欠測値を含む区間データに基づくパラメータの最尤推定について

三浦幹彦*・坂口明男**・成瀬孔基*

1. はじめに

確率変数Xが密度関数 $f(x;\theta)$ に従うものとし，データから未知パラメータ θ を推定する問題で，X の実現値 $x_i(i=1, \cdots\cdots, N)$ そのものは与えられず，x_i が含まれる区間 $(y_i, z_i]$ だけがデータとして得られる場合を考える。たとえば，度数分布表にまとめられたデータから，未知パラメータを推定する場合があげられる。この時，通常は尤度関数

$$L(\theta)=\prod_{i=1}^{N}\{F(z_i)-F(y_i)\} \qquad \text{但し，} F(\cdot)\text{は}f(\cdot)\text{の分布関数}$$

を最大にするような θ を求めて最尤推定値としている（Heitjam (1989)）。しかし，この方法は，区間の一部が重複しているデータに対しては一致推定量を与えてくれない場合がある。鍋谷（1983）は仮谷（1980）のデータを引用し，それを指摘すると同時に，xから区間 $(y, z]$ が決定される確率機構を考慮することにより，そのデータに新たな解析を加えた。それは，確率機構を考慮した対数尤度関数から直接的にパラメータを推定する方法であり，複雑な計算を必要とする。そこで，本論文では，さらに一般的に区間データの一部が欠測している場合も含めて考察を加え，区間 $(y, z]$ が与えられたという条件の下での x の条件付分布を利用した新しい推定法を提案し，その妥当性について検討する。

2. 欠測値を含む区間データからのパラメータの推定

例として，間隔分布 $f(x;\theta)$ に従って生起する独立な事象系列を考える。この時，事象の発生間隔 x は観測できず，図1のように一定時間 t で分割された任意の区間に事象が生起するかしないかだけの情報しか得られないものとする。いま，図のように t_i の区間で事象が生起し，途中の区間で生起せずに，次に区間 t_j $(j>i)$ で生起する場合を考える。ただし，区間 t 内に事象が一つだけ生起するとは限らないものとする。この時，得られる情報は

$$(j-i-1)\ t<x_i\leqq(j-i+1)t$$

* 信州大学繊維学部繊維システム管理学研究室
** 信州大学繊維学部繊維教育実験実習施設

×：事象の生起
x：発生間隔
t：区間幅

図1　事象の生起と発生間隔

であり，データは区間データ $((j-i-1)t, (j-i+1)t]$ として与えられる。ここで，次のような場合を考える。注目した時間に，$N+1$ 個（Nは未知）の事象が生起し，N個の発生間隔 $x_1, x_2, \cdots\cdots, x_N$が出現しているが，このうちのいくつかは区間 t より小さいためある確率で区間データとして観測されず欠測し（この数をmとする），残りの $n=N-m$ 個だけが観測できるものとする。このように，m 個の欠測データがある場合，観測された区間データから分布 $f(x;\theta)$ の未知パラメータ θ を推定する方法について考察を加える。

データの特性を明確にするために表1の度数分布表の形に整理した。表の区間数 (r)

表1　区間データ

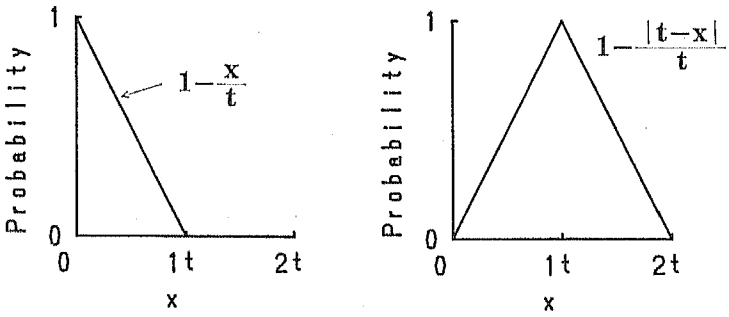
区間数	区間	度数	個々の発生間隔（観測不可能）
0	$(0, t]$	m	$x_{01}, x_{02}, \cdots\cdots, x_{0m}$
1	$(0, 2t]$	n_1	$x_{11}, x_{12}, \cdots\cdots, x_{1n_1}$
2	$(t, 3t]$	n_2	$\cdots\cdots$
⋮	⋮	⋮	⋮
r	$((r-1)t, (r+1)t]$	n_r	$x_{r1}, x_{22}, \cdots\cdots, x_{rn_r}$
⋮	⋮	⋮	⋮
k	$((k-1)t, (k+1)t]$	n_k	$x_{k1}, x_{k2}, \cdots\cdots, x_{kn_k}$

