事件史分析及其应用

SHIJIANSHI FENXI JIQI YINGYONG

杜本峰/著

经济科学出版社
Economic Science Press

事件史分析及其应用

杜本峰　著

经济科学出版社

图书在版编目（CIP）数据

事件史分析及其应用/杜本峰著．—北京：经济科学出版社，2008.5

ISBN 978-7-5058-7218-9

Ⅰ．事… Ⅱ．杜… Ⅲ．统计分析-研究 Ⅳ．C812

中国版本图书馆 CIP 数据核字（2008）第 071591 号

责任编辑：杜　鹏
责任校对：徐领柱
版式设计：代小卫
技术编辑：董永亭

事件史分析及其应用
杜本峰　著
经济科学出版社出版、发行　新华书店经销
社址：北京市海淀区阜成路甲28号　邮编：100036
总编室电话：88191217　发行部电话：88191540
网址：www.esp.com.cn
电子邮件：esp@esp.com.cn
北京汉德鼎印刷厂印刷
永明装订厂装订
880×1230　32开　13.25印张　370000字
2008年5月第1版　2008年5月第1次印刷
ISBN 978-7-5058-7218-9/F·6469　定价：26.00元
（图书出现印装问题，本社负责调换）

前　言

我们处在信息时代，信息时代的基本特征是充满着数据。无论是科学研究工作者或是管理者，每天都需要处理大量的数据。为了能够从数据中探索和发现规律，以便从杂乱无章的数据中得到有助于科学决策的结论并指导实践，具备数据分析和掌握信息的能力显得尤为重要，它是进行科学分析和决策的基础。我们知道，不同的数据其要求的统计分析方法不同，有些数据只能进行分类和构建频数，多使用列联表分析方法；有些数据则可以通过建立统计模型分析。横截面数据、时间序列数据、纵向数据等都要求不同的统计分析方法。横向研究是在一个时间点上收集研究资料，并用于描述调查对象在这一时间点上的状况，或探讨这一时间点上不同变量之间的关系，它通常用于探索性研究和描述性研究，也可用于解释性研究。但横向研究用于解释性研究总是存在一定的局限性，这是因为，原因因素与结果因素在时间上是有先后之别的，要对一个比较长的时间段所发生的现象进行考察，才有可能发现现象之间的因果关系。纵向研究也称纵贯研究，指的是在比较长的时期内的若干个不同时间点收集资料，用于描述观测现象的发展过程，解释现象之间的相互关系，分析观测现象产生的历史背景和社会条件，探讨现象的前后联系，力图揭示观测现象的发展规律和发展趋势。因此，要在数据中探索和发现规律，数据与正确统计分析方法的匹配不可忽略，否则，要么数据的信息未能充分利用，要么由于分析方法使用不当而得到错误的结论。同时，即使相同的数据类型，由于研究问题的不同也需要不同的分析方法，如纵向数据研究事件发生这一问题的方法，使用的就是事件史分析

(event history analysis)[①]。

事件的发生和变化是许多领域所关心的核心问题。如何分析影响人们晋升的因素？这里，晋升即为事件，是指发生在某个时点的某种质的变化，而持续时间表示了在这一观测期内的个体在未晋升这一状态的时间长度，也表示了晋升这一事件发生的时间。我们的目标是，考察所研究的对象中，获得晋升的可能性在多大程度上依赖于文化程度、年龄、性别等因素。在一般的横截面数据分析中，我们多采用如下的分析模型：因变量是“是否获得晋升”，自变量则是年龄、文化程度、收入、党派、原职务级别、性别等。这里对自变量的取值或许有人会武断地取观测期末时的观察值，仅仅因为这是观察期的结束；或许有人会出于某种理论假设，认为应该以观测期中的某年为一个划分点，比较在此以前获得晋升的与以后被晋升之间存在的差异。然而，不管怎样，这些取值方案都会浪费大量的信息，因为这些方案都忽略了取值时点前后的变化情况，即它们在观测期间的取值是在不断变化的，这就涉及如何处理“时变变量”的问题。我们同时也会怀疑这种笼统的分析忽略了第一年就获得晋升的学员与最近刚获得晋升的学员之间存在的较大差异。更何况我们在某一时点测得的解释变量（例如收入）可能不仅不是“能否获得晋升”的原因，反而可能是升职后的结果。

