FURTHER INVESTIGATIONS INTO THE DISCRETE DISTRIBUTIONS WITH JUMPS IN PROBABILITIES

by

G.S. Lingappaiah[1]

Summary

Discrete distributions where the probabilities are affected in the sense "inflated" and "modified" are considered. In the former case, a situation is considered where among k (m + 1) counts, each of (m + 1) counts are inflated at the same rate, while these counts are spread out in a cycle and thereby called as "cyclical inflation." Test procedure for the hypothesis that there is only one rate of inflation, that is, $\lambda_o = \lambda_1 = \ldots = \lambda_k$ is attempted. In the modified case, firstly, set of m counts are considered where each is being misreported as a set of k counts, and hence called "spectral modification." Second case considered under modification is where a set of m counts being reported as a single count and this situation is termed as "ray-convergent modification." In these two cases, estimation of the parameters is being carried out both by maximum likelihood (m.l.) method as well as by Bayes' approach.

I. Introduction

Recently much work is being done in discrete distributions where the probabilities at different counts are affected. One such case is where there are excess of zeroes, thereby reducing the probabilities at other counts. Such a situation termed as "infla-

[1]The author is a professor in the Department of Mathematics, Sir George Williams Campus, Concordia University; Montreal, Canada.

tion" has been dealt with recently by many authors, i.e. Sobic and Szynal [5], Singh [6], Singh [7] and others. Distribution of the sum of variables from such an inflated distribution, as well as the estimation problem is taken up in these works. Similarly, another situation dealt is where a count or counts are misreported as some other count. This case has been termed as "modification" and is dealt by Cohen [1], [2] and Varahamurthy [8]. This author [4] has recently considered a case where a set of k counts are misreported and the estimation of parameters under such circumstances is attempted. What we have done here is to attempt further generalizations of these situations. Firstly treated is the inflated case where among k (m + 1) counts, probabilities at (m + 1) counts are inflated at the same rate. That is, the probabilities at the counts $i_j + ht_j$, h = 0, 1, 2, . . . m are all inflated at the same rate λ_j $(0 < \lambda_j < 1)$ while the probabilities at rest of the counts except at x = 0 are inflated at the rate λ_0 $(0 < \lambda_0 < 1)$. Test procedure for the hypothesis $\lambda_0 = \lambda_1 = \ldots = \lambda_k$ is attempted for a particular case when h = 0. Regarding modification, two cases are considered. Firstly, we deal with a situation where among km counts, each of m counts $i_1, i_2, \ldots i_m$ are misreported as the counts h_i (i = 1, . . . k) and hence termed as "ray-convergent modification." Second situation is where a single count h_j (j = 1, 2, . . . m) is being misreported as a set of counts $l_j \ldots k_j$ and thereby named as "spectral modification." In the case of "ray-convergent modification," estimation is attempted via m.l. method while in the case of "spectral modification," Bayes' approach is used.

2. Inflated Case (Cyclical Inflation)

Here we consider k (m + 1) counts $i_j + ht_j = u_j$, h = 0, 1, 2, . . . m; j = 1, 2, . . . k (i is some arbitrary subscript) and let the probabilities at the counts i_j, $i_j + t_j$, $i_j + 2t_j$, . . . $i_j + mt_j$ be inflated at the rates λ_j $(0 < \lambda_j < 1)$ while the probabilities at all other counts except at x = 0, at the rate λ_0 $(0 < \lambda_0 < 1)$. In this case, the density can be written as

$$1 - \lambda_0 + \lambda_0 P_0 - \sum_{j=1}^{k} \sum_{h=0}^{m} (\lambda_j - \lambda_0) P_{u_j} \text{ if } x = 0$$

$$\lambda_j P_{u_j} \text{ if } x = u_j \qquad (1)$$

$$h = 0, 1, 2, \ldots m$$

$$j = 1, 2, \ldots . k$$

$$\lambda_0 P_x \text{ if } x \neq 0, u_j$$

where $P_o = P(x = 0)$ in the non-inflated (simple) distribution and P_x accordingly.

