随机截尾试验下癌症患者的寿命分布模型的研究*

——基于福建省某医院从2003- 2007年肿瘤患者的调查

张志强　马　骅　王洁丹

内容提要: 本文基于福建省某医院对2003- 2007年肿瘤患者的调查, 按住院人数统计, 对顺位前三类癌症患者的寿命分布模型进行了研究。研究发现: (1)40- 50岁是女性子宫肿瘤发病的高峰期, 并且有逐年增长的趋势; (2)剔除女性患子宫肿瘤人数的影响, 肿瘤发病的高峰期转移到60- 70岁这个年龄段, 表现出其它肿瘤多发于老年人, 而且肺癌、胃癌、肝癌患者人数尤为突出; (3)肺癌、胃癌、肝癌患者的寿命分布属同一分布族, 并且均能通过随机截尾数据下的H-P检验。

关键词: 随机截尾试验; 癌症; 寿命分布
中图分类号: C812　　文献标识码: A　　文章编号: 1002- 4565(2009)12- 0096- 04

Research on Life Distribution Model of Cancer Patients Under Random Censoring Life Test

——based on the Survey of Tumor Patients in a Hospital in Fujian Province from 2003 to 2007

Zhang Zhiqiang　Ma Hua　Wang Jiedan

Abstract: Based on the survey of tumor patients in a hospital in Fujian Province from 2003 to 2007, this paper studied life distribution of cancer patients whose number is at top three. We find: Firstly, high-incidence period of uterine tumor is in 40 to 50 years old women, and incidence rate is increasing in recent years. Secondly, after removing the effect of uterine tumor, high-incidence period of the tumor is in 60 to 70 years old people and the number of patients suffering from lung cancer, gastric cancer, liver cancer are more than others. Lastly, the life distributions of lung cancer, gastric cancer and liver cancer belong to the same distribution family and the null hypothesis is not rejected by H-P test under random censored data.

Key words: random censoring test; cancer; life distribution

一、引言

随着医学的发展以及诊疗技术的提高, 越来越多的癌症被诊断出来, 癌症日益威胁着人类的健康。为提高人类健康水平提供科学依据, 本文探索了癌症患者寿命分布规律。

本文数据来源于福州市某三级甲等综合医院2003年1月至2007年12月所有住院的肿瘤患者(包括良性肿瘤和恶性肿瘤)的原始病案记录。考虑到数据的代表性, 本文依据原始病案记录中的病人来源一项将非本市病人剔除, 使得研究结果能较好地反映福州市癌症患者寿命分布规律。

寿命数据分析从20世纪50年代开始引起人们的关注, 出现了大量的研究成果, 主要应用于医学、生物学、保险学及工业产品的可靠性等领域。在医学领域中, 至今多数研究是依据Ⅰ型截尾数据和Ⅱ型截尾数据, 从研究的问题来看, 主要集中在3个方面, 一是不同疗法或新药治疗癌症患者生存情况的临床研究; 二是探索影响癌症患者生存时间的因素; 三是研究寿命数据的统计分析方法。其主要代表作

* 本文获得福建省自然科学基金计划项目"保险精算中修匀方法的研究及应用"(S0650038)的资助。

有: Glasser(1965) 关于肺癌患者存活时间的影响因素的研究; Prentice(1973) 依据 40 位后期肺癌患者的存活经历数据研究不同治疗方法的疗效问题; Maupas(1997) 关于肝癌治疗疗效的临床研究; Tiku (1981) 对截尾数据下 Weibull 分布的检验问题的探索; Cho (2006) 关于肺癌治疗疗效的临床研究; Vauhkonen(2006) 关于胃癌治疗疗效的临床研究; Kirsten(2008) 对任意截尾数据下生存函数的统计推断的研究等等。

