生命表的构造理论

周江雄　刘建华　黎颖芳　编著

南开大学出版社
中国·天津

图书在版编目(CIP)数据

生命表的构造理论 / 周江雄编著. —天津:南开大学出版社,2001.3
(中国精算师资格考试用书)
ISBN 7-310-01483-9

Ⅰ.生... Ⅱ.周... Ⅲ.寿命表-理论
Ⅳ.R195.3

中国版本图书馆 CIP 数据核字(2000)第 48513 号

出版发行 南开大学出版社
地址:天津市南开区卫津路 94 号
邮编:300071 电话:(022)23508542
出版人 肖占鹏
承 印 天津宝坻第二印刷厂印刷
经 销 全国各地新华书店
版 次 2001 年 3 月第 1 版
印 次 2001 年 3 月第 1 次印刷
开 本 880mm×1230mm 1/32
印 张 12.125
插 页 4
字 数 347 千字
印 数 1—5000
定 价 22.00 元

第一篇

生存模型及其应用

第一章　生存模型及其性质

§1.1　生存模型

1.1.1　生存模型的基本概念

首先我们考察一台在高温实验室运行的空调设备，假设这台设备从时刻 $t=0$ 开始运行，连续运转直至报废（或失效），用 T 表示该设备从时刻 $t=0$ 开始直至报废的时间（或称为失效时间），显然，T 是一个随机变量，我们所关心的是该设备在任意时刻 t（$t\geqslant 0$）仍然运行的概率 $Pr(T>t)$，它是 t 的函数，记为

$$S(t)=Pr(T>t)。\tag{1.1.1}$$

显然有

(1) $T\geqslant 0$；

(2) $S(0)=1$；

(3) $S(t)$ 是 t 的非增函数，且 $\lim\limits_{t\to+\infty} S(t)=0$。

我们称随机变量 T 为报废时间，或称为失效时间，也称为设备从 $t=0$ 开始的“未来寿命”。一般地，总是从某一时刻开始记录某种设备、某类动物、某类人群的“未来寿命”。我们将这一起始时刻记为 $t=0$，开始运行称为初始事件。值得注意的是实际并没有考虑设备在 $t=0$ 之前已使用的时间。下面再讨论一个类似的例子。

考察一群注射了致癌物质的实验动物（如老鼠）的生存状态。注

射在时刻 $t=0$ 进行，观察这些动物的生存时间 T。因为在给定条件下，概率 $Pr(T>t)$ 仅与时间 t 有关，如果不考虑这些动物在 $t=0$ 之前已存活的时间和其他因素对概率的影响，同样可得式(1.1.1)。

上述两个实例都是在给定条件下，不考虑初始时刻 $t=0$ 之前已有年限（或年龄）等因素对生存概率的影响。$S(t)$ 就是我们要详细加以讨论的生存模型。

1.1.2 精算生存模型

1.1.1 节中的两个例子，是机械设备的可靠性问题和临床医学统计学的问题，所研究对象的精确年龄是无足轻重的，甚至可以不知道。然而用于人寿保险和养老金计划的精算生存模型必须考虑所研究对象的精确年龄（自然年龄），这是因为精算师需要研究不同年龄群体的生存概率 $Pr(T>t)$。

首先，考虑一个关于 x 岁（通常 x 取整数）的投保群体的生存模型，像上一节所讲的那样，视签发保单的时刻 $t=0$，我们要讨论的仍然是生存概率 $S(t)=Pr(T>t)$。

众所周知，$x=25$ 与 $x=55$ 所对应的函数 $S(t)$ 显然是不一样的。为了刻画 $Pr(T>t)$ 对 x 的依赖关系，精算师通常将其记为

$$S(t,x)=Pr(T>t)。 \tag{1.1.2}$$

上例中的年龄 x 称为伴随变量，而从选择年龄 x 开始，未来的寿命 t 称为主要变量，这时，精算生存模型为式(1.1.2)。

值得注意的是，年龄不是唯一对生存函数 $S(t)$ 有影响的伴随变量，性别、吸烟与否等因素对未来寿命都有影响，如果考虑这些因素，年龄为 x 岁的男性(m)吸烟者(s)的生存模型为 $S(t,x,\mathrm{m},\mathrm{s})$，其中 x，m，s 均为伴随变量。更一般的生存模型为 $S(t,x_1,x_2,\cdots,x_m)$。其中 x_1，x_2，…，x_m 是对生存有影响的 m 个因素的伴随变量。称这样的模型为选择模型。