为了避免这个问题，我们也许会构造出另一个模型，即把“从毕业到获得晋升的间隔时间”作为因变量，而自变量与上面的模型相同。然而，在对这个模型进行估计时，那些未获得晋升的学员的案例就不能被纳入这一模型，因此，有关这部分人的信息就被浪费了。如果这部分人在样本中所占比例较大，或者当获得晋升的人与未获得晋升的人在某一方面存在系统性偏差的情况下，对样本的估计就会出现较大的偏差。

这里的“事件（event）”代表着一种变化或从一种状态到另一

① 也称为失效时间分析（failure time analysis）或生存分析、风险建模（hazard modeling）统计方法。

种状态的转变，如跨省迁徙、前面例子中的获得晋升等变动；“事件史（event history）”即从对研究对象的观测到事件发生或观测期结束的历史或生存（持续）时间，可以说是样本所经历的一系列事件所构成的纵向性数据。对事件史的理解不仅要考虑是否发生了某种事情，而且要考虑是何时发生这一现象的，不仅关注它们在一个状态持续的时间长短，而且关注从一个状态到另一个状态发生的概率大小或风险；对持续时间长短和事件的发生受一些因素的影响更是表现出了极大的兴趣。这里的“生存时间”一词应作广义的理解，乃指自然界、人类社会或技术过程中某种状态的持续时间。根据 Yamaguchi（1991）的定义，事件史分析专门研究“事件发生的方式及其相关因素”。它是分析纵向性数据统计分析方法的集合。简单地说，就是研究个人特征变量、环境变量或制度性变量在变化的时空中是如何影响一些事件发生的概率。

虽然纵向数据分析这种方法最初来自于生物统计学家，但这种方法和理论的应用不仅仅限于生物学和医学领域，如经济学家在工作和就业的研究中，注意的焦点则是工作变换、晋升、解雇、失业和就业之间的状态变动以及何时发生；管理科学研究中，客户从一种品牌到另一种品牌的转换或重复购买相同品牌的概率，公司、团体或组织的成立和关闭状态间变动以及风险分析，工人在不同职业、行业和社会阶层持续时间及其状态间的变化和影响因素；在政治科学研究中，政府的执政持续时间、自愿者组织的成立或国家政治制度的转变分析，罪犯释放后再次犯罪的可能性及影响因素分析；人口学家所关注的出生、死亡、结婚、离婚和迁移等事件的分析；在医学和公共卫生的应用中，病人在健康和疾病间的转换、死亡、缓解及疾病发作时间等研究。这些关于持续时间（生存时间）的评估和预测以及从一种状态到另一种状态发生的概率大小和风险是这一方法和理论研究的重要内容。

近二十年来，各类研究人员利用事件史建模这一统计分析技术与日俱增，这既不是偶然的趋势，也不是在追赶调查研究和统计分析的时尚，相反，它表明了科学工作者的共同认识：事件史分析是

人们能够利用纵向数据了解所研究现实过程最合适的统计分析技术，在解决有删失数据和时变变量方面具有独到的优越性。

在我国现有的统计分析著作中，很少有讨论纵向数据分析的统计方法，而纵向数据在我国越来越多。即使个别数据分析著作中有事件史分析的内容，也只是一个章节，不是作为中心内容，缺乏关于这种方法的详尽讨论，国内尚未有此类的专著。随着我国科学研究中纵向数据的使用越来越多，提供解决这一问题方法手段的需求也越来越迫切。本书专注于事件史分析这一方法本身的讨论以及模型结果的详细解释，强调纵向数据分析中事件史分析这一方法的应用。

