·系列讲座·

肿瘤流行病学研究资料的统计分析

项永兵

第三讲　调整生存率估计和生存期比较

无论是临床预后因素研究资料的分析，还是大样本流行病学随访资料的分析，常规的做法第一步就是采用 Kaplan- Meier 法或寿命表法估计各个样本的生存率，然后比较它们的差别有无统计学意义。进一步则是做单因素或多因素分析。对研究因素受其它混杂因素影响的分析一般在多变量分析阶段考虑，忽略了单因素分析中的分层调整。本讲讨论的主要问题有两个方面：一是如何估计调整了混杂因素后的生存率；二是在生存期差别比较时如何计算分层调整的检验统计量，进行假设检验。

一、统计方法：

1. 调整生存率估计：笔者在前文（第二讲）[1] 中谈到样本生存率估计的两种方法，即 Kaplan- Meier 法（KM 法）和寿命表法（LT 法）。假定样本数据有 K 个死亡时间点（对应于 KM 法）或区间（对应于 LT 法），$t_k(k=1,\ldots,K)$。以最简单的两组比较为例，样本资料可以整理成下述标准格式。KM 法和 LT 法的生存率估计[1~4]为：

样本数据整理后的 2×2 表

组	死亡	存活	合计
1	d_{1k}	$l_{1k}-d_{1k}$	l_{1k}
2	d_{2k}	$l_{2k}-d_{2k}$	l_{2k}
合计	D_k	$l_{1k}+l_{2k}-D_k$	T_k

$$S_m^{KM}(t_k)=\prod_{t_k<t}\left[(l_{mk}-d_{mk})\Big/l_{mk}\right], \quad (1)$$

$$S_m^{LT}(t_k)=\prod_{t_k<t}\left[(l'_{mk}-d_{mk})\Big/l'_{mk}\right], \quad (2)$$

这里 $m=1,2$。现在我们关心的是，当研究者估计两组生存率时，存在其它影响其预后的因素，即两组在混杂因素的分布方面不平衡。例如估计男女两组的生存率，但他们的年龄构成（分布）不同。假定某混杂因素的层数或水平数为 $j=1,\ldots,J$，则 KM 法调整生存率[5~7]为：

$$S_m^*(t_k)=\prod_{t_k<t}\left\{\sum_{j=1}^{J}\left[f_{jk}\times\frac{l_{mjk}-d_{mjk}}{l_{mjk}}\right]\right\} \quad (3)$$

若为 LT 法，可用 l'_{mjk} 代替上式中 l_{mjk}。f 为时间 t 之前分布于混杂因素各层中的病例数占相应时点不分层前病例数的比例，即 $f_{jk}=l_{jk}/l_{\cdot k}, (j=1,\ldots,J, k=1,\ldots,K)$。

2. 生存期比较：不同样本间生存期（预后）差别的比较，是结合了各个随访时点的生存情况，与某时点率的比较是不能互相代替的。第二讲中已对此有专门讨论。例如男女性 5 年生存率存在差别，但并不能表明两者整个生存期也存在差别；反之亦然。国人对所谓的时序检验（logrank test）的简化式比较熟悉，该统计量应用最多。

暂以两组为例，时序检验统计量的简化式[2,5,10~14]为：

$$\chi_p^2=\frac{\left[\sum_{k=1}^{K}d_{1k}-\sum_{k=1}^{K}E(d_{1k})\right]^2}{\sum_{k=1}^{K}E(d_{1k})}+\frac{\left[\sum_{k=1}^{K}d_{2k}-\sum_{k=1}^{K}E(d_{2k})\right]^2}{\sum_{k=1}^{K}E(d_{2k})} \quad (4)$$

式中 $E(d_{mk})=l_{mk}\times D_{mk}/T_k, m=1,2$。上式又称 Peto logrank 检验统计量。时序检验的 Mantel logrank 检验统计量[5,8,14]稍复杂些，涉及到方差的运算，见下式：

$$\chi_M^2=\left[\sum_{k=1}^{K}d_{1k}-\sum_{k=1}^{K}E(d_{1k})\right]^2\Big/\sum_{k=1}^{K}Var(d_{1k}) \quad (5)$$

