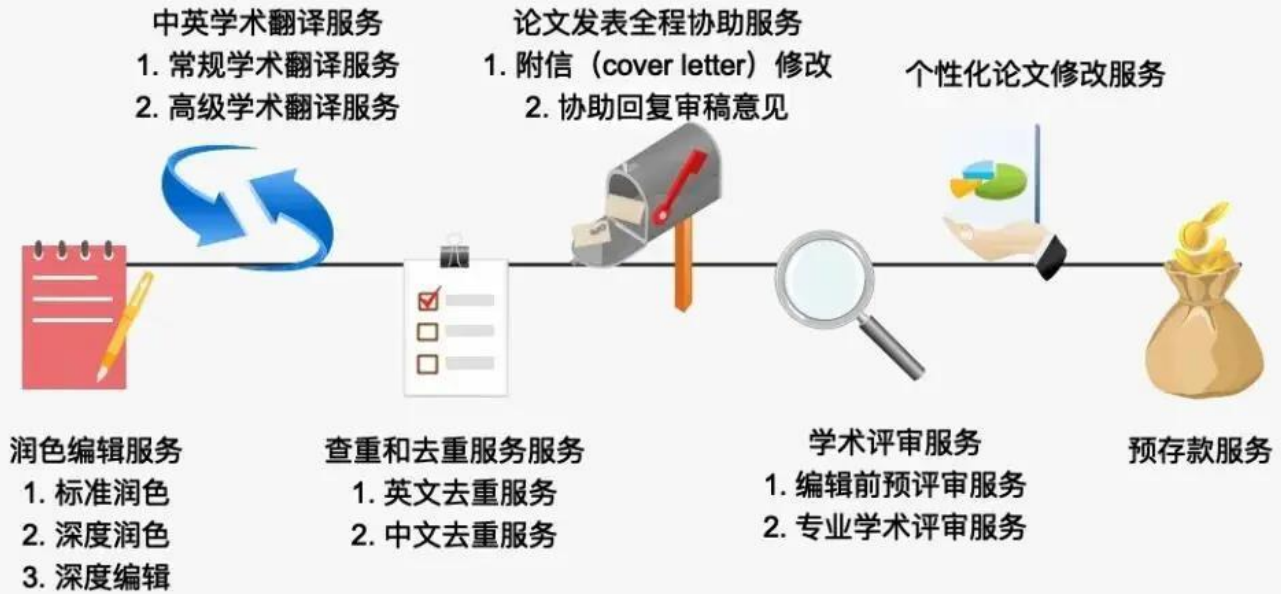


投必得七大服务类型



投必得学术

第二十六讲 R-生存分析：绘制KM生存曲线



投必得论...
已认证的官方帐号

12 人赞同了该文章

在第二十五讲，我们向大家介绍了生产分析的基本概率，并且提到了KM (Kaplan-Meier) 生存曲线，它是生存概率与时间的关系图。那么，这个图形如何在R中实现呢？今天我们将带您来——实现。

1. 安装并加载所需的R包

我们将使用两个R包：

- survival用于计算生存分析
- survminer用于总结和可视化生存分析结果
- 安装软件包

```
install.packages(c("survival", "survminer"))
```

- 加载软件包



2. 示例数据集

我们将使用survival包中提供的肺癌数据。

```
data("lung")
head(lung)
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306     2  74  1      1      90      100     1175      NA
2    3  455     2  68  1      0      90      90      1225     15
3    3 1010     1  56  1      0      90      90        NA     15
4    5  210     2  57  1      1      90      60     1150     11
5    1  883     2  60  1      0     100      90        NA      0
6   12 1022     1  74  1      1      50      80      513      0
```

- inst: 机构代码
- time: 以天为单位的生存时间
- status: 删失状态1 = 删失, 2 = 出现失效事件
- age: 岁
- sex: 性别, 男= 1女= 2
- ph.ecog: ECOG评分 (0 =好, 5 =死)
- ph.karno: 医师进行的Karnofsky评分 (0 = 差, 100 = 好)
- pat.karno: 患者自行进行的Karnofsky评分 (0 = 差, 100 = 好)
- meal.cal: 用餐时消耗的卡路里
- wt.loss: 最近六个月的体重减轻

3. 计算生存曲线: survfit()

目标: 按性别计算生存率。

函数survfit()[在survival包中]可用于计算kaplan-Meier生存估计。主要论包括:

- 使用Surv()函数创建的生存对象
- 数据集中包含了数据内容。

要计算生存曲线, 需要输入R代码如下:

```
fit <- survfit(Surv(time, status) ~ sex, data = lung)
print(fit)
Call: survfit(formula = Surv(time, status) ~ sex, data = lung)
      n events median 0.95LCL 0.95UCL
sex=1 138    112   270     212    310
sex=2  90     53   426     348    550
```