纵观国内外文献, 基于住院病案信息研究癌症患者寿命分布的文献甚少, 但癌症患者生存时间分析在医学研究领域中占有极其重要的地位。通过对癌症患者生存分布模型的研究, 可以对癌症患者的生存分布有总体的了解, 这将有助于预测生存时间, 评价和监测肿瘤的二级预防和三级预防效果。为此, 本文基于福建省某医院 2003 年 1 月至 2007 年 12 月的所有住院病人的住院病案记录, 对肺癌、胃癌、肝癌患者进行了跟踪统计, 获取了关于这三类癌症的随机截尾数据, 在此基础上探索了三类癌症患者的寿命分布模型, 并依据模型获得了如下结论: 肺癌、肝癌发病高峰年龄为 60~ 74 岁, 胃癌发病高峰年龄为 75 岁以上, 中年期即 45~ 59 岁肝癌发病率要远高于肺癌、胃癌的发病率。这与国内其他地区相关文献报道得到了一致的结论。参见文献: 李连弟等(1997) , 贾安华(2001) , 杜其药(2002) , 尹桂成等(2001) , 赵力等(2002) , 顾海雁等(2006) , 华召来等(2006) , 金华(2008) , 等等。

二、基本情况

基于该医院于 2003 年 1 月至 2007 年 12 月的所有住院病人的住院病案记录, 对每年住院的肿瘤患者按 10 岁一组进行分组, 获得了各年度肿瘤患者数占总患者数按年龄分组的频率图, 见图 1。在图 1 中可直观地看到在 40~ 50 岁这一组中, 患肿瘤的频率随时间的推移在不断地增加, 而且各年患肿瘤的频率在这个年龄段始终达到最高; 在 40 岁之前, 患肿瘤的频率随时间的推移都出现了不断下降的趋势; 在 30~ 40 岁和 50~ 60 岁这两组中, 患肿瘤的频率随时间的推移基本不变; 60 岁以后各年度的患肿瘤的频率有所波动。再观察所收集的数据中, 我们发现女性子宫肿瘤发病数较高, 所以考虑剔除女性患子宫肿瘤人数, 再次统计各年度肿瘤患者数占总患者数按年龄分组的频率图, 如图 2 所示。对比图 1 与图 2, 可见患肿瘤的频率在 40~ 50 岁这一年龄段不再是最高, 肿瘤发病的高峰期转移到 60~ 70 岁这个年龄段, 表明 40~ 50 岁是女性子宫肿瘤发病的高峰期, 并且有逐年增长的趋势, 其它肿瘤多发于老年人。而且人们的平均寿命延长了, 肿瘤患者越来越多, 尤其是恶性肿瘤是老年人致病、致亡的主要疾病。依据该医院的统计数据分析可见顺位前三类癌症是肺癌、胃癌、肝癌。我们对这三类癌症患者的寿命分布进行了研究, 研究发现, 肺癌、胃癌、肝癌患者的寿命分布属于同一个分布族, 而且依据所获得的寿命分布可以掌握三类癌症发病率的分布, 为防癌治癌工作提供科学依据。

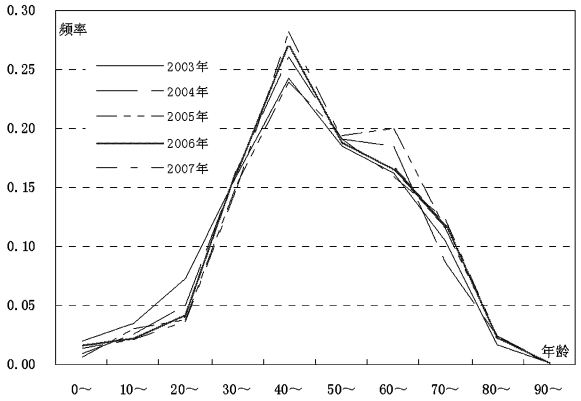

图 **1**　按年龄分组各年患肿瘤的频率图

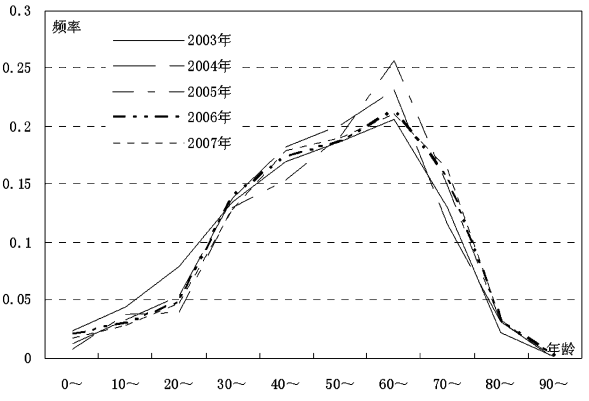

图 **2**　按年龄分组各年患肿瘤的频率图(剔除子宫肿瘤)

