第八章 生存分析与Cox比例风险模型

复旦大学附属肿瘤医院 周支瑞

8.1 生存资料的描述

生存分析方法的正确表达

阅读临床研究文献的时候经常看到如下表述，这其中的统计学方法该何如理解？

- Aims compared treatments with respect to time to LR, time to distant recurrence, RFS, and OS via **log-rank** and **Cox regression methods**. **Log-rank tests** were exact, being based on hypergeometric probabilities. RFS and OS curves were estimated using the **Kaplan-Meier method.**
- Time-to-event curves were estimated by the **Kaplan–Meier method** and compared with the use of a two-sided **log-rank test.**

生存分析的基本概念

- 生存资料(time to event)的统计方法统称为生存分析，(survival analysis)，它是将事件的结局和发生这种结局所经历的时间两个因素综合起来分析的一种统计方法。它能够处理截尾数据，并对整个生存过程进行分析或比较。

生存分析的几个重要名词！

- **1. 终点事件:**

终点事件（terminal event)又称失效事件(failure event)或“死亡”事件(death event），泛指标志某种措施失败或失效的事件或反映治疗效果特征的事件，是根据研究目的确定的。如乳腺癌术后死亡、白血病化疗后复发、肾移植术后的肾衰、术后下床活动等，均可作为“死亡”事件。

- **【举例】RFS该如何定义?**

字面意思是“无复发生存”，那“结局事件”一般定义为：复发+转移+复发同时转移+癌症相关死亡，其余为“删失”

生存分析的几个重要名词 II

➢ 2. 生存时间:

生存时间(survival time)也是一个广义概念，泛指所关心的某现象的持续时间，即随访观察持续的时间，常用符号t表示。

随访资料几种情况【举例】

表 1. 肝癌癌患者术后随访记录

患者编号	观察记录				生存天数 t
	开始日期	终止日期	结局 (因肝癌死亡=1，未出现结局=0)	原因	
1	02-09-03	02-12-29	0	死于肺癌	118+
2	02-09-10	02-12-08	1	转移死亡	90
3	02-09-14	02-12-31	0	研究终止	108+
4	02-08-25	02-11-29	0	失　访	96+
5	02-10-01	02-11-28	0	死于车祸	59+
6	02-10-04	02-12-28	1	复发死亡	86

生存数据分为两种类型：

- [1]．完全数据(complete data)：指从观察起点到发生“死亡”事件所经历的时间。提供了观察对象确切的生存时间。
- [2]．截尾数据(censored data)：亦称截尾值(censored value)或终检值。指从观察起点到发生非“死亡”事件所经历的时间。

截尾原因大致有三种情况：

- [1]. 失访：未继续就诊、拒绝访问或搬迁而失去联系。
- [2]. 死于与研究疾病无关的原因：由于其他原因死亡。
- [3]. 研究终止：研究结束时终点事件尚未发生。

- 截尾数据不能提供完全的信息，真实的生存时间未知，只知道比观察到的截尾时间长，常用符号“+”表示。

总结 生存资料的特点：

- [1]. 有生存结局、生存时间
- [2]. 有不确定数据（截尾数据）
- [3]. 分布呈**指数分布**、**Weibull分布**、**对数正态分布**、**对数logistic分布**等

生存分析的几个重要名词 III

- **3. 死亡概率:**

死亡概率(*probability of death*)表示单位时间段开始存活的个体，在该段时间内死亡的可能性。符号*q*表示。

$$q = \frac{\text{某年内死亡人数}}{\text{某年年初人口数}}$$

生存分析的几个重要名词 IV

- **4. 生存概率:**

生存概率(*probability of survival*)表示单位时间段开始存活的个体，到该段时间结束时仍存活的可能性。符号*p*表示。

$$p = \frac{\text{某年活满一年人数}}{\text{某年年初人口数}}$$

$$p = 1 - q$$

生存分析的几个重要名词 V

- **5. 生存率:**

生存率(*survival rate, survival function*)表示观察对象经历t_k个单位时间段后仍存活的可能性。

- 若无截尾数据，则

$$S(t_k)=P(T>t_k)=\frac{t_k\text{时刻仍存活的例数}}{\text{观察总例数}}$$

$$0\le S(t)\le 1$$

生存分析的几个重要名词 V

- 若有截尾数据，须分时段计算生存概率。假定观察对象在各个时段的生存事件独立，应用概率乘法定理：

$$S(t_k) = P(T > t_k) = p_1 . p_2 \cdots p_k$$

- *Pi*某时段的生存概率，故生存率又称累积生存概率 (Cumulative Probabilityof Survival)。

生存率与生存概率的区别

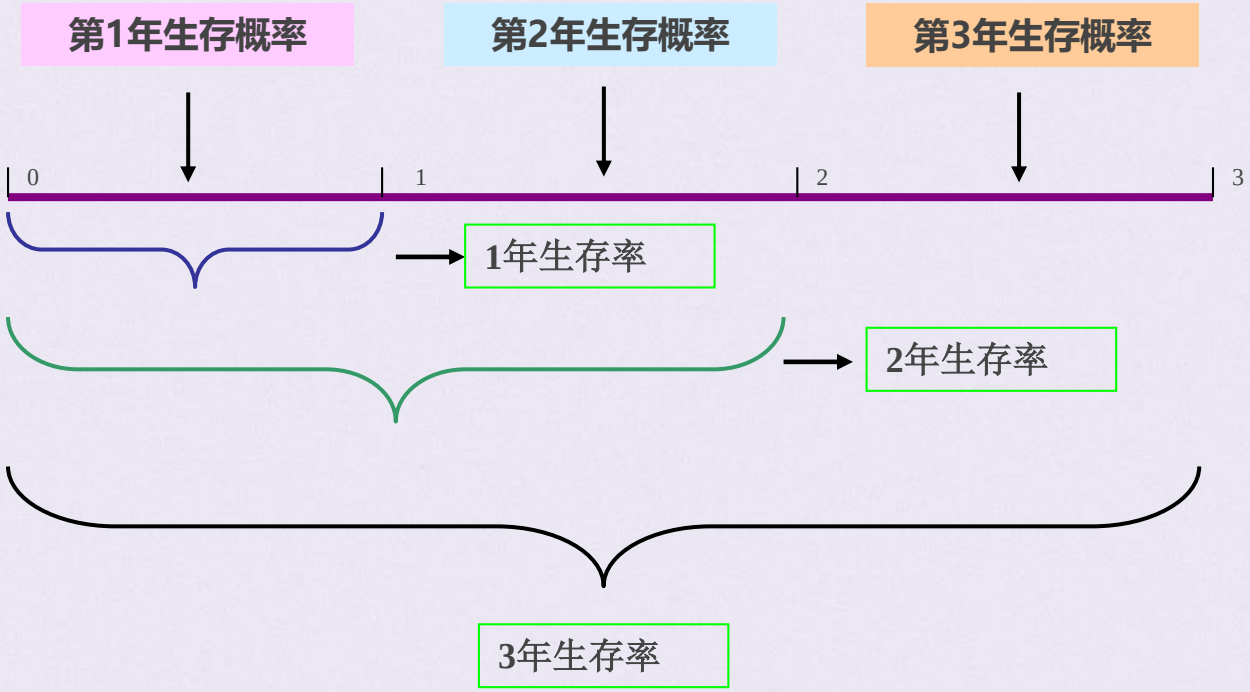


图 1. 生存概率与生存率示意图

生存分析的几个重要名词 VI

- **6. 生存曲线:**

生存曲线(*survival curve*)：生存时间为横轴，将各时点所对应的生存率连接在一起的曲线图。

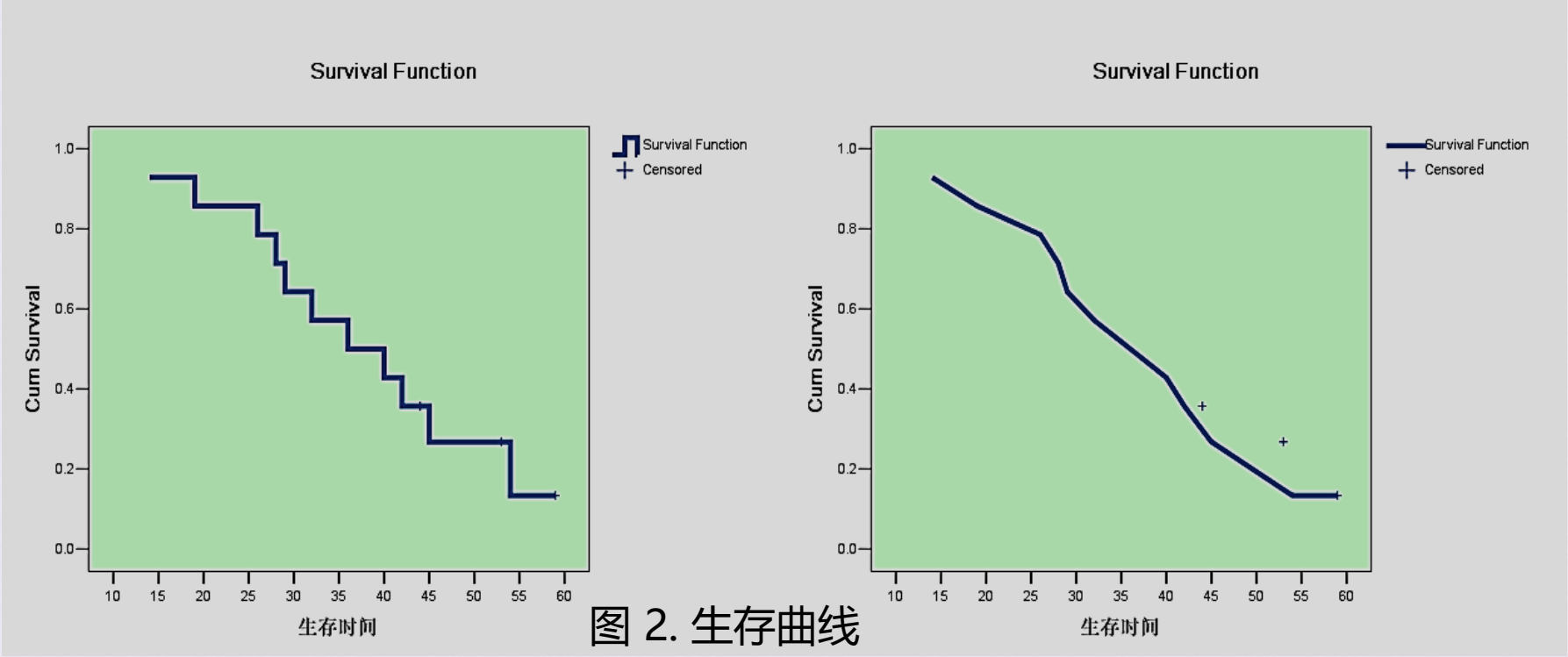


图 2. 生存曲线

所谓生存分析主要包括：

- [1]. 统计描述：
  - 计算生存率、绘制生存率曲线、计算中位生存时间等。
- [2]. 统计推断：
  - 估计总体生存率的可信区间、生存率曲线的比较。

