基于 EM 算法的无失效数据的参数估计[1]

纪志荣　　黄可明
（福州大学数学系，福建 福州　350002）
（E-mail：jzrong666@126.com）

摘　要：本文利用 EM 算法，在无失效数据样本下，引进失效信息，对指数分布的参数进行估计，得到了参数所满足的非线性方程，并利用 EM 算法的收敛性，保证得到的非线性方程解，正是收敛于参数真值的估计。最后对实际数据进行计算，结果合理。

关键词：无失效数据；EM 算法；多项分布
MR(2000)主题分类：62N01；62N02
中图分类号：O213.2

1. 引言

在可靠性试验中，常获得各类截尾数据。在定时截尾试验中，有时会遇到“无失效数据”（即在规定的截尾时间内无产品失效），特别在高可靠、小样本问题中，更容易产生无失效数据，对无失效数据进行研究越来越具有理论和实用价值。

自从文[1]发表以来，对无失效数据的研究逐渐引起了重视，并已经取得了一些成果。文[2]对无失效数据可靠性进展进行了综述。文[3]中综述有错检验模型的研究进展情况，并提醒人们注意检验方法本身的错误，在实际问题中有些工程技术人员认为根据无失效数据直接进行可靠性评估可能会产生“冒进”。考虑到在无失效数据的分析中引进失效信息，作为无失效数据的一个附加信息，在一定程度上弥补数据的不足，从而提高数据分析的精度，文[4]提出了等效失效数的概念；文[5]提出了无失效数据情形下参数的综合估计法。

Dempster 等人提出的 EM 算法是处理不完全数据样本的有效方法，文[6]使用广义线性模型和 EM 算法，对成败型数据下 Weibull 分布的参数进行了讨论。并特别指出，对无失效数据样本，在最后一个试验点引进失效数，不仅符合工程经验，并可保证 EM 算法的收敛性。本文则直接利用 EM 算法，分析中引进失效信息，并充分提取样本潜在信息，对无失效数据下指数分布的参数进行估计，得到了参数所满足的非线性方程，并利用 EM 算法的收敛性，保证得到的非线性方程的解正是收敛于参数真值的估计。最后，对实际数据进行计算分析，结

[1]福建省自然科学基金资助项目（E0310015）
作者简介：纪志荣，女，硕士研究生，研究方向：概率论与数理统计

果符合专家工程经验。

2．数据模型和 EM 算法

在可靠性试验设计中，进行 m 组定时截尾试验，截尾时刻为 $t_i(i=1,2,...,m)$，相应试验样品数为 n_i（$i=1,2,...,m$），到试验结束时，所有样品均无一失效，称 (t_i,n_i) 为无失效数据，其中 $\sum_{i=1}^{m} n_i = n$ 为总的试验样品数。

关于 EM 算法，文[7]中进行了综述。设 X 和 Y 是两个样本空间，且从 X 到 Y 存在一个多对一的映射 $x \to y(x)$，这里 x 为完全数据，它是不能测试到的，我们能测试到的仅是 Y 内的 y，称为不完全数据。设完全数据的密度函数为 $p(x|q)$，其中参数 $q=(q_1,q_2,...,q_P)\in\Theta$，EM 算法如下进行，给定参数 q 的初值 $q^{(0)}$ 后，对 $k=0,1,2...$ 执行以下两步：

E 步：计算 $Q(q|q^{(k)},Y)=E[\ln p(x|q)|q^{(k)},Y]$

M 步：求 $q^{(k+1)}\in\Theta$，使 $Q(q|q^{(k)},Y)$ 极大化，即 $Q(q^{(k+1)}|q^{(k)},Y)=\max_{q} Q(q|q^{(k)},Y)$

将上述 E 步和 M 步进行迭代运算，直至 $\left\|q^{(k+1)}-q^{(k)}\right\|$ 或 $\left\|Q(q^{(k+1)}|q^{(k)},Y)-Q(q^{(k)}|q^{(k)},Y)\right\|$ 充分小为止。

见文[8]根据 EM 算法的一般理论，$p(q^{(k)}|Y)$ 关于 k 是递增的，上述迭代序列 $\{q^{(k)}\}$ 收敛到 $p(q|Y)$ 的极值点。其中，$p(q|Y)$ 为样本的似然函数，$p(q^{(k)}|Y)$ 为第 k 步样本似然函数的估计。

