THE TRUNCATED NON-CENTRAL NEGATIVE BINOMIAL DISTRIBUTION

Phang Yook Ngor
Universiti Teknologi MARA Cawangan Melaka
KM 26, Jalan Lendu, 78000 Alor Gajah, Melaka

Abstract: Zero truncated distribution arises when the zero class cannot be determined. Several approaches have been used in estimating the parameters of truncated negative binomial and truncated Poisson distribution. In this study, Zinlinkas' optimization routine is used to estimate parameters of the truncated negative binomial distribution. In addition, this paper looks into the zero-truncated non-central negative binomial distribution and its application to data fitting. Various estimation methods will be examined and applied to two data sets.

Keywords: Truncated, Non-central negative binomial, Maximum likelihood estimation

INTRODUCTION

Zero truncated distribution arises when the zero class cannot be determined. For example, if chromosome breaks in irradiated tissue (Samford, 1955) [5] can occur only in those cells which are at a particular stage of the mitotic cycle at the time of irradiation, a cell can be demonstrated to have been at that stage only if breaks actually occur. Thus, the distribution of break per cell, cells not susceptible to breakage are indistinguishable from susceptible cells in which no break occur. Likewise, HIV carriers can only be detected three months after infection (the incubation period). Therefore in the distribution of number of HIV carriers per state, there are HIV infected patients who cannot be identified during the incubation period and there will be no differences compared with those who are healthy and not infected

Examples of Zero-truncated Distribution

Several approaches have been used in estimating the parameters of truncated negative binomial and truncated Poisson distribution. Samford (1955) [5] has reported on parameter estimation for the truncated negative binomial. Similar problems on the estimation of truncated distributions have been discussed by David and Johnson (1952) [2]. In view of the difficulty in solving the first two moments and the maximum likelihood equations as mentioned by Samford (1955) [5], Brass (1958) [1] introduced simplified methods of estimation; the efficiency of these methods were also considered. One of the methods is the modification of equations for estimation by moments where he recommended the use of proportion in the first class of the truncated distribution. Another method is modification of the maximum likelihood equations which is found to be less laborious than the procedure required for the maximum likelihood fitting of the truncated negative binomial.

In this study, Zilinkas'(1978) [6] optimization routine is used to estimate parameters of the truncated negative binomial. Other than that, I shall look into the zero-truncated NNB distribution and its application to data fitting. Various estimation methods will be examined and applied to two data sets. The first, second data sets are the truncated data sets collected by the East African Medical Survey in the Kwimba district of Tanganyika territory on the number of children ever born to a sample of mothers over 40 years of age (Table 1), Brass (1958) [1] and an investigation into chromosome breakage where the sample distribution of breaks per cell was obtained (Table 2) by Samford (1955) [5].

Estimating Parameters for Truncated Negative Binomial by Maximizing the Likelihood Function

The probability function of the truncated negative binomial distribution (TNB) is

$$P(x) = \frac{w^k}{1-w^k}\frac{(k+x-1)!}{(k-1)!\,x!}(1-w)^x, \qquad x = 1, 2, \ldots\ldots$$

When estimating the parameters of the TNB distribution using Zilinkas' program, the two parameters are reduced to one using the moment estimate as suggested by Brass (1958) [1], where

$$k = \frac{wm - (n_1/n)}{1-w} \qquad \text{where } w = 1-p,\ \ 0 < p < 1$$

p is maximized in the range of $0 < p < 1$. The obtained results are presented in Tables 1 - 2.

Truncated Non-central Negative Binomial Distribution (TNNB)

The truncated NNBD is a Non-central Negative Binomial (NNB) (Ong, 1979) [4] distribution where the zeros are not recorded.

$$P(k) = e^{-\lambda p} q^{v} p^{k} L_k^{v-1}(-\lambda q)\Big/\left[1 - e^{-\lambda p} q^{v}\right] \qquad k = 1,2,3 \ldots.. \tag{1}$$

The probability generating function is given by

$$\begin{aligned} g(z) &= \sum_{k=1}^{\infty} \Pr(k) z^k \\ &= \frac{1}{1-e^{-\lambda p} q^{v}} \left\{ \sum_{k=0}^{\infty} e^{-\lambda p} q^{v} p^{k} L_k^{v-1}(-\lambda q) - e^{-\lambda q} q^{v} \right\} \\ &= \frac{1}{1-e^{-\lambda p} q^{v}} \left(\left(\frac{q}{1-pz}\right)^{v} e^{\lambda\left(\frac{q}{1-pz}-1\right)} - e^{-\lambda p} q^{v} \right) \end{aligned} \tag{2}$$