$$\sum_{i=1}^{k} n_i = n$$

は事象が生起した後，次に初めて生起するまでの区間数を表す。また，これは区間数0の場合を除いて $t=1$ とおいた時の区間データの中点を表している。

$0<x<t$ の場合は，二つ以上の事象が一つの区間内に生起して区間データとして観測されず欠測するか，二つの区間に分かれて生起し区間データ $(0, 2t]$ が得られるかのいずれかである。欠測する確率は x が0に近い時には大きく，x が t に近づくにつれ小さくなると考えるのが妥当である。そこで，欠測する確率は図2(a)のように x の関数 $1-x/t$ で与えられることを仮定する。また，同様に区間データ $(0, 2t]$ が得られる確率はその中点 t で最も大きく，0あるいは $2t$ に近づくにつれ減少する $1-|t-x|/t$ なる関数（図2(b)）を仮定する。さらに，他の区間データについても同様な関数を仮定する。

ここで，区間幅 t を1としても一般性を失わないので，記号を簡単にするため，以下の議論においては $t=1$ とする。また，分割された区間幅が一定のため表1に示したよ



(a) 欠測する確率　　(b) $(0,\ 2t]$ となる確率

図2　x と区間データとの関係

うに区間データは，それぞれその区間数 r（区間の中点）で表すことができる。ただし，$r=0$ は欠測を表すものとする。この時，これまでの仮定に基づき x から区間データ $(r-1,\ r+1]$ が生ずる確率機構は x が与えられた時の r の条件付密度

$$P(r|x)=\begin{cases}\{1-|r-x|\}\ I_{(r-1,\,r+1)}\ (x), & r=1,2,\cdots\cdots\\ (1-x)\ I_{(0,1)}\ (x), & r=0\end{cases} \qquad (1)$$

ただし $I_{(a,b)}\ (x)$ は定義関数
によって表すことができる。

従って，間隔分布 $f(x;\theta)$ と(1)式から，区間数 r の分布は

$$P(r)=\int_0^\infty f(x;\theta)P(r|x)dx$$

となり，これから欠測の起こる確率は，

$$P(0)=\int_0^1 (1-x)\ f(x;\ \theta)dx \qquad (2)$$

区間 $(r-1,\ r+1]$ の得られる確率は

$$P(r)=\int_{r-1}^{r+1} \{1-|r-x|\}\ f(x;\theta)dx \qquad (3)$$
$$r=1,\ 2,\cdots\cdots\cdots$$

で与えられる。

その結果，実際に観測される r の分布は

$$P_T(r)=\frac{P(r)}{1-P(0)},\ r=1,\ 2,\cdots\cdots$$

となり，観測データ $r_1, r_2, , \cdots\cdots r_n$ による対数尤度関数は

$$\log L(\theta) = -n \log\{1-P(0)\} + \sum_{i=1}^{n} \log P(r_i)$$

で与えられる。ここで，表1のような度数データが得られたとすれば，上式は

$$\log L(\theta) = -n \log\{1-P(0)\} + \sum_{r=1}^{k} n_r \log P(r) \qquad (4)$$

となり，これを最大にするように θ を決めれば最尤推定値を求めることができる。ただし，k は観測された最大の区間数を表す。

3. 指数分布，対数正規分布およびガンマ分布におけるパラメータ推定

(4)式に基づくパラメータ推定の例として，間隔分布 $f(x;\theta)$ が指数分布，対数正規分布およびガンマ分布に従う場合について考察する。

（i） 指数分布の場合

$$f(x;\theta) = \theta e^{-\theta x}, \quad x>0$$

欠測の起こる確率は(2)式から

$$P(0) = 1 - \frac{1}{\theta}(1-e^{-\theta})$$

また，区間数が $r(\neq 0)$ となる確率は(3)式から

$$P(r) = \frac{1}{\theta}(1-e^{-\theta})^2 e^{-\theta(r-1)}, \quad r=1, 2, \cdots\cdots$$

