临床随访资料的统计分析方法

余松林　编著

人民卫生出版社

临床随访资料的统计分析方法

余松林　编著

人民卫生出版社

临床随访资料的统计分析方法

余松林　编著

人民卫生出版社出版

（北京市崇文区天坛西里10号）

河北省遵化人民印刷厂印刷

新华书店北京发行所发行

787×1092毫米16开本　10½印张　240千字

1991年2月第1版　　1991年2月第1版第1次印刷

印数：00,001—3,640

ISBN 7-117-01428-8/R·1429　定价：5.50元

〔科技新书目229—472〕

前　言

病人存活时间的长短是临床随访研究工作者关心的主要课题。经典的生存率计算方法在随访研究中应用已久。近代，随着统计理论的发展与计算技术的进步，有关病人存活时间的统计分析方法也日趋完善。本书系统地介绍了这一领域的研究成果。其中包括非参数及参数法生存率计算，生存率之间的比较，生存时间与影响因素之间的回归分析方法，截尾时间的处理方法等。

本书的主要目的是为临床医师提供一套较完整的统计分析技术，使他们在本书的指导下能充分利用、系统分析和合理解释所取得的随访资料。因此，本书在写作上力图实用，对每一种方法都首先介绍基本原理，再结合实例阐明其应用，最后介绍计算方法。对可以用手工完成的计算过程，详细介绍了计算过程，对比较复杂的计算，也列出了计算公式和计算步骤，供有这方面兴趣的读者参考。

除了临床医学之外，在基础医学和预防医学中也存在时间问题，本书可对这方面的研究工作提供统计方法方面的帮助。本书还可作为医学研究生的统计方法教材。

由于本人的水平有限，书中的错误和缺点在所难免，恳切地希望读者批评指正。

在本书的写作过程中，受到了不少同仁的鼓励，得到了周有尚教授、陈世蓉副教授、刘筱娴副教授的热情支持与指导，得到了尹平、王松、严艾荣、贾桂珍、欧阳宁慧等同志的大力帮助，特别是董明同志为本书的出版付出了辛勤劳动。在此一并致以衷心感谢。

余松林

于同济医科大学

1990年5月15日

目　录

第一章 绪 言

在临床随访研究中，我们着重收集病人出现某种结果(痊愈、复发、失败或死亡等)所经历的时间，以便比较不同治疗措施的远期效应的优劣。我们把病人出现某种结果所经历的这种时间统称为生存时间。在这类研究中，有时我们还收集病人的一些有关因素。以分析哪些因素对延长生存时间有利，哪些因素对延长生存时间不利。

生存时间的统计分析方法起源于19世纪对寿命表的分析。到20世纪初，这一技术已应用于工程学中。在第二次世界大战期间，由于对武器的可靠性的要求，使这一分析方法得到了很大发展，并不断扩展应用到其他研究领域中。近30年来，在临床研究，特别是在不断开展的临床随访研究中，也引进了生存时间分析方法，用来分析病人的随访资料。由于临床研究资料的复杂性，反过来又进一步推动了生存时间分析技术的发展。到目前为止，生存时间分析方法作为统计学的一个分支，已形成了一套完整的体系，包括参数法、非参数法以及回归分析方法等。

本书以临床随访研究中取得的生存时间资料为代表，讨论这类资料的统计分析方法，并用多种回归模型分析影响生存时间长短的因素。

第一节 生存时间

从狭义的角度来说，生存时间（survival time）是患某种疾病的病人从发病到死亡所经历的时间跨度。从广义的角度来说，可以把生存时间定义为从某种起始事件到达某种终点事件所经历的时间跨度。生存时间资料在医学及生物学研究中是很常见的。例如，临床上急性白血病病人从治疗开始到复发为止之间的缓解期；冠心病病人在两次发作之间的时间间隔；已作输卵管结扎的妇女行输卵管吻合术后至受孕的时间间隔；在流行病学研究中，从开始接触危险因素到发病所经历的时间等。从上述例子可以看出，生存时间不仅仅指病人从发病开始到死亡所经历的时间跨度，也可以是任何两个事件之间的时间间隔。因此，在计算生存时间时，要有明确规定的时间起点和终点以及关于时间的测度单位（如小时、日、月、年等作为测度单位），否则就不便于分析和比较。例如，我们可以把下列事件作为起始事件和终点事件：

