生物统计原理

生存分析

王强

April 24, 2020

南京大学生命科学学院

Outline

生存分析

定义

生存分析	是统计学的分支, 用于分析一个或多个事件发生前的预期持续时间 (Survival analysis).
事件	死亡、疾病发生、疾病复发、疾病进展或机械系统故障等 (Event).
时间	从观察期开始 (如手术或开始治疗) 到 1) 事件发生、2) 研究结束、3) 失去联系或退出研究的时间 (Time).

- 工程学 - 可靠性分析
- 经济学 - 持续时间分析
- 社会学 - 事件历史分析
- Time-to-Event (TTE) Analysis

数据

对于癌症

- 总生存 (Overall survival, OS)
  - Time from surgery to death
- 无进展生存 (Progression-free survival, PFS)
  - Time from start of treatment to progression
  - Time from response to recurrence

分析方法

- 描述生存时间
  - Kaplan–Meier 曲线
  - 生存函数
  - 生命表
- 比较生存时间
  - 秩和检验
  - 风险比
- 变量对生存的影响
  - Cox 比例风险回归

Kaplan–Meier curve

Basic Kaplan–Meier curve

PatientID	FollowUp	EventType	Scenario
Sarah	6	1	A
Joel	12	1	A
James	18	1	A
Daniel	24	1	A
Eric	30	1	A
Kaitlynn	36	1	A
Kelly	42	1	A
Morgen	48	1	A
Brandie	54	1	A
Luke	60	1	A

随访	Follow up, 可以用任何时间间隔, 天、月、年
事件类型	Event type, 1 表示事件发生, 在 Overall Survival 里, 代表死亡
分组	Scenario, 类别变量, 也可以是分组后的连续变量
分层	Strata, 除了分组的作用外, 有的软件会加上一些其它的意义
拟合	Fit

R 代码 i

```r
# R packages
library(tidyverse)
library(survival)

# Init data
dataset <- tibble(
    PatientID = c("Sarah", "Joel", "James", "Daniel", "E
    FollowUp  = c(1, 2, 3, 4, 5, 6, 7, 8, 9, 10) * 6,
    EventType = c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1),
    Scenario  = c("A", "A", "A", "A", "A", "A", "A", "A"
)
# dataset <- read_tsv("basic.tsv")

```

R 代码 ii

```r
# Define what the time column is dataset$FollowUp
#   and the event column is dataset$EventType
# Fit the Kaplan--Meier curve to this data `Scenario`
fit <- survfit(
    Surv(FollowUp, EventType) ~ Scenario,
    data = dataset
)

# Figure
plot(fit)
```

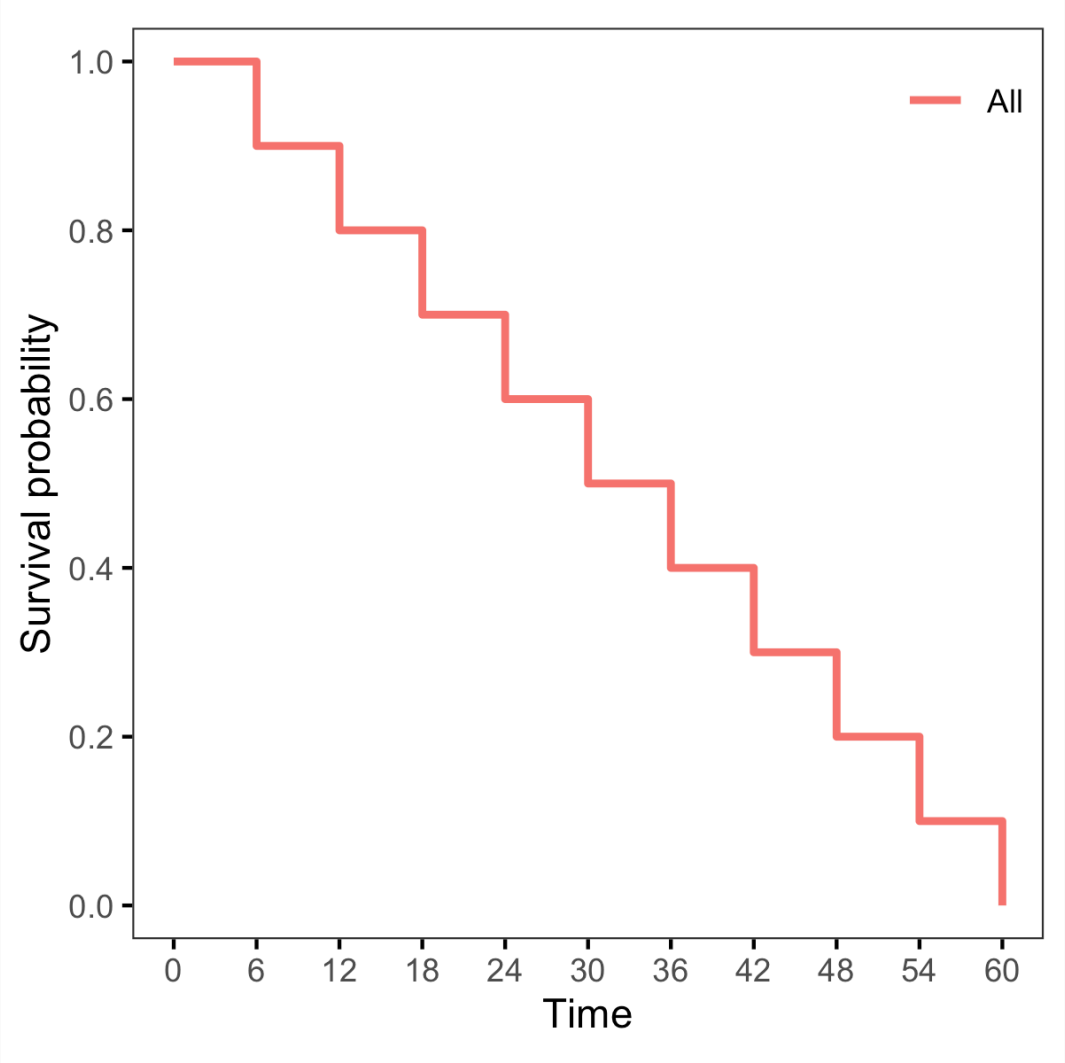


Figure 1. KM - basic

Kaplan–Meier curve with censored data

PatientID	FollowUp	EventType	Scenario
Sarah	6.0	1	A
Joel	12.0	1	A
James	18.0	0	A
Daniel	24.0	0	A
Eric	30.0	1	A
Kaitlynn	36.0	1	A
Kelly	37.2	1	A
Morgen	48.0	0	A
Brandie	54.0	1	A
Luke	60.0	0	A

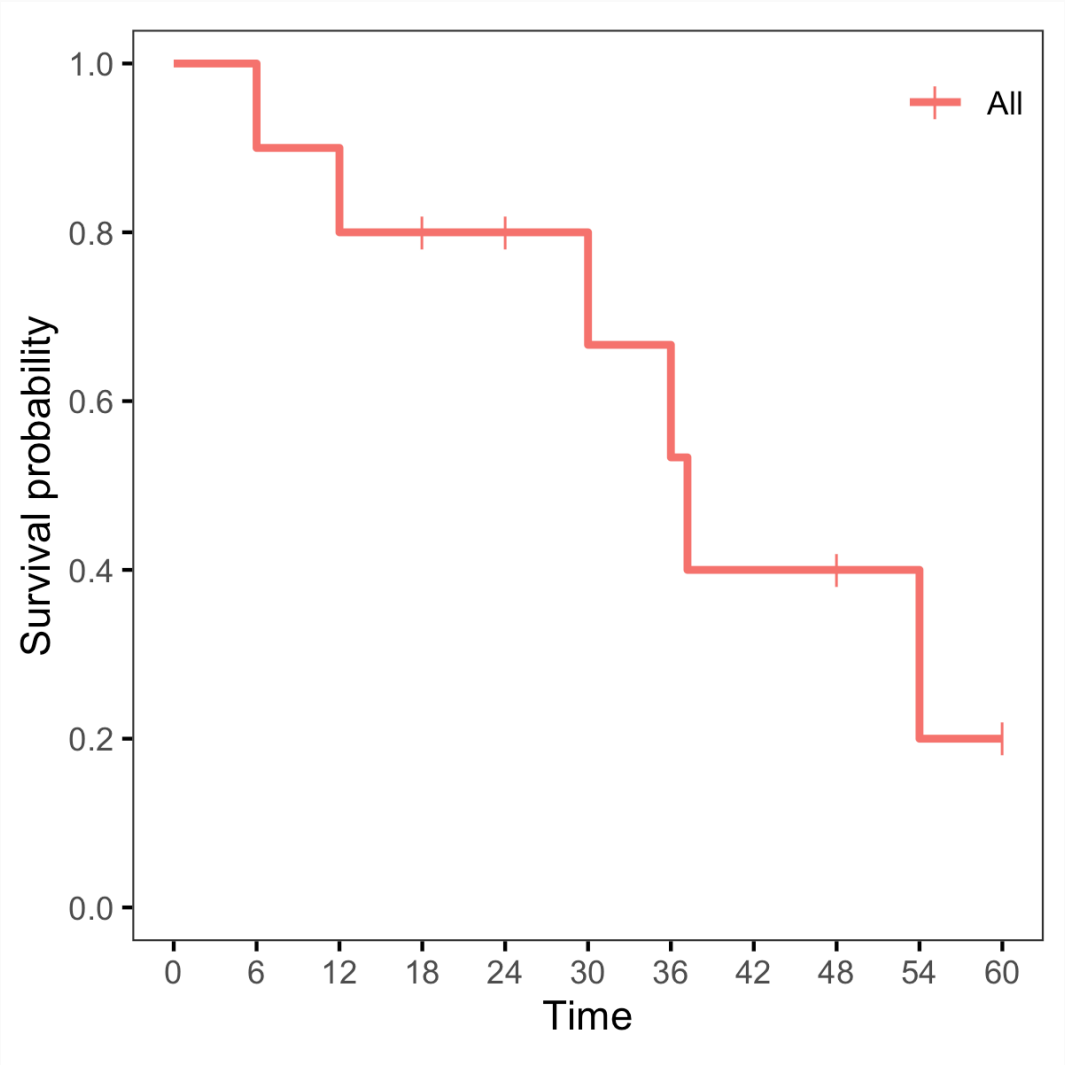


Figure 2. KM - censored

时间

从观察期开始到 1) 事件发生、2) 研究结束、3) 失去联系或退出研究的时间 (Time).

删失

如果一个受试者在观察时间内没有发生事件, 则被描述为被删失 (Censored/Censoring). 删失之后, 不观察或不了解该受试者的信息, 它可能发生也可能不发生事件.

Remove censored data?

感觉删失数据用处不大/不可靠/破坏完整性, 把它们去掉行不行?

