卫生统计方法与应用进展

主　编　饶克勤

第2卷

人民卫生出版社

卫生统计方法与应用进展

第2卷

主　编　饶克勤

副主编　孙　高　徐勇勇　陈　峰

编　委（按姓氏笔画排序）

王　彤　山西医科大学卫生统计学教研室
王心旺　广州医学院公共卫生与全科医学学院
刘　沛　东南大学公共卫生学院
孙　高　中国医科大学公共卫生学院
孙振球　中南大学湘雅医学院公共卫生学院
余松林　华中科技大学同济医学院公共卫生学院
李　康　哈尔滨医科大学公共卫生学院
陈　峰　南京医科大学流行病与卫生统计学系
陈平雁　南方医科大学生物统计学系
陈启光　东南大学公共卫生学院
易　东　第三军医大学卫生统计学教研室
柳　青　中山大学公共卫生学院
饶克勤　卫生部统计信息中心
夏结来　第四军医大学卫生统计学教研室
徐勇勇　第四军医大学卫生统计学教研室
章扬熙　辽宁省疾病预防控制中心
董景五　北京协和医院世界卫生组织疾病分类合作中心

人民卫生出版社

图书在版编目（CIP）数据

卫生统计方法与应用进展　第2卷/饶克勤主编.
—北京：人民卫生出版社，2008.1
ISBN 978-7-117-09381-1

Ⅰ.卫…　Ⅱ.饶…　Ⅲ.卫生统计-统计方法
Ⅳ.R195.1

中国版本图书馆CIP数据核字（2007）第168770号

卫生统计方法与应用进展

第2卷

主　　编：饶克勤
出版发行：人民卫生出版社（中继线010-67616688）
地　　址：北京市丰台区方庄芳群园3区3号楼
邮　　编：100078
网　　址：http://www.pmph.com
E - mail：pmph@pmph.com
购书热线：010-67605754　010-65264830
印　　刷：北京新丰印刷厂
经　　销：新华书店
开　　本：889×1194　1/16　　**印张**：20.75
字　　数：620千字
版　　次：2008年1月第1版　2008年1月第1版第1次印刷
标准书号：ISBN 978-7-117-09381-1/R·9382
定　　价：47.00元

序　言

从20世纪20年代起，统计学的理论与方法日益广泛地被生物医学研究工作者所应用。随着流行病学、基因组学、蛋白质组学、代谢组学、药物开发、计算机和信息等学科的迅猛发展，促使了统计学与这些学科的交叉融合，并且对生物统计学、医学统计学和卫生统计学研究人员提出了很多实践中的新课题。为了解决这些课题，统计学家在对经典统计理论研究和认识更加深化的基础上，不断探索和发展统计的新理论和新方法。

第2卷内容重点介绍近年来在医学卫生研究中所应用的新理论和新方法，全卷共分十章，每章内容相对独立。内容包括Cox比例风险模型、生物信息分析统计方法、非经典条件下的回归分析方法、结构方程模型、广义估计方程和多水平模型、Bootstrap方法、Permutation检验、Monte Carlo方法、数据挖掘、Bayes统计方法。每章在介绍方法的基础上大多附有实际例子，使读者了解方法的意义和实际应用。

下面简要介绍第2卷各章内容：

在**Cox比例风险模型的发展与应用**一章中，首先介绍Cox回归模型的基本方法，即在等比例条件成立时，不同协变量情况下模型的构建、参数的估计以及参数的解释；然后介绍检验等比例条件是否成立的图示法、参数识别法、残差分析法；再介绍非比例风险模型的拟合方法，包括模型中加入协变量与时间的交互项、根据协变量不同进行分段拟合和分层拟合等三种方法；最后介绍有序重复事件资料的AG模型、总时间模型、WLW模型和PWP模型等四种模型。所有这些方法均结合实际资料的分析，并给出了相应的SAS分析程序。

在**生物信息分析统计方法**一章中，首先回顾生物信息学的现状和前景，阐述生物信息学的生物内涵和信息学内涵；然后探讨生物序列比较的方法：数据库搜索Blast工具的应用、序列比较中相似性分析（包括全局相似性和局部相似性）及其统计学检验；再介绍基因芯片的统计分析方法，包括基于基因芯片的数据挖掘及可视化和基因转录调控网络分析等；最后介绍蛋白质序列模式和序列结构域模式的分析方法（频数表法和权值矩阵法）。

众所周知，经典的线性回归模型由于其形式简单，计算不很复杂，且参数意义明确，实施方便，已在医疗卫生研究中被广泛应用并取得许多成果。但是由于经典的线性回归模型要求因变量的条件分布服从正态、方差齐、与自变量的线性关系等使用条件，如果实际问题不满足这些条件而一味地套用经典模型，则将导致误用或滥用该统计方法。事实上，非经典条件下也有相应的回归分析方法并在不断发展中，在**非经典条件下的若干回归分析方法**一章中，介绍了针对误差项非正态分布情况下的稳健线性回归模型（robust linear regression models）、因变量存在不确定取值的截取回归模型（censored regression models）以及自变量与因变量关系非线性甚至未知函数关系时的非参数回归与广义可加模型GAM（generalize additive models），并通过实例介绍了这些方法的应用和软件实现。

