· 系列讲座 ·

肿瘤流行病学研究资料的统计分析

项永兵

第二讲　生存率置信区间估计和比较

生存率的分析和比较在国内虽然应用广泛，但有些方面却不够深入，或有些方法实际应用较少。例如生存率置信区间的估计、多样本时点生存率比较、调整生存率估计、生存期差别比较的分层调整、单因素生存分析中危险度的估计、相对生存率的估计和统计学检验、期望寿命的估计等。本文将集中讨论生存率置信区间的估计方法，以及两组或多组生存率统计学检验的方法。

一、统计方法：

1. 生存率估计：生存率的估计方法主要是 Kaplan- Meier 乘积限估计（product limit estimate）[1,3~7]和寿命表法(life table method)[2~7]。前者主要用在小样本量的生存资料分析，而后者常用在大样本量的随访资料分析。Kaplan-Meier 法可简称 KM 法或 PL 法，而寿命表法则简称为 LT 法。用 $n(t_i)$、$d(t_i)$、$w(t_i)$分别表示在时间或随访区间 t 的期初存活病例数、期内死亡数、期内终检数，KM 法生存率 $\hat{S}(t_i)$估计为

$$\hat{S}_{KM}(t_i)=\prod\left[\frac{n(t_i)-d(t_i)}{n(t_i)}\right],i=1,2,\cdots,m \qquad (1)$$

而 LT 法估计公式为

$$\hat{S}_{LT}(t_i)=\prod\left[1-\frac{d(t_i)}{n(t_i)-w(t_i)/2}\right],i=1,2,\cdots,m \qquad (2)$$

生存率标准误 $se(\hat{S}(t_i))$计算公式[8]为

$$se(\hat{S}(t_i))=\hat{S}(t_i)\sqrt{\sum_{i=1}^{m}\frac{1-\hat{S}(t_i)}{n(t_i)-d(t_i)}}\text{或}$$

$$se(\hat{S}(t_i))\approx\hat{S}(t_i)\sqrt{\sum_{i=1}^{m}\frac{d(t_i)}{n(t_i)[n(t_i)-d(t_i)]}} \qquad (3)$$

2. 置信区间估计[5,6,9~11]：根据是否需要转换，置信区间估计分为两类：非转换方法和转换法。前者包括下述的方法一、二，后者为方法三至五。假定 $\hat{S}(t_i)$、$\hat{Var}[\hat{S}(t_i)]$表示生存率的估计值及其 Greenwood 方差[8]估计值，并设时间区间为 m，$Z_{\alpha/2}$为标准正态分位数。

方法一（经典法）：

$$\hat{S}(t_i)\pm Z_{\alpha/2}\times\sqrt{\hat{Var}[\hat{S}(t_i)]},i=1,2,\cdots,m \qquad (4)$$

方法二（校正法）：

$$\frac{N'}{N'+Z_{\alpha/2}}\left[\hat{S}(t_i)+\frac{Z^2_{\alpha/2}}{2N'}\pm Z_{\alpha/2}\times V\right],i=1,2,\cdots,m \qquad (5)$$

式中：

$$N'=\frac{\hat{S}(t_i)\times[1-\hat{S}(t_i)]}{\hat{Var}[\hat{S}(t_i)]},$$

$$V=\sqrt{\frac{\hat{S}(t_i)\times[1-\hat{S}(t_i)]}{N'}+\frac{Z^2_{\alpha/2}}{4(N')^2}}$$

方法三（反正旋转换）：

$$sin^2\left\{sin^{-1}[\hat{S}(t_i)]\pm\frac{Z_{\alpha/2}\times\sqrt{\hat{Var}[\hat{S}(t_i)]}}{2\sqrt{\hat{S}(t_i)\times[1-\hat{S}(t_i)]}}\right\} \qquad (6)$$

方法四（log(-log)转换）：

$$EXP\left\{-EXP\left\{[\log(-\log\hat{S}(t_i))]\pm\frac{Z_{\alpha/2}\times\sqrt{\hat{Var}[\hat{S}(t_i)]}\times[\log\hat{S}(t_i)]}{\hat{S}(t_i)}\right\}\right\} \qquad (7)$$

