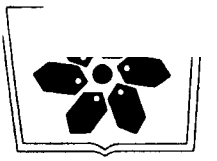
现代数学基础丛书 98

生存数据统计分析

王启华 著

科学出版社
www.sciencep.com



中国科学院科学出版基金资助出版

现代数学基础丛书 98

生存数据统计分析

王启华 著

科学出版社

北 京

内 容 简 介

本书主要系统介绍生存分布函数估计、概率密度估计、失效率估计、包含平均寿命作为特例的一类均值泛函估计及其统计性质，介绍与之相关的统计方法(如鞅重抽样方法、估计方程方法、点过程鞅方法、经验似然方法等)及有关的应用成果；介绍两样本检验及处理差异统计推断方法，介绍随机删失回归分析及比例风险回归统计推断方法、理论及应用.

本书适合作高等院校数学和统计专业的高年级大学生、研究生教材，也适合大学教师、科研人员以及应用工作者阅读参考.

图书在版编目(CIP)数据

生存数据统计分析/王启华著. —北京：科学出版社，2006
(现代数学基础丛书；98)
ISBN 7-03-016454-7

Ⅰ.生… Ⅱ.王… Ⅲ.生存率-统计分析(数学) Ⅳ.R195.3

中国版本图书馆 CIP 数据核字(2005) 第 131809 号

责任编辑：陈玉琢／责任校对：陈丽珠
责任印制：钱玉芬／封面设计：王 浩

科学出版社 出版
北京东黄城根北街16号
邮政编码：100717
http://www.sciencep.com

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销
*
2006 年 1 月第 一 版 开本：B5(720×1000)
2006 年 1 月第一次印刷 印张：18
印数：1—3 000 字数：336 000

定价：45.00 元

(如有印装质量问题，我社负责调换〈环伟〉)

《现代数学基础丛书》序

对于数学研究与培养青年数学人才而言，书籍与期刊起着特殊重要的作用．许多成就卓越的数学家在青年时代都曾钻研或参考过一些优秀书籍，从中汲取营养，获得教益．

20 世纪 70 年代后期，我国的数学研究与数学书刊的出版由于文化大革命的浩劫已经破坏与中断了十余年，而在这期间国际上数学研究却在迅猛地发展着．1978 年以后，我国青年学子重新获得了学习、钻研与深造的机会．当时他们的参考书籍大多还是 50 年代甚至更早期的著述．据此，科学出版社陆续推出了多套数学丛书，其中《纯粹数学与应用数学专著》丛书与《现代数学基础丛书》更为突出，前者出版约 40 卷，后者则逾 80 卷．它们质量甚高，影响颇大，对我国数学研究、交流与人才培养发挥了显著效用．

《现代数学基础丛书》的宗旨是面向大学数学专业的高年级学生、研究生以及青年学者，针对一些重要的数学领域与研究方向，作较系统的介绍．既注意该领域的基础知识，又反映其新发展，力求深入浅出，简明扼要，注重创新．

近年来，数学在各门科学、高新技术、经济、管理等方面取得了更加广泛与深入的应用，还形成了一些交叉学科．我们希望这套丛书的内容由基础数学拓展到应用数学、计算数学以及数学交叉学科的各个领域．

这套丛书得到了许多数学家长期的大力支持，编辑人员也为其付出了艰辛的劳动．它获得了广大读者的喜爱．我们诚挚地希望大家更加关心与支持它的发展，使它越办越好，为我国数学研究与教育水平的进一步提高作出贡献．

杨　乐

2003 年 8 月

前　言

生存分析是近几十年来发展起来的、对生存数据进行统计分析的一门学科. 近二三十年来生存分析受到国内外统计学家的关注, 研究异常活跃. 统计学家们提出了很多有实用价值的方法并建立了一套系统的理论, 然而这些理论成果大多都散布在文献中, 已有的有关著作要么定位于应用, 要么定位于理论, 所介绍的大多是十年之前的成果. 为了介绍一些重要的最新成果, 同时为了将方法、理论与应用有机地结合起来, 使广大读者对这一领域有比较系统的了解, 并从中得到一些科学研究的启发, 我们决定写这本书.