三、肺癌、胃癌、肝癌患者的随机截尾数据的获得

依据福建省某医院 2003 年 1 月至 2007 年 12 月的所有住院病人的住院记录, 对肺癌胃癌、肝癌患者进行了跟踪统计, 获取了关于这三类癌症的随机截

尾数据。但是由于我们无法确定首次在试验观察期 2003 年 1 月至 2007 年 12 月内出现的时间是否是该患者被第一次确诊为癌症的时间，于是本文在每个观察到的患者的存活天数上加上该患者首次在试验观察期出现时已经存活的天数（依据患者的出生日期）就可获得该患者的实际寿命。在这 5 年期间共观察到肺癌患者 112 人，其中存活的 38 人，死亡的 74 人；胃癌患者 54 人，其中存活的 21 人，死亡的 33 人；肝癌患者 87 人，其中存活的 23 人，死亡的 64 人。下一节我们依据所获得的随机截尾数据进行统计推断，以探索这三类癌症的寿命分布。

四、肺癌、胃癌、肝癌患者的寿命分布模型

（一）寿命分布的选择

指数分布、Weibull 分布和对数正态分布是常用的三大寿命分布。指数分布的优点是无论对于何种类的截尾数据其参数的估计均较其他两种模型简单，但由于其危险率函数为常数及指数分布的无记忆性，使其在生物医学中的应用受到限制。其次，由于对随机截尾数据取对数后，其寿命频数分布图的对称性不理想，所以本文不考虑使用对数正态分布。Weibull 分布是适用于某一局部失效而引起全部机能停止的现象。这一性能显然适合描述患有癌症的病人的寿命分布。同时本文利用 Kaplan-Meier 乘积限估计以及 Weibull 分布的生存函数 $S(t)$ 对 Weibull 分布的选择理由进行了验证。

对于 Weibull 分布的生存函数 $S(t)$，有

$$S(t) = \exp[-(\lambda t)^{\beta}],$$

从而

$$\log(-\log(S(t))) = \beta\log t + \beta\log\lambda.$$

对各 $\log t$ 的值描出 $\log(-\log(S(t)))$ 形成图形，若 Weibull 分布是合适的，做出的该图形应近似为一条直线。反之则不合适。其中 $S(t)$ 利用 Kaplan-Meier 乘积限方法获得，$S(t)$ 估计为：

$$\hat{S}(t) = \prod_{t_{(i)} \leqslant t,\ t_{(i)}\text{非截尾的}} \frac{n-i}{n-i+1} \quad (t_{(i)} < t \leqslant t_{(i+1)}).$$

由于三类癌症数据的 $\log(-\log(\hat{S}(t)))$ 散点图的图形与一条直线有较好的近似，因此选择 Weibull 分布是合适的。

（二）随机截尾数据下肺癌、胃癌、肝癌患者的寿命分布模型的参数估计

考虑到极值分布与 Weibull 分布有直接的关系，即当 T 服从 Weibull 分布且有密度函数时，$X = \log T$ 就服从 $b = \beta^{-1}$ 和 $u = -\log\lambda$ 的极值分布。在数据分析时使用对数寿命时间常常是方便的。所以本文利用极大似然估计方法首先估计极值分布的参数，然后利用极值分布与 Weibull 分布参数之间的关系，从而获得 Weibull 分布的参数估计。

Weibull 分布的危险率函数为：

$$h(t) = \lambda\beta(\lambda t)^{\beta-1}.$$

这里 $\lambda > 0$ 和 $\beta > 0$ 都是参数。其密度函数和生存函数分别为

$$f(t) = \lambda\beta(\lambda t)^{\beta-1}\exp[-(\lambda t)^{\beta}],\ t > 0$$

和

$$S(t) = \exp[-(\lambda t)^{\beta}],\ t > 0$$

易知该分布的 r 阶原点矩 $E(X^r)$ 为 $\lambda^{-r}\Gamma(1 + r/\beta)$，这里

$$\Gamma(k) = \int_0^{\infty} u^{k-1}e^{-u}du,\ k > 0$$

为 Gamma 函数，从而其均值和方差分别为 $\lambda^{-r}\Gamma(1+r/\beta)$ 和 $\lambda^{-2}[\Gamma(1+2/\beta) - \Gamma(1+1/\beta)^2]$。