本书是国家统计局2006年重点研究项目2006B03的研究成果，同时也是国家自然科学基金（70773116）的阶段性成果之一。

希望本书能在推动我国社会科学研究方法的发展、提高不同学科领域研究人员应用量化方法的能力尤其是纵向数据分析方面提供微薄之力。作者才疏学浅，恳切得到读者的批评指正。

杜本峰

2008年3月于世纪城时雨园

目　录

第 1 章

引　论

许多人发誓戒烟后永不再抽，一些人成功了，而许多人故态复萌又开始吸烟。人们关注的是：什么时间最易复发？是刚刚戒烟后身体反应最强烈的阶段，还是几周后？当某种信念或社会支持消失后，谁又最可能复发？像这些关于事件发生及时机的问题贯穿于整个社会和行为科学。例如，犯罪学家关注的是累犯、研究犯罪、逮捕、牺牲等事件；经济与管理学家在工作和就业的研究中，注意的焦点则是工作变换、晋升、解雇和退休等；政治学家则关注暴乱、革命、政府的和平更迭等事件；人口学家则关注出生、死亡、结婚、离婚和迁徙等事件；医学家则关心病人访问医生的次数、病人的持续时间及药物治疗对病人持续时间的影响；等等。

然而，尽管这些问题无处不在，但一般的统计学并没有讨论处理这些问题的方法。要想研究事件发生这类问题，我们必须要用新的方式来思考这些数据，我们不仅关注研究对象在一个状态持续的时间长短，而且关注是否从一个状态转换到另一个状态及何时转换，关注从一个状态到另一个状态发生的概率大小或风险，并对持续时间长短和影响事件发生的因素更是表现出了极大的兴趣。

本章我们将讨论描述事件史分析这一问题的核心概念和术语、基本数据形态、基本特性与目的等内容。

1.1　事件史分析

根据 Yamaguchi（1991）的定义，事件史分析专门研究“事件

发生的方式及其相关因素”。事件史分析方法，是运用离散状态（discrete state）、连续时间（continuous time）的随机模型，来分析纵贯性数据的统计分析方法的集合（Mayer et al.，1990）。

这里的时间是指从观测个体开始到事件发生为止所持续的时间，可以是年、月、周或天等，或者是指当事件发生时观测个体的年龄。目标事件可能发生一次，如高中毕业或第一次生育，也可能在观测期重复发生，如开始工作、停止工作及买房、卖房等，事件也可能是个人可以控制或个人不能控制的，如流产、着火等。

由于这种数据的分析方法最早应用于人类生命的研究中，感兴趣的事件是死亡，因此，这一统计术语常常寓意着一种不祥之兆，这导致了许多不正确的认识，即该方法只有在研究消极事件时是适用的，如疾病、累犯、离婚和吸毒等，然而，该方法对于积极事件（如结婚、毕业、生育）和中性事件（买车等）同样是有效的。

为了确定一个问题是否需要使用事件史分析，其最简单的判断是“是否和何时检验”。当我们研究的兴趣是事件是否发生或何时发生时，或许需要使用事件史分析方法。

何谓事件？这是我们必须首先定义的问题。一个事件代表一种变化或从一种状态到另一种状态的转变，也就是从初始状态到目的状态的转变。我们之所以使用“状态”这一术语，是因为它可以应用到多个学科。例如，一名新录用的教师从事教学（状态 1）直到他或她离开教学岗位或学校（状态 2）；一名没有喝酒（状态 1）的人到开始饮酒（状态 2）；在婚姻史的研究中，一个可能的事件是“初婚”，可以定义为从初始状态（即未婚）到目的状态（结婚）的转化。其他可能的事件是：离婚、成为寡居者、再婚。状态的划分对于定义可能的事件是非常重要的。如果仅有已婚和未婚两种状态，就不可能定义上面的事件，在这种情况下，唯一可以定义的事件是结婚和婚姻的解除。有时我们也把事件称为失效。