作者尽管希望能在这本书中做到将方法、理论与应用有机结合, 但本书仍偏于方法与理论的介绍. 这样做的主要原因是介绍生存分析的方法与应用的书相对较多, 而系统且详细介绍方法与理论的书并不多见; 另一个原因是学习方法与理论对希望进入这一领域的读者是至关重要的. 只有首先掌握必要的理论知识, 才有可能真正理解并进一步研究生存分析的方法, 并对其理论进行进一步探索. 而对于一个应用工作者来说, 系统的理论基础知识能帮助他们更好地掌握对生存分析方法的运用, 增强他们对生存分析的方法在实际应用中的灵活性, 帮助他们理解生存分析中一些统计思想的科学性.

这本书既不同于一般的教材, 也不同于一般的专著. 作者在写这本书时, 注重系统介绍生存分析的一些基本方法与理论, 对一些基本的重要的定理给出详细的证明, 并适当介绍方法应用, 使读者在系统掌握必要的基础知识的同时, 又能掌握必要的理论研究技术和方法应用上的技巧, 从而体现了一般教材的特点. 但这本书并不局限于介绍基本方法、理论及其应用, 我们还介绍了一些最新成果, 其中包括作者本人最近的一些工作, 从而体现专著的特性. 在这本书中, 我们还注重介绍一些问题研究的发展过程及发展方向, 从而便于从事科学研究的读者了解科学研究的发展规律. 我们也注意介绍一些相关成果的来源或出处, 希望能使这本书成为广大读者的一本科研指导书.

本书主要系统介绍生存分布函数估计、概率密度估计、失效率估计、包含平均寿命作为特例的一类均值泛函估计及其统计性质, 介绍与之相关的统计方法 (如鞅重抽样方法、估计方程方法、点过程鞅方法、经验似然方法等) 及有关的应用成果; 介绍两样本检验及处理差异统计推断方法、介绍随机删失回归分析及比例风险回归统计推断方法与理论. 尽管其中有一部分内容在其他书中已有介绍, 但在本书介绍这些内容仍是必要的, 因为这保证了本书对这一领域介绍的完整性和系统性. 应该指出的是第 6 章内容主要是 20 世纪 50~60 年代的成果, 是比较经典的内容. 作者在写这一章时, 没有直接追索这些成果的原文, 而是直接参考 Miller (1981), Lee

(1992), Anderson (1993) 及 Klein 与 Moeschberger (1997) 等著作.

据作者所知, 本书中的一些内容如有关概率密度估计、失效率估计、一类均值泛函估计、两样本处理差异推断及随机删失半参数回归等内容是已有书中尚未介绍的. 此外, 本书还介绍了一些新的处理随机删失数据的方法, 如缺重抽样方法、估计方程方法、经验似然方法及局部回归方法等. 既注意该领域的基础知识介绍, 又反映其最新的进展, 是作者在写这本书时试图体现的特色.

本书面向大学数学及统计专业的高年级学生、研究生、大学教师、科研人员以及广大的应用工作者.

本书中一些插图是我的博士生孙志华所完成的, 在此向她表示感谢! 由于作者水平有限, 书中一定还有不少的谬误, 恳请同行及广大读者批评指正.

王启华

2005 年 4 月

于中国科学院数学与系统科学研究院

目　录

引　言

生存分析是近二三十年发展起来的数理统计新分支, 它是根据医学、生命科学、可靠性工程、保险等科学研究中的大量实际问题提出的, 它可以广义地认为是对生存时间 (非负随机变量) 的一类统计分析技术, 主要研究随机删失数据的统计分析. 随机删失是生命科学、医药追踪研究、可靠性寿命试验及其他一些实际问题中常常碰到的一种重要类型的统计数据. 其理论与方法不仅能应用于生命科学、医药卫生、可靠性工程, 而且在保险数学、犯罪学、社会学、市场学、环境科学、航空航天等高科技领域都有广泛的应用前景.

