〈Research Paper〉

Journal of the Koran
Statistical Society
Vol. 15, No. 1, 1986

Multivariate Modified Discrete Distributions

G.S. Lingappaiah*

ABSTRACT

In this paper, multivariate discrete distribution is dealt with, where a set of r distinct counts are misreported as another set of r counts. First, the variance for the one variable marginal case is expressed in the form of an inverted parabola. Next, for the multivariate negative binomial case, elements of the covariance matrix are evaluated with reference to asymptotic distributions. Finally, for the same case of multivariate negative binomial, Bayesian estimates of the parameters and of the modification rates are provided.

1. Introduction

Modified discrete distributions have been extensively studied recently. For example, Cohen(1960, a, b, c) dealt with the modification in terms of a single misreported count in the models based on binomial and Poisson distributions. Parikh and Shah(1969) considered the same problem with reference to power series distribution. Lingappaiah (1978, 1979) generalized modification to more than one misreported counts and includes many kinds of generalizations involving many sets of misreported counts. Varahamurthy (1967) also dealt with modification based on the Poisson model. Williford and Bingham (1979) treated the same problem from the Bayesian point of view. Most of the above works are related to the univariate case. What is being done in this paper is to generalize this modification to the multivariate discrete distributions where a set of r counts is misreported as another set of r counts. Variance for the marginal case of one variable is put in the form of an inverted parabola when all the modification rates are the same

* Department of Mathematics, Concordia University, Montreal, Quebec H3G 1M8, Canada

and the ratio of modified variance to that of simple(non-modified) case is evaluated for a special case. Also for the asymptotic distributions, elements of $(r+k)\times(r+k)$ covariance matrix (k parameters and r modification rates) are evaluated for the multivariate negative binomial distribution. Finally, for the same case of multivariate negative binomial, Bayesian estimates of parameters and of modification rates are put in the closed forms.

2. Modified model and variance

Suppose r counts $(i_1, i_2, ..., i_k)$ $i=1, 2..., r$ in a k-dimensional discrete distribution are misreported as another set of r distinct counts $(i_1', ..., i_k')$, $i=1, 2..., r$, then the general model can be expressed as

$$P(x_1, ..., x_k)=\begin{cases} P(i)+\lambda_i P(i') & \text{if } x_1=i_1, ..., x_k=i_k \\ (1-\lambda_i)P(i') & \text{if } x_1=i_1', ..., x_k=i_k' \\ & i=1, 2, ..., r \\ P(l) & \text{if } x_1=l_1, ..., x_k=l_k \end{cases} \tag{1}$$

where $i_m, i_m', l_m=0, 1, 2, ...,$ $m=1, 2, ..., k$, $i=1, 2, ..., r$, and $i_m\neq i_m'\neq l_m$ [that is, r counts $(i_1, ..., i_k)i=1, 2, ..., r$ are distinct from r counts $(i_1', ..., i_k')i=1, 2, ..., r$ and the counts $(l_1, ...l_k)$ represent all the remaining distinct counts other than the above $2r$ counts]. In (1),

$$\begin{aligned} &P(i)=P(i_1, ..., i_k)=P(x_1=i_1, ..., x_k=i_k), \\ &P(i')=P(i'_1, ..., i_k')=P(x_1=i_1', ..., x_k=i_k') \text{ and} \\ &P(l)=P(l_1, ..., l_k)=P(x_1=l_1, ..., x_k=l_k). \end{aligned} \tag{2}$$

From (1), we can write

$$P(x_1)=\begin{cases} P(i_1)+\lambda_1 P(i_1') & \text{if } x_1=i, \\ (1-\lambda_i)P(i_1') & \text{if } x_1=i_1' \\ P(l_1) & \text{if } x_1=l_1 \end{cases} \tag{3}$$

$i=1, 2, ..., r, i_1\neq i_1'\neq l_1$ and $i_1, i_1', l_1=0, 1, 2, ...$

Again on the x_1 axis, points $i_1 (i=1, ..., r)$ are distinct from the points $i_1' (i=1, ..., r)$. From (3), we have

$$\mu_1=m_1+\sum_{i=1}^{r}\lambda_i a_i P(i_1') \tag{4}$$

where $a_i=i_1-i_1'$, μ_1 is the mean of the modified distribution and m_1 is that of simple

(non-modified) case respectively. Again from (3), we get

$$\mu_2' = m_2' + \sum_{i=1}^{r} a_i b_i \lambda_i P(i_1') \tag{5}$$

with $b_i = i_1 + i_1'$ and μ_2', m_2' are the second raw moments for the modified and simple cases respectively. From (4) and (5) with $\lambda = \lambda_1 = \ldots = \lambda_r$, we have,

$$\mu_2 = m_2 + \lambda B - \lambda^2 C^2 - 2\lambda m_1 C \tag{6}$$

with $C = \sum a_i P(i_1')$, $B = \sum a_i b_i P(i_1')$ and $\mu_2, m_2 = \sigma_1^2$ are the variances for the modified and the simple distribution respectively.

