

第十四章 生存分析

第一节 | 生存分析的基本概念

一句话总览

生存分析研究的不是“会不会发生”，

而是：

什么时候发生 + 在没发生之前如何利用信息。

一、生存时间 (Survival Time) —— “时间”是主角

1 生存时间的广义定义

**生存时间 (survival time) **指：

从一个明确的观察起点，
到某个终点事件发生所经历的时间长度。

关键不在“生死”，而在事件。

2 起点与终点必须明确 (考试高频)

- 观察起点可以是：
 - 发病时间
 - 初次确诊时间
 - 接受治疗时间
 - 开始暴露时间
- 终点事件可以是：
 - 死亡
 - 复发
 - 疾病发生
 - 治疗失败 / 有效反应

◆ 核心要求：

起点、终点、时间单位必须事先规定清楚
否则生存时间不可比。

3 为什么“时间”如此重要？

因为下面这些问题，普通二分类分析回答不了：

- 谁更早复发？
- 同样都复发了，谁拖得更久？
- 截止随访时还没复发的人怎么办？

👉 这正是生存分析存在的根本理由。

二、生存数据 (Survival Data) ——为什么不能直接丢进 t / χ^2 ？

1 两大类生存数据

(1) 完全数据 (complete data)

- 终点事件已发生
- 生存时间可准确观测

(2) 删失数据 (censored data)

- 终点事件 尚未发生或无法观测
- 只知道：

真实生存时间 \geq 已观察时间

◆ 删失 \neq 缺失

表 14-1 乳腺癌患者生存资料原始记录表

编号	age	type	node	size	start	end	time	status
1	56	1	0	2	2003-09-06	2004-05-19	8	1
2	32	1	1	2	2006-02-09	2008-02-01	24	1
3	81	2	0	2	2006-02-17	2009-04-24	38	1
4	44	2	0	1	2003-09-30	2009-04-25	67	1
5	32	1	1	2	2005-12-22	2008-07-01	30	0
6	35	1	1	2	2001-07-20	2003-07-21	24	0
7	40	1	0	2	2003-07-25	2004-11-15	15	0

NO

(李康,贺佳主编, 2024, p. 145) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

删失数据包含信息，不能随意丢弃。

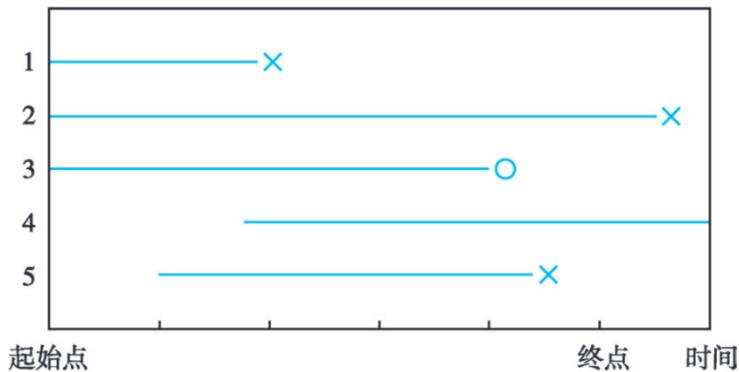


图 14-1 生存时间记录取示意图

(李康,贺佳主编, 2024, p. 146) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

2 删失数据的三种类型（要会区分）

① 左删失 (left censoring)

- 起始事件已发生

- 但具体发生时间未知
例：患者记不清既往发病时间

② 右删失 (right censoring) 【最常见】

- 起点已知
- 终点未发生或未观察到
例：
 - 随访结束仍未死亡
 - 失访
 - 因其他原因退出研究

③ 区间删失 (interval censoring)

- 只知道事件发生在某个时间区间内
- 不知道确切时点

👉 本章重点：右删失数据

3 删失产生的常见原因

1. 随访结束时事件尚未发生
 2. 失访（搬迁、拒访等）
 3. 死于与研究无关的其他原因
 4. 严重不良反应提前终止观察
-

4 删失数据的表示方式（细节但常考）

- 常在时间后标记 “+”
- 含义是：

真实生存时间未知，但至少大于该时间

例如：

- $24 \rightarrow$ 完全数据
 - $24+ \rightarrow$ 右删失数据
-

5 生存数据的本质特点 (必须会总结)

生存数据 = 生存结局 + 生存时间 + 可能删失

并且：

- 生存时间 **通常不服从正态分布**
- 传统统计方法直接失效

👉 这就是为什么需要**生存分析方法**。

三、生存分析的常用统计指标 (核心三件套)

(一) 生存率 / 生存函数 $S(t)$

1 定义 (非常重要)

生存率 (survival rate) 或生存函数：

$$S(t) = P(T > t)$$

含义：

观察对象生存时间超过 t 的概率

2 无删失时的估计 (直观版)

$$\hat{S}(t) = \frac{t \text{ 时刻仍存活的例数}}{\text{观察总例数}}$$

⚠ 现实中很少成立，因为几乎总有删失。

3 有删失时的核心思想（一定要懂）

- 将时间划分为多个时段
- 每一段计算：
 - “这一段活下来”的条件概率
- 生存率是逐段相乘的累积结果

$$S(t_k) = p_1 p_2 \cdots p_k = S(t_{k-1}) p_k$$

👉 因此：

生存率又称
累积生存概率 (cumulative survival probability)