生存分析的起源可能归结于几个世纪之前对死亡表的研究及半个世纪前开始的工程研究. 二战引起了人们对武器可靠性的兴趣, 且这一兴趣一直持续到战后乃至今天的武器及商业产品上. 以前大部分关于工程应用方面的统计研究主要集中在参数模型. 而在过去的 20 年中, 医学研究中的临床试验快速发展, 使生存统计分析方法研究的重点转移到非参数. 本书将着重非参数的介绍, 特别是近一二十年来这一领域的新进展. 以下我们介绍一些有关的概念, 这可能有利于读者在阅读这本书之前对生存分析有一个大致的了解.

1. 生存时间

生存时间可以广泛地定义为一给定的事件发生的时间. 这个事件可以是疾病的发生、一种处理 (治疗) 的反应、病情复发或死亡. 因此, 生存时间可以是无肿瘤时间, 从一种治疗开始到有反应的时间, 缓解时间长度或出现死亡的时间. 生存数据可以包括生存时间、对治疗的反应以及与反应、生存及疾病发生有关的病人特征. 生存数据不仅出现在生物医学中, 而且出现在工业可靠性、社会科学和商业研究中. 在这些领域生存数据的例子是: 可靠性工程中电子设备 (元件或系统) 的寿命, 犯罪学中重罪犯人的假释时间, 社会学中首次婚姻的持续时间, 它也可以不是时间, 它可以是汽车车轮转动的圈数, 也可以是市场学中报纸或杂志的篇幅和订费, 甚至可能是保险公司在某一索赔案中所付的保险费等.

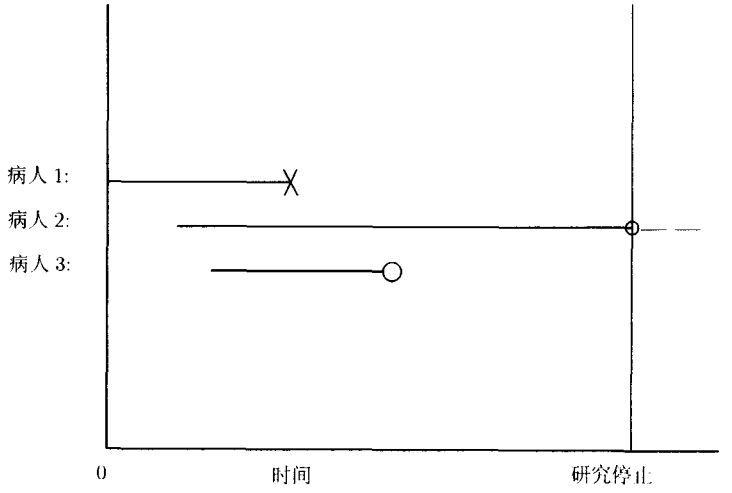
2. 随机删失数据

设 $T_1, T_2, \cdots, T_n$ 是非负独立同分布表示寿命的随机变量, 其分布函数为 F; C_1, $C_2, \cdots, C_n$ 是非负独立同分布表示删失的随机变量, 具有分布函数 G. 在随机右删失模型中, 我们不能完全观察到 T_i, 而仅能观察到

$$X_i = \min(T_i, C_i), \qquad \delta_i = I[T_i \leqslant C_i], \quad i = 1, 2, \cdots, n,$$

其中 $I[\cdot]$ 表示某事件的示性函数. 显然, δ 包含了删失信息.