医学研究中，有很多指标是不可直接测量的（又称为潜变量），有时这些指标互相之间还可能存在直接或间接的联系甚至有因果关系。要科学地评价这些不可直接测量指标，如果应用传统的多因素分析方法就存在局限性。Jöreskog等人在传统的因子分析即探索性因子分析的基础上发展出证实性因子分析，并且将因子分析方法与通径分析相结合而发展出一种多变量的新的统计方法即结构方程模型分析方法。在**结构方程模型**一章中，介绍了这种分析方法是应用在对某些客观存在的现象，根据专业理论提

出事物中特定的内在结构假设，用统计中的假设检验方法检验所提出的假设是否成立，并且检验这种对总体结构所作的假设与实际收集的数据资料的符合程度。因此，结构方程模型分析是一种具有证实性或称为验证性的数据分析方法。这种分析方法已经大量地应用到教育学、心理学、社会学和行为科学等学科的研究中。近年来，结构方程模型作为一种新兴的统计分析方法，在生存质量评价、临床试验疗效评价以及中医证候研究等医学领域中也得到了成功应用。

广义估计方程(GEEs)是 Liang & Zeger(1986)在广义线性模型的基础上提出的，用于分析纵向观察资料分析的一种统计分析方法。在**广义估计方程和多水平模型**一章中，首先介绍广义估计方程的定义、参数估计，讨论几种常见的作业相关矩阵，包括重复测量的数值变量资料、二分类资料和事件数资料的广义估计方程的建立和分析以及作业相关矩阵的选择；然后介绍多水平模型。1986 年英国教育统计学家 H. Goldstein 首先提出多水平模型并专门用于处理具有多层次或多水平结构的资料。本章通过实例介绍多水平 logistic 模型、多水平 probit 模型及重对数模型、多水平 Poisson 模型、多类结果及有序结果的多水平 logistic 回归、多元重复测量资料的多水平模型；最后，就广义估计方程与多水平模型的区别和联系、缺失数据的处理、软件实现等问题进行了讨论。

Bootstrap 方法作为一种再抽样方法，最早是作为类似刀切法(jackknife)的偏差校正和模型验证的方法提出的。经过近年来积极的理论探讨和应用实践，目前已广泛应用到统计推断的几乎所有领域。在**Bootstrap 方法及其应用**一章中，首先介绍其基本思想的基础，再比较这种方法在参数估计和假设检验中与传统方法之间的不同和优势所在，其中重点介绍了三种 Bootstrap 区间估计的方法，并进而介绍了目前生物医学领域几个成功应用 Bootstrap 的实例，以说明该方法应用的特点和精妙所在。最后讨论了 Bootstrap 应用中应该考虑的诸如对资料的要求、重复数的多少、估计的偏差诊断等方面的问题，以及目前偏差校正方面的进展。

在**Permutation 检验及其应用**一章中，首先简要回顾 Permutation 检验的发展简史，概述其基本思想和实施步骤；然后以实例为载体，以资料的设计类型为主线，较为详尽地介绍 Permutation 检验在一元统计分析中的应用，并简要介绍 MRPP 程序在多元分析中的应用；再结合生物医学在新时期的发展，介绍 Permutation 在微阵列数据分析和新药临床试验等级资料等效性检验中的应用；最后就“Permutation”的含义、“Permutation 检验”的特点、“统计量的构造”、“模拟次数的选择”、应用前景等问题展开了深入探讨。

随着计算机技术的飞速发展，随机模拟越来越显示出其简单、有效的特点。在实际工作中，有许多科学问题难于或者不可能通过解析处理求得其数值解，只能进行模拟处理，此时模拟实验既可为知识探求提供方向，又是验证假设、模型是否成立的有效途径。Monte Carlo 方法是最常用的随机模拟方法之一。本卷专门在**Monte Carlo 方法及其在医学中的应用**一章中进行介绍。

蒙特卡洛方法(Monte Carlo)是指利用随机抽样方法得到数值问题近似解的一类方法，其核心是随机模拟。目前，Monte Carlo 模拟广泛应用于几乎所有的研究领域。按其是否涉及随机过程，可以将 Monte Carlo 方法的应用分为两类：第一类属于确定性的数学问题，例如多重积分的计算、线性代数方程组的求解、矩阵计算等；第二类属于随机性问题。由于 Monte Carlo 能够比较逼真地描述具有随机性质事物的特点及物理实验过程。从这个意义上讲，Monte Carlo 方法可以部分代替一些医学实验，甚至可以得到医学实验难以得到的结果。用 Monte Carlo 方法解决实际问题，可以直接从实际问题本身出发，而不从方程或者数学表达式出发，它有直观、形象的特点。作为一种重要的随机模拟方法，Monte Carlo 方法近年来受到医学研究工作者的关注，出现了一些较成功的应用，如饮食暴露评价、临床试验模拟、卫生经济预测等。本章首先通过引入随机模拟的概念介绍蒙特卡洛方法的产生，然后介绍 Monte Carlo 的基本概念、方法步骤和适用范围。在介绍 Monte Carlo 方法的收敛性和误差时，也讨论了减少方差的技巧和 Monte Carlo 方法的优缺点。最后介绍常用的 Monte Carlo 抽样方法以及在医学上的应用。

