# MISCLASSIFIED SIZE-BIASED MODIFIED POWER SERIES DISTRIBUTION AND ITS APPLICATIONS

Anwar Hassan, Peer Bilal Ahmad, Srinagar

(Received August 24, 2006, revised April 3, 2008)

*Abstract.* A misclassified size-biased modified power series distribution (MSBMPSD) where some of the observations corresponding to $x = 2$ are misclassified as $x = 1$ with probability $\alpha$, is defined. We obtain its recurrence relations among ordinary, central and factorial moments and also for some of its particular cases like the size-biased generalized negative binomial (SBGNB) and the size-biased generalized Poisson (SBGP) distributions. We also discuss the effect of the misclassification on the variance for MSBMPSD and illustrate an example for size-biased generalized negative binomial distribution. Finally, an example is presented for the size-biased generalized Poisson distribution to illustrate the results, and a goodness of fit test is also done using the method of moments.

*Keywords*: misclassification, size-biased modified power series distribution, raw moments, central moments, factorial moments, variance ratio, inverted parabola, generalized Poisson, generalized negative binomial

*MSC 2000*: 62E15, 62E10, 60E05

## 1. Introduction

In certain experimental investigations involving discrete distributions external factors may induce a measurement error in the form of misclassification. For instance, a situation may arise where certain values are erroneously reported; such a situation termed as modified or misclassified has been studied by Cohen ([4], [5], [6]) for the Poisson and the binomial random variables, Jani and Shah [16] for modified power series distribution (MPSD) where some of the values of one are sometimes reported as zero, and recently by Patel and Patel ([18], [19]) in the case of generalized power series distribution (GPSD) and MPSD for a more general situation where sometimes the value $(c + 1)$ is reported erroneously as $c$.

Cohen [4] altered data from Bortkiewicz's [3] classical example on deaths from the kick of a horse in the Prussian Army, to illustrate the practical application of his results. He assumed that twenty of 200 given records which should have shown one death were in error by reporting no deaths. The same example was considered by Williford and Binghan [24].

In this paper, we are concerned with the situation where sometimes the value two is reported erroneously as one in relation to size-biased MPSD. As we know, weighted distributions arise when the observations generated from a stochastic process are not given equal chance of being recorded; instead, they are recorded according to some weight function. When the weight function depends on the lengths of the units of interest, the resulting distribution is called length biased. More generally, when the sampling mechanism selects units with probability proportional to some measure of the unit size, the resulting distribution is called size-biased. Such distributions arise in the life length and were studied by various authors (see [1], [21], [9], [10], [11], [12], [13], [14]).

Gupta [8] defined the MPSD with the probability function given by

$$P_1[X=x] = a(x)\frac{(g(\theta))^x}{f(\theta)}, \quad x \in T \tag{1.1}$$

where $a(x) > 0$ and $T$ is a subset of the set of non-negative integers, $g(\theta), f(\theta)$ are positive, finite and differentiable. In the case $g(\theta)$ is invertible it reduces to Patil's [20] GPSD and if in addition $T$ is the entire set of non-negative integers it reduces to the power series distribution (PSD) given by Noack [17]. The class of distributions (1.1) includes among others the GNB and GP distributions.

Gupta [8] obtained the mean ($\mu(\theta)$) and the variance ($\mu_2$) of (1.1) given by

$$\mu(\theta) = \frac{g(\theta)f'(\theta)}{f(\theta)g'(\theta)}, \tag{1.2}$$

$$\mu_2 = \frac{g(\theta)}{g'(\theta)}\frac{\partial\mu(\theta)}{\partial\theta}. \tag{1.3}$$

A size-biased MPSD is obtained by taking the weight of (1.1) as $X$, given by

$$P_2[X=x] = \frac{b(x)(g(\theta))^x}{f^*(\theta)} \tag{1.4}$$

where $b(x) = xa(x)$ and $f^*(\theta) = \mu(\theta)f(\theta)$.

As stated above we have studied the situation such that the probabilities in the distribution (1.4) are modified by a constant quantity $\alpha$ ($0 \leqslant \alpha \leqslant 1$) by increasing the probability of one value of the variable.

In this paper we are concerned with the situation where sometimes the value two is reported erroneously as one in relation to SBMPSD. In Section 2, we obtain the recurrence relations between the raw moments, the central moments and the factorial moments of misclassified SBMPSD. In Section 3, the effect of the misclassification on the variance is considered. To illustrate the situation under consideration some of its particular cases like the size-biased generalized negative binomial (SBGNB) and the size-biased generalized Poisson (SBGP) distributions are studied in Sections 4 and 5. The method of moments for estimation of parameters of size-biased GPD and misclassified size-biased GPD is discussed in Section 6. Finally, an example is presented for the size-biased generalized Poisson distribution (SBGPD) to illustrate the results in Section 7.

## 2. Misclassified size-biased modified power series distribution

Suppose $X$ is a random variable having the SBMPSD (1.4) from which a random sample is drawn. Assume that some of the observations corresponding to $x = 2$ are erroneously reported as $x = 1$, and let this probability of misclassifying be $\alpha$. Then the resulting distribution of $X$, the so called misclassified size-biased modified power series distribution (MSBMPSD), can be written in the form

$$P_3[X = x] = \begin{cases} g(b(1) + \alpha b(2)g)/f\mu, & x = 1, \\ (1-\alpha)b(2)g^2/f\mu, & x = 2, \\ b(x)g^x/(f\mu), & x \in S \end{cases} \tag{2.1}$$

where $S = T/\{1, 2\}$ is a subset of the set $I$ of non-negative integers not containing one and two, $b(x), f = f(\theta), g = g(\theta), \mu = \mu(\theta)$ are stated above and $0 \leqslant \alpha \leqslant 1$, the $\alpha$ being the proportion of misclassified observations. It is interesting to note that for $\alpha = 0$ the distribution (2.1) reduces to simple SBMPSD (1.4). Further, if $g(\theta)$ is invertible, it reduces to the misclassified size-biased GPSD and in addition if $T$ is regarded as an entire set $I$ of non-negative integers, it will be called the misclassified size-biased PSD.