默认情况下, 函数print()显示生存曲线的简短统计内容。它输入了观察值, 事件数, 中位生存率和其置信区间。

如果要显示生存曲线的更完整统计内容, 请键入以下内容:

```
summary(fit)
# 查看完整的生存表格
summary(fit)$table
```



4. 访问survfit()的返回值

函数survfit()返回变量列表，包括以下组件：

- n：每条曲线中的对象总数。
- time：曲线上的时间点。
- n.risk：在时间t处有风险的受试者人数
- n.event：在时间t发生的事件数
- n.censor：在时间t退出事件而不发生风险的删失者的数量
- lower,upper：曲线的置信度上限和下限。
- strata：表示曲线估计的分层。如果strata不为NULL，则结果中有多条曲线。strata的水平（一个因子）会是曲线的标签。

可以按以下方式生成一个可以显示所有组件的数据表格：

```
d <- data.frame(time = fit$time,
                 n.risk = fit$n.risk,
                 n.event = fit$n.event,
                 n.censor = fit$n.censor,
                 surv = fit$surv,
                 upper = fit$upper,
                 lower = fit$lower
                )

head(d)
```

	time	n.risk	n.event	n.censor	surv	upper	lower
1	11	138	3	0	0.9782609	1.0000000	0.9542301
2	12	135	1	0	0.9710145	0.9994124	0.9434235
3	13	134	2	0	0.9565217	0.9911586	0.9230952
4	15	132	1	0	0.9492754	0.9866017	0.9133612
5	26	131	1	0	0.9420290	0.9818365	0.9038355
6	30	130	1	0	0.9347826	0.9768989	0.8944820

5. 可视化生存曲线

我们将使用功能ggsurvplot()[在Survminer包中]生成两组受试者的生存曲线。它可以显示的内容有：

- 使用参数conf.int = TRUE，显示生存函数的95%置信区间。
- 当选择选项risk.table时，将会显示按时间划分的处于风险中的人的数量和/或百分比。risk.table的允许值包括：
 - TRUE或FALSE，指定是否显示风险表。默认值为FALSE。
 - “absolute” 或 “percentage”：分别显示按时间划分处于风险中的受试者的绝对数或百分比。使用 “abs_pct” 同时显示绝对数字和百分比。
- 使用参数pval = TRUE，将显示对数秩检验（Log-Rank test）的p值。
- 使用参数surv.median.line，将显示中位生存时间点水平/垂直线。允许的值包括c（ “none” , “hv” , “h” , “v” ）之一。v: 垂直, h: 水平。

#按分层更改图形颜色，线型等

```
ggsurvplot(fit,
            pval = TRUE, conf.int = TRUE,
```

▲ 赞同 12 ▼ ● 添加评论 ➦ 分享 ♥ 喜欢 ★ 收藏 ...

```
linetype = "strata", # 根据分层更改线型
surv.median.line = "hv", # 同时显示垂直和水平参考线
ggtheme = theme_bw(), # 更改ggplot2的主题
palette = c("#E7B800", "#2E9FDF"))#定义颜色
```

Kaplan-Meier图可以解释如下：

横轴（x轴）表示以天为单位的时间，纵轴（y轴）表示生存的可能性或生存的人口比例。线代表两组 / 层的存活曲线。曲线中的垂直下降表示事件。曲线上的十字叉表示此时患者删失。

- 在起点时，生存概率为1.0（或100%的参与者还活着）。
- 在时间250，性别= 1的存活概率约为0.55（或55%），性别= 2的存活概率约为0.75（或75%）。
- 性别= 1的中位生存时间约为270天，性别= 2的中位生存期约为426天，这表明性别= 2的生存期高于性别= 1

可以使用以下代码获得每组的中位生存时间：

```
summary(fit)$table
```

	records	n.max	n.start	events	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
sex=1	138	138	138	112	325.0663	22.59845	270	212	310
sex=2	90	90	90	53	458.2757	33.78530	426	348	550

每组的中位生存时间代表生存概率S(t)为0.5的时间。



性别= 1（男性）的中位生存时间为270天，而性别= 2（女性）则为426天。与男性相比，女性肺癌似乎具有生存优势，即能生存更长时间。但是，要评估此差异是否具有统计学显著性，需要进行正式的统计检验，该问题将在下一部分中进行讨论。

注意，置信区间在曲线的尾部很宽，很难进行有意义的解释。这可以通过以下事实来解释：在实践中，通常有一些患者在随访快结束时删失明显。因此，在随访结束之前，缩短x轴的显示范围是明智的（Pocock等，2002）。