方法五（logit 转换）：

$$\frac{\hat{S}(t_i)}{(L-1)\times\hat{S}(t_i)+L}\quad\frac{L\times\hat{S}(t_i)}{(L-1)\times\hat{S}(t_i)+1} \qquad (8)$$

其中

$$L=EXP\left[\frac{Z_{\alpha/2}}{\sqrt{N'\times\hat{S}(t_i)\times[1-\hat{S}(t_i)]}}\right]\text{或}$$

$$L=EXP\left[\frac{Z_{\alpha/2}\times\sqrt{\hat{Var}[\hat{S}(t_i)]}}{\hat{S}(t_i)\times[1-\hat{S}(t_i)]}\right]$$

而 N'同方法二。

3. 生存率比较[3,4,7,13]：两个时点样本生存率的比较可采用 u 检验。统计量 u 可由下式计算，再按标准正态分布推断

$$u=\frac{|\hat{S}_1(t_i)-\hat{S}_2(t_i)|}{\sqrt{\hat{Var}[\hat{S}_1(t_i)]+\hat{Var}[\hat{S}_2(t_i)]}}。 \qquad (9)$$

上式即 u 检验公式，大家比较熟悉。对于多个样本（例如 K 组）时点率的比较可采用下述公式，计算卡方统计量

作者单位：上海市肿瘤研究所流行病学研究室　200032

$$\chi^2=\sum_{k=1}^{K}W_k(t_i)[\hat{S}_k(t_i)-\bar{S}(t_i)]^2=\sum_{k=1}^{K}\frac{[\hat{S}_k(t_i)-\bar{S}(t_i)]^2}{Var[\hat{S}_k(t_i)]},\tag{10}$$

其中 $\hat{S}_k(t_i)$、$W_k(t_i)$为比较的 K 个样本生存率中第 k 个生存率及其方差的倒数，后者也相当于权数。$\bar{S}(t_i)$则为加权平均生存率，取下式

$$\bar{S}(t_i)=\sum_{k=1}^{K}[W_k(t_i)\times\hat{S}_k(t_i)]\Big/\sum_{k=1}^{K}W_k(t_i)$$
$$=\left[\sum_{k=1}^{K}\frac{\hat{S}_k(t_i)}{Var[\hat{S}_k(t_i)]}\right]\Big/\sum_{k=1}^{K}\frac{1}{Var[\hat{S}_k(t_i)]}。$$

无效假设条件下 χ^2 服从自由度为 $K-1$ 的卡方分布。

二、实例说明：首先看小样本临床流行病研究资料的 KM 法分析。表 1 为利用 Freireich 白血病临床试验研究数据[12]，估计了试验和对照两组病人的 KM 法生存率。表 2 列出了试验组各年生存率的五种 95%置信区间。表 3 是某地某年女性乳腺癌 5 年随访资料的寿命表生存率估计。而各年生存率 95%置信区间的估计列于表 4。

表 1 Freireich 白血病临床试验数据(生存时间单位：周)的 Kaplan-Meier 乘积限估计

对照组						6-MP 治疗组					
随访时点 i	期初观察人数 $n(t_i)$	期内死亡数 $d(t_i)$	时点生存率 $s(t_i)$	累积生存率 $S(t_i)$	标准误 $se[S(t_i)]$	随访时点 i～	期初观察人数 $n(t_i)$	期内死亡数 $d(t_i)$	时点生存率 $s(t_i)$	累积生存率 $S(t_i)$	标准误 $se[S(t_i)]$
1～	21	2	0.9048	0.9048	0.0641	6～	21	3	0.8571	0.8571	0.0764
2～	19	2	0.8947	0.8095	0.0857	7～	17	1	0.9412	0.8067	0.0869
3～	17	1	0.9412	0.7619	0.0929	10～	15	1	0.9333	0.7529	0.0963
4～	16	2	0.8750	0.6667	0.1029	13～	12	1	0.9167	0.6902	0.1068
5～	14	2	0.8571	0.5714	0.1080	16～	11	1	0.9091	0.6275	0.1141
8～	12	4	0.6667	0.3810	0.1060	22～	7	1	0.8571	0.5378	0.1282
11～	8	2	0.7500	0.2857	0.0986	23～	6	1	0.8333	0.4482	0.1346
12～	6	2	0.6667	0.1905	0.0857						
15～	4	1	0.7500	0.1429	0.0764						
17～	3	1	0.6667	0.0952	0.0641						
22～	2	1	0.5000	0.0476	0.0465						
23～	1	1	0.0000	0.0000	0.0000						

