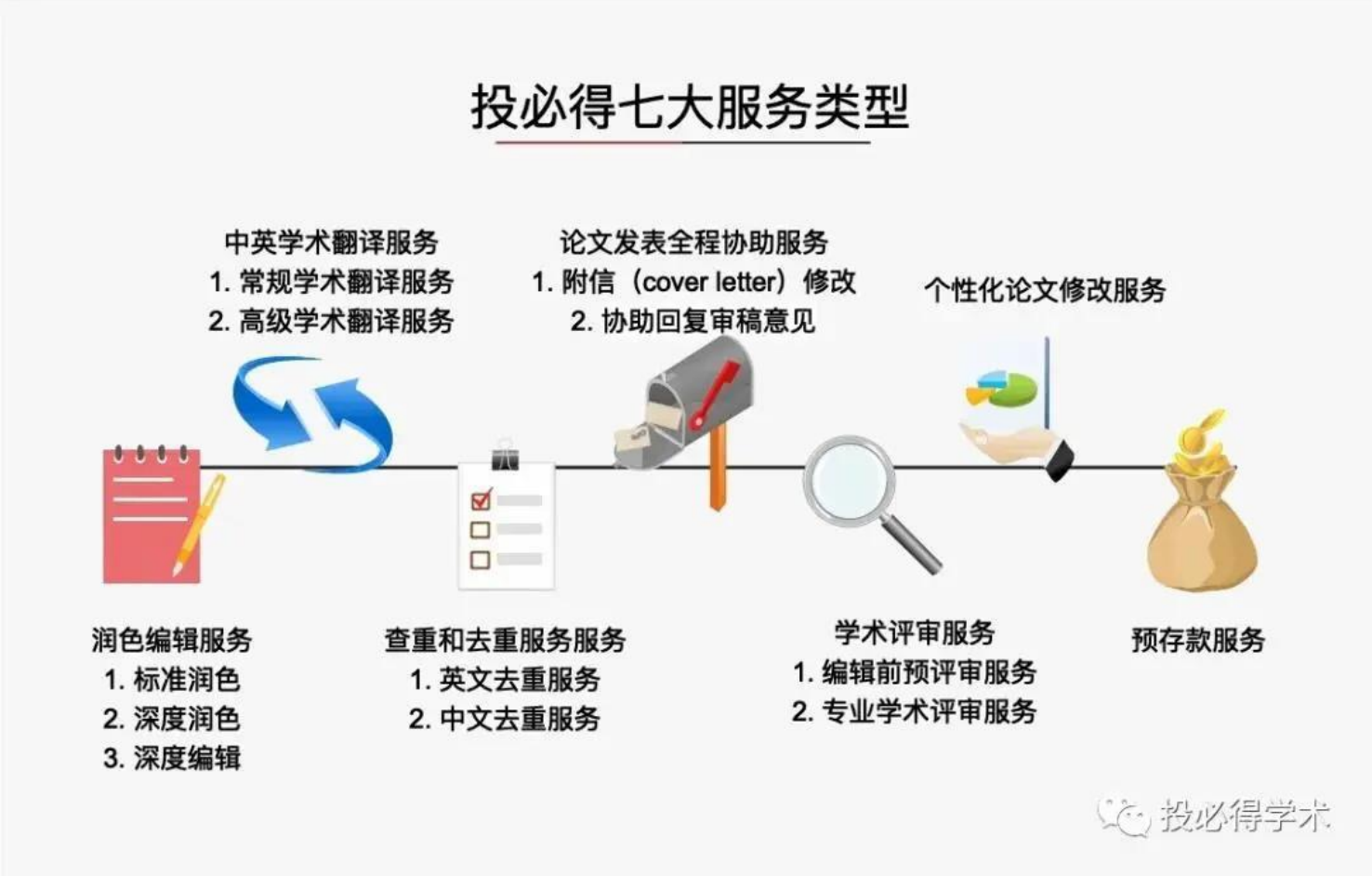


第二十五讲 R语言 生存分析基础概念

投必得论...
已认证的官方帐号

从第二十五讲开始，我们将进入高级统计学的学习，其中包括生存分析、聚类分析、主因子分析及机器学习（回归分析、高级建模等）。首先，我们将从生存分析开始。

1. 生存分析（survival analysis）

生存分析是对一个或多个非负随机变量进行统计推断，研究生存现象和响应时间数据及其统计规律的一门学科。它是一种既考虑结果又考虑生存时间的统计方法，并可充分利用截尾数据所提供的不完全信息，对生存时间的分布特征进行描述，对影响生存时间的主要因素进行分析。生存分析不同于其它多因素分析的主要区别点：生存分析考虑了每个观测出现某一结局的时间长短。

生存分析可用于许多领域，例如：

- 用于患者生存时间分析的癌症研究，如，研究某种药物的疗效，手术后的存活时间，某件机器的使用寿命等。
- 社会学研究中的“事件-历史分析”，如，出狱犯人第一次犯罪，失业人员第一次找到工作，
- 工程学中用于“故障-时间分析”，如，产品的失效。

