系 统 科 学 与 数 学
J. Sys. Sci. & Math. Scis.
28(6) (2008, 6), 641–648

利用区间删失数据的分布函数估计及其收敛速度*

丁 邦 俊

(华东师范大学统计系, 上海 200062)

摘要 讨论任意 m 点均匀分布 ($m \geq 3$) 的情况，用 m 点均匀分布的累积分布函数去逼近连续总体的分布函数，在适当的条件下，证明了用区间删失数据估计分布函数具有收敛速度 $O(n)^{-\frac{2}{9}}$.

关键词 区间数据，随机逼近，收敛速度.

MR(2000) 主题分类号 62E17

1 引 言

在寿命问题研究中，由于客观原因，有时不能获得准确的寿命数据 X, 而只能知道 X 是否位于某一时间区间内，这种数据称为区间删失数据 (简称区间数据). 一般地，设第 i 次样本为 $(U_i, V_i, \delta_i, \gamma_i)$, $i = 1, 2, \cdots, n$, 其中 U_i, V_i 为观察到的数据 (且假设 $P(U_i \leq V_i) = 1$), δ_i, γ_i 为示性函数. 区间数据研究的主要问题是：如何从观察到的这些区间数据去估计 X 的分布函数 (或分布函数的泛函), 文 [1] 讨论了，在一定的条件下，所求的分布函数是一个泛函方程的解，但该解难以求出. 文 [2] 介绍了这方面的一些研究进展，其中比较重要的结果是：在适当的条件下，分布函数的估计量具有强相合性. 他们还对这个问题提出猜想：在适当的条件下，分布函数估计量的收敛速度可能是 $O\big(\frac{\ln\ln n}{n}\big)^{-\frac{1}{3}}$, 文 [3] 中也有这方面的介绍. 但是，这个问题至今未能得到令人满意的结果. 文 [4] 讨论了最简单的两点均匀分布的情况. 本文把文 [4] 的结果推广到任意 m 点均匀分布 ($m \geq 3$), 并用 m 点均匀分布去逼近连续总体的分布，在适当的条件下，给出了分布函数估计量的收敛速度.

2 关于 m 点分布的情况

2.1 一些假设和说明

本文仍采用文 [5] 中的假设 2.1 、假设 2.3 、假设 2.4 和假设 2.5, 但将其中的假设 2.2 改为如下的假设 2.2.

* 沪港合作资金资助项目.
收稿日期：2007-01-30.

假设 2.2 随机变量 X 在 $(0, M_1)$ 内等概率只取 m 个点： $x_1,\ x_2, \cdots, x_m$, 且它们都是未知的，即 $P(X = x_1) = P(X = x_2) = \cdots = \frac{1}{m}$, 但 $x_1, x_2, \cdots, x_m$ 都未知.

设 $x_1 < x_2 < \cdots < x_m$, 用 $\widehat{x}_1$ 表示 x_1 的估计，用 $x_1^{(k)}$ 表示对 x_1 的第 k 次估计，类似地理解 $\widehat{x}_i$ 和 $x_i^{(k)}$, 记 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$, $Z^{(k)} = \left(x_1^{(k)}, x_2^{(k)}, \cdots, x_m^{(k)}\right)^{\mathrm{T}}$.

2.2 用随机逼近法求未知向量的估计

当 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 已知时，对 m 个点的离散分布；受文 [4] 的启发，考虑如下的自相合方程组

$$\widehat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha^i(x_j)p_j}{\sum\limits_{k=1}^{m} \alpha^i(x_k)p_k}, \quad j = 1, 2, \cdots, m. \tag{2.1}$$

其中 α_j^i 或 $\alpha^i(X_j)$ 是一个示性函数，其定义为 $\alpha_j^i = \alpha^i(X_j) = I\{X_j \in (X_L^i, X_R^i]\}$, 在 (2.1) 式中，当 $p_1 = p_2 = \cdots = \frac{1}{m}$ 时，就有

$$\widehat{p}_j = \frac{1}{n}\sum_{i=1}^{n} \frac{\alpha^i(x_j)}{\sum\limits_{k=1}^{m} \alpha^i(x_k)}, \quad j = 1, 2, \cdots, m. \tag{2.2}$$

