ESTIMATION OF PARAMETERS OF A MIXTURE OF TWO OR MORE POISSONIAN POPULATIONS FROM A CENSORED SAMPLE

By Naunihal Singh

Defence Science Laboratory, Delhi

(Received: November, 1959)

1. Introduction

In practice the problem of estimation of parameters of a heterogeneous population[1] is encountered on many occasions, specially in the production of items on a large scale by a single machine or a group of machines producing similar types of articles, where the chance of an item to be defective may be attributed not only to one cause but several causes, for example, the defective material, defective machines, human error, etc. The Poisson distribution, as is well known, is the appropriate mathematical model for representing such type of data as the number of defects in a manufactured product. It is, thus, of interest to the quality control engineers and research workers who are often confronted with the problem like that of estimating means and variances of the number of defects, particularly when samples are either truncated or censored.[2] To simplify this an attempt has been made in this paper to work out certain devices for determining the means and the variances from a censored sample taken from a mixture of two Poissonian sub-populations mixed in an unknown proportion, $p_1 : p_2$ where p_2 is the complement of p_1. The more general case has also been considered in Section 6. The causes here are assumed to be independent.

The estimates of the parameters have been obtained by the method of maximum likelihood. The solution of the equations is arrived at with the help of iterative process by using *Tables of Poisson Distribution* by Tosio Kitagawa.[3] The information matrix is given for evaluating the standard errors of the estimates. To illustrate the practical application of the results a numerical example is added at the end of the paper.

2. The Population Model

Let the parent population be composed of two sub-populations mixed in the proportion mentioned above, each following the Poisson distribution. The variate x_i (i = 1, 2) of the i-th sub-population is assumed to have the cumulative distribution function,

$$\Phi_i(c) = 1 - \sum_{x=c+1}^{\infty} \phi_i(x)$$

where

$$\phi_i(x) = \frac{e^{-m_i} m_i^x}{x!}, \quad 0 \leq x \leq \infty. \tag{2.1}$$

If p_1 be the proportion of units pertaining to $\Phi_1(c)$ and p_2 to $\Phi_2(c)$, then the population model may be written as

$$\Phi(c) = p_1\Phi_1(c) + p_2\Phi_2(c) \tag{2.2}$$

and the density function

$$\phi(c) = p_1\phi_1(c) + p_2\phi_2(c) \tag{2.3}$$

and also

$$A_i = 1 - \Phi_i(c) \tag{2.4}$$

$$A = 1 - \Phi(c) \tag{2.5}$$

where A is the probability that a unit falls outside the range $(0 - c)$, c is the censored point.

3. The Maximum Likelihood Estimation

Given a random sample of size n, the probability density of r_1 units belonging to $\Phi_1(c)$, r_2 units belonging to $\Phi_2(c)$ and $(n - r)$ units falling outside the range $(0 - c)$, may be written as

$$P(r_1, r_2, (n-r)/n) = \frac{n!}{r_1!\, r_2!\, (n-r)!} [p_1\Phi_1(c)]^{r_1} [p_2\Phi_2(c)]^{r_2} A^{n-r};$$
$$r = r_1 + r_2. \tag{3.1}$$

The conditional density of ordered observations $x_{i_1}, x_{i_2}, \ldots x_{ir_i}$, given r_i is

$$P(x_{i_1}, x_{i_2}, \ldots x_{ir_i}/r_i) = \frac{r_i! \prod_{j=1}^{r_i} \phi_i(x_{ij})}{[\Phi_i(c)]^{r_i}}. \tag{3.2}$$

The likelihood function may now be written as

$$L = K \prod_{j=1}^{r_1} \phi_1(x_{1j}) \prod_{j=1}^{r_2} \phi_2(x_{2j})\, p_1^{r_1} p_2^{r_2} A^{n-r}, \qquad (3.3)$$

where

$$K = \frac{n!}{(n-r)!}. \qquad (3.4)$$

Taking log of (3.3) and differentiating with respect to m_1, m_2 and p_1 we get

$$\left.\begin{aligned} \frac{\partial \log L}{\partial m_1} &= -r_1 + \frac{r_1 \bar{x}_1}{m_1} + \frac{n-r}{A} p_1 \phi_1(c) \\ \frac{\partial \log L}{\partial m_2} &= -r_2 + \frac{r_2 \bar{x}_2}{m_2} + \frac{n-r}{A} p_2 \phi_2(c) \\ \frac{\partial \log L}{\partial p_1} &= \frac{r_1}{p_1} - \frac{r_2}{p_2} - \frac{n-r}{A} [\Phi_1(c) - \Phi_2(c)] \end{aligned}\right\}. \qquad (3.5)$$