(二) 中位生存期 (Median Survival Time)

1 定义

50% 个体尚存活时对应的时间

也称：

- 半数生存期
-

2 如何理解与比较？

- 中位生存期 **越长** → 预后 **越好**
- 中位生存期 **越短** → 预后 **越差**

◆ 实际获取方式：

从生存曲线上
找到 $S(t)=0.5$ 对应的时间

3 为什么常用“中位”而不是“均值”？

因为：

- 生存时间分布偏态
 - 均值容易被极端值拉偏
 - 中位数更稳健
-

(三) 风险函数 / 危险率 $h(t)$

1 定义 (理解 > 公式)

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

直觉翻译：

已经活到 t 的人，
在 t 时刻“瞬间发生终点事件”的速率。

2 离散时间下的直观近似

当 $\Delta t = 1$ ：

$$h(t) \approx P(t \leq T)$$

即：

t 时刻仍存活者，在下一个单位时间内死亡的概率

3 风险函数的形态 (非常有解释力)

- **常数风险**
→ 死亡速率不随时间变化
- **递增风险**
→ 随时间推移风险加大 (如癌症进展)
- **递减风险**
→ 越活越 “安全” (如手术早期风险高)

👉 这也是后续 Cox 模型的核心思想来源。

四、这一节的“整体逻辑框架”

为什么要生存分析?

→ 因为有时间 + 有删失

核心研究对象是什么?

→ 生存时间 T

三大核心指标:

→ 生存率 $S(t)$

→ 中位生存期

→ 风险函数 $h(t)$

传统方法为什么不行?

→ 无法正确处理删失与时间信息

五、考试级一句话总结 (可直接写)

生存分析用于研究从规定的观察起点到终点事件发生所经历的时间，其数据特点是同时包含生存结局、生存时间并可能存在删失数据。常用统计指标包括生存率 (生存函

数)、中位生存期和风险函数，其中生存率表示个体生存时间超过某一时刻的概率，风险函数描述在已存活至某时刻条件下终点事件发生的瞬时速率。

第二节 | 生存曲线及比较

一句话总览

生存曲线解决“怎么看”，
生存率比较解决“差在哪、差得显不显著”。

一、生存率估计方法的整体框架（先定位）

生存率估计分为两大类：

- **参数法**：假定生存时间分布形式（如指数、Weibull）
- **非参数法**：不对分布作假定（更常用）

非参数法中最常见的两种：

方法	适用情形
寿命表法	大样本、时间分组
Kaplan–Meier 法 (K–M)	精确生存时间、小样本或大样本均可

👉 本节核心：**Kaplan–Meier 法**。

二、Kaplan–Meier 生存曲线 (K–M 曲线)

① 方法本质（一定要抓住）

Kaplan–Meier 法是一种**非参数生存率估计方法**，其核心思想是：

**在每一个“真实发生终点事件的时间点”上，
计算一次条件生存概率，
再将这些概率累乘。**

❖ 关键词：
只在“死亡/复发”等事件发生时更新生存率。

2 生存率的计算逻辑（结构理解）

在第 i 个事件时间点 t_i ：

- n_i : 该时刻前仍处于风险集中的人数
- d_i : 该时刻发生终点事件的人数

则该时段的条件生存概率为：

$$p_i = 1 - \frac{d_i}{n_i}$$

累计生存率为：

$$\hat{S}(t_i) = \prod_{j \leq i} p_j$$

❖ 删失个体的处理原则：

**只在风险集中“减人数”，
不计入死亡数，
也不改变该时刻的生存率。**

3 生存率标准误（Greenwood 公式）

Kaplan-Meier 生存率的近似标准误为：

$$SE[\hat{S}(t_i)] = \hat{S}(t_i) \sqrt{\sum_{t_j \leq t_i} \frac{d_j}{n_j(n_j - d_j)}}$$

◆ 理解重点：

- 误差是逐事件时间点累积的
 - 随时间推移，SE 通常变大（风险集变小）
-

4 生存率的置信区间

常用正态近似：

$$\hat{S}(t_i) \pm z_{\alpha/2} SE[\hat{S}(t_i)]$$

⚠ 注意一个考试级细节：

在生存时间末端，
该区间可能超出 [0,1]，
实务中可采用对数或 log-log 变换修正。

5 生存曲线的绘制与解读

- 横轴：随访时间
- 纵轴：生存率
- 曲线形态：阶梯状下降

◆ 直观解读规则：

- 曲线高、下降缓慢 → 生存率高、生存期长
 - 曲线低、下降陡峭 → 生存率低、生存期短
-

6 中位生存期在 K-M 曲线中的定位

当生存曲线纵轴 $S(t) = 0.5$ 时，
对应的横轴时间
即中位生存期（半数生存期）。

例如：

图中肿瘤 $> 5 \text{ cm}$ 组的中位生存期 ≈ 23 个月。

三、生存率的比较（生存曲线的统计检验）

本质问题：
多条生存曲线
是否来自同一总体？

四、Log-rank 检验（最常用）

1 基本思想（非常重要）

在原假设 H_0 下：

各组在任何时间点的生存率相同。

因此：

- 可以根据：
 - 各组在该时刻的风险集人数
 - 全体死亡数
 - 计算：
 - **理论死亡数**
 - 再与：
 - **实际死亡数**进行比较。
-

2 统计量结构 (会认就行)

