

生存分析理论及其在研究 生命表中的应用

杨凤翔 王顺庆 徐海根 李邦庆

(兰州大学数学系)

摘 要

本文以生存分析理论中的生存函数 $S(t)$ 、累积死亡率函数 $F(t)$ 、死亡密度函数 $f(t)$ 、危险率函数 $\lambda(t)$ 以及与之相应的估计量 $\hat{S}(t)$ 、 $\hat{F}(t)$ 、 $\hat{f}(t)$ 、 $\hat{\lambda}(t)$ 为理论基础,对生命表进行研究,同时为估计这种误差,给出了它们的方差近似公式,并进行应用检验,说明生存分析理论在生态学中有着广泛应用前景^[1-5]。

关键词: 生存分析。

一、生存分析的基本数学理论

1. 基本函数

生物群体的生存分析研究中重要的一项是死亡率(简记C.D.R.)^[6]。总死亡率是一生物群体中总生物体的平均死亡率。公式为:

$$C.D.R. = D_*/N_* \quad (1)$$

式中 $D_* = \sum_i d_{*i}$, $N_* = \sum_i n_{*i}$ 分别是生物群体的死亡总数与生物群体的总数量。 d_{*i} 表示生物群体进入年龄区间 $[t_{i-1}, t_i]$ 的死亡数, n_{*i} 表示进入该发育期的活体数。于是, 年龄区间 $[t_{i-1}, t_i]$ 的平均死亡率为:

$$M_{*i} = d_{*i}/n_{*i} \quad (2)$$

死亡数是死亡率和进入该发育期的活体数的乘积:

$$d_{*i} = n_{*i} \cdot M_{*i} \quad (3)$$

将(3)式代入(1)式得到

$$C.D.R. = \sum_i \frac{n_{*i}}{N_*} \cdot M_{*i} \quad (4)$$

这里求和是对所有年龄组进行的。由此, C.D.R. 是年龄别死亡率的加权平均值, 以各区间实际生物体数比例 n_{*i}/N_* 为权数。可见, C.D.R. 是综合反映死亡经历的指标。是年龄区间 $[t_{i-1}, t_i]$ 的死亡率 M_{*i} 和进入区间的生物群体数量比例 n_{*i}/N_* 两者的函数。作为年龄组死亡率的加权平均, C.D.R. 受生物群体构成的影响。(4)比(1)式意义更深刻。

$$\text{存活率为} \quad 1 - M_{*i} \quad (5)$$

根据这一思想, 定义一随机变量 T 及函数 $S(t)$, t 为发育期, T 为存活期。则

本文于1989年2月6日收到。

$$S(t) = P(T > t), \quad 0 \leq t < \infty \quad (6)$$

$S(t)$ 是 t 的非增函数, 且有 $S(0) = 1$, $S(\infty) = 0$, 把 $S(t)$ 叫做生存率函数。

$$\text{再定义} \quad F(t) = P(T \leq t) = 1 - P(T > t) = 1 - S(t) \quad (7)$$

$F(t)$ 称为积累死亡率函数, 即生物群体到时刻 t 的积累死亡率。死亡密度函数 $f(t)$ 可表示为

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t} = F'(t)$$

从而 $F(t) = \int_0^t f(u) du$ ($F(0) = 0$, $F(\infty) = 1$), 且有

$$S(t) = 1 - F(t) = \int_t^{+\infty} f(u) du \quad (8)$$

利用条件概率 $P(t < T \leq t + \Delta t | T \geq t)$ 来代替 $P(t < T \leq t + \Delta t)$, 则有

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T \geq t)}{\Delta t} \quad (9)$$

称 $\lambda(t)$ 为危险率函数。

由于 $P(t < T \leq t + \Delta t | T \geq t) = \frac{P(t < T \leq t + \Delta t)}{P(T \geq t)}$, 故有

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t)}{\Delta t \cdot S(t)} = \frac{f(t)}{S(t)}$$

或者 $\lambda(t) = -\frac{d}{dt}[\ln S(t)]$, 积分得:

$$S(t) = \exp - \int_0^t \lambda(u) du \quad (10)$$

可见, $S(t)$ 、 $f(t)$ 、 $F(t)$ 、 $\lambda(t)$ 之间可互算, 且构成生存分析理论的基本函数。为生存分析在研究生命表的应用提供了理论依据。