On equating each of (3.5) to zero, the estimating equations are attained as

$$\left.\begin{aligned} \hat{m}_1 &= \frac{r_1 \bar{x}_1 A}{r_1 A - (n-r)\, p_1 \phi_1(c)} \\ \hat{m}_2 &= \frac{r_2 \bar{x}_2 A}{r_2 A - (n-r)\, p_2 \phi_2(c)} \\ \hat{p}_1^{\,2} &- \left\{ \frac{rA}{(n-r)\,[\Phi_1(c) - \Phi_2(c)]} + 1 \right\} \\ \hat{p}_1 &+ \frac{r_1 A}{(n-r)\,[\Phi_1(c) - \Phi_2(c)]} = 0 \end{aligned}\right\} \qquad (3.6)$$

where $\hat{m}_1$, $\hat{m}_2$ and $\hat{p}_1$ are the corresponding estimates of m_1, m_2 and p_1, the population parameters.

The solution of equations (3.6) is illustrated in Section 5.

4. Precision of Estimates

To determine the asymptotic values of variances and covariances of the estimates the information symmetric matrix $I_{(\hat{m}_1, \hat{m}_2, \hat{p}_1)}$ is given below which on inversion will give us the standard errors of the estimates:

$$I_{(\hat{m}_1, \hat{m}_2, \hat{p})} = nx \begin{bmatrix} p_1 h_{11} & \frac{1}{A} h_{12} & \frac{1}{A} h_{13} \\ .. & p_2 h_{22} & \frac{1}{A} h_{23} \\ .. & .. & h_{33} \end{bmatrix} \tag{4.1}$$

where

$$h_{11} = \left[\frac{1}{m_1}\left(1 - \frac{\phi_1(c)}{\Phi_1(c)}\right) - \left(1 - \frac{m_1}{c}\right)\phi_1(c-1) + \frac{p_1}{A}\phi_1{}^2(c)\right]. \tag{4.2}$$

$$h_{22} = \left[\frac{1}{m_2}\left(1 - \frac{\phi_2(c)}{\Phi_2(c)}\right) - \left(1 - \frac{m_2}{c}\right)\phi_2(c-1) + \frac{p_2}{A}\phi_2{}^2(c)\right]. \tag{4.3}$$

$$h_{33} = \left[\frac{\Phi_1(c)}{p_1} + \frac{\Phi_2(c)}{p_2} + \frac{\{\Phi_1(c) - \Phi_2(c)\}^2}{A}\right]. \tag{4.4}$$

$$h_{12} = p_1 p_2 \phi_1(c)\,\phi_2(c). \tag{4.5}$$

$$h_{13} = [\Phi_2(c) - \Phi_1(c)]\,p_1\phi_1(c) - A\phi_1(c). \tag{4.6}$$

$$h_{23} = [\Phi_2(c) - \Phi_1(c)]\,p_2\phi_2(c) + A\phi_2(c). \tag{4.7}$$

5. A Numerical Example

A random sample of size, $n = 4981$ is taken from *Tables of Poisson Distribution* by Tosio Kitagawa[3] which is constituted of two subsamples with $m_1 = 5$, $m_2 = 7$. Let the subsample (1) be due to some known cause and the subsample (2) due to some unknown cause. The censored point c is taken at 8. The data of Tables I and II yield the following:

$r_1 = 2314$; $r_2 = 1819$; $r = 4133$

$r_1\bar{x}_1 = 10837$; $r_2\bar{x}_2 = 10474$; $n - r = 848$

$\bar{x}_1 = 4{\cdot}68$; $\bar{x}_2 = 5{\cdot}76$; $p_1 = 0{\cdot}46$

Table I

Defects due to known cause

x_1	1	2	3	4	5	6	7	8
f_1	84	211	351	439	439	366	261	163

TABLE II

Defects due to unknown cause

x_2	1	2	3	4	5	6	7	8
f_2	16	56	130	226	319	373	373	326

Starting with the first set of approximations given above, the subsequent five approximations are determined from equations (3.6) by the use of iterative process. The results are given in Table III.