生存分析基本方法：

- [1]. 非参数法: 不论资料是什么样的分布类型，只根据样本提供的顺序统计量对生存率进行估计，常用**乘积极限法和寿命表法**。
- [2]. 参数法: 假定生存时间服从于特定的参数分布，根据已知分布的特点对影响生存的时间进行分析，常用指数分布法、Weibull分布法、对数正态回归分析法和对数Logistic回归分析法。
- [3]. 半参数法: 介于参数法和非参数法之间，一般属多因素分析方法，用于探讨生存过程的主要影响因素，其经典方法是**Cox比例风险回归模型**。

随访研究的几个问题 I -- 随访时间

➢ 1. 开始随访的时间：

入(出)院时间、确诊时间、开始治疗时间等可作为随访开始的时间。如乳腺癌的乳腺切除术后第一天或出院日、白血病化疗后缓解出院日等，也可规定开始治疗日为随访开始时间。

随访研究的几个问题 I -- 随访时间

- **2. 随访结局和终止随访时间:**
- 随访的结局可能有以下几种:

(1) 死亡：泛指处理措施失败的事件。如肿瘤化疗后的复发、肾移植因肾衰或与之有关的原因而死亡等。终止随访时间为“死亡”时间。

(2) 失访：拒绝随访、失去联系或中途退出等。终止随访时间为最后一次时间。

(3) 死于与研究疾病无关的原因：终止随访时间为死亡时间。

(4) 研究终止。研究终止时观察对象仍然存活。终止随访时间为研究终止时间。

随访研究的几个问题 I -- 随访时间

- **3. 影响生存的有关因素：**

如患者年龄、病情、病程、术前健康等情况，以便分析这些因素对生存率的影响。

随访研究的几个问题 II -- 随访方式

➢ 1. 规定时点入组，统一分组，同一起点开始随访

全部观察对象同时接受处理措施，观察到最后一例出现结果或事先规定的随访截止时间。

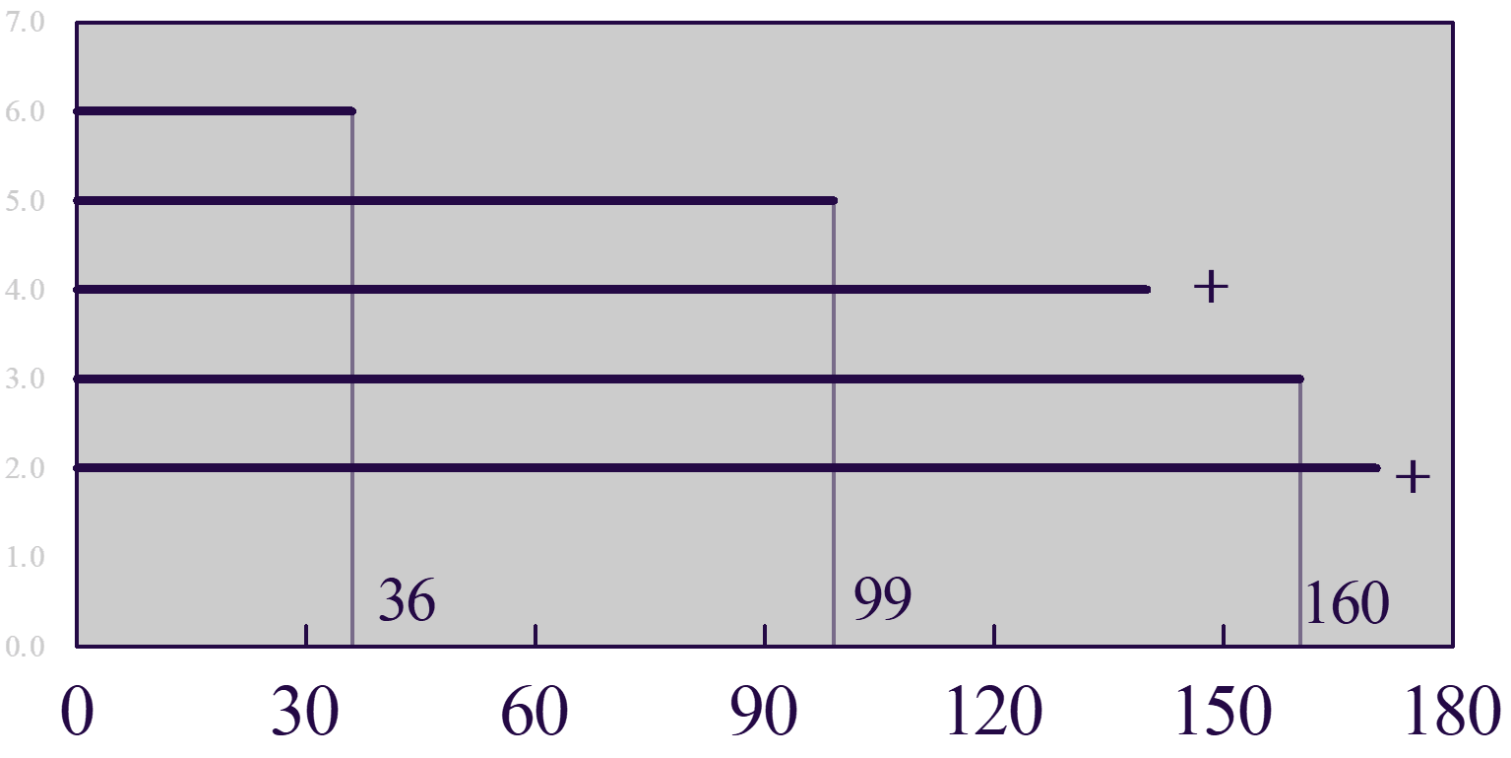


图 3. 随访资料常见形式示意图

随访研究的几个问题 II -- 随访方式

2. 研究对象入组时间不定，随时入组，即刻随访

观察对象在不同时间接受处理措施，完成一定数量随访病例或按事先规定的时间停止随访

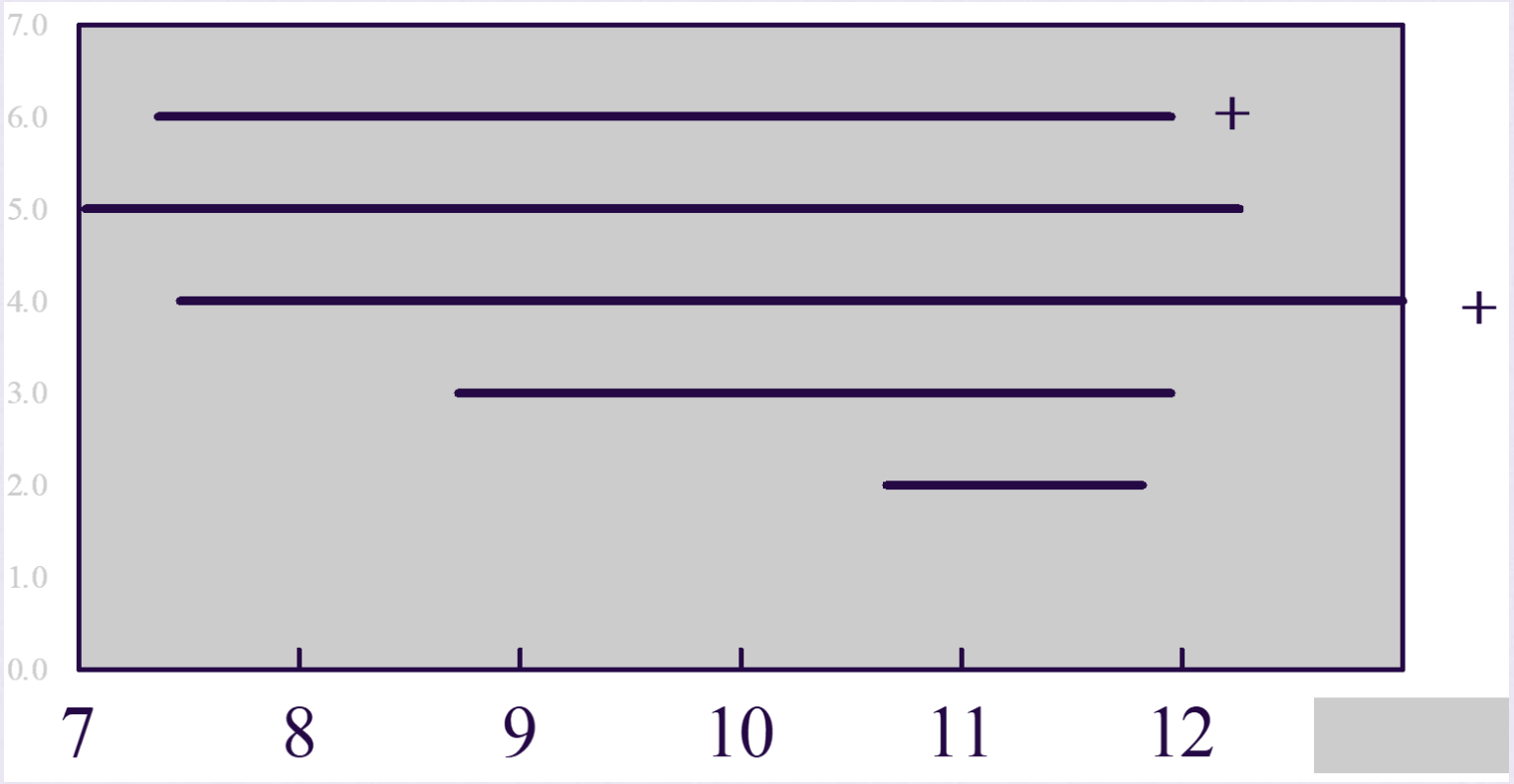


图 4. 随访资料常见形式示意图

生存率的估计与生存曲线 -- *Kaplan-Meier I*

➢ 1. 乘积极限法:

- ➢ 乘积极限法(product-limit estimate)又称Kaplan- Meier法，适用于分组生存资料的分析，**需要已知每例患者的生存时间与状态**

- ➢ 【案例】14例膀胱肿瘤 < 3.0cm患者和16例膀胱肿瘤 ≥ 3.0患者的生存时间（月）如下，试估计两组各时点生存率及其标准误、各时点总体生存率的95%可信区间、中位生存时间，并绘制生存曲线。

Kaplan-Meier【案例】

肿瘤<3.0cm	14	19	26	28	29	32	36	40	42	44+	45	53 +	54	59 +		
肿瘤≥3.0cm	6	7	9	10	11	12	13	20	23	25	27	30	34	37	43	50