数据挖掘涉及的学科领域和方法很多，这门新兴的边缘学科结合了统计学、机器学习、模式识别、智

能数据库、知识获取、人工智能、专家系统、数据可视化及高性能计算等领域的知识。本卷第九章是**数据挖掘技术及其应用**。数据挖掘也常称为知识发现(knowledge discovery)，是从存放在数据库、数据仓库或其他信息库中的大量数据中挖掘有用知识，在数据中发现模式的过程。数据挖掘是运用统计学、人工智能、机器学习、数据库技术等方法发现数据的模型和结构、发现有价值的关系或知识的一门交叉学科。从庞大的观察数据集中提炼并分析出不可轻易察觉或断言的关系，最后给出一个有用的并可以理解的结论。

本卷最后一章是**贝叶斯**(Bayes)**统计方法应用**。贝叶斯统计学是统计学研究中的一个重要学派，它与经典统计学派(又称频率学派)在信息利用、统计方法乃至统计推断的哲学层面上均存在重大差异。值得注意的是，近年来，由于现代计算机技术的发展较好地解决了高维积分这一限制贝叶斯统计应用的"瓶颈"问题，使得贝叶斯统计不论在理论研究还是在应用方法上都取得了较经典统计学更快的发展。相对于发达国家，我国无论在贝叶斯统计理论研究还是应用研究上均要落后和迟缓一些。但近年来，国内学者已逐步认识到这一理论的重要性，高度关注贝叶斯统计方法的引进、介绍、研究和应用工作。本章首先通过介绍 Bayes 定理、Bayes 统计对信息的利用、先验分布的选择与确定，使读者对贝叶斯统计学及基本概念有所了解；然后介绍贝叶斯统计推断方法以及贝叶斯统计学与经典统计学的区别和联系；考虑到 Markov Chain Monte Carlo(MCMC)方法和 WinBugs 软件是近年来广泛应用于贝叶斯统计推断中的技术和工具，对其进行了重点介绍；最后介绍贝叶斯统计学在医学上的应用。我们热切期望我国卫生统计工作者重视对贝叶斯统计的研究和应用，以赶上当前国际统计学发展的潮流。

随着自然科学与社会科学的发展，对统计学工作者提出了很多实践中迫切需要解决的问题，这也促进了统计学理论与方法的发展。计算机科学及统计软件的日益发展又为统计学的发展提供了有效处理数据的工具。本卷由于篇幅有限，只能介绍目前最常见的几种新方法。限于我们的水平，书中有错误之处敬请读者不吝赐教。

饶克勤　陈启光　陈　峰　刘　沛

2007 年 9 月

前　言

21世纪是知识经济日新月异的时代，信息的广泛传播和交流，知识的不断更新，以及计算机技术的广泛应用，越来越多的现代新理论、新方法应用到卫生统计学领域，使卫生统计学原理、方法及应用得到了很大的发展。卫生统计学作为一门方法学与应用学科，随着医学科学的迅猛发展和多学科研究工作的开展，日益受到人们的重视和关注。卫生统计学早已不仅限于一般的统计描述与分析，还包括卫生信息化建设、健康统计、卫生管理与卫生服务统计、统计分类与代码、医疗保险统计、药物临床统计、基因挖掘与分子生物学中的统计方法、循证医学中的统计方法、生命质量测评中的统计方法等多种统计学理论和方法。为了向广大读者介绍卫生统计学理论、方法和应用的国内外最新知识和信息，以适应各相关学科、专业学生及工作人员学习与应用卫生统计学的需要，促进学术交流，由卫生部统计信息中心牵头，《中国卫生统计》杂志编辑部组织全国活跃在卫生统计科研与教学领域的中青年骨干专家学者编写了《卫生统计方法与应用进展》一书。

《卫生统计方法与应用进展》第2卷重点介绍近年来在医学卫生研究中所应用的卫生统计的新理论和新方法，全卷共有10章，各章相对独立。内容包括Cox比例风险模型、生物信息分析统计方法、非经典条件下的回归分析方法、结构方程模型、广义估计方程和多水平模型、Bootstrap方法，Permutation检验、Monte Carlo方法、数据挖掘、Bayes统计方法等。每章在介绍方法的基础上大多附有实际例子和计算机程序，使读者了解方法的意义和实际应用。本书各章内容相对独立，具有一定的实用性、先进性和新颖性，期望本书能够不断更新基层卫生统计工作者、科研人员以及硕士、博士研究生的卫生统计知识，提高卫生统计水平，对我国的卫生统计理论、方法和应用的交流有所推动和促进。