Equation (6) represents an inverted parabola in λ and can be put as

$$\left[y - 1 - \frac{(B - 2m_1 C)^2}{4\sigma^2 C^2} \right] = -\frac{C^2}{\sigma^2} \left[\lambda - \frac{B - 2m_1 C}{2C^2} \right]^2 \tag{7}$$

with $y = \mu_2/\sigma_1^2$. Now consider the multivariate negative binomial distribution given by

$$f(\underset{\sim}{x}, \underset{\sim}{\theta}) = f(x_1, \ldots, x_k ; \theta_1, \ldots, \theta_k) = \frac{\Gamma(s + x_1 + \ldots + x_k)}{\Gamma(s) x_1! x_2! \ldots x_k!} \theta_1^{x_1} \ldots \theta_k^{x_k} A^s \tag{8}$$

where $A = 1 - \theta_1 - \ldots - \theta_k$, $0 < \theta_i < 1, i = 1, \ldots, k$, $\theta_1 + \ldots + \theta_k < 1, s > 0$, $x_j \geq 0$.

Let $k = 3$, $s = 2$ and the counts be

i	$(i) = (i_1, i_2, i_3)$	$(i') = (i_1', i_2', i_3')$
1	(0, 0, 0)	(2, 2, 2)
2	(1, 1, 1)	(3, 3, 3)

(8a)

From (8a), we have $a_1 = a_2 = -2, b_1 = 2, b_2 = 4$.

If $\theta_1 = .1$, $\theta_2 = .2$, $\theta_3 = .3$, then we have from (8), $m_1 = s\theta_1/A = .5$, $\sigma_1^2 = s\theta_1(A + \theta_1)/A^2 = .625$, $P(1_1') = .00361$, $P(2_1') = .00058$, $C = -.00838$, $B = -.01908$ and with these (7) reduces to

$$(y - 1.6521)^2 = -(.00011)(\lambda + 76.1843)^2 \tag{9}$$

If $2m_1 C < B < 2C(m_1 + C)$, we have the vertex of the parabola in the region $0 \leq \lambda \leq 1$ and we get a substantial part of parabola in (7). Otherwise, we get only a part of right or the left arm of the parabola.

3. Asymptotic properties

Consider again (8) and sample size n. Then the likelihood function can be written as

$$L(x_1, \ldots, x_k ; \theta_1, \theta_2, \ldots, \theta_k) = L(\underset{\sim}{x}, \underset{\sim}{\theta}) =$$

$$= \prod_{i=1}^{r} [P(i) + \lambda_i P(i')]^{n(i)} [(1 - \lambda_i) P(i')]^{n(i)} \prod [P(l)]^{n(l)} \tag{10}$$

where $P(i)$, $P(i')$, $P(l)$ are as in (2), $n(i) = n(i_1, \ldots, i_k)$, $n(i') = n(i_1', \ldots, i_k')$, and

$n(l)=n(l_1,\ldots,l_k)$ are the number of observations at i, i' and l respectively. Product Π is on all counts except $2r$ counts i and i' $(i=1, 2,\ldots,r)$. From (10), we get

$$D_{\theta_m}=[(\partial/\partial\theta_m)\ \log\ L]=\sum_{i=1}^{r} n(i)\left[\frac{i_m P(i)+\lambda_i P(i') i_m'}{\theta_m Q(i)}\right]$$

$$+\sum_{i=1}^{r}\left[\frac{i_m' n(i')+\sum l_m n(l)}{\theta_m}\right]-\left[\frac{ns}{A}\right] \tag{11}$$

where $Q(i)=P(i)+\lambda_i P(i')$ and $\sum$ is on all tuples except $2r$ tuples i and i'. And from (8), we have

$$[(\partial/\partial\theta_m)f(\underset{\sim}{x},\ \underset{\sim}{\theta})]=\left(\frac{x_m}{\theta_m}-\frac{s}{A}\right) f(\underset{\sim}{x},\underset{\sim}{\theta})\quad m=1,2,\ldots,k.$$

