

常见寿命数据类型及生命表的编制方法

宛新荣, 王梦军, 王广和, 钟文勤*

(农业虫害综合治理研究国家重点实验室, 中国科学院动物研究所, 北京 100080)

摘要:生命表是描述种群死亡过程的有用工具。介绍了 4 种常见的寿命数据类型: 寿终数据、右删失数据、左删失数据和区间型数据特征及其相应的数据分析处理方法即生命表法、乘积限估计和 Turnbull 估计法。对生命表法和乘积限估计法应用上的特点进行了比较, 同时还对特殊的寿命数据类型——截断数据做了简要介绍。

关键词:生命表; 删失数据; 截断数据; 生命表法; 乘积限估计; Turnbull 估计

Some types of life data and methods for constructing life table

WAN Xin-Rong, WANG Meng-Jun, WANG Guang-He, ZHONG Wen-Qin (State Key Lab of Integrated Management of Pest Insects & Rodents, Institute of Zoology, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Life table is a convenient tool for depicting the mortality process of a population. The characteristics of four common types of life data, i. e. complete data, right-censored data, left-censored data, and interval data are introduced. Three methods for constructing life table, i. e. Life Table Method, Product-Limit Estimation, and Turnbull Estimation regarding to these types of life data are presented. Comparisons are made on the features of the Life Table Method and the Product-Limit Estimation in life table constructing procedure. A special type of life data, the truncated data, is introduced.

Key words: life table; censored data; truncated data; life table method; product-limit estimation; turnbull estimation

文章编号: 1000-0933(2001)04-0660-05 中图分类号: Q141 文献标识码: A

生命表(Life table)是描述种群死亡过程的有用工具。其最初出现在人口统计学,尤其是在人寿保险事业上,用于估计人的期望寿命。在早期有关生命表研究的文献中,有关人的生命表很多,而讨论动植物生命表的则较少^[1,2]。近年来,生命表受到研究人员的普遍重视,越来越多的科学工作者着手动植物生命表的编制工作^[3~10]。

生命表分为动态生命表(Dynamic life table)和静态生命表(Static life table)两大类。受取样方法的限制,静态生命表的数据类型为单一的寿终数据(Complete data),分析方法也简单;相反,动态生命表的数据类型较复杂,除了寿终数据外,还兼有删失型数据(Censored data)和截断型数据(Truncated data)。目前多数学者主要沿用经典的无删失机制的生命表法编制生命表,所处理数据均为寿终数据,这样,两类生命表在实际编制方法上不存在任何差别。

随着科学技术的发展,人类对客观世界的认识不断地深入,传统的生命表编制方法也遇到了许多新问题。近年来,由于删失机制(Censoring)和截断机制(Truncation)的引进,使寿命数据类型多样化并更加接近于客观实际,在此基础上发展出许多新的统计方法,形成许多新的理论^[11]。遗憾的是,这些先进的统计方

基金项目: 中国科学院知识创新重大项目(KSCX1-08)、国家自然科学基金(39870136)及中国科学院内蒙古草原生态系统定位研究站资助。

收稿日期: 1999-02-08; 修订日期: 1999-09-30

作者简介: 宛新荣(1969~),男,安徽人,博士,副研究员。主要从事动物种群生态学研究。

* 通讯联系人

法却不能及时地为大多数从事生命表研究的学者所掌握和应用。国内外一些著名的生态学教科书,也均未提到上述具有删失机制的数据处理方法^[1,2,12~19],这在很大程度上制约了新的统计方法的推广与应用。因此,作者在总结一些相关文献的基础上,对常用寿命数据类型及其分析方法做简要介绍。

1 常用寿命数据类型

根据 Nelson 的定义^[20],常见寿命数据可分为 4 种类型(见图 1):寿终数据(Complete data)又叫完全寿命数据,即一个个体的确切寿命已知。自然情况下,一般数据多为此种类型。在静态生命分析法中所有个体的寿命数据均按寿终数据处理,这是寿命数据最重要的一种类型。