表 2. 膀胱肿瘤<3.0cm组生存率及标准误的计算

生存时间 t	死亡数 d_t	期初病例数 n_t	截尾数 c_t	死亡概率 q_t	生存概率 p_t	生存率 $S(t)$	生存率标准误 $SE[S(t)]$
14	1	14	0	1/14=0.0714	0.9286	0.9268	0.0688
19	1	13	0	1/13=0.0769	0.9231	0.8572	0.0935
26	1	12	0	1/12=0.0833	0.9167	0.7858	0.1097
28	1	11	0	1/11=0.0909	0.9091	0.7144	0.1207
29	1	10	0	1/10=0.1000	0.9000	0.6429	0.1281
32	1	9	0	1/9=0.1111	0.8889	0.5715	0.1323
36	1	8	0	1/8=0.1250	0.8750	0.5001	0.1336
40	1	7	0	1/7=0.1429	0.8571	0.4286	0.1323
42	1	6	0	1/6=0.1667	0.8333	0.3571	0.1281
44	0	5	1	0/5=0.0000	1.0000	0.3571	0.1281
45	1	4	0	1/4=0.2500	0.7500	0.2678	0.1233
53	0	3	1	0/3=0.0000	1.0000	0.2678	0.1233
54	1	2	0	1/2=0.5000	0.5000	0.1339	0.1130
59	0	1	1	0/1=0.0000	1.0000	0.1339	0.1130

生存率的估计与生存曲线 -- *Kaplan-Meier II*

- [1]. 生存时间t：由小到大排列，遇非截尾和截尾值相同，截尾值排后。
- [2]. 死亡数dt：与生存时间t对应。注意：截尾值对应的个体未发生“死亡”事件，故死亡数为0。
- [3]. 期初病例数nt，表示恰好在该时刻以前的病例数。如n29为10，表示恰好在29月时点前有10人存活。
- [4]. 死亡概率qt，表示t月前的观察对象恰好在t月时点死亡的概率。

生存率的估计与生存曲线 -- *Kaplan-Meier II*

- [5]. 生存概率pt，表示t月前的观察对象恰好在t月时点存活的概率。
- [6]. 生存率$S(t)$。表示该人群恰好活过t时刻的概率。它为小于和等于 t时刻的各时点生存概率的乘积。
- [7]. 生存率的标准误$SES(t)$ 。

$$SE[S(t)] = S(t)\sqrt{\frac{1-S(t)}{n_t - d_t}}$$

生存率的估计与生存曲线 -- *Kaplan-Meier III*

- **总体生存率**的可信区间计算
- 假定生存率近似服从正态分布，某时点总体生存率的（1 - a）%可信区间，公式为：

$$S(t) \pm u_{\alpha/2} SE[S(t)]$$

- 本例28月总体生存率的95%可信区间：

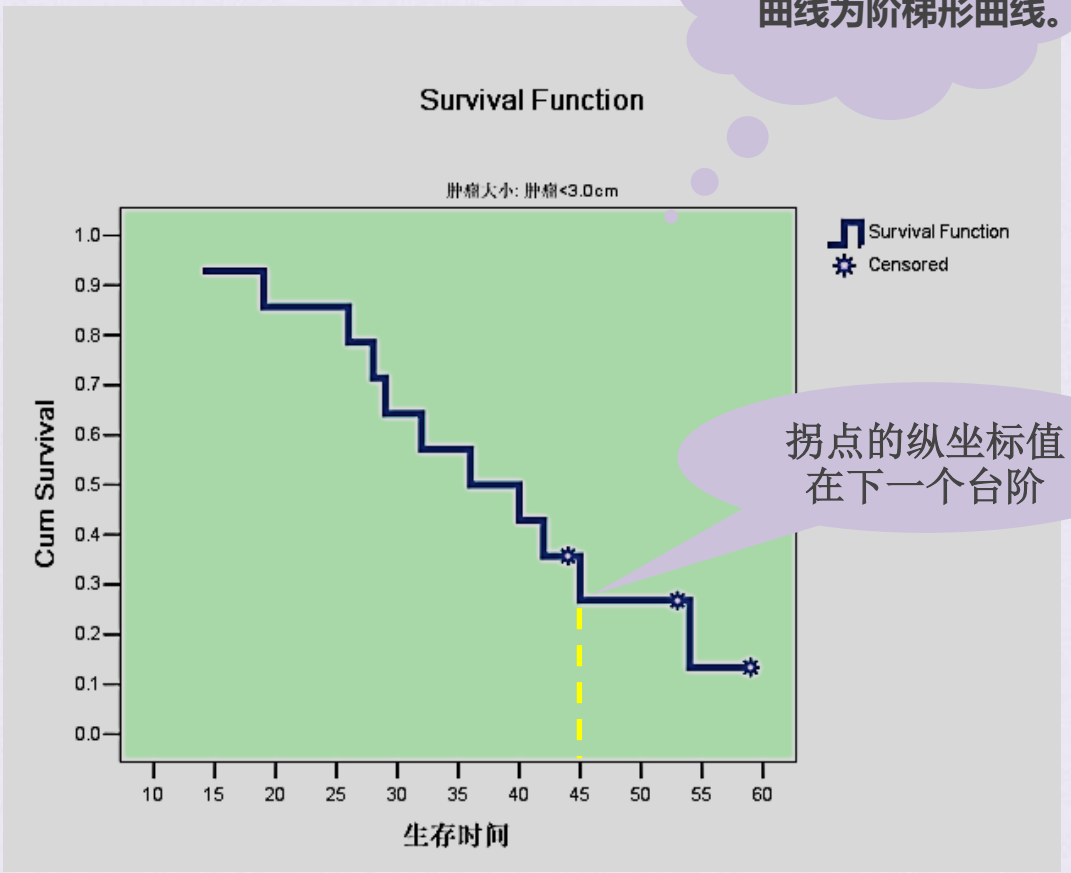
$$0.7144 \pm 1.96 \times 0.1207$$

- 即膀胱肿瘤<3.0cm患者28月生存率的95%可信区间为47.78%~95.10%。生存曲线尾部的生存率不适合于用该法计算总体生存率的可信区间。

生存率的估计与生存曲线 -- *Kaplan-Meier III*

死亡时点生存率，Kaplan-Meier法生存曲线为阶梯形曲线。

曲线高度&下降坡度



拐点的纵坐标值在下一个台阶

图 5. 肿瘤<3.0cm组生存曲线

生存率的估计与生存曲线 -- *Kaplan-Meier III*

- **中位生存时间计算:**
  - 由表2可见，中位生存时间估计在36月。
  - 采用线性内插法计算：找到与生存率50%相邻的上下两个生存率及其生存时间，利用线性比例关系求解中位生存时间（后面有举例）。
  - 若生存率0.5处所对应的曲线与*X*轴平行，则中位生存时间不止一个。
- **若各时间点生存率均大于50%，则无法估计中位生存时间。**