在自相合方程 (2.2) 中，当 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 已知时，可以估计出 $(\widehat{p}_1, \widehat{p}_2, \cdots, \widehat{p}_m)$, 反过来，当 $(p_1, p_2, \cdots, p_m)$ 已知时，从理论上讲也可以估计出 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$. 下面在 $p_1 = p_2 = \cdots = p_m = \frac{1}{m}$ 基础上, 利用随机逼近法, 从 (2.2) 式出发, 讨论 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 的估计. 估计 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 的步骤如下：

第 1 步 取 $Z = (x_1, x_2, \cdots, x_m)^{\mathrm{T}}$ 的初值估计 $Z^{(1)} = \left(x_1^{(1)}, x_2^{(1)}, \cdots, x_m^{(1)}\right)^{\mathrm{T}}$.

第 2 步 取实数列 $\{a_n\}, \{c_{jn}\}$ $(j = 1, 2, \cdots, m)$ 满足文 [4] 中 (2.7) 式和 (2.8) 式.

第 3 步 对 X 作两次独立观察, 第 1 次得到确定区间 $(X_L^1, X_R^1]$, 计算示性函数 $\alpha^1(x_j^{(1)} + c_{j1})$ 的值，并用下式估计 p_j $(j = 1, 2, \cdots, m)$

$$\widehat{p}_j(x_j^{(1)} + c_{j1}) \widehat{=} \widehat{p}_j\left(x_1^{(1)}, x_2^{(1)}, \cdots, x_{j-1}^{(1)}, x_j^{(1)} + c_{j1}, x_{j+1}^{(1)}, \cdots, x_m^{(1)}\right) = \frac{\alpha^1(x_j^{(1)} + c_{j1})}{\sum\limits_{r=1}^{m} \alpha^1(x_r^{(1)} + c_{j1})}.$$

第 2 次得到确定区间 $(X_L^2, X_R^2]$, 计算相应的 $\alpha^2\left(x_j^{(1)} - c_{j1}\right)$ 的值，并用下式估计 p_j

$$\widehat{p}_j\left(x_j^{(1)} - c_{j1}\right) \widehat{=} \widehat{p}_j\left(x_1^{(1)}, x_2^{(1)}, \cdots, x_{j-1}^{(1)}, x_j^{(1)} - c_{j1}, x_{j+1}^{(1)}, \cdots, x_m^{(1)}\right) = \frac{\alpha^1\left(x_j^{(1)} - c_{j1}\right)}{\sum\limits_{r=1}^{m} \alpha^1\left(x_r^{(1)} - c_{j1}\right)}.$$

第 4 步 取

$$x_j^{(2)} = x_j^{(1)} + \frac{a_1}{c_{j1}}\left[\widehat{p}\left(x_j^{(1)} + c_{j1}\right) - \widehat{p}\left(x_j^{(1)} - c_{j1}\right)\right],$$

记 $Z^{(2)} = \left(x_1^{(2)}, x_2^{(2)}, \cdots, x_m^{(2)}\right)^{\mathrm{T}}$.

第 5 步 重复以上第 3 步至第 4 步，对 X 作 $n = 2k$ 次独立观察后，由递推关系

$$x_j^{(k+1)} = x_j^{(k)} + \frac{a_k}{c_{jk}}\left[\widehat{p}\left(x_j^{(k)} + c_{jk}\right) - \widehat{p}\left(x_j^{(k)} - c_{jk}\right)\right], \tag{2.3}$$

可以得到 $x_j^{(k+1)}$ $(j=1,2,\cdots,m)$, 记

$$Z^{(k+1)}=\big(x_1^{(k+1)},x_2^{(k+1)},\cdots,x_m^{(k+1)}\big)^{\mathrm{T}}. \tag{2.4}$$

根据以上步骤, 可以得向量序列 $Z^{(1)},Z^{(2)},\cdots,Z^{(k)},\cdots$, 下面将证明, 在适当的条件下, 它们以概率 1 收敛到 Z (见定理 2.1).

