北京大学数学教学系列丛书

研究生
数学基础课教材

生存分析与可靠性

陈家鼎　编著

北京大学出版社

北京大学数学教学系列丛书

生存分析与可靠性

陈家鼎 编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

生存分析与可靠性/陈家鼎编著. —北京:北京大学出版社,2005.11
(北京大学数学教学系列丛书)
ISBN 7-301-05372-X

Ⅰ.生…　Ⅱ.陈…　Ⅲ.生存率-统计分析(数学)-高等学校-教材
Ⅳ.R195.3

中国版本图书馆 CIP 数据核字(2005)第 107691 号

书　　名:生存分析与可靠性
著作责任者:陈家鼎　编著
责 任 编 辑:刘　勇
标 准 书 号:ISBN 7-301-05372-X/O·0526
出 版 发 行:北京大学出版社
地　　址:北京市海淀区成府路 205 号　100871
网　　址:http://cbs.pku.edu.cn
电　　话:邮购部 62752015　发行部 62750672　理科编辑部 62752021
电 子 信 箱:zpup@pup.pku.edu.cn
排　版　者:北京高新特打字服务社　82350640
印　刷　者:北京大学印刷厂
经　销　者:新华书店
890 毫米×1240 毫米　A5　13.875 印张　410 千字
2005 年 11 月第 1 版　2005 年 11 月第 1 次印刷
印　　数:0001—4000 册
定　　价:22.00 元

序　言

自 1995 年以来，在姜伯驹院士的主持下，北京大学数学科学学院根据国际数学发展的要求和北京大学数学教育的实际，创造性地贯彻教育部“加强基础，淡化专业，因材施教，分流培养”的办学方针，全面发挥我院学科门类齐全和师资力量雄厚的综合优势，在培养模式的转变、教学计划的修订、教学内容与方法的革新，以及教材建设等方面进行了全方位、大力度的改革，取得了显著的成效。2001 年，北京大学数学科学学院的这项改革成果荣获全国教学成果特等奖，在国内外产生很大反响。

在本科教育改革方面，我们按照加强基础、淡化专业的要求，对教学各主要环节进行了调整，使数学科学学院的全体学生在数学分析、高等代数、几何学、计算机等主干基础课程上，接受学时充分、强度足够的严格训练；在对学生分流培养阶段，我们在课程内容上坚决贯彻“少而精”的原则，大力压缩后续课程中多年逐步形成的过窄、过深和过繁的教学内容，为新的培养方向、实践性教学环节，以及为培养学生的创新能力所进行的基础科研训练争取到了必要的学时和空间。这样既使学生打下宽广、坚实的基础，又充分照顾到每个人的不同特长、爱好和发展取向。与上述改革相适应，积极而慎重地进行教学计划的修订，适当压缩常微、复变、偏微、实变、微分几何、抽象代数、泛函分析等后续课程的周学时。并增加了数学模型和计算机的相关课程，使学生有更大的选课余地。

在研究生教育中，在注重专题课程的同时，我们制定了 30 多门研究生普选基础课程（其中数学系 18 门），重点拓宽学生的专业基础和加强学生对数学整体发展及最新进展的了解。

教材建设是教学成果的一个重要体现。与修订的教学计划相配合，我们进行了有组织的教材建设，计划自 1999 年起用 8 年的时间

修订、编写和出版40余种教材，这就是将陆续呈现在大家面前的《北京大学数学教学系列丛书》。这套丛书凝聚了我们近十年在人才培养方面的思考，记录了我们教学实践的足迹，体现了我们教学改革的成果，反映了我们对新世纪人才培养的理念，代表了我们新时期的数学教学水平。

经过20世纪的空前发展，数学的基本理论更加深入和完善，而计算机技术的发展使得数学的应用更加直接和广泛，而且活跃于生产第一线，促进着技术和经济的发展，所有这些都正在改变着人们对数学的传统认识。同时也促使数学研究的方式发生巨大变化。作为整个科学技术基础的数学，正突破传统的范围而向人类一切知识领域渗透。作为一种文化，数学科学已成为推动人类文明进化、知识创新的重要因素，将更深刻地改变着客观现实的面貌和人们对世界的认识。数学素质已成为今天培养高层次创新人才的重要基础。数学的理论和应用的巨大发展必然引起数学教育的深刻变革。我们现在的改革还是初步的。教学改革无禁区，但要十分稳重和积极；人才培养无止境，既要遵循基本规律，更要不断创新。我们现在推出这套丛书，目的是向大家学习。让我们大家携起手来，为提高中国数学教育水平和建设世界一流数学强国而共同努力。