起始事件	终点事件
疾病确诊	死亡
治疗开始	痊愈
症状缓解	疾病恶化
接触毒物	出现毒性反应
接触危险因素	发病

为方便起见，在进行一般讨论时，我们以疾病确诊作为起始事件，而把死亡作为终点事件。

在实际工作中，我们经常遇到两类生存时间数据，下面分别加以说明。

一、完全数据

在随访工作中，当观察到了某病人的明确结局时，该病人所提供的关于生存时间的信息是完整的。我们把达到了明确结局的病人的生存时间数据称为完全数据，并用符号 t 表示完全数据。由完全数据组成的样本资料称为完全资料（complete data）。

二、截尾数据

在随访工作中，由于某种原因未能观察到病人的明确结局，所以，不知道该病人的确切生存时间，就像该病人的生存时间在未到达规定的终点之前就被截尾了，因此称之为截尾数据。虽然截尾数据所提供的关于生存时间的信息是不完全的，不知其真正能生存多长时间，但是这类数据还是提供了部分信息。它告诉我们该病人至少在已经经历的时间长度内没有死亡，其真实的生存时间只能长于我们现在观察到的时间而不会短于这个时间。我们用符号 t^+ 表示截尾数据。

产生截尾现象的原因大致有以下几个方面：

1. 病人失访。由于搬迁而失去联系，或由于其他原因死亡而未能观察到规定的终点。

2. 病人的生存期超出了研究的终止期。例如，研究计划规定只对病人随访 5 年，但有的病人的生存期超过了 5 年；或者由于病人进入研究的时间较晚，虽然对他的随访期未满 5 年，但已到了研究的截止时间。

3. 在动物实验中，有时预先规定观察期限。虽然有一部分动物在到达实验终止日期时尚未出现规定的终止事件，但仍停止实验；或者当出现了预先规定的终止事件的动物数后实验也停止。这一部分残存动物的生存时间就是截尾数据。包含截尾数据的样本资料称为截尾资料（censored data）或不完全资料（uncomplete data）。

我们用图解法来说明临床试验中常见的关于生存时间的完全数据与截尾数据。图 1.1 表示 4 名病人依次进入观察并退出试验的情况。试验预定的观察期为 5 年。第 1 号病人在 1981 年 1 月 1 日确诊而进入观察，到 1982 年 12 月 31 日死亡，生存时间为 2 年，系完全数据，记 $t_1 = 2$ 年。第 2 号病人在 1982 年 1 月 1 日确诊而进入观察，到 1985 年底观察截止时仍未死亡，故这名病人的生存时间为 4 年，系截尾数据，记为 $t_2^+ = 4$ 年。类似地，第 3 号病人的生存期为 3 年，系完全数据，记为 $t_3 = 3$ 年；第 4 号病人的生存期为 1 年，系截尾数据，记为 $t_4^+ = 1$ 年。

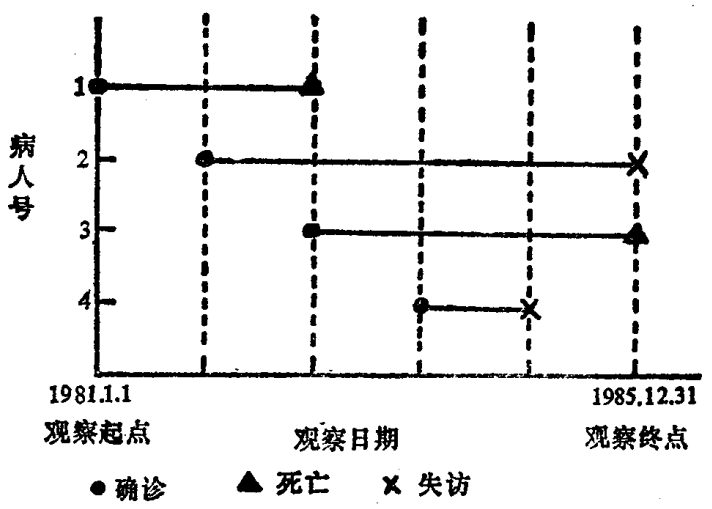


图 1.1 临床随访研究中的完全数据与截尾数据

三、生存时间资料的整理

设总共观察了 n 例病人的生存时间，记第 i 名病人的生存时间为 t_i，则全部 n 例病人的生存时间可以记为

$t_1, t_2, \cdots, t_n$

如图 1.1 中的生存时间（年）按病人号的顺序排列为

$2, 4^+, 3, 1^+$

如果将这些生存时间按由小到大的顺序排列，则得到一个有序的生存时间序列为

$$t_{(1)} \leqslant t_{(2)} \leqslant \cdots \leqslant t_{(n)}$$

图 1.1 中的生存时间有序系列为：$1^+, 2, 3, 4^+$。下面介绍几个不同类型的生存时间的例子。

例 1.1　20 名行输卵管结扎术的妇女经峡部-峡部吻合术后的受孕时间（月）为：1，1，2，3，3，4，4，4，6，6，8，9，9，10，11，12，13，15，17，18。