《卫生统计方法与应用进展》第2卷由人民卫生出版社出版发行。刘延龄教授、陈启光教授、徐勇勇教授和陈峰教授参与了书稿的审定工作，提出了许多宝贵的意见。本书秘书组胡建平、郭海强、曲波三位同志在文稿的组织编辑修改过程中付出了辛勤劳动。在本书出版之际，谨向曾对书稿审读、指导和提供意见的各位专家，向给予本书出版以关怀和支持的诸多同道表示深深的谢意。

由于编者水平所限及编写时间匆忙，在全书的内容、方法和编排上难免有不当之处，真诚地希望不吝指正。

饶克勤

2007年9月

目　录

第一章　Cox比例风险模型的发展与应用 …… 1
第一节　Cox比例风险模型 …… 1
一、Cox比例风险模型的结构 …… 1
二、参数估计 …… 2
三、关于参数的解释 …… 5
四、变量的不同编码方式对参数估计值的影响 …… 6
第二节　比例风险性质的判别 …… 8
一、比例风险的性质 …… 8
二、比例风险的生存概率曲线识别法 …… 9
三、比例风险的参数识别法 …… 12
四、比例风险的残差分析法 …… 12
第三节　非比例风险的Cox模型配合 …… 18
一、配合协变量与时间交互作用模型(时依系数法) …… 18
二、配合带时依协变量的Cox模型(分段模型) …… 21
三、非比例风险的分层分析法 …… 24
第四节　多次事件的生存分析 …… 26
一、多次事件的资料结构 …… 26
二、各种整理模式下的模型结构 …… 27
三、多次事件资料的模型配合过程 …… 28
四、多种事件的分析 …… 39

第二章　生物信息分析统计方法 …… 42
第一节　生物信息学概述 …… 42
一、生物信息学研究现状与发展趋势 …… 43
二、生物信息学的生物内涵 …… 44
三、生物信息学的信息学内涵 …… 46
四、生物信息学研究和发展中的交叉学科和大科学特点 …… 50
第二节　序列比较方法 …… 52
一、数据库搜索简介 …… 52
二、序列相似性定义 …… 56
三、序列类似性的统计显著性 …… 59
四、算法的敏感性与准确度(选择性) …… 62
五、有空隔配准的BLAST程序与位置特异的迭代BLAST程序 …… 63
第三节　基因芯片的统计分析方法 …… 67
一、基因芯片 …… 67

二、基于基因芯片的数据挖掘及可视化 …… 68
三、基因转录调控网络分析 …… 72
第四节 蛋白质序列模式和序列结构域摸式 …… 74
一、基准序列(序列模式):标纹、标志、指纹和位点 …… 74
二、序列结构与模式匹配方法 …… 75

第三章 非经典条件下的若干回归分析方法 …… 77
第一节 稳健回归方法 …… 77
一、稳健统计的基本理论 …… 78
二、稳健回归方法进展 …… 81
三、应用实例及软件实现 …… 86
第二节 截取回归模型 …… 88
一、Tobit 模型概述 …… 88
二、Tobit 模型的异方差性和非正态性 …… 91
三、应用实例及软件实现 …… 95
第三节 非参数回归与广义可加模型 …… 98
一、非参数回归的基本方法 …… 99
二、偏倚-方差权衡和光滑参数的选择 …… 103
三、可加模型 …… 105
四、广义可加模型 …… 107
五、应用实例及软件实现 …… 112

第四章 结构方程模型 …… 118
第一节 前言 …… 118
第二节 结构方程模型中的几个基本概念 …… 119
第三节 结构方程模型中的两类子模型 …… 119
第四节 路径图及 SEM 的协方差结构 …… 121
第五节 结构方程模型的分析步骤 …… 123
第六节 结构方程模型中的模型识别 …… 123
第七节 结构方程模型分析软件 …… 125
第八节 结构方程模型参数估计 …… 125
第九节 结构方程模型的拟合度评价 …… 126
第十节 结构方程模型的修正 …… 127
第十一节 应用实例 …… 127

第五章 广义估计方程和多水平模型 …… 153
第一节 广义估计方程 …… 153
一、GEE 模型简介 …… 153
二、几种常见的组内相关矩阵 …… 154
三、GEE 的参数估计 …… 155
四、GEE 在生物医学领域中的应用 …… 156
五、其他应用 …… 160
第二节 多水平模型 …… 160

