生存分析與可靠性引論

陳家鼎 著

安徽教育出版社

SCFXYKKXYL

陳家鼎 著

生存分析與可靠性引論

安徽教育出版社

（皖）新登字03号

生存分析与可靠性引论

安徽教育出版社出版发行

（合肥市金寨路283号）

新华书店经销　　安徽新华印刷厂印刷

*

开本850×1168　1/32　印张10　字数180 000

1993年3月第1版　　1993年7月第1次印刷

印数 1000

ISBN7—5336—1148—9/G·1593

定价：5.30元

序　言

产品寿命的评定与预测是现代可靠性理论的主要研究对象，生物和人的寿命的评估与预测是生物学和医学中重要的研究领域。这两方面的研究都涉及到基础科学和技术科学的许多领域，其中数学方法起着重要的作用。从第二次世界大战以来，这两方面的研究有了长足的发展，形成了许多共同的数学理论和方法，对工程实践和医疗实践有很大的促进意义。

从数学角度来看，上述两方面的研究都是对一个或多个非负随机变量(寿命的数学抽象物)进行统计分析。这种统计分析已形成现代数理统计学的重要分支——生存分析。

生存分析(又称残存分析)就是根据寿命试验或寿命调查得到的数据，对产品或生物(包括人)的寿命进行分析和推断。它含有许多实用的方法和丰富的理论。随着工程实践和医疗实践的发展，不断有新的统计方法出现，应用范围越来越广。

1988年8月，中国科学院系统科学所和北京大学概率统计系在北京大学联合举办全国数理统计研究生学习班。笔者承担了“生存分析”课的讲授工作。在讲稿的基础上，笔者编写了“残存分析”(北京大学油印讲义)，作为北大概率统计专业研究生的教材，教了三届学生。本书是在该讲义的基础上进行较大的修改与扩充而成的。

本书是从实用的角度介绍生存分析与可靠性统计的基本方法，并从数学理论的角度对这些方法的理论根据进行解释与论证。这是为概率统计专业研究生编写的教材。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

限于篇幅只对一部分定理写出严格的证明。读者如对那些未写出的数学证明感兴趣，可查看有关的参考文献。本书还注意介绍生存分析与可靠性统计的现代发展动向，特别叙述了笔者和北京大学寿命与可靠性研究组在不完全数据情形下的最大似然估计和置信限方面的研究成果。本书没有去总结生存分析各方面的进展。例如，不完全数据情形下的假设检验理论(特别是两样本的比较)讲得很不充分。又如贝叶斯方法在书中没有论述。贝叶斯方法无疑是重要的，但先验分布的确定常引起争议，如何在实际工作中使用贝叶斯方法仍是值得研究的课题，笔者对此缺乏研究，没有真知灼见，只好在书中不叙述了。

笔者相信，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“数学证明”，则要求读者预先具备测度论与现代概率论的基础知识，并受过相应的数学训练。

由于笔者水平有限，加上编写时间仓促，本书一定有不少缺点、错误，欢迎专家和读者批评指正。

陈家鼎

于北京大学概率统计系

1992年9月

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(也叫残存分析)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这个或这些非负随机变量进行统计推断。非负随机变量常解释成寿命(技术产品的寿命或生物、人的寿命)。所以，生存分析也就是寿命数据的分析。生存分析对产品可靠性的评定、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、生物学上有广泛的应用价值，日益受到人们的重视。

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。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，也可以是其它的度量单位。如轮胎的寿命用里程，跑20万公里坏了就算寿命是20万公里。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。

在实际工作中常常遇到所谓寿命检验问题。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。

怎样刻划非负随机变量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$S(t)$，称之为可靠性函数或可靠度。当然$F(0-)=0$，$S(0-)=1$.

在实际工作中遇到的“寿命”多是连续型的，即T有分布密度。

刻划T的特征数主要有：

平均寿命μ　　($\mu=ET$，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)：

$$\lambda(t)=\lim_{\Delta t\to 0+}\frac{1}{\Delta t}P(T\leqslant t+\Delta t|T>t)$$

(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：

$$\lambda(t)=\frac{f(t)}{1-F(t)}=\frac{f(t)}{S(t)}\qquad(t\geqslant 0),\qquad(1.1)$$