Similarly, we have for $m\neq q$

$$\left[(\partial^2/\partial\theta_m\partial\theta_q)\ (\log L)\right]=D^2_{\theta_m\theta_q}=\sum_{i=1}^{r}\frac{\lambda_i n(i) P(i) P(i')}{\theta_m\theta_q Q^2(i)}\ [(i_m'-i_m)(i_q'-i_q)]-ns/A^2 \tag{12}$$

and

$$D^2_{\theta_m}=\left[(\partial^2/\partial\theta^2_m)\ (\log L)\right]=\sum_{i=1}^{r}\frac{n(i)}{\theta_m^2 Q^2(i)}\ [\{P(i) i_m(i_m-1)+P(i') i_m'(i_m'-1)\}\{Q(i)\}$$

$$-\{i_m P(i)+\lambda_i i_m' P(i')\}^2]+\left[-\sum_{i=1}^{r} n(i') i_m'/\theta_m^2-\sum l_m n(l)/\theta_m^2-\frac{ns}{A^2}\right] \tag{13}$$

Similarly, we have

$$D_{\lambda_i}=[(\partial/\partial\lambda_i)\ (\log L)]=\frac{n(i) P(i')}{Q(i)}-\frac{n(i')}{1-\lambda_i}. \tag{14}$$

Also from $D_{\lambda_i}=0$, it follows

$$\hat{\lambda}_i=\frac{n(i) P(i')-n(i') P(i)}{P(i')[n(i)+n(i')]}, \tag{15}$$

and from (14) we have, using $D_{\lambda_i}=0$,

$$D^2_{\lambda_i}=[(\partial^2/\partial\lambda_i^2)\log L]=-n(i')[P(i)+P(i')]/(1-\lambda_i)^2 Q^2(i) \tag{16}$$

and also

$$D^2_{\theta_m\lambda_q}=[(\partial^2/\partial\theta_m\partial\lambda_q)\ (\log L)]=n(q)[P(q) P(q')\ (q_m'-q_m)]/\theta_m Q^2(q) \tag{17}$$

where $q\neq m$, $q=1, 2,\ldots, r$, $m=1, 2,\ldots, k$.

Example : Consider the case of three variate negative binomial ($k=3$) in (8), and let the two counts be as in (8a). Now we have the following;

(i)	(i_1, i_2, i_3) = count	$P(i)$	
(1)	$(1_1, 1_2, 1_3) = (0, 0, 0)$	$P(1) = c_1 A^s$	(17a)
(2)	$(2_1, 2_2, 2_3) = (1, 1, 1)$	$P(2) = c_2\theta_1\theta_2\theta_3 A^s$	
(i′)	(i_1', i_2', i_3') = count	$P(i')$	
(1′)	$(1_1', 1_2', 1_3') = (2, 2, 2)$	$P(1') = c_3\theta_1^2\theta_2^2\theta_3^2 A^s$	(17b)
(2′)	$(2_1', 2_2', 2_3') = (3, 3, 3)$	$P(2') = c_4\theta_1^3\theta_2^3\theta_3^3 A^s$	

In (17a), (17b), c_1, c_2, c_3, c_4 are constant functions of s.

$$Q(1) = P(1) + \lambda_1 P(1') \qquad A(2) = P(2) + \lambda_2 P(2') \tag{17c}$$

Variance-covariance matrix is 5×5 in size corresponding to the parameters θ_1, θ_2, θ_3 and λ_1, λ_2. We provide a few typical elements of this 5×5 matrix below, by which others can be easily evaluated. From now on, $n(1)$, $n(2), n(1'), n(2')$ are the number of observations at counts (1), (2), (1′), (2′) respectively.

From (16), we have

$$D^2_{\lambda_1} = -n(1')[P(1)+P(1')]/(1-\lambda_1)^2 Q(1) \text{ and} \tag{17d}$$

$$D^2_{\lambda_1\lambda_2} = 0 \tag{17e}$$

In (12), let $m=1$, $q=3$, then

$$D^2_{\theta_1\theta_3} = \frac{\lambda_1 n(1) P(1) P(1')}{\theta_1\theta_3 Q^2(1)}[(1_1'-1_1)(1_3'-1_3)] + \frac{\lambda_2 n(2) P(2) P(2')}{\theta_1\theta_3 Q^2(2)}[(2_1'-2_1)(2_3'-2_3)] - \frac{ns}{A^2} \tag{17f}$$

In (17f), $(1_1'-1_1) = (1_3'-1_3) = 2 = (2_1'-2_1) = (2_3'-2_3)$.