From (1), we have

$$\mu'_1 \doteq m + A \qquad (2)$$

where $A = \sum_{j,h} (\lambda_j - \lambda_o)(u_j) P_{u_j}$ and μ'_r is the r^{th} raw-moment of the inflated distribution while m'_r is the corresponding moment for the simple distribution ($m'_1 = m$). Similarly, obtaining μ'_2, we have,

$$y = \left[\lambda_0 + \lambda_0 (1 - \lambda_0) \frac{m^2}{\sigma^2}\right] + \frac{1}{\sigma^2}\left[B^2 - 2m \lambda_0 A - A^2\right] \qquad (3)$$

where $\sigma^2 = m_2$, $y = \mu_2/\sigma^2$ and $B = \sum_{j,h} (\lambda_j - \lambda_0)(u_j^2) P_{u_j}$. If all the inflated rates are the same, then the second part of (3) on the right hand side variables and the first part is an inverted parabola when graphed on (y, λ_o) axes, and truncated at right at $\lambda_0 = 1$. In general y depends partly on the signs of $(\lambda_j - \lambda_0)$. Now we proceed towards the test procedure for the hypothesis $\lambda_0 = \lambda_1 = \ldots = \lambda_k$. Now, from (1), we have the corresponding likelihood function, on the lines of [1] and [8] and can be written as

$$(Q_0)^{n_0} \prod_{j=1}^{k} \prod_{h=0}^{m} (\lambda_j P_{u_j})^{n_{u_j}} \prod_{t \neq 0, u_j} (\lambda_0 P_t^{n_t}) \qquad (4)$$

Where $Q_0 = 1 - \lambda_0 + \lambda_0 P_0 - \sum_{j,h} (\lambda_j - \lambda_0) P_{uj}$ and from (4) we have $D\lambda_0$ and $D\lambda_j$ respectively (where $D\lambda = \partial \log L/\partial\lambda$) as

$$\frac{-n_0 (1 - P_0 - \bar{P})}{Q_0} + \frac{n - \bar{n}_s - n_0}{\lambda_0} = 0 \tag{5}$$

$$\frac{-n_0 \sum_{h=0}^{m} P_{uj}}{Q_0} + \frac{n_{uj}}{\lambda_j} = 0 \tag{6}$$

with $\bar{P} = \sum_{j,h} P_{uj}$, $\bar{n}_s = \sum_{j,h} n_{uj}$. We consider below a simple case with $h = 0$. *θ known.* (5) and (6) respectively, now are,

$$\frac{-n_0 (1 - P - P_0)}{Q} + \frac{n - n_s - n_0}{\lambda_0} = 0 \tag{7}$$

$$\frac{-n_0 P_{ij}}{Q} + \frac{n_{ij}}{\lambda_j} = 0 \tag{8}$$

where $Q = 1 - \lambda_0 + \lambda_0 P_0 - \sum_{j=1}^{k} (\lambda_j - \lambda_0) P_{ij}$ and $P = \sum_j P_{ij}$, $n_s = \sum_j n_{ij}$. Taking all such k equations in (8), we get

$$\hat{\lambda}_j = n_{ij} (1 - \lambda_0 B)/(n_0 + n_s) P_{ij} \tag{9}$$

where $B = 1 - P_0 - P$. Now using (7), we get

$$\hat{\lambda}_0 = A/nB \text{ where } A = n - n_s - n_0. \tag{10}$$

Now, if we wish to test the hypothesis $H_0: \lambda_0 = \lambda_1 = \ldots = \lambda_k$, that is, there exists only one rate of inflation, we have for the estimate of λ_0 under H_0 as

$$(n - n_0)/n\,(1 - P_0) \qquad (11)$$

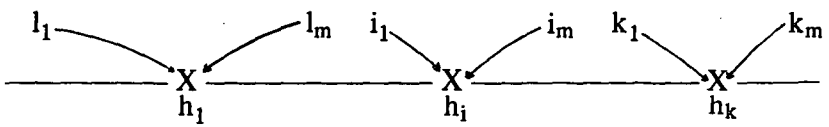
and hence the likelihood criterion $\lambda = L(\hat{w})/L(\hat{\Omega})$ where w has just an element λ_0 while Ω has k + 1 elements $\lambda_0, \lambda_1, \ldots \lambda_k$. Then from (9), (10), (11), we have

$$\lambda = \prod_{j=1}^{k} \frac{P_{i_j}}{n_{i_j}/n}^{n_{i_j}} \frac{1 - n_0/n}{1 - P_0}^{n - n_0} \frac{nB}{A}^{A} \qquad (12)$$

Tests can be carried out for either very low values or high values of λ.

3. Modified Case

3(a). Ray-convergent modification. Here we deal with first of two situations under modified case. First one is, where a set of km counts are considered to be misreported. That is, each of m counts $i_1, \ldots i_m$ are reported as count h_i, i = 1, 2, . . . k.