下面，再考察一个特殊情况，即在初始事件发生的时刻 $t=0$ 时，$x=0$，根据式(1.1.2)，生存模型为 $S(t,0)$，也可简写为 $S(t)$，我们注意到在时刻 t，被观察者的自然年龄也刚好是 t 岁，在不致引起

混淆的情况下，用 x 代替 t，此时的生存模型为 $S(x)$，对应的随机变量用 X 表示，它表示新生婴儿的死亡年龄(或称为未来生命随机变量)，且 $S(x)=Pr(X>x)$。这里，$S(x)$ 是新生婴儿的生存模型，即所谓的总量模型。

1.1.3 生存模型的其他形式

在实际应用中，常见的是由表格给出的生存模型，称为表格生存模型。但是表格只给出了 $x=0, 1, \cdots$ 处对应的 $S(x)$ 的值，而当 x 不是整数时，就无法得知 $S(x)$ 对应的值。通常又给出在相邻整数间 $S(x)$ 的表达式的假设，诸如常值死力，Balducci，Gompertz，Weibull，Makeham 假设等，当这些假设运用于整个模型时，对所有 $x\geqslant 0$ 的值，$S(x)$ 均可求出。精算表格生存模型已有一百多年的历史，它就是我们熟知的生命表（或死亡表）。

生存模型的另一种形式是含有伴随变量的生存模型 $S(t,x_1,x_2,\cdots,x_m)$。在以后的讨论中，我们经常把某种假设附加于含伴随变量的生存模型，由于这些假设中通常含有待估参数，因此，这样的生存模型又称为含伴随变量的参数模型。

如前所述，我们把生存模型定义为 $S(t)$[或 $S(x)$]，$S(t)=Pr(T>t)$，它是随机变量 T 的概率分布。在以后的章节，将深入地讨论上述生存模型的性质。如果存在一个用 $S(t)$ 表示的可供运作的实用的特殊生存模型，就可建立一个估计或接近于可运作的特殊模型，这一过程称为对 $S(t)$ 的估计，通常用 $\hat{S}(t)$ 表示。我们将依样本数据的性质、研究的目标和选择各种分布假设等不同情况，用各种方法来估计 $S(t)$，这是本书讨论的主要问题。

一般而言，精算师和人口统计学家研究较大的样本空间，我们称之为横向研究。在这一研究过程中，首先确定一个研究团体——对其生存状态进行研究的独立人群，这一研究团体可以是某个城市或国家的人口，或者是某家人寿保险公司的保单持有者，也可以是参与养老金计划的成员等。还要选取一个观察期，在这期限开始的时候，已有许多观察对象是研究团体的成员，并且从一开始就接受观察，另一些

观察对象将在观察期内任意时刻加入该研究团体，同时也可能有一些研究对象在死亡之前退出该研究团体。在观察期内进入和退出的行为称为迁移。通过适当的分类和数据处理，尤其是根据某一指定的程序对所观察的死亡人数进行分类，我们将从这些样本数据中估计出每个年龄的死亡概率。这些概率值就构成了一个过渡性的表格生存模型。

与大样本的横向研究相反，大多数临床研究是建立在小样本基础上的，我们称为纵向研究。它不是选择一个时间区间并观察在该时间区间上的死亡人数，而是选取一个研究群体，并研究这一个群体中的个体由开始时刻直至死亡的状态。当生存的时间较短时，通常采用这种方法。在许多情况下，临床研究将会出现研究整个群体的情况，这个群体可以是注射过致癌物的老鼠；或者是正在进行某种特殊治疗的人群；也可以是等待检验的灯泡样品……在所有情况下，对每一研究的个体均从开始时刻 $t=0$ 开始观察，为了进行此项研究，观察者必须具备不让任何研究个体在死亡之前消失的这种控制能力，因此，这种研究有时又称为可控数据研究。