注：1. 时点生存率的计算：如对照组 1 周时点生存率：1～2/21＝0.9048，2 周：1～2/19＝0.8947，余依次类推。

2. 累积生存率的计算：仍以对照组为例，1 周＝0.9048，2 周＝0.9048×0.8947＝0.8095；再如 5 周＝0.9048×0.8947×0.9412×0.8750×0.8571＝0.5714，余依次类推。

表 2 对照组病人 Kaplan-Meier 生存率五种置信区间的估计值(95%CI)

时间区间 i	经典法		校正法		反正旋转换		log(-log)转换		logit 转换	
	下限	上限	下限	上限	下限	上限	下限	上限	下限	上限
1～	0.7792	1.0303	0.7225	0.9618	0.8937	0.9153	0.6700	0.9753	0.6887	0.9761
2～	0.6416	0.9775	0.6093	0.9140	0.7830	0.8347	0.5689	0.9239	0.5885	0.9266
3～	0.5797	0.9441	0.5578	0.8850	0.7281	0.7941	0.5194	0.8933	0.5397	0.8973
4～	0.4650	0.8683	0.4617	0.8201	0.6212	0.7107	0.4253	0.8250	0.4467	0.8321
5～	0.3598	0.7831	0.3731	0.7476	0.5193	0.6228	0.3380	0.7492	0.3597	0.7599
8～	0.1732	0.5887	0.2153	0.5834	0.3327	0.4305	0.1831	0.5778	0.2032	0.5975
11～	0.0925	0.4789	0.1464	0.4913	0.2471	0.3259	0.1166	0.4818	0.1343	0.5076
12～	0.0225	0.3584	0.0860	0.3907	0.1653	0.2170	0.0595	0.3774	0.0734	0.4115
15～	−0.0068	0.2925	0.0600	0.3361	0.1250	0.1617	0.0357	0.3212	0.0468	0.3614
17～	−0.0303	0.2208	0.0382	0.2775	0.0847	0.1063	0.0163	0.2613	0.0239	0.3113
22～	−0.0435	0.1387	0.0228	0.2123	0.0436	0.0518	0.0033	0.1970	0.0067	0.2714

两样本生存率的统计学检验比较简单。例如某恶性肿瘤两个期别的 5 年生存率为 0.7792(0.0059)、0.6021(0.0576)，括号内为其标准误。由(9)式计算的统计量 u 值为 3.0586，则两者的 5 年生

存率的差异有高度统计学意义。

同样是该恶性肿瘤，其三种分化程度的5年生存率分别为0.8402(0.0166)、0.7127(0.0397)、0.3894(0.1046)。先计算加权平均生存率，其估计值为0.8122；再由(10)式计算卡方统计量。χ^2为25.4649，自由度为2，则三组5年生存率的差异有高度统计学意义。

表3　某地某年女性乳腺癌生存率分析的寿命表(五年随访资料，生存时间：年)

时间区间 i	期初观察人数 $n(t_i)$	期内死亡人数 $d(t_i)$	期内终检人数 $w(t_i)$	校正人数 $n'(t_i)$	区间生存率 $s(t_i)$	生存率 $S(t_i)$	生存率标准误 $se[S(t_i)]$
0～	817	95	0	817.0	0.8837	0.8837	0.0112
1～	722	92	1	721.5	0.8725	0.7710	0.0147
2～	629	54	11	623.5	0.9134	0.7043	0.0160
3～	564	34	138	495.0	0.9313	0.6559	0.0169
4～	392	22	132	326.0	0.9325	0.6116	0.0182