一、多水平模型简介 …… 160
二、多水平模型的参数估计 …… 162
三、多水平 logistic 模型 …… 163
四、多水平 probit 模型及余重对数模型 …… 164
五、多水平 Poisson 模型 …… 164
六、多类结果及有序结果的多水平 logistic 回归 …… 167
七、多元重复测量资料的多水平模型 …… 167
第三节 广义估计方程与多水平模型的正确应用 …… 169
一、GEE 中作业相关矩阵的选择 …… 169
二、关于缺失数据 …… 170
三、GEE 与多水平模型的比较 …… 170
四、GEE 与多水平模型的软件实现 …… 170

第六章 Bootstrap 方法及其应用 …… 172
第一节 发展简史 …… 172
第二节 基本思想 …… 172
第三节 与传统方法的比较 …… 173
一、Bootstrap 区间估计 …… 174
二、Bootstrap 假设检验 …… 177
第四节 在生物医学领域的应用 …… 177
一、主成分的可信区间估计 …… 177
二、可加性 logistic 回归模型参数的估计 …… 178
三、临床试验中生物等效性检验 …… 180
第五节 Bootstrap 方法的正确应用 …… 181
一、Bootstrap 方法的资料要求 …… 181
二、Bootstrap 的误差与自举样本数的确定 …… 181
三、Bootstrap 的刀切法诊断 …… 182
四、Bootstrap 法的偏差校正 …… 182

第七章 Permutation 检验及其应用 …… 184
第一节 发展简史 …… 184
第二节 基本思想和实施步骤 …… 184
一、基本思想 …… 184
二、实施步骤 …… 185
第三节 Permutation 检验与传统方法的比较 …… 185
一、在一元分析中的应用 …… 185
二、在多元分析中的应用 …… 194
第四节 在生物医学领域中的应用 …… 197
一、微阵列数据分析中的应用 …… 197
二、临床试验资料分析中的应用 …… 198
第五节 Permutation 检验的正确应用 …… 199
一、Permutation 含义和特点 …… 199
二、检验统计量与模拟次数 …… 200

三、应用前景 …… 201

第八章 Monte Carlo 方法及其在医学中的应用 …… 203
第一节 简介 …… 203
第二节 Monte Carlo 方法的基本思想 …… 203
一、Monte Carlo 方法的基本原理 …… 203
二、Mote Carlo 方法的一般步骤 …… 204
三、一个简单的例子 …… 205
四、Monte Carlo 方法的适用范围 …… 205
第三节 Monte Carlo 方法的收敛性和误差 …… 206
一、Monte Carlo 方法的收敛性 …… 206
二、Monte Carlo 方法的误差 …… 206
三、减少方差的一些技巧 …… 207
四、Monte Carlo 方法的优缺点 …… 208
第四节 随机数和伪随机数 …… 209
一、随机数及其性质 …… 209
二、产生随机数的方法 …… 209
三、伪随机数的独立性和均匀性 …… 210
四、伪随机数的产生方法 …… 210
第五节 常用的 Monte Carlo 抽样方法 …… 211
一、连续型变量的抽样方法 …… 211
二、离散型变量的抽样方法 …… 212
三、特殊的抽样方法 …… 213
四、多维随机变量的抽样 …… 214
五、关于正态分布的抽样 …… 215
第六节 Monte Carlo 方法在医学上的应用 …… 216
一、回归分析中的应用 …… 216
二、饮食暴露评价 …… 217
三、生物医学现象(过程)的直接模拟 …… 218
四、疾病预防与监测中抽样方案的考查 …… 219
五、药物的临床实验 …… 219
六、应用中的注意事项 …… 219

第九章 数据挖掘技术及其应用 …… 221
第一节 数据挖掘概述 …… 221
一、数据挖掘的定义和范畴 …… 221
二、数据挖掘的特点 …… 223
三、数据挖掘算法的基本要求 …… 223
四、数据挖掘的过程 …… 223
第二节 概念描述 …… 230
一、概念描述的生成过程 …… 230
二、概念分层 …… 230
三、数据泛化 …… 230

第三节 数据挖掘基础数学理论 …… 232
一、基于概率论和数理统计的数据挖掘 …… 232
二、模糊理论 …… 237
三、粗糙集理论 …… 240
四、不确定性理论的关系 …… 244
第四节 数据挖掘最优化理论 …… 245
一、模拟退火算法 …… 245
二、人工神经元模型 …… 247
三、进化算法(evolutionary algorithm) …… 249
四、蚁群算法(ant colony alogrithm) …… 253
五、支持向量机 …… 254
六、SA、ANN、EA、ACA、SVM 的比较 …… 260
第五节 分类方法 …… 260
一、基于数理统计的分类算法 …… 261
二、基于机器学习的分类算法 …… 265
第六节 聚类方法 …… 272
一、聚类分析概述 …… 272
二、聚类处理的数据结构 …… 273
三、相似性测度 …… 274
四、聚类算法种类 …… 275
五、典型聚类方法 …… 276
第七节 关联规则 …… 284
一、基本概念 …… 285
二、关联规则挖掘算法 …… 286
三、基于兴趣度的关联规则挖掘 …… 289