If $m=3$ in (13), then

$$\begin{aligned} D^2_{\theta_3} = {} & \frac{n(1)}{\theta_3^2 Q^2(1)}[\{P(1)(1_3)(1_3-1)+P(1')(1_3')(1_3'-1)\}\{Q(1)\} \\ & -\{(1_3)P(1)+\lambda_1(1_3')P(1')\}^2] \\ & + \frac{n(2)}{\theta_3^2 Q(2)}[\{P(2)(2_3)(2_3-1)+P(2')(2_3')(2_3'-1)\}\{Q(2)\} \\ & -\{(2_3)P(2)+\lambda_2(2_3')P(2')\}^2] \\ & -\left[\frac{n(1')(1_3')}{\theta_3^2}+\frac{n(2')(2_3')}{\theta_3^2}\right]-\left[\frac{\sum l_m n(l)}{\theta_3^2}\right]-\left[\frac{ns}{A^2}\right] \end{aligned} \tag{17g}$$

In (17g), $(1_3)=0$, $(1_3')=2$.

Again, let $m=3$, $q=2$, in(17), then

$$D^2_{\theta_3\lambda_2}=n(2)\,[P(2)\,P(2')\,(2_3'-2_3)\,]/\theta_3 Q^2(2). \tag{17h}$$

Other elements of the matrix follow from those of (17a) to (17h).

4. Bayesian Estimates

From (10), we have

$$L(\underset{\sim}{x},\underset{\sim}{\theta})=\prod_{i=1}^{r}\left[\sum_i(\Omega_i)\,\lambda_i^{n(i)-t_i}\,(1-\lambda_i)^{n(i')}\right](C_0)\,(A^{ns})\cdot\left[\prod_{m=1}^{k}\theta_m^{v_m}\right] \tag{18}$$

where $\sum_i=\sum_{t_i=0}^{n(i)}$, $\Omega_i=\binom{n(i)}{t_i}$,

$$C_0=\prod_{i=1}^{r}\left[\frac{\Gamma(s+i_1+\ldots+i_k)}{\Gamma(s)\,i_1!\cdots\cdots i_k!}\right]^{n(i)}\left[\frac{\Gamma(s+i_1'+\ldots+i_k')}{\Gamma(s)\,i_1'!\cdots\cdots i_k'!}\right]^{n(i')}\cdot\Pi\left[\frac{\Gamma(s+l_1+\ldots+l_k)}{\Gamma(s)\,l_1!\cdots\cdots l_k!}\right]^{n(l)},$$

product Π is on all tuples except $2r$ tuples i and i', and

$$v_m=\sum_{i=1}^{r}i_m t_i+\sum_{i=1}^{r}i_m'[n(i')+n(i)-t_i]+\sum l_m n(l)\quad m=1,2,\ldots,k.$$

Now take the prior for θ's as Dirichlet's prior

$$g(\underset{\sim}{\theta})=g(\theta_1,\ldots,\theta_k)=\left(\prod_{i=1}^{k}\theta_i^{f_i-1}\right)\,A^{f_{k+1}-1}/B(f_1,\ldots,f_k\,;\,f_{k+1}) \tag{19}$$

where

$$B(f_1,\ldots,f_k\,;\,f_{k+1})=\frac{\Gamma(f_1)\ldots\Gamma(k)\,\Gamma(f_k+f_1)}{\Gamma(f_1+\ldots+f_k+f_{k+1})}.$$

Similarly take prior for λ_i's as

$$g(\underset{\sim}{\lambda})=g(\lambda_1,\ldots,\lambda_r)=\prod_{i=1}^{r}\left[\lambda_i^{d_i-1}(1-\lambda_i)^{e_i-1}/B(d_i,e_i)\right]. \tag{20}$$

From (18), (19) and (20), we get the Bayesian estimates of θ_m's as

$$\hat{\theta}_m=\frac{\prod_{i=1}^{r}\sum_i(\Omega_i)B(h_i,\,g_i)\;B(w_1,\ldots,w_m+1,\ldots,w_k\,;\,w_{k+1})}{\prod_{i=1}^{r}\sum_i(\Omega_i)B(h_i,\,g_i)\;B(w_1,\ldots,w_m,\ldots,w_k\,;\,w_{k+1})} \tag{21}$$

where $w_j=v_j+f_j$, $j=1,2,\ldots,k$, $w_{k+1}=ns+f_{k+1}$, $h_i=n(i)-t_i+d_i$ and $g_i=n(i')+e_i$.