となる。従って対数尤度関数は(4)式から

$$\log L(\theta) = n \log(1-e^{-\theta}) - \theta \sum_{r=1}^{k} n_r (r-1)$$

で与えられる。この場合，最尤推定量は簡単に求まり

$$\hat{\theta} = \log \frac{\bar{r}}{\bar{r}-1}$$

となる。ただし，

$$\bar{r} = \frac{\sum_{r=1}^{k} n_r r}{n}$$

である。また，この場合は，区間内の事象の生起数がポアソン分布に従うことを利用すれば，区間内に事象が一つも生起しない確率は $e^{-\theta}$，一つ以上生起する確率は $1-e^{-\theta}$ であり，区間数 r が幾何分布

$$P_T(r)=(1-e^{-\theta})\ e^{-\theta(r-1)},\quad r=1,\ 2,\cdots\cdots$$

に従うことからも上記の結果を得ることができる。

(ii) 対数正規分布の場合

$$f(x;\theta)=\frac{1}{\sqrt{2\pi}\sigma x}e^{-\frac{(\log x-\mu)^2}{2\sigma^2}},\ x>0$$

ただし，$\theta^t=(\mu,\ \sigma)$

欠測する確率および区間数が r となる確率は，それぞれ(2)，(3)，式から

$$P(0)=\Phi(c_{1,0})-\varphi\Phi(c_{1,1}) \tag{5}$$

$$\begin{aligned}P(r)&=(1-r)\{\Phi(c_{r,0})-\Phi(c_{r-1,0})\}+\varphi\{\Phi(c_{r,1})-\Phi(c_{r-1,1})\}\\&\quad+(1+r)\{\Phi(c_{r+1,0})-\Phi(c_{r,0})\}-\varphi\{\Phi(c_{r+1,1})-\Phi(c_{r,1})\}\\&=(1-r)H_{r,r-1,0}+(1+r)H_{r+1,r,0}+\varphi(H_{r,r-1,1}-H_{r+1,r,1})\end{aligned} \tag{6}$$

となる。ただし，$\Phi(\cdot)$ は標準正規分布の分布関数を表し

$$\varphi=e^{\mu+\frac{1}{2}\sigma^2}$$

$$c_{ij}=\frac{\log\ i-\mu}{\sigma}-j\sigma \tag{7}$$

$$H_{i,j,k}=\Phi(c_{ik})-\Phi(c_{jk}) \tag{8}$$

を表す。これを(4)式に代入すれば，対数尤度関数が得られる。

鍋谷は表2に示した仮谷のデータに対して対数正規分布を仮定しパラメータの最尤推定を行っている。この例では，欠測データがなく区間〔0，6〕の度数は $m=57$ であることが知られている。また，実際には区間幅が $t=6$ であるが，最尤推定量の性質から $t=1$ として推定を行い，得られた推定量 $\tilde{\mu}$，$\tilde{\sigma}$ を

$$\hat{\mu}=\log(t)+\tilde{\mu},\ \hat{\sigma}=\tilde{\sigma}$$

と変換すればよい。この場合の対数尤度関数は(4)式を

$$\log\ L(\theta)=m\ \log\ P(0)+\sum_{r=1}^{k}n_r\ \log\ P(r)$$

表2 仮谷のデータへの分布のあてはめ

区間（単位月）	齲蝕発病歯数	対数正規分布理論度数	ガンマ分布理論度数
0～6	57	44.66	52.66
0～12	127	170.46	159.96
6～18	205	194.72	184.79
12～24	119	107.56	117.72
18～30	73	58.84	58.03
24～36	22	24.02	24.97
30～42	5	11.73	9.88
計	608		