此例中的生存时间均为完全数据，并已按由小到大的顺序整理。

例 1.2　23 名行输卵管结扎术的妇女经壶腹部-壶腹部吻合术后的受孕时间（月）为：1，3，5，5，5，6，6，6，7，8，10，10，14^+，17，19^+，20^+，22^+，26^+，31^+，34，34^+，44，59。

此例为有序生存时间资料，并包括有截尾数据。可在随访期内任何时点上发生的截尾类型称随机截尾（random censoring）。本例中的截尾就是随机截尾。

例 1.3　15 只雌性大白鼠接触毒物 DHG 后观察 12 周，其生存时间为：4，6，8，9，9，10，10，10，11，12，12，12^+，12^+，12^+，12^+。

此例是一个定时截尾数据的资料，所有到第 12 周未死亡的动物的生存时间都属于截尾数据。这种类型的截尾又称为右截尾（right censoring）。在某些情况下，对右截尾资料的统计处理要比对随机截尾资料的统计处理容易一些。

当观察的病例数较多时，可按一定的时间区间分段整理。设在第 i 个时间区间 $[t_i, t_{i+1})$ 开始时的病人数为 n_i，在此区间内的死亡数为 d_i，则在第 $i+1$ 个区间开始时的病人数为 $n_{i+1} = n_i - d_i$。分组资料的整理情况见例 1.4 中表 1.1 的第(1)至第(3)列。

例 1.4　169 名急性白血病病人的缓解期（年）资料经分组整理列于表 1.1 中。

表1.1　169名急性白血病病人的分组资料及生存函数计算

生存时间(年) [起*—止] t_i—	期初观察病人数 n_i	期内死亡例数 d_i	概率密度函数 $\hat{f}(t_i)$	生存概率(生存率) $S(t_i)$	风险函数 $h(t_i)$
(1)	(2)	(3)	(4)	(5)	(6)
0—	169	108	0.639	0.361	0.639
1—	61	35	0.207	0.154	0.574
2—	26	14	0.083	0.071	0.538
3—	12	5	0.030	0.041	0.417
4—	7	3	0.018	0.024	0.429
5—	4	2	0.012	0.012	0.500
6—	2	1	0.006	0.006	0.500
7—8	1	1	0.006	0.000	1.000

* 以就诊时间作为起点。

第二节　描述生存时间分布规律的函数

生存时间资料看来比较简单，但是却不适合于用一般的正态分布规律来进行描述。这是因为第一：生存时间资料的分布往往不呈正态分布，而是呈偏态分布；第二：对截尾数据的处理比较复杂。故近代发展了一套系统的关于生存时间资料的统计分析方法，以便能正确概括生存时间资料的特点。

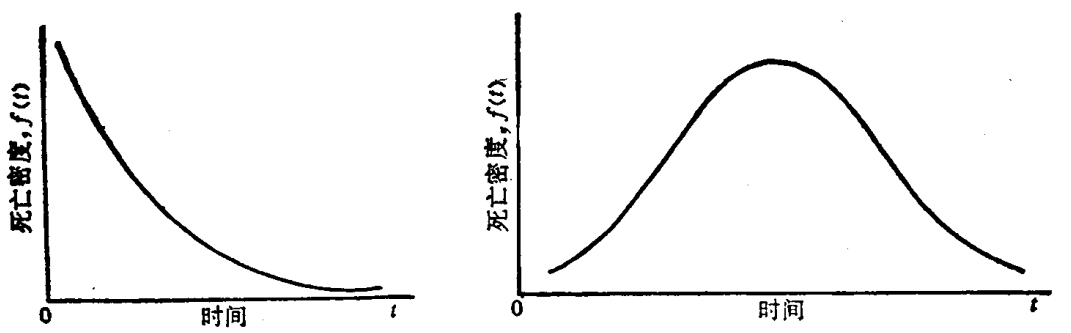
概率密度函数、死亡概率、生存概率和风险函数通常是用来描述生存时间分布规律的四个函数，下面对这几种函数的性质、意义以及它们之间的关系分别加以介绍。