显然　$\lambda(t)=-S'(t)/S(t)$，于是有

$$S(t)=\exp\left\{-\int_0^t\lambda(u)du\right\}\qquad(1.2)$$

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，

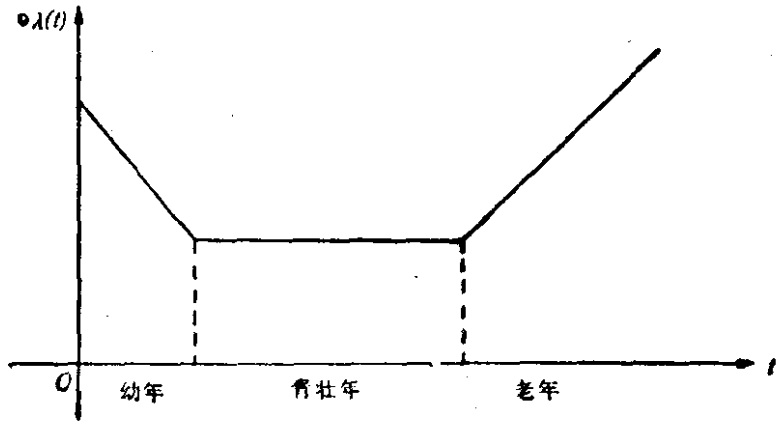

图1.1

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$\lambda(t)$大致如下图(呈浴盆形)。

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)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。

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出终结。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。

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，对几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(注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

[注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(censoring)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(truncation)型数据比较少见。本书只讨论含有删失的数据。

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$S(t)$(或其它有关量)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,才建立起种种科学的数据处理方法。

在工程上有许多数据的获得是基于事先有意识安排好的试验,试验过程受到严格的控制。这时的"删失机制"不言而喻,一目了然。工程上最常见的常规寿命试验有下列四种:

(1) 定时截尾试验 (又叫Ⅰ型截尾试验);

(2) 定数截尾试验 (又叫Ⅱ型截尾试验);

(3) 混合型截尾试验;

(4) 有替换的寿命试验(又分定时的,定数的及更一般的序贯的。)

分别介绍如下.

定时截尾试验:从一个总体(比如一大批产品)中随机抽出几个同时进行寿命试验,试验进行到预先给定的时间T_0为止.这个方案叫做定时截尾方案,简称(n, T_0)型方案。此方案下的寿终数据是$t_1, t_2, \cdots, t_\tau$, $(0\leqslant\tau\leqslant n)$,右删失数据有$n-\tau$个,即$t_{\tau+1}^{+}$

$=\cdots=t_n^+=T_0$. 注意，这里τ是随机变量。

定数截尾试验：从一个总体中随机抽出n个同时进行寿命试验，试验进行到恰好出现第r个寿终时停止。这个方案叫做定数截尾方案，简称(n, r)型方案.$(1\leqslant r\leqslant n)$。此方案$F$的寿终数据有$r$个：$t_1, \cdots, t_r$，$(t_1\leqslant\cdots\leqslant t_r)$，有$n-r$个右删失数据：$t_{r+1}^+=\cdots=t_n^+=t_r$. 注意这种寿命试验的停止时间是随机变量。

混合型截尾试验：从一个总体中随机抽取n个同时进行寿命试验，试验的截止时间ζ是这样规定的：事先给定正数T_0及正整数$r(r\leqslant n)$，若在时刻T_0之前能观测到r个寿终则ζ等于恰好出现第r个寿终的时刻，否则$\zeta=T_0$. 这个方案叫做(n, r, T_0)型方案。显然，(n, r, ∞)型方案就是(n, r)型方案，(n, n, T_0)型方案就是(n_1, T_0)型方案，(n, n, ∞)就是完全寿命试验方案。可见(n, r, T_0)型方案范围很广，它首先由苏联数学家Б.В.ГНЕДЕНКО等人于1965年提出，见[11]. 陈家鼎[12]和涂冬生[28]对这个方案进行了系统的研究。

对(n, r, t_0)型方案，数学上可描述如下. 设n个个体的寿命是$X_1(\omega), X_2(\omega), \cdots, X_n(\omega)$（它们是概率空间$(\Omega, \mathscr{F}, P)$上相互独立同分布的随机变量）. 设它们的次序统计量是$X_1^{(n)}\leqslant X_2^{(n)}\leqslant\cdots\leqslant X_n^{(n)}$.（即将$X_1, \cdots, X_n$按从小到大的次序重新排列）. 试验的停止时间是

$$\zeta=\min(X_r^{(n)}, T_0)$$

有替换的寿命试验：将n个个体同时投入试验，一旦出现个体寿终，则立刻换上新的个体参加试验（在整个试验过程中同时保持有n个个体在试验），试验进行到“一定时刻”为止，这个“一定时刻”可以是一个常数或具有某种特性的“随机变量”。有替换的寿命试验在工程上也是重要的，著名的美国军用标准781

(1987)(我国已引进，见[45])就是采用有替换的试验方案。有替换的试验方案也叫更新型试验方案。这种方案的特性涉及到随机过程的统计推断，我们将在第七章中进行专题讨论。

