

21
世纪统计学系列教材

彭王 非伟 编著

生存分析



中国人民大学出版社

21 世纪统计学系列教材

生存分析

彭 非 王 伟 编著

中国人民大学出版社

图书在版编目(CIP)数据

生存分析 / 彭非, 王伟编著.
北京: 中国人民大学出版社, 2004
(21 世纪统计学系列教材)

ISBN 7-300-05956-2 / F·1900

I. 生…

II. ①彭…②王…

III. 生存率 - 统计分析 - 高等学校 - 教材

IV. R195.4

中国版本图书馆 CIP 数据核字(2004)第 096610 号

21 世纪统计学系列教材

生存分析

彭 非 王 伟 编著

出版发行	中国人民大学出版社		
社 址	北京中关村大街 31 号	邮政编码	100080
电 话	010-62511242(总编室)		010-62511239(出版部)
	010-82501766(邮购部)		010-62514148(门市部)
	010-62515195(发行公司)		010-62515275(盗版举报)
网 址	http://www.crup.com.cn		
	http://www.ttrnet.com (人大教研网)		
经 销	新华书店		
印 刷	北京宏伟双华印刷有限公司		
开 本	787×965 毫米 1/16	版 次	2004 年 10 月第 1 版
印 张	33	印 次	2004 年 10 月第 1 次印刷
字 数	595 000	定 价	38.00 元

版权所有 侵权必究

印装差错 负责调换

《21 世纪统计学系列教材》编委会

编委会主任 易丹辉

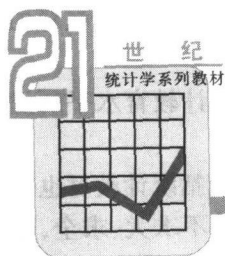
编委会委员 (按姓氏笔画排序)

尹德光 冯士雍 张尧庭

陈希孺 吴喜之 赵彦云

柯惠新 袁 卫 倪加勋

顾 岚 袁寿庄 耿 直



总 序

改革开放以来,高等统计教育有了很大的发展。随着课程设置的不断调整,有不少教材出版,同时也翻译引进了一些国外优秀教材。作为培养我国统计专门人才的摇篮,中国人民大学统计学系自 1952 年创建以来,走过了风风雨雨,一直坚持着理论与应用相结合的办学方向,培养能够理论联系实际、解决实际问题的高层次人才。随着新知识和网络时代的到来,我们在教学科研的实践中,深切地感受到,无论是自然科学领域、社会科学领域的研究,还是国家宏观管理和企业生产经营管理,甚至在人们的日常生活中,信息需求量日益增多,信息处理技术更加复杂,作为信息技术支柱的统计方法,越来越广泛地应用于各个领域。

面对新的形势,我们一直在思索,课程设置、教材选择、教学方式等怎样才能使学生适应社会经济发展的客观需要。在反复酝酿、不断尝试的基础上,我们决定与统计学界的同仁,共同编写、出版一套面向 21 世纪的统计学系列教材。

这套系列教材聘请了中科院院士、中国科技大学陈希孺教授,上海财经大学数量经济研究院张尧庭教授,中国科学院数学与系统科学研究所冯士雍研究员等作为编委。他们长期任中国人民大学的兼职教授,一直关心、支持着统计学的学科建设和应用统计的发展。中国人民大学应用统计科学研究中心 2000 年已成为国家级研究基地,这些专家是首批专职或兼职研究人员。这一开放性研究基地的运作,

将有利于提升我国应用统计科学研究的水平,也必将进一步促进高等统计教育的发展。

这套教材是我们奉献给新世纪的,希望它能够为促进应用统计教育水平的提高增添一份力量。这套教材力求体现以下特点:

第一,在教材选择上,主要面向经济类统计学专业。选材既包括统计教材也包括风险管理与精算方面的教材。尽管名为统计学系列教材,但并不求大、求全,而是力求精选。对于目前已有的内容较为成熟、适合教学需要、公认的较好的教材,并未列入本次出版计划。

第二,每部教材的内容和写作,注意广泛吸收国内外优秀教材的成果。教材力求简明易懂、内容系统和实用,注重对统计方法思想的阐述,并结合大量实际数据和实例说明统计方法的特点及应用条件。

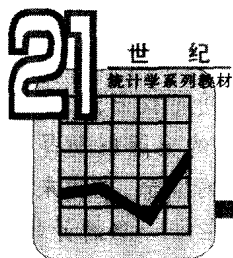
第三,强调与计算机的结合。为着力提高学生运用统计方法分析解决问题的能力,教材所涉及的统计计算,要求运用目前已有的统计软件。根据教材内容,选择使用 SAS、SPSS、TSP、STATISTICA、EViews、MINITAB、Excel 等。