For four points in (8a) with $(l_1,l_2,l_3)=(4,4,4)$, $n=10$, $r=2$, $k=3$, $s=2$, $n(l)=n(l_1,l_2,l_3)=4$ and noting $n(i)=(i_1,i_2,i_3)$, $n(i')=n(i_1',i_2',i_3')$ and $n(1)=1$, $n(2)=2$, $n(1')=2$, $n(2')=1$, and $f_1=f_2=f_3=f_4=d_1=d_2=e_1=e_2=2$, we have from (21), Bayesian estimate of θ_1 as

$$\hat{\theta}_1=\frac{\sum_1\sum_2\binom{1}{t_1}\binom{2}{t_2}B(3-t_1,4)\,B(4-t_2,3)\,B(T+1,\,T,\,T\,;\,22)}{\sum_1\sum_2\binom{1}{t_1}\binom{2}{t_2}B(3-t_1,4)\,B(4-t_2,3)\,B(T,\,T,\,T\,;\,22)} \tag{22}$$

where $T=33-2t_1-2t_2$.

Again from (18), (19) and(20), we have the Bayesian estimate of λ_i as

$$\hat{\lambda}_i=\frac{\prod_{\substack{j=1\\ j\neq i}}^{r}[\sum_j(\Omega_j)B(h_j,g_j)]\sum_i(\Omega_i)B(h_i+1,g_i)B(w_1,\ldots,w_k\,;\,w_{k+1})}{\prod_{i=1}^{r}\sum_i(\Omega_i)B(h_i,g_i)B(w_1,\ldots,w_k\,;\,w_{k+1})}. \quad (23)$$

From (23), we get for the same data,

$$\hat{\lambda}_1=\frac{\sum_1\sum_2\binom{1}{t_1}\binom{2}{t_2}B(4-t_1,4)B(4-t_2,3)B(T,T,T\,;\,22)}{\sum_1\sum_2\binom{1}{t_1}\binom{2}{t_2}B(3-t_1,4)B(4-t_2,3)B(T,T,T\,;\,22)}. \quad (24)$$

Comments : In our analysis, we have considered only simple case where a set of r distinct counts $(i_1,\ldots,i_k)$ $i=1,2,\ldots,r$ are misreported as another set of r distinct counts $(i_1',\ldots,i_k')$, $i=1,2,\ldots,r$. If we take $i_j'=i_l'$ or $i_j=i_l$, $j,l=1,2,\ldots,k$, analysis will be slightly more complex. Similarly, we may consider m sets of r counts each being misreported as another m sets of r counts each. Further variation may be when each of these m sets have $r_1,\ldots,r_m$ counts instead of the same number of r counts in each set.

Again as another extension, we may consider either in the case of a single set or in the case of m sets, that each set of counts being misreported as another single count. That is, for example, all the r counts $(i_1,\ldots,i_k)$, $i=1,2,\ldots,r$ may be misreported as a single count $(i_1',\ldots i_k')$, $i=1$. All these generalizations may make algebra slightly complex but the main structure of the analysis remains the same.

References

(1) Cohen A.C. (1960a) Estimating the parameters of a modified Poisson distribution. *Journal of American Statistical Association*, Vol. 55, pp. 139~143.

(2) Cohen, A.C. (1960b) Misclassified data from a binomial population. *Technometrics*, Vol. 2, pp. 109~113.

(3) Cohen, A.C. (1960c) Estimation in the Poisson distribution when the sample values $(c+1)$ are sometimes erroneously as c. *Annals of Institute of Statistical Mathematics*, Vol. 9, pp. 181~193.

(4) Lingappaiah, G.S. (1977) On some discrete distributions with varying probabilities. *Egyptian Statistical Journal*, Vol. 21, pp. 1~15.

(5) Lingappaiah, G.S. (1978) Further investigations into the discrete distributions with jumps in probabilities. *Philippine Statistician*, Vol. 27, pp. 26~35.

(6) Lingappaiah, G.S. and Patel, I.D. (1979) On the modified and inflated discrete distributions of general type. *Gujarat Statistical Review*, Vol. 6, No. 2, pp. 50~60.

(7) Parikh, N.T. and Shah, S.M. (1969) Misclassification in power series distribution in which the value one is sometimes reported as zero. *Journal of Indian Statistical Association*, Vol. 7, pp. 11~19.

(8) Varahamurthy, K. (1967) A modified Poisson distribution. *Portugalai Mathematica*, Vol. 26, pp. 319~328.

(9) Williford, W.O. and Bingham, S.F. (1979) Bayesian estimation of the parameters in two modified Poisson distributions. *Communications in Statistics*, Vol. A8(13), pp. 1315~1326.