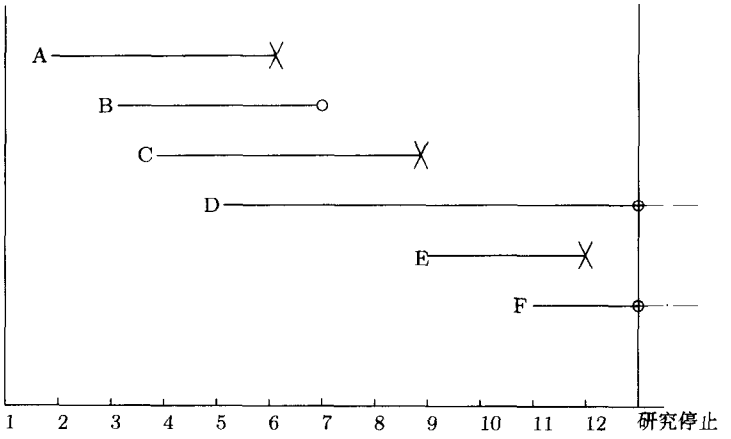
下图表示随机删失数据的例子：



这里, 第一个病人在 $t=0$ 时刻进入研究, 但在研究停止前某时刻死亡; 第二个病人在研究开始后某时刻进入研究, 但在研究结束时, 该病人仍然活着, 因而产生一个删失观察; 第三个病人在研究开始后进入研究, 但在研究结束之前退出试验, 因而产生另一个删失观察.

这种删失通常发生在医学研究及临床试验中. 在大部分临床研究试验中, 研究时间通常是固定的, 且病人通常在这段时间内的不同时刻进入研究. 一些人在研究结束前可能死亡, 这部分人生存时间已知; 其他人可能在研究结束之前退出试验或失去跟踪, 或在研究结束时仍然活着. 对中途退出或失去跟踪的病人生存时间至少是从进入研究到失去联系这段时间; 对仍然活着的病人, 生存时间至少是从进入研究到研究结束这段时间, 这两种观察就是删失观察. 既然进入试验时间可能不同, 因而被删失的时间也可能不同. 例如, 假设 6 个有急性白血病的病人, 先后分别进入研究时间为 1 年临床试验, 假设 6 个病人获得治疗并得到缓解. 缓解时间见下图, 病人 A, C, E 获得缓解的时间分别是二月、四月和九月, 而复发时间分别是 4 和 5 及 3 个月以后. 病人 B 在第三个月的开始获得缓解, 但 4 个月以后失去跟踪, 对 B 缓解时间至少是 4 个月. 病人 D 和 F 分别在第五、第十一个月开始获得缓解, 且在研究结束时没有复发, 因此他们的缓解时间因此至少是 8 个月和 3 个月. 这 6 个病人的缓解时间分别是 4, 4+, 5, 8+, 3 及 2+ 个月, 其中带 "+" 号数据表示删失数据.

应当指出这里所介绍的随机删失只是随机右删失, 其他随机删失还有左删失、双向删失及区间删失等, 但在此不做一一介绍, 因为本书主要集中在随机右删失方面的内容.



3. 生存时间的函数

生存时间的特征通常用下面 3 个函数来刻画：(1) 生存分布函数; (2) 概率密度函数; (3) 失效率函数. 应当指出：这 3 个函数在数学上是等价的. 设非负随机变量 T 有密度 $f(t)$ 和分布函数 $F(t)$, 则生存函数 $\bar{F}(t)$ 定义为 $\bar{F}(t)=1-F(t)$, 失效率为 $\lambda(t)=f(t)/(1-F(t))$. 本书将探讨这些生存时间函数的估计及其统计性质.