2.3　向量序列 $\{Z^{(k)}\}$ 收敛性的证明

引理 2.1　设 $F_r=\sigma[u_i,v_i,\delta_i,\gamma_i;i=1,2,\cdots,r]$ $(r=1,2,\cdots)$, $y_j=x_j^{(k)}\pm c_{jk}$, 记

$$\begin{aligned}M_j^{(k)}(y_j)&\widehat{=}M_j^{(k)}\big(x_1^{(k)},x_2^{(k)},\cdots,x_{j-1}^{(k)},y_j,x_{j+1}^{(k)},\cdots,x_m^{(k)}\big)\\&=E\big[\widehat{p}_j\big(x_1^{(k)},x_2^{(k)},\cdots,x_{j-1}^{(k)},y_j,x_{j+1}^{(k)},\cdots,x_m^{(k)}\big)|F_{2k-2}\big].\end{aligned} \tag{2.5}$$

若密度函数 $h(u,v)$ 满足：对任意 v, 有

$$\int_0^{x_1}h(u,v)\mathrm{d}u\geq(m-1)\Big|\int_{x_j}^{y_j}h(u,v)\mathrm{d}u\Big|,\quad j\leq m-1;$$

对任意 u, 有

$$\int_{x_m}^{M_1}h(u,v)\mathrm{d}v\geq(m-1)\Big|\int_{x_j}^{y_j}h(u,v)\mathrm{d}v\Big|,\quad j\geq 2.$$

则在 $x_{l-1}<\widehat{x}_l<x_{l+1}$ $(l=1,2,\cdots,m)$ 的条件下, 对每一个 j, 有

1) 当 $y_j<x_j$ 时, $M_j^{(k)}(y_j)$ 单调递增;

2) 当 $y_j>x_j$ 时, $M_j^{(k)}(y_j)$ 单调递减.

证　见文 [3] 附录 D.

引理 2.1 为下面的假设 2.6 提供了必要条件.

假设 2.6　对给定的 j $(j=1,2,\cdots,m)$, 由引理 2.1 定义的 $M_j^{(k)}(y)$ 和由 (2.4) 式定义的序列 $\{x_j^{(k)}\}$, 存在与 k 无关的常数 L 及常数 C_0, 对任意 $c\in(0,C_0)$, 有

$$E\big\{\big[M_j^{(k)}\big(x_j^{(k)}+c\big)-M_j^{(k)}\big(x_j^{(k)}-c\big)\big]\big(x_j-x_j^{(k)}\big)\big\}\geq cL_jE(x_j^{(k)}-x_j)^2. \tag{2.6}$$

定理 2.1　若引理 2.1 的条件以及假设 2.1 至假设 2.6 都成立, 则有 $Z^{(k)}\to Z$, a.s.

证　先证 $B_n=O(n^{-(2\eta-\varepsilon)})$, 取

$$a_n=\frac{1}{n^{1-\varepsilon}},\quad \varepsilon\in\Big(0,\frac{1}{5}\Big);\quad c_{jn}=n^{-\frac{1}{2}+\eta_j},\quad \eta_j\in\Big(\frac{2}{3}\varepsilon+\frac{1}{6},\frac{1}{2}-\varepsilon\Big).$$

