保险精算丛书

生存模型

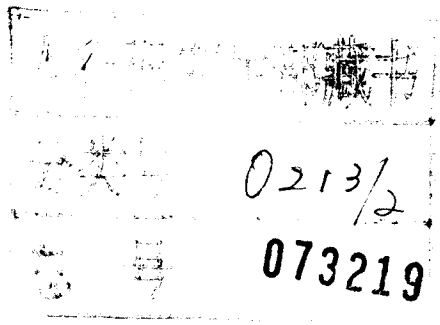
D. 伦敦 著　　陈子毅 译

上海科学技术出版社

《保险精算丛书》

生 存 模 型

Dick London 著
陈 子 毅 译



上海科学技术出版社

内 容 简 介

本书综合描述了生存模型的特征及由样本估计该类模型的统计过程。书中介绍了精算师、人口统计学家和医学统计学家等使用的若干处理方法。本书还提供了许多例子和习题以及大量的文献书目以供进一步的研究学习。

Dick London
Survival Models
and Their Estimation
(Second Edition)
ACTEX Publications
Winsted and Abington, Connecticut

《保险精算丛书》

生 存 模 型

D. 伦敦 著

陈 子 毅 译

上海科学技术出版社出版、发行

（上海瑞金二路 450 号）

新华书店上海发行所经销　　常熟市第四印刷厂印刷

开本 850×1168　1/32　印张 10.75　字数 276 000

1996 年 3 月第 1 版　　1998 年 11 月第 2 次印刷

印数 1 801—3 800

ISBN 7-5323-3941-6/O・199

定价：25.20 元

《保险精算丛书》编委会

《保险精算丛书》前言

保险，作为商品社会中处理风险的一种有效方法，已被全世界所普遍采纳。在现代保险业蓬勃发展的进程中，科学的理论和方法，特别是精确的定量计算，起着十分重要的作用。保险业运营中的一些重要环节，如新险种的设计、保险费率和责任准备金的计算、分保额的确定、养老金等社会保障计划的制定等等，都需要由精算师 (Actuary) 依据精算学 (Actuarial Science) 原理来分析和处理。有鉴于此，许多发达国家都以法律形式规定，保险公司的营业报告必须由精算师签字方为有效。这也是国家对保险业进行调控管理的一种手段。

所谓精算学，实际上是将数学方法应用于金融保险所形成的一套理论体系。它的基础包括精算数学、利息理论、风险理论、人口数学、修匀数学、生存模型和生命表构造等等，还包括一些更专门的内容。这一套理论的重要性和正确性，已经得到国际社会的公认。

在我国，虽然早在 1949 年就由中央人民政府批准成立了中国人民保险公司，但是，由于种种历史原因，在相当长一段时间内我国的保险业发展缓慢，人才培养远不能适应实际需要。特别是精算学的研究和精算人才的培养，未得到应有的重视。在保险业的实际运作中，也很少严格按照精算学的原理办事。这一切都影响了我国保险业的进一步发展及与国际接轨。这种情况已引起保险界、教育界和学术界的注意，正在采取积极措施改变现状。刚刚颁布的《保险法》更明确规定："经营人身保险业务的保险公司，必须聘用经金融监督管理部门认可的精算专业人员，建立精算报告制度。"在此情况下，迫切需要引进国际上先进的精算学

理论，并结合我国的实际加以应用，本丛书就是在这样的背景下翻译出版的。

《保险精算丛书》（第一辑）是由复旦大学数学系、中国人民保险公司上海市分公司（以下简称人保上海分公司）合作翻译的，由上海科学技术出版社出版。全国政协副主席、中科院院士苏步青为丛书题写书名；复旦大学研究生院院长、中科院院士李大潜担任丛书主编；中国人民保险公司上海市分公司总经理何静芝、副总经理钱建中，上海市新闻出版局局长徐福生担任丛书总顾问。上海是我国保险业的发源地之一，历来是保险业的中心。成立于1950年的人保上海分公司，经过45年艰难曲折的发展，业务有了很大开拓，1994年已实现业务收入30亿元人民币，占上海保险市场的80%。根据市场的需要，公司已开办了财产、人身、责任、信用四大类约200多个险种。特别是作为公司主要业务之一的国内人身保险业务，1994年的业务收入已近12亿元。公司所开设的人身险种类也从1982年时的一种，扩展到各种形态的医疗保险、定期和终身保险及责任不同的各种人身意外伤害保险等多个品种，并逐步形成系列化。上海保险市场虽然在不断扩大，但竞争也日趋激烈。特别是一些实力雄厚的国际著名大保险公司的进入，促使国内各保险公司采取有力措施不断提高从业人员的业务素质，包括学习精算知识和培养精算人才。正是由于这样的需要，人保上海分公司决定与复旦大学数学系联手，在上海科学技术出版社的积极支持下，翻译了这套《保险精算丛书》。

