卓越大学出版联盟
学术出版精品工程

实用统计技术丛书

实用生存模型：不完全数据分析

APPLIED SURVIVAL MODELS: ANALYSIS OF INCOMPLETE DATA

李元章　何春雄　著

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS

实用统计技术丛书

实用生存模型：不完全数据分析

李元章　何春雄　著

华南理工大学出版社
SOUTH CHINA UNIVERSITY OF TECHNOLOGY PRESS

内 容 提 要

本书介绍生存分析这一重要的统计学分支．针对生存数据的特点，讨论模型建立、参数估计、模型检验以及用SAS软件分析计算的方法和步骤.

本书适于了解概率论与数理统计基础知识和具有使用计算机软件的基本经验的读者阅读，可作为概率论与数理统计专业的研究生或数学专业高年级本科生的选修课参考教材，也适于用数理统计方法从事社会科学、自然科学的研究的人员参考.

图书在版编目(CIP)数据

实用生存模型：不完全数据分析/李元章，何春雄著．—广州：华南理工大学出版社，2015.9

（实用统计技术丛书）

ISBN 978-7-5623-4082-9

Ⅰ.①实… Ⅱ.①李… ②何… Ⅲ.①统计数据-数据模型-统计分析 Ⅳ.①O212 ②TP311.13

中国版本图书馆CIP数据核字(2013)第273784号

实用生存模型：不完全数据分析

李元章 何春雄 著

出 版 人：韩中伟

出版发行：华南理工大学出版社

（广州五山华南理工大学17号楼，邮编510640）

http://www.scutpress.com.cn E-mail:scutc13@scut.edu.cn

营销部电话：020-87113487 87111048（传真）

策划编辑：詹志青

责任编辑：詹志青

印 刷 者：佛山市浩文彩色印刷有限公司

开 本：787mm×1092mm 1/16 **印张**：12 **字数**：300千

版 次：2015年9月第1版 2015年9月第1次印刷

印 数：1～1200册

定 价：28.00元

前　言

生存分析是统计科学的重要分支，在社会科学和自然科学的研究中都有广泛的应用. 生存分析在医疗科学中，称为存活概率分布理论或存活概率分布分析；在机械系统工程中，称为可靠性理论；在经济学或社会学中，称为“持续期分析”或“期限弹性分析”，等等.

生存分析研究中以“事件”和“寿命”为两个重要的变元. 例如，死亡、失败、损坏、生病、解雇、解除合约等均为事件. 寿命则指试验或记录开始到事件发生的时间. 生存分析研究的内容是回答下列问题：对于给定的时间，在目标总体中“死亡”或“失效”的百分率为多少？造成事件发生的原因是什么？如何分辨多因素对事件发生的影响？如何通过因素的控制来减少或增加事件发生的概率？本书就是围绕这些问题的解决而展开讨论的.

本书中虽然以医疗卫生界应用的例子为主来介绍生存分析，但其方法可以普遍应用于其它领域中有关问题的研究. 例如，在社会科学或经济学中，电话公司研究什么时候电话用户转换所使用的电话局，企业集团研究什么时候雇员会转换雇主，市场研究中研究产品如何更新换代，等等. 在生存分析中，我们不仅仅需要时间及事件是否发生的数据，而且还需要收集其它与其有关的变元数据. 例如，企业集团研究雇员转换雇主问题，企业集团的状况、工作的挑战性、市场对雇员的需求、雇员经理的管理水平等可能都会对雇员是否转换雇主有影响. 人事部门可以通过对雇员和雇员经理做调查来收集数据. 通过对这些变元之间关系的研究及统计分析，发现及预测什么样的雇员有可能辞职，或许人事部门可以提出有效方案去改进企业管理和企业政策，以便改进企业工作，避免因失去称职雇员而造成的损失.