PatientID	FollowUp	EventType	Scenario
Sarah	6.0	1	Actual
Joel	12.0	1	Actual
James	18.0	0	Actual
Daniel	24.0	0	Actual
Eric	30.0	1	Actual
Kaitlynn	36.0	1	Actual
Kelly	37.2	1	Actual
Morgen	48.0	0	Actual
Brandie	54.0	1	Actual
Luke	60.0	0	Actual

PatientID	FollowUp	EventType	Scenario
Sarah	6.0	1	Best
Joel	12.0	1	Best
James	60.0	0	Best
Daniel	60.0	0	Best
Eric	30.0	1	Best
Kaitlynn	36.0	1	Best
Kelly	37.2	1	Best
Morgen	60.0	0	Best
Brandie	60.0	1	Best
Luke	60.0	0	Best

PatientID	FollowUp	EventType	Scenario
Sarah	6.0	1	Worst
Joel	12.0	1	Worst
James	18.0	1	Worst
Daniel	24.0	1	Worst
Eric	30.0	1	Worst
Kaitlynn	36.0	1	Worst
Kelly	37.2	1	Worst
Morgen	48.0	1	Worst
Brandie	54.0	1	Worst
Luke	60.0	1	Worst

PatientID	FollowUp	EventType	Scenario
Sarah	6.0	1	Event
Joel	12.0	1	Event
Eric	30.0	1	Event
Kaitlynn	36.0	1	Event
Kelly	37.2	1	Event
Brandie	54.0	1	Event