极值分布的密度函数和生存函数分别为：

$$f(x) = b^{-1}\exp\left[\frac{x-u}{b} - \exp\left(\frac{x-u}{b}\right)\right],$$

$$-\infty < x < \infty$$

$$S(x) = \exp\left[-\exp\left(\frac{x-u}{b}\right)\right],\ -\infty < x < \infty$$

这里 $b > 0$ 和 $u(-\infty < u < \infty)$ 均为参数。

考虑常用的 I 型截尾数据的情形，这里 T_i 表示寿命时间，L_i 是容量为 n 的样本中第 i 个样本的固定截尾时间。人们能观察到的仅仅是 $t_i = \min(T_i, L_i)$，即观察到 t_i 是寿命时间或是截尾时间。诸 T_i 被假设服从 Weibull 分布，或等价地，$X_i = \log T_i$ 服从参数为 u 和 b 的极值分布。令 $\eta_i = \log L_i$，$x_i = \log t_i$ 和 $\delta_i = 1$ 和 0，要看 $t_i = T_i$ 还是 $t_i = L_i$。

用 $L = \prod_{j=1}^{n} f(t_j)^{\delta_j} S(L_j)^{1-\delta_j}$ 可得似然函数

$$L(u,b) = \prod_{i=1}^{n}\left[\frac{1}{b}\exp\left(\frac{x_i-u}{b} - e^{(x_i-u)/b}\right)\right]^{\delta_i}\left[\exp\left(-e^{(\eta_i-u)/b}\right)\right]^{1-\delta_i}.$$

令 $r = \sum\delta$，它表示观察到的寿命时间的样本个

数, D 表示 $\delta_i = 1$ 的样本(即寿命时间没有被截尾的样本) 组成的集合, 我们有:

$$\log L(u, b) = -r\log b + \sum_{i\in D}\frac{x_i - u}{b} - \sum_{i=1}^{n}\exp\left(\frac{x_i - u}{b}\right).$$

注意到 $\log L(u, b)$ 与 Ⅱ 型截尾场合有完全相同的形式, 这是由于在后者中的 x_r 对 $n-r$ 个没有观察到寿命时间的样本是表示截尾时间。因此极大似然方程可如下写出(要求 $r > 0$)

$$\sum_{i=1}^{n}x_i\exp\left(\frac{x_i}{b}\right)\Big\backslash\sum_{i=1}^{n}\exp\left(\frac{x_i}{b}\right) - b - \frac{1}{r}\sum_{i\in D}x_i = 0$$

$$e^{\hat{u}} = \left(\frac{1}{r}\sum_{i=1}^{n}e^{\frac{x_i}{b}}\right)^{b}$$

以上方程可用迭化法解出 b, 然后算出 $\hat{u}$。从而通过 $b = \beta^{-1}$ 和 $u = -\log\lambda$得到β和λ。具体估计结果见表1。

表 **1**　**3** 种癌症寿命分布模型中的参数估计结果

参数	肺癌	胃癌	肝癌
β	7.8354	7.1236	6.0065
λ	3.6960E-005	3.6016E-005	3.9825E-005

(三) 随机截尾型的拟合优度检验

当存在截尾观测值时, χ^2 拟合优度检验是不适用的。本文利用 Hollander 和 Proschan 所提出的检验方法进行拟合优度检验, 简称 H-P 检验。

设 $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$ 是由不同的有次序的生存时间组成的一个集合且 $t_{(i)}$ 的有些值是截尾的。若截尾的观测值与非截尾的观测值相等, 就把截尾观测值看作大于与之相等的非截尾观测值。设 $S(t)$ 是要考虑的生存函数, 而 $S_0(t)$ 是Weibull分布的生存函数。于是零假设为 H_0: $S(t) = S_0(t)$。

利用 Kaplan-Meier 乘积限方法, $S(t)$ 估计为

$$\hat{S}(t) = \prod_{\substack{t_{(i)}\leqslant t\\ t_{(i)}\text{非截尾的}}}\frac{n-i}{n-i+1}\quad(t_{(i)} < t \leqslant t_{(i+1)}).$$