TABLE III

No.	Iterated values		
	$\hat{m}_1$	$\hat{m}_2$	$\hat{p}_1$
1	5·19	7·89	0·44
2	4·91	6·71	0·50
3	5·06	7·09	0·50
4	5·03	6·99	0·50
5	5·03	6·99	0·50

The oscillatory and converging tendency of the results, as it looks from Table III, is a sufficient indication to throw light upon the accuracy of the estimates.

6. THE MORE GENERAL POPULATION MODEL

Consider a mixture of M Poissonian sub-populations mixed in proportions $p_1 : p_2 : \ldots p_{M-1} : p_M$, where $0 \leq p_i \leq 1$ and $\sum_{i=1}^{M} p_i = 1$. Let the cumulative distribution function of the i-th sub-population be $\Phi_i(c, m_i)$ which is independent of the proportion p_i. The cumulative distribution function of the parent population then may be written as

$$\Phi(c, m_1, m_2, \ldots m_M) = \sum_{i=1}^{M} p_i \Phi_i(c, m_i). \qquad (6.1)$$

A random sample of size n is drawn from this general mixture and censored at the point c. Let r_i be the number of measured observations belonging to the i-th sub-population, where $\sum_{i=1}^{M} r_i = r$, then the likelihood function may be written as

$$L = K \prod_{i=1}^{M} p_i^{r_i} \prod_{j=1}^{r_1} \phi_1(x_{1j}) \prod_{j=1}^{r_2} \phi_2(x_{2j}) \ldots \prod_{j=1}^{r_M} \phi_M(x_{Mj}) A^{n-r} \quad (6.2)$$

where

$$A = 1 - \Phi(c, m_1, m_2, \ldots m_M) \quad (6.3)$$

and

$$K = \frac{n!}{(n-r)!}. \quad (6.4)$$

Following the procedure adopted in Section 3, the estimating equations are

$$\left.\begin{aligned} \hat{m}_i &= \frac{r_i \bar{x}_i A}{r_i A - (n-r) p_i \phi_i(c)}, \quad i = 1, 2, \ldots M \\ \hat{p}_i &= \frac{(n-r)}{A} [\Phi_M(c) - \Phi_i(c)] \\ &\quad + \frac{r_i}{p_i} - \frac{r_M}{p_M}, \quad i = 1, 2, \ldots M-1 \end{aligned}\right\}. \quad (6.5)$$

The required estimates $\hat{m}_i$ and $\hat{p}_i$ can be obtained by solving $2M-1$ equations of (6.5) simultaneously, which will give us $2M-1$ unknowns. The remaining p can be had from the relation $\sum_{i=1}^{M} p_i = 1$.

7. Summary

This paper deals with the problem of estimation of parameters from a heterogeneous population which is constituted of two or more sub-populations each following Poisson distribution. Maximum likelihood equations are derived for obtaining the estimates. Information matrix is given for evaluating the standard errors of the estimates. To illustrate the practical application of the results a numerical example has also been discussed.

8. Acknowledgment

The author is extremely grateful to Dr. R. S. Varma, Director, Defence Science Laboratory, for the permission to publish the paper.

His thanks are also due to Dr. P. V. K. Iyer, for his able guidance in the preparation of this paper and Messrs. D. Ray and P. P. Saigal or useful discussions.

9. REFERENCES

1. Mendenhall, W. and Hader, R. J. — "Estimation of parameters of mixed exponentially distributed failure time distribution from censored life-test data," *Biometrika*, 1958, **45**, 504–20.

2. Cohen, A. C., Jr. .. "Estimation of the Poisson parameters from truncated and censored samples," *J. Amer. Stat. Asso.*, 1954, **49**, 158–68.

3. Kitagawa, Tosio .. *Tables of Poisson Distribution*, 1952.