

生存率置信区间的五种估计方法

项永兵¹ 高玉堂¹ 金凡¹ 杨工² 孙璐¹

摘要 生存率是医学随访研究资料分析中常用的指标,例如适用于小样本资料的 Kaplan-Meier 乘积限估计和大样本资料的寿命表法生存率估计。本文对生存率置信区间的估计方法进行了讨论。主要介绍了五种置信区间的估计方法:经典法(基于 Greenwood 方差公式)、校正法、反正旋转换法、log(-log) 转换法及 logit 转换法。文中给出了两个实例,并就生存率95%置信区间的估计做了详细介绍,还进一步讨论了它们在小样本和大样本随访资料分析中的特点和应用。

关键词 生存率 置信区间 寿命表

Five Methods for the Estimation of Confidence Intervals on Survival Rate Xiang Yong-bing, Gao Yu-tang, Jin Fan. Department of Epidemiology, Shanghai Cancer Institute, Shanghai 200032

Descriptive analysis on medical follow-up data often involves estimation of survival rates, such as Kaplan-Meier product limit estimator or life table survival rate for small or large scale data sets, respectively. Five methods for estimating the confidence intervals (CI) of survival rates were described. These methods for estimation of intervals were the classic (based on Greenwood's formula), correctness, arcsine transformation, log(-log) transformation and logit transformation. Two examples were described in detailed for the calculation of 95% confidence intervals.

Key words Survival rate Confidence intervals Life table estimator

生存率是描述性生存分析中最重要的指标,可估计其置信区间(如95%、99%等)。本文将系统讨论生存率置信区间的估计方法,及其实际应用中的有关问题。

统计背景

用 $\hat{S}(t_i)$ 表示生存率(累积)的 Kaplan-Meier 乘积限估计值或寿命表法估计值^[1], $\text{Var}[\hat{S}(t_i)]$ 表示 $\hat{S}(t_i)$ 方差的 Greenwood 估计值^[2], $Z_{\alpha/2}$ 表示显著性水平(第一类错误)为 α 时的标准正态分位数。并假定随访时间分成 m 个区间: $t_1, \dots, t_m$ 。

方法一(经典法)^[3,4]:

$$\hat{S}(t_i) \pm Z_{\alpha/2} \cdot \sqrt{\text{Var}[\hat{S}(t_i)]}, \quad i = 1, 2, \dots, m。$$

方法二(校正法)^[4,5]:

$$\frac{N'}{N' + Z_{\alpha/2}^2} \cdot [\hat{S}(t_i) + \frac{Z_{\alpha/2}^2}{2N'} \pm Z_{\alpha/2} \cdot V],$$

$i = 1, 2, \dots, m。$

式中

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

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

方法三(反正旋转换)^[3,4]:

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

方法四 [log(-log)] 转换^[3,4]:

$$\exp \{ -\exp \{ [\log(-\log \hat{S}(t_i))] \} \pm \frac{Z_{\alpha/2} \cdot \sqrt{\text{Var}[\hat{S}(t_i)] \cdot [\log \hat{S}(t_i)]}}{\hat{S}(t_i)} \}$$

方法五(logit 转换)^[4]:

$$\text{下限: } \frac{\hat{S}(t_i)}{[(L-1) \cdot \hat{S}(t_i) + L]},$$

$$\text{上限: } \frac{L \cdot \hat{S}(t_i)}{[(L-1) \cdot \hat{S}(t_i) + 1]},$$

式中 $L = \exp \left[\frac{Z_{\alpha/2}}{\sqrt{\hat{S}(t_i) \cdot N' \cdot (1 - \hat{S}(t_i))}} \right]$, 其中 N'

1 上海市肿瘤研究所流行病学研究室 200032

2 浙江医科大学肿瘤研究所

表3 女性乳腺癌样本资料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 $SE_{LT}(t_i)$	生存率 $\hat{S}_{LT}(t_i)$	生存率标准误 $\hat{SE}_{LT}(t_i)$
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-: $0.8837 \times 0.8725 \times 0.9134 = 0.7043$, 等。

表4 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

Kaplan-Meier 生存率及其标准误的估计公式如下:

$$\hat{S}_{KM}(t_i) = \prod \left[\frac{n(t_i) - d(t_i)}{n(t_i)} \right],$$

$$\hat{SE}_{KM}(t_i) = \hat{S}_{KM}(t_i) \cdot \sqrt{\sum \frac{1 - \hat{S}(t_i)}{n(t_i) - d(t_i)}},$$

而寿命表法的估计公式为

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

$$\hat{SE}_{LT}(t_i) =$$

$$\hat{S}_{LT}(t_i) \cdot \sqrt{\sum \frac{1 - \hat{S}(t_i)}{(1 - \hat{S}(t_i)) \cdot \left[n(t_i) - \frac{w(t_i)}{2} \right]}},$$

式中 $\hat{S}_{KM}(t_i)$ 、 $\hat{S}_{LT}(t_i)$ 、 $\hat{S}(t_i)$ 、 $n(t_i)$ 、 $d(t_i)$ 、 $w(t_i)$ 的意义见表1和表3中各栏目的标示。需要说明的是, 小样本资料的生存率的估计是针对每一个观察死亡点而言, 在截尾时间点上生存率不变。而寿命表法是计算各个时间区间的生存率, 截尾数据的贡献要考虑进去, 需计算校正人数用于估计生存率。截尾数据主要指失访、终检病例数。