注：1. 校正人数计算：例如区间0～：817～0/2=817.0，区间1～：722～1/2=721.5。

2. 区间生存率计算：例如区间0～：1～95/817.0=0.8837，

区间1～：1～92/721.5=0.8725，等。

3. 累积生存率计算：例如区间2～：0.8837×0.8725×0.9134=0.7043，等。

表4　某地某年女性乳腺癌生存率五种置信区间的估计值(95%CI)

时间区间 i	经典法		校正法		反正旋转换		log(-log)转换		logit 转换	
	下限	上限	下限	上限	下限	上限	下限	上限	下限	上限
0～	0.8617	0.9057	0.8600	0.9038	0.8814	0.8859	0.8597	0.9038	0.8599	0.9039
1～	0.7422	0.7998	0.7410	0.7985	0.7659	0.7761	0.7406	0.7983	0.7409	0.7985
2～	0.6729	0.7357	0.6721	0.7345	0.6977	0.7108	0.6716	0.7344	0.6720	0.7347
3～	0.6228	0.6890	0.6221	0.6882	0.6484	0.6634	0.6216	0.6879	0.6221	0.6882
4～	0.5759	0.6473	0.5755	0.6465	0.6031	0.6201	0.5749	0.6462	0.5754	0.6466

三、结语：

1. 从统计学角度讲，本文所提到的生存率术语是指观察生存率(observed survival rate，OSR)，也是国内应用最广的一种指标。习惯上称之为"生存率"。相对于观察生存率而言，有相对生存率(relative survival rate，RSR)。在国外RSR是通用的和标准的生存率分析指标。笔者将在后面的文中介绍RSR。

2. 需要注意的是，采用何种方法估计生存率的置信区间，是研究者需要在你的论文或研究报告中提到的。不可含糊或笼统地讲"估计生存率的置信区间"，因为不同的方法结果不同。

3. 不同的计算机软件，其中生存率标准误的估计方法也不同。比较新的统计或流行病学计算机软件，可能已经不再采用老的方法计算标准误，如Greenwood法。同流行病学领域内应用较广泛的EGRET软件，它采用的方法为log(-log)转换法。

4. 生存率置信区间的估计有七种方法，其中有两种很少用。国人至今一直延用经典方法，即Greenwood法。不过它有一定的局限性，且目前不少统计软件其结果输出中的生存率CI估计已不用该法，但很多人并不注意这一点。

在小样本资料分析时，经典法的区间范围较宽，反正旋转换的区间范围较窄，其它三种方法的CI比较接近。而且当样本例数较少时，经典法还会出现不合理的现象，如负值、大于1。对于大样本资料，除反正旋转换的CI区间较窄，其它四种方法均相接近。这是因为上述几种方法均是在大样本(分布)理论基础上推导的缘故。理论上的研究(如Monte Carlo模拟抽样)[10,11]表明，第二至第五种方法均适合于临床小样本资料的分析，它们也正是生物统计学家们所推荐应该采用的方法。

5. 某一时点生存率差异的比较，只能说明两组或多组样本它们在该时点(如五年)上的差异，并不能代表它们整个生存期或随访期内的预后有差别；反之亦然。利用几个假设的生存率曲线图来直观地说明问题。第一种情况是两组时点生存率和生存期的差异都有统计学意义，见图1；第二种情况两组生存期差异有统计学意义，但某一或某些时点生存率的差异可能并没有统计学意义，见图2；图3的情形

则可能是两组生存期差异没有统计学意义，同时他们的某些时点生存率差异也没有统计学意义。实际上，在多变量分析（如 Cox 模型）中经常谈到的“比例危险假设”即是图 1 这种情况。而图 2 和图 3 则是所谓的“非比例危险假设”情形。

