

定时截尾缺失数据下指数分布的统计推断

田 霆 刘次华

(华侨大学数学系, 福建 泉州 362021)

摘要 试验数据缺失是产品寿命试验中经常遇到的情况, 处理起来也比较复杂. 文中在寿命分布为指数分布时, 给出寻求定时截尾寿命试验数据缺失场合下样本分布参数点估计的一种近似方法. 通过大量的 Monte-Carlo 数值模拟试验, 在缺失数据数目不太大的情况下, 参数估计的精度令人满意. 文中还从理论上证明, 可利用枢轴量 $\frac{\bar{m}}{m}$ 的分布对参数 m 作区间估计.

关键词 指数分布, 定时截尾数据缺失, 似然函数, 柯西定理

中图分类号 O 213.2

文献标识码 A

在用统计方法处理实际问题时, 常会遇到数据缺失问题. 譬如在产品寿命试验中, 由于试验设备、观测手段, 或者其他方面的困难造成某些试验数据丢失或未观测到的现象等. 因此, 对不完全数据的处理是统计分析的一个重要领域. 下面表述寿命试验中的缺失问题. 设产品寿命 T 服从指数分布, 其分布函数为

$$F(t; \theta) = 1 - e^{-t/\theta}, \quad t > 0.$$

在上式中, $\theta > 0$ 是平均寿命, $\lambda = 1/\theta$ 为失效率. 取 n 个产品同时参加定时截尾试验, 试验进行到 (τ 是预先给定的正数) 时刻停止. 设在 τ 时刻以前有 r 个产品失效, 记相应的失效时间为 t_1, t_2, \dots, t_r , 总试验时间 $T = \sum_{j=1}^r t_j + (n-r)\tau$. 若由于某种原因造成数据丢失, 不妨设剩下的数据为

$$0 < t_{r_1+1} < \dots < t_{r_1+s_1} < t_{r_2+1} < \dots < t_{r_2+s_2} < \dots < t_{r_k+1} < \dots < t_{r_k+s_k}.$$

其中

$$\{r_1 + 1, \dots, r_1 + s_1, r_2 + 1, \dots, r_2 + s_2, \dots, r_k + 1, \dots, r_k + s_k\} \subset \{1, 2, \dots, n\}. \quad (1)$$

对于试验数据 (1) 的统计分析, 已有相关的研究报道. 文 [1] 给出平均寿命 θ 的 Bayes 估计、近似极大似然估计与无偏估计; 文 [2] 给出了定时截尾数据缺失场合下指数分布参数的 Bayes 估计; 文 [3] 给出了定数截尾缺失数据下, Weibull 分布参数的点估计及区间估计. 同时, 文 [3] 通过大量的 Monte-Carlo 模拟说明, 在样本比较大且缺失数较小的情况下, 其点估计的精确度是令人满意的. 本文给出定时截尾缺失数据下, 指数分布参数的点估计及区间估计.

1 参数的点估计

下面用概率元方法^[4], 求出样本 (1) 的似然函数 $L_n(t | \theta)$, 并以 $t_{r_1+1}, \dots, t_{r_1+s_1}; t_{r_2+1}, \dots, t_{r_2+s_2}; \dots; t_{r_k+1}, \dots, t_{r_k+s_k}$ 分别表示 $T_{r_1+1}, \dots, T_{r_1+s_1}; T_{r_2+1}, \dots, T_{r_2+s_2}; \dots; T_{r_k+1}, \dots, T_{r_k+s_k}$ 的观察值 (T_i 表示第 i 个产品的失效时间). 这时就把数轴分为如下区间, 即 $(-\infty, t_{r_1+1})$, $[t_{r_1+1}, t_{r_1+s_1})$, $[t_{r_1+s_1}, t_{r_2+1})$, $[t_{r_2+1}, t_{r_2+s_2})$, $[t_{r_2+s_2}, t_{r_3+1})$, $[t_{r_3+1}, t_{r_3+s_3})$, $[t_{r_3+s_3}, t_{r_4+1})$, $[t_{r_4+1}, t_{r_4+s_4})$, $[t_{r_4+s_4}, t_{r_5+1})$, $[t_{r_5+1}, t_{r_5+s_5})$, $[t_{r_5+s_5}, t_{r_6+1})$, $[t_{r_6+1}, t_{r_6+s_6})$, $[t_{r_6+s_6}, t_{r_7+1})$, $[t_{r_7+1}, t_{r_7+s_7})$, $[t_{r_7+s_7}, t_{r_8+1})$, $[t_{r_8+1}, t_{r_8+s_8})$, $[t_{r_8+s_8}, t_{r_9+1})$, $[t_{r_9+1}, t_{r_9+s_9})$, $[t_{r_9+s_9}, t_{r_{10}+1})$, $[t_{r_{10}+1}, t_{r_{10}+s_{10}})$, $[t_{r_{10}+s_{10}}, \infty)$; $i = 1, \dots, k, j = 1, \dots, s_i$. 其中, $dt_{r_{i+j}}$ 都充分小. 这样, 有 n 个观察值落在第 $[j]$ 个区间, $\sum_{i=1}^k s_i$ 个落在第 $[j]$ 个区间, $\sum_{i=1}^{k-1} r_{i+1} - r_i - s_i$ 个落