一、概率密度函数

概率密度函数〔probability density function,f(t)〕简称为密度函数。其定义是：一个病人死于从时间 t 到 $t+\Delta t$ 这一小区间内的概率极限。这一函数表示死亡速率的快慢。密度函数的数学表达式为

$$f(t)=\lim_{\Delta t\to 0}\frac{\text{一个病人在区间 }(t,\ t+\Delta t)\text{ 内死亡的概率}}{\Delta t} \qquad (1.1)$$

上式中的Δt是指一段很小的时间区间，$\lim_{\Delta t\to 0}$表示当Δt 小到接近于 0 时的极限。以时间 t 为横轴，密度函数 $f(t)$ 为纵轴所绘制的曲线叫密度曲线。图 1.2 绘出了两种形态的密度曲线。



(1) 死亡速率随时间延长而下降　　(2) 山峰形的死亡速率

图 1.2　两种不同的密度曲线示意图

从密度曲线可以看出不同时间的死亡速率以及死亡高峰时间。例如图 1.2 (1)的曲线表示早期的死亡速率很高，以后随时间延长而下降。图 1.2 (2)的曲线表示死亡速率高峰出现在时间中段，呈现山峰形。在高峰处的死亡人数最多。每种疾病都有其特殊的密度曲线。故密度函数表示在时间 t 上死亡发生的严重程度。

当观察例数较多时，常将生存时间划分为若干个时间区间，每一时间区间可包括一个或多个单位时间长度。此时概率密度函数是用死于单位时间长度内的病人数与观察总病人数之比来估计：

$$\hat{f}(t)=\frac{\text{从时间 }t\text{ 开始的一个时间区间内的死亡病人数}}{\text{观察病人总数}\times\text{该时间区间包含的单位时间数}} \qquad (1.2)$$

对表 1.1 中的资料用式（1.2）计算的各时间区间内的概率密度函数列于该表第(4)

列中。例如$\hat{f}(0)=108/169=0.639$，$\hat{f}(1)=35/169=0.207$，等等。用各时间区间内的中点与该区间内的概率密度函数绘制的多边图见图1.3。

从图1.3可以看出该169名急性白血病人的死亡速率高峰出现在就诊后1年以内，其后逐渐下降。

概率密度函数有两个特点：

1. $f(t)\geqslant 0$。即概率密度函数没有负值。这可以从式（1.2）的计算中反映出来。当在某一时间区间内没有死亡病人时其分子取0，故函数值为0。

2. 在概率密度函数曲线与横轴t之间包含的总面积为1，即从$t=0$到$t=\infty$之间的积分值为1。这一积分公式可写为

$$\int_0^{\infty} f(T)dT=1 \qquad (1.3)$$

式中T为生存时间的积分变量。

这一特点也可从式（1.2）中反映出来。因为分母是观察病人数，分子是各区间内的死亡人数，将全部死亡人数相加必然等于总观察病人数，所以所辖的面积之和为1。

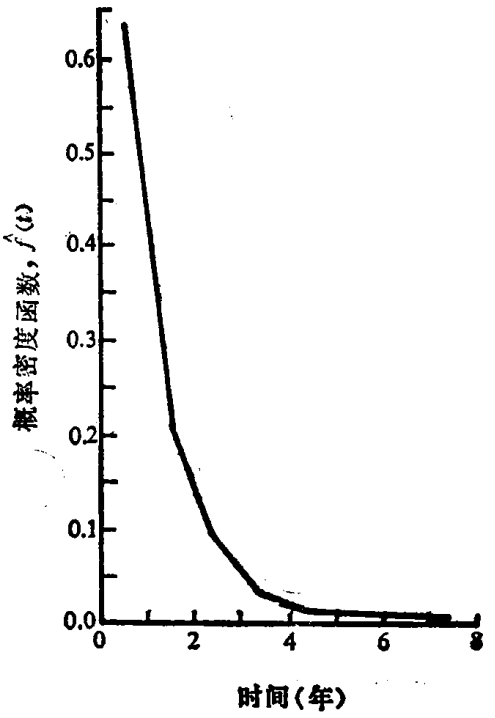


图1.3　表1.1资料的概率密度函数多边图

二、死亡概率

定义为：

$$F(t)=\int_0^{t} f(T)dT \qquad (1.4)$$

死亡概率又称概率函数〔probability function, F(t)〕或累积死亡率，或简称为死亡率。这一函数表示一个病人从开始观察起到时间t为止的死亡概率。它是一个随时间而上升的函数，当t趋向于∞时，死亡概率则趋近于1，此表示该个体最终必然死亡。