本书分为7章. 第1章简要叙述生存分析的研究内容、生存数据的类型以及生存随机变元的各种刻画函数. 第2章介绍生存数据的生存分布函数的经典估计方法，包括K－M估计和区间表估计，以及SAS软件的PROC LIFETEST. 第3章讨论生存函数的比较，主要介绍“风险函数比为常数”假设的检验方法，包括对数序检验、线性序检验和Wilcoxon检验. 第4章介绍生存分析中常用的分布函数类型及其参数的最大似然估计方法，以及SAS中参数估计程序的用法. 第5章讨论生存数据的风险比例模型，在风险函数比与时间无关的前提下，讨论风险函数比与预测变元的关系模型及其参数估计方法和用SAS建立风险比例模型的方法. 第6章介绍用SAS过程LIFEREG进行参数估计和回归

分析的方法．第7章简要讨论重复测量生存数据的处理方法，主要介绍重复测量数据的风险比例模型和一个实例，最后对非风险比例模型(即FRAILTY模型)做简要介绍．

李元章教授曾在美国某研究院、兰州大学、华南理工大学等多地讲授过本书的内容，曾在实际问题的研究中用到本书中涉及的多种模型，并作为华南理工大学的客座教授与何春雄教授合作，在华南理工大学举办实用统计技术系列讲座，本书内容是在讲座“生存数据分析与SAS应用”基础上加工整理而成的．在此我们深深感谢华南理工大学数学学院和研究生院的大力支持，并衷心感谢华南理工大学出版社精诚合作．由于我们水平有限，难免会有错误或不当之处，恳请同行和读者指正．

著　者

2015年9月

目　　录

1 基本概念和模型

生存分析是统计科学的重要分支，在各种领域都有广泛的应用. 比如，对医疗科学中病人的去世、教育行业中学生的中途退学、工业中产品的失效和机器发生故障、保险业中的赔偿、服务行业中的顾客流失、金融业中的银行坏账等事件发生的时间和次数，都可用生存模型来分析. 生存分析在医疗科学中称为存活概率分布理论或存活概率分布分析，在机械系统工程中称为可靠性理论，在经济学或社会学中称为“持续期分析”或“期限弹性分析”，它是衡量利率变动对银行经济价值影响的一种方法.

1.1 生存分析研究的问题

一般地说，生存分析研究“事件”和“寿命”这两个随机变元. 例如，死亡、失败、损坏、生病、解雇、解除合约等均为事件. 寿命则指试验或记录开始到事件发生的时间. 生存分析的目标是回答下列问题：对于给定的时间，在目标总体中“死亡”或“失效”的百分率为多少？(例如，失业率为多少？有百分之几的员工仍然继续工作?)什么是造成事件发生的原因？如何分辨多因素对事件发生的影响？如何通过因素的控制来减少或增加事件发生的概率？

为了回答上述问题，首先需要定义存活时间，在医学领域，死亡是有明确定义的. 但是，如果不是研究生物的死亡或存活，那么必须根据所研究的问题适当地定义什么是“死亡”. 死亡是在生存分析理论中适当定义的事件，或者模糊定义的事件. 生存分析理论中最常用的模型是对于每一个个体，如同医学中的死亡一样，事件仅仅发生一次. 但是，在研究与生存分析有关的课题中，我们会遇到事件重复发生的情形，或者说重复“死亡”的情形. 例如，在研究某种疾病引起的看病或住院、系统设备的维修与保养等问题时，就需要研究事件重复发生的生存分析理论.

生存分析在医学有关问题的研究中有广泛的应用. 生存分析利用统计理论和统计模型的方法，通过实际收集的数据去预测定义好的事件什么时候将会发生；对于一个给定群体，在给定的时间，其死亡率为多少，以及死亡率如何随时间变化。在生存分析中，通常遇到的是离散事件，事件的发生次数通常假设服从二项分布或泊松(Poisson)分布，并用罗吉斯特(Logistic)回归模型或者泊松回归模型进行分析或预测. 例如，对给定的群体，分析事件如何发生，以及到事件发生的时间有多近或多远.