Now the density can be written as

$$f(x) = \begin{cases} (1 - \lambda_j)\,P_{i_j} & \text{if } x = i_j \\ & j = 1, \ldots m \\ & i = 1, \ldots k \\ P_{h_i} + \sum_{j=1}^{m} \lambda_j P_{i_j} & \text{if } x = h_i \\ P_x & \text{if } x \neq h_i, i_j \end{cases} \qquad (13)$$

Corresponding likelihood function can be expressed as

$$\prod_{i=1}^{k} \prod_{j=1}^{m} (1-\lambda_j)\, P_{ij}^{\,n_{ij}}\, P_{h_i} + \sum_j \lambda_j P_{ij}^{\,n_{h_i}} \prod_{t \neq h_i, i_j} (P_t^{n_t}) \qquad (14)$$

and hence $D_{\lambda_j} = 0$ (with $D_\lambda = \partial \text{ Log } L/\partial\lambda$), is

$$\sum_{i=1}^{k} \left[\frac{-n_{ij}}{1-\lambda_j} + \frac{n_{h_i} P_{ij}}{Q_2} \right] = 0 \qquad (15)$$

where $Q_2 = P_{h_i} + \sum_j \lambda_j P_{ij}$.

If $k = 1$, and with all equations for all λ_j's in (15), we have,

$$\sum_j \lambda_j P_{ij} = \frac{n_{n_1} P_s - n_s P_{h_1}}{n_s + n_{h_1}} \qquad (16)$$

where $n_s = \sum_j n_{1_j}$ and $P_s = \sum_j P_{1_j}$. From (16), we get

$$\hat{\lambda}_j = 1 - \frac{n_{1_j}}{P_{1_j}} \frac{(P_{h_1} + P_s)}{(n_{h_1} + n_s)} \qquad (17)$$

Similarly $D_\theta = 0$ gives, (with P' for $\partial P/\partial\theta$)

$$\sum_{j=1}^{m} n_{1_j} \frac{P'_{1_j}}{P_{1_j}} + \frac{n_{h_1}(P'_{h_1} + \sum_j \lambda_j P'_{1_j})}{Q} + \sum_{t \neq 1_j, h_1} n_t \frac{P'_t}{P_t} = 0 \qquad (18)$$

Example: Now as an illustration, consider the famous "horse-kicks" example cited in Cohen [1]. Suppose, we modify the data as follows.

No. of deaths per army corps per year	Original data	Modified data
0	109	105

No. of Deaths per Army Corps per Year	Original Data	Modified Data
1	65	62
2	22	29
3	3	3
4	1	1
5	0	0

That is, counts 0 and 1 are misreported as count 2. Then in our notation $n_0 = n_{l_1} = 105$, $n_1 = n_{l_2} = 62$, $n_{h_1} = 29 = n_2$, $n_3 = 3$, $n_4 = 1$, $n_5 = 0$. Then from (17), we have

$$\hat{\lambda}_1 = 1 - \frac{105}{196}(c) \text{ where } c = 1 + \theta + \theta^2/2$$

$$\hat{\lambda}_2 = 1 - \frac{62}{196}(c/\theta) \tag{19}$$

and (18) gives

$$\frac{75}{\theta} + 29 \quad \frac{\theta + \hat{\lambda}_2}{(\theta^2/2) + \hat{\lambda}_1 + \hat{\lambda}_2 \theta} - 200 = 0 \tag{20}$$

Now using (19) in (20) we can solve for $\hat{\theta}$ and in turn using this $\hat{\theta}$, we can obtain $\hat{\lambda}_1, \hat{\lambda}_2$.

Further, regarding the asymptotic distribution of λ_j 's, we can get most of the elements of the inverse of $(k+1) \times (k+1)$, variance-covariance matrix except $-E(D_\theta^2)$ term which turns out to be slightly messy. For example, with $k = 1$, in (15), and noting further in this case,

$$E(n_{l_j}) = n(1 - \lambda_j) P_{l_j} \tag{21}$$
$$E(n_{h_1}) = nQ$$

we have

$$-E \quad D^2_{\lambda_j} = \frac{nP_{l_j}}{1 - \lambda_j} + \frac{nP^2_{l_j}}{Q}$$

$$-E\ D^2_{\theta\lambda_j} = nP_{lj}\left(P'_{h_l} + \sum_j \lambda_j P'_{lj}\right) \Big/ Q \qquad (22)$$

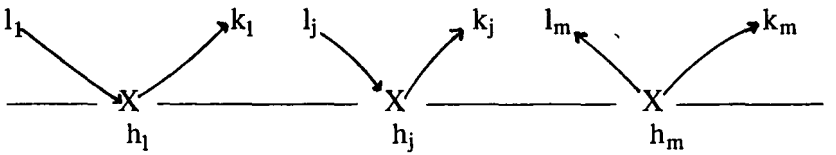
where Q is Q_2, and i = 1.

3 (b). Spectral Modification.