第十章 Bayes 统计方法应用 …… 292
第一节 概述 …… 292
一、Bayes 定理 …… 292
二、Bayes 统计对信息的利用 …… 293
三、先验分布的选择与确定 …… 295
四、Bayes 统计推断 …… 296
五、Bayes 统计学与经典统计学的联系 …… 297
第二节 使用 MCMC 方法解决 Bayes 统计计算问题 …… 298
一、Bayes 统计学所面临的实际困难 …… 298
二、MCMC 方法概述 …… 299
三、使用 MCMC 方法需要考虑的几个实际问题 …… 299
第三节 Bayes 统计分析软件— WinBUGS …… 299
一、构造统计模型 …… 300
二、迭代收敛性的诊断 …… 301
三、WinBUGS 一般操作 …… 303
第四节 应用实例 …… 303
一、对各医院心脏手术死亡率的估计 …… 303

二、一般线性回归 …… 305
三、logistic 回归 …… 308
四、meta 分析 …… 312
五、应用 Cox 回归进行生存分析 …… 314

第一章

Cox比例风险模型的发展与应用

生存分析起源于对死亡的研究，用来分析从出生至死亡之间的寿命长度。可以把生存（或死亡）时间的概念扩展为事件（event）发生所经历的时间，如疾病潜伏期、仪器使用寿命、婚姻维持期等。生存分析方法就是用来研究事件发生的时间规律及其影响因素的一种统计方法。这里的反应变量是某特定事件发生所经历的时间。例如在肿瘤治疗中，对比不同治疗方案（手术、放疗和化疗）的缓解期或生存期的长短。在劳动卫生与职业病研究中，计算从开始职业暴露到发生职业病的潜伏期长短等。由于事件时间的分布往往不是正态的，而且在观察过程中常有失访事例发生，故生存分析形成了一套特定的统计方法。1958年Kaplan和Meier提出了生存概率的非参数估计方法，从而奠定了现代生存分析方法的基础。用于生存分析的回归模型有指数回归、Weibull回归、Gamma回归、对数logistic回归和竞争风险模型等。1972年D. R. Cox提出了比例风险模型，从而使生存分析的多因素分析方法发生了一个质的飞跃。近几十年来，对Cox回归模型的发展和应用研究一直是生存分析中的热点，并取得了很大成绩。本章仅限于介绍Cox比例风险模型的发展与应用。

为便于叙述，本章把出生时间作为生存时间的观察起点，死亡事件作为生存时间的终点，生存时间是从出生到死亡所经历的时间长度。如果中间由于各种原因非因死亡事件而退出观察的个体，称为失访，其所经历的时间长度称为失访时间（censored time）。以下将介绍Cox比例风险模型的结构、比例风险性质的判别、非比例风险的Cox模型配合以及多个事件的生存分析方法。

在生存分析中，常用的统计指标有死亡概率、生存概率、死亡速率和死亡风险四个统计指标。死亡风险又称风险率（hazard rate），风险函数（hazard function）或死亡力（force of mortality），它表示已经活到时间 t 的一个人，在其后的 $t+\Delta t$ 这一极小时间区间内死亡的概率。用公式表示为：

$$h(t)=\frac{f(t)}{1-F(t)}=\frac{f(t)}{S(t)}$$

式中 $F(t)$ 为在时间 t 时的死亡概率，$S(t)=1-F(t)$ 表示一个人能活到时间 t 的概率。$f(t)$ 为在时间 t 时的死亡速率。Cox比例风险模型就是用回归模型建立风险函数与有关因素之间的关系。

第一节　Cox比例风险模型

一、Cox比例风险模型的结构

假定检测两个人的生存类型，他们的不同死亡风险是由于他们对影响死亡风险大小的相关因素具有不同的暴露水平之故。Cox比例风险模型建立风险函数与协变量之间的回归关系，把风险函数构造为协变量的对数线性函数。

记当一组协变量 X 处于基础状态下的一个人，在时间 t 的风险函数为 $h_0(t)$，对于协变量 X 不等于基础状态的任何其他人的风险函数记为 $h(t)$，Cox 比例风险模型表示 $h(t)$ 为 $h_0(t)$ 的 $\exp(\beta X)$ 倍。用公式表达为：

$$h(t)=h_0(t)\exp(\beta X) \tag{1-1}$$

这里 $h_0(t)$ 称为基准风险函数，是一个未加规定的函数。β 为模型的待估参数。协变量 X 处于基础状态可以是 0 状态（如危险因素的 0 暴露水平），也可以规定为平均状态（如平均血糖、平均血压等，因为实际上这些因素不可能处于 0 状态）。