在上面的讨论中总是把寿命看成是一个随机变量，把全部数据看成是这个总体中一些个体的寿命的观测值(包括删失值)，没有去考虑这些个体的寿命如何依赖于种种自身的或环境的条件。换句话说，我们没有去考察个体的寿命如何依赖于某些参变量(或叫协变量)。但在大量实际工作中，研究寿命对其它因素或变量的依赖关系十分重要。例如在医学研究中考察某种手术后病人的生存时间，这个时间的长短一般与病人的年龄，以及病人作手术前的多项生理指标有关。又如考察玻璃电容器的使用寿命，它依赖于使用温度和电压等。这种对寿命起影响的变量，一般称为协变量(*concomitant variable*)。在技术产品的加速寿命试验中，协变量叫做加速变量或加速因子。

协变量可能不只一个，以下常用$X=(x_1, \cdots, x_p)'$表示协变量(向量)。我们把寿命变量T看成是响应变量(输出变量)，把协变量看成是回归自变量(输入变量)，要找T与X的关系式就是一个回归分析问题。可以设想有：

$$T=\varphi(x_1, \cdots, x_p)+\varepsilon,$$

其中ε是随机项，$E\varepsilon=0$，$\varphi(\cdot)$是回归函数。回归分析的中心问题是找φ，使得ε的方差尽可能的小。这里的φ常有特定的形式，并含有未知的成分，需要我们从“寿命数据”出发加以确定。但寿命数据常含有删失的情况。这种删失数据的回归分析就比通常的回归分析复杂得多，内容也丰富得多，也是生存分析领域里要研究的基本问题。这是现代统计学中方兴未艾的研究领域，本书第五章要进行初步探讨。

上面说过，生存分析与可靠性研究中常遇到删失数据。在进行统计分析时除了对“删失机制”必须有认识外，还应该对寿命T这个总体的已有知识进行收集和概括。和普通统计学一样，生存分析的方法也主要分为两大类。当对寿命总体知之甚少或毫无所知时，采用非参数方法。当总体的分布类型已知，只是若干个(有限个)参数未知时，采用参数性方法。此外还有半参数模型及相应的半参数方法，这是比较新的分支，方兴未艾。

我们的指导思想是：如果被研究的总体能纳入参数模型，又有成熟的参数性方法进行处理，那最好不过了；如果总体不能纳入参数模型，但可纳入半参数模型，又有比较成熟的半参数方法可供使用，则也很好；如果总体很难纳入参数或半参数模型（或者勉强纳入了，但无成熟的方法进行统计分析），则采用非参数方法。

要积极采用参数性方法，就需对总体的分布类型有认识。各种具体问题中碰到的寿命分布一般是互不相同的。但大量实际情况表明，常遇到的寿命分布只有很少几个类型：指数分布、威布尔分布、对数正态分布、Γ分布、广义Γ分布、截尾正态分布等。其中最常见的是前三种。

§2 常见的寿命分布

(一)威布尔(weibull)分布

称随机变量T服从威布尔分布，如果它的分布函数$F(t)$是这样的：

$$F(t)=\begin{cases}0 & t\leqslant 0\\ 1-e^{-(\frac{t}{\eta})^m} & t>0\end{cases} \qquad (2.1)$$

其中$m>0$，$\eta>0$，m叫形状参数，η叫刻度参数。（注2）

易知威布尔分布的密度函数是：

$$f(t)=\begin{cases}0 & t\leqslant 0\\ \frac{m}{\eta}t^{m-1}e^{-(\frac{t}{\eta})^m} & t>0\end{cases}$$

形状参数m是最重要的参数，它的值决定了密度函数曲线的形状。

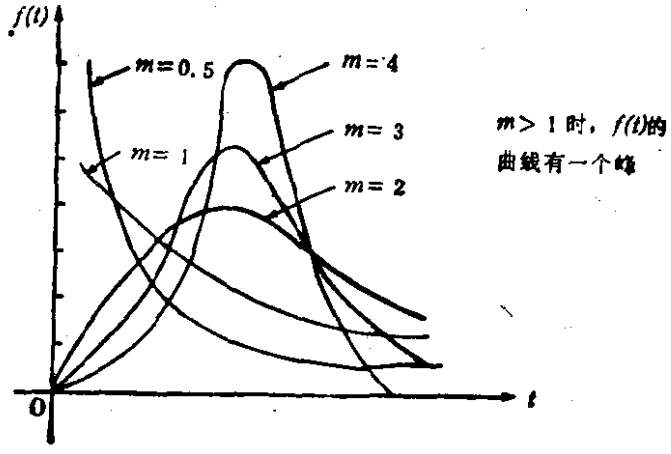

图2.1

刻度参数η的意义见下图，这是$m=2$，η取不同值时$f(t)$的图形。

[注2] 更一般的威布尔分布函数的定义是：

$$F(t)=\begin{cases}0 & t\leqslant r\\ 1-e^{-(\frac{t-r}{\eta})^m} & t>r\end{cases} \qquad (2.2)$$

当$r=0$时，上式变成正文中的(2.1).(2.1)，叫做两参数威布尔分布，(2.2)叫做三参数威布尔分布，后者在应用中比较少见，我们下面讨论威布尔分布就是指两参数情形(2.1)。