生存分析理论可以应用于很多其它不同的领域. 例如，在社会科学或经济学中，电话公司研究什么时候电话用户转换所使用的电话局，企业集团研究什么时候雇员会转换雇主，市场研究中研究产品如何更新换代，等等. 在生存分析中，我们不仅仅需要收集时间及事件是否发生的数据，而且还需要收集其它与其有关的变元数据. 例如，企业集团研究

雇员转换雇主问题，企业集团的状况、工作的挑战性、市场对雇员的需求、雇员经理的管理水平等可能都会对雇员是否转换雇主有影响．人事部门可以通过对雇员和雇员经理做调查来收集数据．通过对这些变元之间关系的研究及统计分析，发现及预测什么样的雇员有可能辞职，或许人事部门可以提出有效方案去改进企业管理和企业政策，以便改进企业工作，避免因失去称职雇员而造成的损失．由于数据的限制，生存分析在社会科学中的应用不如在医疗卫生领域中应用广泛．但是，我们相信，生存分析一定会在社会科学和经济学中得到广泛应用．本书中虽然以医疗卫生界应用的例子为主来介绍生存分析，但其方法可以普遍应用于其它领域中有关问题的研究.

1.2 生存数据

生存分析的结果是估计患某种疾病的人在患病以后能够存活一定时间的概率．生存时间可以从确诊该疾病的时间或者从实施某种处理措施(如手术或化学疗法)的时间开始算起，或者从使用某一种服务的开始时间(如电话服务、汽车保险服务)算起．生存数据不同于通常的统计数据，需要给以特别的处理，因此需要特别的统计方法和特别的统计术语去描述它的特性.

1.2.1 生存数据中的事件与数据删失

常用生存数据中事件的例子有：

①学院里学生中途不学某一课程；

②大学或高中的学生中途退学；

③公司中职员的离职或被解雇；

④病人旧病复发；

⑤高危病人的病情恶化；

⑥戒烟者重新吸烟；

⑦刑满释放人员重新犯罪；

⑧信用卡失效；

⑨职员从某公司离职；

⑩患者从某种病痛中解脱.

在上述的例①至例⑧中，到事件发生的时间越长越好，例⑩中则是时间越短越好．例⑨中事件发生的时间长短的好坏很难确定，它取决于雇工离开公司的原因，光荣退休、晋升等不应该与开除、因病劝退等同样看待．如果研究因为哮喘病而退役的课题，那么因服务到期、晋升等原因而离开的时间，即使它们很短，也不能视为事件发生的时间，也不应该删除这种数据。生存分析定义这种数据为删失(censore)数据．为了一致起见，统称到某一时间没有事件发生为该个体“存活”，有事件发生为该个体“死亡”．生存分析研究群体在给定时间的存活率或死亡率.

如前所述，在生存分析中要明确所研究的问题中的事件．在有些情况下，也许仅仅对某种原因引起的结果感兴趣．例如，要研究吸烟与肺癌致死的关系，则“死亡”事件就定

义为患肺癌而死亡．此时，需要分析在不同的群体中各有多少人患肺癌而死，群体中从患病到死亡的时间有多长．对于那些因其它原因而死，或迁移到其它地区而失去跟踪的人的数据均视为删失数据．因此，可以看到，除了那些在给定时间内事件发生的数据之外，其它的数据可以有很大的不同．这可能有各种不同的情况，例如：

- 研究的目的是死亡事件是否发生，但目前患者仍健在．
- 一位患者的信息在没有看到痊愈或死亡时已经丢失．
- 其它原因导致所关心的事件不会发生．例如，在研究两种治疗前列腺癌的方案时，我们关心的是病人是否因癌症去世，但患者也许因其它原因(比如车祸)去世．
- 一位患者的信息在我们所关心的事件发生之前已中断．这也许因为医患双方的协议被一方违约或者协议的特殊条款所致．
- 一位雇员离开岗位去读书．