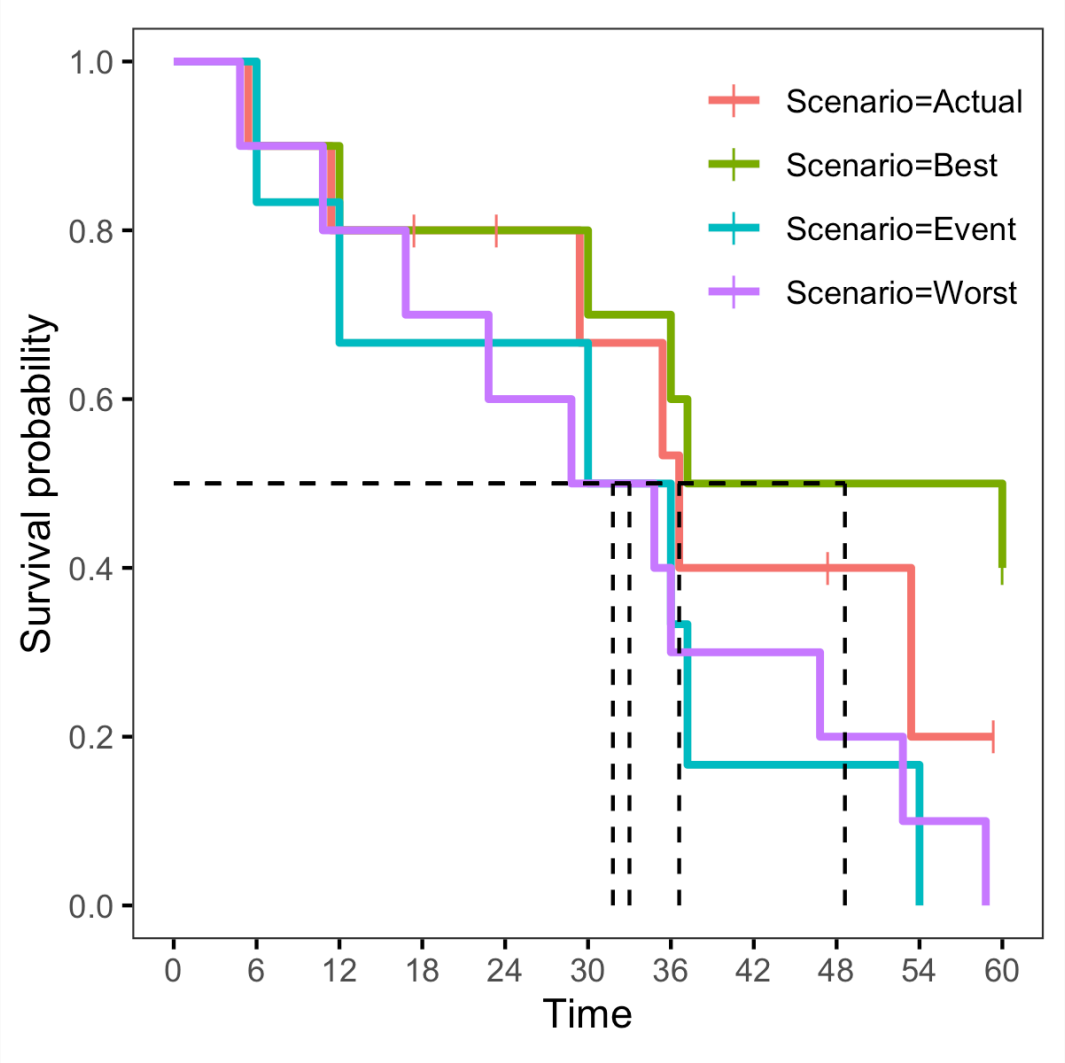


Figure 3. KM - actual/best/worst

Importance of confidence intervals

PatientID	FollowUp	EventType	Scenario
Brandie	54	1	A
Luke	60	0	A

PatientID	FollowUp	EventType	Scenario
Brandie	54	0	B
Luke	60	1	B

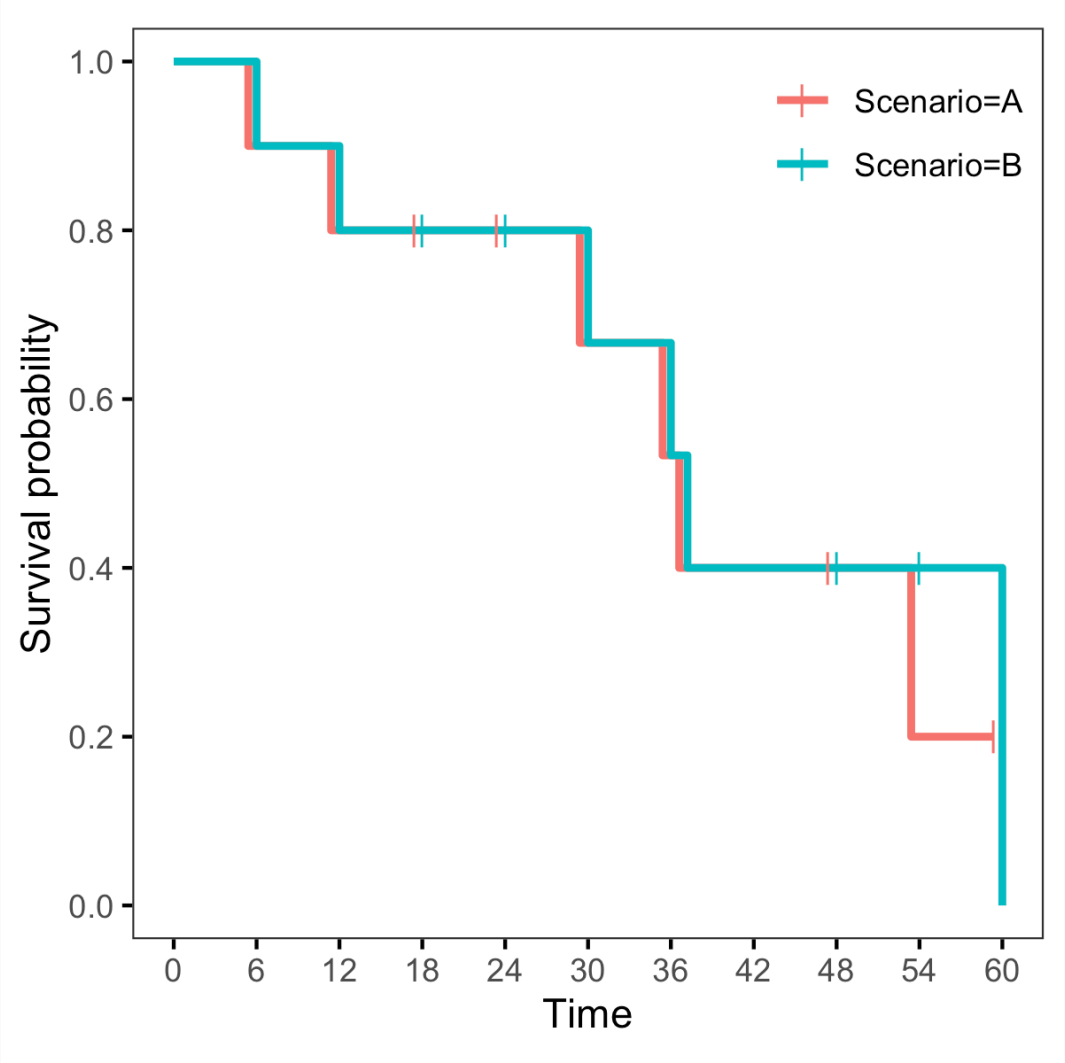


Figure 4. KM - scenarios

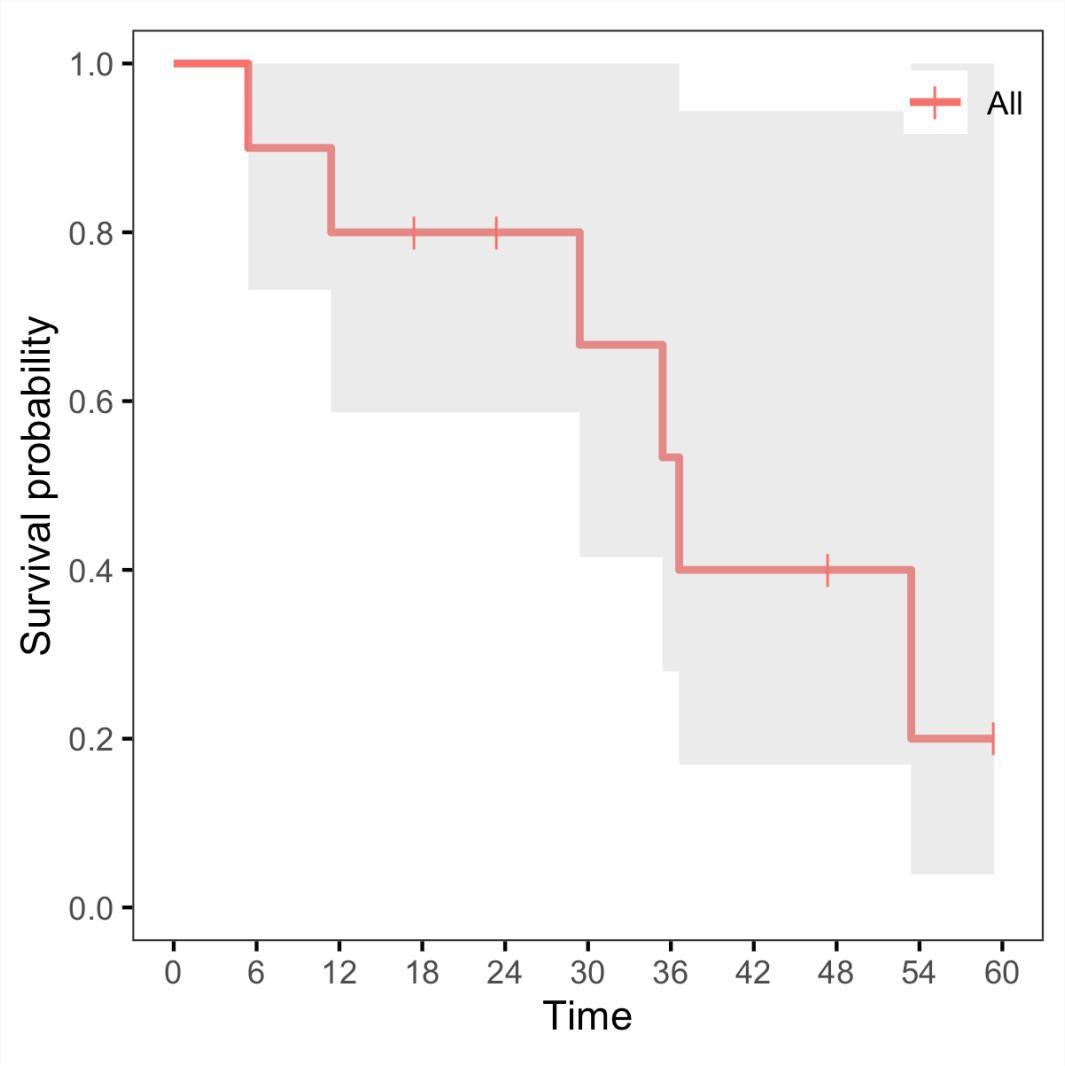


Figure 5. KM - CI

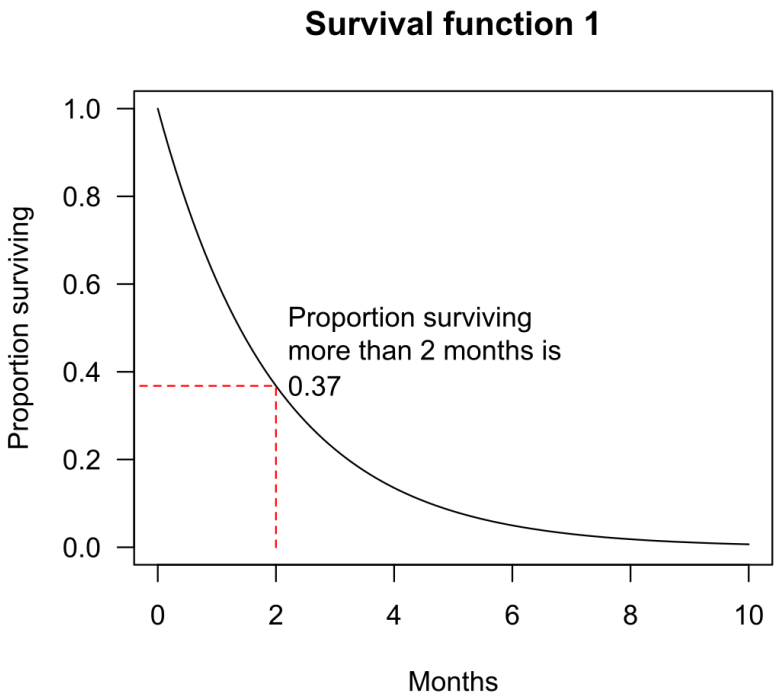
生存函数与风险比

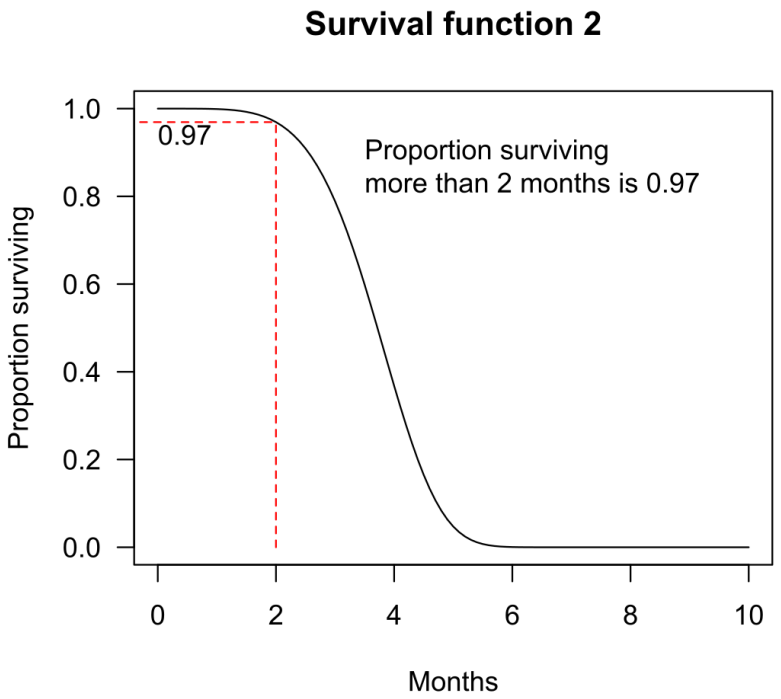
生存函数

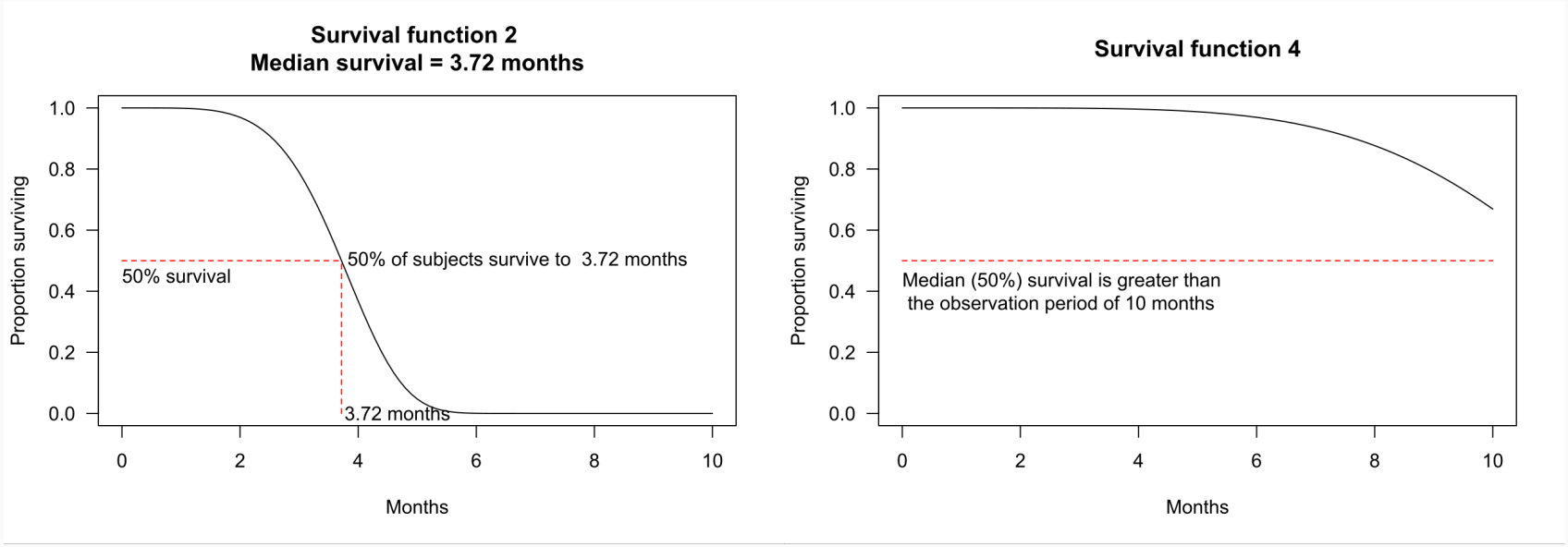
受试者在任何指定的时间内存活下来的概率

$$S(t) = Pr(\{T > t\}) = \int_t^\infty f(u)\,du$$

$S(t)$: 生存函数


Survival function 1
Proportion surviving
1.0
0.8
0.6
0.4
0.2
0.0
Proportion surviving
more than 2 months is
0.37
0
2
4
6
8
10
Months


Survival function 2
Proportion surviving
1.0
0.8
0.6
0.4
0.2
0.0
0.97
Proportion surviving
more than 2 months is 0.97
0
2
4
6
8
10
Months


Survival function 2
Median survival = 3.72 months
Proportion surviving
1.0
0.8
0.6
0.4
0.2
0.0
50% survival
50% of subjects survive to 3.72 months
3.72 months
0
2
4
6
8
10
Months
Survival function 4
Proportion surviving
1.0
0.8
0.6
0.4
0.2
0.0
Median (50%) survival is greater than
the observation period of 10 months
0
2
4
6
8
10
Months

Kaplan—Meier estimator

KM 方法是最常见的估计生存函数的方法

$$\widehat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

t_i: 至少有一个事件发生的时间

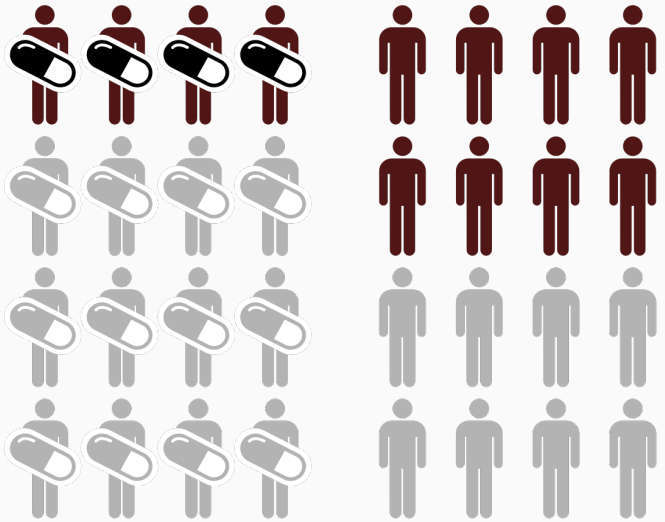
d_i: 在时间 t_i 时发生的事件 (如死亡) 的数量

n_i: 在时间 t_i 前存活的 (未发生事件或删失的) 个体数量

风险比 HR

相对危险

暴露组与未暴露组中事件发生率的比值 (Relative Risk, risk ratio, RR).



风险率　　单位时间内发生的事件数占被试总体的百分比 (Hazard rate).

$$h(t) = \lim_{\Delta t \to 0} \frac{\text{observed events in interval}[t, t + \Delta t]/N(t)}{\Delta t}$$

风险比　　指一个解释变量的两个水平所描述的条件对应的风险率之比 (Hazard Ratio).