三、生存概率

生存概率又称生存函数〔survival function, S(t)〕，表示一个病人的生存时间长于t的概率，故又简称生存率。其意义正好和死亡率相反。它的计算公式为

$$S(t)=\int_t^{\infty} f(T)dT=1-F_{(t)} \qquad (1.5)$$

在实际工作中生存率是用生存时间长于t的病人数对总病人数的比例来估计的，为

$$\hat{S}(t)=\frac{\text{生存时间长于 } t \text{ 的病人数}}{\text{观察病人总数}} \qquad (1.6)$$

生存概率的特点是：

1. 观察起点即$t=0$时的生存率为$S(0)=1$。

2. 当观察期无限长（$t \to \infty$）时的生存率为0，即$S(\infty) = 0$。

对于表1.1中的资料用式（1.6）计算的生存率 $\hat{S}(t)$ 列于该表中第(5)列。例如 $\hat{S}(t > 1) = (169 - 108)/169 = 0.361$，$\hat{S}(t > 2) = (169 - 108 - 35)/169 = 0.154$，等等。

以时间 t 为横轴，生存率$S(t)$ 为纵轴绘制生存率曲线图。例如用表1.1中第(5)列的估计生存率绘制的生存率曲线见图1.4。

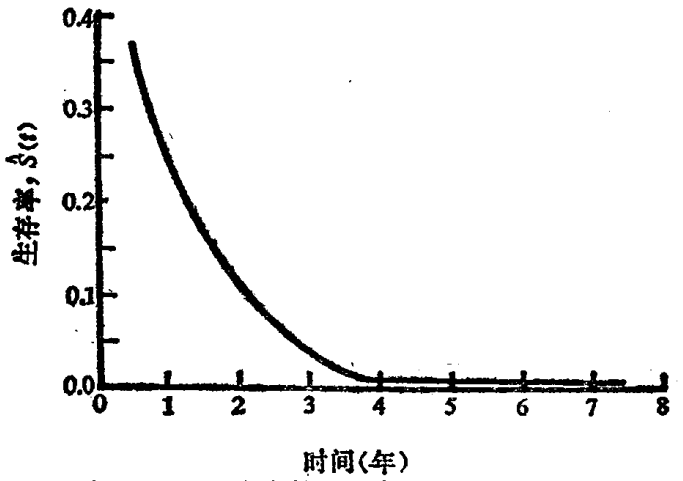


图 1.4　169名急性白血病病人的生存率曲线

生存率曲线是一条下降的曲线。下降的坡度越陡表示生存率越低或生存时间越短。曲线变化的趋势是否接近直线形态表示死亡速率变化的稳定程度。直线程度越高表示各时期的死亡速率越稳定，曲线的坡度变化越多表示各时期的死亡速率越不稳定。从图1.4可以看出，该组白血病病人在头3年内的死亡速率很高，以后则趋于平稳。

利用生存函数可以计算中位生存时间，即生存率为0.50时的时间$t_{(0.50)}$。由于一般生存时间资料呈偏态分布，用算术平均数作为代表值不够理想。而中位数则能较好地反映偏态分布资料的集中趋势。也可用其他百分位数来描述一组生存时间资料的分布范围。

四、风险函数

风险函数〔hazard function, h(t)〕表示一个生存到时间 t 的病人，在从 t 到 $t + \Delta t$这一非常小的区间内死亡的概率极限。计算公式为

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{在时间 } t \text{ 生存的病人死于区间}(t + \Delta t)\text{的概率}}{\Delta t} \qquad (1.7)$$

因为计算这一函数时，用到了生存到时间 t 这一条件，故又称为一个生存到时间 t 的病人在时间 t 的瞬时死亡率或条件死亡速率。其他术语还有条件死亡密度函数、死亡力或年龄别死亡速率等，它表示老化或耗损过程，在生存分析中起有重要作用。

实际工作中，$h(t)$可用在时间区间〔t_i，t_{i+1}）内的死亡人数对该区间开始时的病人数之比来估计（用单位时间长度来表示），即

$$\hat{h}(t) = \frac{\text{死于区间 } t_i \text{至 } t_{i+1} \text{内的病人数}}{\text{在 } t_i \text{生存的病人数} \times \text{该区间包含的单位时间数}} \qquad (1.8)$$

