed at $0.2,0.2$ and 0.3 respectively. From Figure $4(a, b, c, d)$, we see that the CMMEs of all the four parameters perform consistently worse than the MLEs and EPGEs with respect to SBias. Both MLEs and EPGEs seems to have very similar performance. Also from parts (e, f, g, h) in Figure 4 concerning the SMSE, we notice that the MLEs of all the parameters perform
consistently better than the other two type of estimators. We observe similar performance of the three types of estimators, when sample size is increased to 100 (Figure 14).


Fig. 5. Plots of the SBias and SMSE of the CMMEs and MLEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ (from ZOTIG distribution) by varying $\pi_{2}$ for fixed $\pi_{1}=\pi_{3}=0.2$ and $p=0.3$ and $n=25$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(d) Comparisons of SBiases of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively. (e)-(h) Comparisons of SMSEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively.

In our second scenario which is presented in Figure 5, we vary $\pi_{2}$ keeping $\pi_{1}, \pi_{3}$ and $p$ fixed at $0.2,0.2$ and 0.3 respectively. We observe some interesting things in these plots. From Figure 5 we see that MLE and EPGE of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ are performing better upto $\pi_{2}=0.4$ with respect to SBias but after that performance of all the estimators are same. However we see that the MLEs of all four parameters perform better than their CMME and EPGE counterparts with respect to SMSE.

In Figure 15, when sample size is 100 , we observe that CMMEs are consistently worse w.r.t both SBias and SMSE and EPGEs are performing marginally better than MLEs.


Fig. 6. Plots of the SBias and SMSE of the CMMEs and MLEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ (from ZOTIG distribution) by varying $\pi_{3}$ for fixed $\pi_{1}=\pi_{2}=0.2$ and $p=0.3$ and $n=25$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(d) Comparisons of absolute SBiases of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively. (e)-(h) Comparisons of SMSEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively.

In the third scenario which is presented in Figure 6, we vary $\pi_{3}$ keeping $\pi_{1}, \pi_{2}$ and $p$ fixed at $0.2,0.2$ and 0.3 respectively. Here also we observe similar results as the second case of ZOTIG distribution, MLEs and EPGEs for all the four parameters are uniformly outperforming CMMEs with respect to SBias up to $\pi_{3}=0.4$. Also as before MLEs uniformly outperform CMMEs and EPGEs of all the four parameters with respect to SMSE. But as sample size increases to 100, we observe that EPGEs are performing marginally better than MLEs (Figure 16).

Thus from our simulation study it is evident that MLE has an overall better performance than CMME and EPGE for all the Generalized Inflated Geometric models that we have considered. So in the next chapter, we consider an example where we fit an appropriate GIG model to the Swedish fertility data set.

## 5. Application of GIG Distribution

In this section, we consider the Swedish fertility data presented in Table 1 and try to fit a suitable GIG model. Since our simulation study in the previous section suggests that the MLE has an overall better performance, all of our estimations of model parameters are carried out using the maximum likelihood estimation approach. While fitting the Inflated Geometric models, parameter estimates of some mixing proportions $\left(\pi_{i}\right)$ came out negative. So we need to make sure that all the estimated probabilities according to the fitted models are non-negative. We tried all possible combinations of GIG models and then compared each of these Inflated Geometric models using the Chi-square goodness of fit test, the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC). While performing the Chi-square goodness of fit test, the last three categories of Table 1 are collapsed into one group due to small frequencies. More details of our model fitting is presented below.

First, we try with single-point inflation at each of the four values ( $0,1,2$ and 3 ). In this first phase, an inflation at 2 seems most plausible as it gives the smallest AIC and BIC values 4188.794 and 4198.924 respectively. However the p-value of the Chi-square test is very close to 0 , suggesting that this is not a good model. Next, we try two-point inflations at $\{0,1\},\{0,2\},\{0,3\},\{1,2\}$, etc. At this stage, $\{2,3\}$ inflation seems most appropriate going by the values of AIC (3947.064) and BIC ( 3962.258 ). But p -value of the Chi-square test close to 0 again makes it an inefficient model.

In the next stage, we try three-point inflation models, and here we note that a GIG with inflation set $\{0,2,3\}$ significantly improves over the earlier $\{2,3\}$ inflation model (i.e., TTIG). This ZTTIG model significantly improves the p -value (but still close to 0 ) while maintaining a low AIC and BIC of 3841.169 and 3861.428 respectively. The main reason for low p -value is that this model is unable to capture the tail behavior. The estimated value of the parameters are (with $k_{1}=0, k_{2}=2, k_{3}=3$ ): $\hat{\pi}_{1}=-0.2340517, \hat{\pi}_{2}=0.2816816, \hat{\pi}_{3}=0.1376353$, and $\hat{p}=0.4068477$ using the Maximum Likelihood Estimation approach. Interpretation of the negative $\hat{\pi}_{1}$ is very difficult in this case, perhaps it can be thought of as a deflation point.

