Mixture of K Normal Distributions by Dyar's Law

Sang-Un Yun*

1. Interoduction

The problem considered in this paper can be defined as follows. Consider observations $x_1, x_2, \ldots, x_n$ which are assumed to come from a mixed population of the density function,

$$f(x) = \sum_{k=1}^{m} p_k f_k(x)$$

where m is the number of subpoulations and p_k is the proportion of subpopulation k such that $\sum_{k=1}^{m} p_k = 1$, $0 < p_k < 1$, and where $f_k(x)$ is the normal density function with mean $a^{(k-1)}u$ and variance $a^{2(k-1)}\sigma^2$. The problem then is to estimate on the basis of the observations $x_1, x_2, \ldots, x_n$ the unknown parameters a, u, σ and p_k, $k = 1, 2 \ldots, m$.

The application of this formulation can be found in entomology where $x_1, x_2, \ldots, x_n$ represent the length of the larva, and m is the number of larval stages and $a^{k-1}u$ is the mean size of larva at the k^{th} stage with variance $a^{2(k-1)}\sigma^2$. Dyar's law has been confirmed for a wide variety of insect species (Forbes, 1953: Hoxie and Wellso, 1974). An Algorithm to estimate the parameters is suggested based on the following scheme. After the observations are clustered into m groups, the estimates of a and u are found by least square (LS) method, and the MLE of σ^2 is obtained as function of the estimates of a, u and p_k. An example from a real experiment is demonstrated in Section 4.

* KOREA INSTITUTE FOR DEFENSE ANALYSES, C.P.O. BOX 3089

2. Estimation of Parameter

To reduce the difficulty of the problem, the estimates of the parameters are derived for a simpler problem in which it is assumed that the observations of each group after clustering are the sample from single population. As one can easily recognize from the characteristics of the problem, the estimates of a and u are compromising well with the assumption while there exists the obvious bias in the estimate of σ^2 such that $\hat{\sigma}^2<\sigma^2$.

It will be assumed that m cut-off points for clustering are given, $T_1, T_2, \ldots, T_{m-1}$ such that nearly all of the sample of the k^{th} subpopulation lie between T_{k-1} and T_k where $T_0=-\infty$, $T_m=\infty$. Thus we write the observations:

$$X'=(x_1,x_2,\ldots,x_n)$$
$$=(x_{11},x_{12},\ldots,x_{1n_1},\ x_{21},\ldots,x_{2n_2},\ x_{31},\ldots,x_{mn_m})$$

where

$$\sum_{k=1}^{m} n_k=n \quad ,$$

and $(x_{k1},x_{k2},\ldots,x_{kn_k})$ is the observation of k^{th} group. It will hereafter be assumed that k always ranges from $k=1$ to m and that i always ranges from $i=1$ to n_k when not specified.

Suppose that the observations of k^{th} group is the sample from the k^{th} subpopulation only. Then the density function of x_{ki} is $N(a^{k-1}u,\ a^{2(k-1)}\sigma^2)$. By taking the log-transformation of the random variable, it can be easily shown that $\ln x_{ki}-(k-1)\ln a$ has the same distribution for all k with mean $\ln u$. Accordingly, we have the linear regression function.

$$ln\ x_{ki}=lnu+(k-1)lna+\epsilon_{ki}. \tag{1}$$

After the LS estimates of lnu and lna are obtained we have the estimates of u and a by taking exponential of them such that

$$\hat{a}=\prod_k\left(\prod_i x_{ki}\right)^{-(k-1)n-N_1'N},$$
$$\hat{u}=\prod_k\left(\prod_i x_{ki}\right)^{-(k-1)N_1+N_2'N}, \tag{2}$$

where

$$N_1=\sum_k (k-1)n_k,$$

$$N_2=\sum_k (k-1)^2 n_k,$$

$$N=nN_2-N_1^2$$

The MLE of σ^2 is obtained from the derivative of log likelihood function;

$$\hat{\sigma}^2=\frac{1}{n}\sum_k\sum_i (x_{ki}-\hat{a}^{(k-1)}\hat{u})^2/\hat{a}^{2(k-1)}. \tag{3}$$

However, in practice, the cut-off points T_k are not known, nor is n_k. Therefore the following iteration scheme is suggested to resolve this problem.