で置き換えて考えればよいことになる。ここで

$$\frac{\partial \log L(\theta)}{\partial \theta}=0$$

を解くことになるが，それは大変な計算を必要とし，鍋谷は通常の Newton-Raphson 法では計算が成功しなかったため，疑似 Newton-Raphson 法の一種を適用して，初期値$(\hat{\mu}_0,\ \hat{\sigma}_0)=(3.2,\ \sqrt{0.8})$ の下に，推定値 $(\hat{\mu},\ \hat{\sigma})=(2.4100, 0.6130)$ を得ている。

(iii) ガンマ分布の場合

$$f(x;\lambda,\ \gamma)=\frac{\lambda^{\gamma}}{\Gamma(\gamma)}x^{\gamma-1}e^{-\lambda x},\ x>0$$

但し，$\Gamma(\gamma)=\int_0^{\infty}x^{\gamma-1}\ e^{-x}dx$

欠測する確率および区間数が r となる確率は，それぞれ(2)，(3)式から

$$P(0)=J(\lambda,\ \gamma)-\gamma/\lambda\cdot J\ (\lambda,\gamma+1) \tag{9}$$

ただし，

$$J(x,\ p)=\frac{1}{\Gamma(p)}\int_0^{x}t^{p-1}\ e^{-t}dt,\ p,x>0$$

$$P(r)=(1-r)\ G(\lambda r,\ \lambda\ (r-1),\ \gamma)+\frac{\gamma}{\lambda}G(\lambda r,\ \lambda(r-1),\ \gamma+1)$$

$$+(1-r)G(\lambda(r+1),\ \lambda r,\ \gamma)-\frac{\gamma}{\lambda}G(\lambda(r+1),\ \lambda r,\ \gamma+1)$$

また，

$$G(x,\ y,\ z)=J(x,\ z)-J(y,\ z)$$

これらを(4)式に代入すれば対数尤度関数が得られるが，それから最尤推定量を求めるには，対数正規分布の場合と同様に大変な計算を必要としかなりの困難が伴う。

4. 事象生起の間隔 x の条件付分布を利用した最尤推定

3節に示したように，表1のようなデータを扱う場合，間隔分布 $f(x;\theta)$ が指数分布のような特別な場合を除いて，(4)式から求まる対数尤度関数を直接用いて θ の最尤推定量を求めるのは困難な場合が多い。そこで，本論文では次のような別の推定法を提案する。

はじめに，表1の右端の欄に示したように，実際には観測できない $n+m$ 個の間隔データをそれぞれ $x_{01}, x_{02}, \cdots\cdots, x_{0m}, x_{11}, \cdots\cdots, x_{kn_k}$ とする。ただし，$n=n_1+n_2+\cdots\cdots+n_k$ である。x から区間データ $(r-1, r+1]$ が生ずる確率機構として前と同様に x が与えられた時の r の条件付分布 ((1)式) を仮定すれば，逆に区間データが与えられた時の x の条件付分布 $P(x|r)$ は

$$P(x|0)=\frac{(1-x)f(x;\theta)\,I_{(0,1)}(x)}{P(0)}, \tag{10}$$

$$P(x|r)=\frac{\{1-|r-x|\}f(x;\theta)\,I_{(r-1,r+1)}(x)}{P(r)}$$

$$r=1, 2, \cdots\cdots$$

となる。

従って，区間の度数 m, $n_i(i=1, 2, \cdots\cdots, k)$ が与えられた時の $n+m$ 個の間隔データ x の条件付同時分布は，(10)式から

$$f(x_{01}, x_{02}, \cdots, x_{0m}, x_{11}, \cdots, x_{kn_k}|m, n_1, n_2, \cdots, n_k)$$
$$=\prod_{i=1}^{m}\frac{f(x_{0i})f(0|x_{0i})}{P(0)}\cdot\prod_{j=1}^{n_1}\frac{f(x_{1j})f(1|x_{1j})}{P(1)}\cdot\cdots\cdots\cdot\prod_{s=1}^{n_k}\frac{f(x_{ks})f(k|x_{ks})}{P(k)} \tag{11}$$

で与えられる。ここで，$n_1, \cdots\cdots, n_k$ の同時分布は多項分布

$$P(n_1, n_2, \cdots, n_k;\theta)=\frac{n!}{n_1!n_2!\cdots n_k!}\left\{\frac{P(1)}{1-P(0)}\right\}^{n_1}\cdots\left\{\frac{P(k)}{1-P(0)}\right\}^{n_k} \tag{12}$$