右删失数据(Right-censored data) 当一个个体的确切寿命不知道,只知道其寿命大于某个值。如生命数据收集过程中由人为因素或其它不可避免的因素造成的数据收集中断。例如:取样过程中有意去除一部分的个体(removals);重捕操作中的捕捉丧失(Losses on traps)或追踪个体迁出样地而中断追踪,这些个体的实际寿命实际上大于上述事件发生时该动物的年龄,但其确切值无从知道。右删失数据在野外调查中极为常见,虽然比例不高,但在种群存活率估计时必须考虑这一类型数据,否则影响计算结果的正确性。

左删失数据(Left-censored data) 左删失数据定义为一个个体的确切寿命不知道,只知道其寿命小于某个值。在某种条件下,例如人们知道动物的初生群数目,但在第一次取样时即有部分个体已死亡,这些个体的具体死亡时间是不知道的,只知道它们的寿命要小于这个时间段的长度。因此,这部分早期死亡的个体可视为左删失数据。

区间型数据(Interval or Group data):即确切寿命不知道,仅知道其寿命介于两个值之间,这种数据类型可理解为在左端右删失,而在右端左删失。区间型数据在实际中也很多,尤其在按一定间隔调查取样的数据收集过程中。在某种意义上说,区间型数据类型含义最广,它实际包括了上述 3 种寿命数据类型^[11,20]。

2 常见生命数据的非参数统计分析方法

针对不同的寿命数据类型,也有不同的分析处理方法。常用的有 3 种非参数统计方法,如生命表法(Life table method)、乘积限估计(Oproduct-limit estimation)和 Turnbull 估计(Turnbull estimation),下面分别介绍这 3 种方法。

2.1 生命表法

生命表法可以处理寿终数据、右删失数据及区间型数据。具体地说,它只需要知道在每个寿命区间里有多少个寿终数据或右删失数据,而不一定要知道其具体寿终时间或右删失时间。它是历史最悠久而目前仍大量使用的生命表统计方法。关于生命表估计法,有一些变体^[21],但其基本原理都相似,下面给出标准生命表的估算方法:

区间 I_j ;死亡数 D_j ;右删失数 R_j ;存活数 N_j ;死亡率 $= D_j / (N_j - R_j / 2)$;区间存活率 $P_j = 1 - D_j / (N_j - R_j / 2)$;存活率 $S(t_i) = P_1 P_2 \cdots P_j$ 。

其中, I_j 表示某一时间间隔(区间), D_j 表示区间 I_j 中的寿终个体数, R_j 表示区间 I_j 中的右删失个体数(传统分析方法往往忽略此项,而按死亡处理), N_j 表示经历区间 I_j 后仍然存活的个体数。

2.2 乘积限估计法

乘积限估计法(P-L 估计)是由 Kaplan 和 Meier 于 1958 年首次提出的^[22],因此又叫 K-M 估计。此法只针对于寿终数据和右删失数据,区间型数据不能应用此法。其估计方法如下:

①将所有寿命数据由小到大排列。若一个右删失数据和一个寿终数据相等时,则寿终数据要排列在前;当 t_i 为寿终数据时,令 $\delta(i)=1$,当 t_i 是右删失数据时,令 $\delta(i)=0$,即有 $t_1 \leq t_2 \leq \cdots \leq t_i \leq \cdots \leq t_n$ 。其中 n 为全部寿命数据个数。

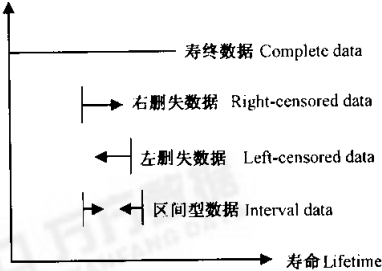


图 1 常见寿命数据类型示意图

Fig. 1 Illustration for common types of life data

②生存函数的估计值 $S(t_i) = \prod_{t(i) \leq t} ((n-i)/(n-i+1))^{1-\delta(i)}$; 这里约定 $\prod_{\phi} = 1, \phi$ 为空集。