$$\begin{aligned}B_{n+1}&=E(Z^{(n+1)}-Z)^{\mathrm{T}}(Z^{(n+1)}-Z)\\&=B_n+\sum_{j=1}^m\frac{2a_n}{c_{jn}}E[(\widehat{p}_j(x_j^{(n)}+c_{jn})-\widehat{p}_j(x_j^{(n)}-c_{jn}))(x_j^{(n)}-x_j)]\\&\quad+\sum_{j=1}^m\frac{a_n^2}{c_{jn}^2}E(\widehat{p}_j(x_j^{(n)}+c_{jn})-\widehat{p}_j(x_j^{(n)}-c_{jn}))^2\end{aligned}$$

$$
\begin{aligned}
&= B_n + \sum_{j=1}^{m} \frac{2a_n}{c_{jn}} E\Big\{(x_j^{(n)} - x_j) E[(\widehat{p}_j(x_j^{(n)} + c_{jn}) - \widehat{p}_j(x_j^{(n)} - c_{jn}))|F_{2k-2}]\Big\} \\
&\quad + \sum_{j=1}^{m} \frac{a_n^2}{c_{jn}^2} E(\widehat{p}_j(x_j^{(n)} + c_{jn}) - \widehat{p}_j(x_j^{(n)} - c_{jn}))^2 \\
&\le B_n + \sum_{j=1}^{m} \frac{2a_n}{c_{jn}} (-c_{jn}L) E(x_j^{(n)} - x_j)^2 + \sum_{j=1}^{m} \frac{a_n^2}{c_{jn}^2} \\
&\le B_n(1 - 2a_nL) + \sum_{j=1}^{m} \frac{a_n^2}{c_{jn}^2} \\
&= B_n\Big(1 - \frac{2L}{n^{1-\varepsilon}}\Big) + \sum_{j=1}^{m} \frac{1}{n^{2(1-\varepsilon)-(1-2\eta_j)}} \\
&\le B_n\Big(1 - \frac{2L}{n^{1-\varepsilon}}\Big) + \frac{m}{n^{1+2(\eta-\varepsilon)}}, \quad \eta = \min(\eta_1, \eta_2, \cdots, \eta_m).
\end{aligned}
$$

由文 [4] 引理 2.1, 有

$$
\overline{\lim_{n\to\infty}} n^{2\eta-\varepsilon} B_n \le \frac{m}{2L}. \tag{2.7}
$$

所以

$$
B_n = O(n^{-(2\eta-\varepsilon)}). \tag{2.8}
$$

其次，证明存在 $\{Z^{(k)}\}$ 的子序列 $\{Z^{(k_j)}, j = 1, 2, \cdots\}$ 以概率 1 收敛到 Z. 选取 β 满足

$$
(2\eta - \varepsilon)^{-1} < \beta < \Big(\frac{1}{2} - \eta + \varepsilon\Big)^{-1}. \tag{2.9}
$$

对自然数 $k = 1, 2, \cdots$, 定义 $n_k = [k^\beta]$, 由契贝晓夫不等式

$$
P(\|Z^{(n_k)} - Z\| \ge \delta) \le \frac{E\|Z^{(n_k)} - Z\|^2}{\delta^2} = \frac{O(k^{-\beta(2\eta-\varepsilon)})}{\delta^2} = O(k^{-\beta(2\eta-\varepsilon)}). \tag{2.10}
$$

由 (2.9) 式， $\beta(2\eta - \varepsilon) > 1$, 所以 $\sum k^{-\beta(2\eta-\varepsilon)} < \infty$, 从而 $Z^{(n_k)} \to Z$ a.s.

最后，证明 $\{Z^{(n)}\}$ 以概率 1 收敛到 Z. 对 $n_k \le n < n_{k+1}$, 有

$$
\begin{aligned}
\|Z^{(n)} - Z\| &\le \|Z^{(n)} - Z^{(n_k)}\| + \|Z^{(n_k)} - Z\| \\
&\le \sum_{s=n_k}^{n_{k+1}} \|Z^{(s+1)} - Z^{(s)}\| + \|Z^{(n_k)} - Z\| \\
&\le \sum_{s=n_k}^{n_{k+1}} \sqrt{\sum_{j=1}^{m} \frac{a_s^2}{c_{js}^2}} + \|Z^{(n_k)} - Z\| \\
&\le \sum_{s=n_k}^{n_{k+1}} \sqrt{\frac{m}{s^{1+2\eta-2\varepsilon}}} + \|Z^{(n_k)} - Z\|.
\end{aligned} \tag{2.11}
$$