收稿日期 2005-06-21

作者简介 田 霆 (1972-), 男, 助教, 硕士, 主要从事产品可靠性的研究. E-mail: tianting_1972928@sohu.com

基金项目 福建省自然科学基金资助项目 (Z0511027)

在第 $[j-1, j]$ 个区间, $r - (r_k + s_k)$ 个落在第 $[j-1, j]$ 个区间, $n - r$ 个落在第 $[j, j+1]$ 个区间. 记 $s = \sum_{i=1}^k s_i$, $s(t) = \sum_{i=1}^k t_{r_i+s_i}$, $m_i = r_{i+1} - r_i - s_i$, $t_{r_{k+1}} = t_0 = s_0 = t_0 = 0$, $r_{i+j} = r$. 据多项分布 $T_{r_1+1}, \dots, T_{r_1+s_1}; T_{r_2+1}, \dots, T_{r_2+s_2}; \dots, T_{r_k+1}, \dots, T_{r_k+s_k}$ 的概率元, 则为

$$L(t_{r_1+1}, \dots, t_{r_1+s_1}; t_{r_2+1}, \dots, t_{r_2+s_2}; \dots, t_{r_k+1}, \dots, t_{r_k+s_k}) dt_{r_1+1} \dots dt_{r_k+s_k} = \\ C [F(t_{r_1+1})]^{r_1} \cdot \prod_{i=1}^k [f(t_{r_i+j})]^{m_i} [F(t_{r_{i+1}+1}) - F(t_{r_i+s_i})]^{m_i} \cdot \\ [F(t_{r_k+1}) - F(t_{r_k+s_k})]^{r-r_k-s_k} \cdot [1 - F(t_{r_k+1})]^{n-r} dt_{r_1+1} \dots dt_{r_k+s_k} + 0(dt_{r_1+1}) \dots + 0(dt_{r_k+s_k}).$$

上式中, C 是与 t 无关的常数. 将式子两边约去 $dt_{r_1+1} \dots dt_{r_k+s_k}$ 后, 再让 $dt_{r_1+1} \dots dt_{r_k+s_k}$ 都趋于零. 最后, 得到 $T_{r_1+1}, \dots, T_{r_1+s_1}; T_{r_2+1}, \dots, T_{r_2+s_2}; \dots, T_{r_k+1}, \dots, T_{r_k+s_k}$ 的联合密度为

$$L(t) = C [1 - e^{-\frac{t_{r_1+1}}{m}}]^{r_1} \cdot \prod_{i=1}^k \exp^{-s(t)/m_i} [e^{-\frac{t_{r_i+s_i}}{m_i}} - e^{-\frac{t_{r_{i+1}+1}}{m_i}}]^{m_i} \cdot [e^{-\frac{t_{r_k+s_k}}{m}} - e^{-\frac{t_{r_{k+1}+1}}{m}}]^{r-r_k-s_k} \cdot [e^{-\frac{t_{r_k+1}}{m}}]^{n-r}.$$