1. Determine n_k by looking at the data or from the prior information.
2. Estimate a and u using (2).
3. Estimate n_k by the midpoint of two population means, i.e., find n_k such that

   $$x_{n_k}<(a^{k-1}+a^k)u/2<x_{n_k+1}.$$

   Repeat steps 2 and 3 until they converge.
4. Estimate σ^2 from (3), and obtain the estimates of p_k as $\hat{p}_k=n_k/n$.

3. Properties of Estimates

To investigate the properties of the estimates in the previous section the expectations and variances of the estimates are derived.

Let T_k be the dividing point between the k^{th} and $(k+1)^{ti}$ group, i.e., from step 3 of the Algorithm.

$$T_k=(a^k+a^{k-1})u/2, \quad k=1, 2, \ldots, m-1. \tag{4}$$

Let $T_m=\infty$, and

$$\begin{aligned} z_1&=(T_k-a^{k-1}u)/(a^{k-1}\sigma)\\ &=u(a-1)/(2\sigma),\\ z_2&=-(T_k=a^k u)/(a^k\sigma)\\ &=u(a-1)/(2a\sigma), \end{aligned}$$

and write

$$\Phi_1=\Phi(z_1),\quad \phi_1=\phi(z_1),$$
$$\Phi_2=\Phi(z_2),\quad \phi_2=\phi(z_2).$$

Suppose that the populations are separated enough to assume that the sample from k^{th} subpopulation is only compounded with the $(k+1)^{th}$ and $(k-1)^{th}$ group, and not compounded with any other groups. One can easily see that this assumption is not unrealistic in practice. Accordingly, we assume that

$$\Phi\{(T_k-a^{(k+1)}u)/(a^{(k+1)}\sigma)\}=0,$$
$$\phi\{(T_k-a^{k+1}u)/a^{(k+1)}\sigma)\}=0,$$
$$\Phi\{(T_{k+1}-a^{k-1}u)/(a^{(k-1}\sigma)\}=1,$$
$$\phi\{(T_{k+1}-a^{k-1}u)/(a^{k-1}\sigma)\}=0,\quad k=1,2,\ldots,\quad m-1.$$

Thus, the moments of the random variables of each group become

$$E(x_{1.})=\int_{-\infty}^{T_1}\left[\frac{p_1x}{\sqrt{2\pi}\sigma}\exp\{-(x-u)^2/(2\sigma^1)\}+\frac{p_2x}{\sqrt{2\pi a\sigma}}\exp\{-(x-au)^2/(2a^2\sigma^2)\}\right]dx/CP_1$$
$$=[\{p_1\Phi_1+p_2a(1-\Phi_2)\}u-(p_1\phi_1+p_2a\phi_2)\sigma]/CP_1$$
$$E(x_{k.})=a^{k-2}[\{(-p_{k-1}+p_ka)\Phi_1+a(p_k-p_{k+1}a)(\Phi_2-1)+p_k^-{}_1\}u+\{(p_k^-{}_1-p_ka)\phi_1+(p_k-p_{k+1}a)a\phi_2\}\sigma]/CP_k,\qquad(6)$$
$$k=2,3,\ldots,m-1,$$
$$E(x_{m.})=a^{m-2}[\{p_{m-1}(1-\Phi_1)+p_ma\Phi_2\}u+(p_{m-1}\phi_1+p_ma\phi_2)\sigma]/CP_m$$

where

$$CP_1=p_1\Phi_1+p_2(1-\Phi_2),$$
$$CP_k=(-p_{k-1}+p_k)\Phi_1+(p_k-p_{k+1})(\Phi_2-1)+p_{k-1},\qquad(7)$$
$$k=2,3,\ldots,m-1,$$
$$CP_m=p_{m-1}(1-\Phi_1)+p_m\Phi_2.$$