Now we consider a case which is slightly opposite to 3(a). Here we deal with a situation where a single count has been misreported as other counts. That is, the count h_j is being modified as $1_j, 2_j, \ldots k_j$; j = 1, . . . m. Hence the density function can be written as

$$P_{i_j} + \lambda_i P_{h_j} \quad \text{if } x = i_j, j = 1, \ldots m \quad i = 1, 2, \ldots k$$

$$(1 - \lambda_1 - \lambda_2 \ldots - \lambda_k) P_{h_j} \quad \text{if } x = h_j \qquad (23)$$



In this case, for the purpose of estimating λ_j 's, we resort to Bayes approach with a restriction $\sum_i \lambda_i < 1$.

For the estimation purposes, in the case of generalized distributions, Bayes' method seems to be more favourable than that of m.l. method. Especially so, when we are dealing with some situations like "inflation" or "modification."

In [3], this author has tried this approach for the estimation problem in the case of generalized distributions. Now, in this case, we can write the density as

(24)

$$\prod_{j=1}^{m} \prod_{i=1}^{k} P_{i_j} + \lambda_i P_{h_j}{}^{n_{ij}} \left(1 - \sum_i \lambda_i\right) P_{h_j}{}^{n_{h_j}} \prod_{t \neq i_j, h_j} (P_t^{n_t})$$

and now taking the prior for λ_i's as Dirichlet's distribution

$$f(\lambda_1, \ldots \lambda_k) = \frac{\lambda_1^{\delta_1 - 1} \ldots \lambda_k^{\delta_k - 1} (1 - \lambda_1 \ldots - \lambda_k)^{\delta - 1}}{B(\delta_1, \ldots \delta_k; \delta)}$$

$$0 < \lambda_i < 1, i = 1, \ldots k, \Sigma \lambda_i < 1 \qquad (25)$$

and $B(\delta_1, \ldots \delta_k; \delta) = \Gamma(\delta_1)\, \Gamma(\delta_2) \ldots \Gamma(\delta_k)\, \Gamma(\delta) / \Gamma(\delta_1 + \ldots + \delta_k + \delta)$

Using (24), (25), we have

$$f(n_{i_j}, n_{h_j}) = \prod_{j=1}^{m} \prod_{i=1}^{k} \sum_{r_{ij}=0}^{n_{ij}} \binom{n_{ij}}{r_{ij}} P_{ij}^{n_{ij} - r_{ij}} P_{hj}^{n_{hj} + r_{ij}}$$

$$= \qquad (26)$$

$$\cdot \Gamma(\delta + n_h)\, \Gamma(r_{i\cdot} + \delta_i) / \Gamma(r + \delta_0 + \delta + n_h)$$

$$\cdot 1/B(\delta_1, \delta_2 \ldots \delta_k; \delta)$$

where $\delta_0 = \delta_1 + \ldots \delta_k$, $r = \sum_{i=1}^{k} r_{i\cdot}$, $r_{i\cdot} = \sum_j r_{ij}$, $n_h = \sum_j n_{hj}$

From (26), we get the Bayes' estimate of λ_i as (27)

$$E(\hat{\lambda_i}) = \frac{\prod_{i=j} \sum r_{ij}^{n_{ij}} \quad \frac{P_{hi}}{P_{ij}}^{r_{ij}} \frac{\Gamma(r_{i\cdot} + \delta_i + 1)\, \Gamma'_{t \neq i}(r_{t\cdot} + \delta_t + 1)}{\Gamma(r + \delta_0 + \delta + n_h + 1)}}{(a)};$$

where (a) is exactly the numerator of (27) except $t = i$. That is, now the Gamma functions are replaced by

$$\prod_{i=1}^{k} \Gamma(r_i + \delta_i)/\Gamma(r + \delta_0 + n_h) \cdot$$

(The Γ' in (27) is a product of $(k - 1)$ gamma functions.)

4. References

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

[2] Cohen, A.C. (1960) 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.

[3] Lingappaiah, G.S. (1976) Effect of outliers in the estimation of parameters. *Metrika,* Vol. 23, pp. 27-30.

[4] Lingappaiah, G.S. (1976) On Some discrete distributions with varying probabilities. (Submitted.)

[5] Sobic, Lusza, and Szynal, Dominic. (1974) Some properties of inflated binomial distribution. *Canadian Mathematical Bulletin,* Vol. 17, pp. 609-611.

[6] Singh, S.N. (1963) A note on inflated Poisson distribution. *Journal of Indian Statistical Association, Vol. 1, pp.* 140-144.

[7] Singh, M.P. (1965) Inflated binomial distribution. *Journal of Scientific Research* (Benares Hindu University), Vol. 16, pp. 87-90.

[8] Varahamurthy, R. Krishna. (1967) A modified Poisson distribution. *Portugaliae Mathematica,* Vol. 26, pp. 319-328.