に従うので，m個のデータが欠測したという条件の下での $n+m$ 個の間隔データの同時分布は，(11)，(12)式から

$$f(x_{01}, x_{02}, \cdots, x_{kn_k}|m)=\frac{n!}{n_1!\ n_2!\cdots n_k!}\{P(0)\}^{-m}\ \{1-P(0)\}^{-n}$$
$$\times\prod_{i=1}^{m}f(x_{0i})f(0|x_{0i})\prod_{j=1}^{n_1}f(x_{1j})f(1|x_{1j})\cdots\cdots\prod_{s=1}^{n_k}f(x_{ks})f(k|x_{ks}) \tag{13}$$

のようになる。

さらに，欠測データの数mの分布として負の二項分布

$$P(m;\theta)=\binom{n+m-1}{m}\{P(0)\}^m\ \{1-P(0)\}^n \tag{14}$$

を仮定すれば，m, x_{01}, x_{02}, $\cdots$, x_{0m}, x_{11}, $\cdots$, x_{kn_k} に基づく対数尤度関数は(13), (14)式から

$$\log L(\theta)=\log\binom{n+m-1}{m}+\sum_{i=1}^{m}\{\log f(x_{0i})+\log(1-x_{0i})\}$$
$$+\sum_{i=1}^{k}\sum_{j=1}^{n_i}[\log f(x_{ij})+\log\{1-|i-x_{ij}|\}] \tag{15}$$

と表すことができる。

これは，観測データ n_1, n_2, $\cdots$, n_k とパラメータの初期値 θ_0 を与えた時の(15)式で表される対数尤度の期待値 ($m, x_{01}, \cdots, x_{0m}, \cdots, x_{kn_k}$ に関する)

$$E[\log L(\theta)|\theta_0,\ n_1,\ n_2,\cdots,\ n_k] \tag{16}$$

を最大にするステップをくり返す EM アルゴリズムにより解くことができる。3節で示した例について，本論文で提案する方法を述べる。

(i) 指数分布の場合

(15), (16)式から

$$E[\log L(\theta)|\theta_0,\ n_1,\ n_2,\cdots,\ n_k]$$
$$=\text{const}+\{E[m|\theta_0]+n\}\ \log\theta-\theta\{E[m|\theta_0]\ E[x_{0i}|\theta_0]$$
$$+n_1E[x_{1i}|\theta_0]+\cdots+n_kE[x_{ki}|\theta_0]\}$$

となる。ただし，

$$E[m|\theta_0]=C_1(\theta_0)=\sum_{m=0}^{\infty}mP(m|\theta_0)$$
$$=n\frac{\theta_0}{1-e^{-\theta_0}}\{1-\frac{1}{\theta_0}(1-e^{-\theta_0})\}$$

$$E[x_{0i}|\theta_0]=C_2(\theta_0)=\int_0^1\frac{1}{P(0)}x(1-x)\theta_0\ e^{-\theta_0x}dx$$
$$=\frac{\frac{1}{\theta_0}+\frac{1}{\theta_0}e^{-\theta_0}-\frac{2}{\theta_0^2}(1-e^{-\theta_0})}{1-\frac{1}{\theta_0}(1-e^{-\theta_0})}$$

$$E[x_{ri}|\theta_0]=C_3(r,\ \theta_0)$$

$$=\int_{r-1}^{r+1}\frac{1}{P(r)}\{1-|r-x|\}\ x\theta_0\ e^{-\theta_0 x}dx$$
$$=r+\frac{2}{\theta_0}-\frac{1+e^{-\theta_0}}{1-e^{-\theta_0}}$$

である。

従って

$$\frac{\partial E[\log L(\theta)|\theta_0,\ n_1,\cdots,\ n_k]}{\partial\theta}=0$$

より，最尤推定値は

$$\theta^{(i)}=\frac{C_1(\theta^{(i-1)})+n}{C_1(\theta^{(i-1)})C_2(\theta^{(i-1)})+n_1C_3(1,\ \theta^{(i-1)})+\cdots+n_kC_3(k,\ \theta^{(i-1)})}$$