2.3 Turnbull 估计

如果在一组数据中,既含有寿终数据和右删失数据,又含左删失数据,那么如何估计生存函数呢?前文介绍的生命表法和乘积限估计法均无能为力,需要发展新方法。Turnbull^[23]首先就分组数据的情形提出一种方法,在此基础上 Tsai and Crowley^[24]和 Chang and Yang^[25]就非分组数据的情形也给出了估计法。两种估计方法在数学上是相似的,统称 Turnbull 估计,下面简要介绍其估算步骤:

(1)利用寿终数据分割数据为一些区间 I_j ,设在寿命数据中的最小值为 t_1 ,而 t_1 不一定要为寿终数据;除去 t_1 后剩下的互异寿终数据为 $t_2 < \dots < t_m$,记 $t_{m+1} = \infty, I_j = (t_j, t_{j+1}), (j = 1, \dots, m)$ 这样将 $[t_1, \infty)$ 分为 m 个区间 $I_1, I_2, \dots, I_m$ 。

(2)计算各区间上的统计量 d_j, λ_j 和 μ_j ,其中, d_j 表示在区间 I_j 中寿终数据的个数; λ_j 表示区间 I_j 中的右删失数据个数; μ_j 表示 I_j 中左删失数据个数。

(3)找出函数 $L(\theta_1, \theta_2, \dots, \theta_m) = \prod_{j=1}^m (\theta_{j-1} - \theta_j)^{d_j} \times \theta_j^{\lambda_j} \times (1 - \theta_j)^{\mu_j} \quad (\theta_0 \equiv 1)$

在闭区域“ $1 \geq \theta_1 \geq \theta_2 \geq \dots \geq \theta_m \geq 0$ ”上的最大值点 $(\theta_1, \theta_2, \dots, \theta_m)$ 。

下面介绍常规的求极值方法^[11]:

①设 $d_1 + \mu_1 > 0, \lambda_m > 0$; 那么这个最大值点是下列方程组的唯一解

$$\begin{cases} -d_i/(\theta_{i-1} - \theta_i) + d_{i+1}/(\theta_i - \theta_{i+1}) + \lambda_i/\theta_i - \mu_i/(1 - \theta_i) = 0; (\text{其中 } i = 1, 2, \dots, m-1) \\ -d_m/(\theta_{m-1} - \theta_m) + \lambda_m/\theta_m - \mu_m/(1 - \theta_m) = 0; (1 > \theta_1 > \dots > \theta_m > 0; \theta_0 = 1) \end{cases}$$

②如果 $d_1 + \mu_1 > 0, \lambda_m = 0$; 那么这个最大值点是下列方程组的唯一解

$$\begin{cases} -d_i/(\theta_{i-1} - \theta_i) + d_{i+1}/(\theta_i - \theta_{i+1}) + \lambda_i/\theta_i - \mu_i/(1 - \theta_i) = 0; (\text{其中 } i = 1, 2, \dots, m-2) \\ -d_{m-1}/(\theta_{m-2} - \theta_{m-1}) + \bar{\lambda}_{m-1}/\theta_{m-1} - \mu_{m-1}/(1 - \theta_{m-1}) = 0; (1 > \theta_1 > \dots > \theta_{m-1} > 0; \theta_0 = 1) \end{cases}$$

其中 $\bar{\lambda}_{m-1} \approx \lambda_{m-1} + d_m$, 注意 $\bar{\lambda}_{m-1} > 0$ 。