$$\chi^2 = \frac{(\sum d_{ki} - \sum T_{ki})^2}{\sum V_{ki}}, k = 1, \dots, g$$

其中：

- d_{ki} : 第 k 组在 t_i 的实际死亡数
- $T_{ki} = \frac{n_{ki}d_i}{n_i}$: 理论死亡数
- V_{ki} : 对应方差
- 自由度:

$$v = g - 1$$

判断规则：

- χ^2 小 \rightarrow 实际 \approx 理论 \rightarrow 生存率相同
- χ^2 大 \rightarrow 偏离明显 \rightarrow 生存曲线有差异

3 Log-rank 的“敏感区间”

对远期差异更敏感

(晚期死亡权重更大)

五、Breslow (广义 Wilcoxon) 检验

虽然你这段文字主要提到它的性质对比，但这是考试常问点：

- **Breslow 检验：**
 - 对近期死亡差异更敏感
- **Log-rank 检验：**
 - 对远期死亡差异更敏感

原因直觉：

风险集人数 n_i 随时间减少，
Breslow 给早期较大的 n_i 更高权重。

六、两种检验的共同前提（非常关键）

各组生存曲线应满足比例风险假设，
且生存曲线不能明显交叉。

⚠ 重要提醒（常被忽略）：

当生存曲线发生交叉时，
不适合做整体生存率比较
(log-rank / Breslow 都不合适)。

七、这一节的“判断级流程图”

① 有无删失？

→ 有：K-M 法

② 要不要画生存曲线？

→ 必须（任何生存分析第一步）

③ 比较几组？

→ ≥ 2 组：做 log-rank / Breslow

④ 差异主要在什么时候？

→ 近期：Breslow

→ 远期：Log-rank

⑤ 曲线是否交叉？

→ 是：慎用整体比较

八、考试级一句话总结 (可直接写)

Kaplan-Meier 法是一种非参数生存率估计方法，通过在各个终点事件发生时刻计算条件生存概率并累乘得到生存率，其生存曲线呈阶梯状。生存曲线的比较通常采用 log-rank 检验或 Breslow 检验，其中 log-rank 检验对远期生存差异更敏感，Breslow 检验对近期差异更敏感，两者均要求各组生存曲线满足比例风险关系且无明显交叉。

【例 14-2】某研究者收集了 15 例乳腺恶性肿瘤直径小于或等于 2cm 的患者和 21 例肿瘤直径大于 5cm 患者术后的生存资料，定义从手术后到患者复发的时间为生存时间(月)，试估计两组的复发率。

肿瘤直径 $\leq 2\text{cm}$:	10	10^+	13	18	25^+	29	30	33	46	50^+	54
	68 ⁺	71	88 ⁺	95 ⁺							
肿瘤直径 $> 5\text{cm}$:	5	9	13	13	14	15	19	20	21	22	24
	25	26	27	28	32	47	52	54	60	86	

以肿瘤小于等于 2cm 组为例，计算步骤如下：

- (1) 将生存时间(t_i)由小到大顺序排列，若完全数据与删失数据相同，则删失数据排在完全数据之后，见表 14-2 第(2)栏。
- (2) 列出时间区间 $[t_i, t_{i+1})$ 上的复发数 d_i 和删失数 c_i ，见表 14-2 第(3)、(4)栏。
- (3) 计算恰在每一时刻 t_i 之前的生存人数，即期初例数 n_i 。计算时应减去小于 t_i 的复发数和删失数，即 $n_i = n_{i-1} - d_{i-1} - c_{i-1}$ ，见表 14-2 第(5)栏。
- (4) 计算各时间区间上的复发概率 q_i 和生存概率 p_i ，见表 14-2 第(6)、(7)栏。

$$q_i = \frac{d_i}{n_i}, \quad p_i = 1 - q_i \quad (14-4)$$

- (5) 计算生存率 $\hat{S}(t_i)$ 。活过某时刻 t 的生存率是其对应的各时点条件生存率的连乘积，即

$$S(t_i) = \prod_{j=1}^i p_j = S(t_{i-1}) p_i \quad (14-5)$$

值得注意的是，具有删失数据的条件死亡率为 0，而其条件生存率必为 1，所对应的生存率必然与前一个非删失值的生存率相同。各患者的生存率见表 14-2 第(8)栏。如乳腺肿瘤小于等于 2cm 患者 18