2. 估计量公式

生存时间已达到时间 t 的生物种群, 在时刻 t 的瞬时死亡率 q 及能继续存活概率 P , 对每个时刻 t_i , 由于当时存活的总数为 n_i , 且其中含有 d_i 个在 t_i 时刻死亡, 而其余 $n'_i = n_i - d_i$ 个能继续存活。因此, q_i 和 P_i 可分别用以下两个式子来估算:

$$\hat{q}_i = d_i / n_i, \quad \hat{P}_i = 1 - \hat{q}_i \quad (11)$$

特别当没有生物个体在时刻 t_i 死亡时, $\hat{q}_i = 0$, $\hat{P}_i = 1$ 。

根据 P_i 定义可知, 对任意时间 t , 所求的生存率 $S(t)$ 应是所有在 t 以前的 t_i 所对应的 $\hat{P}_i$ 的乘积, 因此, $S(t)$ 可用下式来估计:

$$\hat{S}_i = \hat{S}(t_i) = \hat{P}_1 \hat{P}_2 \cdots \hat{P}_i \quad \hat{P}_i = \hat{S}_{i-1} \cdot \hat{P}_i \quad (12)$$

据此式则有: $\hat{F}_i = \hat{F}(t_i) = 1 - \hat{S}_i$ (13)

$$\hat{f}(t_{i+1}) \approx \frac{\hat{S}_{i-1} - \hat{S}_i}{h_i} = \frac{\hat{S}_{i-1}}{h_i} \hat{q}_i$$

$$= \frac{1}{h_i} (1 - \hat{P}_i) \hat{P}_{i-1} \cdots \hat{P}_1 \quad (14)$$

$$\hat{\lambda}(t_{mi}) = \hat{f}(t_{mi}) / \hat{S}(t_{mi}) \approx \frac{\hat{2}q_i}{h_i(1 + \hat{P}_i)} = \frac{d_i}{h_i(n_i - d_i/2)} \quad (15)$$

($i = 1, 2, \dots, k$), 这里 $t_{mi} = \frac{1}{2}(t_{i-1} + t_i)$, $h_i = t_i - t_{i-1}$ 是第 i 个发育期的时间区间 $[t_{i-1}, t_i]$ 的中点与长度。其长度不都等。以上是 4 个函数的估计量公式, 是制作生命表的基本公式。

二、生命表的格式

1. 用频率值 $\hat{P}_i, \hat{q}_i$ 来作相应的概率估计, 生物生命表可列成如下格式(见表 1)。 n_i, d_i 都是实际观察数据。

2. $\hat{S}(t_i), \hat{F}(t_i), \hat{f}(t_{mi}), \hat{\lambda}(t_{mi})$ 的方差估计。

表 1 生存分析在研究生命表中的应用

Table 1 Survival analysis and its application to life table

时间区间 (t_{i-1}, t_i)	区间 中点 t_{mi}	区间 长度 h_i	进入该 区间数 n_i	死亡 数 d_i	该区间内		生存率 $\hat{S}_i$	累积 死亡率 $\hat{F}_i$	死亡密度 $\hat{f}(t_{mi})$	危险率 $\hat{\lambda}(t_{mi})$
					死亡率 $\hat{q}_i$	存活率 $\hat{P}_i$				
(t_0, t_1)	t_{m1}	h_1	n_1	d_1	$\hat{q}_1$	$\hat{P}_1$	$\hat{S}_1$	$\hat{F}_1$	$\hat{f}(t_{m1})$	$\hat{\lambda}(t_{m1})$
(t_1, t_2)	t_{m2}	h_2	n_2	d_2	$\hat{q}_2$	$\hat{P}_2$	$\hat{S}_2$	$\hat{F}_2$	$\hat{f}(t_{m2})$	$\hat{\lambda}(t_{m2})$
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.
(t_{k-1}, t_k)	t_{mk}	h_k	n_k	d_k	$\hat{q}_k$	$\hat{P}_k$	$\hat{S}_k$	$\hat{F}_k$	$\hat{f}(t_{mk})$	$\hat{\lambda}(t_{mk})$
(t_k, ∞)	—	—	n_{k+1}	d_{k+1}	1	0	0	1	—	—