③若 $d_i + \mu_1 = 0, \lambda_m > 0$; 那么这个最大值点是下列方程组的唯一解

$$\begin{cases} -d_i/(\theta_{i-1} - \theta_i) + d_{i+1}/(\theta_i - \theta_{i+1}) + \lambda_i/\theta_i - \mu_i/(1 - \theta_i) = 0; (\text{其中 } i = 2, 3, \dots, m-1) \\ -d_m/(\theta_{m-1} - \theta_m) + \lambda_m/\theta_m - \mu_m/(1 - \theta_m) = 0; (1 > \theta_2 > \dots > \theta_m > 0; \theta_1 \approx 1) \end{cases}$$

④如果 $d_1 + \mu_1 = 0, \lambda_m = 0$; 那么这个最大值点是下列方程组的唯一解

$$\begin{cases} -d_i/(\theta_{i-1} - \theta_i) + d_{i+1}/(\theta_i - \theta_{i+1}) + \lambda_i/\theta_i - \mu_i/(1 - \theta_i) = 0; (\text{其中 } i = 2, \dots, m-1) \\ -(d_{m-1}/(\theta_{m-2} - \theta_{m-1})) + \bar{\lambda}_{m-1}/\theta_{m-1} - \mu_{m-1}/(1 - \theta_{m-1}) = 0; (1 > \theta_2 > \dots > \theta_{m-1} > 0; \theta_1 \approx 1) \\ \bar{\lambda}_{m-1} \approx \lambda_{m-1} + d_m \end{cases}$$

(4)生存函数估计值为

$$\begin{cases} S(t) = 1, \text{当 } t \in [0, t_1] \\ S(t) = \theta_j, \text{当 } t \in [t_j, t_{j+1}], (j = 1, 2, \dots, m) \end{cases}$$

2.4 由生存函数编制生命表

从上面的分析可以看到,3 种估计方法的共同点为估算生存函数 $S(t_i)$ 。根据 t_i 与 $S(t_i)$,就可以进行生命表的编制工作。生存函数 $S(t_i)$ 表现为下降的阶梯函数^[26,27],其在水平线上(即存活率不变)无死亡事件,而在垂直下降处(间断点)有死亡发生。与此相对应,生命期望(Lifespan)函数 e_i 则表现为连续下降(斜率为 -1)、在间断点作垂直跃升的锯齿状折线(图 2),它的间断点与生存函数的间断点相吻合。

根据一系列的时间 t_i 及生存函数 $S(t_i)$ 数值,就可以估算其生命期望,步骤如下:

$$\Delta t_i = t_{i+1} - t_i; T_i = \sum_{j=i}^n S(t_j) \Delta t_j; e_i = T_i / S(t_i)$$

其中 e_i 为生命期望,表示年龄为 t_i 的个体,平均还能存活多长时间^[1]。

3 生命表法与乘积限估计法在应用上的特点比较

在上述 3 种估计方法中,Turnbull 估计数据处理很复杂,计算量大;而生命表法与乘积限估计法则相对容易得多。左删失数据在生物学上应用实例并不多见,在此着重讨论生命表法与乘积限估计法在应用上的特点。

(1)当原始寿命数据全为寿终数据和右删失数据时,使用乘积限估计更为精密;当生命表法中的步长(Step)等于 1 时,生命表法完全等价于乘积限估计法,只是更烦琐。

(2)乘积限估计法不能处理区间型数据类型。而生命表法虽然能够处理区间型数据,但并非所有的区间型数据都能用生命表法处理。一般要求满足两个条件:

①动物同时出生,没有时间上的差异;②在每个取样期(动物存活状态调查)中,所有动物均有相同的调查间隔。如果符合上述条件,取得的区间型数据可能类似下列情形:

$[0,10],[10,20],[20,30],[30,45],[45,60],[60,80],\dots$

这样的区间型数据则可以用生命表法计算,但野外条件下能同时满足上述两个条件几乎很少,那么实际观测到的区间型数据更可能如下所示:

$[0,10],[5,15],[8,18],[3,16],[7,23],[9,33],[14,46],\dots$

这样的原始寿命数据根本无法用生命法估算,这表明,生命表法在估算处理区间型数据时,会表现出很大的局限性。

(3)考虑到野外收集到的区间型寿命数据的复杂性和统计方法的局限性,一个合理的建议是将区间型数据进行适当的转换,将其转换为寿终数据,再选择生命表法或乘积限估计法。例如,区间型数据为 $[0,10]$ 时,可转换成寿终数据为 5;区间型数据为 $[3,7]$ 时,可转换为寿终数据为 5。显然,区间越小,估计偏差越小,可靠性程度越高;区间越大,偏差相对增大,因此,此类估计法只能在一定范围内使用^[20]。在数据收集过程中,应力求缩短取样间隔以提高寿命数据的估算精度和可信度。

4 截断数据简介

除了已列举的一些常用的寿命数据类型及其估计方法外,还存在一些特殊的数据类型如截断数据(Truncated data)。其典型情形是:设 Y 是直线上的集合,当随机变量 x 的取值属于 Y 时, x 的值可观测到;当 x 取值不属于 Y 时,则什么数据也观测不到。

在动物种群的寿命数据收集过程中,刚出生的动物往往很难观测到。例如某些穴居的啮齿动物,其幼仔在出窝前的寿命数据(死亡数据)几乎无法观测,只有待其出窝后才能观测到。从这个角度考虑,出窝前这段时间的存活过程实际上不具可观测性,则这批寿命数据即具有截断数据的性质;但是,如果能够推断其初群数目,则这批出窝前死亡数据可视为左删失数据(左删失时间为平均出窝日龄),可按 Turnbull 估计法处理。

截断数据的另外一个来源是在调查动物种群的存活过程时,不同个体进入调查范围的时间或年龄上有所差异,一些个体并非在一开始(出生或孵化)就进入调查取样范围,而是在某个年龄后才进入调查范围^[27,28]。例如某些迁入个体即属于这种情况。此外,由于在实验过程中扩大调查范围而半途加入的个体也可视为左截断数据(Left-truncated data)。显然,左截断数据具有同样的有前提:个体在出生到截断年龄(开始进入调查时的年龄)存活率为 100%,其本身不能真实反映截断年龄以前的种群存活过程,其有效代表时间只能从左截断时刻开始。因此,与删失型数据相同,在实际应用中必须将截断数据与一般数据区别对待,否则将影响结果的准确性。诚然,截断型数据可以排除在分析处理之外而不影响统计分析的正确性(注意删失型数据则不能作排除处理)。然而,对野生动物而言,获取一个个体的寿命数据是如此的不易,往往面

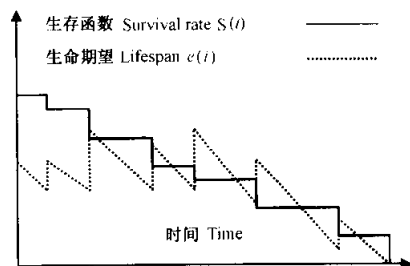


图 2 生存函数 $S(t)$ 、生命期望 $e(i)$ 与(存活)时间的关系

Fig. 2 Relationships of the cumulative survival rate $S(t)$, lifespan $e(i)$ and survival time