Finally, we fitted the full $\{0,1,2,3\}$ inflated model. We obtained the maximum likelihood estimates of the model parameters as $\hat{\pi}_{1}=-2.60853103, \hat{\pi}_{2}=-0.92017233, \hat{\pi}_{3}=-0.04498528, \hat{\pi}_{3}$ $=0.02737037$, and $\hat{p}=0.59520603$. The AIC and the BIC values for this model are 3800.688 and 3826.012 respectively. Which is significantly lower than all the previous models. Also p-value of the Chi-square test is 0.810944 , thus rendering this Zero-One-Two-Three inflated Geometric (ZOTTIG) model to be a very good fit. For the sake of completeness, we have also fit a discrete Lindley distribution, proposed by Bakouch et al. (2014) [1], whose probability mass function is given in the Appendix (A.1), to the data and obtained the parameter estimate as $\hat{\theta}=0.614$. Also the AIC and BIC are 4353.098 and 4358.163 respectively. Moreover, $p$-value of the Chi-square test is almost 0 . Recently, this data set was also modeled using a Generalized Inflated Poisson distribution. It was observed that the Zero-Two-Three inflated Poisson (ZTTIP) model is giving best fit to the data with an AIC and BIC value of 3805.485 and 3769.226 respectively. Also the Chi-square test produced a test statistic and p-value of 2.085 and 0.720 respectively. Though both ZOTTIG and

ZTTIP models seem to be performing reasonably well for this data set, but based on the AIC and the Chi-square test, the ZOTTIG model seems to be a better fit. We have included a plot of regular geometric model, discrete Lindley model and the ZOTTIG model in Figure 7. It is evident from this plot that the ZOTTIG model is performing way better than the regular geometric and discrete Lindley model for the Swedish fertility data.


Fig. 7. Plot of the observed frequencies compared to the estimated frequencies from the regular Geometric, discrete Lindley and the Zero-One-Two-Three Inflated Geometric model.

## 6. Conclusion

This work deals with a general inflated geometric distribution (GIG) which can be thought of as a generalization of the regular Geometric distribution. This type of distribution can effectively model datasets with elevated counts. We have discussed some properties of this distribution and also outlined the parameter estimation procedure for this distribution using the method of moments estimation, empirical probability generating function bases methods and the maximum likelihood estimation techniques. Simulation studies were also performed and we found that MLEs performed better than the corrected MMEs and EPGEs in estimating the model parameters with respect to the standardized bias (SBias) and standardized mean squared errors (SMSE). While performing the simulation, we observed that for certain ranges of the inflated proportions in the GIG models, the MMEs and EPGEs lie outside the parameter space. Nonetheless, we selected all permissible values and compared the overall performance of the CMMEs, EPGEs and MLEs for three special cases of GIG. Different GIG models were used for analyzing the fertility data of Swedish women and was compared with a discrete Lindley model. It was found that the Zero-One-Two-Three inflated

Geometric (ZOTTIG) model is a good fit. Because of the extra parameter(s), the GIG distribution seems to be much more flexible in model fitting than the regular geometric distribution.

## Acknowledgment

The authors are grateful to the editor and two referees for providing valuable comments and suggestions, which enhanced this work substantially.

## References

[1] H. S. Bakouch, M. A. Jazi, and S. Nadarajah, A new discrete distribution, Statistics 48 (2014), no. 1, 200-240.
[2] M. Begum, A. Mallick, and N. Pal, A Generalized Inflated Poisson Distribution with Application to Modeling Fertility Data, Thailand Statistician 12 (2014), 135-159.
[3] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, A limited memory algorithm for bound constrained optimization, SIAM Journal on Scientific Computing 16 (1995), no. 5, 1190-1208.
[4] R. Chen, A survillance system for congenital malformations, Journal of the American Statistical Association 73 (1978), no. 362, 323-327.
[5] F. Eggenberger and G. Pólya, Über die statistik verketetter vorgäge, Zeitschrift für Angewandte Mathematik und Mechanik 1 (1923), 279-289 (German).
[6] W. Feller, An introduction to probability theory and its applications, 3rd ed., vol. 1, John Wiley \& Sons, Inc, New York, 1968.
[7] W. Feller, An introduction to probability theory and its applications, 3rd ed., vol. 2, John Wiley \& Sons, Inc, New York, 1971.
[8] P. Jagers, How many people pay their tram fares?, Journal of the American Statistical Association 68 (1973), no. 344, 801-804.
[9] N.L. Johnson, S. Kotz, and A.W. Kemp, Univariate discrete distributions, 3rd ed., John Wiley \& Sons, Inc, Hoboken, New Jersey, 2005.
[10] C.D. Kemp and A.W. Kemp, Rapid Estimation for Discrete Distributions, The Statistician 37 (1988),no. 3, 243-255.
[11] M. Melkersson and D.O. Rooth, Modeling female fertility using inflated count data models, Journal of Population Economics 13 (2000), no. 2, 189-203 (English).
[12] G.P. Patil, M.T. Boswell, S.W. Joshi, and M.V. Ratnaparkhi, Dictionary and classified bibliography of statistical distributions in scientific work: Discrete models, vol. 1, International Co-operative Publishing House, 1984.
[13] E.C. Pielou, Runs of one species with respect to another in transects through plant populations, Biometrics 18 (1962), no. 4, 579-593.
[14] E.C. Pielou, Runs of healthy and diseased trees in transects through an infected forest, Biometrics (1963), 603-614.
[15] D.D. Wackerly, W. Mendenhall, and R.L. Scheaffer, Mathematical statistics with applications, 7th ed., Cengage Learning, 2008.

