生物统计

蒋冰

学号：201322130002

1 模型和方法

1.1 生存分析

生存分析（survival analysis）也称之为风险模型（hazard model）或持续模型（duration model）, 是一种根据实验或调查数据, 对生物、人以及具有类似于生存规律的其它事物的生存时间进行分析和推断的统计方法。目前生存分析已广泛应用于生物统计、医学、金融和工业工程等多个领域, 近年来也逐渐被运用到城市交通领域的研究中。比如用来研究交通事故的清理时间 [13 鄄 14], 基于活动的出行行为 [15 鄄 16], 机动车持有时间及报废时间 [17 鄄 18], 城市道路混合交通行为 [19 鄄 20], 航班的延误时间 [21], 高速公路危险区域交通冲突发生的风险率 [22] 等.。

广义的生存时间指生物体存活的时间, 或所关心的某种现象的持续时间。道路拥堵持续时间是指从交通拥堵现象产生开始, 一直到拥堵结束为止的持续时间, 属于广义生存时间范畴，可运用生存分析方法来对其进行研究。令 T 代表生存的时间（寿命）, 它是一个非负的随机变量, 令 $f(t)$ 表示 T 的概率密度函数，P 代表概率，T 的分布函数为 $F(t) = P(T \leqslant t) = \int_0^t f(x)dx$。

生存函数 S(t)，也叫生存率，表示生存时间大于 t 的概率，它的表达式为 $S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(x)dx$。

生存分析中，T 可以用危险率函数 h(t) 来描述。危险率函数也叫条件生存率，它指病人从患病起持续了 t 时间后没有结束，但在接下来的一段很小的时间 Δt 内结束的概率，可表示为 $h(t) = \lim\limits_{\Delta t \to \infty} \dfrac{P(t \leqslant T < t + \Delta t | T \geqslant t)}{\Delta t} = \frac{f(t)}{S(t)} = \frac{-dlnS(t)}{dt}$。它表示死亡速率的大小。如以 t 为横坐,f(t) 为纵坐标作出的曲线称为密度曲线, 由曲线上可看出不同时间的死亡速率及死亡高峰时间。纵坐标越大, 其死亡速率越高, 如曲线呈现单调下降, 则死亡速率越来越小, 如呈现峰值, 则为死亡高峰。

1.2 估计生存函数的非参数方法

估计生存函数的方法主要有参数方法和非参数方法两种。但当分布类型未知时, 非参数方法的效率更高。由于当前并不知道 **** 服从何种参数分布, 因此, 本文运用非参数方法来估计拥堵持续时间的生存函数.。乘积限方法是最常用的非参数估计方法，以 *** 的持续时间为研究对象，假定共有 n 个拥堵持续时间

样本,，这些样本的拥堵持续时间共有 k() 个不同的取值, 由于没有删失数据, 直接将它们从小到大排序 t1 < t2 < ⋯< tk . 令 dj 为 tj 这一单位时段内拥堵结束的样本数,nj 为时刻 tj 之前仍拥堵的样本数. 生存函数 S(t) 的乘积限估计可用下式表示，(4) 这里的 j 是满足不等式 tjt 的任何值, 估计量 S(t) 是由数项乘积构成, 乘积中的每一项 (nj-dj)/nj 为 tj 时段的生存概率, 它指的是在时刻 tj 之前拥堵没有结束的样本 (nj 个) 中, 在 tj 之后拥堵仍没有结束 ((nj - dj) 个) 的比例.