张继平

2002年5月18日

于北京大学蓝旗营

前　言

生物和人的生存时间的评估与预测是生物学和医学的重要研究任务，产品(或系统)的使用寿命的评定与预测是现代可靠性理论的主要研究对象.这两方面的研究都涉及基础科学和技术科学的许多领域，其中数学方法，特别是统计方法，起着重要作用.从数学角度来看，不管是生物学、医学领域还是工程技术领域，都是对一个或多个非负随机变量(生存时间或寿命)进行统计分析和推断.这种统计分析和推断已形成现代数理统计学的重要分支——生存分析.

生存分析就是根据试验或调查得到的数据，对生物和人的生存时间或产品的寿命进行分析和推断.和其他的数学方法一样，生存分析的方法和理论有广泛的应用，不限于生物学、医学和工程科学，而且可应用于社会学、心理学、经济学、保险精算学等等."生存时间"或"寿命"应作广义的理解，乃是指自然界、人类社会或技术过程中某种状态的持续时间.生存分析含有许多实用的方法和丰富的理论.随着医疗实践和工程实践及其他领域的推动，不断有新的统计方法出现，应用范围越来越广.

本书作者多年来为北京大学概率统计系的研究生开设"生存分析与可靠性"课程.本书是在十年前作者编写的教材《生存分析与可靠性引论》的基础上进行改写与扩充而成的，与前书有很大不同.本书在写作过程中注意到下列各点.

(1) 对生存分析与可靠性的基本概念和主要内容进行系统论述.在叙述概念和统计方法时力求数学上严格清楚，对数学定理都给出准确的陈述，但限于篇幅只对大部分定理写出了严格的证明，还有一些定理未给出证明.

(2) 鞅方法是研究生存分析的重要手段.本书叙述了鞅方法的入门知识，重点叙述了鞅方法如何用于研究 PL 估计的统计性质，但限于篇幅，没有介绍鞅的中心极限定理及其在大样本研究中的应用.

(3) 对不完全数据情形的最大似然估计进行系统论述.对最大似

然估计的存在性和强相合性给予很大关注，但限于篇幅未讨论这种估计的渐近分布（个别情形除外）.

（4）指数分布在质量与可靠性研究中居重要地位，有关的试验方案与数据类型多种多样，国际标准化组织（ISO）和我国颁布的一些质量与可靠性标准也是针对指数分布的，因此，特专列一章论述指数分布情形的多种寿命试验方案与数据处理方法.

（5）努力概括近十年来在生存分析与可靠性领域的某些重要进展，重点介绍了北京大学概率统计系在可靠性领域的一些主要研究成果，但限于篇幅，对于生存分析与医学、生物学相关的某些重要研究方向（如传染病模型、捕获与再捕获模型）未能介绍.

（6）注意突出重点. 某些非基本的内容或过于冗长的推理用小字排印或加 * 号标示，但有些非小字排印的内容在课堂上也可不讲，留给学生课外阅读（例如最大似然估计的可测性）.

作者相信，只要读者学过高等数学和初等概率统计，就不难阅读和掌握本书中所叙述的生存分析与可靠性的概念和方法. 若要学习和掌握所有定理的证明，则要求读者预先具备测度论和现代概率论的基本知识，并受过相应的数学训练. 关于 Lebesgue-Stieltjes 积分（简称 L-S 积分）的补充知识已写在附录里，这是一般测度论书中未讲而在生存分析研究中要用到的知识.

本书是概率统计专业和应用数学专业的研究生教材，也可供有关专业的高年级大学生和实际工作部门的研究人员阅读.