期望值 $E(d_{1k})$ 算法同上，而方差为：

$$Var(d_{1k})=\left[D_k\times(l_{1k}+l_{2k}-D_k)\times l_{1k}\times l_{2k}\right]\Big/\left[T_k^2\times(T_k-1)\right]。$$

若 $D_k=1$（即每个死亡时间点上仅有一例死亡发生），可以简化为：

$$E(d_{1k})=l_{1k}/T_k, Var(d_{1k})=(l_{1k}\times l_{2k})/T_k^2。$$

多组比较时，（4）式很容易扩展，计算仍然不复杂。但（5）式将变得更复杂，即

$$\chi^2=S'V^{-1}S \quad (6)$$

其中 S 为 1 至 $m-1$ 组的实际死亡数与期望死亡数之差的矩阵，S' 为其转置矩阵。V^{-1} 为 V 的逆矩阵，而 V 即是方差与协方差矩阵。

类似于调整生存率，当要控制混杂因素时，可按

作者单位：上海市肿瘤研究所流行病学研究室　200032

混杂因素分层,计算各层的检验统计量;然后合并做卡方检验,自由度为比较组数减 1。

3. 趋势检验[2,13]:当研究者所感兴趣的因素存在某种自然顺序时,如年龄从小到大、病期从 iv 期到Ⅰ期等,还可进行趋势检验。以判明生存率是否存在随着研究因素顺序的变化也存在一种趋势性变化。假定某因素有 g 个等级,$g = 1, \ldots, G$。趋势检验(trend test)公式如下:

$$\chi^2_T = \left[\sum_{g=1}^{G}\left[g \times \left(O_g - E_g\right)\right]\right]^2 \Big/ \left\{\sum_{g=1}^{G}\left(g^2 \times E_g\right) - \left[\sum_{g=1}^{G}\left(g \times E_g\right)\right]^2 \Big/ \sum_{g=1}^{G} E_g\right\} \quad (7)$$

其中 O 和 E 分别代表实际与期望死亡数。根据卡方分布,自由度取 1 做假设检验。

二、实例说明:表 1 是某地 1972 年与 1980 年两年的结肠癌随访资料。这是一个大样本随访资料,随访时间以年为单位。随访截止日期在 1987 年年底,故寿命表区间数定为 8。采用寿命表法估计两组资料的生存率列在表中。很显然,1980 年各年生存率均高于 1972 年。考虑到两个年份肿瘤病人的年龄分布可能不同,即可按年龄分层,然后估计调整生存率。为了简单起见,年龄分为两层:年龄≤59 岁和年龄≥60 岁。可按上面所述方法(公式 3)计算调整生存率,结果列于表 2。第 15 和 16 栏为 1972 和 1980 年两年的调整(累积)生存率估计值。表中 11 至 14 栏的计算可以看下面的实例,以 f_{11} 为例,$f_{11} = (170.0 + 233.5)/809.5 = 0.4985$。以 s_{11}(表中第 13 栏)为例,$s_{11} = 0.4985 \times (1 - \frac{61}{170.0}) + 0.5015 \times (1 - \frac{80}{154.0}) = 0.5606$。与表 1 未调整生存率相比,1972 年的调整生存率比调整前小,而 1980 年比调整前大,似乎拉大了两个年份生存率之间的差别。这正是分层调整的作用。

表 1 某地 1972 年与 1980 年结肠癌随访资料的寿命表生存率分析

随访时间	期初病例数		期内死亡数		截尾病例数		区间别生存概率		累积生存率	
(年)	1972	1980	1972	1980	1972	1980	1972	1980	1972	1980
0~	324	486	141	200	0	1	0.5648	0.5881	0.5648	0.5881
1~	183	285	42	53	0	0	0.7705	0.8140	0.4352	0.4787
2~	141	232	25	27	0	0	0.8227	0.8836	0.3580	0.4230
3~	116	205	18	19	0	0	0.8448	0.9073	0.3025	0.3838
4~	98	186	5	12	0	0	0.9490	0.9355	0.2870	0.3590
5~	93	174	10	14	0	1	0.8925	0.9193	0.2562	0.3301
6~	83	160	2	6	0	14	0.9759	0.9608	0.2500	0.3171
7~	81	140	6	5	75	135	0.8621	0.9310	0.2155	0.2952