以上是对生存模型研究的两种思路，我们将会在以后的章节详细地加以阐述。

§1.2 T 的分布函数

在上一节中，已经定义了生存模型 $S(t)=Pr(T>t)$，式中的随机变量 T 表示失效时间（或死亡时间），由随机变量 T 确定的函数 $S(t)$ 也称为生存分布函数，它是概率分布的一种特殊形式，且有 $S(0)=1, S(+\infty)=0$。

在概率论中定义的分布函数

$$F(t)=Pr(T\leqslant t)。\tag{1.2.1}$$

实际上是累积分布函数，显然有

$$F(t)=1-S(t)。\tag{1.2.2}$$

并且 $F(0)=0, F(+\infty)=1$。

但由于讨论生存函数的需要，我们着重讨论$S(t)$。对于连续型随机变量 T，概率密度函数为

$$f(t)=\frac{\mathrm{d}}{\mathrm{d}t}F(t)=-\frac{\mathrm{d}}{\mathrm{d}t}S(t),\quad (t\geqslant 0), \tag{1.2.3}$$

从而有

$$F(t)=\int_0^t f(y)\mathrm{d}y, \tag{1.2.4}$$

$$S(t)=\int_t^{+\infty} f(y)\mathrm{d}y。 \tag{1.2.5}$$

显然有

$$\int_0^{+\infty} f(y)\mathrm{d}y=1。 \tag{1.2.6}$$

概率密度函数 $f(t)$表示开始时刻 $t=0$ 的实体在时间 t 失效（或死亡）的密度，或者称 $f(t)$是在时间 t 的非条件死亡密度。

在生存到时间 t 的条件下，在时间 t 处的瞬间死亡密度称为时间 t 处的危险率（或称为危险率函数），记为$\lambda(t)$。显然$\lambda(t)$是在生存到时间 t 的条件下的死亡密度，从而有

$$\lambda(t)=\frac{f(t)}{S(t)}。 \tag{1.2.7}$$

值得注意的是式(1.2.7)和式(1.2.3)分别为随机变量 T(死亡时间)的危险率函数和概率密度函数，它们都是在时间 t 失效密度的瞬时度量。它们的区别是，$\lambda(t)$是以生存到时间 t 为条件的，而$f(t)$是非条件的（只是给定条件 $t=0$)。

由于$f(t)=-\frac{\mathrm{d}}{\mathrm{d}t}S(t)$，那么

$$\lambda(t)=-\frac{\frac{\mathrm{d}}{\mathrm{d}t}S(t)}{S(t)}=-\frac{\mathrm{d}}{\mathrm{d}t}\ln S(t), \tag{1.2.8}$$

等式两边积分得

$$\int_0^t \lambda(y)\mathrm{d}y=-\ln S(t) \tag{1.2.9}$$

或

$$S(t)=\exp[-\int_0^t \lambda(y)\mathrm{d}y]。 \tag{1.2.10}$$

累积危险函数定义为

$$\Lambda(t)=\int_0^t \lambda(y)\mathrm{d}y=-\ln S(t), \tag{1.2.11}$$

则

$$S(t)=\mathrm{e}^{-\Lambda(t)}。 \tag{1.2.12}$$

在$(0,+\infty)$上$f(t)$可积的条件下，连续随机变量T的一阶矩为

$$E(T)=\int_0^{+\infty} t\cdot f(t)\mathrm{d}t, \tag{1.2.13}$$

分部积分可得

$$E(T)=\int_0^{+\infty} S(t)\mathrm{d}t。 \tag{1.2.14}$$

同样可得T的二阶矩为

$$E(T^2)=\int_0^{+\infty} t^2\cdot f(t)\mathrm{d}t。 \tag{1.2.15}$$

上述积分应存在。

由此T的方差为

$$\mathrm{Var}(T)=E(T^2)-[E(T)]^2。 \tag{1.2.16}$$

如果$Pr(T>y)=Pr(T\leqslant y)=\frac{1}{2}$，则称$y$为随机变量$T$的期中值（简称未来寿命随机变量$T$的中值）。

显然

$$S(y)=F(y)=\frac{1}{2}。 \tag{1.2.17}$$

以上我们考察了随机变量T和T的数字特征以及它们之间的关系。实际上，在用生存分布函数$S(x)(x\geqslant 0)$表示的精算生存模型中，这些特征以及它们之间的关系依然存在，只是使用的符号有所不同罢了，如危险率也称为死力，用μ_x表示，而不是用$\lambda(x)$，即

$$\mu_x=-\frac{\frac{\mathrm{d}}{\mathrm{d}x}S(x)}{S(x)}=-\frac{\mathrm{d}}{\mathrm{d}x}\ln S(x)。 \tag{1.2.18}$$