4. 随机删失回归

随机删失回归主要包含随机删失线性回归、非参数回归、半参数回归及 Cox-回归等. 现以线性回归为例说明. 设有线性模型 $Y_i=X_i^{\mathrm{T}}\beta+\epsilon_i(i=1,2,\cdots,n)$, 其中 β 是参数向量, ϵ_i 是均值为 0 且方差有限的随机误差变量, X_i 是协变量向量, Y_i 是反映变量 $(i=1,2,\cdots,n)$. 在完全观察下, 最小二乘法可用于估计参数 β. 然而在随机删失下, Y_i(通常是某寿命随机变量 T_i 的对数) 可能被另一随机变量 C_i 所删失, 使 Y_i 不能被完全观察, 而观察到的数据是 (Z_i,δ_i,X_i), 其中 $Z_i=\min(Y_i,C_i),\delta_i=I[Y_i\leqslant C_i],i=1,2,\cdots,n$. 这种反映变量被随机删失的回归就称为随机删失回归. 显然, 对这种随机删失回归, 标准方法比如线性回归分析中的最小二乘法等, 就不能直接应用, 于是如何利用随机删失数据对回归模型进行统计分析, 正是随机删失回归分析所要探讨的内容.

5. 本书的范围

本书共分 8 章. 第 1 章首先介绍如何利用随机删失数据构造生存分布函数的 Kaplan-Meier 乘积限估计, 然后再介绍它的渐近性质, 包括相容性、强相合性及其收敛速度、鞅方法、鞅表示、渐近表示、弱收敛与强逼近定理, 介绍 Edgeworth 展开与 bootstrap 方法及 bootstrap 逼近方面的结果; 第 2 章介绍概率密度估计及其渐近性质, 包括概率密度估计的强相合性、渐近正态性、均方收敛速度、一些矩不

等式、光滑 bootstrap 方法及窗宽选择等; 第 3 章介绍各种形式的失效率估计及其渐近性质, 包括渐近正态性, 强相合收敛速度, 一些不等式及窗宽选择等; 第 4 章介绍平均寿命、一类均值泛函估计及其统计性质, 介绍点过程鞅方法及经验似然方法; 第 5 章介绍对照差在不同模型下的估计方法及估计的渐近性质; 第 6 章从生存分布与失效率的角度介绍比较两个总体及多个总体的检验方法; 第 7 章介绍回归模型及相关的回归分析方法, 介绍线性、非线性及半参数部分线性回归的估计方法及统计推断的理论; 第 8 章介绍比例风险回归, 介绍估计方法及估计的渐近性质, 介绍模型检验及点过程鞅方法在比例风险模型研究中的应用等.

第 1 章　生存分布函数估计

Kaplan 与 Meier(1958) 在随机删失下提出了生存函数的一种所谓的乘积限估计, 它在生存分析中的地位相当于完全观察下的经验分布函数, 然而它的构造及其统计特性的研究要比经验分布函数复杂得多. 近二三十年来人们致力于这一估计的研究, 获得很多重要的成果. 本章只介绍其中最重要和最有代表性的内容. 下面在给出这一估计之前, 先介绍生存分布函数.

§1.1　生存分布函数

设 T 表示生存时间, $F(t)=P(T\leqslant t)$ 表示 T 的分布函数, 则 $\bar{F}(t)=1-F(t)$ 定义 T 的生存分布函数, 它实际上是个体生存时间长于 t 的概率. 易知, $\bar{F}(t)$ 是非增函数, 且 $\bar{F}(0)=1, \bar{F}(+\infty)=0$.

函数 $\bar{F}(t)$ 也叫累积生存率, 它的图形叫做生存曲线. 陡峭的生存曲线表示低的生存概率, 见图 1-1-1(a); 较平坦的曲线表示高的生存概率, 见图 1-1-1(b).