因此，事件史分析的第一步是定义人们想区别的状态。状态是我们要描述的一类因变量，它是我们想解释的一种动态变化，在每

个时点上，每个人占据一个状态。在多数情况下，观测个体可能仅经历两个可能的状态，即工作与失业；但在有些情况下，观测个体可能经历三个或更多的状态，例如，在婚姻史的分析中，一般有四个状态，即未婚、已婚、已离婚和寡居，这些可能的状态集有时又称为状态空间。

大多数事件史模型考虑的是单一事件、两状态（一个起始状态，一个终止状态）过程。例如，假如我们对女性初育时机的差别感兴趣，在这种情况下，事件就是生育第一个孩子，这一事件可定义为从无孩子初始状态到有孩子这一目的状态的转变。这种情况称为单一无重复事件，这里的“单一”这个术语表示从没有孩子这一初始状态开始观测，仅有一种事件发生；术语“无重复”是指事件只能发生一次。如果终止状态不止一个，我们称这些模型为多状态模型。拥有一个初始状态而有两个或两个以上终止状态的模型又称为竞争事件或竞争风险问题。例如，一个家庭主妇可能变成“失业的”（意指进入“找工作”的状态），或者开始成为“全职”和“兼职雇用”。图 1－1 和图 1－2 描述了多状态多事件的过程，即观测个体在几个不同的状态之间重复地变动。

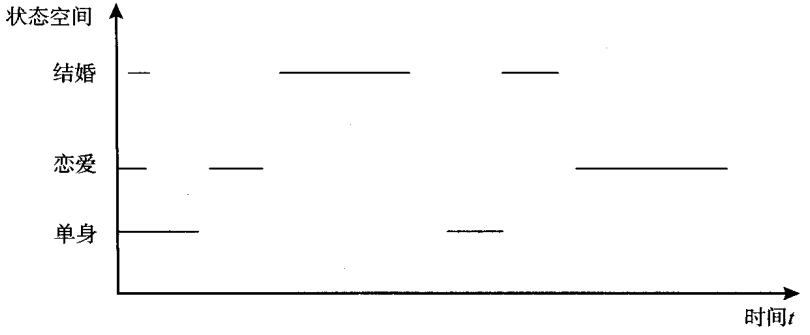

图 1－1 多状态多事件过程

在图 1－2 中，A 显示妇女可处于两个状态：一是单身母亲；二是非单身母亲。于是，所有妇女从状态 1（非单身母亲）开始，

一些妇女转换到状态 2（单身母亲），成为单身母亲的这些人也可能退出这个状态，返回到状态 1，等等。B 代表了另一种情景，妇女从未婚和没有孩子开始了她的生命历程（此时状态标记为 10），可以转换到后继的状态，此时，转换到单身母亲有几种可能性，如从未婚无孩子状态 10 转变到未婚初育状态 11，从已婚有孩子状态 21 转变到状态 11，从状态 22 转换到状态 12，以及从状态 23 到状态 13 的转变，都代表了未婚和已婚妇女之间的变化。