从上述分析结果可以看出下面几点:

(1) 临床随访资料: 5种方法估计的95%置信区间之间有一定差别。其中校正法、log(-log)转换和logit转换法比较接近, 而经典法的区间范围较宽, 反正旋转换法的区间范围又较窄, 而且经典法估计范围出现了不合理现象, 如大于1或负值。(2) 人群随访资料: 5种方法估计的置信区间相当接近, 仅反正旋转换法的区间范围较窄。(3) 从理论上讲, 上述5种方法均是在大样本(分布)理论的基础上推导的。正因为如此, 在人群随访资料分析中, 几种方法的估计值之间都非常接近。人群随访资料一般样本量较大, 但临床随访资料的样本较小, 各种方法的结果之间的差别表现的就明显些。(4) 本文的分析结果提示我们, 在临床小样本资料分析时, 要慎用置信区间估计的经典法。进一步从理论上说明或探讨此类问题, 国外已有作者做了一些工作(采用 Monte Carlo 模拟抽样研究法)^[4,5], 它们的结果表明, 第2至第5种方法仍然适用于小样本资料的分析。

结 语

笔者就生存率置信区间的估计,介绍了5种方法(笔者编有计算程序)。国人对方法1较熟悉,也是人们较常采用的方法。但正如前文所述,对于临床小样本随访资料的分析,生物统计学家们推荐采用第2至第5种方法^[4,5]。所以一些流行病学或生物统计微机软件(如EGRET),其中所采用的方法为log(-log)转换法。这是国人今后在资料分析中值得注意的问题。参照生存率95%置信区间的估计,亦可很方便地给出其它百分位置信区间(如90%、99%等)的估计值。

参 考 文 献

- 1 Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 1958, 33:457.

- 2 Greenwood M. The error of sampling of the survivorship tables (Reports on Public Health and Medical Subjects, No. 33, Appendix 1). London: HM Stationery Office, 1926.
- 3 Afifi AA, Elashoff RM, Lee JJ. Simultaneous non-parametric confidence intervals for survival probabilities from censored data. *Statistics in Medicine*, 1986, 5:653.
- 4 Anderson JR, Bernstein L, Pike MC. Approximate confidence intervals for probabilities of survival and quantiles in life-table analysis. *Biometrics*, 1982, 38:407.
- 5 Rothman KJ. Estimation of confidence limits for the cumulative probabilities of survival in life table analysis. *J Chron Dis*, 1978, 31:557.
- 6 Freireich EJ. The effect of 6-mercaptopurine on the duration of steroidinduced remission in acute leukemia. *Blood*, 1963, 21:699.
- 7 项永兵,屠基陶. 恶性肿瘤生存数据的分析比较. *中国肿瘤情报*, 1991, 2:32.

(收稿:1994-09-22 修回:1994-11-10)

一起麻疹爆发的流行病学调查分析及病毒唑治疗疗效观察

王志荣¹ 孟宪红²

1994年4月6日至5月17日甘肃省张掖市梁家墩乡清凉寺村发生一起小儿麻疹爆发流行。在全村易感儿童514人中共发病71人,总罹患率13.81%。其中男37例,女34例,年龄5月龄至12岁。易感儿中5~7月龄(未到接种年龄)14人,其中发病7人,罹患率50%。如在易感人群和发病人群中均除外未到接种年龄者,则该两个人群的疫苗接种率分别为97.08%和88.73%,疫苗保护率为87.37%。全部病例按临床表现分为典型病例33例和非典型病例38例。

将全部病例分为治疗组42例和对照组29例。两组患儿均设家庭病床隔离观察治疗。全部病例中除合并肺炎者使用抗生素外,治疗组给予病毒唑治疗。剂量为10~15mg/kg,每日1次静滴,疗程4~6天。对照组只给予对症治疗和一般支持疗法。结果治疗组平均退热时间、退疹时间和总病程均比对照组明显缩短,差异有非常显著性($P<0.001$)。

本组病例有一个特点值得注意,即在未到接种年龄的5~7月龄组14人中竟有7人发病,致使该年龄组罹患率高达50%,远居其他各年龄组之首。这可能由于该年龄组正值从母体获得的先天免疫已经减弱乃至消失,而计划免疫尚未开始的空白阶段。此点似可提示,在麻疹流行地区,对该年龄组小儿应在接触感染后2日内给予自动免疫,或在整个潜伏期内给予被动免疫加以保护。或者应该考虑将免疫接种时间提前到5月龄开始。

病毒唑是一种广谱抗病毒药,对30多种RNA和DNA病毒有抑制作用。麻疹病毒是一种RNA病毒,用病毒唑治疗麻疹是一种新的尝试。本文治疗结果表明,病毒唑治疗麻疹疗效满意。

(收稿:1995-02-28 修回:1995-04-20)

1 甘肃省张掖市医院二门诊 734000

2 沈阳医学院学报编辑部