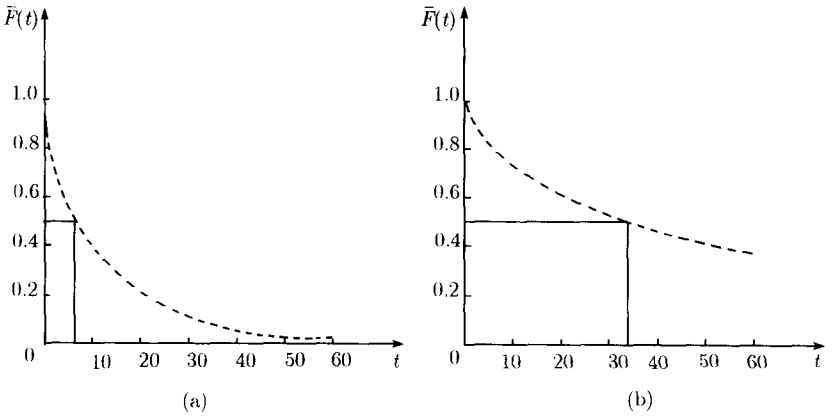


图 1-1-1　两个生存曲线的例子

在实践中, 如果数据被完全观察, 生存函数可用生存时间长于 t 者所占的比例来估计:

$$\hat{\bar{F}}(t)=\frac{\text{生存时间长于 } t \text{ 的病人数}}{\text{病人总数}}, \tag{1.1.1}$$

这里 $\hat{\bar{F}}(t)$ 表示 $\bar{F}(t)$ 的估计. 当数据有删失时, 式 (1.1.1) 的分子一般不能确定. 例如考虑下面生存数据: 4, 6, 6+, 10+, 15, 20, 其中带 "+" 号的数据表示是删失数据. 利用式 (1.1.1) 可得 $\hat{\bar{F}}(5) = 5/6 = 0.833$, 但不能得到 $\hat{\bar{F}}(11)$, 因为生存时间长于 11 的病人是不知道的, 第三个病人或第四个病人的生存时间可能长于 11 也可能小于 11. 因此, 一旦有删失数据, 用式 (1.1.1) 估计 $\bar{F}(t)$ 是不合适的. 于是在随机删失下构造 $\bar{F}(t)$ 的合适估计是本章的内容.

构造 $\bar{F}(t)$ 主要有两种方法. 第一种方法是生命表分析法, 这种方法适合于样本量很大 (例如数以千计) 或数据是按区间分组等情形; 第二种方法是 Kaplan 与 Meier(1958) 所提出的估计生存函数的乘积限方法. 由于计算机使用越来越广泛, 这个方法可用于小样本、中样本及大样本等各种情形. 生命表估计与乘积限估计实质上是一样的. 很多作者也把乘积限估计称作寿命表估计, 二者的差别是: 乘积限估计是基于一个一个的数, 而寿命表估计是基于按区间的分组数据, 因而乘积限估计是寿命表估计在各个区间只含一个观察值时的一种特殊情形. 由于人们主要致力于乘积限估计的研究, 并获得一系列丰富的成果, 因而在此只介绍乘积限估计方法.

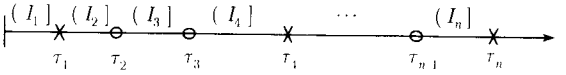
§1.2　估计的定义与计算

设 $T_1, T_2, \cdots, T_n$ 是非负独立同分布表示寿命的随机变量, 其分布函数为 F; $C_1, C_2, \cdots, C_n$ 是非负独立同分布表示删失的随机变量, 具有分布函数 G. 在随机右删失模型中, 我们不能完全观察 T_i, 而仅能观察到

$$X_i = \min(T_i, C_i), \qquad \delta_i = I[T_i \leqslant C_i], \quad i = 1, 2, \cdots, n,$$

显然 X_i 有分布函数 $H(t) = P(X \leqslant t) = 1 - (1 - F(t))(1 - G(t))$. Kaplan 与 Meier (1958) 针对这一随机删失数据提出了生存分布 $\bar{F}(t)$ 的乘积限估计. 下面我们介绍乘积限估计的构造.