(李康,贺佳主编, 2024, p. 147) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

个月时生存率为 $0.862 \times 0.917 = 0.790$, 30 个月时生存率为 $0.711 \times 0.889 = 0.632$, 以此类推。

(二) 计算生存率的标准误

Greenwood 生存率标准误近似计算公式为

$$SE[\hat{S}(t_i)] = \hat{S}(t_i) \sqrt{\sum_{t_j \leq t_i} \frac{d_j}{n_j(n_j - d_j)}}, \quad j=1, 2, \dots, i \quad (14-6)$$

如 $SE[\hat{S}(t_4)] = 0.790 \times \sqrt{\frac{1}{15 \times (15-1)} + \frac{0}{14 \times (14-0)} + \frac{1}{13 \times (13-1)} + \frac{1}{12 \times (12-1)}} = 0.108$

根据上述计算的生存率及其标准误可估计总体生存率的置信区间。其方法是采用正态分布的原理,用下式来进行估计。

$$\hat{S}(t_i) \pm z_{\alpha/2} SE[\hat{S}(t_i)] \quad (14-7)$$

采用该式时,生存时间末端值的置信区间可能会出现超出 [0,1] 范围的不合理情况,此时可采用对数变换的方法进行计算,具体详见相关书籍。

(李康,贺佳主编, 2024, p. 148) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 14-2 乳腺肿瘤直径 $\leq 2\text{cm}$ 组生存率计算表

序号 <i>i</i>	时间/月 <i>t_i</i>	复发数 <i>d_i</i>	删失数 <i>c_i</i>	期初例数 <i>n_i</i>	复发概率 <i>q_i</i> (6)=(3)/(5)	生存概率 <i>p_i</i> (7)=1-(6)	生存率 $\hat{S}(t_i)$ (8)	标准误 $SE[\hat{S}(t_i)]$ (9)
(1)	(2)	(3)	(4)	(5)				
1	10	1	0	15	0.067	0.933	0.933	0.064
2	10 ⁺	0	1	14	0.000	1.000	0.933	0.064
3	13	1	0	13	0.077	0.923	0.862	0.091
4	18	1	0	12	0.083	0.917	0.790	0.108
5	25 ⁺	0	1	11	0.000	1.000	0.790	0.108
6	29	1	0	10	0.100	0.900	0.711	0.123
7	30	1	0	9	0.111	0.889	0.632	0.132
8	33	1	0	8	0.125	0.875	0.553	0.137
9	46	1	0	7	0.143	0.857	0.474	0.139
10	50 ⁺	0	1	6	0.000	1.000	0.474	0.139
11	54	1	0	5	0.200	0.800	0.379	0.140
12	68 ⁺	0	1	4	0.000	1.000	0.379	0.140
13	71	1	0	3	0.333	0.667	0.253	0.139
14	88 ⁺	0	1	2	0.000	1.000	0.253	0.139
15	95 ⁺	0	1	1	0.000	1.000	0.253	0.139

(三) 绘制生存曲线

以随访时间为横轴,生存率为纵轴,将各个时间点所对应的生存率连接起来的一条曲线即为生存曲线。Kaplan-Meier 法对所有死亡时点估计生存率,其生存曲线呈阶梯式的变化。曲线高、下降平缓表示高生存率或较长生存期;曲线低、下降陡峭表示低生存率或较短生存期。图 14-2 为肿瘤小于或等于 2cm 和肿瘤大于 5cm 组(计算表略)的生存曲线,结果显示前者的生存率高于后者。

生存曲线纵轴生存率为 50% 时,所对应横轴生存时间即半数生存期。从图 14-2 中可以直观地看出肿瘤大于 5cm 组的半数生存期大约为 23 个月。

(李康,贺佳主编, 2024, p. 148) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

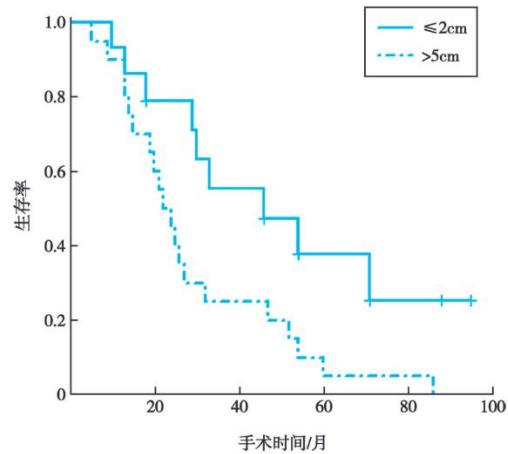


图 14-2 乳腺肿瘤 $\leq 2\text{cm}$ 和乳腺肿瘤 $> 5\text{cm}$ 组生存曲线

(李康,贺佳主编, 2024, p. 149) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

【例 14-3】根据例 14-2 资料,试比较两种不同肿瘤直径患者术后的生存率有无差别?