因为 $n_{k+1}-n_k=O(k^{\beta-1})$, 由 β 的选取 (2.9) 知

$$\sum_{s=n_k}^{n_{k+1}}\sqrt{\frac{m}{s^{1+2\eta-2\varepsilon}}}=O(k^{\beta-1})\sqrt{m}k^{-\beta(\frac{1}{2}+\eta-\varepsilon)} \tag{2.12}$$

$$=O\big(n^{(\varepsilon-\eta+\frac{1}{2})-\frac{1}{\beta}}\big)=o(1). \tag{2.13}$$

将 (2.13) 式代入 (2.11) 式，当 $k\to\infty$ 时， $\|Z^{(n)}-Z\|\to 0$ a.s.

实际上，类似文 [4] 中的 (2.39) 式， (2.11) 式可以更精确地表示成

$$\|Z^{(n)}-Z\|\le O(k^{\beta-1})\sqrt{m}k^{-\beta(\frac{1}{2}+\eta-\varepsilon)}+o(n^{-\varphi}) \quad \text{a.s.} \tag{2.14}$$

其中 φ 满足 $\frac{1-8\varepsilon}{3}>\varphi>\frac{[1-8\varepsilon-\sigma(1+\varepsilon)]}{3}>0$.

与文 [4] 中收敛速度的讨论类似，可得下列定理.

定理 2.2 在引理 2.1 的条件以及假设 2.1 至假设 2.6 都成立的条件下，由 (2.4) 定义的向量序列 $\{Z^{(n)}\}$ 的收敛速度满足 $\|Z^{(n)}-Z\|=o(n^{-\frac{1}{3}+\sigma})$. 其中 σ 为一个可以趋于零的正数.

由于 X 的分布函数是 $F(t)=\sum\limits_{t_j\le t}\frac{1}{m}$, 因此可定义 $F(t)$ 的估计为 $\widehat{F}_n(t)=\sum\limits_{\widehat{t}_j\le t}\frac{1}{m}$. 定理 2.1 和定理 2.2 给出了这个估计量的性质.

3　关于连续分布的情况

3.1　一些假设和说明

假设 3.1 设随机变量 X 的密度函数 $f(x)$ 未知，满足 $A_1\ge f(x)\ge A_0>0$, $f(x)$ 的支撑是 $[0,M_1]$, 其中 M_1 为已知的， A_0 和 A_1 都是常数.

假设 3.2 随机向量 (U,V) 的密度函数 $h(u,v)$ 未知，满足 $A_2\ge h(u,v)>0$, $(u,v)\in D_H$, $D_H=\{(u,v):0<u<v<M_2\}$, 其中 M_2 可以未知但要求 $M_2>M_1$, A_2 为常数.

引理 3.1 若假设 3.1 成立，则对任意给定的自然数 m, 存在一个离散的随机变量 Y, 其分布列为 $P(Y=t_i)=\frac{1}{m}$ $(i=1,2,\cdots,m)$, 且它的分布函数 F_Y 与 X 的分布函数 F_X 满足 $\sup\limits_{t\in(0,\infty)}|F_X(t)-F_Y(t)|=\frac{1}{m}$.

3.2　分布函数 $F(x)$ 的估计及其性质

设 X 是一个连续的随机变量，与随机向量 (U,V) 独立，且满足假设 3.1 和 3.2, 对固定的自然数 m, 考虑 X 的分布函数的估计问题.