3. 无失效数据情形下参数的估计

在无失效数据样本下，可以验证，不能直接求得参数的极大似然估计。 如果记产品的实际寿命为 $x_i(i=1,2,...,n)$，不仅丢失样本信息，利用 EM 算法，也可验证,得不到参数的估计。所以本文从另一个角度，把无失效数据转换成区间型数据，充分提取样本潜在信息，并利用文[5][6]中的思想，在最后截尾时刻，引进失效信息，应用 EM 算法的收敛性，得到参数的估计。

设无失效数据为 (t_i,n_i) $(i=1,2,...,m)$，在最后一个截尾时刻 t_m 赋予其失效数 r，通常，

令 $r=\frac{1}{2}$，见文[6],这样符合工程经验。则在时刻 t_m 未失效样本数为 $n_m-\frac{1}{2}$。记从所研究的总体中抽取的 n 件产品在时间区间 $(t_{i-1},t_i]$ 中实际失效的件数为 $x_i(i=1,2,...,m+1)$，其中，令 $t_0=0$，$t_{m+1}=+\infty$。把已知数据 (t_i,n_i) 作为不完全数据，并令 $Y=(n_1,n_2,...,n_m)$。

显然，$X=(x_1,x_2,...,x_{m+1})$ 为离散型随机变量，服从多项分布，密度函数为

$$p(x|q)=\frac{n!}{\prod_{i=1}^{m+1}x_i!}\prod_{i=1}^{m+1}[F(t_i|q)-F(t_{i-1}|q)]^{x_i}$$

其中，$\mathrm{F}(x|q)$ 是总体的分布函数，执行 EM 算法的两步：

E 步：

$$\begin{aligned}Q(q|q^{(k)},Y) &= E\left(\ln p(x|q)\middle|q^{(k)},Y\right)\\ &= E(\ln n!-\sum_{i=1}^{m+1}\ln x_i!+\sum_{i=1}^{m+1}x_i\ln(F(t_i|q)-F(t_{i-1}|q))\middle|q^{(k)},Y)\\ &= \ln n!-\sum_{i=1}^{m+1}E(\ln x_i!|q^{(k)},Y)+\sum_{i=1}^{m+1}\ln(F(t_i|q)-F(t_{i-1}|q))E(x_i|q^{(k)},Y)\end{aligned}\qquad(1)$$

M 步：将 $Q(q|q^{(k)},Y)$ 分别对 $q_l(l=1,2,...,p)$ 求导，以求出使 $Q(q|q^{(k)},Y)$ 极大化的点

$$q^{(k+1)}=(q_1^{(k+1)},q_2^{(k+1)},...,q_p^{(k+1)})$$

显然(1)式中第一项，第二项与 q 无关。

令 $t_i=F(t_i|q)-F(t_{i-1}|q)$，则有

$$\frac{\partial Q(q|q^{(k)},Y)}{\partial q_l}=\sum_{i=1}^{m+1}\frac{\partial t_i}{\partial q_l}\frac{1}{t_i}E(x_i|q^{(k)},Y)=0\qquad l=1,2,...,p\qquad(2)$$

满足(2)式的 q 即为所要寻找的 $q^{(k+1)}$，这样就形成了 $q^{(k)}\to q^{(k+1)}$ 的一次迭代，反复利用(1)，(2)这两个迭代公式就可得到参数的估计。

上述迭代过程中，关键是求出 $E(x_i|q^{(k)},Y)$。为此，得到如下定理：

定理 在无失效数据模型 (t_i,n_i)，$i=1,2,...,m$ 下，在时刻 t_m 引进失效数 r，则

$$E(x_1 \mid q^{(k)}, Y) = r\frac{t_1^{(k)}}{F(t_m \mid q^{(k)})}$$

$$E(x_i \mid q^{(k)}, Y) = \sum_{j=1}^{i-1} n_j \frac{t_i^{(k)}}{1-F(t_j \mid q^{(k)})} + r\frac{t_i^{(k)}}{F(t_m \mid q^{(k)})} \qquad i = 2,...,m$$

$$E(x_{m+1} \mid q^{(k)}, Y) = \sum_{j=1}^{m-1} n_j \frac{t_{m+1}^{(k)}}{1-F(t_j \mid q^{(k)})} + (n_m - r)$$

其中，$t_i^{(k)}$ 为 t_i 的第 k 步迭代值。

证明：设 $n_j\,(j=1,2,...,m)$ 个样本在 $(t_{i-1},t_i](i=1,2,...,m+1)$ 中实际失效的个数为 x_{ij}， n_m 个样本中引入失效数 r，在 $(t_{i-1},t_i]$ 中实际失效的个数为 r_i，则

$$n_j = \sum_{i=j}^{m+1} x_{ij}\ ;\quad x_1 = r_1\ ;\quad x_i = \sum_{j=1}^{i-1} x_{ij} + r_i\ ,\quad i = 2,...,m\ ;\quad x_{m+1} = \sum_{j=1}^{m-1} x_{(m+1)j} + (n_m - r)$$