中位生存时间计算方法 -- 图示法

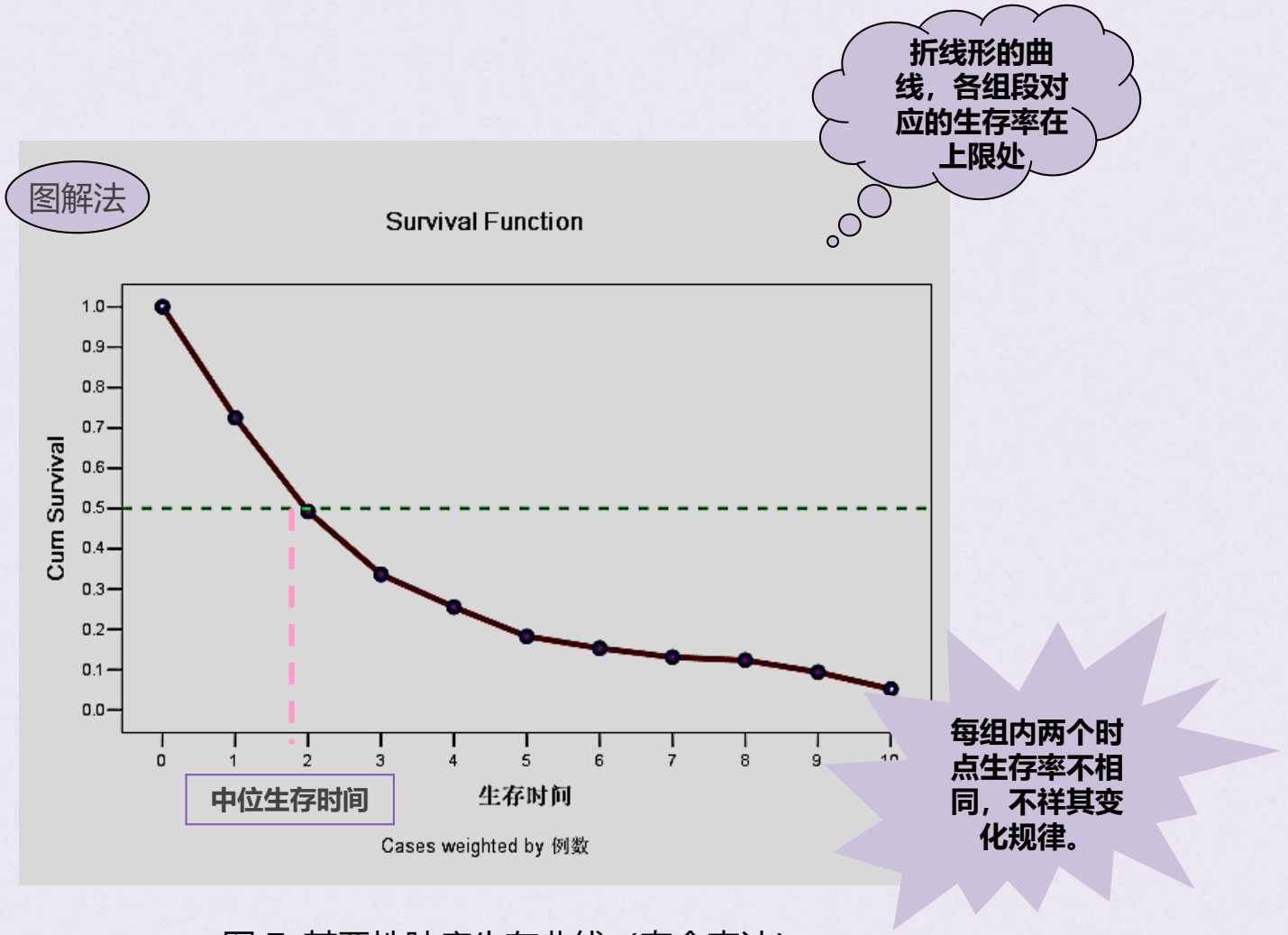


图 7. 某恶性肿瘤生存曲线（寿命表法）

中位生存时间计算方法 -- 线性内插法

- 线性内插法计算方法：

$$(2-3):(2-t)=(0.5562-0.4198):(0.5562-0.5)$$

$$t=2-\frac{(2-3)(0.5562-0.5)}{0.5562-0.4198}=2.41$$

生存率的估计与生存曲线 -- *Kaplan-Meier III*

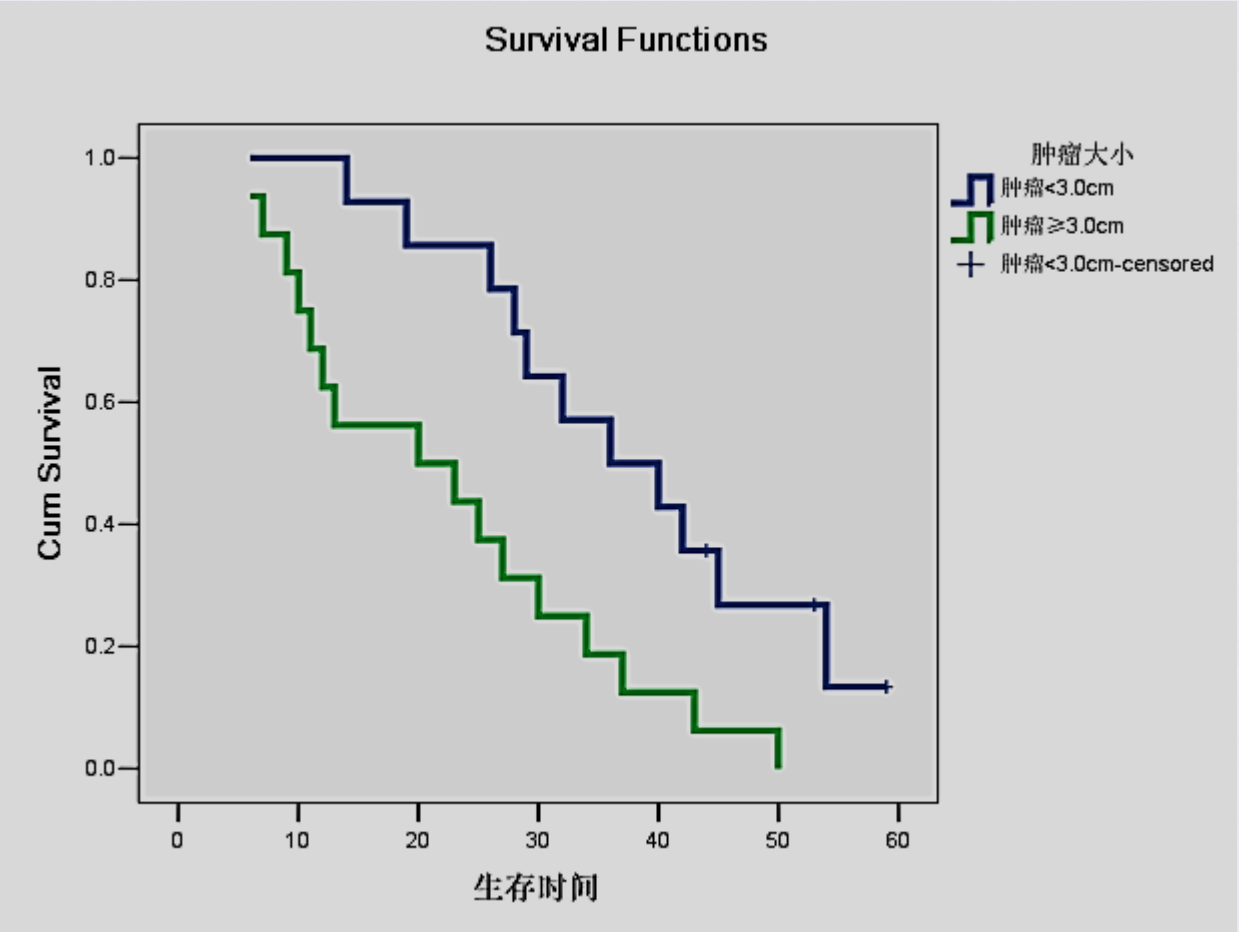


图 6. 肿瘤<3.0cm组和肿瘤≥3.0cm生存曲线

生存率的估计与生存曲线 -- 寿命表法/

- **2. 寿命表法**
- 适用于未分组的生存资料，不需要知道每例患者的生存时间与状态。
  - ①. 实际工作中，随访结果常常没有每个观察对象确切的生存时间，只能获得按随访时间分段的资料。
  - ②. 当样本较大(如 $n \geq 50$)时，采用乘积极限法估计生存率及其标准误往往较为繁琐。

生存率的估计与生存曲线 -- 寿命表法/

【案例】收集374名某恶性肿瘤患者随访资料，取时间区间均为1年，结果间下表，试估计生存率及其标准误，中位生存时间，并绘制生存曲线。

大样本、分组（生存时间粗略）、截尾

表 3. 某恶性肿瘤患者随访资料

序号	确诊后年数	期内死亡数	期内截尾数	期初病例数
1	0~	90	0	374
2	1~	76	0	284
3	2~	51	0	208
4	3~	25	12	157
5	4~	20	5	120
6	5~	7	9	95
7	6~	4	9	79
8	7~	1	3	66
9	8~	3	5	62
10	9~10	2	5	54

生存率的估计与生存曲线 -- 寿命表法//

- **生存率的计算:**
  - 1. 确诊后年数t～：“0～”表示从确诊日起不满一年，“1～”表示确诊后1年至不满2年，依次类推。
  - 2. 期内死亡数dt：指期内死于某恶性肿瘤的人数。
  - 3. 期内截尾数ct：泛指具有截尾数据的人，表示随访已满t年，但在未满t+1月期间失访的人。
  - 4. 期初观察例数n’ t：指时刻t以前的人数。
  - 5. 期初有效例数nt：相当于实际观察人时数。在各年年初观察人数中减去同年截尾数的一半。

生存率的估计与生存曲线 -- 寿命表法//

- **生存率的计算 续:**
  - 6. 死亡概率qt，指活满t年的病人在t + 1年内死亡的的概率。
  - 7. 生存概率pt，指活满t年的病人在t + 1年内存活的概率。
  - 8. 生存率S(t)，表示活过t年的概率。它为小于和等于 t 时刻的各时点生存概率的乘积。
  - 9. 生存率的标准误SES(t) 。

生存率的估计与生存曲线 -- 寿命表法///

表 4. 寿命表法估计生存率计算表

确诊后年数 *t*	期内死亡数 *dt*	期内截尾数 *ct*	期初病例数 n'_t	期初有效例数 *nt*	死亡概率 *qt*	生存概率 *pt*	生存率 *S(t)*	生存率标准误 *SE[S(t)]*
0~	90	0	374	374.0	90/374.0=0.2406	0.7594	0.7594	0.0221
1~	76	0	284	284.0	76/284.0=0.2676	0.7324	0.5562	0.0257
2~	51	0	208	208.0	51/208.0=0.452	0.7548	0.4198	0.0255
3~	25	12	157	151.0	25/151.0=0.1656	0.8344	0.3503	0.0248
4~	20	5	120	117.5	20/117.5=0.1702	0.8298	0.2907	0.0239
5~	7	9	95	90.5	7/90.5=0.0773	0.9227	0.2682	0.0235
6~	4	9	79	74.5	4/74.5=0.0537	0.9463	0.2538	0.0233
7~	1	3	66	64.5	1/64.5=0.0155	0.9845	0.2499	0.0233
8~	3	5	62	59.5	3/59.5=0.0504	0.9496	0.2373	0.0232
9~10	2	5	54	51.5	2/51.5=0.0388	0.9612	0.2281	0.0232

生存分析对生存资料的基本要求：

- (1) 样本由随机抽样方法获得，并应有足够的数量；
- (2) 死亡例数不能太少(≥30)；
- (3) 截尾值比例不能太大；
- (4) 生存时间尽可能精确到天数，因为多数生存分析方法都在生存时间排序的基础上作统计处理的，即使是小小的舍入误差，也可能改变生存时间顺序而影响结果。

Kaplan-Meier法估计案例

melanom　　恶性黑色素瘤后的生存数据

描述

该数据集包含 205 行、7 列，包含了患恶性黑色素瘤病人术后的生存数据，由 Odense 大学医院的 K.T.Drzewiecki 收集。

用法

```
melanom
```

格式

该数据框包含如下列：

no 数值向量，病人编号。
status 数值向量，编码表示生存状态：1：死于黑色素瘤，2：存活，3：死于其他原因。
days 数值向量，观测时间。
ulc 数值向量编码，ulceration；1：存在，2：缺失。
thick 数值向量，肿瘤厚度（1/100 mm）。
sex 数值向量编码：1：女性，2：男性。

Kaplan-Meier法估计案例 代码


```
> library(survival)
> library(ISwR)
> attach(melanom)
> names(melanom)
> Surv(days, status==1)
> survfit(Surv(days, status==1)~1)
> surv.all <- survfit(Surv(days,status==1)~1)
> summary(surv.all)
> plot(surv.all)
> surv.bysex <- survfit(Surv(days,status==1)~sex)
> plot(surv.bysex)
> plot(surv.bysex, conf.int=T, col=c("black","gray"))

# log-rank
> survdiff(Surv(days,status==1)~sex)
> survdiff(Surv(days,status==1)~sex+strata(ulc))

# cox regression
> summary(coxph(Surv(days,status==1)~sex))
> summary(coxph(Surv(days,status==1)~sex+log(thick)+strata(ulc)))
> plot(survfit(coxph(Surv(days,status==1)~ log(thick)+sex+strata(ulc))))
```

LifeTable寿命表法案例

- **例15-2** 某医生比较两种药物治疗HIV感染患者后的生存时间t(月)，试用寿命表法分析患者生存时间

LifeTable 寿命表法案例 代码


```
> hmohiv<-read.table("hmohiv.csv", sep=",", header = TRUE)
> attach(hmohiv)
> head(hmohiv)
> library(KMsurv)
> library(nlme)
> t6m<-floor(time/6)
> tall<-data.frame(t6m, censor)
> die<-gsummary(tall, sum, groups=t6m)
> total<-gsummary(tall, length, groups=t6m)
> rm(t6m)
> ltab.data<-cbind(die[,1:2], total[,2])
> detach(hmohiv)
> attach(ltab.data)

> lt=length(t6m)
> t6m[lt+1]=NA
> nevent=censor
> nlost=total[,2] - censor
> mytable<-lifetab(t6m, 100, nlost, nevent)
> mytable[,1:5]
> plot(t6m[1:11], mytable[,5], type="s", xlab="Survival time in every 6 month",
ylab="Proportion Surviving")
```