设 $\eta = \varphi(\xi)$ 是随机变量 ξ 的任意可微函数, 且 $E(\xi) = \mu, Var(\xi) = \sigma^2$, 将 η 在 μ 的近旁内进行泰勒展开, 舍去 $\xi - \mu$ 的高次项, 则 $\eta \approx \varphi(\mu) + \varphi'(\xi)(\xi - \mu)$ 。因为 $E(\xi - \mu) = 0$ 得 $E(\eta) \approx \varphi(\mu)$ 。可推得

$$Var(\eta) \approx E[\varphi'(\xi)(\xi - \mu)]^2 = (\varphi'_\xi(\xi))^2 \sigma^2 \quad (16)$$

又对于多个随机变量 $\eta = \varphi(\xi_1, \xi_2, \dots, \xi_k)$, 且设

$$E(\xi_i) = \mu_i, Var(\xi_i) = \sigma_i^2, Cov(\xi_i, \xi_j) = \sigma_{ij}$$

($i \neq j$), 则有

$$Var(\eta) = \sum_{i=1}^k \left(\frac{\partial \varphi}{\partial \xi_i} \right)^2 \sigma_i^2 + \sum_{\substack{i \neq j \\ i=1, j=1}}^k \left(\frac{\partial \varphi}{\partial \xi_i} \right) \cdot \left(\frac{\partial \varphi}{\partial \xi_j} \right) \sigma_{ij} \quad (17)$$

特别当 ξ_i 互不相关时, $\sigma_{ij} = Cov(\xi_i, \xi_j) = 0$ ($i \neq j$)

有

$$Var(\eta) \approx \sum_{i=1}^k \left(\frac{\partial \varphi}{\partial \xi_i} \right)^2 \sigma_i^2 \quad (18)$$

由(12)式得

$$\frac{\partial \hat{S}_i}{\partial \hat{P}_j} = -\frac{\hat{S}_i}{\hat{P}_j} \quad (j \leq i) \quad (19)$$

又根据二项分布的方差公式 $Var[\hat{P}_i] = \frac{1}{n_i} \hat{P}_i \hat{q}_i$ 及互不相关性 $Cov(\hat{P}_i, \hat{P}_j) = 0$ ($i \neq j$) 得

$$Var(\hat{S}_i) \approx \sum_{j=1}^i \left(\frac{\hat{S}_i}{\hat{P}_j} \right)^2 \frac{\hat{P}_j \hat{q}_j}{n_j} = \hat{S}_i^2 \sum_{j=1}^i \frac{d_j}{n_j (n_j - d_j)} \quad (20)$$

$$Var(\hat{F}_i) = Var(\hat{S}_i) \quad (21)$$

又据(14)式有

$$Var[\hat{f}(t_{n_i})] \approx [\hat{f}(t_{n_i})]^2 \left(\frac{\hat{P}_i}{n_i \hat{q}_i} + \sum_{j=1}^{i-1} \frac{\hat{q}_j}{n_j \hat{P}_j} \right) \quad (22)$$

$$(i = 1, 2, \dots, h)$$

将(15)式对 d_i 求导得

$$\frac{d\hat{\lambda}(t_{n_i})}{d d_i} = [\hat{\lambda}(t_{n_i})]^2 \frac{h_i}{n_i \hat{q}_i^2}, \text{ 且 } Var(d_i) = n \hat{P}_i \hat{q}_i \text{ 并代入(16)式得:}$$

$$Var[\hat{\lambda}(t_{n_i})] \approx [\hat{\lambda}(t_{n_i})]^4 \left(\frac{h_i^2 \hat{P}_i}{n_i \hat{q}_i^3} \right)^2 \quad (23)$$

根据公式(20)一(23), 可计算得到主要的方差值(见表3)。

从生命表还可以对平均生存期 μ 及中数生存期 $\bar{\mu}$ 进行估计, 其估计量为:

$$\hat{\mu} = \int_0^{\infty} \hat{S}(t) dt \approx \sum_{i=1}^k \hat{S}(t_{n_i}) h_i = \sum_{i=1}^k \frac{f(t_{n_i})}{\hat{\lambda}(t_{n_i})} h_i$$