をくり返すことにより求めることができる。この場合は，3節で示した推定法の方が簡単である。

(ii) 対数正規分布の場合

$$\begin{aligned}&E[\log L(\theta)|\theta_0,\ n_1,\ n_2,\cdots,\ n_k]\\&=\text{const}-\{E[m|\theta_0]+n\}\ \log\sigma\ -\frac{1}{2\sigma^2}\{E[m|\theta_0]\\&\times E[(\log x_{0i}-\mu)^2|\theta_0]+n_1E[(\log x_{1i}-\mu)^2|\theta_0]+\cdots\\&+n_kE[(\log x_{ki}-\mu)^2|\theta_0]\}\end{aligned}$$

であるから，$\beta_0(\theta_0)=E[m|\theta_0]$,

$$\beta_1(\theta_0)=E[\log x_{0i}|\theta_0],\ \beta_2(\theta_0)=E[(\log x_{0i})^2|\theta_0],$$
$$\alpha_1(r,\ \theta_0)=E[\log x_{ri}|\theta_0],\ \alpha_2(r,\ \theta_0)=E[(\log x_{ri})^2|\theta_0]$$

とおけば，上式はさらに

$$\begin{aligned}&-(\beta_0+n)\log\sigma-\frac{1}{2\sigma^2}[\{\beta_0\beta_2+\sum_{r=1}^{k}n_r\alpha_2(r,\ \theta_0)\}\\&-2\{\beta_0\beta_1+\sum_{r=1}^{k}n_r\alpha_1(r,\ \theta_0)\}\mu+(\beta_0+n)\mu^2]\end{aligned}$$

となり（付録A．1参照）最尤推定値は，

$$\mu^{(i)}=\frac{1}{\beta_0+n}\{\beta_0\beta_1+\sum_{r=1}^{k}n_r\alpha_1(r,\ \theta^{(i-1)})\}$$

$$\sigma^{2(i)}=\frac{1}{\beta_0+n}\{\beta_0\beta_2+\sum_{r=1}^{k}n_r\alpha_2(r,\theta^{(i-1)})\}-\mu^{(i)2}$$

をくり返すことによって求めることができる。3節で示した方法と比較すれば，この方法がいかに簡潔であるかは明らかである。また，m が既知の場合（鍋谷の例に対応する）は上式において β_0 をmに変えるだけでよい。

この方法に基づき Pascal プログラムを作成し，表2のデータを用いてパラメータの推定値を求めた。鍋谷と同じ初期値，同じ収束条件を用いた結果，13回の繰り返しで収束し，$(\hat{\mu},\ \hat{\sigma})=(2.4099,\ 0.6130)$ が得られた。計算には富士通 FMR70HX（数値演算プロセッサー付）を用い処理時間は約2秒であった。その結果得られた密度関数のグラフを図3に示した。このように，3節の方法では大型計算機を用いて困難な計算を行う必要があるのに対し，本論文で提案した方法は，パーソナルコンピュータを用いて簡単に推定値を求めることができる。