表 2 某地 1972 年与 1980 年结肠癌病例随访资料调整生存率的估计(按年龄分为两层)

随访区间	合计校正人数	1972 年				1980 年				分布于混杂因素各层中病例数占总例数的比例		调整生存概率(区间别)		调整累积生存率(累积)	
		年龄≤59 岁		年龄≥60 岁		年龄≤59 岁		年龄≥60 岁							
(年)	N'	n'	d	n'	d	n'	d	n'	d	≤59	≥60	1972	1980	1972	1980
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
0~	809.5	170.0	61	154.0	80	233.5	74	252.0	126	0.4985	0.5015	0.5606	0.5913	0.5606	0.5913
1~	468.0	109.0	20	74.0	22	159.0	26	126.0	27	0.5726	0.4274	0.7679	0.8148	0.4305	0.4817
2~	373.0	89.0	14	52.0	11	133.0	13	99.0	14	0.5952	0.4048	0.8207	0.8846	0.3533	0.4261
3~	321.0	75.0	11	41.0	7	120.0	7	85.0	12	0.6075	0.3925	0.8439	0.9091	0.2982	0.3874
4~	284.0	64.0	1	34.0	4	113.0	5	73.0	7	0.6232	0.3768	0.9459	0.9363	0.2820	0.3627
5~	266.5	63.0	5	30.0	5	107.5	5	66.0	9	0.6398	0.3602	0.8892	0.9211	0.2508	0.3341
6~	236.0	58.0	0	25.0	2	98.5	0	54.5	6	0.6631	0.3369	0.9731	0.9629	0.2440	0.3217
7~	116.0	31.0	4	12.5	2	49.0	4	23.5	1	0.6897	0.3103	0.8614	0.9305	0.2102	0.2994

注:N' 及 n':校正人数= 期初病例数- 截尾病例数/2;d:期内死亡人数

表 3 为两年份比较的 logrank 检验统计量的计算步骤,同时给出了相对危险度的估计值(危险度或危险比的估计方法[16]将是下一讲的主要内容)。按式(4)、(5)计算的统计量为 3.2190($P = 0.0728$)、4.6470($P = 0.0311$)。相应地相对危险度估计值为 1.1617、1.2356。有趣的是,按 Mantel 氏统计量判断,两组生存期差别有统计学意义。而 Peto 统计量显示差别没有统计学意义。后者小于前者。两者的差别及有关问题将在后文讨论。按年龄分层后,可以分别统计各层中两个年份的寿命表。有关数据列于表 4 中。分层后统计量分别为 4.2362($P = 0.0396$)、6.2394($P = 0.0125$),相对危险度为 1.1836、1.2939。表中

还给出了各层的统计量计算步骤。无论是 Mantel 氏，还是 Peto 氏统计量都大于分层前。而且即使按照 Peto 氏统计量，两组生存期差别也有统计学意义。分层前后的统计量在值上的变化及意义与生存率调整前后的结果相一致。

趋势检验用在 3 及 3 个有序等级以上的情况，限于篇幅（至少 3 组数据），此处不在举例。读者可按公式（7）进行，注意自由度取值 1。计算上并不复杂。

表 3　某地 1972 年与 1980 年结肠癌病例生存期差别比较的 logrank 检验

随访区间	合计校正	合计死亡	实际死亡人数		期望死亡人数	
（年）	人　数	人　数	1972	1980	1972	1980
0~	809.5	341	141	200	136.48	204.52
1~	468.0	95	42	53	37.15	57.85
2~	373.0	52	25	27	19.66	32.34
3~	321.0	37	18	19	13.37	23.63
4~	284.0	17	5	12	5.87	11.13
5~	266.5	24	10	14	8.38	15.62
6~	236.0	8	2	6	2.81	5.19
7~	116.0	11	6	5	4.13	6.88

统计量计算：

Mantel´s logrank: $O_1=249, E_1=227.84$, $Var(O_1)=96.36$*, $\chi^2=4.6470$, RR= 1.24

Peto´s logrank: $O_2=249, E_2=227.84, O_2=336, E_2=357.16$, $\chi^2=3.2190$, RR= 1.16