在所有这些情形中，都是不知道事件发生的时刻．如果没有生存分析的知识和方法，研究者也许只是简单地删除这些数据，但这显然丢失了许多有用的信息．在所有这些情形中，我们知道事件发生次数超过某个数目的时刻．例如，一位患者在观测的三年中是存活的，而后因其移居国外而不知他存活了多长时间，但他至少存活了三年，这种个体的数据称为“右删失数据”．如果一个个体在时刻 T 还是存活的，也就是该个体至少存活 T 时间，那么观测时间 T 称为“右删失”的；生存时间也可能是“左删失”的，这种情况发生在一个个体至少已经死亡了多长时间；另外，一个个体的死亡数据也许是“区间删失”的，如果仅仅知道他在某个时间区间内死亡了．

关于删失数据，下面将详细讨论．虽然生存分析中也有大量研究是针对左删失和区间删失数据的，而多数生存分析仅研究右删失数据的处理方法，因为右删失数据最为常见．在 SAS 生存分析的三个程序中有两个用来处理右删失数据，右删失数据在医学研究中最常见．

1.2.2　生存数据中的寿命与数据的完全性

生存数据中的寿命可以用任何时间单位来量度，分钟、小时、天、月、年、公里、周期，以及任何可以比较的与长度或时间类似的量度．由于时间在医疗卫生科学数据中是普遍应用的度量，因此通常利用术语“常态到事件发生时间(time-to-event)”来表述上述所有人或产品的寿命．

在生存分析中，存在多种不同类型的生存数据．如果“死亡”或“失败”已经发生，且其常态到事件发生时间也确实观测到了，这种数据称为“完全”的．除此之外，其它的数据均是不完全的而且具有不同的类型，它们仅仅提供其常态到事件发生时间或人及产品寿命的部分信息．由于数据类型的不同，所使用的统计分析方法也不同．对于那些已经“死亡”的试验对象或者已经失效的产品或系统，我们知道这个试验对象存活了多少时间，这个产品或系统工作了多少年、多少小时．但是，对于在给定时间(例如 100 天)那些仍然存活的对象，我们只知道其存活时间大于 100 天，也许仅仅存活 101 天，也许存活 100 年．对于产品或系统也是如此，如果我们在 1000 小时观测到它们仍然工作，我们仅仅知道它们的寿命大于 1000 小时．它们的寿命也许仅仅 1000 小时零 1 分钟，也许是 2000 小时或更长．此外，在试验中，也许因某种不可抗拒的原因，试验对象停止了试验，或者

无法确定试验开始的时间，这种情况下就无法确定试验对象或者产品的准确寿命．因此，在生存分析中我们要处理不完整或者不确定的数据.

1.2.3 生存数据的类型

收集数据是使用统计分析不可缺少的步骤．统计分析要对数据作统计推断及预测．数据的质量直接影响统计分析的结果及其可靠性．生存数据包含完全数据及不完全数据．因此保证数据质量及选择适当的模型是进行准确统计推断的前提．与通常的统计数据不同，我们不仅仅要考虑数据的随机性和代表性，而且要考虑数据的完全性.

1. 完全数据

样本的完全数据提供了个体的信息，如学生性别、年龄、分数等．在研究到事件发生的时间的课题中，如果事件发生了，也就是说个体死亡了、系统失败了，而且还观测到从试验开始到失败的时间，也就是说生存数据由两个随机变元(T, δ)描述，这里T是从试验开始到失败的时间，δ取1或0．δ取0表示不存活或者失败．在图1－1中，对给出的观测起始时间、终止时间和观测时间区间，个体或产品A、B和D是完全数据，但C不是.

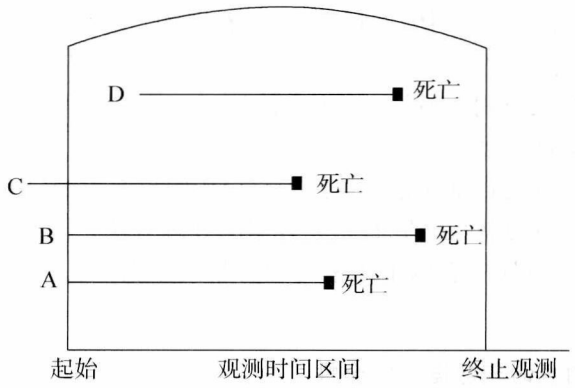

图1－1 生存数据的完全性示意图

2. 不完全数据

在很多情况下，数据是不完全的，例如图1－1中的个体C，我们不知道患者什么时候开始接触某种有害物质或病菌，在我们开始观测时，患者已经不是初诊，这个数据是不完全的．所有不完全数据统称为“删失数据”．通常有左删失、右删失及区间删失等三种删失数据.