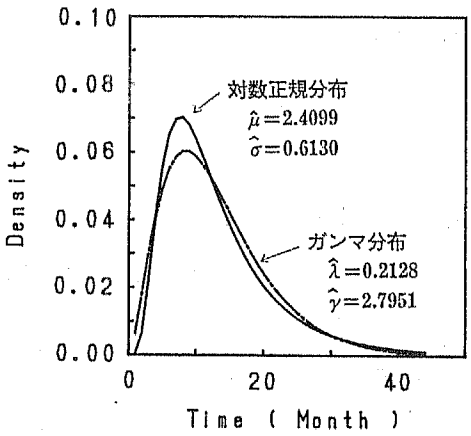


図3　仮谷のデータから得られた密度関数

また，得られた推定値を用いて理論度数を計算した結果を表2に示した。単純に

$$X^2=\sum\frac{(\text{実測度数}-\text{理論度数})^2}{\text{理論度数}}$$

を計算した。結果は27.96であった。

(iii)　ガンマ分布の場合

$$\begin{aligned}&E[\log(\theta)|\theta_0,\ n_1,\cdots,n_k]\\&=(E[m|\theta_0]+n)\ \{\gamma\log\lambda-\log\Gamma(\gamma)\}\\&+(r-1)\{E[m|\theta_0]\ E[\log x_{0i}|\theta_0]\\&+n_1E[\log x_{1i}|\theta_0]+\cdots+n_kE[\log x_{ki}|\theta_0]\}\\&-\lambda\{E[m|\theta_0]E[x_{0i}|\theta_0]\\&+n_1E[x_{1i}|\theta_0]+\cdots+n_kE[x_{ki}|\theta_0]\}\end{aligned}$$

ここで

$$\begin{aligned}&\beta_3(\theta_0)=E[m|\theta_0],\ \beta_4(\theta_0)=E[x_{0i}|\theta_0],\\&\beta_5(\theta_0)=E[\log x_{0i}|\theta_0],\ \alpha_3(r,\ \theta_0)=E[x_{ri}|\theta_0],\\&\alpha_4(r,\ \theta_0)=E[\log x_{ri}|\theta_0]\end{aligned}$$

（付録A．2参照）

とおけば，方程式

$$\lambda^{(i)}=\frac{\beta_3+n}{\{\beta_3\beta_4+\sum_{r=1}^{k}n_r\alpha_3\ (r,\ \theta^{(i-1)})\}}\ \gamma^{(i)}$$

$$(\beta_3+n)\{\log\lambda^{(i)}-\Psi(\gamma^{(i)})\}$$
$$+\{\beta_3\beta_5+\sum_{r=1}^{k}n_r\alpha_4(r,\ \theta^{(i-1)})\}=0$$

ただし，$\Psi(\cdot)$ はディガンマ関数を表す。

を繰り返しにより解くことで，最尤推定値を求めることができる。mが既知の場合は，β_3 をmで置き換えればよい。表2のデータのモデルとしてガンマ分布を仮定して，ここに示した方法でパラメータの推定を試みた。計算は $t=1$ として行い，得られた推定値 $\tilde{\gamma}$, $\tilde{\lambda}$ を $\hat{\gamma}=\tilde{\gamma}$, $\hat{\lambda}=\tilde{\lambda}/t$ により変換した。その結果 $(\hat{\lambda},\ \hat{\gamma})=(0.2128,\ 2.7951)$ が得られた。対数正規の場合と同様に処理時間は約2秒であった。推定値を代入した密度関数のグラフを図3に示した。また，対数正規分布の場合と同様に理論度数を計算し表2に示した。X^2 の値は16.00となり，表2のデータに対してはガンマ分布の方が適合性が良いことが考えられた。

文　　献

Heitjan, D.F.: Statistical Science, 4, 164-183(1989)
仮谷太一：応用統計学，9，73-81(1980)
鍋谷清治：応用統計学，12，59-67(1983)

付　　録

A.1　対数正規分布の場合

$$\beta_0(\theta_0)=n\cdot\frac{P(0)}{1-P(0)}$$

ただし，$P(0)$ は本文(5)式で表されるものである。

$$\beta_1(\theta_0)=\int_0^1\log x\cdot f(x|0)dx=\frac{1}{P(0)}(a_{10}-b_{10})$$

ただし，

$$a_{ij}=\mu_0H_{ij0}-\sigma_0h_{ij0},$$
$$b_{ij}=(\mu_0+\sigma_0^2)\varphi H_{ij1}-\sigma_0\varphi h_{ij1},$$
$$h_{ijk}=\phi(c_{ik})-\phi(c_{jk})$$

を表す。また $\phi(\cdot)$ は標準正規分布の密度関数であり，c_{ij}, H_{ijk} は本文中(7)，(8)式で定義されている。

$$\alpha_1(r,\ \theta_0)=\int_{r-1}^{r+1}\log x\cdot f(x|r)dx$$
$$=q_1(r)/P(r)$$

$$\alpha_2(r,\ \theta_0)=\int_{r-1}^{r+1}(\log x)^2\cdot f(x|r)dx$$
$$=q_2(r)/P(r)$$

