学术期刊论文生存被引次数的定义与应用

张中文 徐天和† 董秀芬 高 永
滨州医学院《中国医院统计》编辑部,264003,山东烟台

摘 要 通过对《编辑学报》和《中国科技期刊研究》2000 年第 1 期发表的论文被引情况的统计分析,总结论文被引的规律性。将生存分析方法引入论文被引次数的研究,提出学术期刊论文生存被引次数的概念,并给出论文"寿终""复活""复活率"的概念。采用 Kaplan-Meier 法对生存被引次数的生存率进行估计。结合实例对所提出的生存被引次数的合理性、应用前景及存在的问题进行了讨论。

关键词 学术期刊;生存分析;生存被引次数;Kaplan-Meier 估计

Definition and application of survival citation frequency of academic journals // ZHANG Zhongwen, XU Tianhe, DONG Xiufen, GAO Yong

Abstract This paper summarizes the rule of paper´s citation frequency through analyzing the papers published in the first issue of Acta Editologica and Chinese Journal of Scientific and Technical Periodicals in 2000. From the perspective of survival analysis, this paper proposes survival citation frequency of academic journal and puts forward the definitions of death, resurrection, revival rate of papers, and estimates survival function of survival citation frequency using Kaplan-Meier method. At last, we discuss the rationality, application prospect and the problems of survival citation frequency.

Keywords academic journal; survival analysis; survival citation frequency; Kaplan-Meier estimator

Author's address Editorial Department of Chinese Journal of Hospital Statistics, Binzhou Medical University, 264003, Yantai, China

期刊论文是科研人员发表研究成果的主要方式,因而其发表的论文的数量和质量就成为衡量该科研人员研究能力、学术水平和贡献的主要指标[1]。当前,评价期刊论文水平最常用的指标是期刊影响因子,但众所周知,期刊影响因子反映的是期刊所刊载论文的平均水平。与之相比,在特定的来源期刊群中,用单篇论文的被引次数作为论文的学术质量的评价标准显然更为合理;然而,论文被引往往在论文发表一段时间后才开始进入被引的高峰期,一些高水平论文被引可以持续很多年,论文在未来还有可能被引用,更重要的是不同论文被引次数的时间分布存在较大差异。单纯以某个时间段论文的被引次数反映论文的学术质量的说服力有限;因而,通过某种方法了解一篇学术期刊论文的"最终被引次数"就变得非常有意义。

本研究从生存分析的角度,提出学术期刊论文生存被引次数的概念,用以描述学术期刊论文的"最终被引次数",进而为学术期刊、科研机构、科研人员的评价工作服务。

1 学术期刊论文的生存被引次数的概念

生存分析方法是当前统计学研究的热点,生存分析是对 1 个或多个非负随机变量进行统计分析,即根据观测到的数据对 1 个或者多个非负随机变量进行统计推断。非负随机变量常用来表示自然界、人类社会或技术过程中某种状态的持续时间[2]。这里的"时间"是个抽象的概念,在实际应用中,人们除了可以用它表示时间以外,还可以用它表示一个轮胎可以使用的里程,也可以表示一个打印机硒鼓可以打印多少张纸等。生存分析的一个显著的特点也是它的最大优势就在于它可以处理所谓的删失数据。删失是指在进行观测或调查时,一个个体的确切寿命无法获得,而只知道它的一个范围。其中,最常见的删失为右删失,即只知道个体寿命大于某个常数。

通过分析大量学术论文被引次数的分布,可以发现学术论文被引是存在统计规律性的。一部分论文没有被引或者被引次数很少以后就不再被引用了,很大一部分论文在发表多年以后,尤其是已经有几年没有被引用以后,被再次引用的可能性也变得很小;与此同时,还存在这样一批论文:尽管已经发表多年,近年来却仍然被数次引用。对于那些已经很少被引用的论文,我们可以用它现在的被引次数来近似表示它的"最终被引次数",而对于那些仍在继续被引用的论文,则需要通过统计推断方法去了解它们的统计规律性。