临样本量不够的困境;同时,截断数据含有用信息(从截断年龄开始),运用先进的统计学方法,便可以对截断数据加以分析利用。

关于截断数据的统计分析处理方法,需要涉及更多的数学内容,本文不作专门介绍。有兴趣者请参阅文献[27~31]。

参考文献

- [1] 孙儒泳. 动物生态学原理. 北京: 北京师范大学出版社, 1992. 130~146.
- [2] Krebs C J. *Ecology* (fourth edition). Harper and Row, Publishers. New York, 1996. 168~195.
- [3] 梁杰荣, 孙儒泳. 根田鼠生命表和繁殖的研究. *动物学报*, 1985, **31**(2): 170~177.
- [4] 王学高, 戴克华. 高原鼠兔自然寿命研究. *兽类学报*, 1989, **9**(1): 56~62.
- [5] 李进华, 王岐山, 李明. 短尾猴种群生态学研究 III 年龄结构和生命表. *兽类学报*, 1995, **15**(1): 31~35.
- [6] 杨光, 周开亚, 高利安, 常青. 江豚生命表和种群动态的研究. *兽类学报*, 1998, **18**(1): 1~7.
- [7] Connell J H. A predator-prey system in the marine intertidal region, *Balanus glandula* and several predatory species of *Thais*. *Ecological Monographs*, 1970, **40**: 49~78.
- [8] Crisp M D and Lange R T. Age structure distribution and survival under grazing of arid zone shrub *Acacia burkittii*. *Oikos*, 1976, **27**: 86~92.
- [9] Leverich W J and Levin D A. Age-specific survivorship and reproduction in *Phlox drummondii*. *American Naturalist*, 1979, **113**: 881~903.
- [10] Hartman J D. Age determination, age structure, and longevity in the mole, *Scalopus aquaticus* (Mammalia: Insectivora). *Journal of Zoology*, 1995, **237**: 107~122.
- [11] 陈家鼎. 生存分析与可靠性引论. 合肥: 安徽教育出版社, 1982. 1~43.
- [12] 华东师范大学等. 动物生态学. 北京: 人民教育出版社, 1982.
- [13] 丁岩钦. 昆虫数学生态学. 北京: 科学出版社, 1994. 153~214.
- [14] 伊藤嘉昭, 郭祥光, 等译. 动物生态学研究法. 北京: 科学出版社, 1986.
- [15] 伊藤嘉昭. 动物生态学. 北京: 科学出版社, 1982. 32~49.
- [16] Price P W. *Insect ecology*. John Wiley and Sons, New York, 1975. 124~148.
- [17] Begon M, Harper J L and Townsend C R. *Ecology* (third edition). Blackwell Science, Ltd., 1996. 147~168.
- [18] Begon M, Mortimer M and Thompson D J. *Population Ecology*. Blackwell Science, 1996.
- [19] Gutierrez A P. *Applied Population Ecology*. John Wiley and Sons, Inc. New York, 1996.
- [20] Nelson W. *Applied life data analysis*. John Wiley & Sons, New York, 1982. 1~14.
- [21] Breslow N E and Crowley J. A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, 1974, **2**: 437~481.
- [22] Kaplan E L and Meier P. Nonparametric estimation from incomplete observation. *Journal of the American Statistical Association*, 1958, **53**: 457~481.
- [23] Turnbull B W. Nonparametric estimation of a survivorship function with doubly Censored data. *Journal of the American Statistical Association*, 1974, **69**: 169~173.
- [24] Tsai W Y and Crowley J. A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *Ann. Stat.*, 1985, **13**: 1317~1334.
- [25] Chang M N and Yang G L. Strong Consistency of nonparametric estimator of the survival function with doubly Censored data. *Annals of Statistics*, 1987, **15**: 1536~1547.
- [26] Miller R G JR. *Survival Analysis*. John Wiley & Sons, New York, 1981. 1~212.
- [27] Tsai W Y, Jewell N P and Wang M C. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 1987, **74**(4): 883~886.
- [28] Wang M C, Jewell N P and Tsai W Y. Asymptotic properties of the Product Limit estimation under random truncation. *Ann. Stat.*, 1986, **14**(4): 1597~1605.
- [29] Turnbull B W. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 1976, **38**: 290~295.
- [30] Pan W and Chappell R. Estimating survival curves with Left-truncated and Interval-censored data under monotone hazards. *Biometrics*, 1998, **54**: 1053~1060.
- [31] Joly P, Gombast A D and Letenneur L. A Penalized likelihood approach for arbitrarily Censored and Truncated data: Application to age-specific incidence of dementia. *Biometrics*, 1998, **54**: 185~194.