习惯上用 $\dot{e}_0$ 表示随机变量 X 的一阶矩，即

$$\dot{e}_0 = E(X) = \int_0^{+\infty} x \cdot f(x)\mathrm{d}x。 \tag{1.2.19}$$

因为 $\dot{e}_0$ 是 X 的非条件期望，因此也称为出生婴儿未来寿命的完全期望。

对于选择生存模型 $S(t,x)$，t 为随机变量的值，x 为选择年龄，那么，T 的期望值 $E(T,x)$ 给出了 x 岁的人群的未来预期寿命（或生命期望），用 $\dot{e}_{[x]}$ 表示，它的危险率函数用 $\mu_{[x]+t}$ 表示，并且

$$\mu_{[x]+t} = -\frac{\frac{\mathrm{d}}{\mathrm{d}t}S(t,x)}{S(t,x)} = -\frac{\mathrm{d}}{\mathrm{d}t}\ln S(t,x)。 \tag{1.2.20}$$

显然，以上 X 或 T 的矩都是非条件的。条件矩及其他的条件度量的定义将在§1.4中给出，它们的标准精算符号将在第二章中介绍。

§1.3 参数生存模型举例

本节研究一些可作为生存模型的非负连续型概率分布，在实际应用中，对于给定的经验样本数据，需要分析哪些分布更为合适，因此，对于每一类分布，就要讨论其作为生存模型的适宜性。

1.3.1 均匀分布

均匀分布是仅有两个参数的分布，其概率密度函数为

$$f(t) = \begin{cases} \dfrac{1}{b-a} & t \in [a,b]; \\ 0, & \text{其他}。 \end{cases}$$

如果 $a=0$，b 就是区间的长度，同时也是当 $f(t)>0$ 时 t 的最大值。如果将均匀分布视为一生存模型，常用希腊字母 ω 表示这个参数，则密度函数为

$$f(t) = \begin{cases} \dfrac{1}{\omega} & t \in [0,\omega]; \\ 0, & \text{其他}。 \end{cases} \tag{1.3.1}$$

显然，这时均匀分布有以下性质：

$$F(t)=\int_0^t f(y)\mathrm{d}y=\frac{t}{\omega}, \tag{1.3.2}$$

$$S(t)=1-F(t)=\int_t^{\omega} f(y)\mathrm{d}y=\frac{\omega-t}{\omega}, \tag{1.3.3}$$

$$\lambda(t)=\frac{f(t)}{S(t)}=\frac{1}{\omega-t}, \tag{1.3.4}$$

$$E(T)=\int_0^{\omega} t\cdot f(t)\mathrm{d}t=\frac{\omega}{2}, \tag{1.3.5}$$

$$\mathrm{Var}(T)=E(T^2)-[E(T)]^2=\frac{\omega^2}{12}。 \tag{1.3.6}$$

在时间区间较长的情况下，将生存模型视为均匀分布是不合适的。但不管怎样，这是 Abraham 和 de Moivre 于 1724 年提出的表示生存模型的第一个连续型概率分布。

1.3.2 指数分布

这是一个普遍运用的单参数分布，其生存分布函数为

$$S(t)=\mathrm{e}^{-\lambda t},\quad t\geqslant 0,\lambda>0, \tag{1.3.7}$$

其概率密度函数为

$$f(t)=-\frac{\mathrm{d}}{\mathrm{d}t}S(t)=\lambda\mathrm{e}^{-\lambda t}, \tag{1.3.8}$$

危险率函数为

$$\lambda(t)=\frac{f(t)}{S(t)}=\lambda, \tag{1.3.9}$$

λ 为常数，精算教材中常称指数分布为“常力”分布。

例 1.3.1 对于指数分布，证明

$$E(T)=\frac{1}{\lambda}, \tag{1.3.10}$$

及

$$\mathrm{Var}(T)=\frac{1}{\lambda^2}。 \tag{1.3.11}$$

证 $$E(T)=\int_0^{+\infty} t\cdot f(t)\mathrm{d}t=\int_t^{+\infty}\lambda\mathrm{e}^{-\lambda t}\mathrm{d}t,$$