In this section we obtain the mean and the variance and establish certain recurrence relations for the raw, the central and the factorial moments of the distribution (2.1). Here the notation with $*$ corresponds to size-biased MPSD.

**2.1 Mean of the distribution.** The mean of (2.1) is obtained as

$$\text{Mean} = \mu_1' = \frac{g(b(1) + \alpha b(2)g)}{f\mu} + \frac{2(1-\alpha)b(2)g^2}{f\mu} + \sum \frac{xb(x)g^x}{f\mu}$$

where $\sum$ stands for the sum over $x \in S$ here and onwards. Also, from (2.1) we have

$$f\mu = g(b(1) + \alpha b(2)g) + (1-\alpha)b(2)g^2 + \sum b(x)g^x.$$

Differentiating w.r.t. $\theta$ and multiplying both sides by $g/g'$, after simplification we get the mean of (2.1) as

$$\text{Mean} = \mu_1' = \mu_1'^* - (\alpha b(2)g^2)/f\mu \tag{2.2}$$

where $\mu_1'^* = \mu + f/f' \cdot \partial\mu/\partial\theta$ is the mean of (1.4) and $f' = (\partial/\partial\theta)f(\theta)$, $g' = (\partial/\partial\theta)g(\theta)$.

**2.2 Recurrence relation among raw moments.** The $r$-th raw moment of (2.1) is given as

$$\mu_r' = \frac{1}{f\mu}\Big[g(b(1) + \alpha b(2)g) + 2^r(1-\alpha)b(2)g^2 + \sum x^r b(x)g^x\Big]. \tag{2.3}$$

Differentiating (2.3) w.r.t. $\theta$ and simplifying we get

$$\mu_{r+1}' = \frac{g}{g'}\frac{\partial\mu_r'}{\partial\theta} + \alpha(\mu_r' - 1)\frac{g^2 b(2)}{f\mu} + \mu_1'\mu_r'. \tag{2.4}$$

Higher moments can also be obtained with $r = 2, 3, \ldots$. From (2.3) it is easy to establish a relation between the $r$-th moment $(\mu_r')$ of the misclassified size-biased MPSD (2.1) and the $r$-th moment $\mu_r'^*$ of the size-biased MPSD (1.4) as

$$\mu_r' = \mu_r'^* + \alpha(1 - 2^r)\frac{b(2)g^2}{f\mu} \tag{2.5}$$

from which the higher order moments $(r > 1)$ $\mu_2', \mu_3', \ldots$ are obtained.

**2.3. Recurrence relation among central moments.** The $r$-th central moment of (2.1) is given as

$$\mu_r = \frac{1}{f\mu}\Big[(1-\mu_1')^r g(b(1) + \alpha b(2)g) + (1-\alpha)(2-\mu_1')^r b(2)g^2 + \sum (x-\mu_1')^r b(x)g^x\Big].$$

Differentiating w.r.t. $\theta$ and simplifying, we get

$$\mu_{r+1} = \Big(\frac{g}{g'}\Big)\Big[\frac{\partial\mu_r}{\partial\theta} + r\mu_{r-1}\Big(\frac{\partial\mu_1'}{\partial\theta}\Big)\Big] + \alpha b(2)g^2[\mu_r - (1-\mu_1')^r]/f\mu. \tag{2.6}$$

Putting $r = 1$ in (2.6) and noting that $\mu_0 = 1$ and $\mu_1 = 0$, we get the variance $(\mu_2)$ of (2.1) as

$$\mu_2 = \frac{g}{g'}\left(\frac{\partial \mu_1'}{\partial \theta}\right) + \alpha(\mu_1' - 1)b(2)\frac{g^2}{f\mu}. \tag{2.7}$$

Using (2.7) in (2.6) we obtain a recurrence relation for central moments as

$$\mu_{r+1} = \frac{g}{g'}\frac{\partial \mu_r}{\partial \theta} + r\mu_{r-1}\mu_2 + \alpha b(2)g^2\left[\frac{\mu_r - (1-\mu_1')^r - r\mu_{r-1}(\mu_1' - 1)}{f\mu}\right] \tag{2.8}$$

where $r = 2, 3, \dots$ for higher order moments.

**2.4 Recurrence relation among factorial moments.** The $r$-th factorial moment of (2.1) is obtained as

$$\mu_{[r]}' = \frac{1^{[r]}g(b(1) + \alpha b(2)g)}{f\mu} + \frac{2^{[r]}(1-\alpha)b(2)g^2}{f\mu} + \frac{\sum x^{[r]}b(x)g^x}{f\mu}. \tag{2.9}$$

Differentiating (2.9) w.r.t. $\theta$ and using the identity $x.x^{[r]} = x^{[r+1]} + rx^{[r]}$, after a simplification we get

$$\mu_{[r+1]}' = \frac{g}{g'}\frac{\partial \mu_{[r]}'}{\partial \theta} + [\mu_{[1]}'^{*} - r]\mu_{[r]}' - \frac{\alpha 1^{[r]}b(2)g^2}{f\mu} \tag{2.10}$$

where $\mu_{[1]}'^{*}$ is the first factorial moment of the size-biased MPSD(1.4).