当时间区间长度按1单位（1年、1月、1周等）分组时，则有 $t_{i+1} - t_i = 1$。此处 t_i为区间 i 的起始时点，t_{i+1} 为终止时点（也是第 $i+1$ 区间的起始时点）。

风险函数随时间延长可以表现为递增、递减或其他种波动形式。当风险函数为常数时，表示没有随时间而加速死亡的情况。如果$h(t)$随时间而上升，则表示条件死亡速率随时间而加速。反之若$h(t)$随时间而下降，则表示条件死亡速率随时间而减速。

图1.5绘出了5种不同的风险函数曲线。其中$h_1(t)$是一种上升曲线，表示有随年龄而加速的趋势，老年人的年龄别死亡率就属于这种类型。$h_2(t)$是一条下降曲线，表示有随年龄而减速的趋势，幼儿期的年龄别死亡率就属于这种类型。$h_3(t)$是一种平稳型没有随年龄增加而加速死亡的情况。人口从少年直到中年阶段的年龄别死亡率保持在低水平上就属于这种类型。$h_4(t)$是一种盘形曲线，可以把$h_1(t)$、$h_2(t)$及$h_3(t)$看成为它的阶段曲线。人口的寿命曲线就属于这种类型。$h_5(t)$是另一类型的曲线。很多疾病在不同发展阶段上的死亡率变化特点都可以用$h_5(t)$型的曲线来描述。如急性白血病病人经过治疗后维持一段时间的低复发率，继之复发率上升，但有少数人的缓解期很长，病情稳定。因此通过了复发率高峰时区后，其复发率反而较低。麻疹病人在发病一周左右的死亡率最高也属于这种类型。