分析步骤如下:

1. 建立检验假设,确定检验水准

$H_0: S_1(t)=S_2(t)$, 即两种不同肿瘤直径的患者术后生存曲线相同

$H_1: S_1(t)\neq S_2(t)$, 即两种不同肿瘤直径的患者术后生存曲线不同

$\alpha=0.05$

2. 计算统计量 χ^2 值

(1) 时间排序:将两组生存时间混合后由小到大统一排序,对删失数据的处理同前。 n_{1i}, n_{2i} 分别表示两组观察患者数, $n_i=n_{1i}+n_{2i}$; d_{1i}, d_{2i} 分别表示两组在不同时间点上的复发数, $d_i=d_{1i}+d_{2i}$; c_{1i}, c_{2i} 分别表示两组在不同时间点上的截尾例数。计算结果列于表 14-3 中。

表 14-3 两种肿瘤直径患者术后生存曲线比较的 log-rank 检验计算表

序号 <i>i</i>	时间/月 <i>t_i</i>	肿瘤 $\leq 2\text{cm}$ 组					肿瘤 $> 5\text{cm}$ 组					合计	
		<i>n_{1i}</i> (3)	<i>d_{1i}</i> (4)	<i>c_{1i}</i> (5)	<i>T_{1i}</i> (6)	<i>V_{1i}</i> (7)	<i>n_{2i}</i> (8)	<i>d_{2i}</i> (9)	<i>c_{2i}</i> (10)	<i>T_{2i}</i> (11)	<i>V_{2i}</i> (12)	<i>n_i</i> (13)	<i>d_i</i> (14)
1	5	15	0	0	0.417	0.243	21	1	0	0.583	0.243	36	1
2	9	15	0	0	0.429	0.245	20	1	0	0.571	0.245	35	1
3	10	15	1	0	0.441	0.247	19	0	0	0.559	0.247	34	1
4	10	14	0	1	0.000	0.000	19	0	0	0.000	0.000	33	0

(李康,贺佳主编, 2024, p. 149) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

续表

序号	时间/月	肿瘤≤2cm 组						肿瘤>5cm 组						合计	
		i	t_i	n_{1i}	d_{1i}	c_{1i}	T_{1i}	V_{1i}	n_{2i}	d_{2i}	c_{2i}	T_{2i}	V_{2i}	n_i	d_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)		
5	13	13	1	0	0.813	0.467	19	1	0	1.188	0.467	32	2		
6	14	12	0	0	0.800	0.463	18	2	0	1.200	0.463	30	2		
7	15	12	0	0	0.429	0.245	16	1	0	0.571	0.245	28	1		
8	18	12	1	0	0.444	0.247	15	0	0	0.556	0.247	27	1		
9	19	11	0	0	0.423	0.244	15	1	0	0.577	0.244	26	1		
10	20	11	0	0	0.440	0.246	14	1	0	0.560	0.246	25	1		
11	21	11	0	0	0.458	0.248	13	1	0	0.542	0.248	24	1		
12	22	11	0	0	0.478	0.250	12	1	0	0.522	0.250	23	1		
13	24	11	0	0	0.500	0.250	11	1	0	0.500	0.250	22	1		
14	25	11	0	1	0.524	0.249	10	1	0	0.476	0.249	21	1		
15	26	10	0	0	0.526	0.249	9	1	0	0.474	0.249	19	1		
16	27	10	0	0	0.556	0.247	8	1	0	0.444	0.247	18	1		
17	28	10	0	0	0.588	0.242	7	1	0	0.412	0.242	17	1		
18	29	10	1	0	0.625	0.234	6	0	0	0.375	0.234	16	1		
19	30	9	1	0	0.600	0.240	6	0	0	0.400	0.240	15	1		
20	32	8	0	0	0.571	0.245	6	1	0	0.429	0.245	14	1		
21	33	8	1	0	0.615	0.237	5	0	0	0.385	0.237	13	1		
22	46	7	1	0	0.583	0.243	5	0	0	0.417	0.243	12	1		
23	47	6	0	0	0.545	0.248	5	1	0	0.455	0.248	11	1		
24	50	6	0	1	0.000	0.000	4	0	0	0.000	0.000	10	0		
25	52	5	0	0	0.556	0.247	4	1	0	0.444	0.247	9	1		
26	54	5	1	0	1.250	0.402	3	1	0	0.750	0.402	8	2		
27	60	4	0	0	0.667	0.222	2	1	0	0.333	0.222	6	1		
28	68	4	0	1	0.000	0.000	1	0	0	0.000	0.000	5	0		
29	71	3	1	0	0.750	0.188	1	0	0	0.250	0.188	4	1		
30	86	2	0	0	0.667	0.222	1	1	0	0.333	0.222	3	1		
31	88	2	0	1	0.000	0.000	0	0	0	0.000	0.000	2	0		
32	95	1	0	1	0.000	0.000	0	0	0	0.000	0.000	1	0		
合计	—	—	9	6	15.695	7.110	—	21	0	14.305	7.110	—	30		