将 $[e^{-\frac{t_{r_k+s_k}}{m}} - e^{-\frac{t_{r_{k+1}+1}}{m}}]^{r-r_k-s_k}$ 化为 $[e^{-\frac{t_{r_k+s_k}}{m}} - e^{-\frac{t_{r_{k+1}+1}}{m}}]^{m_k}$, 归并到 $[e^{-\frac{t_{r_i+s_i}}{m_i}} - e^{-\frac{t_{r_{i+1}+1}}{m_i}}]^{m_i}$ 项中. 则上式可化为

$$L(t) = C \cdot \exp^{-s(t)/m} \prod_{i=1}^k [e^{-\frac{t_{r_i+s_i}}{m_i}} - e^{-\frac{t_{r_{i+1}+1}}{m_i}}]^{m_i} [1 - e^{-\frac{t_{r_1+1}}{m}}]^{r_1} [e^{-\frac{t_{r_k+1}}{m}}]^{n-r} = \\ C \cdot \exp^{-s(t)/m} \prod_{i=0}^k [e^{-\frac{t_{r_i+s_i}}{m_i}} - e^{-\frac{t_{r_{i+1}+1}}{m_i}}]^{m_i} [e^{-\frac{t_{r_1+1}}{m}}]^{n-r}$$

C 为参数 m 无关的常数. 则在上式中, $\ln L(t) = \ln C - s(t)/m + \sum_{i=0}^k m_i \ln [e^{-\frac{t_{r_i+s_i}}{m_i}} - e^{-\frac{t_{r_{i+1}+1}}{m_i}}] - (n-r)$

一. 令 $m = \frac{1}{m}$, 则可化为

$$\ln L(m) = \ln C + s(t)/m - \sum_{i=0}^k m_i \ln [e^{-m t_{r_i+s_i}} - e^{-m t_{r_{i+1}+1}}] - m(n-r),$$

$$\frac{\partial \ln(m)}{\partial m} = \frac{s}{m} - s(t) - \sum_{i=0}^k m_i \frac{t_{r_i+s_i} e^{-m t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-m t_{r_{i+1}+1}}}{e^{-m t_{r_i+s_i}} - e^{-m t_{r_{i+1}+1}}} - (n-r).$$

令 $\frac{\partial \ln L(m)}{\partial m} = 0$, 则可得方程为

$$\frac{s}{m} - \sum_{i=0}^k m_i \frac{t_{r_i+s_i} e^{-m t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-m t_{r_{i+1}+1}}}{e^{-m t_{r_i+s_i}} - e^{-m t_{r_{i+1}+1}}} - s(t) - (n-r) = 0. \quad (2)$$

引理 1 方程(2)有唯一解.

证明 将方程(2)变形为

$$\frac{s}{m} - \sum_{i=0}^k m_i \frac{t_{r_i+s_i} e^{-m t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-m t_{r_{i+1}+1}}}{e^{-m t_{r_i+s_i}} - e^{-m t_{r_{i+1}+1}}} = s(t) + (n-r).$$

令 $g(m) = \frac{s}{m} - \sum_{i=0}^k m_i \frac{t_{r_i+s_i} e^{-m t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-m t_{r_{i+1}+1}}}{e^{-m t_{r_i+s_i}} - e^{-m t_{r_{i+1}+1}}} = \frac{s}{m} - \sum_{i=0}^k m_i \frac{t_{r_{i+1}+1} e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - t_{r_i+s_i}}{e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - 1}$. 于是, 有

$$g'(m) = -\frac{s}{m^2} - \sum_{i=0}^k m_i \frac{-t_{r_{i+1}+1} e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} (t_{r_{i+1}+1} - t_{r_i+s_i}) [e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - 1]}{[e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - 1]^2} = \\ \sum_{i=0}^k m_i \frac{[t_{r_{i+1}+1} e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - t_{r_i+s_i}] e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} (t_{r_{i+1}+1} - t_{r_i+s_i})}{[e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - 1]^2} = \\ -\frac{s}{m^2} - \sum_{i=0}^k m_i \frac{t_{r_{i+1}+1} e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} (t_{r_{i+1}+1} - t_{r_i+s_i}) - t_{r_i+s_i} e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} (t_{r_{i+1}+1} - t_{r_i+s_i})}{[e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - 1]^2} = \\ -\frac{s}{m^2} - \sum_{i=0}^k m_i \frac{e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} (t_{r_{i+1}+1} - t_{r_i+s_i})^2}{[e^{-m(t_{r_{i+1}+1} - t_{r_i+s_i})} - 1]^2} < 0.$$