本书初稿承蒙北京大学郑忠国教授的仔细审阅，他指出了许多不当之处并提出了宝贵的改进意见，修改稿采纳了他的意见，这对提高本书的质量起了重要作用，笔者在此向郑忠国教授表示深深的谢意. 房祥忠同志也对初稿提出了许多宝贵意见，我在修改稿中也都予以采纳；北京大学出版社的刘勇同志指出了原稿中许多不妥或不规范之处，并花很大精力帮助作者进行修改，在此也向他们二位表示感谢.

由于作者水平有限，书中存在的缺点和谬误一定不少，欢迎读者和专家批评指出.

陈家鼎

2004 年 9 月于北京大学承泽园

目　录

第一章　生存分析与可靠性的基本概念

§1　生存分析与有关的统计问题

什么是生存分析？简单地说，生存分析(survival analysis)就是对一个或多个非负随机变量进行统计分析，即根据观测到的数据对一个或多个非负随机变量进行统计推断．非负随机变量常用来表示自然界、人类社会或技术过程中某种状态的持续时间．一种最常见的情况是，用非负随机变量表示“寿命”(技术产品的寿命或生物、人的寿命)，因而生存分析可看成是对寿命进行研究，是对寿命数据进行分析．生存分析对产品寿命的评估、人和生物寿命的研究、手术后人的寿命的预测等等都十分重要，因而生存分析的理论和方法在工程上及医学、生物学上都有广泛的应用价值，日益受到人们的重视．读者还应注意，虽然本书常用“寿命”一词，但对“寿命”应作广义的理解，是指某种状态的持续时间，因而所述的理论和方法适用范围很广．

对于产品来说，我们总是希望它质量可靠，使用寿命长．这里“产品”二字作广义的理解，可以是元件、部件或整机、系统．(工程上对产品有细致的划分，见何国伟的著作[14]．)什么是产品的寿命？一件产品从开始使用它的时刻算起，到它损坏(或规定的功能丧失)不能使用需要予以修理或换新的时刻为止，所经历的总“时间”就是它的寿命．例如，这只灯泡用了 1250 小时就坏了，则该灯泡的寿命为 1250 小时．要注意的是，寿命的单位不一定是时间，也可以是其他的度量单位．如轮胎的寿命用里程，跑 20 万千米坏了就算寿命是 20 万千米．寿命的单位也可以是次数，例如金属软管就用它能承受的脉冲次数来度量其寿命．

在工程上，与寿命有关的重要概念是可靠性．什么是可靠性呢？

定义 1.1　产品在规定的条件下和规定的时间内完成规定功能的概率叫做产品的**可靠度**．

若将这个定义里的“概率”二字改为抽象的能力，则这个定义就变为产品的“**可靠性**”(reliability)的定义. 应该指出，这里所说的“规定条件”包括环境条件，即所有内部与外部条件，如温度、湿度、辐射、磁场、电场、冲击、振动或其组合等.

从定义看出，产品的可靠度与时间因素有关(当然，在有些问题里时间因素不重要，可以略去)，因而与产品的寿命有关. 当然，产品的寿命也是相对于一定的使用条件而言的，是指在一定的使用条件下产品具有规定功能的“持续时间”. 若用 T 表示产品在规定条件下的寿命，t_0 是给定的正数，若我们希望产品在时间区间$[0,t_0]$上具有规定的功能，则概率 $P(T>t_0)$便是产品的可靠度，t_0 就是所谓的“任务时间”(mission time).

在实际工作中常常遇到所谓寿命统计问题. 例如某工厂生产了一批轴承，试验了若干个产品，记录了它们的寿命. 于是就要问：这批产品的寿命怎么样？对于这批轴承，能否有把握说它的寿命至少多长？满足要求的产品占的百分比是多少？等等. 这些典型的寿命检验问题就是生存分析里要研究的.

在医疗实践中常常需要鉴定药品的疗效. 如安眠药，就需要了解一些失眠患者服用某种安眠药后的睡眠延长时间，从而推断该种安眠药的医用价值. 又如严重的心脏病患者在决定是否愿意接受手术治疗时，他们首先关心的是：所在医院以前对类似的病人施加手术后，那些病人还能活多久？这就需要对历史数据进行分析. 这些也是生存分析里要研究的.

生存分析不是孤立地研究某个个体的寿命，而是研究一批个体的寿命. 任何个体的寿命多长带有偶然性，而一批个体的寿命多长就有一定的规律性. 我们用 T 表示任何个体的寿命，把 T 看成随机变量，T 的值依赖于个体.