感谢中国人民大学出版社的同志们,他们怀着发展我国应用统计科学的热情和提高统计教育水平的愿望,经过反复论证,使这套教材得以出版。感谢参与教材编写的同行专家、统计学系的教师。愿大家的辛勤劳动能够结出丰硕的果实。我们期待着与统计学界的同仁,共同创造应用统计辉煌的明天。

易丹辉

2000 年 8 月

于中国人民大学



前言

生存分析是研究生存现象和响应时间数据及其统计规律的一门学科。该学科在生物学、医学、保险学、可靠性工程学、人口学、社会学、经济学等方面都有重要应用。目前这方面的教材,国内还不太多,而且大多局限于生存分析的某一特定应用领域。在现有的几本教科书中,或者偏重于工程学,或者偏重于生物学和医学。上述教材的共同特点是通过生存分析的方法体系来说明其基本应用范围,而不是围绕生存分析的问题类型及特点来寻找理论方法。

在总结前人工作的基础上,我们结合了近来一些方面的最新研究进展情况,如对截尾和双删失数据以及对任意删失和截尾数据的生存函数估计;对单调平均剩余寿命的估计方法;离散生命表的平滑方法;对生存分布比较的常用检验方法,两个平均剩余寿命函数比较的非参数检验;利用线性回归模型对平均生存函数的估计方法;含有随机右删失数据且在删失值遗漏情况下,指数分布的统计推断问题;存在多维固定协变量情况下的加法危险模型的拟合优度检验;多变量生存变量危险梯度的概念,以及多维平均剩余寿命的定义、边界平均剩余寿命的概念;利用加权偏似然得分方程对多变量比例危险模型的回归参数进行估计;利用估计方程对边界模型中的比例危险参数进行假设检验等诸方面内容。这些内容构建成现在这本教材的主要框架体系。

为避免向读者抽象地介绍生存分析的方法和模型,我们没有按照统计方法的体系来组织本教材的内容,而是按照生存分析所研究问题的类型及采用的相应方法来建立本教材的内容体系。也就是说,我们是面对生存分析的问题找方法,而不是按照方法体系来组织内容。换言之,突出应用和解决生存分析的实际情况是本教材的一个特点。为此,我们根据生存分析所研究的问题和使用的概念、方法来构造本教材的框架,既保证内容充实,又限于简单的记流水账。在每一章的开始,我们对将要探讨的问题作了一个概要的说明,以便于读者对一章的内容有一个总体的把握和了解;在每一章的后面,我们提出了留给读者思考的问题(当然有些问题可能不在本书考虑的范围),以帮助读者检验是否掌握了全章的主要内容。目的是使本教材在使用过程中能发挥更大的作用,既让读者学到相关知识,又启发读者进行积极的思考。

在编写本教材的过程中,我们试图用通俗易懂的语言和实例,来介绍生存分析所涉及的内容、概念及其常用的统计理论和方法。在引入模型和方法的时候,注重对其产生的思想背景的介绍。由于目前计算机技术的应用已经相当普及,在很多统计软件中都包括有生存分析的算法,所以我们着重从实际应用的角度出发,结合实际案例,力争阐明方法的应用背景和解决的实际问题,在相关部分对有关的软件工具进行了介绍,而没有把重点放在方法或模型的理论推导上。

由于生存分析本身的理论和方法内容还处于不断发展和完善之中,我们在编写本教材的过程中采用了开放式的结构。即首先根据我们提出的生存分析所面临的问题来设计本书的结构,然后我们对现有的相关理论成果进行整理归纳,并对有待进一步解决的问题提出我们的思考和建议。因此在本书开始时我们即对生存分析中需要解决的问题给出了一个整体的框架,随后整个教材中的内容都围绕这个框架展开。

为适应不同领域读者的需要,本书在例子的选择上尽可能考虑取材于来自多个不同学科的案例,在本书中我们既选取了工程学(可靠性)、医学和生物学的例子,也选取了社会学(人口学)、经济学等方面的例子。希望读者能够从这些来自不同学科领域的案例中,更全面地把握和了解生存分析方法的用途。

此外,本书在介绍相关的概念和方法之前,还尽量给出其所解决或研究问题与方法的产生背景,以帮助读者对所学习的概念和方法有更完整的认识和更多的启发。

本书的具体内容包括以下几个部分:

第一部分由前三章组成,在介绍一些典型例子的基础上,重点介绍生存数据的类型及特征,常用的生存分布的统计模型,以及表征生存变量特征的有关函数

及性质。

第二部分包括第4章至第11章,主要针对不同的数据类型(删失和截尾)、分组数据及含有协变量的变量,介绍其生存分布,或相关参数的估计和检验问题所使用到的方法。具体来说,第4章至第7章主要介绍非参数估计和检验方法,第8章至第10章主要介绍半参数模型的估计和检验方法,在第11章中,将向读者介绍参数估计和检验方法。