对(1-1)式等式两边取对数得

$$\log(h(t))=\log(h_0(t))+\beta X=\alpha(t)+\beta X \tag{1-2}$$

式中 $\alpha(t)=\log(h_0(t))$。如果规定 $\alpha(t)=\alpha$，(1-1)式即为指数回归模型；如果规定 $\alpha(t)=\alpha t$，(1-1)式即为 Gompertz 回归模型；如果规定 $\alpha(t)=\alpha\log(t)$，(1-1)式即为 Weibull 回归模型。由于 Cox 回归模型对 $h_0(t)$ 未作任何规定，从而使模型具有很大的灵活性。

把(1-1)式右边的 $h_0(t)$ 移到等式左边得到相对风险 HR 为：

$$HR=\frac{h(t)}{h_0(t)}=\exp(\beta X) \tag{1-3}$$

二、参 数 估 计

用偏似然函数法求解参数估计值。令 $t_i(i=1,2,\cdots,n)$ 为第 i 例观察对象的生存时间，在生存时间无重复的条件下，把一组观察对象的生存时间顺序化并表示为：

$$t_1<t_2<\cdots<t_n$$

同时将已对生存时间顺序化的一组观察对象的相应协变量记为：

$$X_1,X_2,\cdots,X_n$$

定义恰在第 $i(i=1,2,\cdots,n)$ 个观察对象死亡之前的危险集为 $R_i=h_0(t)[\exp(\beta X_i)+\exp(\beta X_{i+1})+\exp(\beta X_{i+2})+\cdots+\exp(\beta X_n)]$，构造第 i 个观察对象死亡的偏似然函数分量为

$$l_i=\frac{h_0(t)\exp(\beta X_i)}{\sum_{j=n}^{i}h_0(t)\exp(\beta X_j)}=\frac{h_0(t)\exp(\beta X_i)}{h_0(t)\sum_{j=n}^{i}\exp(\beta X_j)}=\frac{\exp(\beta X_i)}{\sum_{j=n}^{i}\exp(\beta X_j)} \tag{1-4}$$

对全部观察对象构造的偏似然函数为

$$L=\sum_{i=1}^{n}l_i=\sum_{i=1}^{n}\frac{\exp(\beta X_i)}{\sum_{j=n}^{i}\exp(\beta X_j)} \tag{1-5}$$

当生存资料中存在有失访数据时，该失访个体在失访前包括在危险集内，但不包括在失访后的危险集内，也不能构造偏似然函数分量。令指示变量：

$$\delta_i=\begin{cases}1 & \text{如果第 } i \text{ 例观察对象死亡}\\ 0 & \text{如果第 } i \text{ 例观察对象失访}\end{cases}$$

这时的偏似然函数的构造为

$$L=\prod_{i=1}^{n}l_i^{\delta}=\prod_{i=1}^{n}\left[\frac{\exp(\beta X_i)}{\sum_{j=n}^{i}\exp(\beta X_j)}\right]^{\delta_i} \tag{1-6}$$

通常对偏似然函数取自然对数后用 Newton-Raphen 解法求出参数 β 的估计值 b 及其渐近方差 $V(b)$。

在实际工作中经常遇到在同一时间点上有多个死亡事件发生的情况。当在同一时间点上有多个死亡事件时，其偏似然函数的构造比较复杂。例如在时间点 t_i 上有 3 个死亡事件(A、B、C)，记为 t_{iA}、t_{iB} 和 t_{iC}，但无法分出它们的先后顺序。对这种持相等时间点的多个事件的偏似然函数分量的构造方法有精确法、Breslow 近似法和 Efron 近似法三种。简单介绍如下：

1. 精确法　这时需考虑这3个死亡事件的所有排列数，即3!=6。这6种排列为：ABC，BAC，ACB，CAB，BCA，CBA。用 j 表示第 j 种排列数，表示为

编号 j	1	2	3	4	5	6
排列	ABC	BAC	ACB	CAB	BCA	CBA

根据概率理论，这时的偏似然函数分量 l_i 是这6种排列的概率之：

$$l_i = \sum_{j=1}^{6} Pr(j) \tag{1-7}$$

等式右边的 $Pr(j)$ 表示编号 j 的概率。如

$$Pr\ (1)=\left(\frac{e^{\beta X_{iA}}}{e^{\beta X_{iA}}+e^{\beta X_{iB}}+e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)\left(\frac{e^{\beta X_{iB}}}{e^{\beta X_{iB}}+e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)\left(\frac{e^{\beta X_{iC}}}{e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)$$

$$Pr\ (2)=\left(\frac{e^{\beta X_{iB}}}{e^{\beta X_{iA}}+e^{\beta X_{iB}}+e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)\left(\frac{e^{\beta X_{iA}}}{e^{\beta X_{iA}}+e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)\left(\frac{e^{\beta X_{iC}}}{e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)$$

$$Pr\ (6)=\left(\frac{e^{\beta X_{iC}}}{e^{\beta X_{iA}}+e^{\beta X_{iB}}+e^{\beta X_{iC}}+\cdots+e^{\beta X_n}}\right)\left(\frac{e^{\beta X_{iB}}}{e^{\beta X_{iA}}+e^{\beta X_{iB}}+\cdots+e^{\beta X_n}}\right)\left(\frac{e^{\beta X_{iA}}}{e^{\beta X_{iA}}+\cdots+e^{\beta X_n}}\right)$$