比例风险模型, Proportional hazards model

例题: HR

研究人员调查了卡介苗再接种是否会降低儿童死亡率. 在西非的几内亚比绍进行了一项随机对照试验. 在 19 个月大时, 共有 1437 名儿童被随机选择接受卡介苗再接种, 1434 名儿童被选为对照组 (无再接种). 所有受试者在出生时就已接种过卡介苗, 且在接种当天对结核菌素的反应性较低或无反应, 且无严重疾病.

儿童从 19 个月到 5 岁都接受了随访. 在随访期间, 有 77 名儿童死亡. 与对照组相比, 再接种疫苗的儿童的死亡风险比 (HR) 为 1.2 (95%置信区间为 0.77 至 1.89).

以下哪项陈述是对的?

1. 与对照组相比, 再接种疫苗的儿童在随访期间的任何时间死亡风险增加了 20%.
2. 再接种组的风险率 (hazard rate) 在随访期间假设为恒定.
3. 对照组儿童的生命长度与再次接种疫苗的儿童相比增加了 20%.
4. 可以推断, 与对照组相比, 接种了再接种疫苗的儿童在 10 岁之前的死亡危险率为 1.2.

样例数据 aml

ID	time	status	x
12	5	1	Nonmaintained
13	5	1	Nonmaintained
14	8	1	Nonmaintained
15	8	1	Nonmaintained
1	9	1	Maintained
16	12	1	Nonmaintained
2	13	1	Maintained
3	13	0	Maintained
17	16	0	Nonmaintained
4	18	1	Maintained
5	23	1	Maintained
18	23	1	Nonmaintained
19	27	1	Nonmaintained
6	28	0	Maintained
20	30	1	Nonmaintained
7	31	1	Maintained
21	33	1	Nonmaintained
8	34	1	Maintained
22	43	1	Nonmaintained
9	45	0	Maintained
23	45	1	Nonmaintained
10	48	1	Maintained
11	161	0	Maintained

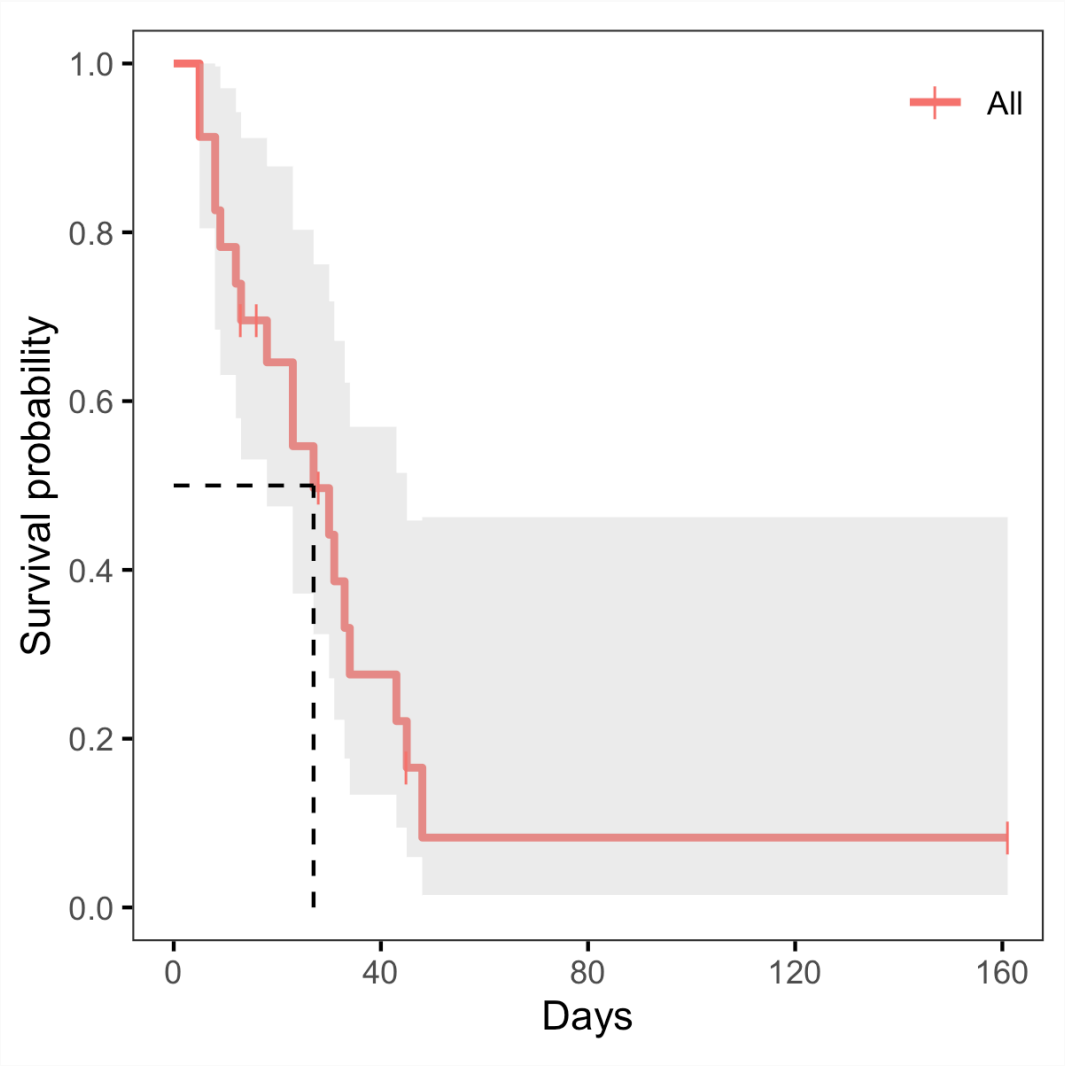


Figure 6. AML - survival

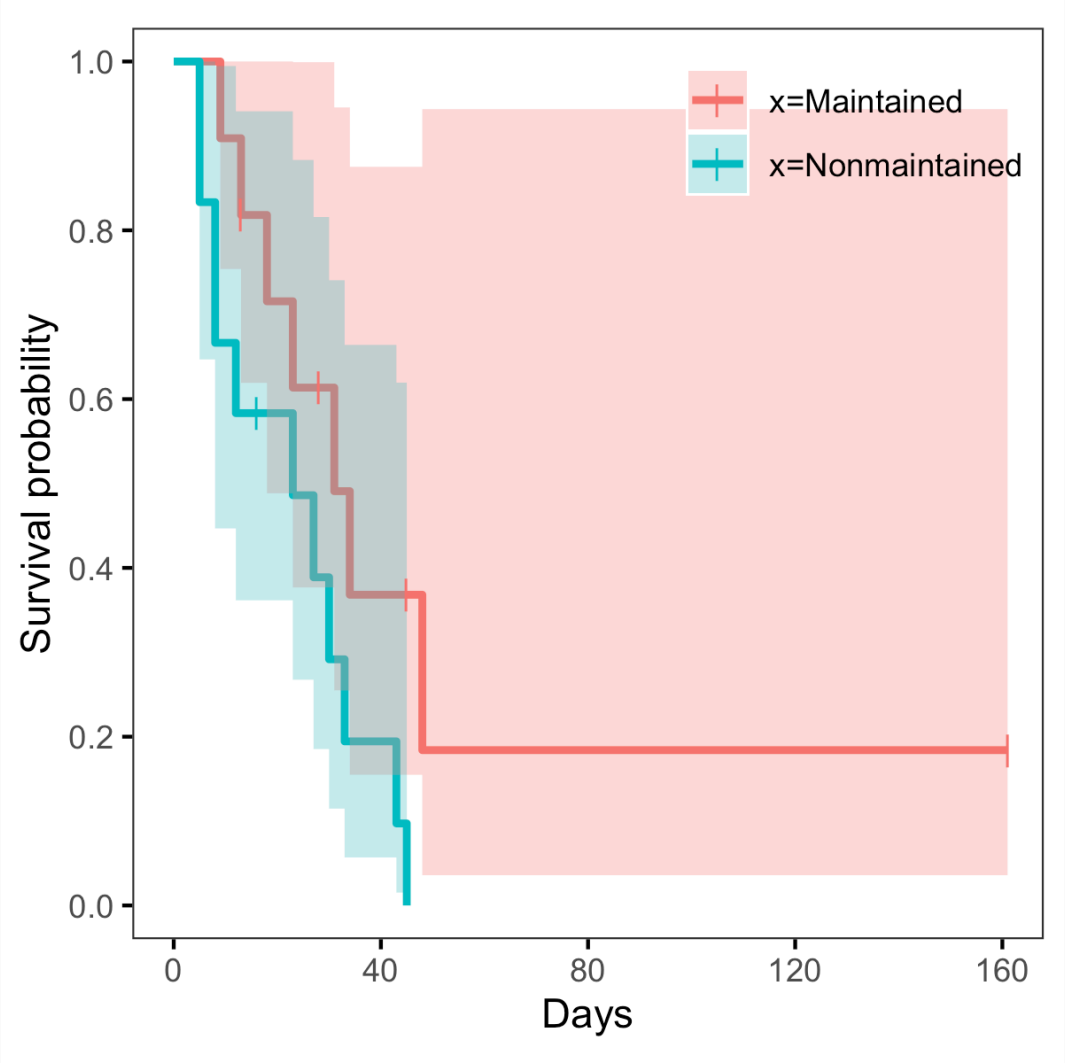


Figure 7. AML - maintenance

秩和检验

CMH Logrank test.

在每个不同的死亡时间构建一个 2×2 的表, 比较两组之间的死亡率与风险人数, 然后用 Cochran–Mantel–Haenszel 检验将这些表格合并. 有时也被叫作 Mantel–Cox 检验.

```r
diff <- survdiff(
    Surv(time, status) ~ x, data = aml,
    rho=0
)

pchisq(diff$chisq, length(diff$n)-1, lower.tail=F)
```

```
## [1] 0.06534
```

化疗在 aml 里的风险比

```
hr <- (diff$obs[1]/diff$exp[1]) /
    (diff$obs[2]/diff$exp[2])
cilow <- exp(log(hr) -
    qnorm(0.975) *
    sqrt(1/diff$exp[1]+1/diff$exp[2]))
cihigh <- exp(log(hr) +
    qnorm(0.975) *
    sqrt(1/diff$exp[1]+1/diff$exp[2]))

cat(str_interp(
"HR = $[.2f]{hr}\n95% CI: $[.2f]{cilow}-$[.2f]{cihigh}"
))
```

```
## HR = 0.44
## 95% CI: 0.17-1.11
```

生命表

生命表 是指每一个年龄段的人在下一个生日前死亡的概率 (Life table).