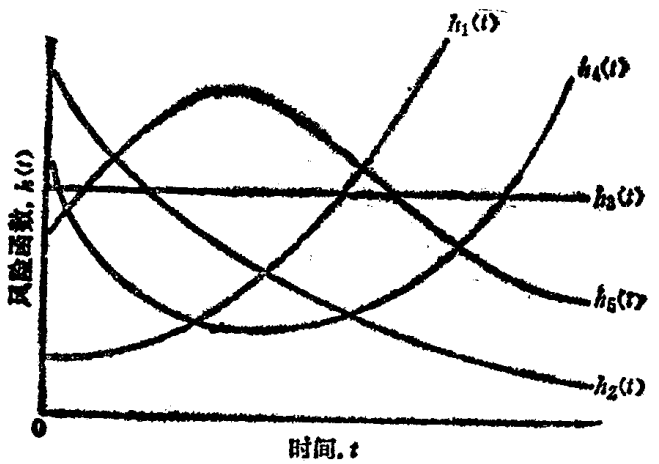


图1.5 不同风险函数曲线示意图

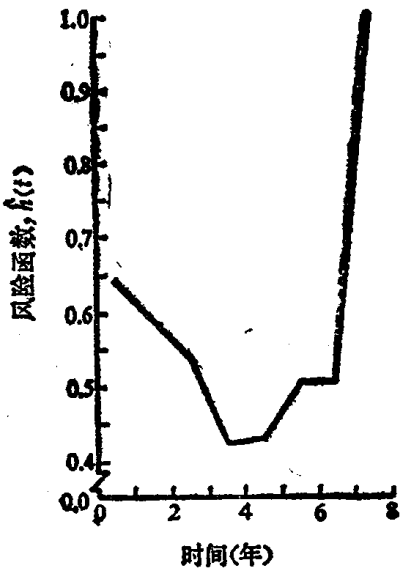


图1.6 169名白血病病人的风险函数曲线

例1.4资料用式（1.8）计算的风险函数值列于表1.1中第(6)列。用此计算结果绘制的风险函数曲线见图1.6。

最后可以把概率密度函数、生存概率及风险函数三者之间的关系表示为

$$h(t)=f(t)/S(t) \tag{1.9}$$

这三个函数分别从不同的角度描述了生存时间资料的分布特点。

第三节 生存时间分布函数的进一步讨论*

一、生存时间T为连续的情况

用T表示某个病人的生存时间，其取值等于或大于0。由于各个病人的生存时间不同，所以T为一随机变量。如果T的取值是连续的，则称T为连续型的随机变量。用t

* 初学者可以跳过这一节的内容

表示T的某一特定取值，用$f(t)$表示T在取值为t时的概率密度函数，其相应的分布函数定义为

$$F(t)=\int_0^t f(T)dT \tag{1.10}$$

$F(t)$表示随机变量T的取值小于或等于t的概率。这一函数表示一个病人在特定时间t之前的死亡概率。

一个病人能够至少生存到特定时间t的生存概率为

$$S(t)=1-F(t)=\int_t^\infty f(T)dT \tag{1.11}$$

由于有$f(t)=dF(t)/dt=F'(t)$，故生存概率与概率密度函数之间具有关系式

$$f(t)=\frac{d}{dt}[1-S(t)]=-S'(t) \tag{1.12}$$

上式中的$S'(t)$表示生存概率的导数。故概率密度函数是生存概率的导数的负值。

关于风险函数的定义为

$$h(t)=\lim_{\Delta t\to 0}\frac{Pr(t\leqslant T\leqslant t+\Delta t\mid T=t)}{\Delta t}=\frac{f(t)}{S(t)} \tag{1.13}$$

由于有$f(t)=-S'(t)$，故上式可表示为

$$h(t)=-\frac{d}{dt}\ln S(t) \tag{1.14}$$

上式中的ln表示自然对数。利用$S(0)=1$并对上式积分得到

$$\ln S(t)=-\int_0^t h(T)dT \tag{1.15}$$

从而有

$$S(t)=\exp[-\int_0^t h(T)dT] \tag{1.16}$$

将上式代入式（1.13）得到

$$f(t)=h(t)\exp[-\int_0^t h(T)dT] \tag{1.17}$$

我们定义累积风险函数［cumulative hazard function $H(t)$］为

$$H(t)=\int_0^t h(T)dT \tag{1.18}$$

故式（1.16）又可表示为

$$S(t)=\exp[-H(t)] \text{ 或 } H(t)=-\ln[S(t)] \tag{1.19}$$

当$t=0$时，有$S(0)=1$，$H(0)=0$；当$t=\infty$时，有$S(\infty)=0$，$H(\infty)=\infty$。

二、生存时间T为离散的情况

如果将一组病人的生存时间分为若干段，或者只对生存时间取离散的整数值时，随机变量T的取值t_1，t_2，…，t_k的序列表示为

$$0 \leqslant t_1 < t_2 < \cdots < t_k$$

记在时点t_j（$j=1, 2, \cdots, k$）的分布函数为$p(t_j)$

$$p(t_j) = P_r(T = t_j) \qquad j = 1, 2, \cdots, k \tag{1.20}$$

相应的生存概率为

$$S(t) = Pr(T \geqslant t) = \sum_{j: t_j \geqslant t} p(t_j) \tag{1.21}$$

这时的风险函数定义为

$$h(t_j) = Pr(T = t_j | T \geqslant t_j) = \frac{p(t_j)}{S(t_j)} \qquad j = 1, 2, \cdots, k \tag{1.22}$$

由于有$p(t_j) = S(t_j) - S(t_{j+1})$，故上式又可表示为

$$h(t_j) = 1 - \frac{S(t_{j+1})}{S(t_j)} \qquad j = 1, 2, \cdots, k \tag{1.23}$$

或写为

$$\frac{S(t_{j+1})}{S(t_j)} = 1 - h(t_j) \qquad j = 1, 2, \cdots, k \tag{1.24}$$

式（1.24）等号左边部分为条件生存概率，根据概率原理有

$$S(t) = \prod_{j: t_j < t} [1 - h(t_j)] \tag{1.25}$$

可以类似于连续型的情况，对在离散型情况下的累积风险函数$H(t)$定义为

$$H(t) = -\ln S(t) \tag{1.26}$$

式中的$\omega(t)$为式（1.25）所规定。必须说明的是，在离散情况下的$H(t)$，并不真正等于$\sum_{j: t_j < t} h(t_j)$，只是对连续型情况下关于累积风险函数这一定义的近似。

第二章　生存率的非参数估计法

用生存率$S(t)$表示一名病人生存时间长于t的概率。如果t以年为计算单位，则$S(t)$就是t年生存率。生存率的计算方法有两种：非参数估计法与参数估计法。在非参数法中又可分为乘积—极限法（product-limit method）和寿命表法（lifetable method）。前法适用于观察例数较少而不需分组的资料，后法适用于观察例数较多而分组的资料。下面将分别加以介绍。

第一节　乘积—极限法

乘积—极限法（以下简称PL法）为 Kaplan 和 Meier 于1958年首先提出，故又称 Kaplan-Meier 法。此法是利用条件概率及概率乘法原理来计算生存率。首先用一个简单例子对其基本计算方法加以解释。假设在1984年初开始对10名病人进行随访，观察到当年内死亡6人，1985年内又死亡3人，故这10名病人有4人活满1年，有1人活满2年；在1985年初又接收20名病人进行随访，在当年内死亡15人，即有5人活满1年。全部观察在1985年终截止。现在利用上述资料计算2年生存率。