(1)右删失数据(right censored data)

生存分析中最常见的是右删失数据，如图1－2所示，个体E一直存活或一直工作直至观测结束；个体F虽然死亡或失去工作状态，但是其事件是在观测结束后发生的；个体G在试验未结束前离开，在离开前未有死亡或失败的事件发生．所有上述三种情况均属于右删失数据．因此，右删失数据可以概括为，如果试验个体在观测终止前(或离开试验前)一直存活或工作，其失败或死亡的发生在观测终止时间(或离开试验)之后，或在观测终止时间没发生失败或死亡．因此，描述右删失数据的数据对为(T, δ)，这里，

$$\begin{cases} T = \min\{\text{观测终止时间，个体离开试验时间}\} - \\ \qquad \max\{\text{个体试验开始时间，观测开始时间}\}; \\ \delta = 1. \end{cases}$$

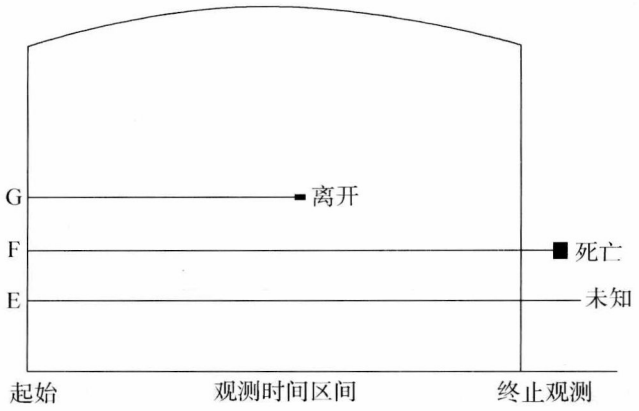

图 1－2　右删失数据示意图

(2)区间删失数据(interval censored data)

区间删失数据只反映出个体在某个时间区间内已经死亡，但不能确定具体的死亡时刻. 这种数据往往来自对研究个体没有实施持续的观测或监控。例如，对 5 个个体每隔 5 小时观测一次，那么观测到的死亡或存活只是在观测时刻的状态。特别地，如果对某个体在第 10 小时观测时仍正常工作，但在第 15 小时观测时已不能正常工作，这说明“死亡”发生在第 10 小时到第 15 小时之间. 换言之，我们只有个体在哪个区间内死亡的信息(参见图 1－3).

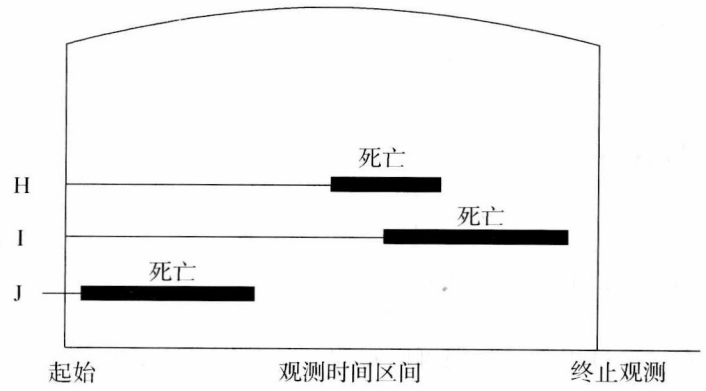

图 1－3　区间删失数据示意图

(3)左删失数据(left censored data)

如图 1－4 中的个体 C，只知道该个体在某时刻前已经失效. 比如，我们知道该个体在第 10 小时之前已经失效，但不知道具体何时失效. 换言之，失效发生在 0～10 小时之间. 这也等同于起始时间为 0 的区间删失数据.