复旦大学数学系不仅在数学的基础理论研究方面成就卓著，而且历来重视数学在国民经济中的应用，并取得多项重大研究成果。近年来，他们为了拓宽数学应用的领域，又开辟了精算学研究的新方向，并进行了大量的实际工作。他们在数学系研究生和本科生中开设了有关精算的课程和专题讨论，努力培养精算人才；他们还与各大保险公司合作，从事保险精算实际课题的研究，招收应用数学（保险）大专班，举办面向社会的保险精算培训班，培

训了一批人员参加 A.S.A (北美精算师学会准会员) 资格考试 (该项考试的上海考点就设在复旦大学内), 并于第一期考试中取得通过率超过 90% 的优异成绩。与人保上海分公司合作翻译这套《保险精算丛书》，不仅是复旦数学系理论和实践相结合的一项新的举措，也是他们面向社会培养国家急需的精算人才的重要措施。

"保险精算丛书" (第一辑) 共六本，分别为：

《利息理论》， S.G. 凯利森 著，尚汉冀译；

《风险理论》， N.L. 鲍尔斯 著，郑韫瑜、余跃年译；

《精算数学》， N.L. 鲍尔斯 著，余跃年、郑韫瑜译；

《人口数学》， R.L. 布朗 著，郑培明译；

《修匀数学》， D. 伦敦 著，徐诚浩译；

《生存模型》， D. 伦敦 著，陈子毅译。

所依据的原书均是北美精算师学会 (Society of Actuaries) 为其准会员 (A.S.A) 资格考试所指定的教材和参考书，具有一定的权威性。阅读这套丛书，不论对读者了解和掌握精算学基本原理并应用于保险业实践，还是对读者准备参加 A.S.A 资格考试 (该项考试在中国的北京、上海、天津、长沙等地已设有考点), 均会有很大帮助。

保险精算在我国是一项刚刚起步的新事物，这套丛书是高等院校、保险公司和出版社三方共同合作，编写翻译出版学术水平较高、填补国家缺门的专业书籍的一种有益的探索。我们热诚希望广大读者提出宝贵意见，以利于我们改进工作，做好这套丛书的出版工作，促进保险精算事业在中国的发展。

编者谨识

1995 年 11 月于上海

第一部分　生存模型的性质和特征

本书的主题是生存模型的统计估计和被估模型的分析。

但在我们讲述估计思想前，首先必须深入了解生存模型自身，这便是本书前三章的意图。

第一章介绍了生存模型的一般概念，给出了全书的一个总的概貌。

第二章介绍了生存模型的符号及其意义，给出了若干个分布的例子，这些分布在参数生存模型中将会用到。

第三章介绍了传统表格式生存模型，也就是生命表的性质和特征。这一章着重说明借助于假设的死亡分布而建立的生命表具有第二章中参数模型的类似性质。

第一章　引　论

§1.1　何谓生存模型

生存模型 是一类特殊随机变量的概率分布。

假定一台空调机在室温很高的实验室中运转。空调机开始工作时相当于时间 $t=0$, 一般说来，我们感兴趣的是未来任何时间 t 空调机仍然运转着的概率。我们用 $S(t)$ 记这样一个概率。

我们在这里考虑的随机变量 T 表示一个研究对象从 $t=0$ 到它失效的时间，因此常称为 **失效时间随机变量** 。如果 T 是失效时间，那么在时间 t 该研究对象仍然运行的概率等于失效时间迟于 (在数学上即是大于) t 的概率。也就是

$$S(t)=Pr(T>t). \tag{1.1}$$

由 T 的性质容易知道 $T\geq 0$, $S(0)=1$, $S(t)$ 是非增函数。我们假设 $\lim\limits_{t\to\infty}S(t)=0$ 。