The j-th factorial moment is

$$\frac{\partial^j g(z)}{\partial z^j}\bigg|_{z=1} = \frac{1}{1-e^{-\lambda p} q^{v}} \frac{\partial^j}{\partial z^j}\left[\left(\frac{q}{1-pq}\right)^{v} e^{\lambda\left(\frac{q}{1-pq}-1\right)}\right]\bigg|_{z=1}$$

$$\mu'_{[z]} = j!\left(\frac{p}{q}\right)^{j} L_j^{v-1}(-\lambda)\Big/\left[1 - e^{-\lambda p} q^{v}\right] \tag{3}$$

A recurrence formula for $\mu'_{[j]}$ is found to be

$$\mu'_{[j+1]} = \frac{p}{q}(2j+v+\lambda)\mu'_{j} - \frac{p}{q}(j+v-1)j\mu'_{[j-1]} \tag{4}$$

The three term probability recurrence relation is

$$(k+1)P(k+1) = (2k+v+\lambda q)pP(k) - p^2(k+v-1)P(k-1) \quad k \geq 2 \tag{5}$$

where

$$P(1) = e^{-\lambda p} q^{v} p(v+\lambda q)\Big/\left(1 - e^{-\lambda p} q^{v}\right) \qquad \text{and}$$

$$P(2) = \left[(2+v+\lambda q)(v+\lambda q) - v\right]P(1)\big/\left(2(v+\lambda q)\right)$$

The first three factorial moments of the truncated NNB are

$$\mu'_{[1]} = \frac{p}{q}(\mu+\lambda)\Big/\varphi, \qquad \mu'_{[2]} = 2!\left(\frac{p}{q}\right)^{2} L_2^{v-1}(-\lambda)\Big/\varphi, \qquad \mu'_{[3]} = 3!\left(\frac{p}{q}\right)^{3} L_3^{v-1}(-\lambda)\Big/\varphi$$

where $\varphi = 1 - e^{-\lambda p} q^{v}$

David and Johnson (1952) [2] mentioned that the use of estimates which is less than the maximum efficiency is justifiable only if they are directly obtainable as explicit solutions of easily constructed equation. In the truncated NNB the three factorial moment given above do not provide explicit solution. David and Johnson abandoned the moment estimates due to the inefficiency caused by introducing the third moment. They recommended the use of MLE for use in all cases. In the following I shall discuss some methods based on pseudo-maximum likelihood estimation, minimum chi-square and MLE using maximization by NELMIN in estimating the parameters of the truncated NNB.

MLE Based upon Pseudo-MLE Using the Third and Forth Factorial Moments

From (4) we obtain

$$\mu'_{[3]} = \frac{p}{q}\left[(4+\nu+\lambda)\mu'_{[2]} - \frac{p}{q}(1+\nu)2\mu'_{[1]}\right] \qquad (6)$$

$$\mu'_{[4]} = \frac{p}{q}\left[(6+\nu+\lambda)\mu'_{[3]} - \frac{p}{q}(2+\nu)3\mu'_{[2]}\right] \qquad (7)$$

From (7), we get

$$\lambda = \left(\frac{q}{p}\right)\frac{\kappa_{[3]}}{\kappa_{[2]}} + \left(\frac{p}{q}\right)\frac{\kappa_{[1]}}{\kappa_{[2]}}2(1+\nu) - (4+\nu) \qquad (8)$$

Substituting (8) into (6), we have

$$\left(\frac{p}{q}\right)^2\left[2(1+\nu)\frac{\mu'_{[1]}\mu'_{[3]}}{\mu'_{[2]}} - 3(2+\nu)\mu'_{[2]}\right] + \frac{p}{q}\left(2\mu'_{[3]}\right) + \frac{\mu'^{2}_{[3]}}{\mu'_{[2]}} - \mu'_{[4]} = 0 \qquad (9)$$

Let $w = \frac{p}{q}$, $a = 2(1+\nu)\frac{\mu'_{[1]}\mu'_{[3]}}{\mu'_{[2]}} - 3(2+\nu)\mu'_{[2]}$, $b = 2\mu'_{[3]}$, $c = \frac{\mu'^{2}_{[3]}}{\mu'_{[2]}} - \mu'_{[4]}$ and the results of (9) are given by

$$w = \left[-b \pm \sqrt{b^2 - 4ac}\right]\Big/2a \qquad (10)$$

Since $w > 0$, we take $w = \left[-b + \sqrt{b^2 - 4ac}\right]\Big/2a$ provided $b^2 > 4ac$. (10) is express in term of ν.