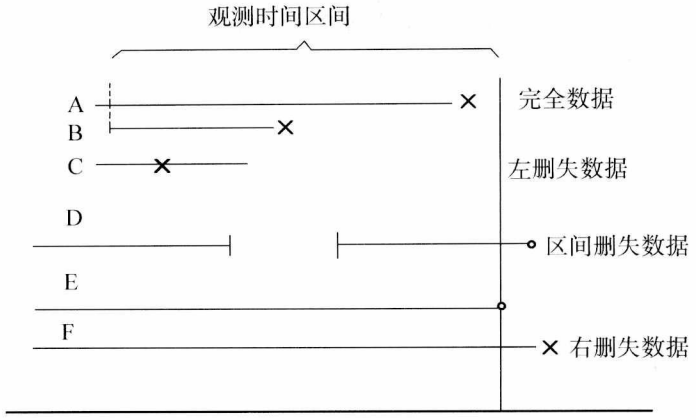

图 1-4 删失数据示意图

SAS 允许在一个数据集中含有以上每种数据．换言之，一个数据集中可以有完全数据、右删失数据、区间删失数据和左删失数据.

1.2.4 生存随机变元的表达方式

首先，对于任何一个生存随机变元都有一个起始点，任何事件发生的寿命都要从这个起点开始计算．假设共有 n 个元件，用 n 个数据对 $(T_1, U_1), (T_2, U_2), \cdots, (T_n, U_n)$ 记它们的生存时间或删失时间，其中 T_i 为第 i 个元件的寿命，U_i 为观测区间的长度，$i=1, 2, \cdots, n$.

对于右删失的情形，定义随机变元

$$X_i = \min\{T_i, U_i\}, \quad i = 1,2,\cdots,n.$$

对于这种情形，在试验过程中，我们可以观测到的数据是 X_i，而不是 T_i 或 U_i，所以定义事件的失效示性函数为

$$\delta_i = \begin{cases} 1 & \text{当 } T_i \leqslant U_i \\ 0 & \text{当 } T_i > U_i \end{cases}$$

和事件的删失示性函数为

$$c_i = \begin{cases} 0 & \text{当 } T_i \leqslant U_i \\ 1 & \text{当 } T_i > U_i \end{cases}.$$

对于左删失的情形，定义随机变元

$$Y_i = \max\{T_i, U_i\}, \quad i = 1,2,\cdots,n.$$

由于是左删失，在试验过程中，我们可以观测到的数据是 Y_i，所以定义事件的失效示性函数为

$$\varepsilon_i = \begin{cases} 1 & \text{当 } U_i \leqslant T_i \\ 0 & \text{当 } U_i > T_i \end{cases}$$

和事件的删失示性函数为

$$d_i = \begin{cases} 0 & \text{当 } U_i \leqslant T_i \\ 1 & \text{当 } U_i > T_i \end{cases}.$$

对于区间删失变元，我们可以观测到的是时间段(L_i,R_i)，如果事件在此区间发生，那么$T_i\in(L_i,R_i)$.

删失随机变元与失效随机变元的独立性

通常我们希望T_i与U_i是相互独立的随机变元，也就是数据是否删失与该事件是否发生无关. 然而，在实际问题中很难做到这一点，此时所建立的模型及估计会产生偏差. 例如，如果我们研究大学生两年的退学率，那么$U_i=2$，对这样的删失数据，删失随机变元与失效随机变元显然是独立的. 但是，如果研究癌症病人对不同药物的存活率，那么通常会有病人因为某种原因退出试验，譬如病情加重或药物反应，在这种情况下，T_i和U_i不再是相互独立的，因为U_i包含了病人存活状况的信息.

关于删失随机变元与失效随机变元的取值情况，经常遇到的情形有：

(1)所有个体的删失数据均相同，并在试验开始前就已知. 例如，研究大学生两年内的退学状况.

(2)所有未失效个体的删失数据均相同，但在试验开始前未知，需在试验中决定. 例如，小白鼠动物实验，在第某个事先决定(如第6只)的小白鼠产生肿瘤后就停止.