如果 T 是一个研究对象从 $t=0$ 到它失效的时间，那么 T 也是该研究对象从 $t=0$ 开始计算的将来存活时间。许多人比较随意地交替使用 T 的等价定义，即“失效时间”或“将来存活时间”。尽管对读者来说概念一致可能方便些，但对上述概念的交替使用是不会有困难的。

在前面的例子中，有一点是重要的，那就是我们感兴趣的是从发生最初事件的 $t=0$ 开始该制冷机的存活期。本例中，最初事件是空调机在实验条件下的起动。特别应当注意的是，我们并不关心那时候的空调机的实际“年龄”。

作为第二个例子，考虑注射了致癌物质的实验动物的生存研

究。注射致癌剂就是 $t=0$ 的最初事件。我们感兴趣的是从注射致癌剂开始这些动物作为时间函数的生存形式。

在上述两个例子中，我们感兴趣的随机变量是 T, 即失效时间。不论是有生命的还是无生命的研究对象，其在 $t=0$ 时的“年龄”及其在失效时的“年龄”对我们来说都不感兴趣。这甚至是完全不可能知道的。理由是我们关心的是在研究条件下作为时间函数的失效的可能性，而不是研究对象达到的“日历年龄”。正基于这点，我们使用符号 $S(t)$ 。

就绝大部分情况而言，失效 (或死亡) 从最初事件开始作为时间的函数总是适当地表现出来，这当然不是研究对象的“日历年龄”, 正如上述两个例子所表明的，这类情况包括机器设备或实验室动物。另一方面，与人类生存形式相联系的情况，尤其是那些精算师直接感兴趣的情况通常是认可失效 (死亡) 的发生是与其达到的年龄相关连的。这类认可到达的日历年龄的情况将在下一节深入研究。

在某些情况下，到达日历年龄的研究群体的性质较之从初始事件开始的时间的性质显得不很重要。我们确应弄清楚这类情况。例如，考虑已被诊断患某种疾病且已开始治疗的某些人的生存研究。如果患病的那天记为 $t=0$, 以及我们相信健康条件 (和治疗方案) 对生存 (或死亡) 影响程度之大使年龄变得实际上不重要了，那么我们可能感兴趣于测定的那些人的生存期是仅从 $t=0$ 起计的时间函数，在那样的情况下，我们仍用 $S(t)$ 作为生存函数。要注意的是，进入这个研究的不同的人在 $t=0$ 时会有他们各自的初始事件，即许多不同的日历年龄。

§1.2 精算生存模型

§1.1 中的 3 个例子主要引起可靠性工程师，医学统计学家和生物统计学家的兴趣。我们强调这样的思想，即在那些例子中，

研究中的每个个体的实际日历年龄是不重要的，甚至是完全不可能知道的。

对比之下，主要用于保险或养老金方案操作的精算师生存模型却承认这个日历年龄，因为我们相信，这里的生存期是年龄的函数。我们将考察两种形式的精算生存模型。

1.2.1 选择模型

考虑这样一个生存模型，其用于年龄为 x (假设是一整数) 为保险保障而挑选来的人的保险计算。我们看到，保险签约就是前面所说的定义在 $t=0$ 时的初始事件，因而一般地说模型给出了时刻 t 仍然活着的概率。例如，如果我们仍然想用 $S(t)$ 函数的话，那么 $S(10)$ 就给出了在时刻 $t=10$ 时存活的概率 (可能是以年来度量)。我们当然也会赞同当 $t=0$ 时，$x=25$ 与 $x=55$ 会使 $S(10)$ 有不同的值。换句话说，在那样的情况下仅仅 $S(t)$ 是不满足我们需要的。我们需要的 $S(t)$ 还在某种程度上依赖于 $t=0$ 时的 x 值。在那种情形下，我们将使用符号 $S(t;x)$ 。

本文中，选择的年龄 x 称为 *相伴变量* 。选择后，我们仍然视 t 为我们感兴趣的主要变量，但必须在我们的生存模型中以某种方式 $S(t;x)$ 来反映出 x 。