Similarly the second moments of the random variables are

$$E(x_1.^2)=[p_1(-C_1\phi_1+C_3\Phi_1)+p_2\{-C_2\phi_2+a^2C_3(1-\Phi_2)\}]/CP_1,$$
$$E(x_k.^2)=a^{2(k-2)}[(p_{k-1}-a^2p_k)(C_1\phi_1-C_3\Phi_1)+(p_k-a^2p_{k+1})\{C_2\phi_2-a^2C_3(1-\Phi_2)\}+p_{k-1}C_3]/CP_k,$$

$$k=2, 3, \ldots, m-1, \tag{8}$$

$$E(x_{m\cdot}^{2}) = a^{2(m-2)}[p_{m-1}\{C_1\phi_1+C_3(1-\Phi_1)\}+p_m(C_2\phi_2+a^2C_3\Phi_2)]/CP_m,$$

where

$$C_1=\sigma u(a+3)/2,$$
$$C_2=a\sigma u(3a+1)/2,$$
$$C_3=\sigma^2+u^2.$$

The expected values and uariances of estimates are obtained utilizing the Taylor's Theorem of Kendall and Stuart(1977, page 246). From $\hat{a}$ and $\hat{u}$ of (2),

$$E(\hat{a}) = g_a\left[1+\sum_{k=1}^{m}\frac{n_k\{(k-1)n-N_1\}}{2N\{E(x_{k\cdot})\}^2}\left\{\frac{(k-1)n-N_1}{N}-1\right\}V(x_{k\cdot})\right]+0(n^{-1}),$$

$$v(\hat{a}) = g_a{}^2\sum_{k=1}^{m}n_k\left\{\frac{(k-1)n-N_1}{NE(x_{k\cdot})}\right\}^2\mathrm{Var}(x_{k\cdot})+0(n^{-1}),$$

$$E(\hat{u}) = g_u\left[1+\sum_{k=1}^{m}\frac{n_k\{-(k-1)N_1+N_2\}}{2N\{E(x_{k\cdot})\}^2}\left\{\frac{-(k-1)N_1+N_2}{N}-1\right\}V(x_{k\cdot})\right]$$
$$+0(n^{-1}),$$

$$V(\hat{u}) = g_u{}^2\sum_{k=1}^{m}n_k\left\{\frac{-(k-1)N_1+N_2}{NE(x_{k\cdot})}\right\}^2V(x_{k\cdot})+0(n^{-1}),$$

and from $\hat{\sigma}$ of (3),

$$E(\hat{\sigma}^2) = \frac{1}{n}\sum_{k=1}^{m}n_k\left[E(x_{k\cdot})-\{E(\hat{a})\}^{-1}E(u)\right]^2/\{E(\hat{a})\}^{2(k-1)}+0(n^{-1}),$$

$$V(\hat{\sigma}^2) = \frac{2}{n}\sum_{k=1}^{m}g'_{x_k}V(x_{k\cdot}+g'_aV(\hat{a})+g_u{}'V(\hat{u})+2g_a{}'g_u{}'Cov(\hat{a},\hat{u})+0(n^{-1})$$

where the moments, $E(x_{k\cdot})$ and $E(x^2{}_{k\cdot})$, are from (6) and (8), and

$$V(x_{k\cdot}) = E(\mathrm{x}^2{}_{k\cdot})-\{E(x_{k\cdot})\}^2,$$

and where

$$g_a = \prod_{k=1}^{m}\{E(x_{k\cdot})\}n_k{}^{-(k-1)n-N_1{}'N}$$

$$g_u = \prod_{k=1}^{m}\{E(x_{k\cdot})\}n_k{}^{-(k-1)N_1+N_2{}'N}$$

$$g'_{x_k}=n_kE(x_{k\cdot})-\{E(\hat{a})\}^{k-1}E(\hat{u})^2/\{E(a)\}^{4(k-1)}$$

$$g'_a=\frac{4(k-1)^2}{n^2}\left[\sum_{k=1}^{m}n_kE(x_{k\cdot})\left\{E(x_{k\cdot})-\{E(\hat{a})\}^{k-1}E(\hat{u})\right\}/\{E(\hat{a})\}^{2k-1}\right]^2$$