1.3 比较多个生存分布的非参数方法

一般而言，不同时间和空间的道路拥堵持续时间会有所差异。许多非参数检验方法都可以用来比较多个生存分布, 本文以应用广泛的 Cox-Mantel 方法来检验 ********** 持续时间生存函数分布的差异性。以比较两个不同断面的拥堵持续时间分布是否一致为例，假设 1 和 2 表示不同的道路断面，x1，x2,⋯，xn1 是断面 1 的 n1 个拥堵持续时间样本，y1，y2,⋯，yn2 是断面 2 的 n2 个拥堵持续时间样本。假设 ** 和 **2 的观测值分别是来自生存函数为 $S_1(t)$ 和 $S_2(t)$ 的样本，其原假设和备择假设分别是 H_0：$S_1(t) = S_2(t)$（断面 1 和断面 2 的分布一致）

H_1：$S_1(t) \neq S_2(t)$（断面 1 和断面 2 的分布不同）

设 $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ 是两组合在一起后不相同的死亡时间，$m_{(i)}$ 是死亡时间等于 $t_{(i)}$ 的个数，即 $t_{(i)}$ 的重复次数。因此 $\sum_{i=1}^{k} m_{(i)} = n_1 + n_2$。

设 R(t) 是时间 t 的风险集，即死亡时间或删失时间至少是 t 的所有个体组成之集合。设 n_{1t} 和 n_{2t} 分别是 R(t) 中对应于处理组 1 和处理组 2 的个体数。在 $R(t_{(i)})$ 中的个体数是 $r(i) = n_{1t(i)} + n_{2t(i)}$，令 $U = n_2-$，(6) 摇摇 I = 移 k i = 1 m(i)(r(i) - m(i)) r(i) - 1 A(i)(1 - A(i))，(7) 这里 A(i) 是 R(t(i)) 中属于第 2 组的个体所占的比例，令摇摇 C = U/ I 。(8) 可以证明 C 近似服从标准正态分布。因此若 Z > Z1 -琢/ 2，则拒绝原假设，认为断面 1 和断面 2 的分布不一致，两者有显著差异。

2 数据和样本提取

数据来源于 A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa

本文的数据提取自一个眼科学的临床实验，其主要目的是检验给视网膜炎着色的病人补充不同的维生素 A 用以预防病人实力损失。视力损失由视网膜功能损失所测定，它是由仪器 ERG（电子视网膜成像仪）上 30Hz 上的振幅减少 50% 以上表征。对正常人，ERG 30Hz 的振幅范围是 >50 V（微伏）。而在有视网膜炎着色的病人上，ERG 30Hz 的振幅范围通常 <10 V，及常 <1 V。约有 50% 的病人在 ERG 30Hz 上的振幅接近 0.05 V 时就失明，不到 10% 的病人 EFG 30Hz 振幅解决 1.3 V（这是这次临床试验中病人的平均 ERG 振幅）。这次临床实验中病人被随机分到 4 个处理组之一。

组 1：接受 15000IU（痕量）的维生素 A 及 3IU 的维生素 E。

组 2：接受 75IU（痕量）的维生素 A 及 3IU 的维生素 E。

组 3：接受 15000IU（痕量）的维生素 A 及 400IU 的维生素 E。

组 4：接受 75IU（痕量）的维生素 A 及 400IU 的维生素 E。

为方便阐述，有把上面 4 组改称为 A 组、对照组、AE 组和 E 组。我们要比较不同组中病人治疗失败的比例（“失败”是值 ERG 的 30Hz 振幅损失 50%）。病人取自 1984 年至 1987 年的记录，且跟踪到 1991 年 9 月。因为追踪停止实在同一个时刻终止的，故而不同病人的跟踪时间长度各不相同。早登记的病人可以跟踪达 6 年，而记录较晚的病人则只有 4 年。另外，有些病人有些病人在追踪挺值钱就推出研究了在追踪挺值钱就推出研究了，而这些人却又未曾“失败”。把由于死亡，其他疾病、可能的药物副作用，或不愿意继续下去等称为“退出”者。

3 实证分析

3.1 截尾数据的生存函数和危险率的估

为方便阐述，有把上面 4 组改称为 A 组、对照组、AE 组和 E 组。我们要比较不同组中病人治疗失败的比例（“失败”是值 ERG 的 30Hz 振幅损失 50%）。

4 实证分析

4.1 乘积极限法 (Product-Limit Method)