一般性的精算方法只是简单地对每一 x 值有分别的 $S(t)$ 。对每一选择的年龄，生存模型仅被看作为 t 的函数，但在任何给定 x 的情况下使用的合适模型总依赖 x 。

选择的年龄并不是唯一对生存有影响的相伴变量，另一个重要的相伴变量是性别，这个相伴变量也可能被考虑，因其取男性或女性而分离了生存模型。还有一个例子是吸烟者和不吸烟者。如果一个生存模型考虑所有这三个相伴变量，也考虑选择后的主要变量时间 t, 那么我们可以用符号 $S(t;x,m,s)$ 来描述一个选择年龄 x 的男性吸烟者的合适模型。

我们在本书的稍后部分将深入讨论这些有相伴变量的生存模型。现在我们主要是引出这样的思想，即与从初始事件以来的时

间所不同的另一因素能影响生存，在初始事件的年龄 (选择的) 是这一想法的一个明显的例子。

1.2.2 综合模型

接下来我们考虑一特殊情况，即定义在 $t=0$ 时的初始事件是某人的实际出生日，他的生存概率由 $S(t)$ 给出。一般地用 x 表示到达年龄，那样我们看到当 $t=0$ 时 $x=0$, 所以到达年龄与消逝的时间事实上是一致的。于是既可以用 x 也可以用 t 作为主要变量，按通常的惯例是用 x 。生存概率将由 $S(x), x \geq 0$ 给出，而 $S(0)=1$, 当 $x \to \infty$ 时， $S(x) \to 0$ 。

这时，随机变量 X 显然是 *死亡年龄*, 或者说是 *将来寿命随机变量*, 正如 T 是死亡时间 (或失效时间) 或将来寿命随机变量。

显然，$S(x)$ 和 $S(t)$ 实际上是同一函数 (但是不同于 $S(t;x)$) 。因而不论我们是研究 $S(t)$ 还是 $S(x)$ 的数学性质，实际上是同时涉及了它们两者。但是，我们在本书中还是使用两个符号，在使用 x 和使用 t 时有明确的区分。

概括起来说，当我们关注其生存概率的研究对象的日历年龄不是一个重要因素时，那么生存期可视作仅是自某初始事件以来的时间的函数，这时我们使用 $S(t)$ 。在一类特殊情形中，也就是自实际日历诞生日算起的人的生存形式研究中，我们使用 $S(x)$ 。$S(t)$ 和 $S(x)$ 都是单变量函数。在研究 $t=0$ 时年龄为 x 的人的生存形式时，这里到达的年龄被认为会影响未来生存期，故我们用 $S(t;x)$ 。这三种情况是最常见的。

§1.3 生存模型种类

至今我们只涉及到了函数 $S(t)$ (或 $S(x)$) 的概念，并未涉及具体的形式。既然我们谈及 T 是一维的随机变量，那么读者心中可以有理由期望它会有一个特定的分布，像指数分布或正态分布，那是些频繁出现的分布。当生存概率 $S(t)$ 由一个 t 的确定的

函数给出的话，我们称 $S(t)$ 是参数型的。取这个名称是因为 $S(t)$ 的值依赖于一个或若干个参数，当然也依赖于 t 。例如，如果 T 是指数分布的随机变量， $S(t) = e^{-\lambda t}$, 则 λ 就是生存模型的参数。在下一章我们将回顾若干特殊的参数模型。

一般说来，精算生存模型并非可归于参数型的。根据经验数据描绘的 $S(x)$ 的形状太复杂以至不能用简单的一个参数的分布 (如均匀分布或指数分布) 来适当的描述。较好的描述是用双参数的 Gompertz 分布或 Weibull 分布，而用三参数的 Makeham 分布来描述甚至是更好。(这五个分布将在 §2.3 再次提及。)

尽管这些分布作为 *参数生存模型* 有某些应用，然而精算界更倾向于用表格模型，而不是参数模型。表格生存模型中对某一选择的 x, $S(x)$ 的值用数字 (通常是整数) 表出。这就构成一张 *数字表格*, 表格模型之名称由此而来。