三、根田鼠生命表的应用研究

现将文献〔2〕中所提供的根田鼠种群的实测数据用生存分析理论制成生命表(表2)及雌性根田鼠种群的生存函数的估计量的方差值(表3), 作出相应的死亡密度函数 $\hat{f}(t)$ 曲线图1, 危险率函数 $\hat{\lambda}(t)$ 曲线图2, 对根田鼠种群进行生存分析研究。

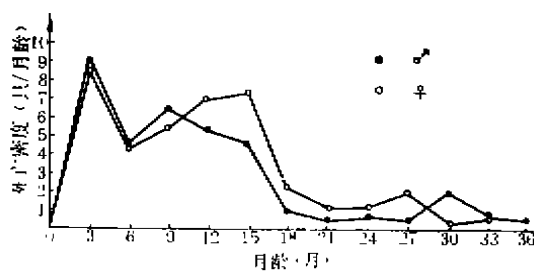


图 1 死亡密度函数曲线
Fig.1 Function curve of mortality density

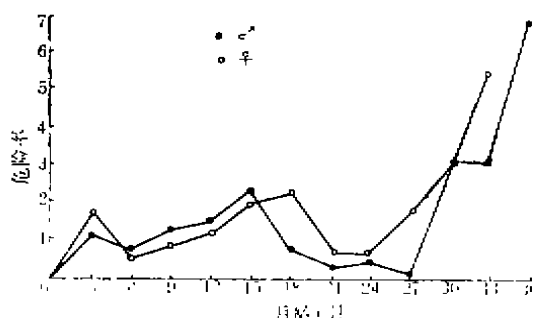


图 2 危险率函数曲线
Fig.2 Function curve of hazard rate

表 2 根田鼠生命表
Table 2 Life table of root vole *Microtus oeconomus*

月龄	性别	存活数 (个)	死亡数 (个)	死亡率 (%)	存活率 (%)	生存率 函数	死亡率 函数	死亡密度 函数	危险率 函数
0—3	雌	105	26	0.2476	0.7524	0.7524	0.2476	0.0825	0.1651
	雄	124	33	0.2661	0.7339	0.7339	0.2661	0.0887	0.1023
3—6	雌	79	13	0.1646	0.8354	0.6286	0.3714	0.0413	0.0624
	雄	91	16	0.1758	0.8242	0.6049	0.3951	0.0430	0.0642
6—9	雌	66	16	0.2424	0.7576	0.6277	0.3723	0.0508	0.0880
	雄	75	23	0.3067	0.6933	0.4186	0.5814	0.0618	0.1208
9—12	雌	50	16	0.3200	0.6800	0.4269	0.5731	0.0670	0.1214
	雄	52	19	0.3654	0.6346	0.2656	0.7344	0.0510	0.1490
12—15	雌	34	17	0.5000	0.5000	0.2134	0.7866	0.0712	0.1984
	雄	33	17	0.5152	0.4848	0.1288	0.8712	0.0456	0.2313
15—18	雌	17	5	0.2941	0.7059	0.1507	0.8493	0.0209	0.2222
	雄	16	3	0.1875	0.8125	0.1046	0.8954	0.0081	0.0690
18—21	雌	12	2	0.1667	0.8333	0.1255	0.8745	0.0084	0.0651
	雄	13	1	0.0769	0.9231	0.0966	0.9034	0.0027	0.0267
21—24	雌	10	2	0.2000	0.8000	0.1004	0.8996	0.0084	0.0727
	雄	12	2	0.1667	0.8333	0.0805	0.9195	0.0054	0.0606
24—27	雌	8	4	0.5000	0.5000	0.0502	0.9498	0.0167	0.1852
	雄	10	1	0.1000	0.9000	0.0724	0.9276	0.0027	0.0035
27—30	雌	4	3	0.7500	0.2500	0.0126	0.9874	0.0126	0.3333
	雄	9	6	0.6667	0.3333	0.0241	0.9759	0.0161	0.3373
30—33	雌	1	1	1.0000	0.0000	0.0000	1.0000	0.0042	0.5333
	雄	3	2	0.6667	0.3333	0.0080	0.9920	0.0054	0.3333
33—36	雄	1	1	1.0000	0.0000	0.0000	1.0000	0.0027	0.6667