(2) 计算理论数及其方差:在 H_0 成立的条件下,计算各组理论(期望)复发数及其方差。如生存时间为 5 个月时,肿瘤≤2cm 组的观察人数为 15,两组合计观察人数及复发人数分别为 36 和 1,则其理论复发数和方差为

(李康,贺佳主编, 2024, p. 150) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

肿瘤 $> 5\text{cm}$ 组的理论复发数和方差为

$$T_{21} = \frac{21 \times 1}{36} = 0.583, \quad V_{21} = \frac{21}{36} \times \left(1 - \frac{21}{36}\right) \times \left(\frac{36-1}{36-1}\right) \times 1 = 0.243$$

其余类推,结果见表 14-3 的第(6)、(7)和(11)、(12)栏。

(3) 计算统计量

由表 14-3 中两组总的实际复发数和理论复发数以及总的方差估计值,根据式(14-8),按肿瘤 $\leq 2\text{cm}$ 组计算得到统计量为

$$\chi^2 = \frac{(9 - 15.695)^2}{7.110} = 6.304$$

同理,或按肿瘤 $> 5\text{cm}$ 组计算可得

$$\chi^2 = \frac{(21 - 14.305)^2}{7.110} = 6.304$$

3. 确定概率,作出统计推论

查 χ^2 界值表(附表 7)得 $\chi^2_{0.05,1} = 3.84$, $P < 0.05$,按 $\alpha = 0.05$ 水准,拒绝 H_0 ,接受 H_1 ,可认为两条生存曲线不同,肿瘤直径 $\leq 2\text{cm}$ 患者的生存率高于肿瘤直径 $> 5\text{cm}$ 患者。

(李康,贺佳主编, 2024, p. 151) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

第三节 | Cox 回归

一句话总览

Cox 回归解决的是:

在存在删失、时间非正态的前提下,

多个因素如何共同影响“事件发生的速率”。

一、为什么需要 Cox 回归? (Why)

在生存分析中,多因素分析**不能用**:

- 多元线性回归
→ 生存时间不服从正态分布
- Logistic 回归
→ 只能分析“是否发生”，忽略“什么时候发生”

而生存数据的本质是：

时间 + 结局 + 删失

👉 因此需要一种模型：

- 同时利用 **生存时间信息**
- 正确处理 **右删失**
- 不强行假定生存时间分布

这就是 **Cox 比例风险回归模型 (PHREG)**。

二、Cox 回归模型的基本形式 (Model)

1 模型表达式 (一定要认得)

$$h(t, X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m)$$

含义逐项拆解：

- $h(t, X)$ ：
给定协变量 X 时，个体在 t 时刻的**风险函数**
- $h_0(t)$ ：
基础风险函数
(所有 $X = 0$ 时的风险，随时间变化，但不指定形式)
- β_j ：
回归系数 (待估参数)

👉 核心思想：

协变量通过指数函数成倍地“放大或缩小”基础风险。

2 线性化后的解释形式

$$\ln \left[\frac{h(t, X)}{h_0(t)} \right] = \beta_1 X_1 + \cdots + \beta_m X_m$$

👉 Cox 回归线性的不是时间，也不是风险本身，而是：

对数风险比 (log hazard ratio)

三、模型参数的意义：HR 是 Cox 的“语言核心”

1 回归系数与风险比 (HR)

设 X_j 在两个水平 c_1, c_0 下（其他变量相同）：

$$\ln HR_j = \beta_j(c_1 - c_0) \Rightarrow HR_j = \exp [\beta_j(c_1 - c_0)]$$

2 二分类自变量 (最常见)

若：

$$X_j = \begin{cases} 1, & \text{暴露} \\ 0, & \text{非暴露} \end{cases}$$

则：

$$HR_j = \exp (\beta_j)$$

3 HR 的判断规则 (必须条件反射)

β_j HR 解释

= 0 HR = 1 无影响

> 0 HR > 1 危险因素

< 0 HR < 1 保护因素

👉 解释永远要加一句限定词：

在控制其他协变量后。

💡 例子：化疗是否降低复发风险

给定：

$$\beta = -0.380$$

则：

- 接受化疗：

$$h_1(t) = 0.68 h_0(t)$$

- 未接受化疗：

$$h_2(t) = h_0(t)$$

风险比：

$$HR = \exp(-0.38) = 0.68$$

👉 化疗将复发风险降低约 32%

(或未化疗者风险是化疗者的 1.47 倍)

四、参数估计与假设检验 (How)

1 为什么不能用普通 MLE?