* $Var(O_1)$ 为方差估计值（见正文中公式的说明）

三、结语：

1. 本文对生存率与预后因素分析中控制混杂因素的分层技术进行了系统介绍。与多变量分析技术相比，它的特点是简单、易懂、易学，而且手工即可运算，无需具备计算机及相应的计算机程序。

2. 需要注意的几个问题：首先是 logrank 检验统计量的两种公式之间的差别。生物统计学家已经证明，由(4)式估计的统计量（即 Peto 氏 logrank 检验统计量）总是小于(5)式（即 Mantel 氏 logrank 统计量）。Peto 氏法在统计检验上较为保守[11,13]。前面的例子也说明了这个问题。其次，当分析的预后因素水平数较多时，简化式在计算上有一定的优势，计算方便。因为(4)式不涉及到方差的运算，而(5)式要计算方差，尤其是多组比较时还要计算协方差。第三个问题是，若要控制 2 或 2 个以上的混杂因素，仍然可以按本文所述方法。只是分层越来越多，因而样本量要足够大。否则，有些层次例数会太少。而且随着研究因素与混杂因素的水平数增加，层数越来越多。第四点是，当研究因素为多个水平，且存在某种自然顺序时，可以计算趋势检验统计量。

表 4　结肠癌生存期比较的分层分析（分层 logrank 检验）

随访区间（年）	第一层：年龄≤59 岁						第二层：年龄≥60 岁					
	实际死亡数			合计校正人数	期望死亡数		实际死亡数			合计校正人数	期望死亡数	
	1972	1980	合计		1972	1980	1972	1980	合计		1972	1980
0~	61	74	135	403.5	56.88	78.12	80	126	206	406.0	78.14	127.86
1~	20	26	46	268.0	18.71	27.29	22	27	49	200.0	18.13	30.87
2~	14	13	27	222.0	10.82	16.18	11	14	25	151.0	8.61	16.39
3~	11	7	18	195.0	6.92	11.08	7	12	19	126.0	6.18	12.82
4~	1	5	6	177.0	2.17	3.83	4	7	11	107.0	3.50	7.50
5~	5	5	10	170.5	3.70	6.31	5	9	14	96.0	4.38	9.63
6~	0	0	0	156.5	0.00	0.00	2	6	8	79.5	2.52	5.48
7~	4	4	8	80.0	3.10	4.90	2	1	3	36.0	1.04	1.96

统计量计算：

第一层：Mantel´s logrank: $O_1=116, E_1=102.30, Var(O_1)=46.08$*, $\chi^2=4.0745$, RR= 1.35

Peto´s logrank: $O_1=116, E_1=102.30, O_2=134, E_2=147.70$, $\chi^2=3.1063$, RR= 1.25

第二层：Mantel´s logrank: $O_1=133, E_1=122.49, Var(O_1)=47.90$, $\chi^2=2.3073$, RR= 1.25

Peto´s logrank: $O_1=133, E_1=122.49, O_2=202, E_2=212.51$, $\chi^2=1.4222$, RR= 1.14

合　并：Mantel´s logrank $O_1=249, E_1=224.79, Var(O_1)=93.97$, $\chi^2=6.2394$, RR= 1.29

Peto´s logrank: $O_1=249, E_1=224.79, O_2=336, E_2=360.21$, $\chi^2=4.2362$, RR= 1.18

* 同表 3

3. 除了本文介绍的 logrank 检验, 还有一些非参数统计方法可以采用。如 Gehan 氏 Wilcoxon 型检验(主要用在两组比较)、Prentice 氏 Wilcoxon 型检验、Breslow 氏 Kruskal Wallis H 检验(多组比较)、Tarone- Ware 广义 logrank 检验等。这里简单介绍一下 Gehan 比分检验(score test)[17,18] 统计量的计算, 它也是生存分析中常用假设检验统计量, 用下式计算: $U=\sum_{k=1}^{K}[a_k\times d_k]/\left[\frac{n_1 n_2}{(n_1+n_2)(n_1+n_2-1)}\sum_{m=1}^{2}\sum_{k=1}^{K}[a_k^2\times d_k]\right]$。 (8)