仍然有争论，说这数字表格还是可以说成描述了一个参数模型，其中每个数字是一个参数。这种说法有一定的逻辑性，但根据定义，我们还是称它为表格模型。

如果一个表格模型仅对 $x = 0, 1, \cdots$ 给出了 $S(x)$ 的值，那么该模型就无法回答含有 x 小数值的任何问题。我们也许可以说这样的生存模型是不完全的。为了克服这个不足，通常假定 $S(x)$ 的形式是非离散的。当这种称为死亡分布的假定附加给表格模型时， $S(x)$ 就对所有的 x $(x \geq 0)$ 有定义，于是由参数模型而引发的任何计算都能从带分布假定的联合表格模型中求得。

精算表格生存模型的存在已有很长时间了，它被称为 *生命表* 或 *死亡表*, 这显然决定于称呼者是乐观者还是悲观者。鉴于 $S(x)$ 是非离散的，通常可有三种假定：线性 (最为常见), 指数型和双曲型。

因为这种传统的表格生存模型的重要性，我们将在第三章中在非离散的 $S(x)$ 的 3 种通常假定下对模型作很详细的介绍。

精算师在使用我们的选择模型 $S(t; x)$ 时差不多总是用表格形

式。1980 年 Tenenbein 和 Vanderhoof 在他们的文章中已考察了 $S(t;x)$ 的参数形式的发展 [61] 。许多人也按此思路研究 $S(t;x)$, 包括在医学环境下的更为一般的 $S(t;z)$, 此处的 z 是与相伴变量相应的数值向量。(例如可见 Elandt-Johnson 和 Johnson 著作的十三章 [21] 及其它的参考文献。) 我们将在第八章简要地叙述带相伴变量的参数模型的问题。

§1.4 估 计

至今，我们说生存模型是一记为 $S(t)$ (或 $S(x)$) 的函数，它给出了在时刻 t (或年龄 x) 的生存概率，其或者是变量的解析函数，或者是在非离散型假定下的数字表格。我们的意图是在下两章中继续研究生存模型的这两种形式的性质和特征，所以必须清楚地理解它们。我们将继续本书的一项主要工作，即建立可用的具体的模型。我们认为客观存在着一个由 $S(t)$ (或 $S(x)$) 表示的理性、潜在、有效的生存模型，因而我们建立的具体的模型只是那个 "真正的" 有效模型的一个近似，或者说是一个 *估计* 。于是我们说估计 $S(t)$, 且一般地用 $\hat{S}(t)$ 来表示这个估计。

根据样本数据的性质和研究方案设计，各种方法将被用来估计 $S(t)$, 各种 *分布假设* 也将被利用。

重要的是要注意生存模型两种基本形式的 "估计" 的含义。对表格模型，我们通常是估计生存期超过单位区间 (一般是一年) 的条件概率，并由此产生 $S(t)$ 的估计。("条件区间概率" 的含义将在下两章中阐明。) 尽管这方法有少量的变分，它仍是我们得到 $\hat{S}(t)$ 的最通常的途径。表格模型的估计将在第四至七章中讨论。

作为一个对比，对参数生存模型，由估计假设的 $S(t)$ 函数形式的未知参数而产生 $\hat{S}(t)$, 这较之表格模型需估计的量相当地少，我们只要直接从样本数据估计这些参数就行了。常见的生存模型的估计首先给出一条件区间概率估计序列，然后使它们 "符

合”挑选出来的参数形式。在第八章将讨论这两种方法和参数模型的假设检验。

§1.5 研究方案设计

本书中我们将论述医学统计学家、精算师和人口统计学家所信奉的生存模型估计的方法。本节中我们简要地叙述一下他们各自涉及的研究方案的习惯设计。

就绝大部分情况而言，可以说精算师和人口统计学家涉及的是大样本研究，而大多数医学研究基于的是小样本。此外，一般说来精算学和人口统计学研究用的是 *横截面设计*，而许多医学统计研究用的是 *纵剖面设计* 。

1.5.1 横截面研究

在这研究方案设计中，我们首先定义 *研究群体*，即一个可看作是相同人的群体，研究他们的生存形式 (即估计他们潜在有效的生存模型) 。一个城市、国家或民族的全部人口 (人口统计学)，一家人寿保险公司的保单持有人或一个养老金计划的成员 (精算学) 等都是研究群体的例子。