用符号$S(1)$表示1年生存率，$S(2)$表示2年生存率，$S(2|1)$表示在生存1年条件下再能生存1年的条件概率。根据概率乘法规则（即无条件概率是各条件概率的连乘积）有

$$S(2)=S(1)\cdot S(2|1) \tag{2.1}$$

按照上面的假设资料计算的估计生存率为

$$\hat{S}(1)=(4+5)/(10+20)=0.30$$

$$\hat{S}(2|1)=1/4=0.25$$

利用式（2.1）得到估计的2年生存率$\hat{S}(2)$为

$$\hat{S}(2)=0.30\times 0.25=0.075$$

将式（2.1）化成一个通式，即从随访开始（$t_0=0$）到生存时间长于t_i年的生存率$S(t_i)$的计算公式为

$$\begin{aligned} S(t_i)&=S(1|0)\cdot S(2|1)\cdot S(3|2)\cdot\cdots\cdot S(t_i|t_{i-1}) \\ &=S(t_{i-1})S(t_i|t_{i-1}) \end{aligned} \tag{2.2}$$

上式中的$S(1|0)$为开始进入随访后生存时间长于1年的条件概率，$S(t_i|t_{i-1})$为在已生存t_{i-1}年条件下再生存1年到达生存t_i年的条件概率。下面将式（2.2）用于实际资料的计算。

一、用PL法计算完全资料的生存率

前已提到，所谓完全资料是指所有被观察者的生存时间都是完全数据，资料中无截尾数据。如第一章中例1.1的资料即完全资料。我们用这一例子来说明PL法的计算步骤。

1. 设资料总共有n名病人（本例中$n=20$，由于有的病人生存时间相等，故可归纳为k个不同的生存时间），第i名病人的生存时间为t_i（$i=1,2,\cdots,k$）。将这n名病人的生存时间按由小到大的顺序重新排列为

$$t_{(1)}\leqslant t_{(2)}\leqslant\cdots\leqslant t_{(k)}$$

该例中全部生存时间的排列顺序见表2.1中的第(1)列。为简便起见，以下仍用t_i表示有序时间$t_{(i)}$。

2. 计算在每一时间t_i上的死亡人数d_i。本例为受孕人数。例如在第1个月有2人受孕，在第2个月有1人受孕，……。到第18个月以后所有随访对象都已受孕。见表2.1

表2.1 完全资料的生存率计算表（例1.1资料：峡部-峡部吻合术后的受孕时间）

顺序号 i	时间（月） t_i	在t_i受孕例数 d_i	恰在t_i前未受孕例数 n_i	条件受孕率 $\hat{F}(t_i\|t_{i-1})$	条件未受孕率 $\hat{S}(t_i\|t_{i-1})$	未受孕率（生存率） $\hat{S}(t_i)$	$\frac{d_i}{n_i(n_i-d_i)}$	$\Sigma\frac{d_i}{n_i(n_i-d_i)}$	$SE[\hat{S}(t_i)]$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1	1	2	20	0.100	0.900	0.900	0.0056	0.0056	0.0673
2	2	1	18	0.056	0.944	0.850	0.0033	0.0088	0.0802
3	3	2	17	0.118	0.882	0.750	0.0078	0.0167	0.0969
4	4	3	15	0.200	0.800	0.600	0.0167	0.0333	0.1097
5	6	2	12	0.167	0.833	0.500	0.0167	0.0500	0.1118
6	8	1	10	0.100	0.900	0.450	0.0111	0.0611	0.1112
7	9	2	9	0.222	0.778	0.350	0.0317	0.0929	0.1067
8	10	1	7	0.143	0.857	0.300	0.0238	0.1167	0.1025
9	11	1	6	0.167	0.833	0.250	0.0333	0.1500	0.0968
10	12	1	5	0.200	0.800	0.200	0.0500	0.2000	0.0894
11	13	1	4	0.250	0.750	0.150	0.0833	0.2833	0.0798
12	15	1	3	0.333	0.667	0.100	0.1667	0.4500	0.0671
13	17	1	2	0.500	0.500	0.050	0.5000	0.9500	0.0487
14	18	1	1	1.000	0.0	0.000	0.0	—	—