Again in view of (2.2), this can be put in the form

$$\mu_{[r+1]}' = \left(\frac{g}{g'}\right)\left(\frac{\partial \mu_{[r]}'}{\partial \theta}\right) + [\mu_{[1]}' - r]\mu_{[r]}' + \frac{\alpha(\mu_{[r]}' - 1^{[r]})b(2)g^2}{f\mu} \tag{2.11}$$

where $r = 2, 3, \dots$ and $\mu_{[1]}', \mu_{[2]}'$ are $\mu_{[1]}' = \mu_1'$ and $\mu_{[2]}' = \mu_2' - \mu_1'^2$ where $\mu_1'$ and $\mu_2'$ are obtained from (2.4) with $r = 0$ and $r = 1$.

## 3. Variance comparison

In this section we discuss how the variance of (2.1) is effected when the reporting observations are erroneously misclassified. From (2.5) we have

$$\mu_1' = \mu_1'^* - \alpha b(2) g^2 / f\mu, \tag{3.1}$$

$$\mu_2' = \mu_2'^* - 3\alpha b(2) g^2 / f\mu. \tag{3.2}$$

Hence, the variance of the distribution (2.1) is given by

$$\mu_2 = \mu_2^* + \alpha\Big(2\mu_1'^* - 3 - \frac{\alpha b(2) g^2}{f\mu}\Big)\frac{b(2)g^2}{f\mu}$$

where $\mu_2^* = (g/g' \cdot \partial\mu_1'^*/\partial\theta)$ is the variance of the size-biased MPSD (1.4).

This gives a relation between the variance $\mu_2$ of (2.1) and the variance $\mu_2^*$ of (1.4) as

$$\mu_2 = \mu_2^* + \varphi(\alpha, \theta) \tag{3.3}$$

where $\varphi(\alpha, \theta)$ is a function of $\alpha$ and $\theta$ only given by

$$\varphi(\alpha, \theta) = \alpha\Big[2\mu_1'^* - 3 - \frac{\alpha b(2) g^2}{f\mu}\Big]\frac{b(2)g^2}{f\mu}. \tag{3.4}$$

The above relation between the two variances shows that the variance of the size-biased MPSD has been affected by a term $\varphi(\alpha, \theta)$ and is due to reporting the observations erroneously. We note that this misclassification has a reasonably moderate effect on the variance and can be easily seen from the graph (see Figure 1). This suggests that the variance of the distribution increases due to misclassification. Now if we take the ratio of the two variances we have

$$Z = \frac{\mu_2}{\mu_2^*} = 1 + \frac{\varphi(\alpha, \theta)}{\mu_2^*}. \tag{3.5}$$

This ratio $Z$ shows that $\mu_2$ may be equal to, greater than or less than $\mu_2^*$ depending upon the value of $\alpha$ and $\theta$; that is, on the term $\varphi(\alpha, \theta)$. This term would be useful in studying the effect of misclassification on the variance. This effect can be studied in two ways:

i) When the value of $\theta$ is fixed.

ii) When the value of $\alpha$ is fixed.

Here we discuss the case (i) only. Similarly the case (ii) can be dealt with. Thus when the value of $\theta$ is fixed, the ratio $Z$ will be a function of $\alpha$ only. That is to say $Z = \varphi(\alpha)$. Hence we have

$$\frac{\partial Z}{\partial \alpha} = \left[\frac{b(2)g^2}{f\mu}\left(2\mu_1'^* - 3 - \frac{2\alpha b(2)g^2}{f\mu}\right)\right]/\mu_2^* \tag{3.6}$$

and

$$\frac{\partial^2 Z}{\partial \alpha^2} = -2\left(\frac{b(2)g^2}{f\mu}\right)^2/\mu_2^* < 0.$$

From this we note that the curve $Z = \varphi(\alpha)$ seems to be concave towards the origin. Equating $\frac{\partial Z}{\partial \alpha}$ to zero gives the value of $\alpha$ as

$$\alpha = (2\mu_1'^* - 3)/(2b(2)g^2/f\mu) \tag{3.7}$$

and hence the maximum value of $Z$ becomes

$$Z_{\max} = 1 + \frac{(2\mu_1'^* - 3)^2}{(4\mu_2^*)}. \tag{3.8}$$

It is clear from (3.8) that the ratio $Z_{\max}$ is always greater than unity. This indicates that the variance has been increased due to the misclassification and it has reasonably a moderate effect on the variance (see Figure 1). Thus we conclude that the variance of the distribution increases due to the erroneous way of reporting the observations. Further, it would be interesting to see that this ratio, when graphed on the $(\alpha, Z)$ axis, will provide an invertible parabola very concave to origin having its vertex at the point

$$\left[\frac{(2\mu_1'^* - 3)}{(2b(2)g^2/f\mu)};\ 1 + \frac{(2\mu_1'^* - 3)^2}{(2\sqrt{\mu_2^*})^2}\right], \tag{3.9}$$

when the value of $\theta$ is fixed. This parabola would be more useful in studying the behavioral effect of misclassification on the variance. Of course, one has to treat separately various possible specific cases with respect to their situations being observed in practice.

## 4. Some applications

We illustrate here the situation under consideration defined by (2.1) in which some of the observations corresponding to the value two are sometimes reported erroneously as one, for some of its special cases like the size-biased generalized negative binomial distribution (SBGNBD), the size-biased generalized Poisson distribution (SBGPD), and apply them to the results obtained in Sections 2 and 3.

**4.1 Misclassified size-biased GNBD.** Jain and Consul [15] defined the generalized negative binomial distribution as

$$P_4[X = x] = \frac{m\Gamma(m+\beta x)}{x!\,\Gamma(m+\beta x - x + 1)}\theta^x(1-\theta)^{m+\beta x - x}; \; x \in T \tag{4.1}$$

where $0 < \theta < 1$; $m > 0$; $|\beta\theta| < 1$.