LifeTable 寿命表法 SPSS软件实现 案例

表 19.2　114 例男性胃癌患者术后生存情况

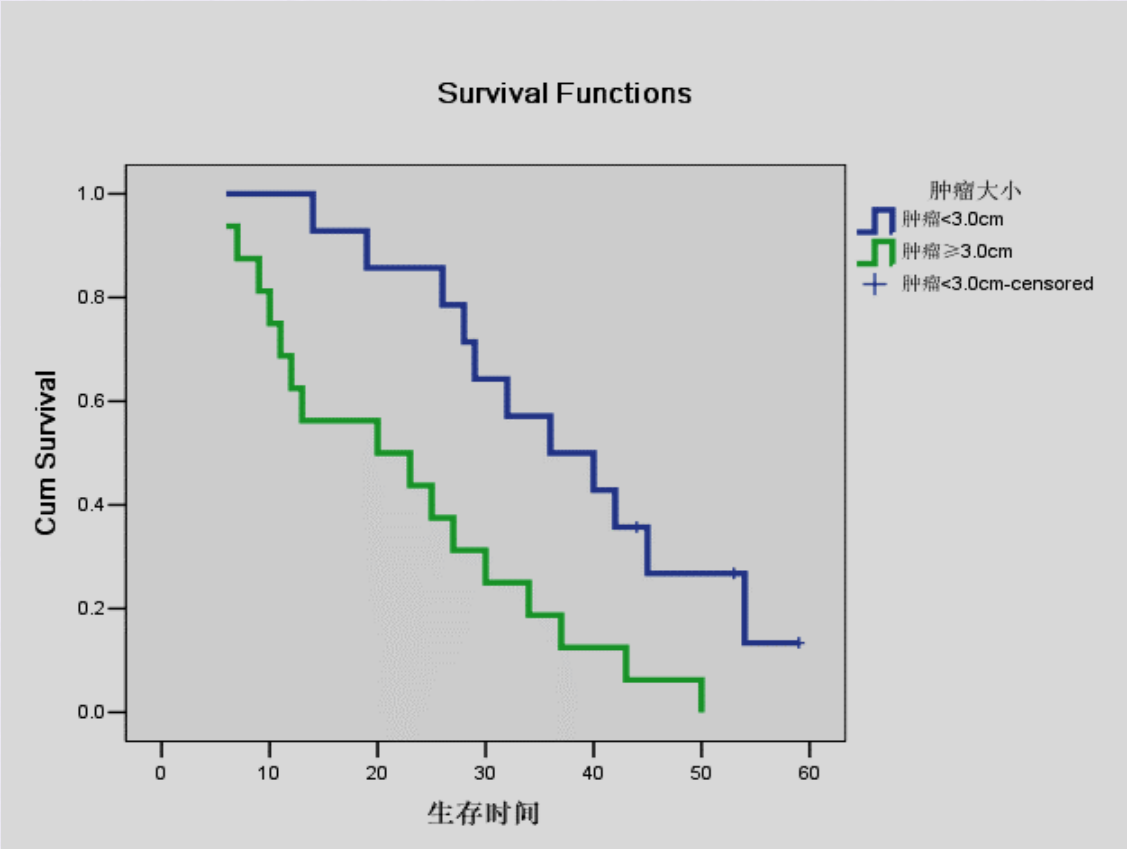
术后年数	0 ~	1 ~	2 ~	3 ~	4 ~	5 ~	6 ~	7 ~	8 ~	9 ~	10 ~ 11
期间失访人数	5	4	1	0	2	2	2	1	0	1	1
期间死亡人数	3	9	10	22	2	8	12	10	5	3	11

8.2 组间生存曲线的比较

生存曲线的比较 -- 秩和检验?

- 【案例】比较14例膀胱肿瘤<3.0cm和16例膀胱肿瘤≥3.0的生存曲线是否有差别。

肿瘤<3.0cm	14	19	26	28	29	32	36	40	42	44+	45	53 +	54	59 +		
肿瘤≥3.0cm	6	7	9	10	11	12	13	20	23	25	27	30	34	37	43	50



生存曲线的比较 -- Log-rank检验

- 对数秩检验，非参数检验法，其零假设为两总体生存曲线相同，但检验过程一般不估计生存率，而利用死亡数和死亡率函数作统计推断。
- 基本思想：当H0成立时，根据t时点的死亡率，计算出各组的理论死亡数，则检验统计量：

$$\chi^2 = \frac{(A_g - T_g)^2}{V_g} \qquad V_g = \sum \frac{n_{gi}}{ni}(1-\frac{n_{gi}}{ni})(\frac{n_i - d_i}{n_i - 1})d_i$$

亦可用公式：

$$\chi^2 = \sum \frac{(A-T)^2}{T}$$

检验统计量 χ^2 **近似服从** ν=（组数－1） **的** χ^2**分布**

生存曲线的比较 -- Log-rank检验

- 【案例】两条生存曲线比较步骤：
  - H_0：两总体的生存曲线位置相同
  - H_1：两总体的生存曲线位置不同
  - $a = 0.05$
  - （1）将两组资料混合后统一按生存时间（t）排序：n1i、n2i分别表示两组观察病例数，ni = n1i + n2i。
  - （2）分别列出各组在时间t的期初例数n_{gi} 和d_{gi}，两组合计的期初例数n和死亡例数 d_i。

生存曲线的比较 -- Log-rank检验

- （3）计算各组在时间t上的理论死亡T_{gi}：$T_{gi}=\frac{n_{gi}d_i}{n_i}$
- 各时间 t上都对应一个四格表，以第一个6（月）为例：

表 5. 理论死亡数计算表（以第一个6月为例）

组别	死亡数	未死亡数	合计
肿瘤<3.0cm	0	14	14
肿瘤≥3.0cm	1	15	16
合计	1	29	30

生存曲线的比较 -- Log-rank检验

- （4）计算各组合计的实际死亡数和理论死亡

表 6. 肿瘤<3.0cm和肿瘤≥3.0cm生存曲线比较的log-rank检验计算表

序号	时间	肿瘤<3.0cm				肿瘤≥3.0cm				合计	
		n_{1i}	d_{1i}	T_{1i}	V_{1i}	n_{2i}	d_{2i}	T_{2i}	V_{2i}	n_i	d_i
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
1	6	14	0	0.4667	0.2489	16	1	0.5333	0.2489	30	1
2	7	14	0	0.4827	0.2497	15	1	0.5172	0.2497	29	1
3	9	14	0	0.5000	0.2500	14	1	0.5000	0.2500	28	1
…											
合计	—	—	11	17.5416	5.8064	—	16	9.4584	5.8064	—	27

生存曲线的比较 -- Log-rank检验

- （5）计算 χ^2 统计量

肿瘤<3.0cm：$\chi^2=\frac{(11-17.5416)^2}{5.8064}=7.37$　　$\nu=1$

肿瘤≥3.0cm：$\chi^2=\frac{(16-9.4584)^2}{5.8064}=7.37$　　$\nu=1$

- 查界值表，0.005<*P*<0.010，按*a*=0.05水准，拒绝H_0，接受H_1，可认为两条生存曲线位置不同，肿瘤<3.0cm患者生存率高于肿瘤≥3.0cm患者。

生存曲线的比较 -- Log-rank检验

- Log-rank检验应用及注意事项：
- [1]. 相对死亡比（relative death ratio）：实际死亡数A与理论死亡数T之比，则相对危险度（relative risk，RR）估计值为两组相对死亡比率之比。肿瘤<3.0cm患者与肿瘤≥3.0cm患者相比

$$RR = \frac{R_1}{R_2} = \frac{A_1 / T_1}{A_2 / T_2} = \frac{11/17.5416}{16/9.4584} = 0.37$$

- 即肿瘤<3.0cm患者死亡风险是肿瘤≥3.0cm患者死亡风险的37%；肿瘤≥3.0cm患者对肿瘤<3.0cm患者RR= 2.69，即肿瘤≥3.0cm 患者死亡风险是肿瘤<3.0cm患者死亡风险的2.69倍 。

生存曲线的比较 -- Log-rank检验

- [2]. log-rank检验：用于整条生存曲线的比较，若比较两条生存曲线某时点的生存率，如2年生存率，按下式

$$u=\frac{S_1(t)-S_2(t)}{\sqrt{SE^2[S_1(t)]+SE^2[S_2(t)]}}$$

- 【案例】数据，肿瘤<3.0cm患者组和肿瘤≥3.0cm患者组2年生存率分别是0.8572和0.4375，标准误分别为0.0935和0.1240，

$$u=\frac{0.8572-0.4375}{\sqrt{0.0935^2+0.1240^2}}=2.70$$

- *P*<0.01，两组间2年生存率差别有统计学意义。

生存曲线的比较 -- Log-rank检验

- [3]. 若比较多个时点生存率，检验水准应取Bonferroni校正，

即 $\alpha' = \alpha / k$ 其中*k*为比较次数，以保证总的Ⅰ型错误概率不超过 α 。

- [4]. log-rank检验：单因素分析，应用条件是除比较因素外，影响生存率的各混杂因素组间均衡可比，否则采用Cox比例风险回归模型。

- [5]. 对数秩检验也可用于三组生存曲线的比较。

生存曲线的比较 -- Log-rank检验

- [6]. 由对数秩检验过程可知，若每一时点A组死亡率都高一点(生存率低一点)，则检验结果必然为A不同于B。因此，在比较的两条生存率曲线无交叉时，直接用对数秩检验是合适的。反之，就需进一步分析原因，了解是否存在混杂因素的影响。

生存曲线的比较 -- Log-rank检验案例1

- **例15-3** 观察两组卵巢腺癌患者的病程天数如下。请用乘积极限法进行描述，并比较两组的生存期差异有无统计学意义，并作生存率曲线。如表15-2 所示。

表 15-2　　病程天数表

A 组（低恶性高分化癌）：	28	29	175	195	309	377+	393+	421+	447+	452	709+	744+	770+	1106+	1206				
B 组（高恶性低分化癌）：	34	88	137	199	280	291	299+	300+	309	351	358	369	370	375	382	392	429+	451	1119+

Log-rank检验案例1 代码


```
> #install.packages("survival")
> library(survival)
> example15_3 <- read.table ("example15_3.csv", header=TRUE, sep=",")
> attach(example15_3)
> total  <- survfit(Surv(t, censor)~1)
> summary(total)
> plot(total)
> separate  <- survfit(Surv(t, censor)~group)
> summary(separate)
> plot(separate, lty = c('solid','dashed'), col=c('black','blue’ ), xlab='survival
time in days',ylab='survival probabilities')
> legend('topright', c('Group A',' Group B'), lty=c('solid','dashed’ ),
col=c('black','blue'))
> survdiff(Surv(t, censor)~group)
```

Log-rank检验案例2 代码


```
> library(coin) #载入coin包
> data(glioma) #加载数据集“glioma”

> library(survival) #载入生存分析survival包
> g3 <- subset(glioma, histology == ‘Grade3’) #提取数据集的子集

> fit <- survfit(Surv(time, event)~group, data = g3) #拟合生存函数
> plot(fit, lty = c(2,1), col = c(2,1)) #画生存曲线
> legend('bottomright', legend = c('Control','Treatment'), lty = c(2,1), col = c(2,1))

> survdiff(Surv(time, event)~group,data = g3) #两组比较 survdiff {survival}

> logrank_test(Surv(time, event)~group, data = g3, distribution ="exact")
> logrank_test(Surv(time, event)~group|histology, data = glioma, distribution =
approximate(B = 1000)) #两组比较,coin包 logrank_test函数 #SurvivalTests {coin}
```