按照生存分析的观点,如果我们将一篇论文不再被引用称为"寿终"的话,学术期刊论文的"最终被引次数"也就是它"寿终"时的被引次数。对于那些仍然被引用的论文,我们则认为学术期刊论文的"最终被引次数"删失了。

不同于一般的生存分析问题,一篇学术论文的"寿终"是不容易定义的;因为即使发表了很多年一次也没有被引用的论文并不排除在以后有可能被引用,即本来我们认为已经"寿终"的论文还是有可能"复活"的。

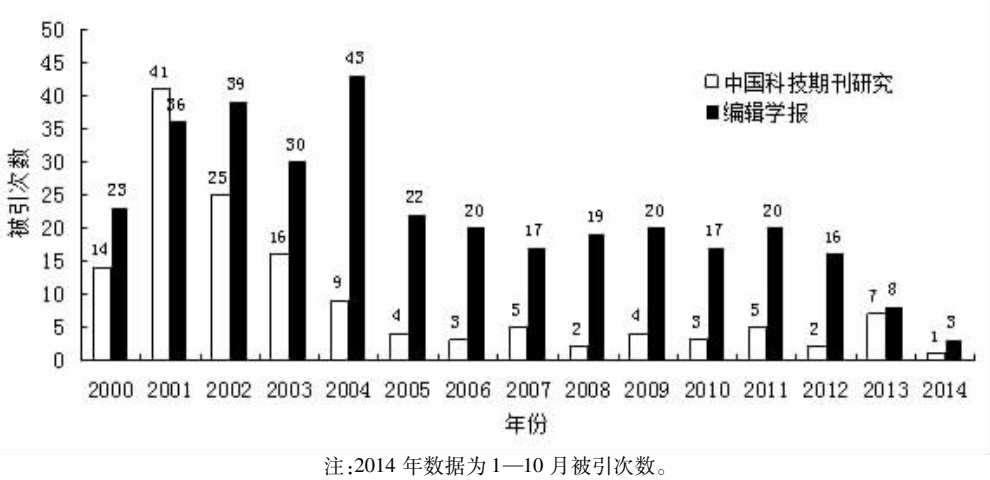
† 通信作者

在此,我们提出一个学术期刊“寿终”论文复活率的概念,即按照我们定义的“寿终”标准,已经“寿终”的论文复活的概率,在实际应用中,可以用已经“寿终”的论文复活的总次数与已经寿终论文的寿终总次数之比来近似表示。这里之所以采用寿终总次数,而不是论文的总篇数,是考虑到有的论文可能多次“寿终”。按照某个学术期刊论文“寿终”的标准,对于复活率不超过某个定值的论文(本文采用 10%),我们称其“寿终”时的被引次数为生存被引次数。

2　应用实例

2.1　资料来源与方法　本文数据来自 CNKI 的中国学术期刊网络出版总库和中国引文数据库,检索《编辑学报》和《中国科技期刊研究》2000 年第 1 期发表的论文,排除报道、说明等非学术论文内容,共得到《编辑学报》论文 29 篇,《中国科技期刊研究》论文 30 篇。检索所有论文从发表至 2014 年 10 月的被引次数与被引年份,所有数据均录入 2 次,并进行比对,对于不一致的数据,重新查找数据源,予以更正。原始数据的录入与整理在 Excel 中进行。生存被引次数的分析采用 SPSS 13.0 统计软件。

2.2　论文被引次数的年度分布　《编辑学报》和《中国科技期刊研究》2000 年第 1 期发表的论文,其被引次数的年度分布见图 1。在 CNKI 中按发表日期排序的《编辑学报》2000 年第 1 期前 5 篇论文的被引次数年度分布情况见图 2。