其次是 *观察期* 的选择。这个时期一开始，作为研究群体成员的许多人就置于观察之下。在观察期间，可能有其他人参加到研究群体中来，也会有一些人可能未死亡就退出了研究群体。这样一种在研究期间的进出活动被称为 *迁移* 。通过对数据特别是观察到的死亡数的适当分类和整理，按一定的估计方法，寿命，具体说是死亡，其概率就可由样本数据来估计。这些概率构成了一个暂时的表格生存模型。

这种研究方案设计的主要优点在于估计是建立在一个时间无限的动态群体的新近经验上，在这种群体中一般说来到达死亡的时刻是相当漫长的。

1.5.2 纵剖面研究

与大样本的横截面研究相对照，医学统计学家的研究设计经常是完全不同的。这些研究者不是选择一个时间的区间，观察该区间的死亡数，而是选择一研究群体并纵向地追踪该群体的经历至将来。一般地说，走到死亡的时间如果相当短的话，这项研究设计是可行的。

在许多场合，医学统计学家用的一种设计我们称之为 *群体整体设计*。在这种研究设计中，选择一个研究单位 — 群体，或者说最初的群体，他们可以是点亮的灯泡，注射过药物的鼠，或者经某种医术治疗的人。在所有场合，每个研究单位都指定开始时间为 $t=0$ 。要注意每个研究单位各自的 $t=0$ 时刻可以是同一日历日 (像点亮的灯泡或注射过药物的鼠的研究很可能是这样)，也可以是不同的日期 (对人群来说很可能如此) 。对于后者来说，从初始选择开始，可常规地处理时间，不用顾及 $t=0$ 发生的那个日期 (这一点将在第四章的例子中进一步阐明) 。

一旦选择了初始群体，就保持对其进行观察，直到全体死亡 (失效)，并记录下每个死亡的时间。为了进行这类研究，有一点是清楚的，就是观察者必须有能力控制这个研究群体，从一定意义上说，在未死亡 (失效) 前，该研究单位是不允许有任何成员“失踪”的。鉴于这个理由，这样的研究有时也称为 *控制数据研究* 。由于这类研究的特性，对于健康寿命的大样本研究它们就不可行，这时可使用横截面设计。

有时，为了避免过分长的研究时期，纵剖面研究也可在全体研究对象未死亡之前就终止。这时我们进行的是 *不完整数据* 的研究，而不是 *完整数据* 的研究。通常有两种终止研究的方法。如果研究终止在预先确定的一个固定的日期，那么我们说数据被 *截尾* 了。另一方面，如果研究进行到一个预先确定的观察到死亡的个数而终止，那么我们说数据被 *截断* 了。区分截尾和截断的主要根据在于对估计结果的事后分析。

§1.6 总结和预习

在本章我们为生存模型性质和特性的研究提供了一个舞台，这个舞台同时也提供给本书主题的生存模型估计的研究。这一章的目的是给出生存模型的概念，描述生存模型的两种基本形式，并介绍估计这种模型的简要的倾向性思路。

第二章我们将寻求概念性生存模型的数学性质，并仔细考察若干特殊的参数模型。第三章我们将研究表格生存模型 (生命表)。

第四章至第九章构成了本书的第二部分，将涉及从样本数据估计生存模型的理论。

第四章是在一个完整数据样本的便利的环境下提出生存模型估计的问题，而此后的五章则考虑了不完整数据样本。

第五章讲研究设计分析，以找出我们遇到的估计问题的类型。

第六章讲矩估计，包括传统的精算方法。

第七章讲最大似然估计，包括乘积极限估计。

第八章讲由完整或不完整数据样本估计参数生存模型。还包括那些估计模型的假设检验和考虑相伴变量的简要讨论。

第九章讲从综合人口数据估计表格生存模型的特殊问题，这是人口统计学领域的一个分支。

第十章和第十一章构成了本书的第三部分，论述了生存模型估计在精算中的应用。

第十章讲在大样本精算研究中处理数据的方法。

第十一章最后谈一些各有特色的实际话题。