乘积极限法又称为积限法或 PL 法, 它是由统计学家 Kaplan 和 Meier 于 1958 年首先提出的, 因此又称为 Kaplan-Meier 法, 它是一种利用条件概率及概率的乘法原理计算生存率及其标准误的方法。

设 S(t) 表示 t 年的生存率,$S(t_i|t_{i-1})$ 表示活过 t_{i-1} 年后再活过 t_i 年的条件概率。根据条件概率的性质有 $S(t_2) = S(t_1)S(t_i|t_{i-1})$。

病人取自 1984 年至 1987 年的记录，并跟踪到 1991 年 9 月。因为追踪停止是在同一个时刻终止的，故而不同病人的跟踪时间长度各不相同。早等级的病人可以跟踪 6 年，而记录较晚的病人则只有 4 年。另外，有些病人在 1991 年 9 越以前就推出研究了，而这些人却又未曾“失败”。“退出”者是由于死亡，其他疾病，可能的药物副作用，或个人原因不愿意继续下去。我们把这些在随访蒸汽内未到达疾病终点的病人的数据称为失访或截尾观察（censored observation）。

设 S_{i-1} 是个体存活到 t_{i-1} 时没有失访的人数，设在 t_i 时仍存活 S_i 人，失败 d_i 人，失访 l_i 人，于是有 $S_{i-1} = S_i + d_i + l_i$，因此可以估计 t_{i-1} 时的存活着到时刻 t_i 时的生存概率为 $1 - \frac{d_i}{S_{i-1}} = 1 - \frac{d_i}{S_i+d_i+l_i}$，$t_i$ 时的失访者在 $t > t_i$ 是不会对生存函数构成影响。所以在 t_i 上的生存概率 $S(t_i)$ 的 Kaplan-Meier 估计量公式为：$\hat{S}(t_i) = \left(1 - \frac{d_1}{S_0}\right)\left(1 - \frac{d_2}{S_1}\right)\cdots\left(1 - \frac{d_i}{S_{i-1}}\right)$，$i = 1, 2, \cdots, k$

在引入 $Var\{ln[\hat{S}(t)]\} = \sum_{j=1}^{i} \frac{d_j}{S_{j-1}(S_{j-1} - d_j)}$，于是可以解得在 t_i 上的生存概率 $S(t_i)$ 的双侧 100%×(1-)CI 即为 (e^{c_L}, e^{c_R})，则

$c_L = ln[\hat{S}(t_i) - z_{1-\frac{\alpha}{2}} se\{ln[\hat{S}(t_i)]\}]$
$c_R = ln[\hat{S}(t_i) + z_{1-\frac{\alpha}{2}} se\{ln[\hat{S}(t_i)]\}]$
一般地，在 t_i 时危险率为 $\hat{h}(t_i) = \frac{d_i}{S_{i-1}}$

4.2 对数 -秩检验

4.3 Wilcoxon 检验

4.4 Likelihood ration(LR) 检验

4.5 比较危险率模型 (Proportaional-Hazards Model)

在比较危险率模型中，危险率可以表示成
$h(t) = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)$
其中 $x_1, x_2, \cdots, x_k$ 是一组独立变量，而 $h_0(t)$ 是在基准状态下在 t 时刻上的基准危险率，它代表所有独立变量全取值 0 是的危险率。做出零假设为 H_{j0}：$\beta_j = 0$，备择假设为 H_{j1}：$\beta_j \neq 0$，计算检验统计量 $z = \frac{\hat{\beta_j}}{se(\hat{\beta_j})}$，当显著性水平为 α 时，当 $z < z_{\frac{\alpha}{2}}$ 或 $z > z_{1-\frac{\alpha}{2}}$ 时，则拒绝 H_{j0}；若 $z_{\frac{\alpha}{2}} \leqslant z \leqslant z_{1-\frac{\alpha}{2}}$，则接受 H_{j0}。
比例危险率模型可以写成 $ln\left[\frac{h(t)}{h_0(t)}\right] = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

4.6 总体分析

5 结论