Table 9. John Graunt's Life Table (1662)

Age Interval	Proportion Deaths in Interval	Proportion Surviving until start of Interval
0–6	0.36	1.00
7–16	0.24	0.64
17–26	0.15	0.40
27–36	0.09	0.25
37–46	0.06	0.16
47–56	0.04	0.10
57–66	0.03	0.06
67–76	0.02	0.03
77–86	0.01	0.01

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
5	23	2	0.913	0.0588	0.8049	1
8	21	2	0.8261	0.079	0.6848	0.996
9	19	1	0.7826	0.086	0.631	0.971
12	18	1	0.7391	0.0916	0.5798	0.942
13	17	1	0.6957	0.0959	0.5309	0.912
18	14	1	0.646	0.1011	0.4753	0.878
23	13	2	0.5466	0.1073	0.3721	0.803
27	11	1	0.4969	0.1084	0.324	0.762
30	9	1	0.4417	0.1095	0.2717	0.718
31	8	1	0.3865	0.1089	0.2225	0.671
33	7	1	0.3313	0.1064	0.1765	0.622
34	6	1	0.2761	0.102	0.1338	0.569
43	5	1	0.2208	0.0954	0.0947	0.515
45	4	1	0.1656	0.086	0.0598	0.458
48	2	1	0.0828	0.0727	0.0148	0.462

Figure 8. Life table for `aml`

样例数据 lung

inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
3	306	2	74	1	1	90	100	1175	NA
3	455	2	68	1	0	90	90	1225	15
3	1010	1	56	1	0	90	90	NA	15
5	210	2	57	1	1	90	60	1150	11
1	883	2	60	1	0	100	90	NA	0
12	1022	1	74	1	1	50	80	513	0
7	310	2	68	2	2	70	60	384	10
11	361	2	71	2	2	60	80	538	1
1	218	2	53	1	1	70	80	825	16
7	166	2	61	1	2	70	70	271	34

R 的 Surv 函数可以接受三种指标样式

1. `1/0` (`1` = event)
2. `1/2` (`2` = event)
3. `TRUE/FALSE` (`TRUE` = event)

对于一个受试者 i:

1. 事件时间 T_i
2. 删失时间 C_i
3. 事件指标 δ_i:
   - 如果事件发生, $\mathbf{1}$ (i.e. $T_i \leq C_i$)
   - 如果删失, $\mathbf{0}$ (i.e. $T_i > C_i$)
4. 观察时间 $Y_i = \min(T_i, C_i)$