## Appendix A.

The probability mass function of the discrete Lindley distribution is given by

$$
P(X=k \mid \theta)=p(k)= \begin{cases}\frac{e^{-\theta k}}{1+\theta}\left\{\theta\left(1-2 e^{-\theta}\right)+\left(1-e^{-\theta}\right)(1+\theta k)\right\} & \text { if } k=0,1,2, \ldots  \tag{A.1}\\ 0 & \text { otherwise }\end{cases}
$$

where $\theta>0$.


Fig. 8. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}$ and $p$ (from ZIG distribution) plotted against $\pi_{1}$ for $p=0.2$ and $n=100$. The solid blue line represents the SBias or SMSE of the CMME. The dashed red line represents the SBias or SMSE of the EPGE and the longdashed black line is for the MLE. (a) Comparison of SBias of $\pi_{1}$ estimators. (b) Comparison of SBias of $p$ estimators. (c) Comparison of SMSE of $\pi_{1}$ estimators. (d) Comparison of SMSE of $p$ estimators.


Fig. 9. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}$ and $p$ (from ZIG distribution) plotted against $\pi_{1}$ for $p=0.7$ and $n=100$. The solid blue line represents the SBias or SMSE of the CMME. The dashed red line represents the SBias or SMSE of the EPGE and the longdashed black line is for the MLE. (a) Comparison of SBias of $\pi_{1}$ estimators. (b) Comparison of SBias of $p$ estimators. (c) Comparison of SMSE of $\pi_{1}$ estimators. (d) Comparison of SMSE of $p$ estimators.


Fig. 10. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}$ and $p$ (from ZOIG distribution) by varying $\pi_{1}$ for fixed $\pi_{2}=0.15, p=0.3$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(c) Comparisons of SBiases of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively. (d)-(f) Comparisons of SMSEs of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively.


Fig. 11. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}$ and $p$ (from ZOIG distribution) by varying $\pi_{1}$ for fixed $\pi_{2}=0.15, p=0.7$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(c) Comparisons of SBiases of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively. (d)-(f) Comparisons of SMSEs of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively.


Fig. 12. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}$ and $p$ (from ZOIG distribution) by varying $\pi_{2}$ for fixed $\pi_{1}=0.15, p=0.3$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(c) Comparisons of SBiases of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively. (d)-(f) Comparisons of SMSEs of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively.


Fig. 13. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}$ and $p$ (from ZOIG distribution) by varying $\pi_{2}$ for fixed $\pi_{1}=0.15, p=0.7$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(c) Comparisons of SBiases of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively. (d)-(f) Comparisons of SMSEs of $\pi_{1}, \pi_{2}$ and $p$ estimators respectively.


Fig. 14. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ (from ZOTIG distribution) by varying $\pi_{1}$ for fixed $\pi_{2}=\pi_{3}=0.2$ and $p=0.3$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(d) Comparisons of SBiases of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively. (e)-(h) Comparisons of SMSEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively.


Fig. 15. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ (from ZOTIG distribution) by varying $\pi_{2}$ for fixed $\pi_{1}=\pi_{3}=0.2$ and $p=0.3$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(d) Comparisons of SBiases of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively. (e)-(h) Comparisons of SMSEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively.


Fig. 16. Plots of the SBias and SMSE of the CMMEs, EPGEs and MLEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ (from ZOTIG distribution) by varying $\pi_{3}$ for fixed $\pi_{1}=\pi_{2}=0.2$ and $p=0.3$ and $n=100$. The solid blue line represents the SBias or SMSE of the corrected MME. The dashed red line and longdashed black line represents the SBias or SMSE of the EPGE and MLE respectively. (a)-(d) Comparisons of SBiases of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively. (e)-(h) Comparisons of SMSEs of $\pi_{1}, \pi_{2}, \pi_{3}$ and $p$ estimators respectively.


[^0]:    *Corresponding author.