素与生存时间的联系有无及程度大小，称为生存分析。

在癌症研究中，典型的研究问题如下：

- 某些临床特征对患者生存的影响是什么？
- 一个人生存3年的概率是多少？
- 两组患者的生存率是否存在差异？

我们之后将以癌症研究为例进行解说。

2. 基本内容。

大多数生存分析使用以下方法：

- Kaplan-Meier图可视化生存曲线。
- 对数秩检验以比较两组或更多组的生存曲线间是否存在差异。
- 用Cox比例风险回归描述变量对生存的影响。

3. 生存分析的基本概念

生存分析的基本术语：

事件包括起始事件和实效事件。

起始事件(initial event)：反应生存时间起始特征的事件，如疾病确诊、某种疾病治疗开始等。

失效事件(failure event)：在生存分析随访研究过程中，一部分研究对象可观察到死亡，可以得到准确的生存时间，它提供的信息是完全的，这种事件称为失效事件，也称之为死亡事件、终点事件。

生存时间（survival time）：广义上指某个起点事件开始到某个终点事件发生所经历的时间，度量单位可以是年、月、日、小时等，常用符号t所示。

根据研究对象的结局，生存时间数据可分为两种类型：

完全数据(Completed Data)：从观察起点到发生死亡事件所经历的时间。

不完全数据(Incomplete Data)：生存时间观察过程的截止不是由于死亡事件，而是由其他原因引起的

删失（censoring）

不完全数据分为：删失数据(censored Data)，截断数据(truncated Data)

不完全主要原因：

失访：指失去联系；

退出：死于非研究因素或非处理因素而退出研究；

终止：设计时规定的时间已到而终止观察，但研究对象仍然存活。

删失的表现形式

- 左删失(Left Censoring)：只知道实际寿命小于某数；
- 区间删失(Interval Censoring)：只知道实际寿命在一个时间区间内。

3.1 癌症研究中的生存时间和事件

事件有不同的类型，包括：

- 复发
- 进展
- 死亡

从“对治疗的反应”（完全缓解）到所关注事件发生的时间通常称为生存时间（或事件发生时间）。

癌症研究中两个最重要的措施包括：i）死亡时间；ii）无复发生存时间，对应于对治疗的反应与疾病复发之间的时间。也称为无病生存时间和无事件生存时间。

3.2删失

如上所述，生存分析着眼于开始点直到发生感兴趣事件（复发或死亡）之前的预期持续时间。但是，在研究期间内某些人可能未观察到该事件，从而产生了所谓的删失。

在肿瘤研究中，删失可能以下列方式出现：

1. 患者尚未（在研究期间）经历感兴趣的事件，例如复发或死亡；
2. 在研究期间患者失去随访；
3. 患者经历另一种事件，因此无法进行进一步的随访。

这些类型的删失在癌症研究的生存分析中很常见，并且都为右删失。

4. 生存和风险函数

使用两个相关的概率来描述生存数据：生存概率和风险概率。

生存函数，也被称为幸存者函数S(t)，是从时间起源（例如诊断癌症）到指定的未来时间t内仍然处于生存状态，未发生终点事件的概率。

风险函数，h(t)是在时间t内被观察的个体在该时间发生事件的概率。

请注意，生存函数侧重于没有事件发生，相反，风险函数着重于事件发生。

5. Kaplan-Meier生存估计

Kaplan-Meier（KM）方法是一种非参数方法，用于根据观察到的生存时间估算生存概率（Kaplan和Meier，1958年）。

在ti时的生存概率可以计算为

- S(ti-1) = 在ti-1还活着的概率
- ni = 在ti不久之前还活着的患者人数
- di = 在ti时间点发生的事件数
- t0 = 0， S(0) = 1

估计的生存概率（S(t)）是针对时间点t的函数，每个时间点t都有自己的对应值。我们也可以计算生存概率的置信区间。

KM生存曲线是KM生存概率与时间的关系图，它提供了有用的数据统计，可用于估算度量值，例如中位生存时间。

这一讲主要带大家了解了生产分析的基本概念，下一讲中，我们将开始详细介绍生存分析的r实现。

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~

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 07-23

生存分析　R（编程语言）　数据统计

文章被以下专栏收录

投必得统计分析大讲堂　进入专栏

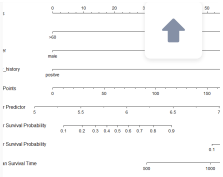
推荐阅读

R语言数据处理120题，终于有人来照顾用R的孩子了

本文来源：和鲸社区优秀创作者 @刘早起本套习题源于 Pandas进阶修炼120题系列。但由于R语言和Pandas有部分差别较大，在尽量不修改原题的基础上制作完成。本项目包含基础、基本数据处理、金...

网贷闲光

第二十七讲 R-生存分析：生存函数的假设检验表于投必得统计...



如何用R语言绘制nc图帅（列线图）发表于

还没有评论

写下你的评论...