It reduces to (1.1) with $a(x) = m\Gamma(m+\beta x)/x!\,\Gamma(m+\beta x - x + 1)$, $g(\theta) = \theta(1-\theta)^{\beta-1}$, $f(\theta) = (1-\theta)^{-m}$.

Suppose $X$ has a size-biased GNBD given by

$$P_5[X = x] = \frac{\Gamma(m+\beta x)(1-\theta\beta)\theta^{x-1}(1-\theta)^{m+\beta x - x}}{(x-1)!\,\Gamma(m+\beta x - x + 1)}; \; x \in T \tag{4.2}$$

which is (1.4) with $b(x) = x \cdot a(x) = m\Gamma(m+\beta x)/(x-1)!\,\Gamma(m+\beta x - x + 1)$ and $\mu(\theta) = m\theta/(1-\theta\beta)$,

$$g(\theta) = \theta(1-\theta)^{\beta-1}, \quad f(\theta) = (1-\theta)^{-m}.$$

Let us assume that the value two is sometimes reported as one in (4.2) and let the probability of misclassifying these observations be $\alpha$. Then the resulting distribution of $X$, the so called misclassified size-biased GNBD, can be defined as

$$P_6[x : \alpha, \theta] = \begin{cases} \dfrac{\theta(1-\theta)^{\beta-1}[m + \alpha m(m+2\beta-1)\theta(1-\theta)^{\beta-1}]}{(1-\theta)^{-m} m\theta/(1-\theta\beta)}, & x = 1, \\[2ex] \dfrac{(1-\alpha)m(m+2\beta-1)[\theta(1-\theta)^{\beta-1}]^2}{(1-\theta)^{-m} m\theta/(1-\theta\beta)}, & x = 2, \\[2ex] \dfrac{m\Gamma(m+\beta x)}{(x-1)!\Gamma(m+\beta x - x + 1)} \dfrac{[\theta(1-\theta)^{\beta-1}]^x}{(1-\theta)^{-m} m\theta/(1-\theta\beta)}, & x \in S \end{cases} \tag{4.3}$$

where $0 < \theta < 1$, $0 \leqslant \alpha \leqslant 1$, $m > 0$ and $|\theta\beta| < 1$.

Using the values of $b(x)$, $\mu(\theta)$, $g(\theta)$ and $f(\theta)$ in the foregoing results we obtain results for (4.3):

(4.4) $$\text{Mean} = \mu_1' = \frac{m\theta}{(1-\theta\beta)} + \frac{(1-\theta)}{(1-\theta\beta)^2} - \alpha\theta(m+2\beta-1)(1-\theta)^{2\beta+m-2}(1-\theta\beta);$$

(4.5)
$$\begin{aligned}\text{Variance} = \mu_2 = {} & \frac{\theta(1-\theta)}{(1-\theta\beta)^4}[m - m\theta\beta - \theta\beta + 2\beta - 1] \\ & + \alpha(m+2\beta-1)\theta(1-\theta)^{\beta-1}[2m\theta(1-\theta\beta) + 2(1-\theta) - 3(1-\theta\beta)^2 \\ & - \alpha\theta(m+2\beta-1)(1-\theta)^{2\beta+m-2}(1-\theta\beta)^3](1-\theta)^{\beta+m-1}(1-\theta\beta)^{-1}.\end{aligned}$$

Recurrence relation among raw moments

(4.6) $$\begin{aligned}\mu_{r+1}' = {} & \theta(1-\theta)(1-\theta\beta)^{-1}\frac{\partial\mu_r'}{\partial\theta} \\ & + \alpha(\mu_r' - 1)(m+2\beta-1)\theta(1-\theta\beta)(1-\theta)^{2\beta+m-2} + \mu_1'\mu_r'.\end{aligned}$$

Recurrence relation among central moments

(4.7) $$\begin{aligned}\mu_{r+1} = {} & \theta(1-\theta)(1-\theta\beta)^{-1}\frac{\partial\mu_r}{\partial\theta} + r\mu_{r-1}\mu_2 + \alpha(m+2\beta-1) \\ & \times \theta(1-\theta\beta)(1-\theta)^{2\beta+m-2}[\mu_r - (1-\mu_1')^r - r\mu_{r-1}(\mu_1' - 1)].\end{aligned}$$

Recurrence relation among factorial moments

(4.8) $$\begin{aligned}\mu_{[r+1]}' = {} & \theta(1-\theta)(1-\theta\beta)^{-1}\frac{\partial\mu_{[r]}'}{\partial\theta} \\ & + [\mu_{[1]}' - r]\mu_{[r]}' + \alpha(m+2\beta-1)\theta(1-\theta\beta)(1-\theta)^{2\beta+m-2}[\mu_{[r]}' - 1^{[r]}].\end{aligned}$$

The variance ratio $Z = \mu_2/\mu_2^*$ becomes

(4.9)
$$\begin{aligned} Z = {} & 1 + \alpha(m+2\beta-1)(1-\theta)^{2\beta+m-3}(1-\theta\beta)^3 \\ & \times \frac{2m\theta(1-\theta\beta) + 2(1-\theta) - 3(1-\theta\beta)^2 - \alpha(m+2\beta-1)\theta(1-\theta)^{2\beta+m-2}(1-\theta\beta)^3}{m - m\theta\beta - \theta\beta + 2\beta - 1}.\end{aligned}$$

For a fixed value of $\theta$, the value $\alpha$ will be

(4.10) $$\alpha = \frac{[2m\theta(1-\theta\beta) + 2(1-\theta) - 3(1-\theta\beta)^2]}{2(1-\theta\beta)^3(m+2\beta-1)\theta(1-\theta)^{2\beta+m-2}}$$

for which the maximum value $Z$ comes out as

$$Z_{\max} = 1 + \frac{[2m\theta(1-\theta\beta) + 2(1-\theta) - 3(1-\theta\beta)^2]^2}{4\theta(1-\theta)(m+2\beta-1-m\theta\beta-\theta\beta)}. \tag{4.11}$$

The invertible parabola when graphed on $(\alpha, Z)$ axis will have its vertex at the point

$$\frac{[2m\theta(1-\theta\beta) + 2(1-\theta) - 3(1-\theta\beta)^2]}{2(1-\theta\beta)^3(m+2\beta-1)\theta(1-\theta)^{2\beta+m-2}}, \quad 1 + \frac{[2m\theta(1-\theta\beta) + 2(1-\theta) - 3(1-\theta\beta)^2]^2}{4\theta(1-\theta)(m+2\beta-1-m\theta\beta-\theta\beta)}. \tag{4.12}$$

Particular Case: If we put $\beta = 1$ we get all the results for the size-biased negative binomial distribution.