第三部分包括第12章和第13章,重点介绍上述三种估计的非正规方法——图方法,以及对于不同估计的诊断、比较与改进方法,包括残差分析和拟合优度检验等方面的内容。

第四部分由第14章独立组成,将向读者介绍多变量生存分布的估计和检验问题及与此有关的一些方法。

此外,对有关生存分析的计算程序的实现问题,我们本打算单列出来,后来考虑与全书内容的衔接连贯方面的问题,仅在相关章节里进行了相应介绍;有关生存分析所涉及的试验设计问题,我们认为超出了本书的范围,所以没有进行介绍,对此有兴趣的读者可以参考有关的资料。

在编著本书时,我们希望达到如下四方面目的:

针对性:针对不同的生存分析问题寻找与之相适应的分析方法。

实用性:对于生存分析方法的论述不是仅仅停留在简单的方法介绍上,而是为了解决生存分析的实际问题来介绍每一种方法的应用前提和应用结果,力争最大限度地使用计算机实现该方法的应用提供方便。

通俗性:本书尽量避免使用生疏偏僻的例子,用相对易于理解的实例来讲解具体方法的应用,在语言的叙述上努力做到深入浅出,通俗易懂,以便不同领域和学科的读者理解和应用本书介绍的方法。

广泛性:除了自然科学和医学之外,本书还选择了一些生存分析在社会科学中应用的例子(如经济学和社会学),尽可能为在这方面有需求的读者提供有价值的参考。

本教材可以作为在生物学、医学、保险精算学、可靠性工程学、人口学、社会学、经济学等学科领域需要使用生存分析方法的教学参考书,也可以作为相关专业的研究生教材。

本教材的两位作者,彭非毕业于比利时鲁汶大学经济与社会学院(1997年5月获得社会科学博士学位),目前在中国人民大学统计学系风险管理与精算教研室任教授,同时兼任中国人民大学应用统计科学研究中心的兼职研究员;王伟毕业于中国科学院系统科学研究所(1994年3月获得运筹学与控制论博士学位),曾在中

国北方交通大学作博士后研究(1996年5月出站),现任中国国防科技信息中心副研究员。全书的写作分工是,由彭非首先提出全书的具体构想和章节安排,并负责第1章至第6章和第11章的编写工作,王伟负责第7章至第10章和第12章至第14章的编写工作。在全书初稿完成后,两位作者一起多次讨论修改并最终完成了全书的定稿。

在本书的编写过程中,中国人民大学统计学系袁卫教授、易丹辉教授对本书的编写给予了很多支持和帮助;中国人民大学信息学院的杨健教授、赵国庆教授和林勇教授对本书的编写和修改定稿提出了不少有益的建议;此外,太平洋财产保险公司精算部的石可,中国出口信用保险公司风险管理部的袁彬,中国工商银行北京市分行的谭诤,现在美国弗吉尼亚大学(University of Virginia)留学的张志勇,在本书前期准备工作的过程中,收集整理了部分相关资料并翻译了一些外文文献;值本书付梓之际,作者对上述诸君的帮助,表示由衷的感谢。

在我们准备这本教材的时候,我们注意到生存分析这门学科还在快速发展,主要体现在如下几个方面:

研究所涉及的领域在不断扩大,特别是一些有关的刊物介绍一些新颖的事例,而解决这些问题的有效工具即是生存分析的方法。

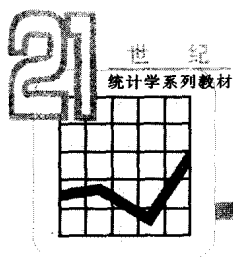
研究方法本身也在不断改进,比如对原有方法的改进,包括估计方法、模型(广义加法模型、广义比例危险模型)以及多变量生存分析等。

涉及多变量的生存分析问题和方法大量出现,正在成为新的发展方向。

因此,我们不希望这本教材成为一种固定的模式,只希望这本教材起到抛砖引玉的作用,我们也将适当的时候,增删改进本教材,希望同仁多提宝贵意见。

编著者

2004年7月



目 录

第一部分 基本生存数据与变量、函数和模型	1
第1章 数据示例及基本问题	3
1.1 前言	3
1.2 生存分析的数据示例	4
1.3 生存分析的基本问题和内容	22
1.4 其他相关问题	23
习题	24
第2章 基本数据与变量类型	25
2.1 本章简介	25
2.2 删失与截尾	25
2.3 分组数据	33
2.4 相关因素(协变量)	35
习题	38
第3章 基本函数和模型	40