由分部积分得

$$E(T)=\int_0^{+\infty} \mathrm{e}^{-\lambda t}\mathrm{d}t=\frac{1}{\lambda},$$

同样可得

$$E(T^2)=\int_0^{+\infty} t^2\lambda \mathrm{e}^{-\lambda t}\mathrm{d}t=2\int_0^{+\infty} t\mathrm{e}^{-\lambda t}\mathrm{d}t=\frac{2}{\lambda^2},$$

所以

$$\mathrm{Var}(T)=\frac{2}{\lambda^2}-\frac{1}{\lambda^2}=\frac{1}{\lambda^2}。$$

由于指数分布的危险率为常数，指数分布多用于非生命物体（如机器、电子器件等）的生存模型。与均匀分布类似，它不适于用作长区间段的人类生存模型，但它可以用于短区间段，如一年。这是因为它有简明的数学表达式，在§2.2中将会详细讨论。我们并不希望把均匀分布或指数分布视为人类生存模型，因而用 T，而不是用 X 作为死亡时间随机变量。但在以后介绍的三个分布中，还将用 X 表示死亡随机变量。

1.3.3 Gompertz 分布

Gompertz 于 1825 年提出将该分布视为人类生存模型，该分布的危险率定义为

$$\lambda(x)=B\cdot c^x,\quad x\geqslant 0,B>0,c>1, \tag{1.3.12}$$

那么，生存分布函数为

$$S(x)=\exp[-\int_0^x \lambda(y)\mathrm{d}y]=\exp[\frac{B}{\ln c}(1-c^x)]。 \tag{1.3.13}$$

其概率密度函数为$\lambda(x)\cdot S(x)$，显然其数学表达式不简洁，且分布的期望$E(X)$也不易求得。

1.3.4 Makeham 分布

Makeham 于 1860 年对 Gompertz 分布进行了修正，其危险率函数为

$$\lambda(x)=A+B\cdot c^x,\quad x\geqslant 0,B>0,c>1,A>-B。 \tag{1.3.14}$$

Makeham 分布假定在任意年龄的部分危险与年龄是相互独立的，这样就在 Gompertz 危险率的基础上加上了一个常数 A。

Makeham 分布的生存分布函数为

$$S(x)=\exp[-\int_0^x (A+B\cdot c^y)\mathrm{d}y]=$$

$$\exp[\frac{B}{\ln c}(1-c^x)-Ax]。 \qquad (1.3.15)$$

显然，这个分布的概率密度函数也不容易进行数学处理，同样对概率、矩等数字特征进行计算也是比较困难的。

1.3.5 Weibull 分布

这一分布的危险率为

$$\lambda(x)=k\cdot x^n,\quad x\geqslant 0,k>0,n>0。 \qquad (1.3.16)$$

其生存分布函数为

$$S(x)=\exp(-\int_0^x k\cdot y^n\mathrm{d}y)=\exp(-\frac{kx^{n+1}}{n+1})。 \qquad (1.3.17)$$

1.3.6 其他分布

在非寿险精算中，对于诸如索赔金额等随机变量，其他的概率分布也是有用的，这些分布包括 Γ 分布、χ^2 分布、正态分布、对数正态分布、Pareto 分布等，不过这些分布都不适合本书所考虑的随机变量，但 χ^2 分布在检验假设参数分布与经验数据的拟合程度时很有效。