即 $g(m)$ 是 m 的严格单调减函数. 令 $g_1(x) = x e^{-mx}$, $g_2(x) = e^{-mx}$, 则 $g_1'(x) = e^{-mx} - m x e^{-mx}$, $g_2'(x) = -m e^{-mx}$. 由柯西定理存在 ξ , $t_{r_i+s_i} < \xi < t_{r_{i+1}+1}$, 有

$$\frac{t_{r_i+s_i} e^{-m t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-m t_{r_{i+1}+1}}}{e^{-m t_{r_i+s_i}} - e^{-m t_{r_{i+1}+1}}} = \frac{(1 - m \xi) e^{-m \xi}}{-m e^{-m \xi}} = \xi - \frac{1}{m}.$$

于是,有

g(m) = \frac{s}{m} - \sum_{i=0}^k m_i (\frac{1}{i} - \frac{1}{m}) = \frac{1}{m} [s + \sum_{i=0}^k m_i] - \sum_{i=0}^k m_i \frac{1}{i} = \frac{r}{m} - \sum_{i=0}^k m_i \frac{1}{i},

进而可得到

\frac{r}{m} = \sum_{i=0}^k m_i \frac{1}{i} + s(t) + (n - r).

由此所得的方程(2)有唯一根,记为 \overline{m} 则

\frac{r}{s(t) + (n - r) + \sum_{i=0}^k m_i t_{r_i+1} + 1} < \overline{m} < \frac{r}{s(t) + (n - r) + \sum_{i=0}^k m_i t_{r_i+s_i} + 1}. \tag{3}

令 m^* = \overline{m}/m,易知 m^* 的分布与参数无关.

2 m^* 的分布

性质 m^* 的分布与参数无关.

证明 由于

\frac{s}{m} - \sum_{i=0}^k m_i \frac{t_{r_i+s_i} e^{-\frac{r}{m} t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-\frac{r}{m} t_{r_{i+1}+1}}}{e^{-\frac{r}{m} t_{r_i+s_i}} - e^{-\frac{r}{m} t_{r_{i+1}+1}}} - s(t) - (n - r) = 0, \\ \frac{s}{m^*} \cdot \frac{1}{m} - \sum_{i=0}^k m_i \frac{t_{r_i+s_i} e^{-m^* \cdot \frac{r}{m} t_{r_i+s_i}} - t_{r_{i+1}+1} e^{-m^* \cdot \frac{r}{m} t_{r_{i+1}+1}}}{e^{-m^* \cdot \frac{r}{m} t_{r_i+s_i}} - e^{-m^* \cdot \frac{r}{m} t_{r_{i+1}+1}}} - s(t) - (n - r) = 0, \\ \frac{s}{m^*} - \sum_{i=0}^k m_i \frac{m t_{r_i+s_i} e^{-m^* \cdot \frac{r}{m} t_{r_i+s_i}} - m t_{r_{i+1}+1} e^{-m^* \cdot \frac{r}{m} t_{r_{i+1}+1}}}{e^{-m^* \cdot \frac{r}{m} t_{r_i+s_i}} - e^{-m^* \cdot \frac{r}{m} t_{r_{i+1}+1}}} - m s(t) - (n - r) m = 0.

由式(2)可知, m^* 为母体 (m=1) 的标准指数分布的参数的估计量,所以 m^* 的分布与参数 m 无关.

由此可知, \frac{\overline{m}}{m} 为一枢轴量,故理论上可以利用枢轴量 \frac{\overline{m}}{m} 的分布对参数 m 作区间估计.这里不作进一步的讨论.

3 Monte-Carlo 模拟实验

关于参数估计的精度,利用式(3)对真值进行了 2 000 次 Monte-Carlo 模拟实验,部分模拟结果如表 1 所示.表中, m(m=\frac{1}{m}) 的区间估计为 [min(m), max(m)].从表中可以看到,当 n 固定时,随着 k 的增大,精度愈高;当 k 很小时,即缺失数太大,参数估计误差偏大.因此,我们应尽量避免数据缺失.总的来说,在缺失数据数目不太大时,参数估计的精度还是令人满意的.