根据第 2 节的讨论，利用样本，可以得到由引理 3.1 定义的离散随机变量 Y 的取值估计 $\widehat{T}=\{t_0,\widehat{t}_1,\cdots,\widehat{t}_{m-1},t_m\}$, 而且由 (2.14) 式，有

$$\max_{1\le j\le m}|\widehat{t}_j-t_j|=O(k^{\beta-1})\sqrt{m}k^{-\beta(\frac{1}{2}+\eta-\varepsilon)}+o(n^{-\varphi})=o(1) \quad \text{a.s.} \tag{3.1}$$

其中常数 φ 满足 $\frac{1-8\varepsilon}{3}>\varphi>\frac{[1-8\varepsilon-\sigma(1+\varepsilon)]}{3}$, 其中 $\sigma=\frac{8\varepsilon}{2-\varepsilon}$.

由于 Y 的分布函数是 $F_Y(t)=\sum\limits_{t_j\le t}\frac{1}{m}$, 因此，可定义 $F_Y(t)$ 的估计量为

$$\widehat{F}_Y(t)=\sum_{\widehat{t}_j\le t}\frac{1}{m}. \tag{3.2}$$

对一切实数 t, 有

$$|F_X(t)-\widehat{F}_Y(t)|\le\frac{1}{m}+2A_1\max_i|\widehat{t}_i-t_i|. \tag{3.3}$$

由 (3.1) 式，对固定的 m, $\lim\limits_{n\to\infty}\sup\limits_{t\in(0,\infty)}|F_X(t)-\widehat{F}_Y(t)|=\frac{1}{m}$.

其次，考虑 m 不为固定数，对给定的样本数 n, 假设 $m=g(n)$, 这时我们希望有下面的关系

$$m=g(n)\to\infty,\quad n\to\infty \tag{3.4}$$

且

$$\max_{1\le j\le m}|\widehat{t}_j-t_j|\to 0,\quad n\to\infty. \tag{3.5}$$

满足 (3.4) 和 (3.5) 式的 m 是否存在呢？答案是肯定的，首先我们有下列定理.

定理 3.1 若假设 3.1 至 3.2 及上节关于估计 $\widehat{T}=\{t_0,\widehat{t}_1,\cdots,\widehat{t}_{m-1},t_m\}$ 的假设都成立, 设样本数为 n, 取 $m=g(n)=n^{\frac{1}{5}}$, $\varepsilon\in(0,\frac{1}{100})$, $\eta=\frac{1}{2}-2\varepsilon$, $\beta=\frac{3}{1+\varepsilon}$, 则有 $|F_X(t)-\widehat{F}_Y(t)|=O(n^{-\frac{1}{5}})$. 其中 $\widehat{F}_Y(t)=\sum\limits_{\widehat{t}_j\le t}\frac{1}{m}$ 由 (3.2) 式定义，是 $F_X(x)$ 的估计.

证 对一切实数 t, 由 (3.3) 式得

$$\begin{aligned}|F_X(t)-\widehat{F}_Y(t)|&\le\frac{1}{m}+2A_1\max_i|\widehat{t}_i-t_i|\\&=n^{-\frac{1}{5}}+2A_1\max_i|\widehat{t}_i-t_i|.\end{aligned} \tag{3.6}$$

所以，我们只需证明 $\max\limits_j|\widehat{t}_j-t_j|=o(n^{-\frac{1}{5}})$ 即可.

依定理 2.1, 有

$$\begin{aligned}B_{n+1}&=E(Z^{(n+1)}-Z)^{\mathrm{T}}(Z^{(n+1)}-Z)\\&\le B_n\Big(1-\frac{2L}{n^{1-\varepsilon}}\Big)+\frac{m}{n^{1+2(\eta-\varepsilon)}}\\&=B_n\Big(1-\frac{2L}{n^{1-\varepsilon}}\Big)+\frac{1}{n^{\frac{9}{5}-6\varepsilon}}.\end{aligned}$$

由文 [4] 引理 2.1 得

$$\overline{\lim_{n\to\infty}}\,n^{\frac{9}{5}-6\varepsilon-(1-\varepsilon)}B_n\le\frac{1}{2L},$$