ただし，$P(r)$ は本文(6)式を表し，

$$q_1(r)=(1-r)\ a_{r,r-1}+b_{r,r-1}+(1+r)\ a_{r+1,r}-b_{r+1,r}$$
$$q_2(r)=(1-r)\ A_{r,r-1}+B_{r,r-1}+(1+r)\ A_{r+1,r}-B_{r+1,r}$$

である。ここで，A_{ij}，B_{ij} はそれぞれ

$$A_{ij}=(\mu_0^2+\sigma_0^2)\ H_{ij0}-2\mu_0\sigma_0 h_{ij0}-\sigma_0^2\{c_{i0}\phi(c_{i0})-c_{j0}\phi(c_{j0})\}$$
$$B_{ij}=(\mu_0^2+2\mu_0\sigma_0^2+\sigma_0^2+\sigma_0^4)\varphi H_{ij1}-2\sigma_0(\mu_0+\sigma_0^2)\varphi h_{ij1}$$
$$-\sigma_0^2\varphi\{c_{i1}\phi(c_{i1})-c_{j1}\phi(c_{j1})\}$$

で表されるものである。

A.2 ガンマ分布の場合

$$\beta_3(\theta_0)=n\cdot\frac{P(0)}{1-P(0)}$$

ただし，$P(0)$ は本文(9)式で表されるものである。

$$\beta_4(\theta_0)=\frac{1}{P(0)}\left\{\frac{\gamma_0}{\lambda_0}J(\lambda_0,\ \gamma_0)-\frac{(\gamma_0+1)\gamma_0}{\lambda_0^2}J(\lambda_0,\gamma_0+2)\right\}$$

$$\beta_5(\theta_0)=\frac{1}{P(0)}\{F_1(1,0)-F_2(1,\ 0)\}$$

ただし，

$$F_1(x,\ y)=\{\Psi(\gamma_0)-\log\ \lambda_0\}\ G(\lambda_0 x,\ \lambda_0 y,\ \gamma_0)$$
$$+\frac{\partial}{\partial\gamma_0}J(\lambda_0 x,\ \gamma_0)-\frac{\partial}{\partial\gamma_0}J(\lambda_0 y,\ \gamma_0)$$

$$F_2(x,\ y)=\frac{\gamma_0}{\lambda_0}[\{\Psi(\gamma_0+1)-\log\lambda_0\}G(\lambda_0 x,\ \lambda_0 y,\ \gamma_0+1)$$
$$+\frac{\partial}{\partial\gamma_0}J(\lambda_0 x,\ \gamma_0+1)-\frac{\partial}{\partial\gamma_0}J(\lambda_0 y,\ \gamma_0+1)]$$

であり，$\Psi(\cdot)$ はディガンマ関数を表す。
また，

$$\alpha_3(r,\ \theta_0)=\frac{1}{P(r)}[(1-r)\frac{\gamma_0}{\lambda_0}G(\lambda_0 r,\ \lambda_0(r-1),\ \gamma_0+1)$$

$$+\frac{(\gamma_0+1)\gamma_0}{\lambda_0^2}G(\lambda_0 r,\ \lambda_0(r-1),\ \gamma_0+2)$$

$$+(1+r)\frac{\gamma_0}{\lambda_0}G(\lambda_0(r+1),\ \lambda_0 r,\ \gamma_0+1)$$

$$-\frac{(\gamma_0+1)\gamma_0}{\lambda_0^2}G(\lambda_0(r+1),\ \lambda_0 r,\ \gamma_0+2)]$$

$$\alpha_4(r,\ \theta_0)=\frac{1}{P(r)}\{(1-r)\ F_1(r,\ r-1)+F_2(r,\ r-1)$$

$$+(1+r)\ F_1(r+1,\ r)-F_2(r+1,\ r)\}$$

である。

Summary

Maximum likelihood estimation from interval data

Mikihiko Miura, Akio Sakaguchi and Yoshiki Naruse

Maximum likelihood estimation from overlapped interval data has computational difficulty in many population models. In order to diminish the difficulty we propose an estimation procedure based on the conditional density of unobserved data given the interval data. Three population models, which are exponential, log-normal and gamma distribution, are considered as examples.