The following steps are used in estimating the parameters:

i) Zilinkas' program is used to search over ν in the range of $0 < \nu < q$ or $q < \nu < \infty$ where

$$q = \left[\frac{\mu'^{2}_{[2]}\mu'^{2}_{[3]}}{\mu'^{2}_{[3]} - \mu'_{[2]}\mu'_{[4]}} - 2\mu'_{[1]}\mu'_{[3]} + 6\mu'^{2}_{[2]}\right]\Big/\left[2\mu'_{[1]}\mu'_{[3]} - 3\mu'^{2}_{[2]}\right]$$

which is obtained from the constraint $b^2 > 4ac$.

(ii) With the obtained ν, p is searched in the range of $0 < p < s/(1+s)$ where

$s = \left[-b_1 - \sqrt{b_1^2 - 4a_1c_1}\right]\Big/2a_1$, $a_1 = 2(1+\nu)\mu'_{[1]}/\mu'_{[2]}$, $b_1 = -(4+\nu)$, $c_1 = \mu'_{[3]}/\mu'_{[2]}$

which is determine from (8) by letting $\lambda > 0$. Solving this inequality will give s and $s_1 = \left[-b_1 + \sqrt{b_1^2 - 4a_1c_1}\right]\Big/2a_1$. s_1 is rejected because it will give rise to inadmissible ν.

(iii) The attained p is used to search for v again over the boundaries $0<v<t$ or $t<v<\infty$ where $t = \left[\mu'_{[3]} + 2\mu'_{[1]}w^2 - 4w\mu'_{[2]}\right] \Big/ \left[w\mu'_{[2]} - 2\mu'_{[1]}w^2\right]$ which is found from (9) by letting $\lambda > 0$.

The steps (ii) and (iii) are repeated until the difference between v_I and v_{i+1} is less than $\varepsilon = 10^{-4}$. In steps (i) and (iii) above, the boundaries for v involve ∞ . In view of this, the transformation $\theta = e^{-v}$ is used. Then the boundaries for v will become $e^{-q} < -\ln(\theta) < 1$ or $0 < -\ln(\theta) < e^{-q}$ and $e^{-t} < -\ln(\theta) < 1$ or $0 < -\ln(\theta) < e^{-t}$.

MLE Base upon Pseudo-MLE Method Using the Third Factorial Moment and Probability P(3).

From (5), we have

$$3P(3) = p(4+v+\lambda q)P(2) - p^2(v+1)P(1) \tag{11}$$

which gives

$$v = \left\{3P(3) + p^2P(1) - (4+\lambda q)pP(2)\right\} \big/ p\left\{P(2) - pP(1)\right\} \tag{12}$$

Substituting (12) into (6) and simplifying we get

$$\lambda = \left(\mu'_{[3]} - \Phi\right) \Big/ \left\{\left(\frac{p}{q}\right)\mu'_{[2]} - \frac{\phi q P(2)}{p(P(2) - pP(1))}\right\} \tag{13}$$

where $$\phi = \left(\frac{p}{q}\right)\mu'_{[2]} - 2\left(\frac{p}{q}\right)^2 \mu'_{[1]}$$ and

$$\Phi = 4\left(\frac{p}{q}\right)\mu'_{[2]} - 2\left(\frac{p}{q}\right)^2 \mu'_{[1]} + \phi\left[3P(3) + p^2P(1) - 4pP(2)\right] \big/ p\left(P(2) - p(P(1)\right)$$

Zilinkas' program is used to search for p .Due to the difficulty in determining the boundaries base on the above equation, p is searched in the range of $0<p<1$. When λ or $v < 0$, λ and v are then set equal to 0.001. Then using the p found , v is searched in the range as given in step (iii) in section 2.1. With this v, p is searched again in the same range as given in step (ii) in section 2.1. The process is repeated until the difference between p_i and p_{i+1} is less than a prescribed ε.

Method of Minimum Chi-square Using Third Factorial Moment and Third Probability

In this section, p is fixed from 0.001 to 0.99 in (13) and then substitute the obtained λ into (12) to get v. Subsequently, we compute the chi-square (χ^2) values with these parameter values. The parameters which are within the feasible range and which provide the minimum χ^2 values are chosen as the estimates for p, λ and v.