截尾数据点 $a_k=S(t_k)-1$, 非截尾数据点 $a_k=S(t_k)+S(t_{k-1})-1$; 同样, 若是截尾数据点, d_k 要代之以截尾例数。利用标准正态分布可做假设检验。上式中的分子即为各组所有时间点上得分的和。多组比较时的公式类似于(6)式。

4. 如何选择使用这些统计方法, 以及在什么条件下采用何种统计量。从理论上讲, 这是生物统计学家探讨的问题。限于篇幅, 本文这里不多探讨。从应用角度出发, 笔者建议使用者在分析生存数据或预后因素时, 在Ⅵ材料与方法㈦中注明所采用的方法或统计量。例如, 不要笼统地只表达为采用Ⅵlogrank 检验㈦或Ⅵ卡方检验㈦因为由两种 logrank 检验公式计算的统计量结果不一样, 由此所下结论自然也不相同。所以注明采用何种统计量的目的, 是告诉读者你的科研结果或结论是源于何种假设检验统计量。

(文献备索)

聚合酶链反应快速检测钩端螺旋体

范 钦 曹东林 刘静宇

为了早期、快速、准确地诊断钩端螺旋体病(钩体病), 本研究利用 PCR 技术, 以问号钩体螺旋体(钩体)高度保守基因的一段序列为引物, 对广东省近年流行的 4 个血清型的典型株进行了扩增, 并对 1997 年 7~ 9 月清远市钩体病流行期 15 份早期血清进行了检测, 现报告如下。

一、材料与方法:

1. 菌株: 问号钩体 4 种血清型(赖型、犬型、秋季型和波摩那型)标准菌株由广东省卫生防疫站提供。

2. 钩体 DNA 提取: 钩体培养物离心后, 加 1% SDS 及终浓度 40μg/ml 的 ProtinaseK 于 37℃水浴裂解菌体, 以等体积酚: 氯仿抽提 3 次, 然后以等体积异丙醇低温沉淀 DNA, 沉淀物用 70% 乙醇洗涤 2 次, 干燥后溶于适量 TE 中备用。

3. 患者血清标本处理: 取血清 50μl 及硅粒液 40μl 加入 900μl 裂解液中〔裂解液配制: 120g GuSCN 加 100ml 0.1mol/L Tris- HCl(pH6.4), 加 22ml 0.2 mol/L EDTA(pH8.0) 和 2.6gTriton X- 100, 搅溶〕, 旋涡 5 秒钟, 室温置 10 分钟, 再旋涡 5 秒。1 200g 离心 15 秒。弃上清, 用冲洗液(120g GuSCN 加 100ml 0.1mol/L Tris- HCl pH6.4)、70% 乙醇各洗涤 2 次, 丙酮洗涤 1 次, 弃上清后于 56℃干燥 10 分钟, 加入 50μl TE 并旋涡, 置 56℃ 10 分钟, 12 000g 离心 2 分钟, 取上清备用。

4. PCR 引物设计及合成: 参照钩体 16S rRNA 基因序列, 设计出一对引物 R_1、R_2, 扩增片段长度为 270bp。引物序列为: R_1: 5㈦- 43 CGCGTCTTAAACA TGCAAGTCAAGC- 3㈦(G+C mol%=48.0%); R_2: 5㈦- 312CCCGTGTTACCTTGACTCT- 3㈦(G+C mol%=52.6%)。

5. PCR 扩增分析及对患者血清标本的检测: 取钩体 DNA 抽提液 5μl(0.1ng) 或患者血清抽提物 40μl 进行扩增反应, 每个循环包括: 95℃解链 1 分钟, 55℃退火 30 秒, 72℃延伸 45 秒, 共 35 个循环。每管取扩增后产物 5μl, 电泳检测扩增结果。

二、结果与讨论: 对钩体病流行期的 15 例疑似患者 PCR 检测结果为 7 例阳性, 与临床确诊的 7 例患者完全符合(先后由当地防疫部门采用血培养及血清学试验确证), 阳性符合率为 100%, 明显高于医院常规方法(MAT) 的检出率(采样时确诊 5 例阳性), 且 PCR 检测中设立了阳性、阴性对照, 保证了结果的可信度, 说明采用的 PCR 检测方法, 快速、简便且准确率高。

作者单位: 广州医学高等专科学校 510315

(收稿: 1998- 04- 16 修回: 1998- 05- 20)