多个样本生存率的统计学检验，国内文献[13]对此有过详细讨论。

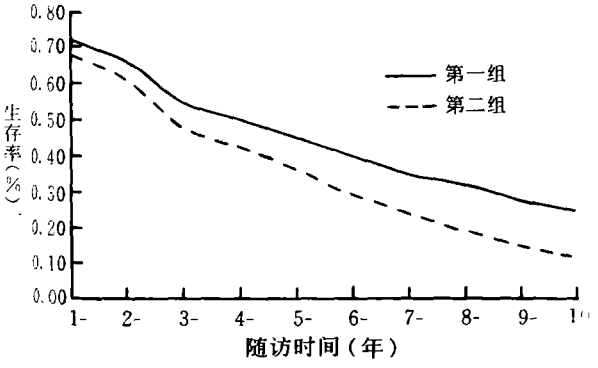

图 1 两组生存率曲线（成比例情形）

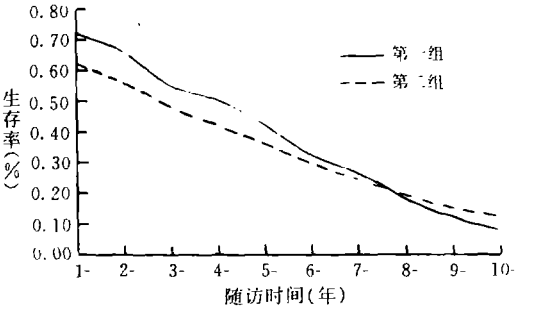

图 2 两组生存率曲线（非比例危险情形）

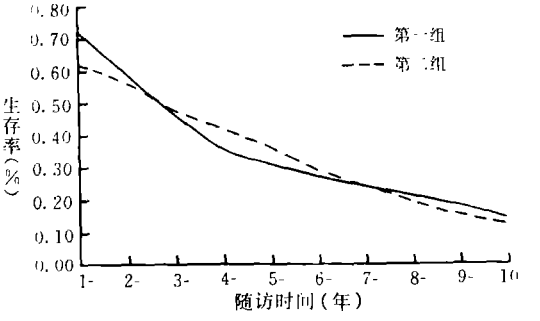

图 3 两组生存率曲线（非比例危险情形）

· 读者福音 ·

发明创新工程 · 一次投资 · 终身受益 · 莫失良机

本刊举办“现代医学发明方法学与临床流行病学培训班”的通知

近年来，现代医学发明方法学及临床流行病学发展很快，现职卫生医务人员有关知识亟待补充，为了科技兴医，培养独创型的人才，促进医学发明的发展，提高临床科研水平，本刊拟举办一期“现代医学发明方法学与临床流行病学培训班”。

一、时间：1998 年 10～12 月，共 3 个月。

二、方法：该班以函授为主，聘请我国知名的医学科研方法学、流行病学专家章扬熙教授等专家任教。采取做习题、发考卷、给标准答案等方法开展教学活动。

三、学习内容：培训班采用专家新编教材，内容新颖，并以实例说明，学以致用。主要内容包括：1. 医学发明的系统分析；2. 医学发明的选题；3. 医学发明的思考方法（系统介绍 10 余种方法）；4. 医学发明的研究方法（应用系统动力学原理来说明）；5. 医学发明的实例（介绍几十个较重要或典型的医疗、保健、药物、管理各方面发明实例）；6. 病因的流行病学研究；7. 疾病诊断研究；8. 病人预后研究；9. 疾病疗效研究（包括时间效应分析）；10. 卫生经济分析与评价。培训班适合各级医院、卫生院、卫生防疫、妇幼保健、医学院校及科研单位的临床、预防、教学、科研人员和卫生行政管理人员参加，非学员一律不售资料，请勿寄款。

四、考核及结业：采用发、收考卷的方式进行考试。及格者发结业证书，并通过本刊公布优秀学员名单，供有关单位和部门使用干部和晋升时参考。

五、报名及学费：报名日期为 1998 年 4 月 10 日～9 月 30 日。请用楷体写明姓名、性别、年龄、职称、单位、详细通讯地址及邮政编码。报名同时邮寄学费（含资料费）180 元（开收据、报销）。一律寄至北京昌平流字五号《中华流行病学杂志》编辑部**尹麻**收（邮政编码 102206）。款到陆续寄资料，本期名额有限，按报名顺序录取，未被录取者一律退款。