- 事件 与 时间 是最重要的
- 其它的数据项常称为 协变量, 基本都是用来分组的

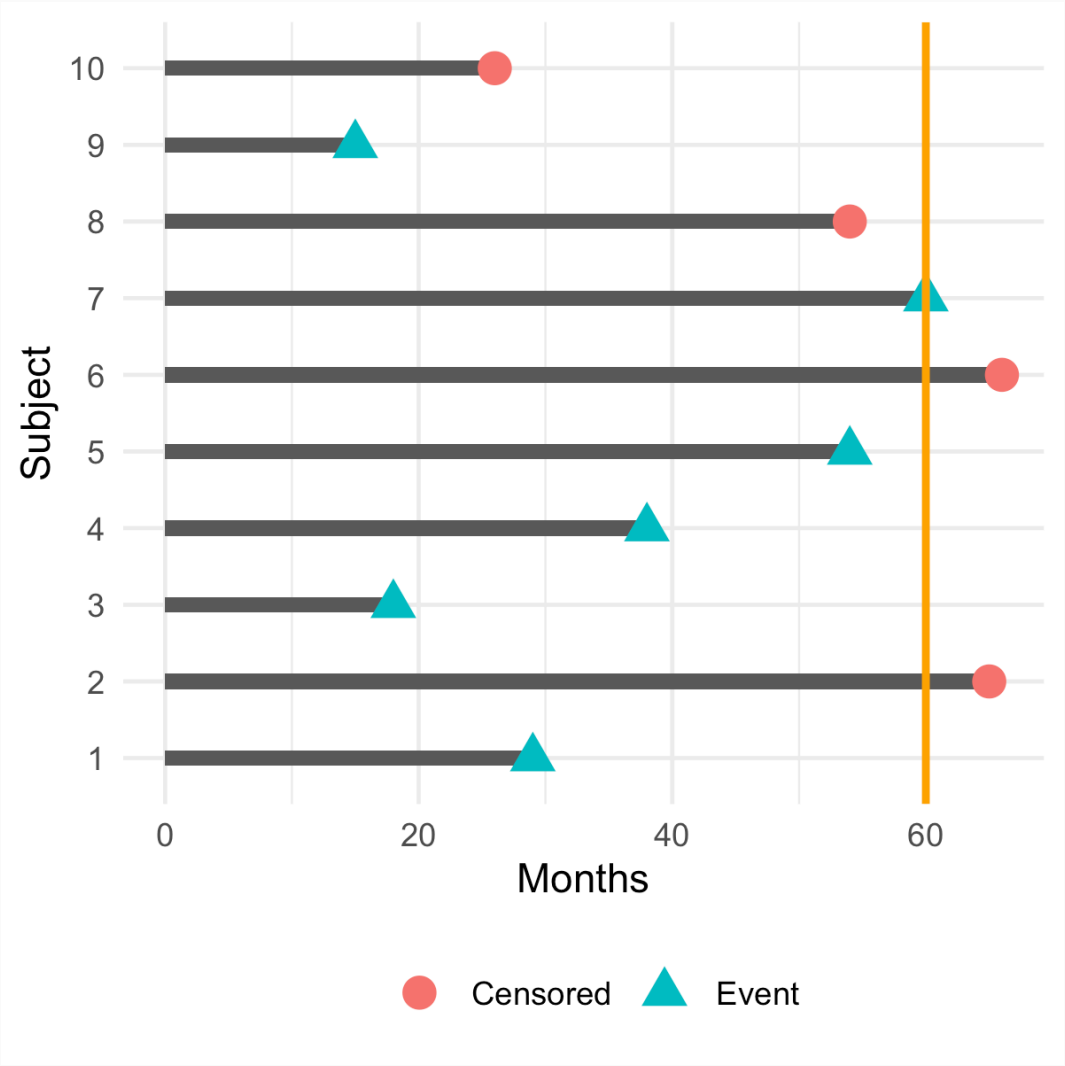


Figure 9. Censored data

观察时间

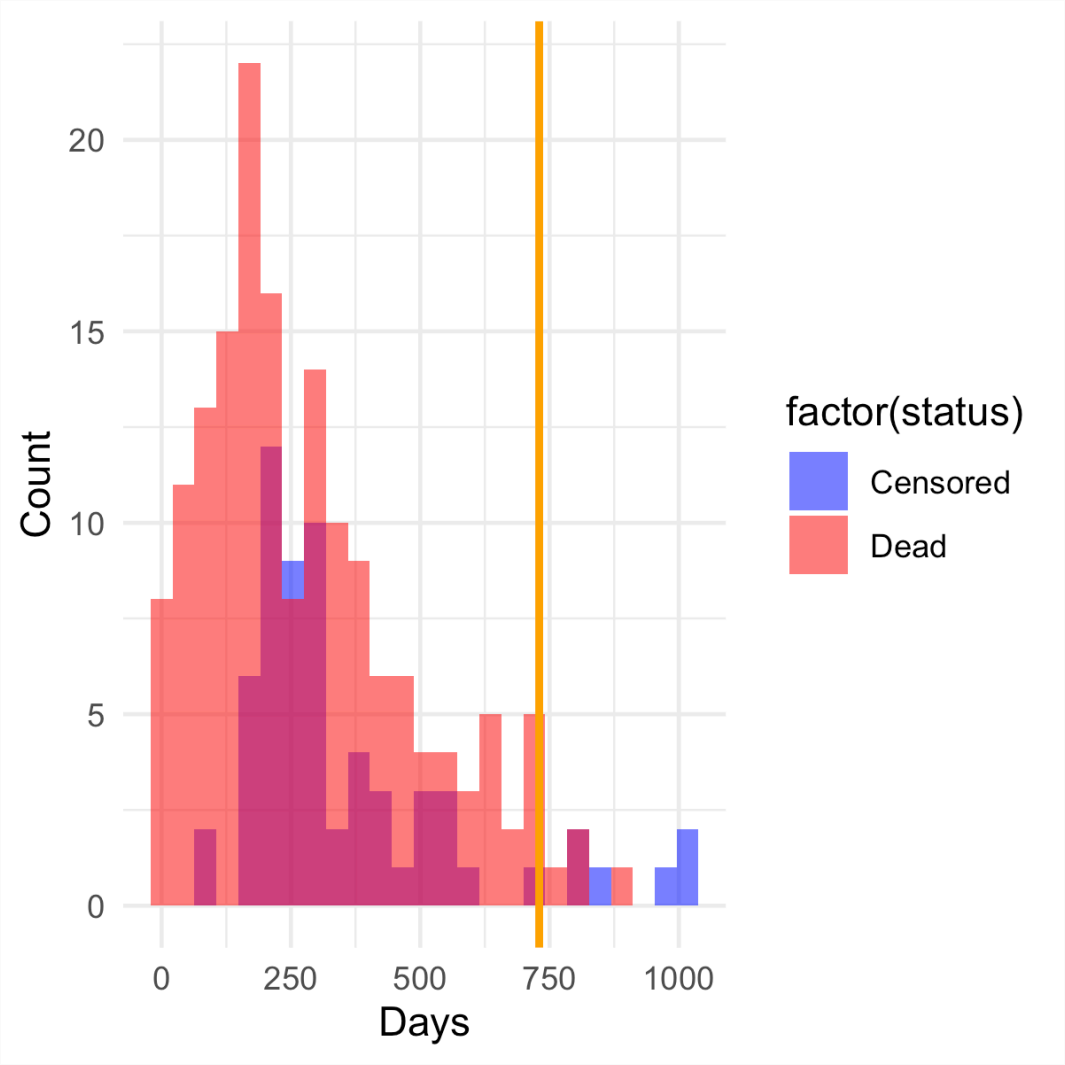


Figure 10. Distribution of follow-up time

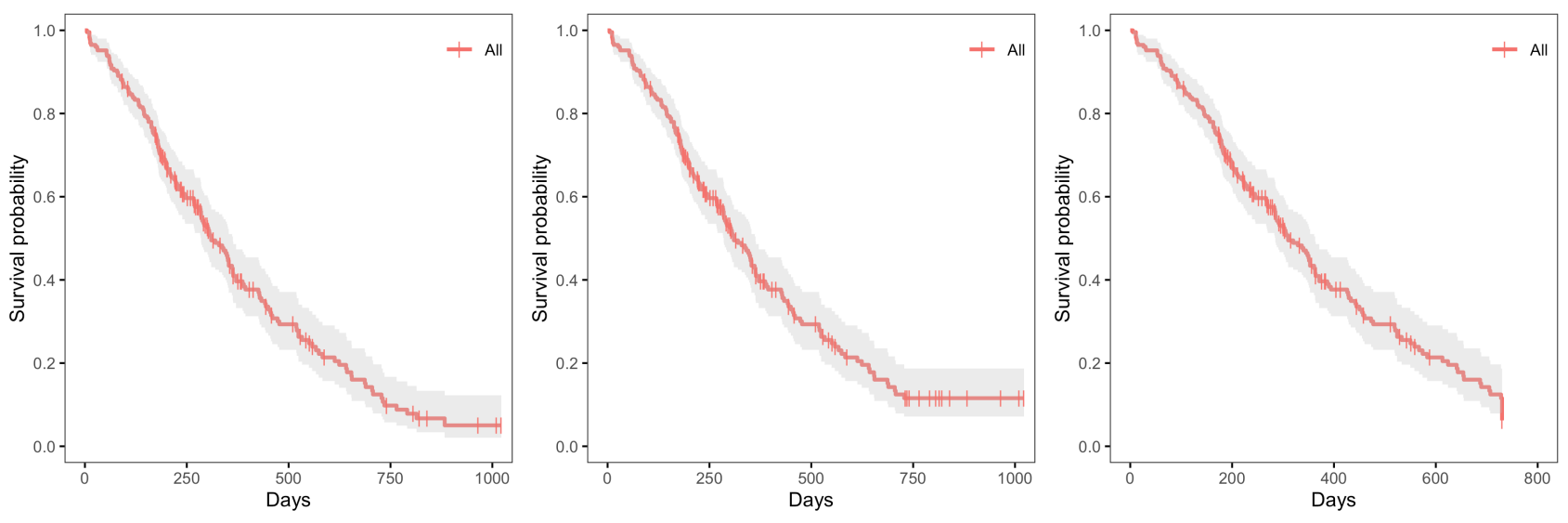


Figure 11. Lung - survival

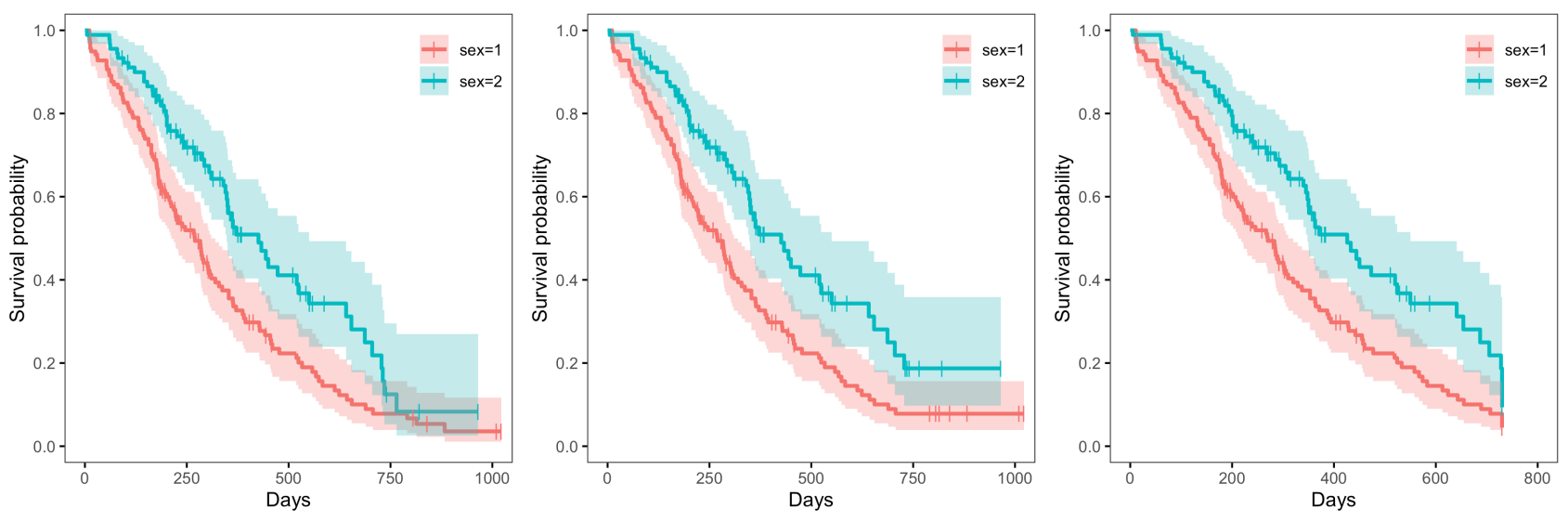


Figure 12. Lung - sex

P = 0.0013; HR = 1.69; 1.25-2.30

P = 0.00059; HR = 1.78; 1.30-2.43

P = 0.00078; HR = 1.73; 1.27-2.35

例题: KM

研究人员调查了专科护士干预是否能降低慢性心力衰竭患者的发病率和死亡率, 采用了随机对照试验的研究设计. 除了常规护理外, 干预措施包括专科护士的家访. 干预的目的是教育患者了解心力衰竭及其治疗. 对照治疗包括单纯的常规护理, 由入院医生和全科医生对患者进行常规管理. 参与者为 165 例因左心室收缩功能障碍而入院的心衰患者. 干预措施从出院前开始, 持续了一年.

任何原因的死亡或因心力衰竭再入院, 作为受试者的主要终点指标.

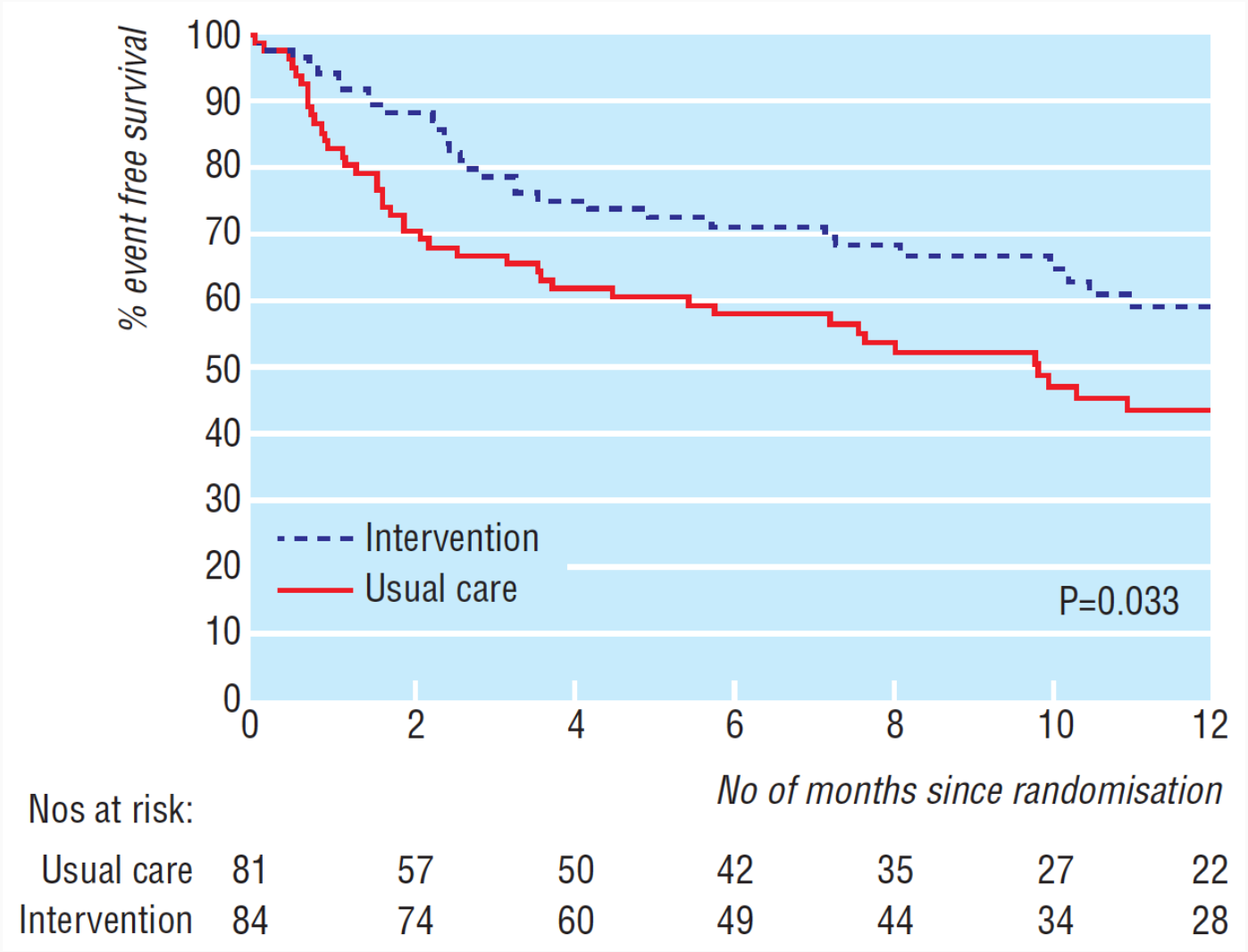


Figure 13. Blue2001

研究人员认为, 经过培训的护士可以改善慢性心力衰竭患者的发病率和死亡率.

干预组和对照组在 12 个月时的生存概率估计值分别为 0.59 和 0.43. Logrank 检验 P=0.033, 治疗组之间的生存时间有显著差异.

可以推断出以下哪种说法?

1. 对照组中约 43% 的人在 12 个月的随访结束时没有经历过主要终点
2. 干预组在 12 个月之后的某个时间经历主要终点的概率约为 0.59
3. 对于任何患者来说, 如果他接受了干预治疗而不是对照治疗, 那么开始治疗后到主要终点所需的时间会更长