(3)所有的U_i都是随机的. 例如，研究某企业职工的年度事故状况.

1.3 生存随机变元的分布及风险函数

1.3.1 生存随机变元的分布

1. 离散分布函数

在概率论中，一个离散概率分布是用一个概率质量函数来刻画的. 于是，如果一个随机变元的概率分布是离散的，则称其为离散型随机变元. 对离散型随机变元X，设其质量函数为f，则有

$$\sum_u P(X=u)=\sum_u f(u)=1. \tag{1.3.1}$$

其中u取遍X所有可能的取值. 于是可假设这样的随机变元仅取有限个或可列多个值. 例如，一群鸟的数量，就可以取$\{0,1,2,\cdots\}$.

2. 分布密度函数

在概率论中，一个概率密度函数(pdf)或一个连续型随机变元的密度函数，是一个描述随机变元取某值附近概率大小的函数. 该随机变元取值在某个区域内概率为其密度函数在该区域上的积分. 设随机变元X有密度函数f，这里f为一个非负 Lebesgue 可积函数，则

$$P(a\leqslant X\leqslant b)=\int_a^b f(x)\mathrm{d}x. \tag{1.3.2}$$

3. 累计分布函数

若随机变元X为离散型的，则其累计分布函数F为

$$F(x)=\sum_{u_j\leqslant x} f(u_j). \tag{1.3.3}$$

若随机变元 X 为连续型的，则其累计分布函数 F 为

$$F(x) = \int_{-\infty}^{x} f(t)\mathrm{d}t. \tag{1.3.4}$$

此时在 f 的连续点 x 处有

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x).$$

直观上，可以认为 $f(x)\mathrm{d}x$ 为随机变元 X 取值落在无穷小区间$[x, x + \mathrm{d}x]$内的概率.

4. 累计生存分布函数

生存函数也称为存活函数或可靠性函数，反映一个随机变元对应一个事件集的性质，通常与系统的死亡或失效的时间相联系. 它刻画该系统生存时间超过某特定时间的概率. 在工程领域中通常称为可靠性函数，而在比较广泛的应用中都称为生存函数(包括人类的存活与死亡). 生存函数也称为累计生存分布函数.

设 T 为一连续型随机变量，在区间$[0, \infty)$上的累计分布函数为 F，则 T 的生存函数或可靠性函数 S 为

$$S(t) = P(T > t) = \int_{t}^{\infty} f(u)\mathrm{d}u = 1 - F(t). \tag{1.3.5}$$

若 T 为离散型的，则其累计生存分布函数 S 为

$$S(t) = \sum_{u_j > t} f(u_j). \tag{1.3.6}$$

1.3.2 风险函数

在生存分析中，一个非常重要的概念就是风险函数. 如果我们把时间分段处理，那么风险函数就是在当前时间段一个仍存活的、处在试验中的个体将要有事件发生的概率密度. 因此，通常这个函数是难以观测到的. 如果一个人在给定的时间段的风险是 2，另一个人的风险为 3，那么可以说，第二个人出“事故”的风险是第一个人的 1.5 倍. 尽管风险函数不能直接观测，但是它是与事件是否发生和何时发生紧密联系的. 风险函数是生存分布分析的基础. 因此，透彻地理解风险函数的意义、特性以及它与生存分布函数的关系，对正确认识和解释生存分析的结果是非常重要的. 风险函数比或其函数，常用来作为回归方程的因变元. 通过对风险函数的回归，估计回归系数并用它来解释不同状态下生存分布的关系，或者我们感兴趣的预测变元对生存分布函数的影响.

1. 风险函数

记随机变元 X 的风险函数为 h，则 $h(t)$ 表示在 t 时刻仍然存活的个体将有事件发生的条件概率密度. 假设我们研究 k 个病人，其发生事件的时间依次为 $t_1 < t_2 < \cdots < t_k$. 每个病人是否有事件发生是相互独立的，因此一个病人从一个区间存活到另一个区间仍然存活的概率等于在每一个区间存活概率的乘积. 在 t_j 存活的概率为 $S(t_j)$，它由在时刻 t_j-的信息 $S(t_j\text{-})$、在 t_j 时刻前夕仍旧存活的个体总数 n_j 以及发生事件(死亡)的个体数目 d_j 所确定.