如何画一幅高水准的生存曲线


```
> library(survival)
> library(survminer)
> fit <- survfit(Surv(time, status) ~ sex, data = lung)

> ggsurvplot(fit,
           pval = TRUE, #在图上添加log rank检验的p值
           conf.int = TRUE,#添加置信区间
           risk.table = TRUE, #在图下方添加风险表
           risk.table.col = "strata", # 根据数据分组为风险表添加颜色
           linetype = "strata", # 改变不同组别的生存曲线的线型
           surv.median.line = "hv", # 标注出中位生存时间
           ggtheme = theme_bw(), # 改变图形风格
           palette = c("#E7B800", "#2E9FDF"))#图形颜色风格
```

画一幅高水准的生存曲线 续


```
> ggsurvplot (
  fit,
  pval = FALSE,
  conf.int = TRUE,
  fun = "cumhaz",
  conf.int.style = "ribbon",  # 设置置信区间的风格
  xlab = "Time in days",   # 设置x轴标签
  break.time.by = 200,     # 将x轴按照200为间隔进行切分
  ggtheme = theme_light(), # 设置图形风格
  risk.table = "abs_pct",  # 在风险表中添加绝对数和相对数
  risk.table.y.text.col = TRUE,# 设置风险表的文字颜色
  risk.table.y.text = FALSE,# 以条柱展示风险表的标签，而非文字
  ncensor.plot = TRUE,      # 展示随访过程中不同时间点死亡和删失的情况
  surv.median.line = "hv",  # 添加中位生存时间
  legend.labs =
   c("Male", "Female"),    # 改变图例标签
  palette =
   c("#E7B800", "#2E9FDF") # 设置颜色
   )
```

画一幅高水准的生存曲线 续


```
> ggsurvplot(fit,
           conf.int = TRUE,
           risk.table.col = "strata",
           ggtheme = theme_bw(),
           palette = c("#E7B800", "#2E9FDF"),
           fun = "cumhaz")
> dev.off()
```

8.3 COX比例风险模型

COX比例风险回归模型简介

- 1972年，英国统计学家 D. R. Cox 博士提出了一种比例风险回归模型（Cox Proportional Hazard Model），简称Cox模型。它可以分析多种因素对生存时间的影响，而且允许有“截尾”存在。是生存分析中最重要的模型之一。
- Cox回归模型主要用于肿瘤和其它慢性病的预后因素分析，也可以用于一般的临床疗效评价和队列的病因探索。

COX模型的基本结构 I

- COX模型不直接考察生存时间与各自变量的关系，而是用风险率作为因变量。

  COX模型的基本结构为：

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

COX模型的基本结构 II

- $h(t,X)$：t时点上m个危险因素起作用时的风险率，即在时间t上的死亡率；
- $h0(t)$：某时间t上当m个危险因素为0时的基准风险率；
- $X=(X_1, X_2,..., X_m)$：与生存时间可能有关的自变量；
- $\beta = (\beta_1, \beta_2,..., \beta_m)$：COX模型的回归系数。

COX模型的基本结构 III

- β_j与h(t,X)之间有如下关系：
  - （1）$\beta_j>0$，则X_j取值越大，h(t,X)的值越大，表示病人死亡的风险率越大；
  - （2）$\beta_j=0$，则X_j取值对h(t,X)无影响；
  - （3）$\beta_j<0$，则X_j取值越大，h(t,X)的值越小，表示病人死亡的风险率越小。

$$h(t,X)=h_0(t)\exp(\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m)$$

COX模型的基本结构 IV

- h(t)和h0(t)成比例，比例系数是：

$$h(t,X)/h_0(t)=\exp(\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m)$$

- 故COX模型又称比例风险模型，将上式两边取自然对数，得：

$$\ln[h(t,X)/h_0(t)]=\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m$$

- 此式与多元线性回归模型非常类似，故有人称COX模型为COX回归。
- 由此式可见β_j的含义是：在其他自变量不变前提下，自变量X_j改变一个单位，引起的死亡风险改变的自然对数值。

COX模型的基本结构 V

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

上式可改写为：

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

- 相对危险度（RR）= $\exp\beta_j (X_{j2}-X_{j1})$，如$X_j$为 0~1 数据，则：RR=$\exp\beta_j$
- RR含义：在其他自变量保持不变前提下，自变量X_j改变一个单位，死亡风险比原水平改变$\exp(\beta_j)$倍。RR是一个与时间无关的变量 。

COX模型的基本结构 VI

- h0(t)分布类型未作任何限定；但h(t)随变量X的变化假定为指数函数exp(bX)。故COX模型为半参数模型。而且h0(t)分布类型未作任何限定，因而应用COX模型不必考虑资料的属于那一种具体的分布。故适用范围广泛，类似于非参数方法，但其检验效率高于非参数模型，接近于参数模型。

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

COX回归模型的构建方法

- 构造偏似然函数，然后用最大似然法求出各参数估计值b_j，须借助计算机完成。

$$h(t,X)/h_0(t)=\exp(\beta_1X_1+\beta_2X_2+\ldots+\beta_mX_m)$$

COX回归模型的主要用途

- （1）建立以多个危险因素估计生存或死亡的 风险模型，并由模型估计对多个危险因素导致死亡的相对危险度（RR)
- （2）用已建立的模型，估计患病后随时间变化的生存率
- （3）用已建立的模型，估计患病后的危险指数, 或预后指数（PI）。

COX回归模型的应用条件

- （1）已知观察对象的生存时间；
- （2）已知观察对象在事先确定的观察时间内，其是否发生某事件的结果；
- （3）自变量可以是计量资料、计数资料、分类资料或等级资料。
- （4）等比例风险（PH）。指在协变量不同状态的病人的风险在不同的时间保持不变。如在研究的10年中，糖尿病人心脏病发作的可能性是非糖尿病人的3倍，无论在第1年，第2年.....等都如此。

COX回归模型的应用条件 -- 等比例风险的验证

- （1）按协变量分组的Kaplan-Meier生存曲线，如生存曲线明显交叉，则不满足PH假定。
- （2）将协变量与时间作为交互项引入模型，如果交互项没有统计学意义，则等比例风险成立，若有统计学意义，则不成立。与时间有关的风险称为非比例风险，采用非比例风险模型分析。

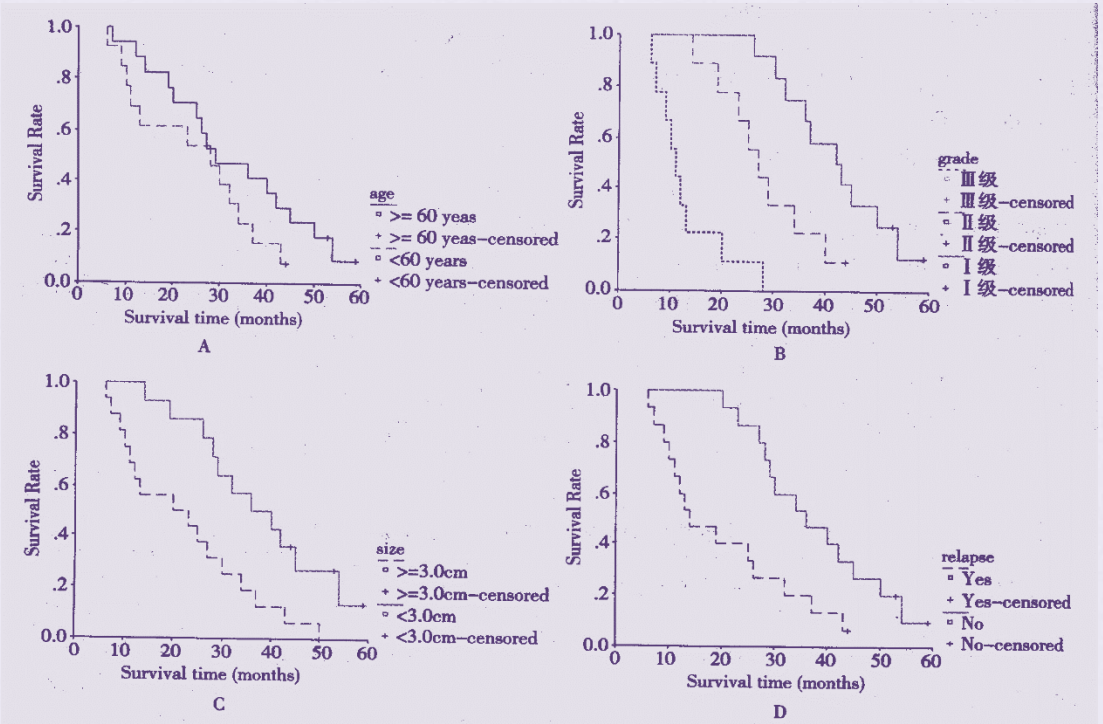


图 16－5　膀胱肿瘤数据 4 个变量的生存曲线

（A 年龄；B 肿瘤分级；C 肿瘤大小；D 是否复发）

COX回归分析的假设检验 I

- （1）Cox回归方程的检验方法：
  - 最大似然比检验（maximumLike-lihood Ratio)-
  - 常用Wald检验、得分检验（Score)

H0: $\beta_1=\beta_2=.....=\beta_p=0$

H1: 各$\beta_j(j=1,2,...,p)$不全为0

COX回归分析的假设检验 II

- （2）Cox回归系数的检验方法：Wald检验
  - 检验统计量为：X2 = $\left(\frac{b_j}{s_{bj_j}}\right)^2$
  - bj为βj的估计值，Sbj为bj的标准误。
  - X2服从自由度　=1的X2分布

H0: βj=0

H1: βj(j=1,2,…,p)≠ 0

COX回归分析的一般步骤 -- 收集资料

- 【案例 】30例恶性膀胱肿瘤患者的生存分析。

- **（1）收集资料**

首先确定观察指标并将其数量化，表1（数量化表），然后收集资料，表2（随访表）。收集到资料后，建立数据文件。（用SPSS或Excel)

表 16－1　膀胱肿瘤患者生存资料变量赋值表

变量 (1)	因素 (2)	分组及赋值 (3)
age	年龄	岁
grade	肿瘤分级	Ⅰ级＝1；Ⅱ级＝2；Ⅲ级＝3
size	肿瘤大小（cm）	<3.0＝0；≥3.0＝1
relapse	是否复发	未复发＝0；复发＝1
start	手术日期	月/日/年
end	终止观察日期	月/日/年
t	生存时间	月
status	生存结局	截尾＝0；死亡＝1

【案例 】30例恶性膀胱肿瘤患者的生存分析

表 16 -2　6 例膀胱肿瘤患者生存资料原始记录表

id	age	grade	size	relapse	start	end	t	status	结局
1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	62	1	0	0	01/10/1996	11/20/2000	59	0	存活
2	64	1	0	0	03/05/1996	08/12/2000	54	1	死亡
3	52	2	0	1	04/09/1996	12/03/1999	44	0	失访
4	60	1	0	0	06/06/1996	10/27/2000	53	0	死于其他
5	59	2	1	0	07/20/1996	06/21/1998	23	1	死亡
6	59	1	1	1	08/19/1996	09/10/1999	37	1	死亡