协变量形式

类别变量	Categorical variable
离散变量	Discrete variable
连续变量	Continuous variable

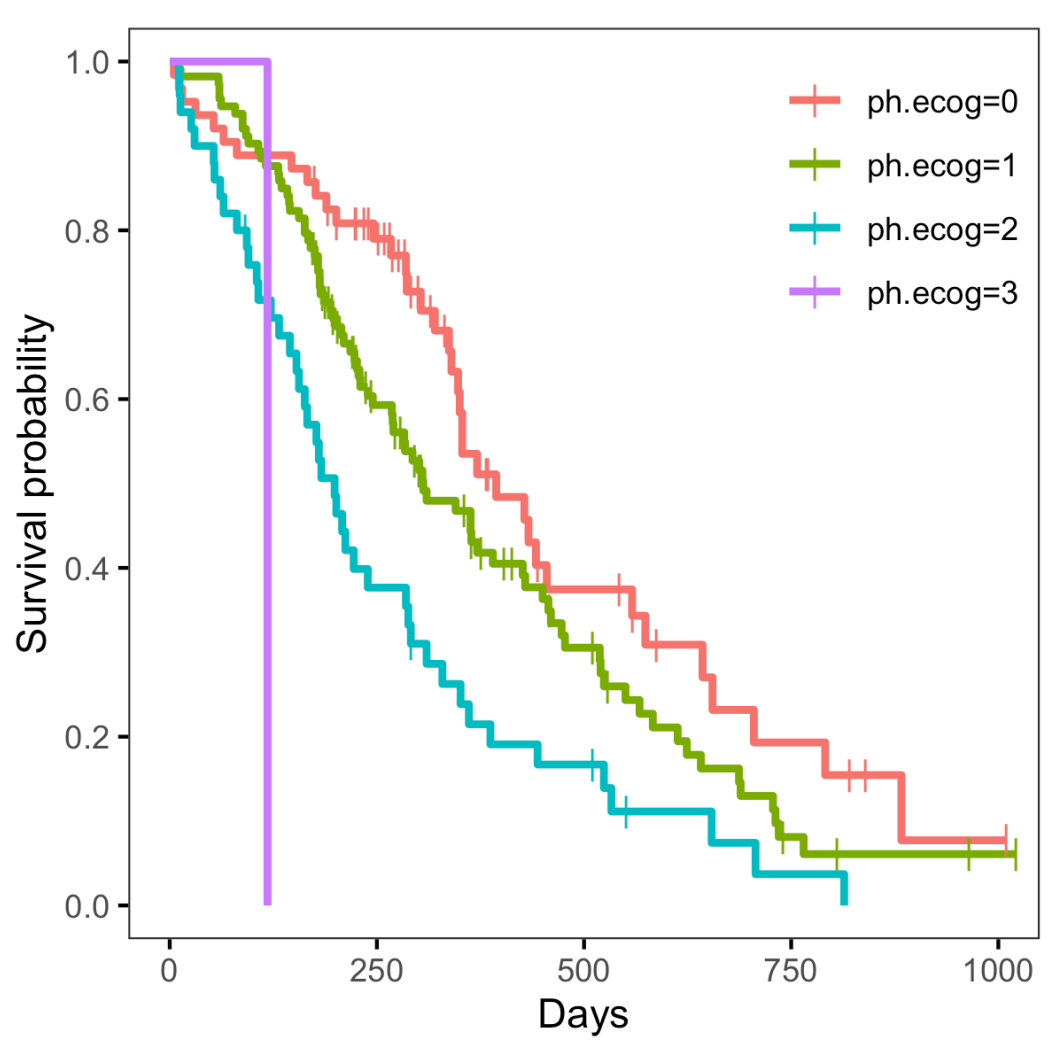


Figure 14. Lung - ph.ecog

P = 6.6e-05; HR = 0.70; 0.49-0.98

Cox 比例风险回归

生存模型　将某类事件发生前的时间与一个或多个可能与该时间量相关的 协变量 联系起来的模型.

回归分析　是一套估计因变量 (通常称为 “结果变量”) 和一个或多个独立变量 (通常称为 “预测因子”、“协变量”或 “特征”) 之间关系的统计过程.

Cox 比例风险模型是一种生存模型, 协变量的独立效应按 风险比 倍增, 即 风险比 恒定.

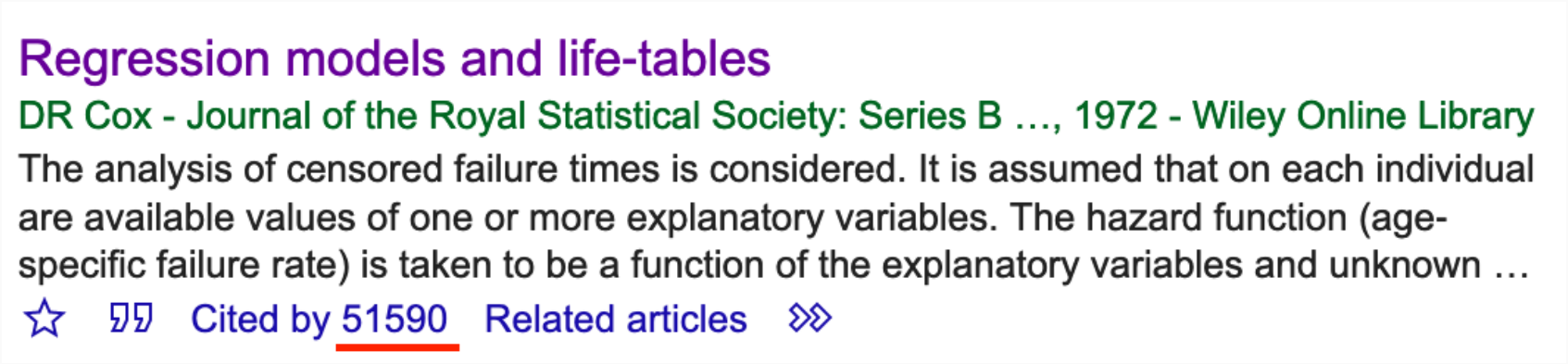


Figure 15. Cox1972

$$h(t) = h_0(t) \times \exp\left(b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k\right)$$

$$h(t) = h_0(t) \times \exp\left(b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k\right)$$

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

$$h(t) = h_0(t) \times \exp\left(b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k\right)$$

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

$$\ln\left(HR\right) = b_1X_1 + b_2X_2 + b_3X_3 + \cdots + b_kX_k$$

单因素 Cox 回归

term	estimate	std.error	statistic	p.value	conf.low	conf.high
inst	0.9908	0.0104	-0.8844	0.38	0.9708	1.0113
age	1.0194	0.0093	2.0575	0.04	1.0009	1.0383
sex	0.5602	0.1711	-3.3875	0.00071	0.4006	0.7833
ph.ecog	1.6061	0.1154	4.1040	4.1e-05	1.2809	2.0139
ph.karno	0.9814	0.0060	-3.1444	0.0017	0.9700	0.9929
pat.karno	0.9794	0.0055	-3.7830	0.00015	0.9689	0.9900
meal.cal	0.9999	0.0002	-0.5841	0.56	0.9994	1.0003
wt.loss	1.0006	0.0062	0.1016	0.92	0.9886	1.0128

多因素 Cox 回归

```
multivariate_cox <- coxph(
    Surv(time, status) ~
        age + sex + ph.ecog + ph.karno + pat.karno,
    data = lung2
)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
age	1.0111	0.0096	1.145	0.25	0.9922	1.0304
sex	0.5472	0.1747	-3.451	0.00056	0.3886	0.7707
ph.ecog	1.6208	0.1899	2.543	0.011	1.1170	2.3516
ph.karno	1.0112	0.0100	1.109	0.27	0.9915	1.0313
pat.karno	0.9885	0.0070	-1.657	0.097	0.9751	1.0021

```r
multivariate_cox <- coxph(
    Surv(time, status) ~
        sex + ph.ecog,
    data = lung2
)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
sex	0.5484	0.1715	-3.504	0.00046	0.3919	0.7674
ph.ecog	1.6270	0.1144	4.256	2.1e-05	1.3003	2.0358