对于数据遵从生存函数 $S_0(t)$ 的零假设的 H-P 的检验统计量是

$$C = \sum_{\text{全体非删截尾观测值}} S_0(t_{(i)})f(t_{(i)}).$$

这里 $f(t_{(i)})$ 是在非截尾的观测值处和在最大的非截尾或截尾的观测值处 Kaplan-Meier 估计的跳跃

$$f(t_{(i)}) = \frac{1}{n}\prod_{j=1}^{i-1}\left(\frac{n-j+1}{n-j}\right)^{1-\delta_{(j)}}.$$

这里当 $t_{(i)}$ 是非截尾时, $\delta_{(i)} = 1$, 当 $t_{(i)}$ 是截尾时, $\delta_{(i)} = 0$, 在零假设下,

$$C^* = \sqrt{n}\left(C - \frac{1}{2}\right)/\sigma.$$

近似地服从标准正态分布, 这里 σ 是 C 的标准偏差的估计值且

$$\sigma^2 = \frac{1}{16}\sum_{i=1}^{n}\frac{n}{n-i+1}[S_0^4(t_{(i-1)}) - S_0^4(t_{(i)})].$$

对检验 H_0: $S = S_0$ ←H_1: $S \neq S_0$, 若 $C^* > Z_{\alpha/2}$ 或 $C^* < -Z_{\alpha/2}$, 则拒绝 H_0。

这里 Z_α 是标准正态分布的 α 百分率上分位点。

依据随机截尾数据可以算得: 对于肺癌, $C^* = -0.2995$; 对于胃癌, $C^* = -0.2087$; 对于肝癌, $C^* = -0.1675$; 对于给定 $\alpha = 0.05$, 以上结果均不能拒绝零假设, 表明使用 Weibull 分布来描述三类癌症患者的寿命分布是合理的。于是依据所获得的三类癌症的寿命分布就可得到其在各年龄段的死亡概率分布, 见表 2。结合已有的医学研究成果可知, 一个人在被诊断患肺癌或肝癌或胃癌后, 该患者最长存活时间不足 3 年。从表 2 可见各年龄段的时间长度都远大于 3 年, 表明各年龄段的死亡率可以近似看作该年龄段的发病率。进而从表 2 的数据中可显示肺癌、胃癌、肝癌患者 45 岁以上者发病率分别达 98.89%、98% 和 93.41%; 60 岁以上者发病率分别达 84.6%、84.92% 和 67.2%, 表明老年是三类癌症的高发期; 中年期(45~ 59 岁) 肝癌发病率要远高于肺癌、胃癌的发病率, 在福寿年(75 岁以上) 胃癌发病率最高, 这一结论为下一步防病治病提供了科学依据。

表 **2**　三类癌症各年龄段的死亡概率分布

癌症	童年(岁)(0~ 6)	少年(岁)(7~ 17)	青年(岁)(18~ 29)	壮年(岁)(30~ 44)	中年(岁)(45~ 59)	老年(岁)(60~ 74)	福寿年(岁)(75 以上)
肺癌	2.79E-09	9.75E-06	6.30E-04	0.016	0.1373	0.4732	0.3728
胃癌	1.39E-08	2.31E-05	0.001	0.019	0.1309	0.4093	0.4398
肝癌	4.33E-07	2.25E-04	0.0053	0.0604	0.2621	0.4597	0.2123

参考文献

[1] Cho K. H, Song YB, Choi IS, et al. A phase Ⅱ study of single-agent gemcitabine as a second-line treatment in advanced non-small cell lung cancer[J]. Jpn J Clin Oncol, 2006 Jan, 36(1): 50- 4.

[2] Glasser, M. Regression analysis with dependent variable censored[J]. Biometrics, 1965(21): 300- 307.

[3] Kirsten D. and Marie D. Smooth inference for survival functions with arbitrarily Censored data[J]. Statist. Med. 2008(27): 5421- 5439.