$$g'_u=\frac{4}{n^2}\left[\sum_{k=1}^{m}n_k\left\{E(x_{k\cdot})-\{E(\hat{a})\}^{k-1}E(\hat{u})\right\}/\{E(a)\}^{k-1}\right]^2$$

$$\mathrm{Cou}(\hat{a},\hat{u})=\sum_{k=1}^{m}\frac{n_kM_k}{\{E(x_{k\cdot})\}^2}(M_k-1)\prod_{k'=1}^{m}\{n_{k'}E(k')\}^{M_{k'}},$$

$$M_k=\{-KN_1+(k-1)n+N_2\}/N$$

4. Example

The method of estimation, described above, is utilized in the following example.

In the Coleoptera (Pissodes nemorensis included) all growth occurs during the larval stage. Due to the fact that certain parts of the body are enclosed by an inflexible exoskeleton as the larva grows it must molt to accomodate its increased size. A given linear measurement of a portion of the rigid exoskeleton (such as width of head capsule) increases in a series of discrete steps. Females of P. nemorensis lay eggs under the bark of susceptible pine trees and the entire larval development occurs there. There is no way to follow the progress of an individual throughout its development since exposure results in death. Table 1 show $n=382$ measurements of individuals of different ages by Atkinson and Foltz of Entomology and Nematology department, University of Florida.

As one can easily see from the Table 1, such is the case when $m=5$, i.e., there are five subpopulations. Following the algorithm given in Section 2, we determine the starting values of n_k as

$n_1=$ 98 up to width 12

$n_2=$ 72 up to width 17

$n_3=$ 46 up to width 22

$n_4 = 30$ up to width 30

$n_5 = 116$ up to width 45.

Table 1. Frequency of head capsule width of *Pissodes nemorensis* larvae reared at 25°C in slash pine bolts, Oct. 1978 to Jan. 1979.

Width Micrometer Units	Frequency	Width	Frequency
8	7	27	10
9	48	28	2
10	35	29	5
11	7	30	1
12	1	31	2
13	14	32	1
14	27	33	4
15	18	34	9
16	7	35	13
17	6	36	21
18	9	37	17
19	17	38	20
20	13	39	14
21	5	40	5
22	2	41	5
23	6	42	2
24	8	43	1
25	8	44	1
26	10	45	1
		TOTAL	382

Table 2. Expected values and variances of the estimates of a,u,σ^2

Parameter	Estimate	Expected Value	Variance
a	1.402	1.401	0.00001
u	9.588	9.598	0.00221
σ^2	0.413	0.422	0.03912

The finale stimates converged after several iterations and the expected values and variances of the estimates of a, u and σ^2 are shown in Table 2. The estimates of n_k' $k=1, 2, \ldots, 5$ are 97, 67, 57, 50, 116 which result in the estimates of p_k such that

$\hat{p}_1=0.254,\ \hat{p}_2=0.175,\ \hat{p}_3=0.136,\ \hat{p}_4=0.131,\ \hat{p}_5=0.304.$

BIBLIOGRAPHY

Forbes, W.T.M. (1953) "Note on Multimodel Curves," *Annals Entological Society of America*, Vol. 46, 221-225.

Hoxie, R.P., and Wellso, S.G. (1974), "Cereal Leaf Beetle Instars and Sex, Defined by Larval Head Capsule Widths," *Annals of Entomological Society of America*, Vol. 67, No. 2, 183-187.

Kendall, M.G. and Stuart, A. (1977), *The Advanced Theory of Statistics*, Vol. I, London: Charles Griffin & Co., 247.