§1.4 条件概率的数字特征和截尾分布

1.4.1 条件概率和密度

直到目前为止，我们仅仅考察了从年龄 $x=0$ 开始的群体的生存概率，这些用 $F(x)$ 或 $S(x)$ 表示的概率是非条件的，因为在 $x=0$ 时

所有被研究的对象都还活着。下面我们将要考察，在已知$x(x>0)$岁被研究的对象还活着，求从年龄$x(x>0)$岁开始仍然生存的概率。

如果某人已生存到 x 岁，他在 n 年后仍生存的概率为 Pr［活到$(x+n)$岁|已生存到 x 岁］，我们把这一条件概率用${}_np_x$ 表示，那么

$$ {}_np_x=\frac{S(x+n)}{S(x)}, \tag{1.4.1} $$

相对应的死亡概率为

$$ {}_nq_x=1-{}_np_x=\frac{S(x)-S(x+n)}{S(x)}。 \tag{1.4.2} $$

如何区分条件概率${}_np_x$ 与非条件概率 $S(n,x)$是非常重要的。这两种情况都需要求 x 岁的人活到$(x+n)$岁的概率，当根据生存模型 $S(x)$确定概率时，它是有条件的，即${}_np_x=\frac{S(x+n)}{S(x)}$。如果所求概率是由$S(t,x)$确定，那么它是非条件的，直接由$S(n,x)$求出，记为${}_np_{[x]}$，以便和${}_np_x$ 区别开来。

在求 x 岁的人在$(x+n)$岁前死亡的概率时，若由$S(x)$确定，它是有条件的，即${}_nq_x=\frac{S(x)-S(x+n)}{S(x)}$；若由$S(t,x)$确定，那么它是非条件的，直接可用$F(n,x)$求得，用${}_nq_{[x]}$表示。

例 1.4.1 根据$S(t,x)$求出所选取的 x 岁的人活到$(x+10)$岁，并在$(x+20)$岁前死亡的概率。

解 先求活到$(x+10)$岁的人在$(x+20)$岁前死亡的概率，用${}_{10}q_{[x]+10}$表示。于是${}_{10}q_{[x]+10}=1-Pr$［在$(x+20)$岁仍生存|活到$(x+10)$岁后］$=1-{}_{10}q_{[x]+10}$，从而有${}_{10}q_{[x]+10}=1-{}_{10}p_{[x]+10}=1-\frac{S(20,x)}{S(10,x)}$。

下面再来考察 x 岁的人在y $(y>x)$岁死亡的概率密度函数。如果将在生存到 x 岁的条件下发生的概率与该条件密度相乘，则求得非条件概率密度$f(y)$，从而条件密度为

$$ f(y|X>x)=\frac{f(y)}{S(x)}。 \tag{1.4.3} $$

最后我们来考察条件危险率函数（或死力函数），即 x 岁的人在

y ($y>x$)岁时死亡的条件死亡率（或死力）。我们知道危险率函数是以生存到某年龄为条件的（不存在非条件死亡率函数），这样，由于 $y>x$。且 μ_y 以生存到 y 岁为条件，那么就省去了“存活到 x 岁”的条件，于是，我们所指的“条件”死亡率显然就是 μ_y。这一直观结果将在下面得到证明。

1.4.2 X 的下截尾分布

当谈到以生存到 x 岁为条件的生存概率时，需要处理随机变量 X 的样本空间子集的分布，即那些超过 x 的 X 的值，这样的分布称为在 x 处截下尾的 X 的分布，这样一来，条件生存概率 ${}_np_x$ 可表达为

$$
\begin{aligned}
{}_np_x &= Pr(X>x+n \mid X>x)= \\
&\quad S(x+n \mid X>x)。
\end{aligned}
\tag{1.4.4}
$$

即 x 岁生存的人在($x+n$)岁后死亡的概率。显而易见，这个概念与“x 岁的人生存到($x+n$)岁”的概率是相同的。由公式(1.4.1)与(1.4.4),可得

$$
S(x+n \mid X>x)=\frac{S(x+n)}{S(x)}。
\tag{1.4.5}
$$

类似地

$$
\begin{aligned}
{}_nq_x &= Pr(X\leqslant x+n \mid X>x)= \\
&\quad Pr(x<X\leqslant x+n \mid X>x)= \\
&\quad F(x+n \mid X>x)。
\end{aligned}
\tag{1.4.6}
$$

比较式(1.4.2)与(1.4.6)，可以证明

$$
\begin{aligned}
F(x+n \mid X>x) &= \frac{S(x)-S(x+n)}{S(x)}= \\
&\quad \frac{F(x+n)-F(x)}{1-F(x)}。
\end{aligned}
\tag{1.4.7}
$$

式(1.4.5)和(1.4.7)都可用普通的概率关系 $P(A \mid B)P(B)=P(A \cap B)$求得。

下面用$f(y \mid X>x)$表示 x 岁的人在$y(y>x)$岁死亡的条件死亡密度函数，即