给定 $(t_j, n_j)\ \ j=1,2,...,m$， x_{ij} **：** $B(n_j, \frac{t_i}{1-F(t_j)})$ ，则 $E(x_{ij}) = n_j\frac{t_i}{1-F(t_j)}$

又 r_i **：** $B(r, \frac{t_i}{F(t_m)})$ ，则 $E(r_i) = r\frac{t_i}{F(t_m)}$

所以 $i=2,...,m$ 时，有

$$\begin{aligned} E(x_i \mid q^{(k)}, Y) &= E(\sum_{j=1}^{i-1} x_{ij} + r_i \mid q^{(k)}, Y) = \sum_{j=1}^{i-1} E(x_{ij} \mid q^{(k)}, Y) + E(r_i \mid q^{(k)}, Y) \\ &= \sum_{j=1}^{i-1} n_j \frac{t_i^{(k)}}{1-F(t_j \mid q^{(k)})} + r\frac{t_i^{(k)}}{F(t_m \mid q^{(k)})} \end{aligned}$$

$$E(x_1 \mid q^{(k)}, Y) = r\frac{t_1^{(k)}}{F(t_m \mid q^{(k)})}$$

$$\begin{aligned} E(x_{m+1} \mid q^{(k)}, Y) &= E(\sum_{j=1}^{m-1} x_{(m+1)j} + (n_m - r)) = \sum_{j=1}^{m-1} E(x_{(m+1)j} \mid q^{(k)}, Y) + (n_m - r) \\ &= \sum_{j=1}^{m-1} n_j \frac{t_{m+1}^{(k)}}{1-F(t_j \mid q^{(k)})} + (n_m - r) \end{aligned}$$

证毕。

下面就无失效数据下指数分布，给出参数的估计。

3.1 指数分布无失效数据参数的估计

设产品寿命 T 服从指数分布，其密度函数为 $f(t)=\begin{cases} l e^{-lt} & t>0 \\ 0 & t\le 0 \end{cases}$，分布函数为 $F(t)=\begin{cases} 1-e^{-lt} & t>0 \\ 0 & t\le 0 \end{cases}$ 此时，参数 $q=l$，$p=1$。令 $r=\frac{1}{2}$，由定理可得

$$E(x_1 \mid l^{(k)},Y)=\frac{1}{2}\frac{F(t_1 \mid l^{(k)})-F(t_0 \mid l^{(k)})}{F(t_m \mid l^{(k)})}=\frac{1}{2}\frac{1-e^{-l^{(k)}t_1}}{1-e^{-l^{(k)}t_m}}$$

$$\begin{aligned} E(x_i \mid l^{(k)},Y) &= \sum_{j=1}^{i-1} n_j \frac{F(t_i \mid l^{(k)})-F(t_{i-1} \mid l^{(k)})}{1-F(t_j \mid l^{(k)})}+\frac{1}{2}\frac{F(t_i \mid l^{(k)})-F(t_{i-1} \mid l^{(k)})}{F(t_m \mid l^{(k)})} \qquad (i=2,...,m) \\ &= \sum_{j=1}^{i-1} n_j \frac{e^{-l^{(k)}t_{i-1}}-e^{-l^{(k)}t_i}}{e^{-l^{(k)}t_j}}+\frac{1}{2}\frac{e^{-l^{(k)}t_{i-1}}-e^{-l^{(k)}t_i}}{1-e^{-l^{(k)}t_m}} \\ &= \sum_{j=1}^{i-1} n_j \frac{t_i^{(k)}}{e^{-l^{(k)}t_j}}+\frac{1}{2}\frac{t_i^{(k)}}{1-e^{-l^{(k)}t_m}} \end{aligned}$$

$$\begin{aligned} E(x_{m+1} \mid l^{(k)},Y) &= \sum_{j=1}^{m-1} n_j \frac{F(t_{m+1} \mid l^{(k)})-F(t_m \mid l^{(k)})}{1-F(t_j \mid l^{(k)})}+(n_m-\frac{1}{2}) \\ &= \sum_{j=1}^{m-1} n_j \frac{e^{-l^{(k)}t_m}}{e^{-l^{(k)}t_j}}+(n_m-\frac{1}{2}) \end{aligned}$$