所以

$$B_n=O(n^{-(\frac{4}{5}-5\varepsilon)}). \tag{3.7}$$

当 $n_k\le n<n_{k+1}$ 时 $\Big($记 $\Delta=\big[\frac{3}{1+\varepsilon}\big]\cdot\big[\frac{2}{5}-\frac{5}{2}\varepsilon-\frac{(1+\varepsilon)^2}{6}\big]>0.5>0\Big)$

$$\begin{aligned}P\big(n^{\frac{\Delta}{\beta}}\|Z^{(n_k)}-Z\|\ge\delta\big)&\le\frac{n^{\frac{2\Delta}{\beta}}E\|Z^{(n_k)}-Z\|^2}{\delta^2}\\&=\frac{O(k^{-\beta(\frac{4}{5}-5\varepsilon)+2\Delta})}{\delta^2}\\&=O(k^{-(1+\varepsilon)}),\end{aligned}$$

即

$$\|Z^{(n_k)} - Z\| = o\big(n^{-\frac{\Delta}{\beta}}\big) \quad \text{a.s.} \tag{3.8}$$

当 $n_k \le n < n_{k+1}$ 时，

$$\begin{aligned}\|Z^{(n)} - Z\| &\le \|Z^{(n)} - Z^{(n_k)}\| + \|Z^{(n_k)} - Z\| \\ &\le \sum_{l=n_k}^{n_{k+1}} \sqrt{\frac{m}{l^{1+2\eta-\varepsilon}}} + \|Z^{(n_k)} - Z\| \\ &= o\Big(n^{-\left(\frac{7}{30}-\frac{17}{6}\varepsilon-\frac{\varepsilon^2}{6}\right)}\Big).\end{aligned}$$

故

$$\max_{1\le j\le m} |\widehat{t_j} - t_j| \le \|Z^{(n)} - Z\| = o\Big(n^{-\left(\frac{7}{30}-\frac{17}{6}\varepsilon-\frac{\varepsilon^2}{6}\right)}\Big) = o\big(n^{-\frac{1}{5}}\big). \tag{3.9}$$

在定理 3.1 中，如果 $g(n) = n^s$, 则 s 有一个变化范围，这就是下面的定理 3.2.

定理 3.2 若定理 3.1 条件成立, 设样本数为 n, 取 $m = g(n) = n^s$, $\varepsilon \in (0, \frac{1}{100})$, $\eta = \frac{1}{2} - 2\varepsilon$, 则有

1) $\max\limits_{1\le j\le m} |\widehat{t_j} - t_j| = o(n^{-\Delta_1})$;

2) $|F_X(t) - \widehat{F}_Y(t)| = O(n^{-s})$, 其中

$$s \in \Big(0, \frac{2}{9}(1 - 9\varepsilon - \varepsilon^2)\Big), \quad \Delta_1 = 1 - 2\varepsilon - \frac{1}{2}s - \frac{1}{3}(1+\varepsilon)(2+\varepsilon). \tag{3.10}$$

$\widehat{F}_Y(t) = \sum\limits_{\widehat{t_j}\le t} \frac{1}{m}$ 是 $F_X(x)$ 的估计.

证　类似 (3.7) 式，有 $B_n = O(n^{-(1-s-5\varepsilon)})$, 所以

$$P\Big(n^{\frac{\Delta}{\beta}}\|Z^{(n_k)} - Z\| \ge \delta\Big) \le \frac{n^{2\frac{\Delta}{\beta}} E\|Z^{(n_k)} - Z\|^2}{\delta^2} = O(k^{-\beta(1-s-5\varepsilon)+2\Delta}).$$

若取 Δ 满足

$$-\beta(1 - s - 5\varepsilon) + 2\Delta = -(1 + 2\varepsilon), \tag{3.11}$$

则 (3.8) 式成立. 另外，若取

$$\beta\Big(\frac{1}{2}s + 3\varepsilon\Big) - 1 = -(\Delta + \varepsilon), \tag{3.12}$$