在时刻 $t_0 = 0$ 时 $S(0) = 1$，因此有

$$S(t_j) = S(t_j\text{-})(n_j - d_j)/n_j. \tag{1.3.7}$$

显然，随机变元 T 的生存函数 S 与风险函数 h 的关系可用式(1.3.8)来刻画，

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leqslant T \leqslant t + \Delta t \mid T \geqslant t)}{\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t \leqslant T \leqslant t + \Delta t)}{\Delta t \times P(T \geqslant t)} \\
&= \frac{f(t)}{S(t)} \\
&= \frac{\mathrm{d}}{\mathrm{d}t}[-\ln(S(t))]. \qquad (1.3.8)
\end{aligned}
$$

风险函数可用来作进一步的统计分析，目前大多用计算机实现. 可以通过计算风险函数来比较不同组的效应，即比较不同组在风险函数上的效应（如不同的药物治疗或控制），该效应用回归模型来估计. 在该回归模型中，将风险函数的对数作为基准风险函数 h_0. Cox 模型是一种医学研究中广泛应用的多变元生存分析模型. 它基于一个基本假设，即风险比为常数，或确切地说，称为 Cox 风险比例模型. 下面将详细讨论.

2. 累计风险函数

设 T 的风险函数为 h，则 T 的累计风险函数定义为

$$
H(t) = \int_0^t h(u)\mathrm{d}u = -\ln(S(t)), \qquad (1.3.9)
$$

因此，累计风险函数 H 与生存函数 S 的关系为

$$
S(t) = \exp(-H(t)). \qquad (1.3.10)
$$

3. 中位生存期(median survival time)

通常，生存分布是非正态、非对称的，且具有右长尾. 在这种情况下，分布的中值要比分布的均值更有意义，因为少数几个长寿命个体对平均存活时间的影响很大. 一种平均存活时间长的药或治疗方法也许仅仅对少数病人效果显著，而对多数人效果不显著.

中位生存期 τ 定义为满足 $S(\tau) = 0.5$的时间.

2 生存数据的非参数估计

在分析完全数据中参数估计是常用的统计方法．本章将讨论如何估计不完全数据分布的非参数估计．由于删失数据存在，我们必须对计算公式进行修正．例如，若4个个体的寿命分别为10小时、20小时、30小时和40小时，那么其平均寿命为25小时．但是，如果它们之中有删失数据，例如10小时是右删失的，那么它的寿命至少为10小时．在这种情况下，4个个体的平均寿命就难以严格计算了，至少大于25小时．

2.1 不完全数据的非参数估计

本节将讨论如何分析右删失数据．首先讨论非参数估计，不假设数据服从任何给定的分布，而通过严格的观测频数估计其概率，然后讨论如何利用SAS进行分析．通常右删失数据包含三类数据：完全数据、试验中间失去的数据以及试验结束时仍然存活的数据．我们不可能也不应该等到所有个体全部失效或死亡后才作数据分析．由于种种原因，我们应该准许试验对象中途退出试验．尽管这些数据是不完全的，它们也包含重要的信息．因此，需要建立同一个模型去描述它们．

右删失数据的概率图

我们先用一个数值例子来说明．

例2.1 假设有5个试验对象，它们的存活时间如表2－1所示，其中，F代表失效，S代表存活．

表2－1 试验对象的存活时间

数据位置	生存状态	存活时间/小时
1	F_1	31
2	S_1	65
3	F_2	150
4	S_2	220
5	F_3	300

我们在研究这些数据时，感兴趣的是失效对象在序列中的排列(按可能的失效时间，从早至晚排序)，然后给出失效序号．显然，试验对象1是已经失效的，它到失效的时间仅为31小时，是所有失效时间中最小的，它的位置是不可以变动的．因此，它的失效序号为1，即F_1的序号为1.