$$\frac{dt_i}{dl}=t_i e^{-lt_i}-t_{i-1}e^{-lt_{i-1}} \qquad i=2,...,m$$

$$\frac{dt_1}{dl}=t_1 e^{-lt_1},\quad \frac{dt_{m+1}}{dl}=-t_m e^{-lt_m}$$

由 $\frac{dQ(l \mid l^{(k)})}{dl}=\sum_{i=1}^{m+1}\frac{dt_i}{dl}\frac{1}{t_i}E(x_i \mid l^{(k)},Y)=0$ 得到

$$\frac{t_1 e^{-lt_1}}{1-e^{-lt_1}}(\frac{1}{2}\frac{1-e^{-l^{(k)}t_1}}{1-e^{-l^{(k)}t_m}})+\sum_{i=2}^{m}(t_i e^{-lt_i}-t_{i-1}e^{-lt_{i-1}})\frac{t_i^{(k)}}{t_i}(\sum_{j=1}^{i-1} n_j e^{l^{(k)}t_j}+\frac{1}{2}\frac{1}{1-e^{-l^{(k)}t_m}})$$

$$+(-t_m)[\sum_{j=1}^{m-1} n_j \frac{t_{m+1}^{(k)}}{e^{-l^{(k)}t_j}}+(n_m-\frac{1}{2})]=0 \qquad (3)$$

(3)式即为由 EM 算法得到的迭代方程，由 EM 算法的一般理论可知(见文[8])，如此循环迭代，得到的序列 $\{l^{(k)}\}$ 是收敛于真值 l 的估计，故方程(3)中可令 $l^{(k)}\to l$，$l^{(k)}$ 用 l 代替，

整理可得

$$1-e^{-\hat{l}t_m}=\frac{\frac{1}{2}t_m}{\sum_{i=1}^{m}n_i t_i} \tag{4}$$

(4)正是参数 l 所满足的非线性方程，如文[8],所求得的解正是收敛于参数 l 的估计。

注意到(4)式左边正是时刻t_m的失效概率，右边是失效时间/总的试验时间，得到的时刻t_m的失效概率恰是经验估计。

4. 实际数据计算分析

本文以文献[4]中给出的某型号发动机的寿命数据为例，阐述上述方法的计算结果。本文同其它文献一样，以专家工程经验作为估计结果好坏的一个方法，即：发动机通过大量试验，都无一失效，故认为其可靠度是相当高的，特别其寿命T超过 1000 秒的概率不会低于 0.95。下表给出无失效数据(t_i, n_i)

i	1	2	3	4	5	6	7	8	9	10	11	12	13
t_i	100.18	109.93	115.01	130.15	150.00	179.94	190.36	250.15	783.00	849.94	870.03	909.77	1450.30
n_i	3	21	2	1	3	8	1	1	4	3	1	1	2

由(4)，可得平均失效率 $\hat{l}=3.2688\times10^{-5}$，与文[4]结果相近；且$\hat{R}(1000)=0.96784$(寿命$T$超过 1000 秒的概率)说明本文结果符合专家工程经验，方法可行。

参 考 文 献

[1]Martz H F, Waller R A. A Bayes Zero-failure(BAZE) Reliability Demonstration Testing Procedure. Journal of Quality Technology. 1979,11(3):128-137.

[2]无失效数据可靠性进展,数学进展,2002,31(1):7-19.

[3]吴喜之,有错检验模型,应用概率统计,1993,9(3):310-318.

[4]无失效数据可靠性参数的综合估计,数学理论与应用,2000,20(3):36-44.

[5]张忠占,杨振海.等效失效数在无失效数据分析中的应用.数理统计与应用概率,1991,6(3):394-404.

[6]王学仁,石磊.成败型寿命试验—GLM 和 E-M 算法,1994,14(1):29-38.

[7]王松桂.EM 算法,应用数学与计算数学, 1983(6):43-48.

[8]茆诗松,王静龙,濮晓龙.高等数理统计,北京:高等教育出版社,1998.7.

Estimation of Parameter

Based on Zero-Failure Data via EM Algorithm

Ji Zhirong Huang Keming

(Department of Mathematics, Fuzhou University, Fuzhou Fujian, 350002)

(E-mail:jzrong666@126.com)

Abstract In this paper, under the zero-failure data, we study parameter estimation for exponential distribution by importing failure information and using the EM algorithm, and obtain nonlinear equation the parameter satisfies, whose solution is to be the convergent estimation to the real value by the convergence property of EM algorithm. Finally, we get the reasonable result through analyzing actual zero-failure data.

Keywords: zero-failure data; EM algorithm; multinomial distribution

MR(2000) Subject Classification: 62N01;62N02

Chinese Library Classification: O213.2