则 (3.9) 式可表示为

$$\max_{1\le j\le m} |\widehat{t_j} - t_j| \le \|Z^{(n)} - Z\| = O(k^{-1+\beta(3\varepsilon+\frac{s}{2})}) + o\big(n^{-\frac{\Delta}{\beta}}\big) = o\big(n^{-\frac{\Delta}{\beta}}\big). \tag{3.13}$$

将 (3.11) 和 (3.12) 代入 (3.13) 得定理的第一部分. 再由 $\frac{\Delta}{\beta} > s$, 有 $s \in (0, \frac{2}{9}\cdot(1 - 9\varepsilon - \varepsilon^2))$. 最后，由 (3.6) 式得

$$\big|F_X(t) - \widehat{F}_Y(t)\big| \le \frac{1}{m} + \frac{2}{M_1}\max_i |\widehat{t_i} - t_i| = n^{-s} + o\big(n^{-\frac{\Delta}{\beta}}\big) = O(n^{-s}).$$

注 1 在定理 3.2 中, 当 $s=\frac{1}{5}$ 时, 则有 $\max\limits_{1\le j\le m}|\widehat{t}_j-t_j|=o(n^{-\frac{1}{5}})$, $|F_X(t)-\widehat{F}_Y(t)|=O(n^{-\frac{1}{5}})$. 这就是定理 3.1 的结论.

注 2 分布函数估计量的收敛速度最快可为 $O\big(n^{-\frac{2}{9}+\sigma}\big)$.

参 考 文 献

[1] Geskus R and Groeneboom P. Asymptotically optimal estimation of smooth functional for interval-censoring, Case 2. *Ann. Statistics*, 1999, **27**: 627–674.

[2] Groeneboom P and Wellner J A. Information Bounds and Nonparametric Maximum Likelihood Estimation. DMV Seminar Band 19, Birkhauser, Basel, 1992.

[3] 丁邦俊. 关于区间数据的分布函数估计. 复旦大学博士学位论文, 2002.

[4] 丁邦俊, 郑祖康. 在区间截断的情况下, 两点分布估计及其收敛速度. 应用概率统计, 2007, **23**(3): 292–302.

[5] 郑祖康, 丁邦俊. 关于区间数据的分布函数估计问题. 应用概率统计, 2004, **20**(2): 119–125.

[6] Derman C. Stochastic approximation. *Ann. Math. Stat.*, 1956, **27**: 879–886.

[7] Efron B and Petrosian V. Nonparametric methods for double truncated data. *JASA.*, 1999, **94**: 824–834.

[8] Frydman H. A note on nonparametric estimation of the distribution function from interval-censored and truncated observation. *J. Roy. Statistics, Soc. B*, 1994, **56**(1): 71–74.

[9] Lee C. An urn model in the simulation of interval censored failure time data. *Statistics & Probability Letter*, 1999, **45**: 131–139.

[10] Topp R and Gomez G. Residual analysis in linear regression models with an interval-censored covariate. *Statistics in Medicine*, 2004, **23**: 3377–3391.

[11] Turnbull B W. The empirical distribution with arbitrarily grouped, censored and truncated data. *J. Roy. Statistics, Soc. B*, 1976, **38**: 290–295.

THE ESTIMATION OF A DISTRIBUTION WITH INTERVAL CENSORED DATA AND ITS CONVERGENCE RATE

DING Bangjun

(*Department of Statistics, East China Normal University, Shanghai* 200062)

Abstract Firstly, the result with binomial distribution is generalized to random variable X with function $P(X=x_i)=\frac{1}{m}$ for $i=1,2,\cdots,m$. With this result, a procedure is proposed to obtain an estimation of distribution function with continuous variable, the convergence rate of the estimation is also obtained.

Key words Interval censored data, stochastic approximation, convergence rate.