**4.2 Misclassified size-biased generalized Poisson distribution.** Consul and Jain [7] defined the generalized Poisson distribution as

$$P_7[X = x] = \frac{\lambda_1(\lambda_1 + \lambda_2 x)^{x-1}\mathrm{e}^{-(\lambda_1+\lambda_2 x)}}{x!}, \; \lambda_1 > 0, \; |\lambda_2| < 1, \; x = 0, 1, 2, \dots. \tag{4.13}$$

Shoukri and Consul [22] modified the form of (4.13) to

$$P_8[X = x] = \frac{(1+\beta x)^{x-1}\theta^x \mathrm{e}^{-\theta(1+\beta x)}}{x!}, \; \theta > 0, \; |\theta\beta| < 1, \; x = 0, 1, 2, \dots \tag{4.14}$$

where $\theta = \lambda_1$ and $\beta = \lambda_2/\lambda_1$. It is a particular case of MPSD (1.1) with

$$a(x) = \frac{(1+\beta x)^{x-1}}{x!}, \quad g(\theta) = \theta\mathrm{e}^{-\theta\beta}, \quad f(\theta) = \mathrm{e}^{\theta}.$$

Now suppose $X$ has a size-biased GPD given by

$$P_9[X = x] = \frac{(1+\beta x)^{x-1}(1-\theta\beta)\theta^{x-1}\mathrm{e}^{-\theta(\beta x+1)}}{(x-1)!}; \; x = 1, 2, \dots \tag{4.15}$$

which is (1.4) with

$$b(x) = xa(x) = \frac{(1+\beta x)^{x-1}}{(x-1)!}, \; \mu(\theta) = \frac{\theta}{(1-\theta\beta)}, \; f(\theta) = \mathrm{e}^{\theta}, \; g(\theta) = \theta\mathrm{e}^{-\theta\beta}.$$

Assume that some of the values two are erroneously reported as one and let $\alpha$ be the probability of misclassifying them. Then the resulting distribution of $X$, the so called misclassified size-biased GPD, can be defined as

$$
(4.16)\qquad P_{10}[X=x]=\begin{cases}(\theta e^{-\theta\beta})[1+\alpha(1+2\beta)(\theta e^{-\theta\beta})]/\dfrac{e^{\theta}\theta}{(1-\theta\beta)}, & x=1,\\ \dfrac{(1-\alpha)[1+2\beta](\theta e^{-\theta\beta})^2}{\theta e^{\theta}/(1-\theta\beta)}, & x=2,\\ \dfrac{(1+\beta x)^{x-1}}{(x-1)!}\,\dfrac{(\theta e^{-\theta\beta})^x}{\theta e^{\theta}/(1-\theta\beta)}, & x\in S\end{cases}
$$

where $S$ was defined earlier; $0\leqslant\alpha\leqslant 1$, $\theta>-1$ and $|\beta\theta|<1$.

Using the values of $b(x)$, $\mu(\theta)$, $g(\theta)$ and $f(\theta)$ in the forgoing results we obtain the results for (4.16):

$$
(4.17)\qquad \text{Mean}=\mu_1'=\frac{1}{(1-\theta\beta)^2}+\frac{\theta}{(1-\theta\beta)}-\alpha(1-\theta\beta)(1+2\beta)(\theta e^{-\theta(2\beta+1)}),
$$

$$
\begin{aligned}
(4.18)\qquad \text{Variance}=\mu_2=&\frac{(2\beta\theta-\theta^2\beta+\theta)}{(1-\theta\beta)^4}\\
&+\alpha[2+2\theta(1-\theta\beta)-3(1-\theta\beta)^2\\
&\qquad\qquad-\alpha(1+2\beta)(1-\theta\beta)^3(\theta e^{-\theta(2\beta+1)})]\\
&\times[(1+2\beta)\theta(1-\theta\beta)^{-1}e^{-\theta(2\beta+1)}].
\end{aligned}
$$

Recurrence relation among raw moments

$$
(4.19)\qquad \mu_{r+1}'=\frac{\theta}{(1-\theta\beta)}\frac{\partial\mu_r'}{\partial\theta}+\alpha(1+2\beta)(1-\beta\theta)\theta e^{-\theta(2\beta+1)}[\mu_r'-1]+\mu_1'\mu_r'.
$$

Recurrence relation among central moments

$$
\begin{aligned}
(4.20)\qquad \mu_{r+1}=&\frac{\theta}{(1-\theta\beta)}\Big(\frac{\partial\mu_r}{\partial\theta}\Big)+r\mu_2\mu_{r-1}+\alpha(1+2\beta)\theta(1-\theta\beta)\\
&\times[\mu_r-(1-\mu_1')^r-r\mu_{r-1}(\mu_1'-1)]e^{-\theta(2\beta+1)}.
\end{aligned}
$$