ML Estimation of Truncated NNBD by Maximizing the Likelihood Function (Algorithm NELMIN)

Optimization routine AS47(O'Neil, 1971, Function Minimization using simplex Procedure) [3] Incorporating Remark ASR11(Chambers, 1976) and Corrigendum AS47(o'Neil, 1971) is used to find the maximum likelihood estimate. The starting values are obtained by solving the third and fourth factorial moment and with $v = 1$. If λ is less than zero, we start with 0.001. In order to ensure that a global maximum is attained, the random start procedure is used.

RESULTS AND DISCUSSIONS

The results show that the truncated NB distribution fitted well to the two data sets. Among all the methods used in estimating parameters for truncated NNB distribution, the method of maximizing the likelihood (Nelmin) provides the best fits for both data sets. Minimum chi-square method gives the best fitting for data set 2. The results are shown in Tables 1 and 2. The poorly fitted data sets are not presented.

Table 1: Number of Children ever Born to a Sample of Mothers over 40 Years of Age (Brass, 1958) [1]

No. of children per mother	No. of mothers	Expected frequency	
		TNBD	TNNBD
		0	4
1	49	48.00	48.96
2	56	61.05	60.07
3	73	60.63	58.95
4	41	51.77	50.47
5	43	39.87	39.35
6	23	28.48	28.63
7	18	19.22	19.74
8	18	12.39	13.04
9	7	7.70	8.31
10	7	4.64	15
11	3	2.73	3.10
12	2	3.52	4.23
Total	340	340.00	340.00
χ^2		10.63	10.51

$\tilde{\lambda}_4 = 2.7442$

$\tilde{p}_0 = 0.4357$ $\tilde{p}_4 = 0.3781$

$\tilde{k}_0 = 4.8393$ $\tilde{v}_4 = 3.4514$

0 = MLE (maximizing likelihood function of NB)

1 = MLE (iteration of PMLE ($\mu'_{[3]}, \mu'_{[4]}$))

2 = MLE (iteration of PMLE ($\mu'_{[3]}$,P(3))

3 = Minimum chi-square

4 = MLE (maximization with NELMIN)

Table 2 : An Investigation into Chromosome Breakage (Samford, 1955) [5]

No. of breaks per cell	No. of cells	Expected frequency				
		TNBD	TNNBD			
		0	1	2	3	4
1	11	10.87	10.81	16.53	9.42	10.79
2	6	6.37	6.23	6.46	5.77	5.84
3	4	4.16	4.09	3.35	4.08	3.80
4	5	2.86	2.84	1.95	3.01	2.68
5	0	2.02	2.04	1.21	2.27	1.96
6	1	1.46	1.49	0.78	1.73	1.48
7	0	1.07	1.10	0.52	1.32	1.14
8	2	0.79	0.82	0.35	1.01	0.88
9	1	0.58	0.62	0.24	0.78	0.69
10	0	0.44	0.47	0.17	0.60	0.55
11	1	0.33	0.35	0.12	0.46	0.43
12	0	0.25	0.27	0.09	0.35	0.34
13	1	0.80	0.87	0.23	1.20	1.42
Total	32	32.00	32.00	32.00	32.00	32.00
χ		2.01	2.16	8.18	3.01	2.98

	$\tilde{\lambda}_1 = 0.1672$	$\tilde{\lambda}_2 = 0.0000$	$\tilde{\lambda}_3 = 0.2673$	$\tilde{\lambda}_4 = 0.1175$
$\tilde{p}_0 = 0.7882$	$\tilde{p}_1 = 0.7546$	$\tilde{p}_2 = 0.7720$	$\tilde{p}_3 = 0.7500$	$\tilde{p}_4 = 0.8071$
$\tilde{k}_0 = 0.4873$	$\tilde{v}_1 = 0.3910$	$\tilde{v}_2 = 0.0101$	$\tilde{v}_3 = 0.0671$	$\tilde{v}_4 = 0.2285$

REFERENCES

1. Brass, W. 1958. Simplified methods of fitting the truncated negative binomial distribution. Biometrika, 45, 59-68

2. David, f. N., and Johnson, N. L.1952. The truncated poisson distribution. Biometric, 8, 275-285

3. O'Neil, R. 1971. Function minimization using a simplex procedure. Algorithm 47. Applied Statistics, 20, 338-345

4. Ong, S. H. and Lee, P. A. 1979. The non-central negative binomial distribution. Biom. J. 21, no. 7, 611-627

5. Samford, M. R. 1955. The truncated negative binomial distribution. Biometrika, 42, 58-69

6. Zilinkas, A. 1978. Algorithm AS 133. Optimization of one-dimensional multimoda functions. Applied Statistics, 27, 367-375