可以使用参数xlim缩短生存曲线，如下所示：

```
ggsurvplot(fit,
            conf.int = TRUE,
            risk.table.col = "strata",
            ggtheme = theme_bw(),
            palette = c("#E7B800", "#2E9FDF"),
            xlim = c(0, 600))
```

注意，可以使用参数fun指定三个经常使用的转换：

- “log”：生存函数的对数转换，
- “event”：绘制累积事件（ $f(y) = 1-y$ ）。也称为累积发生率
- “cumhaz” 绘制累积风险函数（ $f(y) = -\log(y)$ ）

例如，要绘制累积风险函数，请键入：

```
ggsurvplot(fit,
            conf.int = TRUE,
            risk.table.col = "strata",
```

```
palette = c("#E7B800", "#2E9FDF"),  
fun = "event")
```



累积风险常用来估计风险概率。定义为 $H(t) = -\log(\text{survivalfunction}) = -\log(S(t))$ 。

累积风险（ $H(t)$ ）可解释为累积死亡率。换句话说，如果事件是可重复的过程，则它对应为每个个人在时间 t 之前，预期事件发生次数总和。

要绘制累积风险，请键入以下内容：

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

6. Kaplan-Meier寿命表：生存曲线统计内容

如上所述，您可以使用函数summary()获得生存曲线的完整内容：

```
summary(fit)
```

还可以使用功能surv_summary()[在survminer包中]获取生存曲线的统计量。与默认的summary()函数相比，surv_summary()创建一个包含来自survfit结果的数据表。

```
res.sum <- surv_summary(fit)
head(res.sum)
```

	time	n.risk	n.event	n.censor	surv	std.err	upper	lower	strata	sex
1	11	138	3	0	0.9782609	0.01268978	1.0000000	0.9542301	sex=1	1
2	12	135	1	0	0.9710145	0.01470747	0.9994124	0.9434235	sex=1	1
3	13	134	2	0	0.9565217	0.01814885	0.9911586	0.9230952	sex=1	1
4	15	132	1	0	0.9492754	0.01967768	0.9866017	0.9133612	sex=1	1
5	26	131	1	0	0.9420290	0.02111708	0.9818365	0.9038355	sex=1	1
6	30	130	1	0	0.9347826	0.02248469	0.9768989	0.8944820	sex=1	1

函数surv_summary()返回包含以下内容：

- time：曲线上的时间点。
- n.riks：在时间t处有风险的受试者人数
- n.event：在时间t发生的事件数
- n.censor：在时间t退出事件而不发生风险的删失者的数量
- surv：估计等生存概率
- std.err:生存概率的标准误
- lower,upper：曲线的置信度上限和下限



在生存曲线拟合多个变量的情况下，`surv_summary`对象会将变量显示在额外的列里头。这可以按层次或某些因素组合来完成`ggsurvplot`的输出。

`surv_summary`对象还有一个名为“table”的属性，其中包含有关生存曲线的信息，包括具有置信区间的生存中位数以及每条曲线中受试者的总数和事件数。要访问属性“table”，请输入以下命令：

```
attr(res.sum, "table")
```

参考内容：sthda.com

如果您觉得我说的对您有帮助，请点赞让我感到您的支持，您的支持是我写作最大的动力~

ijournal：高颜值的期刊检索网站，助您快速找到理想目标期刊（weixin小程序也上线了哦）

投必得：全专业中英文论文润色编辑助力您的论文快速发表，点击了解业务详情

专栏传送门：

[投必得科研软件安装使用手册](#)；[投必得：SCI期刊介绍与选择](#)；[投必得，教你写论文](#)；[投必得统计分析大讲堂](#)；[投必得科研生活解忧杂货店](#)

发布于 07-24

[生存分析](#) [R（编程语言）](#) [数据分析](#)

文章被以下专栏收录

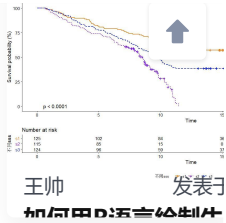
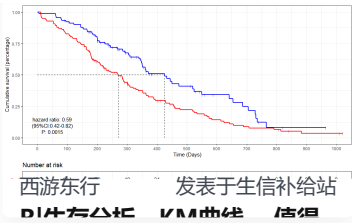
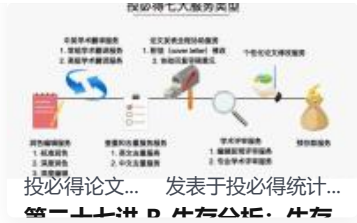


投必得统计分析大讲堂

进入专栏

推荐阅读

▲ 赞同 12 ▼ ● 添加评论 ➦ 分享 ♥ 喜欢 ★ 收藏 ...



还没有评论

写下你的评论...