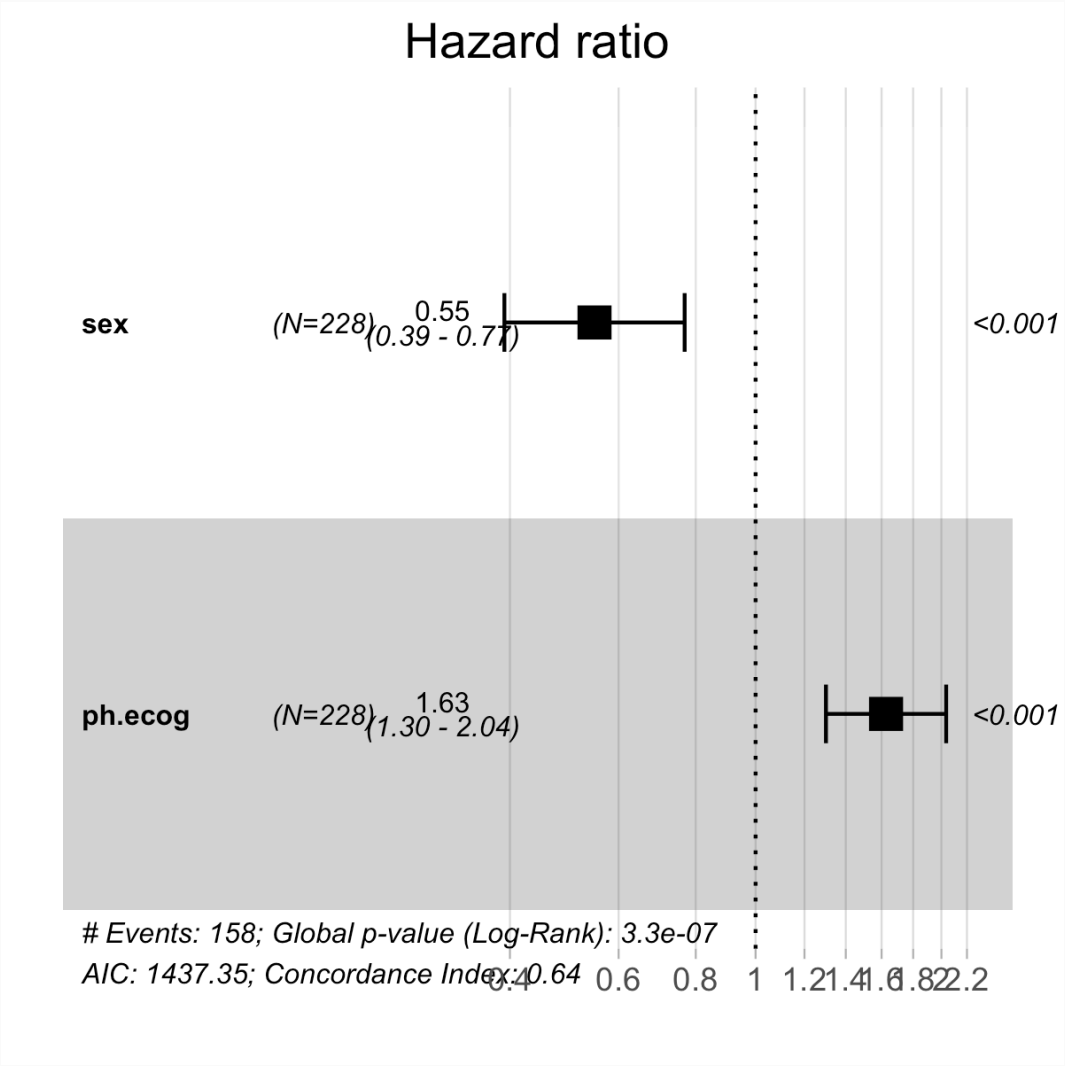


Figure 16. Lung - forest

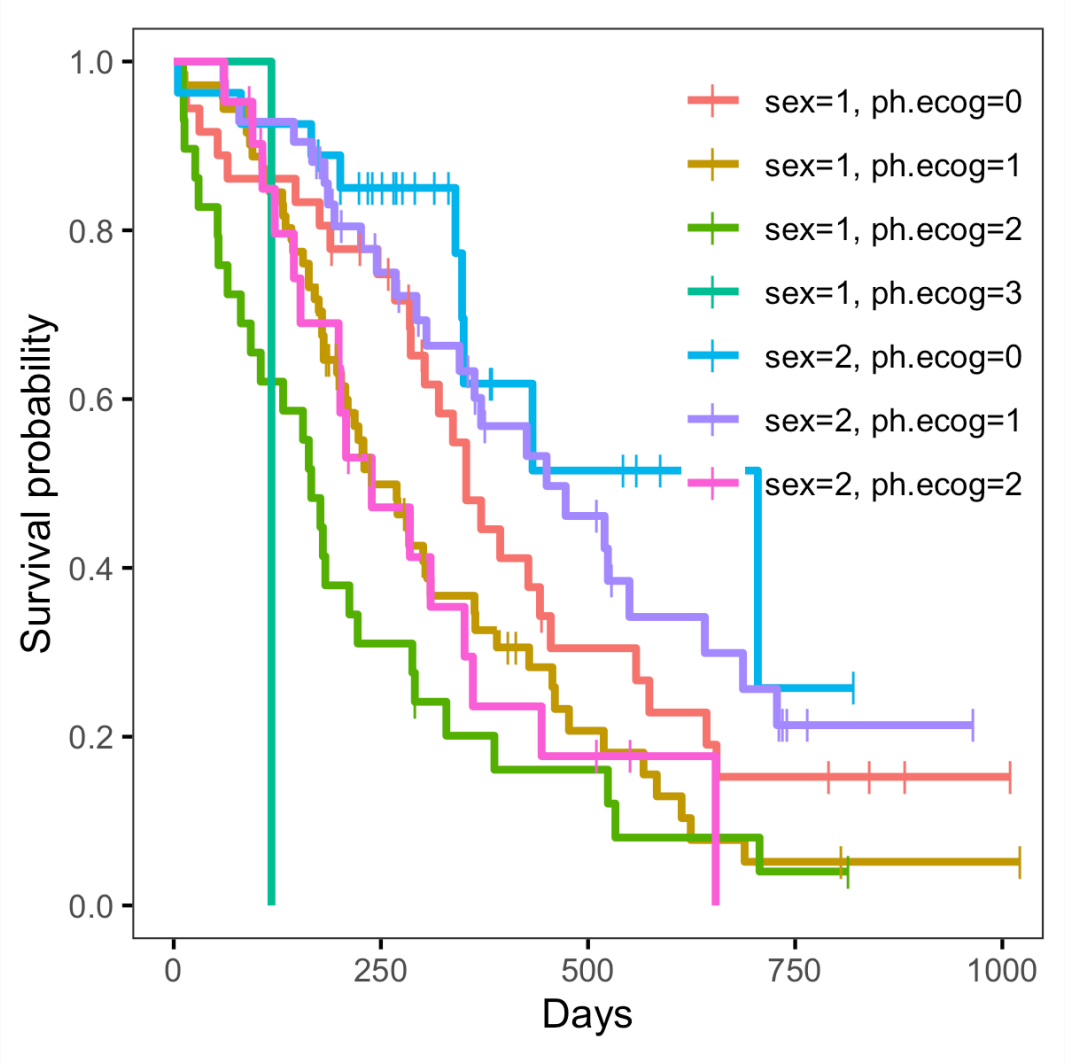


Figure 17. Lung - multivairate

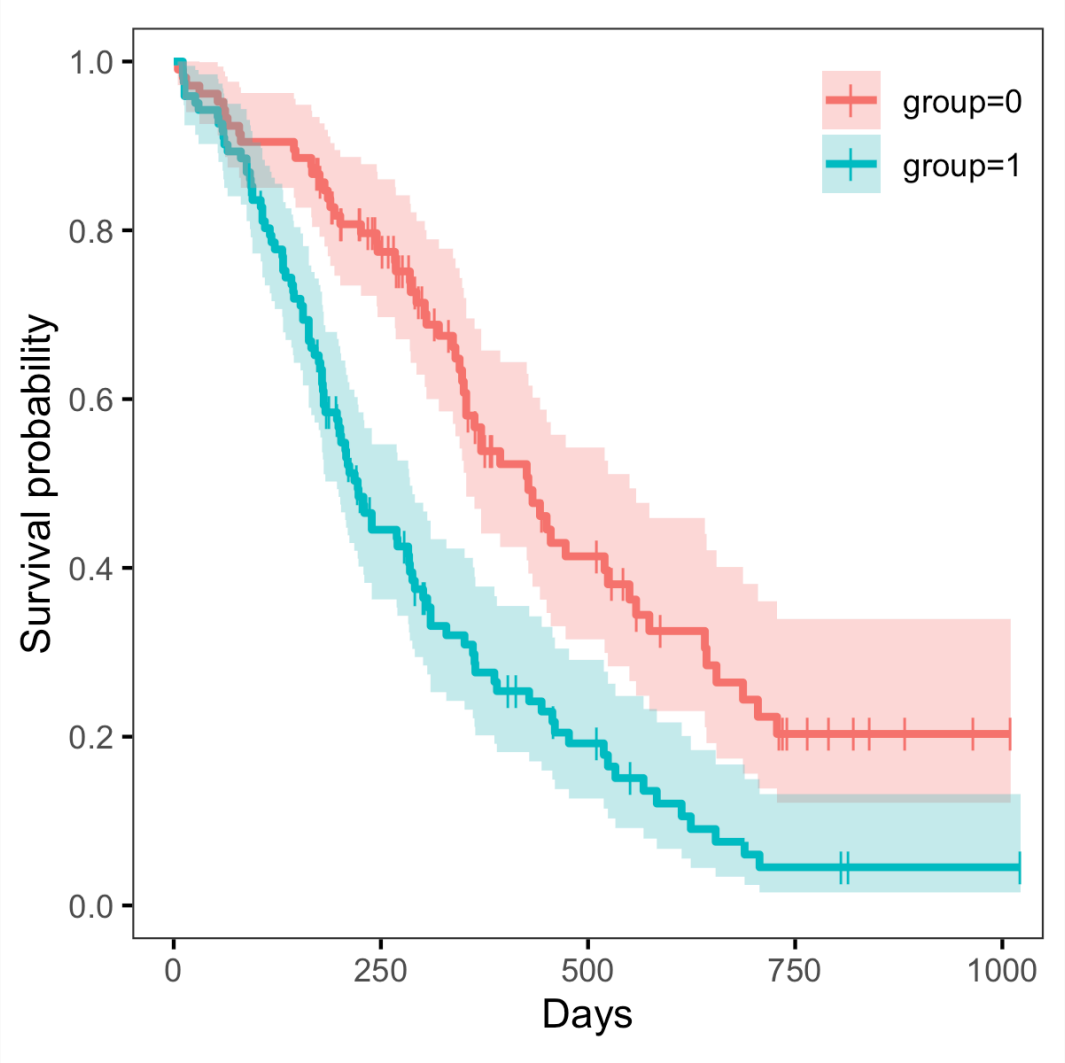


Figure 18. Lung - multivairate - median

组合与单因素对比

term	estimate	std.error	statistic	p.value	conf.low	conf.high
group	2.2026	0.1657	4.765	1.9e-06	1.5918	3.0479
sex	0.5602	0.1711	-3.388	0.00071	0.4006	0.7833
ph.ecog	1.6061	0.1154	4.104	4.1e-05	1.2809	2.0139

```
##      sex ph.ecog
## -0.6008  0.4867
```

ROC AUC

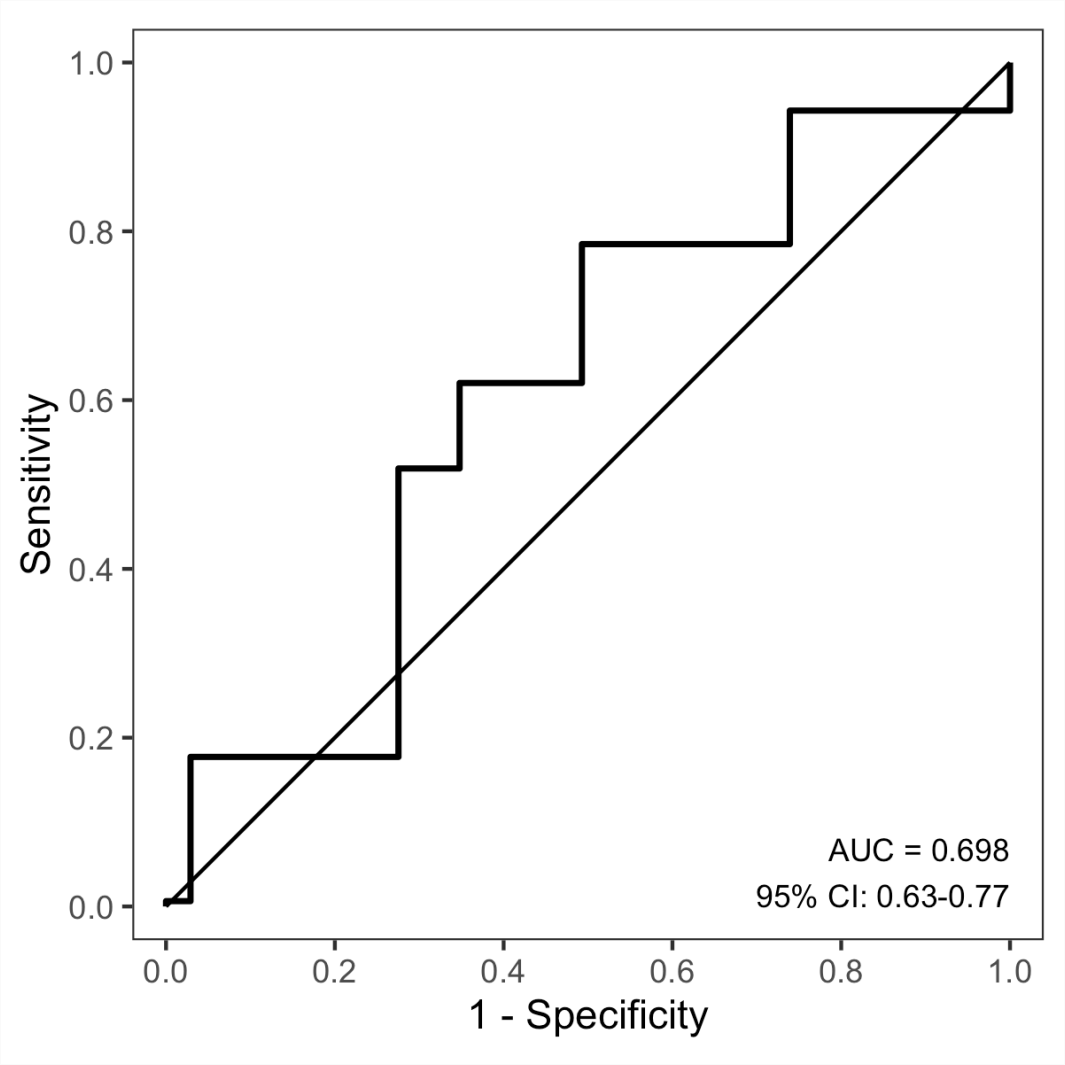


Figure 19. Lung - ROC

总结

- 删失数据会对 KM 曲线产生实质性的影响, 但在拟合曲线时必须包括在内, 在曲线上尽可能地显示删失的数据.
- KM 曲线是最常用的对生存函数的估计
- 秩和检验可以用于检验 KM 曲线间的差异
- 对于协变量, 可以用 Cox 回归来明确它们的影响
- Cox 回归的效果, 除可以用 KM 曲线来定性评判, 还可以用 ROC 曲线的 AUC 值来定量

https://github.com/wang-q/lecture-slides/blob/master/slides/biostat-survival.slides.pdf