怎样刻画非负随机变量 T 的特性呢？用分布函数 $F(t)=P(T\leqslant t)$ 或生存函数 $S(t)=P(T>t)$.

显然 $S(t)=1-F(t)$，它又叫做残存函数. 在工程上常用记号 $R(t)$ 代替 $S(t)$，称之为可靠性函数. 若 t_0 是任务时间(即完成任务所需要的时间)，则 $R(t_0)$就是可靠度.

在实际工作中遇到的"寿命"多是连续型的，即 T 有分布密度.

刻画 T 的特征数主要有：

平均寿命 μ ($\mu=ET$，为 T 的数学期望)；

中位寿命 m ($m=T$ 的中位数)；

寿命方差 σ^2 ($\sigma^2=T$ 的方差).

除了分布函数与生存函数外，另一个重要概念是危险率(失效率、故障率)：

$$\lambda(t)=\lim_{\Delta t\to 0^+}\frac{1}{\Delta t}P(T\leqslant t+\Delta t\,|\,T>t)\quad(\text{当极限存在时}).$$

易知，当 T 的分布密度 $f(t)$ 存在且右连续时，有下列计算公式：

$$\lambda(t)=\frac{f(t)}{1-F(t)}=\frac{f(t)}{S(t)}\quad(t\geqslant 0).\tag{1.1}$$

显然 $\lambda(t)=-S'(t)/S(t)$，于是有

$$S(t)=\exp\left\{-\int_0^t\lambda(u)\mathrm{d}u\right\}.\tag{1.2}$$

如以人的"寿命"为例，危险率 $\lambda(t)$ 大致分成三段. 从出生到青年是第一段，从青年到老年是第二段，从老年到死亡是第三段. 在头一段时间内，随着身体的发育成长，抵抗疾病的能力逐渐增长，函数 $\lambda(t)$ 是下降的；在第二段里身体发育基本完善了，是人一生中精力最充沛的时期，这时 $\lambda(t)$ 可看成常数；到了老年，人的各种机能衰退了，$\lambda(t)$ 是增函数. 因此，整个人生的危险率 $\lambda(t)$ 大致如图 1-1(呈浴盆形).

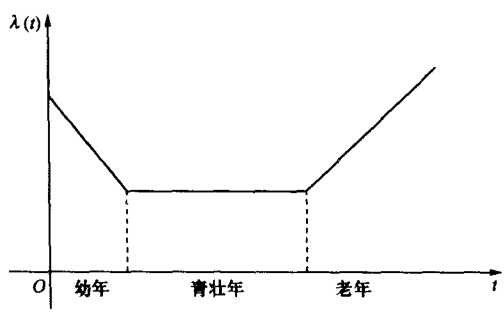

图 1-1

对于各种不同的产品，$\lambda(t)$ 的特点是不一样的，最重要且比较典型

的有三大类.

(1) 会老化的产品,此时 $\lambda(t)$是增函数.

(2) 会新化的产品,此时 $\lambda(t)$是减函数. 这种情况比较少见,一般出现在产品的早期失效期.

(3) $\lambda(t)$恒等于常数,这相当于产品处于随机失效期.

我们指出,在情形(3),寿命分布非常特殊,即所谓指数分布. 实际上,若 $\lambda(t)\equiv\lambda$,从(1.2)式知 $S(t)=e^{-\lambda t}$.

生存分析的头一个基本问题是: 如何根据数据来恰当地估计出 $S(t)$或 $\lambda(t)$?

这就需要对生存分析中碰到的数据类型有认识. 寿命数据有时是有意识地安排试验获得的,有时则是通过现场调查得到的. 可以说,数据一般含有**删失**(censoring)或不精密的特点.

什么是"删失"呢?删失分为"右删失"和"左删失". 在进行观测或调查时,一个个体的确切寿命不知道,但只知寿命大于 L,则称该个体的寿命在 L 是**右删失的**,并说 L 是**右删失数据**;若个体的确切寿命不知道,只知寿命小于 L,则称该个体的寿命在 L 是**左删失的**,并说 L 是**左删失数据**. 常用记号 L^+表示 L 是右删失数据,L^-表示 L 是左删失数据. 右删失的情形在寿命观测中极为常见,左删失的情形出现很少①.