因为：

- $h_0(t)$ 未作任何分布假定
- 无法写出完整似然函数

👉 解决方法：

构造 **偏似然函数 (partial likelihood)**，
对 β 做最大似然估计。

❖ 软件层面你无需操心，**只要知道：**

Cox 回归 \neq 普通 MLE
Cox 回归 = 偏似然 + MLE

2 回归系数的假设检验

原假设统一为：

$$H_0: \beta_j = 0$$

常用三种检验（与 Logistic 完全同构）：

- 似然比检验 (LR)
- Wald 检验
- Score (计分) 检验

共同特点：

- 统计量近似服从 χ^2 分布
- 自由度 = 被检参数个数

❖ 实务与考试共识：

大样本下三种检验结果基本一致。

五、Cox 回归的核心前提：比例风险假定 (PH)

这是 Cox 回归最重要、也是最容易被忽略的前提。

1 什么是比例风险假定？

PH 假定包含两层含义：

1. 任意两个个体的风险比 HR 与时间无关
2. 协变量的效应不随时间改变

👉 换句话说：

曲线可以上下分开，但不能“越跑越不一样”。

2 如何初步判断是否满足 PH？

最直观的方法：

看按协变量分组的 Kaplan-Meier 生存曲线是否交叉

- 不交叉 → PH 假定大致合理
 - 明显交叉 → PH 假定可能被破坏
-

3 不满足 PH 怎么办？（知道方向即可）

可考虑：

- 时依协变量模型 (time-dependent covariates)
- 非比例风险模型
- 分层 Cox 模型

👉 考试层面记住一句话就够：

不满足 PH 假定，不能直接用 Cox。

六、使用 Cox 回归的注意事项（高频雷区）

1 起点与终点必须事先定义

- 起始事件与终点事件是**相对概念**
 - 一旦确定，**研究过程中不能随意更改**
-

2 样本量要求

- 一般 ≥ 40 例
 - 样本量 \approx 协变量数的 **5–20 倍**
 - 两组比较时尽量例数接近
 - 失访、删失过多 → 结果易偏倚
-

3 不能用原始回归系数比较影响大小

原因与 Logistic / 线性回归一致：

单位不同 $\rightarrow \beta_j$ 不可直接比

👉 需使用 **标准化回归系数**。

4 列线图 (Nomogram)

- Cox 回归结果可用于构建列线图
 - 本质：对回归系数标准化并可视化
 - 常用工具：R 的 rms、regplot
-

七、Cox 回归在“方法体系”中的位置

K-M + log-rank

→ 单因素、生存曲线层面比较

Cox 回归

→ 多因素、生存时间层面建模

→ 输出 HR (生存分析版 OR)

八、考试级一句话总结（可直接写）

Cox 比例风险回归模型适用于含删失的生存资料多因素分析，通过构造偏似然函数估计回归系数。模型假定不同个体之间的风险比随时间保持恒定，其回归系数的指数形式为风险比 HR，用于衡量协变量对生存结局发生速率的影响。Cox 回归要求满足比例风险假定，分析前应结合生存曲线判断其合理性。

【例 14-4】为探讨某肿瘤患者手术治疗的预后，某研究者收集了 68 例患者的生存时间、生存结局及影响因素。影响因素包括患者年龄、病理分型、淋巴结转移、肿瘤大小、化疗、绝经，生存时间以月为单位。变量的赋值和收集的相应资料分别见表 14-4 和表 14-5。

(李康,贺佳主编, 2024, p. 152) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 14-4 患者手术治疗的预后影响因素及赋值

因素	变量名	赋值说明
年龄	X_1	岁
病理分型	X_2	非浸润型=0, 浸润型=1
淋巴结转移	X_3	否=0, 是=1
肿瘤大小	X_4	< 3cm=1, 3~5cm=2, > 5cm=3
化疗	X_5	否=0, 是=1
绝经	X_6	否=0, 是=1
生存时间	t	月
生存结局	Y	删失=0, 死亡=1

(李康,贺佳主编, 2024, p. 153) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

表 14-5 68 例患者手术后的生存时间(月)及影响因素

NO	X_1	X_2	X_3	X_4	X_5	X_6	t	Y
1	77	1	1	3	1	1	2	1
2	58	1	0	2	0	1	23	1
3	92	1	0	2	1	1	16	1
4	48	1	1	3	0	1	13	1
5	52	1	1	2	1	1	26	1
6	67	1	0	3	1	0	54	1
7	55	1	1	2	0	1	31	1
...
65	39	1	0	1	0	0	23	0
66	32	1	0	1	0	0	77	0
67	22	0	0	1	0	0	88	0
68	41	1	1	1	0	0	33	0

注:完整数据可通过目录末尾数据包二维码获取。

(李康,贺佳主编, 2024, p. 153) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

对表 14-5 数据,取 $\alpha_{\text{引入}}=0.05$, $\alpha_{\text{剔除}}=0.10$,经逐步选择法对 6 个变量进行筛选,Cox 回归分析结果见表 14-6。

表 14-6 68 例患者 Cox 回归分析结果

变量	$\hat{\beta}$	$SE(\hat{\beta})$	Wald χ^2	P	HR	95%CI(HR)
X_1	0.039	0.017	5.325	0.021	1.039	(1.006, 1.074)
X_3	1.768	0.581	9.272	0.002	5.860	(1.878, 18.290)
X_4	1.479	0.485	9.291	0.002	4.390	(1.696, 11.364)
X_5	1.064	0.471	5.099	0.024	2.897	(1.151, 7.295)