国际统计学会第 57 届大会学术观点综述*

石 婷

国际统计学会(ISI)第 57 届大会于 2009 年 8 月 16~ 22 日在南非德班(DURBAN)召开。应大会组委会邀请,以中国统计学会常务副秘书长隋胜利为团长的中国统计学会代表团一行 6 人参加了会议。黑龙江省统计学会、吉林省统计学会、福建省统计学会、青海省统计学会等分别组团参会,中国统计界(含港澳台)共有 55 名代表参加了盛会。

一、会议概况

国际统计大会自成立 124 年以来,每两年举办一次,这是第一次在非洲召开,来自 130 多个国家(地区)的 2500 多名会员参加了此次大会,参会总人数达到 3000 以上。16 日下午,南非总统雅各布·祖玛(Jacob Zuma)、南非计划部部长 Trevor Manuel' 先生、主办城市 Ethekwini Municipality 市长 Obed Mlaba 先生、南非统计局局长 Pali Lehohla 先生、ISI 主席 Denise Lievesley 教授、ISI 执行董事会成员等出席了开幕式。祖玛总统在开幕式上做了长篇精彩演讲,主旨是统计应面对金融危机挑战,解决现实问题。

祖玛总统指出,世界正面临着许多挑战,包括全球性金融危机、食品安全、贫困、气候变化等,我们聚集于此,正是为了寻求通过合作、全球对话、以及更重要的协调行动来解决这些问题的方案。有效的统计数据和分析将对经济的恢复发生作用,宏观经济和社会信息比以往任何时候都更加重要。

他强调,统计对于社会发展十分重要,政府和决策机构在很大程度上依赖于官方统计数据进行判断和决策,但另一方面,反对党、各类压力集团和非政府组织也利用这些统计数据攻击政府,这样,由于统计数据常常会被质疑。统计工作者的工作就变得十分困难,这一问题可以通过增加与各方的交流和合作,使大家对统计数据和制定的政策都建立起信心而得到解决。在我们看来,任何事情也不能阻止政府统计工作者改进统计收集方法、统计编辑和统计

* 国际统计学会第 57 届大会中国统计学会代表团组:隋胜利、张 、王有捐、石婷、熊友达、孙学光。

[4] Maupas, P., Melniek, J. L. Hepatitis B infection and primary liver cancer[J]. Prog. Med. Virol., 1997(27): 1- 5.
[5] Prentice, R. L. Exponential survival with censoring and explanatory variables[J]. Biometrics, 1973(60): 279- 288.
[6] Tiku, M. L. Testing the two parameter Weibull distribution[J]. Stat. A., 1981(57): 207- 201.
[7] Vauhkonen M, Vauhkonen H, Sipponen P.. Pathology and molecular biology of gastric cancer[J]. Best Pract Res Clin Gastroenterol, 2006, 20(4): 651- 674.
[8] 赵力,张志良,陈毅琳.南京市玄武区肺癌死亡率的趋势分析和预测[J].中国公共卫生,2002,18(5):591- 592.
[9] 金华.基于平均生存时间两种分类方法的比较[J].统计研究,2008,25(1):98- 103.
[10] 尹桂成,茅亚达,刘明,等.2828 例不同病程恶性肿瘤患者生存分析[J].现代预防医学,2001,28(3):318- 320.
[11] 顾海雁,黄萍萍,李申生,等.1995- 2004 年上海市徐汇区居民恶性肿瘤减寿分析[J].上海预防医学,2006,18(4):161- 164.
[12] 华召来,郭国平,周琴,等.扬中市主要恶性肿瘤发病率及生存率分析[J].中国肿瘤,2006,15(11):744- 746.
[13] 李连弟,鲁凤珠.中国恶性肿瘤死亡率 20 年变化趋势和预测分析.中华肿瘤杂志,1997,19(3):15- 9.
[14] 贾安华.贵阳市 1973- 1998 恶性肿瘤死亡变动分析及对策.中国肿瘤,2001,10(9):506.
[15] 魏矿荣.广东省中山市 30 年全死因与恶性肿瘤死因分析.实用预防医学,2002,9(3):193- 5.
[16] 杜其药.成都市 20 年恶性肿瘤死亡动态分析.现代预防医学,2002,29(2).

作者简介

张志强,女,1964 年生,2004 年毕业于南开大学获概率论与数理统计专业博士学位,现为厦门大学副教授,研究方向为统计与精算。

马骅,男,25 岁,四川成都人,厦门大学数学科学学院概率论与数理统计专业,统计与精算方向 2007 级硕士研究生(在读)。

王浩丹,女,24 岁,湖南邵阳人,厦门大学数学科学学院概率论与数理统计专业,统计与精算方向 2007 级硕士研究生(在读)。

(责任编辑:程 周晶)