Recurrence relation among factorial moments

$$
(4.21)\quad \mu_{[r+1]}'=\frac{\theta}{(1-\theta\beta)}\frac{\partial\mu_{[r]}'}{\partial\theta}+[\mu_{[1]}'-r]\mu_{[r]}'+\alpha(1+2\beta)\theta(1-\theta\beta)[\mu_{[r]}'-1^{[r]}]e^{-\theta(2\beta+1)}.
$$

The variance ratio $Z=\mu_2/\mu_2^*$ becomes

$$
\begin{aligned}
(4.22)\qquad Z=&1+\alpha[2+2\theta(1-\theta\beta)-3(1-\theta\beta)^2-\alpha\theta(1+2\beta)(1-\theta\beta)^3e^{-\theta(2\beta+1)}]\\
&\times\theta(1+2\beta)e^{-\theta(1+2\beta)}(1-\theta\beta)^3(2\beta\theta-\theta^2\beta+\theta)^{-1}.
\end{aligned}
$$

For the fixed value of $\theta$, the value of $\alpha$ is

$$\alpha = [2 + 2\theta(1 - \theta\beta) - 3(1 - \theta\beta)^2]/2(1 + 2\beta)\mathrm{e}^{-\theta(1+2\beta)}\theta(1 - \theta\beta)^3 \tag{4.23}$$

for which the maximum value $Z$ is

$$Z_{\max} = 1 + \frac{[2 + 2\theta(1 - \theta\beta) - 3(1 - \theta\beta)^2]^2}{4(2\beta\theta + \theta - \theta^2\beta)}. \tag{4.24}$$

The invertible parabola when graphed on the $(\alpha, Z)$ axis will have its vertex at the point

$$\frac{[2 + 2\theta(1 - \theta\beta) - 3(1 - \theta\beta)^2]}{2(1 + 2\beta)\theta\mathrm{e}^{-\theta(1+2\beta)}(1 - \theta\beta)^3};\ 1 + \frac{[2 + 2\theta(1 - \theta\beta) - 3(1 - \theta\beta)^2]^2}{4[2\beta\theta + \theta - \theta^2\beta]}. \tag{4.25}$$

Particular Cases: If we put $\beta = 1$ we get all the results for the size-biased Borel [2] distribution and if $\beta = 0$ we get results for the size-biased Poisson distribution.

## 5. An illustration

To see the effect of misclassification on the variance the case of SBGNBD (4.2) with $m = 4$ and $\beta = 1$ is considered. It is assumed that some of the observations which correspond to the value two are reported erroneously as value one. The following tables show the computed values of the variance ratio and the maximum variance ratio for different values of $\theta$.

| $\alpha/\theta$ | 0.1 | 0.3 | 0.5 | 0.7 | 0.9 |
|---|---|---|---|---|---|
| 0.1 | 1.0039 | 1.0269 | 1.0070 | 1.0005 | 1.0000 |
| 0.3 | 1.0032 | 1.0793 | 1.0210 | 1.0015 | 1.0000 |
| 0.5 | 0.9913 | 1.1301 | 1.0350 | 1.0024 | 1.0000 |
| 0.7 | 0.9680 | 1.1792 | 1.0489 | 1.0034 | 1.0000 |
| 0.9 | 0.9334 | 1.2267 | 1.0628 | 1.0044 | 1.0000 |
| 1.0 | 0.9119 | 1.2498 | 1.0697 | 1.0049 | 1.0000 |
| | (1) | (2) | (3) | (4) | (5) |

Table 5.1 Variance Ratio values for SBGNBD with $m = 4$ and $\beta = 1$

| $\theta$ | 0.1 | 0.11 | 0.12 | 0.13 | 0.14 |
|---|---|---|---|---|---|
| $\alpha$ | 0.188 | 0.384 | 0.574 | 0.763 | 0.953 |
| $Z_{\max}$ | 1.005 | 1.020 | 1.043 | 1.071 | 1.104 |

Table 5.2 Maximum Variance Ratio ($Z_{\max}$) for different values of $\theta$ ($m = 4$, $\beta = 1$)

Similarly, $Z_{\max}$ can also be obtained for different values of $\theta$ depending upon the suitable values of $m$ and $\beta$.

The variance ratio values from the above table are plotted in Fig. 1 on the $(\alpha, Z)$ axis, which yeilds different invertible parabolas concave to origin for small values of $\alpha$. This concavity of the parabola diminishes slowly as $\theta$ increases with $\alpha$, till the value of $Z$ takes the value unity and there after it becomes a curve of straight line kind. It is seen that for low values of $\theta$, the misclassified variance is significantly smaller than that of the simple case, but for higher $\theta$ values, it increases beyond the simple variance very slowly. This shows that there is a moderate effect on the variance. However, near the end points of the interval $(0, 1)$ of $\theta$ one may not be able to say definitely as these comments refer specifically to the SBGNB distribution, and not to any general misclassified discrete distribution. In fact, one has to look for such effect separately for various possible situations being found in practice.

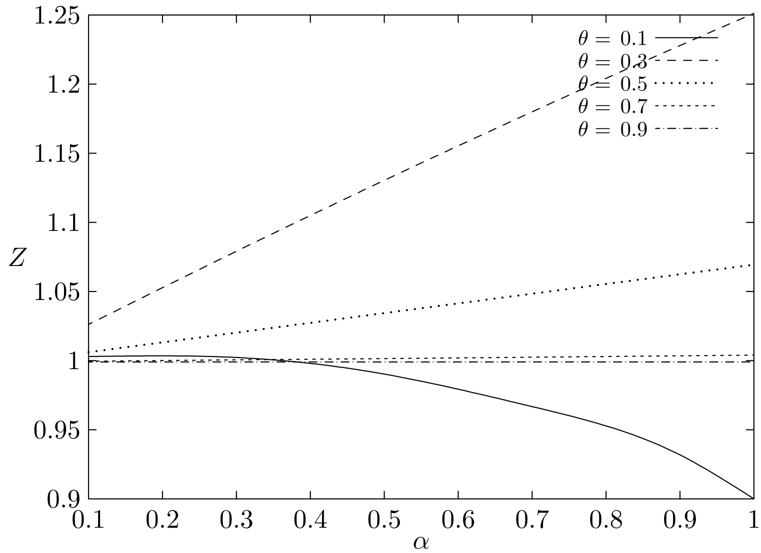


Figure 1. $(\alpha, Z)$-graph for SBGNBD distribution with $m = 4$ and $\beta = 1$.