【案例 】30例恶性膀胱肿瘤患者的生存分析

表 16－8　30 例膀胱肿瘤患者生存资料原始记录表

id (1)	age (2)	grade (3)	size (4)	relapse (5)	start (6)	end (7)	t (8)	status (9)	PI (10)	$S(t)$ (11)
1	62	1	0	0	01/10/1996	11/20/2000	59	0	1.680	0.256
2	64	1	0	0	03/05/1996	08/12/2000	54	1	1.680	0.256
3	52	2	0	1	04/09/1996	12/03/1999	44	0	4.339	0.018
4	60	1	0	0	06/06/1996	10/27/2000	53	0	1.680	0.512
5	59	2	1	0	07/20/1996	06/21/1998	23	1	4.438	0.662
6	59	1	1	1	08/19/1996	09/10/1999	37	1	3.737	0.249
7	63	1	1	0	09/16/1996	10/20/2000	50	1	2.758	0.139
8	62	1	0	0	09/20/1996	09/18/1999	36	1	1.680	0.859
9	50	1	1	0	09/26/1996	03/22/1999	30	1	2.758	0.760
10	26	1	1	1	11/04/1996	05/25/2000	43	1	3.737	0.110
11	43	2	1	0	01/10/1997	11/08/1999	34	1	4.438	0.131
12	62	1	0	0	02/16/1997	11/10/2000	45	1	1.680	0.646
13	67	1	0	0	03/09/1997	08/18/2000	42	1	1.680	0.785
14	70	2	0	0	03/28/1997	07/20/2000	40	1	3.360	0.328
15	56	1	0	1	04/03/1997	11/10/1999	32	1	2.659	0.747
16	85	2	0	1	04/15/1997	11/20/1998	19	1	4.339	0.801
17	65	1	0	1	08/06/1997	09/28/1999	26	1	2.659	0.894
18	54	3	1	1	11/10/1997	12/09/1998	13	1	7.097	0.155
19	62	2	0	0	02/19/1998	07/20/2000	29	1	3.360	0.659
20	52	3	0	0	03/14/1998	07/02/2000	28	1	5.040	0.163
21	63	2	1	0	06/10/1998	09/01/2000	27	1	4.438	0.446
22	50	3	1	1	06/15/1998	04/14/1999	10	1	7.097	0.517
23	83	2	1	1	09/03/1998	09/20/2000	25	1	5.417	0.246
24	61	3	1	0	10/10/1998	06/13/2000	20	1	6.118	0.181
25	57	3	1	1	01/16/1999	12/20/1999	11	1	7.097	0.396
26	63	2	0	1	02/17/1999	04/20/2000	14	1	4.339	0.845
27	72	3	1	1	05/10/1999	05/12/2000	12	1	7.097	0.276
28	56	3	1	1	09/15/1999	06/17/2000	9	1	7.097	0.638
29	73	3	1	1	12/19/1999	07/26/2000	7	1	7.097	0.759
30	54	3	1	1	03/10/2000	09/20/2000	6	1	7.097	0.879

COX回归分析的一般步骤 -- 因子初步筛选

何为 crude HR?
何为adjusted HR?
何为单因素分析?
何为多因素分析?

- **（2）因子初步筛选**

A. 剔除缺失数据较多的因子。

B. 剔除变异几乎为零的因子。

C. 对所有的因子逐个作单因素Cox模型分析，选择有统计意义的变量作多因素COX模型分析。此时的α值可以取稍大些，如α=0.1 。

COX回归分析的一般步骤 -- 拟合多因素模型

- **（3）拟合多因素模型**

  A. 规定检验水准α，初步的探索性研究， 可取 $\alpha = 0.10$ 或 $\alpha = 0.15$；严谨的、证实性研究，取$\alpha = 0.05$ 或 $\alpha = 0.01$ 。

  B. 筛选因子方法：前进法、后退法、逐步法。

- **临床上如何进行自变量的筛选？是否与统计学筛选方法一致？**

COX回归分析的一般步骤 -- 结果解析与评价

- **（4）结果解析与评价**

①. 模型在一定的检验水准α下，入选哪些因素?

②. 入选因素哪些是保护因素，哪些是危险因素?

③. 入选因素哪个对因变量影响（贡献）最大?

30例膀胱癌患者Cox回归分析结果

方程中的变量

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI 用于 Exp(B) 下限	95.0% CI 用于 Exp(B) 上限
步骤 3	grade	1.680	.382	19.385	1	.000	5.367	2.540	11.341
	size	1.078	.460	5.493	1	.019	2.939	1.193	7.242
	relapse	.979	.460	4.525	1	.033	2.662	1.080	6.560

- 采用前进逐步法，在α=0.05水准上，在所分析的4个因素中，入选模型有3个因素：肿瘤分级、肿瘤大小和是否复发为膀胱肿瘤患者独立的影响因素。三者回归系数均为正，为膀胱肿瘤患者死亡的危险因素。

30例膀胱癌患者Cox回归分析结果

方程中的变量

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI 用于 Exp(B) 下限	上限
步骤 3	grade	1.680	.382	19.385	1	.000	5.367	2.540	11.341
	size	1.078	.460	5.493	1	.019	2.939	1.193	7.242
	relapse	.979	.460	4.525	1	.033	2.662	1.080	6.560

- grade 的RR=5.367,即肿瘤分级每增加一个等级，死亡风险增加4.367倍；Size的RR=2.393，肿瘤大于等于3.0cm者，死亡风险是小于3者的2.939倍；Relapse 的RR=2.662,即复发者死亡风险是不复发者的2.662倍。

30例膀胱癌患者Cox回归分析结果

方程中的变量

		B	SE	Wald	df	Sig.	Exp(B)	95.0% CI 用于 Exp(B) 下限	上限
步骤 3	grade	1.680	.382	19.385	1	.000	5.367	2.540	11.341
	size	1.078	.460	5.493	1	.019	2.939	1.193	7.242
	relapse	.979	.460	4.525	1	.033	2.662	1.080	6.560

- grade 、size、 relapse的标准差分别为：0.845、0.507、0.509，其标准回归系数分别是：1.42、0.55、0.50。
- 从标准回归系数来看， grade对生存（或死亡）的影响最大，其次是size，最后是replapse 。

标准回归系数的计算方法

- A. 对原始变量的观察值作标准正态化变换后，在拟合回归方程得到的回归系数，即为标准化回归系数。
- B. 直接用以下公式计算：

$$\beta' = SD \cdot \beta$$

$$SE(\beta') = SD \times SE(\beta)$$

- 式中，SD为自变量X的标准差, SE(β)为β的标准误。

COX回归分析结果评价小结

- 1. 采用前进逐步法，在α=0.05水准上，从所分析的4个因素中，筛选出3个与膀胱肿瘤患者生存有关的因素：grade 、size、 relapse。
- 2. grade 、size、 relapse都是危险因素。
- 3. grade对膀胱肿瘤患者生存的影响最大，relapse的影响最小。

COX回归R语言操作【案例 1】

- 【案例 1】以下数据是一项关于胰脏癌手术中接受放射治疗是否会延长病人生存时间的研究数据。该研究的主要终点为死亡，接受手术被定义为计算生存时间的起点。由于该研究是一项未经随机化的观察性研究，要正确估计术中接受放射治疗提高患者生存时间的效果，还需要考虑对其他因子的效果进行调整。数据的详细说明见下表。

表 1. 胰脏癌术中放疗效果研究变量数据说明

变量名	变量说明	变量类型	分类变量的编码
casno	患者编号		
time	生存时间（月）	连续	
censor	删失	二分类	0：死亡、1：删失
age	手术时的年龄	连续	
trt	处理组别（有无术中放疗）	二分类	0：无术中放疗、1：有术中放疗
sex	性别	二分类	0：男、1：女
bui	占位处	二分类	0：胰脏头部、1：头部以外
ch	胰胆管浸润程度	有序多分类	1：ch0；2:ch1；3:ch2；4：ch3
p	有无腹膜转移	二分类	0：无、1： 有
stage	TNM分类	二分类	3：III期、4：IV期

COX回归R语言操作【案例 1】代码


```
> library(foreign)
> library(survival)
> pancer <- read.spss('pancer.sav')
> pancer <- as.data.frame(pancer)
> head(pancer)
> pancer$censor <- ifelse(pancer$censor=='死亡',1,0)
> pancer$Gender <- as.factor(ifelse(pancer$sex=='男',"Male","Female"))
> pancer$ch <- as.factor(ifelse(pancer$ch=='CH3', "ch","noch"))
> #pancer$ch <- relevel(pancer$ch,ref="CH0") #设置因子的参照水平
> #pancer$ch<- factor(pancer$ch,order=TRUE) #设置为等级变量
> #options(contrasts=c("contr.treatment", "contr.treatment")) #指定等级变量的参照水
平
> #pancer$Gender <- relevel(pancer$Gender,ref='Female’ )
> f<-coxph(Surv(time,censor==1)~age+Gender+trt+bui+ch+p+stage,data=pancer)
> summary(f)
> sum.surv<-summary(f)
> c_index<-sum.surv$concordance
> c_index
```

COX回归R语言操作【案例 2】

- 例15-4 随访25例分别以A 、B 治疗方法治疗的癌症病人的生存情况，资料如表15-3 所示，+为截尾值。1 : 有肾功能损害， 0: 无肾功能损害，请试作Cox 回归分析。

表 15-3　　A、B 治疗方法发生肾功能损害的情况

A 疗 法			B 疗 法		
编　　号	肾功损害	生存日数	编　　号	肾功损害	生存日数
1	1	8	13	1	13
12	0	52	16	1	18
5	1	58	25	1	23
8	1	63	11	0	70
21	1	63	10	0	76
7	0	220	2	0	180
24	0	365	9	0	195
4	0	452	20	0	210
18	0	496	3	0	232
22	0	528^{+}	17	0	300
19	0	560^{+}	23	0	396
15	0	676^{+}	14	0	490^{+}
			6	0	540^{+}

COX回归R语言操作【案例 2】代码


```
> library(survival)
> example15_4  <- read.table ("example15_4.csv", header=TRUE, sep=",")
> attach(example15_4)
> coxmodel  <- coxph(Surv(days, censor)~group)
> summary(coxmodel)
> coxmode2  <- coxph(Surv(days, censor)~group+renal)
> summary(coxmode2)
> anova(coxmodel,coxmode2)
> detach(example15_4)
```

COX回归模型案例3 代码

```
> data('GBSG2', package = 'TH.data')
> head(GBSG2)

> plot(survfit(Surv(time, cens)~horTh,data = GBSG2),lty = c(2,1), col = c(2,1),
mark.time = T)
> legend('bottomright', legend = c('yes','no'), lty = c(2,1), col = c(2,1))

> coxreg <- coxph(Surv(time,cens)~.,data = GBSG2) #构建Cox模型
> summary(coxreg)

> library(party)
> tree <- ctree(Surv(time,cens)~.,data = GBSG2)
> plot(tree)
```