在工程上和医学上还有一种情形是事先规定试验或观测的截止时间 L. 有的个体在试验或观测截止时寿命并未终结,这时称该个体的寿命在 L 被截尾,我们也把这种情况归于"右删失".

什么叫做"不精密"呢? 常常是个体的确切寿命 t 不知道,只知其在 $t^{(1)}$与 $t^{(2)}$之间(即 $t^{(1)}\leqslant t\leqslant t^{(2)}$),这时称$[t^{(1)},t^{(2)}]$是个体寿命的区间型数据. 实际工作中凡是不能连续监测的情况,通常只能得到这种类型的数据. 区间型数据$[t^{(1)},t^{(2)}]$的含义是: 真实寿命属于$[t^{(1)},t^{(2)}]$. 通常假定 $0<t^{(1)}<t^{(2)}<\infty$.

综上所述,在生存分析里碰到的数据有四种类型: 确切寿命数据(又叫寿终数据),右删失数据,左删失数据,区间型数据.

---

① 对"左删失"有不同的定义,有时采用下列定义: 若个体的确切寿命不知道,只知其不超过 L,则称 L 是左删失数据,也记作 L^-. 这种定义在理论研究中有好处.

值得注意的是：确切寿命数据可看成是区间型数据的极端情形(当区间长度变为零时)，右删失数据和左删失数据可看成广义区间型数据(前者是(L,∞)型，后者是$[0,L)$型数据). 这种观点突出了区间型数据的地位，在有些情况下对于处理问题很有好处.

一般说来，对 n 个个体的寿命进行观测(或调查，下同)，得到的数据可表示如下.

(1) 寿终数据(又叫完全寿命数据)：$t_1, t_2\cdots, t_{n_1}$;

(2) 右删失数据：$t^+_{n_1+1}, \cdots, t^+_{n_1+n_2}$;

(3) 左删失数据：$t^-_{n_1+n_2+1}, \cdots, t^-_{n_1+n_2+n_3}$;

(4) 区间型数据：$[t^{(1)}_{n_1+n_2+n_3+i}, t^{(2)}_{n_1+n_2+n_3+i}]$ $(i=1,2,\cdots,n_4)$,

这里 $n_1+n_2+n_3+n_4=n, 0\leqslant n_i\leqslant n, i=1,2,3,4$.

上面是最一般的数据类型①. 怎样分析这些数据呢？由于删失的引入，情况大为复杂. 普通的统计学未讨论这些，它只讨论每个数据都是完全寿命数据的情形(相当于 $n_1=n, n_2=n_3=n_4=0$ 的情形). 生存分析的一大特点，就是讨论含有删失(或区间型)数据的情形，因而发展出许多新的统计方法，形成许多新的理论.

怎样分析这些含有删失的数据呢？这就需要对"删失机制"有足够的认识. 只看见数据，不知道隐藏在数据背后的"删失机制"，不可能对数据有深刻的认识，无法制定出有充分根据的数据处理方法，从而不能对生存函数 $S(t)$(或其他有关量)作出科学的估计和判断.

怎样认识"删失机制"呢？这需要对有关的自然现象或技术过程的特性进行分析. 对于大多数实际工作者来说，在他们的专业范围内所从事的寿命试验(或观测)中，何种"删失机制"在起作用是一清二楚的. 正是在对实际工作中认可的"删失机制"作出的种种假定之下，才建立起

① 还有一类数据，即所谓截断(truncation)型的. 其来源比较特殊. 典型情形是：设 B 是直线上的集合，当随机变量 T 取值属于 B 时，T 的值可观测到；当 T 取值不属于 B 时，什么数据也观测不到. 例如用天文望远镜进行观测时，亮度超过某个值 L 的恒星可观测到，而亮度低于 L 的恒星则观测不到. 要注意的是，在工程上进行的截尾寿命试验里获得的数据却是删失型的. 要看数据的实质，日常生活语言中的"截断数据"多是"右删失数据". 真正的截断型数据比较少见，但在天文学、经济学和医学里有时会遇到. 限于篇幅，本书只讨论含有删失的数据.