## 6. Estimation of misclassified size-biased GPD

The method of moments is considered for estimation of parameters $\alpha$ and $\beta$ when $\theta$ is known in the case of misclassified SBGPD (4.16). Using (4.17) and (4.18), after simplification we get the equation in $\lambda$

$$\lambda^4\mu_2 - 2(1-\lambda) - \lambda\theta - [1+\theta\lambda - 3\lambda^2 + \lambda^2\mu_1'][1+\theta\lambda - \lambda^2\mu_1'] = 0 \tag{6.1}$$

where $1 - \theta\beta = \lambda$.

Replacing $\mu_1'$ and $\mu_2$ by the corresponding sample values $\overline{x}$ and $S^2$ respectively, we get

$$\lambda^4 S^2 - 2(1-\lambda) - \lambda\theta - [1 + \theta\lambda - 3\lambda^2 + \lambda^2\overline{x}][1 + \theta\lambda - \lambda^2\overline{x}] = 0. \tag{6.2}$$

It is a polynomial of degree four that can be solved using the Newton-Raphson method and so an estimate of $\beta$ can be obtained as

$$\hat{\beta} = \frac{1-\hat{\lambda}}{\theta}. \tag{6.3}$$

An estimate of $\alpha$ is obtained from (6.1) as

$$\hat{\alpha} = \frac{1 + \theta\hat{\lambda} - \hat{\lambda}^2\overline{x}}{\hat{\lambda}^3(1+2\hat{\beta})\theta \mathrm{e}^{-\theta(2\hat{\beta}+1)}}. \tag{6.4}$$

**6.1 Estimation of size-biased GPD.** Using the moment method of estimation for the estimation of the parameters of the Size-Biased GPD, i.e., $\alpha$ and $\beta$, we obtain

$$\text{Mean} = \mu_1' = \frac{\theta}{(1-\theta\beta)} + \frac{1}{(1-\theta\beta)^2}, \tag{6.5}$$

$$\text{Variance} = \mu_2 = \frac{2\theta\beta}{(1-\theta\beta)^4} + \frac{\theta}{(1-\theta\beta)^3}. \tag{6.6}$$

For convenience let $1-\theta\beta = \lambda$ or $\theta\beta = 1-\lambda$; then we obtain

$$\mu_1' = \frac{\theta\lambda + 1}{\lambda^2}, \tag{6.7}$$

$$\mu_2 = \frac{2(1-\lambda) + \theta\lambda}{\lambda^4}. \tag{6.8}$$

This yeilds an equation for $\lambda$

$$\mu_2\lambda^4 - \mu_1'\lambda^2 + 2\lambda - 1 = 0. \tag{6.9}$$

Replacing $\mu_1'$ and $\mu_2$ by the corresponding sample values $\overline{x}$ and $S^2$ respectively, we obtain

$$S^2\lambda^4 - \overline{x}\lambda^2 + 2\lambda - 1 = 0. \tag{6.10}$$

It is a polynomial of degree four that can be solved by using the Newton-Raphson method and so an estimate of $\lambda$ can be obtained. An estimate of $\theta$ is then obtained as

$$\hat{\theta} = \frac{\hat{\lambda}^2\overline{x} - 1}{\hat{\lambda}}. \tag{6.11}$$

After estimating $\hat{\theta}$ and $\hat{\lambda}, \hat{\beta}$ is obtained as

$$\hat{\beta} = \frac{1-\hat{\lambda}}{\hat{\theta}}. \tag{6.12}$$

## 7. An illustrative example (goodness of fit)

To illustrate the practical application of results obtained in this paper, data from Singh and Yadav's [22] classical example on the number of households ($f$) having at least one migrant according to the number of migrants ($X$) has been suitably altered. For the purpose of this illustration it has been assumed that ten of the records which should have shown two migrants each were in error by reporting one migrant. Both the original and the altered data for this example are given in Table 7.1. For the original data we fit the size-biased GPD (4.15) and for the altered data we fit the misclassified size-biased GPD (4.16). The moment method of estimation is used for estimation of parameters in both cases. Also in the case of misclassified size-biased GPD we assume the parameter '$\theta$' to be known and take the same value of $\theta$ as obtained from the size-biased GPD.

| **No. of Migrants** $X$ | **Original data** | | | **Altered data** | | |
|---|---|---|---|---|---|---|
| | **Size-Biased GPD** | | | **Misclassified Size-Biased GPD** | | |
| | Observed frequencies | Expected frequencies | Estimates of parameters | Observed frequencies | Expected frequencies | Estimates of parameters |
| 1 | 375 | 377.6 | | 385 | 387.6 | $\theta$ =0.118889 |
| 2 | 143 | 139.8 | $\hat{\theta} = 0.118889$ | 133 | 129.8 | (Known) |
| 3 | 49 | 47.9 | $\hat{\beta} = 1.318793$ | 49 | 47.9 | $\hat{\beta} = 1.318768$ |
| 4 | 17 | 16.3 | | 17 | 16.3 | $\hat{\alpha} = 0.071584$ |
| 5 | 2 | 5.6 | | 2 | 5.6 | |
| 6 | 2 | 1.9 | | 2 | 1.9 | |
| 7 | 1 | 0.7 | | 1 | 0.7 | |
| 8 | 1 | 0.2 | | 1 | 0.2 | |
| **Total** | **590** | **590** | | **590** | **590** | |
| $\chi^2$ | | 0.8322 | | | 0.8372 | |
| d.f. | | 2 | | | 2 | |

Table 7.1 Number of households (f) having at least one migrant according to the number of migrants ($X$) (Singh and Yadav [22])

The estimate $\hat{\beta} = 1.318768$ obtained for the altered data is to be compared with $\hat{\beta} = 1.318793$ obtained when the calculations are based on the original unaltered sample. The estimate $\hat{\alpha} = 0.071584$ is to be compared with $10/143 = 0.069930$, which is the proportion of two's that were misclassified in the process of altering the original data for this illustration.