C-Index计算方法

- 直接从survival包的函数coxph结果中输出，需要R的版本高于2.15.
- 利用函数survConcordance，这种方法和方法1类似，输出的结果相同
- 利用survcomp包，安装这个包我就不在这里赘述了。

```
age <- rnorm(200, 50, 10)
bp <- rnorm(200,120, 15)
d.time <- rexp(200)
cens <- runif(200,.5,2)
death <- d.time <= cens
os <- pmin(d.time, cens)
sample.data <- data.frame(age = age,bp = bp,os = os,death = death)
#让我们看一下生成的例子数据的前6行
head(sample.data)
library(survcomp)
fit <- coxph(Surv(os, death) ~ age + bp, data = sample.data)
cindex <- concordance.index(predict(fit),surv.time = sample.data$os, surv.event =
sample.data$death,method = "noether")
cindex$c.index; cindex$lower; cindex$upper
```

8.3 Cox比例风险模型 -- 列线图绘制

Cox回归模型案例1 列线图及校正曲线绘制 代码

```
> library(foreign)
> library(survival)
> library(rms)
> pancer <- read.spss('pancer.sav')
> pancer <- as.data.frame(pancer)
> head(pancer)
> pancer$censor <- ifelse(pancer$censor=='死亡',1,0)
> pancer$Gender <- as.factor(ifelse(pancer$sex=='男',"Male","Female"))
> pancer$ch <- as.factor(ifelse(pancer$ch=='CH3', "ch","nonch"))
> #pancer$ch <- relevel(pancer$ch,ref="CH0") #设置因子的参照水平
> #pancer$ch<- factor(pancer$ch,order=TRUE) #设置为等级变量
> #options(contrasts=c("contr.treatment", "contr.treatment")) #指定等级变量的参照水
平
> #pancer$Gender <- relevel(pancer$Gender,ref='Female')

> dd<-datadist(pancer)
> options(datadist='dd')
```

Cox回归模型案例1 列线图及校正曲线绘制 代码 续

```
> coxm <- cph(Surv(time,censor==1)~age+Gender+trt+bui+ch+p+stage, x=T, y=T,
data=pancer, surv=T)
> coxm
> summary(coxm)
> surv <- Survival(coxm)
> surv1 <- function(x)surv(1*3,lp=x)
> surv2 <- function(x)surv(1*6,lp=x)
> surv3 <- function(x)surv(1*12,lp=x)

> plot(nomogram(coxm,fun=list(surv1,surv2,surv3),lp= F,funlabel=c('3-Month
Survival probability','6-Month survival probability','12-Month survival
probability'),maxscale=100,fun.at=c('0.9','0.85','0.80','0.70','0.6','0.5','0.4','0.3','0.2','0.1'
)),xfrac=.30)
```

Cox回归模型案例1 列线图及校正曲线绘制 代码 续

```
> library(survival)
> f<-coxph(Surv(time,censor==1)~age+Gender+trt+bui+ch+p+stage,data=pancer)
> summary(f)
> sum.surv<-summary(f)
> c_index<-sum.surv$concordance
> c_index

> cal <- calibrate(coxm, cmethod='KM', method='boot', u=6, m=20, B=1000)
> plot(cal,lwd=2, lty=1, errbar.col=c(rgb(0,118,192,maxColorValue=255)),
xlim=c(0,1), ylim=c(0,1), xlab="Nomogram-Predicted Probabilityof 6 m OS",
ylab="Actual 6 m OS (proportion)", col=c(rgb(192,98,83,maxColorValue=255)))
> lines(cal[,c("mean.predicted","KM")], type="b", lwd=2,
col=c(rgb(192,98,83,maxColorValue=255)), pch=16)
> abline(0,1,lty=3,lwd=2,col=c(rgb(0,118,192,maxColorValue=255)))
```

8.3 Cox比例风险模型 -- 亚组森林图绘制

Cox回归亚组分析森林图案例 代码

Subgroup	No. of patients	IMNI	no IMNI	HR(95%CI)	p value			
Chemotherapy								
NAC	217	103(47.5%)	114(52.5%)	0.451(0.259-0.785)	<0.001	0.45	0.26	0.79
Adjuvant	630	287(45.6%)	343(54.4%)	0.539(0.389-0.747)	<0.001	0.54	0.39	0.75
Molecular subtype								
TNBC	142	68(47.9%)	74(52.1%)	(0.306-1.254)	0.184	0.62	0.31	1.25
other	705	380(54.0%)	324(46.0%)	0.511(0.377-0.693)	<0.001	0.51	0.38	0.69
N stage								
N0	106	52(49.0%)	54(51.0%)	0.257(0.052-1.268)	0.095	0.26	0.05	1.27
N1	317	151(47.6%)	166(52.4%)	0.543(0.316-0.934)	0.027	0.54	0.32	0.93
N2	255	108(42.4%)	147(57.6%)	0.741(0.462-1.189)	0.214	0.74	0.46	1.19
N3	169	104(61.5%)	65(38.5%)	0.460(0.291-0.727)	0.001	0.46	0.29	0.73
T stage								
T1	358	150(41.9%)	208(58.1%)	0.471(0.296-0.747)	0.001	0.47	0.3	0.75
T2	374	182(48.7%)	192(51.3%)	0.568(0.380-0.848)	0.006	0.57	0.38	0.85
T3-4	61	34(55.7%)	27(44.3%)	0.909(0.402-2.053)	0.818	0.91	0.4	2.05
pCR								
pCR	41	20(48.8%)	21(51.2%)	0.311(0.032-2.974)	0.31	0.31	0.03	2.97
Tumor location								
Latera	466	201(43.1%)	264(56.9%)	0.584(0.401-0.850)	0.005	0.58	0.4	0.85
Central/media	381	191(50.1%)	190(49.9%)	0.483(0.322-0.726)	<0.001	0.48	0.32	0.73
Overall	847	382(45.1%)	465(54.9%)	0.532(0.404-0.701)	<0.001	0.53	0.4	0.7

Cox回归亚组分析森林图案例 代码


```
> library(forestplot)
> test_forest <- read.csv('forest_test2.csv',header = FALSE)
> attach(test_forest)
> forestplot(labeltext = as.matrix(test_forest[,1:6]),
           #设置用于文本展示的列，此处我们用数据的前六列作为文本，在图中展示
           mean = test_forest$V7, #设置均值
           lower = test_forest$V8, #设置均值的上限
           upper = test_forest$V9, #设置均值的下限
           is.summary = c(T,T,F,F,T,F,F,T,F,F,F,F,T,F,F,F,T,F,T,F,F,T),
           #该参数接受一个逻辑向量，用于定义数据中的每一行是否是汇总值，若是，则在对应位置设置为TRUE，若否，
则设置为FALSE；设置为TRUE的行则以粗体出现
           zero = 1, #设置参照值，此处我们展示的是HR值，故参照值是1，而不是0
           boxsize = 0.2, #设置点估计的方形大小
           lineheight = unit(10,'mm'),#设置图形中的行距
           colgap = unit(8,'mm'),#设置图形中的列间距
           lwd.zero = 2,#设置参考线的粗细
           lwd.ci = 2,#设置区间估计线的粗细
           lwd.xaxis=2,#设置X轴线的粗细
           xlog=FALSE,#对数刻度
           ci.vertices.height = 0.1,#箭头高度
           clip = c(0.2,1.3),
           grid = FALSE,
           lty.ci = 1,
           col=fpColors(box='#458B00', summary= "#8B008B",lines = 'black',zero = '#7AC5CD'),
           #使用fpColors()函数定义图形元素的颜色，从左至右分别对应点估计方形，汇总值，区间估计线，参考线
           xlab="Hzard Ratio(HR)",#设置x轴标签
           graph.pos = 5)#设置森林图的位置，此处设置为5，则出现在第五列
```

8.4 竞争风险模型

竞争风险的概念

- 在观察某事件发生的时间，如果该事件被其他事件阻碍，即存在竞争风险
- 【举例】：研究骨髓移植对比血液移植治疗白血病的疗效，结局定义为“复发”，假定患者移植后不幸因为移植不良反应死亡，那这些发生移植相关死亡的患者就无法观察到“复发”的终点，也就是说“移植相关死亡”与“复发”存在竞争风险。
- 恶性肿瘤预后研究中这种例子很常见。

竞争风险模型R语言实现

- 数据下载地址： http://www.stat.unipg.it/luca/R/
- 下载后另存 .csv 格式的数据

变量	描述	标签值
Sex	性别	M=男，F=女
D	疾病	ALL，AML
Phase	疾病所处的阶段	CR1，CR2，CR3，Replase
Source	移植类型	BM+PB，PB
Age	年龄	年
Ftime	失败时间（事件发生时长）	月
Status	结局状态	0=删失，1=复发，2=竞争风险事件

竞争风险模型R语言code I


```
> library(foreign)
> bmt <-read.csv('bmtcrr.csv’ )
> head(bmt)
> bmt$D <- as.factor(bmt$D) #把变量“D”转换为因子类型变量

> library(survival)
> library(cmprsk) #加载竞争风险模型的程序包
> library(splines)

> attach(bmt)
> crmod <- cuminc(ftime,Status,D) #构建单因素生存函数
> crmod

> plot(crmod,xlab = '月', ylab = 'CIF’ , col = c('red','blue','orange','forestgreen'))
```

竞争风险模型R语言code II


```
> cov1 <- data.frame(age = bmt$Age,
                     sex_F = ifelse(bmt$Sex=='F',1,0),
                     dis_AML = ifelse(bmt$D=='AML',1,0),
                     phase_cr1 = ifelse(bmt$Phase=='CR1',1,0),
                     phase_cr2 = ifelse(bmt$Phase=='CR2',1,0),
                     phase_cr3 = ifelse(bmt$Phase=='CR3',1,0),
                     source_PB = ifelse(bmt$Source=='PB',1,0))  # 手动设置哑变量

> cov1

> mod1 <-  crr(bmt$ftime, bmt$Status, cov1, failcode=1, cencode=0) # 构建多因素竞
争风险模型
>summary(mod1)

> library(aod)
> wald.test(mod1$var,mod1$coef,Terms = 4:6) # 对模型回归系数进行建设检验
```

主要参考文献

【1】Robert I. Kabacoff 著, 《R语言实战 》(第2版), 人民邮电出版社, 2016

【2】Peter Dalgaard 著, 《R语言统计入门》》(第2版), 人民邮电出版社, 2014

【3】薛毅 陈立萍 著,《R语言实用教程》, 清华大学出版社, 2014

【4】张铁军 陈兴栋 刘振球 著,《R语言与医学统计图形》, 人民卫生出版社, 2018

【5】汪海波 萝莉 汪海玲 著, 《R语言统计分析与应用》， 人民邮电出版社，2018

Thanks!

感谢您的观看！