表 1 部分模拟结果

序号	m			n	r	k	min(m)	max(m)
1	1.0	1.000 0	1.2	20	14	5	1.073 6	1.229 6
2	1.0	1.000 0	1.2	20	15	5	0.910 4	1.009 3
3	1.0	1.000 0	1.2	20	15	5	1.046 0	1.172 1
4	1.0	1.000 0	1.2	20	15	5	1.007 4	1.102 3
5	1.0	1.000 0	1.2	20	12	5	0.830 5	0.867 8
6	1.0	1.000 0	1.2	20	13	6	0.894 6	0.951 5
7	1.0	1.000 0	1.2	20	16	6	1.231 9	1.308 1
8	1.0	1.000 0	1.2	20	10	6	0.600 0	0.655 0
9	1.0	1.000 0	1.2	20	11	6	0.709 3	0.786 1
10	1.0	1.000 0	1.2	20	14	6	1.050 6	1.141 0
11	2.4	0.416 7	2.8	20	16	7	0.447 1	0.523 0

续表								
序号		m		n	r	k	$\min(m)$	$\max(m)$
12	2.4	0.416 7	2.8	20	14	7	0.348 2	0.421 4
13	2.4	0.416 7	2.8	20	11	7	0.275 5	0.311 3
14	2.4	0.416 7	2.8	20	12	7	0.335 9	0.392 0
15	2.4	0.416 7	2.8	20	13	7	0.361 0	0.461 8
16	2.4	0.416 7	2.8	20	13	8	0.372 0	0.432 7
17	2.4	0.416 7	2.8	20	12	8	0.297 9	0.340 4
18	2.4	0.416 7	2.8	20	14	8	0.381 2	0.463 8
19	2.4	0.416 7	2.8	20	11	8	0.245 9	0.321 9
20	2.4	0.416 7	2.8	20	14	8	0.355 8	0.453 6
21	3.0	0.333 3	3.5	20	13	9	0.254 8	0.311 9
22	3.0	0.333 3	3.5	20	13	9	0.267 7	0.400 2
23	3.0	0.333 3	3.5	20	16	9	0.412 2	0.505 6
24	3.0	0.333 3	3.5	20	15	9	0.339 4	0.408 8
25	3.0	0.333 3	3.5	20	11	9	0.206 7	0.270 0
26	3.0	0.333 3	3.5	20	12	8	0.229 7	0.261 4
27	3.0	0.333 3	3.5	20	14	8	0.323 7	0.385 7
28	3.0	0.333 3	3.5	20	15	8	0.347 5	0.404 9
29	3.0	0.333 3	3.5	20	13	8	0.257 7	0.322 0
30	3.0	0.333 3	3.5	20	11	8	0.223 2	0.271 7
31	4.0	0.250 0	4.5	20	13	7	0.192 9	0.228 8
32	4.0	0.250 0	4.5	20	14	7	0.251 4	0.271 8
33	4.0	0.250 0	4.5	20	16	7	0.328 0	0.362 8
34	4.0	0.250 0	4.5	20	14	7	0.240 4	0.269 1
35	4.0	0.250 0	4.5	20	16	7	0.339 5	0.383 6
36	4.0	0.250 0	4.5	20	14	10	0.218 8	0.317 8
37	4.0	0.250 0	4.5	20	14	10	0.226 5	0.283 4
38	4.0	0.250 0	4.5	20	13	10	0.231 1	0.324 1
39	4.0	0.250 0	4.5	20	16	10	0.282 4	0.372 1
40	4.0	0.250 0	4.5	20	13	10	0.219 8	0.289 5

参 考 文 献

1

Balasubranmanian K, Blakrishna N. Estimation for one and two-parameter exponential distributions under multiple type- censoring[J]. Statistical Papers,1992,33:203 ~ 216

2

王乃生,王玲玲.定数截尾数据缺失场合下指数分布参数的 Bayes 估计[J].应用概率统计,2001,8(3):229 ~ 235

3

徐晓岭.定数截尾缺失数据下 Weibull 分布的统计推断[J].应用概率统计,1997,12(4):363 ~ 370

4

茆诗松,王静龙,濮晓龙,等.高等数理统计[M].北京:高等教育出版社,施普林格出版社,2002.35 ~ 39

Statistical Inference for the Exponential Distribution
under Multiply Type- Censoring
Tian Ting Liu Cihua
(Department of Mathematics, Huaqiao University, 362021, Quanzhou, China)

Abstract Mulitply type- censoring is a more general and much more complicated censoring mode. When product life time follows exponential distribution, an approximate method of the point-estimation for the exponential distribution under multiply type- censoring is provided. By the Monte-Carlo simulation, the precision for the point estimation under a large amount of samples with a small amount of lost data is successfully achieved. It is proved theoretically that interval estimation can be acquired by the distribution of the pivot $\frac{\overline{m}}{m}$.

Keywords exponential distribution, failare data of timing censoring, likelihood function, Cauchy theorem