As indicated by the low values obtained by $\chi^2$, the agreement between the observed and expected frequencies both in the original and the altered sample is satisfactory. The value $\chi^2_{(2)} = 0.8372$ for the altered sample is to be compared with $\chi^2_{(2)} = 0.8322$ obtained for the original (unaltered) sample.

Remark. If we consider $\beta = 1.318793$ (as obtained from the size-biased GPD) to be known we obtain the values of $\theta$ and $\alpha$ as $\hat{\theta} = 0.118887$ and $\hat{\alpha} = 0.071584$ using (6.3) and (6.4), which are the same as obtained in Table 7.1.

Acknowledgement. The authors thank very much the anonymous referee and editor for his/her many helpful suggestions, which substantially simplified the paper.

*References*


[1] *S. Blumenthal*: Proportional sampling in life length studies. Technometrics *9* (1967), 205–218.

[2] *E. Borel*: Sur l'emploi du theoreme de Bernoulli pour faciliter le calcul d'un infinite de coefficients. Application au probleme de l'attente a un guichet, Comptes Rendus, Academie des Sciences, Paris, Series A *214* (1942), 452–456.

[3] *L. von Bortkiewicz*: Das Gesetz der kleinen Zahlen. Teubner, Leipzig, 1898.

[4] *A. C. Cohen*: Estimating the parameters of a modified Poisson distribution. Amer. Statist. Assoc. *55* (1960), 139–143.

[5] *A. C. Cohen*: Estimation in the Poisson distribution when sample value of $c+1$ is sometimes erroneously reported as $c$. Ann. Inst. Stat. Math. *9* (1960), 189–193.

[6] *A. C. Cohen*: Misclassified data from binomial distribution. Technometrics *2* (1960), 109–113.

[7] *P. C. Consul, G. C. Jain*: A generalization of the Poisson distribution. Technometrics *15* (1973), 791–799.

[8] *R. C. Gupta*: Modified power series distribution and some of its applications. Sankhya, Series B *36* (1974), 288–298.

[9] *R. C. Gupta*: Some characterizations of discrete distributions. Commun. Stat. A5 *1* (1975), 45–48.

[10] *R. C. Gupta*: Some characterizations of distributions by properties of their forward and backward recurrence times in a renewal process. Scand. J. Stat., Theory Appl. *3* (1976), 215–216.

[11] *R. C. Gupta*: Waiting time paradox and size biased sampling. Commun. Stat., Theory Methods A8 *13* (1979), 601–607.

[12] *R. C. Gupta*: Some characterizations of renewal densities with emphasis in reliability. Math. Operationsforsch. Stat. *15* (1984), 571–579.

[13] *R. C. Gupta, R. C. Tripathi*: A comparison between the ordinary and the length-biased modified power series distributions with applications. Commun. Stat., Theory Methods *16* (1987), 1195–1206.

[14] *R. C. Gupta, R. C. Tripathi*: Statistical inference based on the length-biased data for the modified power series distributions. Commun. Stat., Theory Methods *21* (1992), 519–537.
[15] *G. C. Jain, P. C. Consul*: A generalized negative binomial distribution. SIAM J. Appl. Math. *21* (1971), 501–513.
[16] *P. N. Jani, S. M. Shah*: Misclassification in modified power series distribution in which the value one is sometimes reported as zero and some of its applications. Metron *37* (1979), 121–136.
[17] *A. Noack*: A class of random variables with discrete distributions. Ann. Math. Stat. *21* (1950), 127–132.
[18] *A. I. Patel, I. D. Patel*: Recurrence relations for moments of the so called misclassified distribution. Assam Statistical Review *10* (1996), 107–119.
[19] *A. I. Patel, I. D. Patel*: Misclassification in modified power series distribution and its applications. Assam Statistical Review *15* (2001), 55–69.
[20] *G. P. Patil*: Certain properties of the generalized power series distributions. Ann. Inst. Stat. Math. Tokyo. *14* (1962), 179–182.
[21] *R. L. Scheaffer*: Size biased sampling. Technometrics *14* (1972), 635–644.
[22] *S. N. Singh, R. C. Yadav*: Trends in rural out-migration at household level. Rural Demography. *8* (1971), 53–61.
[23] *M. M. Shoukri, P. C. Consul*: Bayesian analysis of generalized Poisson distribution. Commun. Stat., Theory Methods *18* (1989), 1465–1480.
[24] *W. O. Williford, S. F. Bingham*: Bayesian estimation of the parameters in two modified Poisson distributions. Commun. Stat., Theory Methods A8 *13* (1979), 1315–1326.

*Authors' addresses*: *Anwar Hassan*, P. G. Department of Statistics, University of Kashmir, Srinagar, India, e-mail: anwar.hassan2007@gmail.com, anwar_hassan2007@hotmail.com; *Peer Bilal Ahmad*, P. G. Department of Statistics, University of Kashmir, Srinagar, India, e-mail: peerbilal@yahoo.co.in.