1. 平均死亡率的计算: 雄性为0.283, 雌性为0.272, 两者接近^[2]。雄性 与 雌性 在 $t_{m_i} \in (t_{i-1}, t_i)$ 的瞬间死亡密度 $\hat{f}(t_{m_i})$ 也很接近(见表2, 图1), 9月龄前雄性略高, 12—27月龄雌性略高。死亡密度对田鼠种群死亡过程的动态分析具有重要意义。

2. 本文计算危险率函数 $\hat{\lambda}(t_{m_i})$ (表2) 及图2可以看出, 24月龄以后, 危险率函数值骤增, 这与参考文献[2]的结论一致。

3. 从整体死亡的过程来分析, 根田鼠在0—18月龄内, 死亡密度曲线波动较大, 以后趋向平缓, 这说明0—18月龄内根田鼠种群的生命状态不稳定, 死亡可能性较大。以后生命状态稳定。表3也证明了这一结论的正确性。因此, 加强对18月龄前的实验种群的饲养管理, 避免较大死亡有一定现实意义。

表 3 雌性根田鼠生命表主要方差估计
Table 3 Estimate of female root vole variate

月龄	$V_{\sigma^2}(\hat{S}_i) = V_{\sigma^2}(\hat{F}_i) \times 10^{-2}$	$V_{\sigma^2}(\hat{f}(t_{mi})) \times 10^{-3}$	$V_{\sigma^2}(\hat{\lambda}(t_{mi})) \times 10^{-1}$
0—3	0.1774	0.1970	0.0316
3—6	0.2224	0.1149	0.0032
6—9	0.4128	0.1367	0.0043
9—12	0.3624	0.2378	0.0081
12—15	0.2245	0.2503	0.0164
15—18	0.1676	0.0832	0.0581
18—21	0.1425	0.0346	0.0242
21—24	0.1164	0.0346	0.0251
24—27	0.0806	0.0871	0.0529
27—30	0.0157	0.0614	0.1645

参 考 文 献

- [1] 冯士雍, 1982, 生存分析(I), 数学的实践与认识, (3): 72—80.
 [2] 梁杰荣、孙儒泳, 1985, 根田鼠生命表和繁殖的研究, 动物学报 31(2): 170—176.
 [3] Cox D.R. and Oakes, D., 1984, Analysis of survival data. Chapman and hall, London New York
 [4] Muenchow, G., 1986, Ecological use of failure time analysis Ecology 67(1): 246—250.
 [5] Pyke, D.A and Thompson, J. N., 1988, Statistical analysis of Survival and removal rate experiments. Ecology 67(1): 240—245.
 [6] 蒋庆琅, 1984, 《生命表及其应用》(中译本), 上海翻译出版公司。

THE THEORY OF SURVIVAL ANALYSIS AND ITS APPLICATION TO LIFE TABLE

Yang Feng-Xiang Wang Shun-Qing Xu Hai-Gen Li Bang-Qing

(Department of Mathematics, Lanzhou University)

Life table, a research area with long history, has been frequently employed in bioscience. As a statistical branch dealing with survival data, survival analysis has been recently produced and is in rapid development, which makes life table theory perfect and enables it to reflect survival principles of creatures more clearly and creatively. In terms of statistical methods, Life table analysis belongs to the category of non-parametric method.

This paper based on the theory of survival function $S(t)$, accumulated mortality rates function $F(t)$, mortality density function $f(t)$, and hazard function $\lambda(t)$ and their estimates $\hat{S}(t)$, $\hat{F}(t)$, $\hat{f}(t)$, $\hat{\lambda}(t)$ which are included in survival analysis. Meanwhile, the estimate formulae of variance for these functions are presented in order to value the error of these functions, which enables it convenient to be applied by life table. Lastly, as an example, survival analysis is applied to the data in paper (2), the result to which is in accord with that of paper (2).

Key words: survival analysis.