注:2014 年数据为 1—10 月被引次数。

图 1　论文被引次数的年度分布

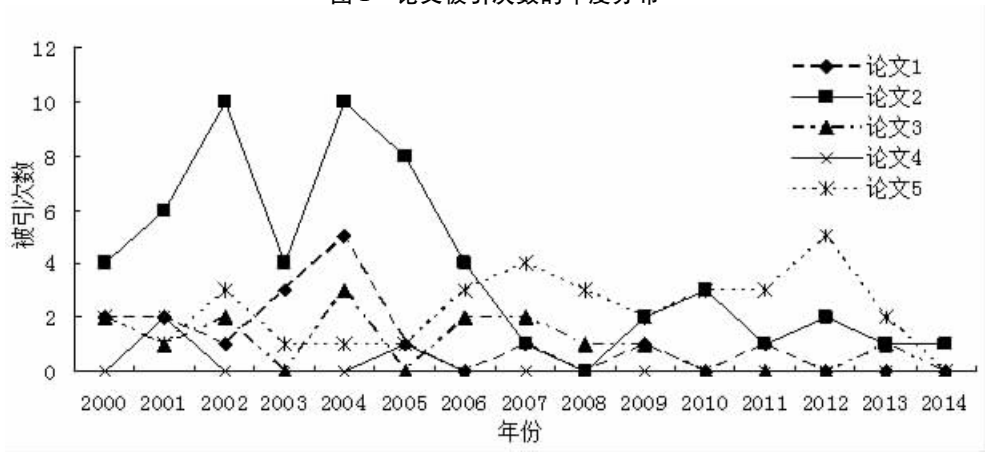


图 2　《编辑学报》2000 年第 1 期 5 篇论文的被引次数年度分布

《编辑学报》和《中国科技期刊研究》2000 年第 1 期发表的论文都在发表后前几年获得了较多的引用,之后被引次数有不同程度的减少并大概维持在一个较为稳定的次数,表现出了明显的规律性。而不同论文的单篇被引次数年度分布差别较大,基本没有表现出统一的规律,同一篇论文不同年份的年度被引次数展现出很大的偶然性。

2.3　论文的“寿终”情况　本文将一篇学术论文超过 4 年不再被引用定义为“寿终”,“寿终”论文的被引次数即为生存被引次数,尚未“寿终”的论文则可以认为生存被引次数删失了。论文“寿终”后如果再次出现被引称为“复活”。《编辑学报》“寿终”的论文出现“复活”2 次,累积寿终 30 次,平均每次“复活”后再次被引 1 次;《中国科技期刊研究》“寿终”的论文出现

“复活”2次,累积“寿终”50次,平均每次“复活”后再次被引2次,“复活”出现次数较少,而且“复活”后再次被引次数不大,总体来看对整体的生存函数影响不大:故可以认为我们定义的“寿终”标准是合理的。

“寿终”次数=[论文连续不被引时间/4],“[ ]”为取整符号,如1篇论文连续0~<4年不被引认定为“寿终”0次,连续4~<8年不被引认定为“寿终”1次,连续8~<12年不被引认定为“寿终”2次,连续12~<16年不被引认定为“寿终”3次,依此类推。《编辑学报》和《中国科技期刊研究》2000年第1期发表论文的“寿终”与“复活”情况见表1。

表1 论文的“寿终”与“复活”情况

期刊名称	论文篇数	“寿终”论文数	删失率/%	复活率/%
编辑学报	29	12	58.62	4.00
中国科技期刊研究	30	20	33.33	6.67

2.4 被引次数生存函数的Kaplan-Meier估计 生存分析不是孤立地研究某篇论文的生存被引次数,而是研究一批论文的生存被引次数。单独一篇论文的生存被引次数存在很大的偶然性,而一批论文的生存被引次数就有一定的规律性。我们用T表示任何个体的生存被引次数,把T看成随机变量,T的值依赖于个体。在了解了T的详细信息以后,进一步可以对个体情况进行推断。仿照一般的生存问题分析,我们用生存函数来刻画T的分布情况。其中生存函数$S(t)=P(T>t)$。采用Kaplan-Meier法对《编辑学报》和《中国科技期刊研究》2000年第1期发表的论文进行估计,计算在SPSS统计软件中进行[3]。经Log Rank(Mantel-Cox)检验得到2种期刊生存被引次数的差别具有统计学意义($\chi^2=5.194$,$P=0.023<0.05$),2种期刊论文生存被引次数的生存曲线见图3。