假若持相等时间点的事件数为5时，则有5!=120种排列。这时的偏似然函数分量 l_i 是这120种排列的概率之并，故其计算极为复杂。在SAS PROC PHREG中用TIES=EXACT完成计算任务。

2. Breslow近似法

前面已定义恰在时间 $t_i(i=1,2,\cdots,n)$ 之前的危险集为 $R_i=h_0(t)[\exp(\beta X_i)+\exp(\beta X_{i+1})+\exp(\beta X_{i+2})+\cdots+\exp(\beta X_n)]$。由于在构造偏似然函数时消除了基准风险函数 $h_0(t)$，故可以重新定义相对危险集为：

$$R_i=\exp(\beta X_i)+\exp(\beta X_{i+1})+\exp(\beta X_{i+2})+\cdots+\exp(\beta X_n) \tag{1-8}$$

在 R_i 中有 n_i 例观察对象，持相等时间死亡例数为 d_i。Breslow建议用下列近似的偏似然函数公式求参数估计值：

$$L=\prod_{i=1}^{n}\left[\frac{\exp[\beta(X_{i1}+X_{i2}+\cdots+X_{id_i})]}{\left[\sum_{j\in R_i}\exp(\beta X_j)\right]^{d_i}}\right]^{\delta_i} \tag{1-9}$$

此法的运算速度较快，对失访例数较少资料的近似程度很好。此法在SAS PROC PHREG中为省略。

3. Efron近似法

以 $d_i=3$ 为例，Efron近似法的偏似然函数分量的计算为：

$$l_i=\left(\frac{e^{\beta X_{iA}}}{\frac{3}{3}(e^{\beta X_{iA}}+e^{\beta X_{iB}}+e^{\beta X_{iC}})+\cdots+e^{\beta X_n}}\right)\times\left(\frac{e^{\beta X_{iB}}}{\frac{2}{3}(e^{\beta X_{iA}}+e^{\beta X_{iB}}+e^{\beta X_{iC}})+\cdots+e^{\beta X_n}}\right)\times \tag{1-10}$$
$$\left(\frac{e^{\beta X_{iC}}}{\frac{1}{3}\left(e^{\beta X_{iA}}+\frac{1}{3}e^{\beta X_{iB}}+\frac{1}{3}e^{\beta X_{iC}}\right)+\cdots+e^{\beta X_n}}\right)$$

此法对失访例数较多资料的近似程度优于Breslow近似法。此法在SAS PROC PHREG中的选项为TIES=Efron。

例1-1　26例Ⅲ期浆液性卵巢上皮癌患者经手术治疗后的生存时间资料见表1-1。事件Event为死亡指示变量，死亡记为Event=1，失访记为Event=0。

表 1-1 26例Ⅲ期浆液性卵巢上皮癌患者经手术治疗后的生存时间

病例号 ID	年龄(岁) age	癌细胞分化度 division	淋巴细胞浸润数 lymph	手术残留灶 resid	生存时间(月) months	死亡 event
1	67	2	8.4	2	1.0	1
2	50	3	5.5	2	2.5	1
3	60	1	2.3	2	4.5	1
4	53	2	5.1	2	7.5	1
5	47	2	13.7	2	9.5	1
6	48	3	9.3	2	11.5	1
7	56	3	33.3	2	12.5	1
8	50	3	5.9	2	14.5	1
9	43	3	4.6	2	15.0	1
10	61	2	19.2	2	15.0	1
11	46	3	4.1	2	17.0	1
12	54	3	3.2	2	18.5	1
13	62	3	3.9	0	24.0	1
14	42	3	4.87	2	24.0	1
15	32	2	9.8	2	25.0	1
16	61	3	11.6	1	32.7	1
17	45	2	29.5	2	36.0	1
18	23	2	9.9	0	36.0	1
19	43	3	8.4	0	43.0	1
20	44	2	9.2	2	44.0	1
21	56	3	8.9	1	46.0	1
22	29	2	19.8	2	69.0	1
23	59	1	10.6	2	70.0	1
24	67	1	14.9	0	83.4	1
25	60	1	13.1	0	83.5	1
26	57	2	16.3	0	156.0	0

用精确法配合Cox比例风险模型的SAS程序如下：

```
DATA ovary;
  INPUT ID age division lymph resid months event @@;
DATALINES;
1 67 2 8.4 2 1.0 1
2 50 3 5.5 2 2.5 1
3 60 1 2.3 2 4.5 1
...
25 60 1 13.1 0 83.5 1
26 57 2 16.3 0 156.0 0
;
PROC PHREG DATA = ovary;
  MODEL months * event(0) = age division lymph resid/TIES = EXACT;
RUN;
```