分析结果显示,年龄、淋巴结转移、肿瘤大小和化疗的回归系数均为正值,说明它们是影响患者术后死亡的危险因素。在年龄、肿瘤大小、化疗保持不变的情况下,有淋巴结转移者的死亡风险是无淋巴结转移者的 5.860 倍;同样,在年龄、淋巴结转移、化疗保持不变的情况下,肿瘤大小每增加一级,死亡风险增加 3.390 倍。其他变量的解释以此类推。

(李康,贺佳主编, 2024, p. 153) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

进一步采用列线图可视化 Cox 回归结果(图 14-3),解读如下:由年龄(X_1)、淋巴结转移(X_3)、肿瘤大小(X_4)和化疗(X_5)的取值可获取对应得分,通过计算各变量的得分总和,可进一步获得 1 年、3 年和 5 年的生存率(注:图中方框表示各变量的类别占比,方框越大说明该类别占比越高;曲线反映了年龄和总得分的分布情况)。如图 14-3 所示,某位患者的年龄(X_1)约为 48 岁,存在淋巴结转移($X_3=1$)、肿瘤大小为 $> 5\text{cm}$ ($X_4=3$)、未接受化疗($X_5=0$),那么该患者的总得分为 213,其 1 年(12 月)生存率为 87.9%,3 年(36 月)生存率为 7.7%,5 年(60 月)生存率为 3.8%。

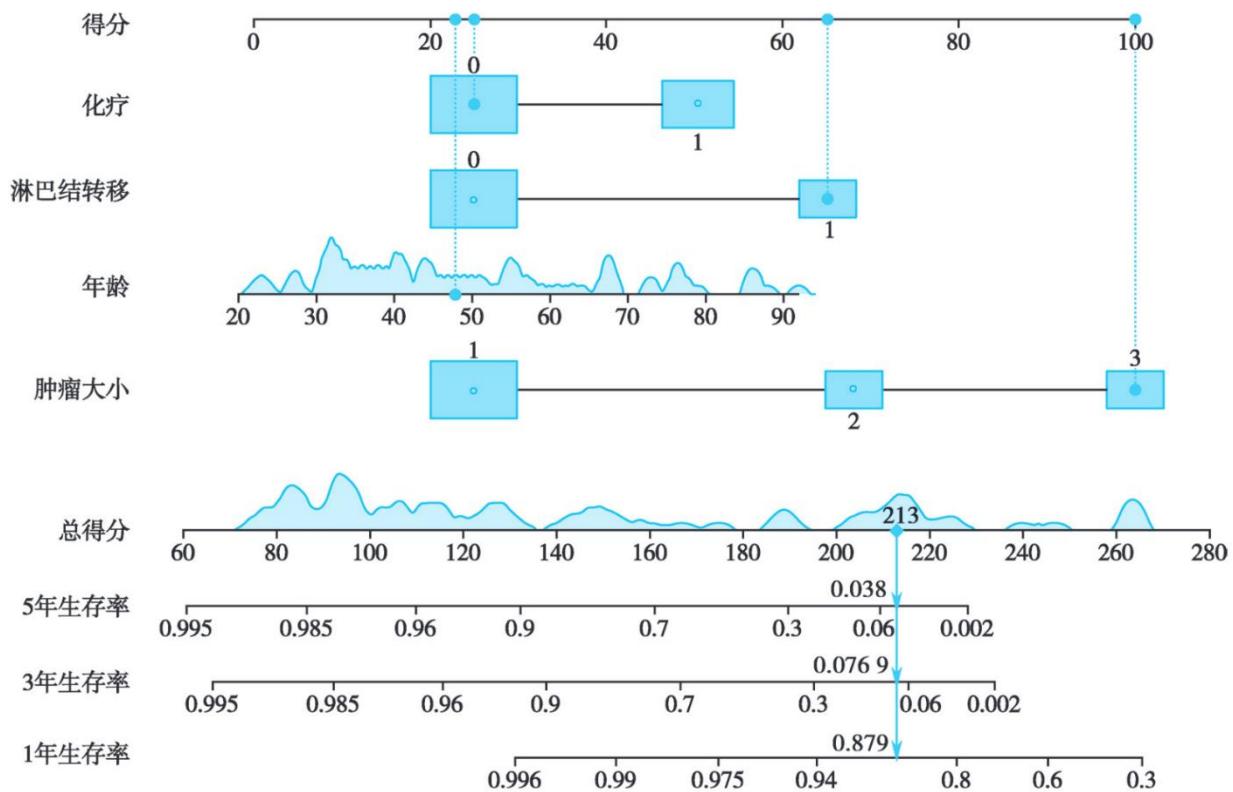


图 14-3 例 14-4 多因素 Cox 分析结果列线图

(李康,贺佳主编, 2024, p. 154) - Please click Zotero - Refresh in Word/LibreOffice to update all fields.

[BIBLIOGRAPHY] Please click Zotero - Refresh in Word/LibreOffice to update all fields