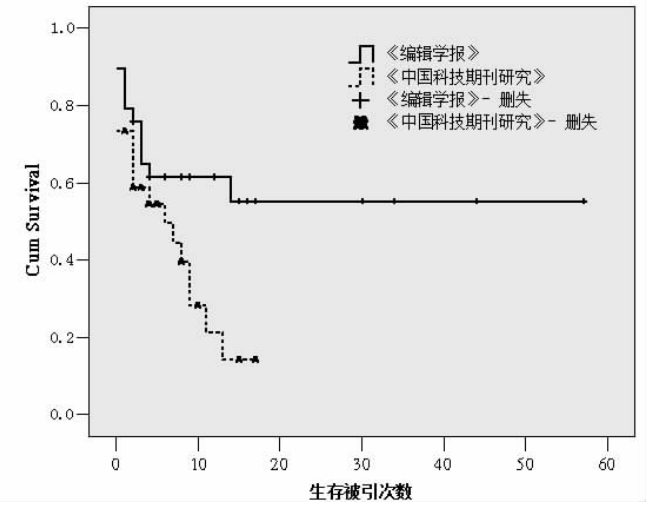


图3 论文生存曲线

由图3和寿命表(由于篇幅受限从略,计算过程中直接由软件得到)可知:被引超过4次的论文,《编辑学报》的论文生存率明显高于《中国科技期刊研究》,即论文被引次数超过4次以后,《编辑学报》论文再次被引的可能性明显高于《中国科技期刊研究》;而且随着被引次数的增加,生存率的差距越来越大。结合论文的年度分布我们发现,尽管论文发表前几年,2种期刊论文的被引次数差别是不大的,但从2004年开始,2种期刊被引次数的差距变的非常明显。这都表明,相比于《中国科技期刊研究》,《编辑学报》发表的论文在获得较高引用以后,更倾向于获得更高的引用,还表明《编辑学报》论文的生存期明显较长。

3 讨论

3.1 生存被引次数用于论文评价的合理性 同一期刊不同论文被引次数的较大差距提示我们,应该用单篇论文被引次数取代影响因子来评价论文的学术质量。由《编辑学报》2000年第1期5篇论文的被引次数年度分布可见,单篇论文的年度分布具有很大的随机性,缺乏明显的趋势与规律,而一篇论文不同年份的学术质量评价结果不应该有太剧烈的变化;所以,采用年度被引次数评价论文的学术质量是不合适的,更不具备深入研究的可能性。与此相对应的,论文生存被引次数反映的是论文的“最终被引次数”,是唯一的、稳定的表示论文被引情况的指标,适用于对论文学术质量的评价。

3.2 生存被引次数的应用前景 由本文结果可以发现,生存被引次数既可以描述单组文献的被引情况,也可用于比较2组文献的最终被引情况,类似地,也可以用于多组文献最终被引情况的比较。这种比较提供的信息非常丰富,应用寿命表以及生存曲线对应于不同的被引次数,都可以给出生存率的值。同时,生存被引次数还可以用于估计高被引论文继续被引的概率。

生存被引次数除了可以用于文献评价外,更广阔的应用前景在于它的分析功能与预测功能;因为生存被引次数的稳定性,研究者可以采用考虑右删失的回归分析法去研究它的影响因素,掌握哪些指标是影响论文被引次数的显著性因素,进而可以通过已知的指标对文献的最终被引情况做出预测,从而辅助论文的审理以及学术评价[4]。当然,具体采用什么样的回归模型有待于进一步研究。

3.3 研究中存在的问题 本研究把生存分析方法引入文献研究,提出了生存被引次数的概念,用以近似表达论文的最终被引次数。生存被引次数有可能出现的“复活”现象,这将导致我们定义的生存被引次数会低估最终被引次数;因而在应用生存被引次数评价论文