我们观察到的数据对是 $(X_1, \delta_1), (X_2, \delta_2), \cdots, (X_n, \delta_n)$, 假设没有"结", 设 $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ 是 $X_1, X_2, \cdots, X_n$ 的次序统计量, 设时间被分割成 n 个区间, 每个区间 I_i 的长度是变量, I_i 区间的右端点 $\tau_i = X_{(i)}$, 如下图:



"o" 表示删失, " × " 表示非删失

设 $\delta_{(i)}$ 是对应于 $X_{(i)}$ 的 δ 值, 即当 $X_{(i)}=X_j$ 时, $\delta_{(i)}=\delta_j$. 设 $\mathcal{R}(t)$ 记在时间 t 的风险集, 即在时刻 t 仍然活着的个体数, 且设

$n_i=\mathcal{R}(X_{(i)})$ 中的个体数,

$d_i=$ 在时刻 $X_{(i)}$ 死亡数,

$p_i=P($活过 $I_i|$ 在I_i的开始活着$)=P(T>\tau_i|T>\tau_{i-1})$,

$q_i=1-p_i$.

在观察没有"结"时, $d_i=1$ 或 0, 由 q_i 与 p_i 的估计

$$\widehat{q}_i=\frac{d_i}{n_i},\quad \widehat{p}_i=1-\widehat{q}_i=\begin{cases}1-\dfrac{1}{n_i}, & \text{若}\quad \delta_{(i)}=1,\\ 1, & \text{若}\quad \delta_{(i)}=0.\end{cases}$$

Kaplan 与 Meier (1958) 所定义的乘积限估计是

$$\widehat{\bar{F}}_n(t)\equiv 1-\widehat{F}_n(t)=\prod_{X_{(i)}\leqslant t}\widehat{p}_i=\prod_{X_{(i)}\leqslant t}\left(1-\frac{1}{n_i}\right)^{\delta_{(i)}}=\prod_{X_{(i)}\leqslant t}\left(1-\frac{1}{n-i+1}\right)^{\delta_{(i)}}. \tag{1.2.1}$$

下面我们用一个例子来帮助理解这一估计. 为方便, 以下有时用 $\widehat{\bar{F}}$ 记 $\widehat{\bar{F}}_n$.

假设有 10 个病人在 1988 年初进入某临床研究, 在 1988 年内有 6 人死亡而 4 人活着. 在这年末又有 20 个病人进入研究. 在 1989 年, 首批进入研究的有 3 人死亡, 还有 1 人活着; 后进入研究的有 15 人死亡, 还有 5 人活着. 假设研究工作于 1989 年底结束, 要求估计生存两年以上的病人所占的比例. 此例中的第一组病人有两年观察时间, 第二组病人只有 1 年观察时间. 一种可能的估计是 $\widehat{\bar{F}}(2)=1/10=0.1$. 这个估计忽略了 20 个病人只观察了 1 年的事实. Kaplan 与 Meier 认为, 第 2 个样本对于估计 $\bar{F}(2)$ 也有作用.

活了两年的病人都可以看做第 1 年活着然后又活了 1 年. 于是

$$\begin{aligned}\widehat{\bar{F}}(2)&=P(\text{第 1 年活着然后再活了 1 年})\\ &=P(\text{病人活了 1 年的条件下活了 2 年})\times P(\text{第 1 年活着}).\end{aligned} \tag{1.2.2}$$

基于式 (1.2.2), 按 Kaplan 与 Meier 的思想, $\bar{F}(2)$ 的估计应该定义如下:

$$\widehat{\bar{F}}(2)=\frac{\text{活了两年的病人数}}{\text{第 1 年活着的病人数}}\times\frac{\text{活过 1 年的病人数}}{\text{占总病人数}}. \tag{1.2.3}$$

对上面所给的数据, 4 个活了 1 年的病人中有 1 人活过两年, 因而式 (1.2.3) 右边第一个比例是 1/4; 10 个从 1988 年年初进入的病人中有 4 人活过了 1 年, 20 个