3.1	本章简介	40
3.2	生存分析的基本函数	41
3.3	生存数据建模常用的参数模型	53
3.4	生存数据的回归模型(半参数模型)	59
3.5	删失和截尾数据似然函数的构造	62
3.6	计数过程	66
	习题	74
第二部分 生存数据的基本分析方法		81
第4章	右删失和左截尾数据基本特征的非参数估计	83
4.1	本章简介	83
4.2	右删失数据的生存函数和累积死亡力函数的估计	84
4.3	生存函数点估计的置信区间	93
4.4	生存函数的置信带	96
4.5	生存时间均值和中位数的点估计与区间估计	103
4.6	对同时存在左截尾和右删失数据的生存函数的估计	107
	习题	110
第5章	其他样本结构下基本数量特征的估计	116
5.1	本章简介	116
5.2	左删失、双删失和区间删失数据的生存函数估计	117
5.3	右截尾数据生存函数的估计	124
5.4	左截尾和区间删失数据的迭代估计	126
5.5	基于截尾和双区间删失数据的自一致估计	128
5.6	团体生命表中生存函数的估计	133
5.7	平均剩余寿命的估计	138
	习题	140
第6章	单变量估计的核平滑方法	146
6.1	本章简介	146
6.2	估计平滑(修匀)的概念	146
6.3	危险率函数的核平滑估计	147
6.4	离散危险函数的平滑估计	157
6.5	超额死亡率的估计	159
6.6	贝叶斯非参数平滑估计方法	168

习题	178
第 7 章 非参数估计的假设检验	180
7.1 本章简介	180
7.2 单样本检验	181
7.3 两个或多个样本的检验	184
7.4 趋势检验	193
7.5 分层检验	196
7.6 Renyi 型检验	200
7.7 其他两样本检验	203
7.8 对生存分布进行比较的检验方法	205
7.9 平均剩余寿命及生存寿命中位数的检验	216
习题	221
第 8 章 固定协变量的半参数比例危险回归	227
8.1 本章简介	227
8.2 比例危险模型的表示	227
8.3 可区分事件时间数据的偏似然估计	229
8.4 存在结时的偏似然估计	234
8.5 局部检验	239
8.6 利用比例危险模型建模	247
8.7 生存函数的估计	253
习题	256
第 9 章 半参数比例危险模型的改进	261
9.1 本章简介	261
9.2 时间相依协变量	261
9.3 分层比例危险模型	272
9.4 左截尾情形下的比例危险模型	278
9.5 时间变化效应的综合(多阶段模型)	280
习题	290
第 10 章 加法危险回归模型	293
10.1 本章简介	293
10.2 固定参数的估计	294
10.3 非参数加法危险模型的最小二乘估计	295
10.4 非参数加法危险模型的检验	306

10.5	回归模型中生存函数的估计	310
	习题	313
第 11 章	参数模型的统计推断	315
11.1	本章简介	315
11.2	含有随机右删失数据的指数分布	316
11.3	威布尔分布	319
11.4	对数罗吉斯蒂分布	325
11.5	其他参数模型	328
	习题	332
第三部分	生存分析的模型检验方法介绍	335
第 12 章	模型检验的图形方法	337
12.1	本章简介	337
12.2	生存函数或危险函数的估计用图	338
12.3	评价 Cox 模型拟合的 Cox-Snell 残差	342
12.4	确定协变量的函数形式: 鞅残差	347
12.5	比例危险率假设的图形检验	351
12.6	偏离残差	365
12.7	检验个别观测值的影响	368
12.8	检验加法模型拟合的图方法	373
12.9	参数模型的图方法	376
12.10	参数模型的诊断方法	381
12.11	模型选择和检验方法	390
	习题	390
第 13 章	生存分析中的拟合优度检验	395
13.1	本章简介	395
13.2	拟合优度检验的一般方法	395
13.3	对于无删失数据情况下的拟合优度检验	398
13.4	对删失数据的拟合优度检验	399
13.5	加法模型的拟合优度检验	401
13.6	参数估计的拟合优度检验	403
13.7	模型识别的假设检验	404
	习题	406

第四部分 多元生存分析方法介绍	407
第 14 章 多变量生存分析	409
14.1 本章简介	409
14.2 脆弱性模型与多变量生存分布	410
14.3 相关性的得分检验	412
14.4 伽玛脆弱性模型的估计	414
14.5 多变量生存数据的边际模型	419
14.6 基于偏似然得分方程的回归估计	421
14.7 有关危险比例的假设检验	423
习题	426
附录	430
附录 A 极大化的数值技术	430
附录 B 基于似然理论的大样本检验	436
附录 C 统计表	440
附录 D 选择的数据集	466
参考文献	494

第一部分

基本生存数据与变量、 函数和模型