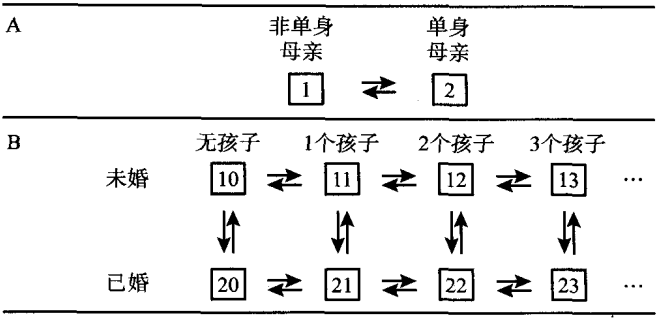

图 1－2　状态间的转换

开始时间是我们必须面对的又一个问题。开始时间是一个观测中的每个观测个体占据一个可能状态的瞬间。如教师受聘的那一天，意指他们将要从事教学工作。在研究观测期，当观测个体从最初状态转移到另一个状态时，他们经历了目标事件，从开始时间直到事件发生的距离，我们称之为事件时间。显然，在研究的每个时点上，并不是每个人都可能经历事件，为了经历一特定事件，人们必须处于所定义事件的初始状态，也就是说，必须处于事件发生的风险之中。处于该特定事件风险或暴露于特定风险的时期，称之为风险期。例如，考虑已婚可能经历一次离婚的人群，此时，只有结婚的人处于离婚的风险中。此外，离婚的风险期是人们处于结婚的这一时期，与此有很大联系的另一个概念是风险集。处于一定时点上的风险集是由所有这样的观测对象所组成的，它们在这个时点上

处于经历事件的风险之中。

事件史分析就是在风险期内没有发生事件的持续时间分析。当我们所感兴趣的事件是“初婚”时，分析者所关心的是初婚没有发生的持续时间。换句话说，也就是观测对象仍处于未婚状态的时间。在实际应用中，正如我们下面将要说明的，事件史模型中的因变量通常不是持续时间而是转换率，所以，事件史分析也可定义为风险期内事件发生概率的分析。如在初婚的例子中，事件史模型关心的是处于未婚状态时期中人们结婚的概率。

事件史分析研究状态间的转换，以及进入和离开特定状态的时间间隔长度。其基本分析框架是状态空间和时间坐标。分析中用到的时间坐标或计时（例如年龄、经验、婚姻存续时间等）的选择必须基于理论上的考虑以及对统计模型的影响。关于时间刻度没有一个普遍适用的准则，同一个事件不同的研究内容也可能使用不同的时间刻度。如关于工作变换这一事件的研究，一种可能是关注于议员的变换，另一种则可能是关注于教师工作的变换，第三种则可能是关注于商业人员工作的变换。政治科学在研究议员的变换时，其时间刻度或许选择2年、3年或5年，它与议员的任期时间有关；而教育方面对教师流动的研究，时间刻度通常是学年，因为大多数教师在学年底选择离开学校；对商业人员流动的研究，时间刻度通常选定为周、月。比较细、精确的时间单位我们称为连续时间，而相对宽的时间区间单位我们称为离散时间。

在本书中，一段持续时间、一段时间、等待时间或持续时间是分析对象（例如一个人）花在特定状态上的时间长度，它们表示的是同一含义，本书中经常互换使用。对于一组可能状态（称为状态空间 γ）的定义要考虑研究对象的本质问题。因此，仔细并基于理论的时间选择和设计状态空间是重要的，因为它们经常是错误描述的来源。尤其是，模型的错误描述可能会因为没有观察到一些重要状态而发生。例如，在分析德国妇女参与劳动市场决定因素的研究中，Blossfeld 和 Rohwer（1997）指出，如果将“被雇用”状

态区分为“全职工作”和“兼职工作”，就可以得出许多实质性结论。这里我们应该注意到所关心的实质性问题焦点的微小变化（可能导致状态空间的重新定义），常常要求对事件史数据进行重新认识。

1.2 删失

在我们前面的讨论中，无论是何时开始数据的收集，也无论观测期多么长，总有一些观测个体没有发生目标事件，有可能未知其事件时间，我们称这类问题为删失。由于删失的不可避免性，它就对我们一般的统计方法提出了挑战，如在删失存在的情况下，如何计算事件时间的平均长度或其他统计量？

简而言之，当我们有一些关于观测个体的持续时间信息而不知道确切的持续时间时，删失就发生了。通常有三种原因使得删失发生：（1）在研究结束之前，某些观测个体还没有经历我们所关心的事件；（2）在研究期间，观测个体丢失；（3）由于其他原因，观测个体退出研究。图 1－3 给出了不同删失类型的一些例子（参见 Yamaguchi，1991；Guo，1993）。横轴表示历史时间，起始和结束分别用 τ_a 和 τ_b 表示。

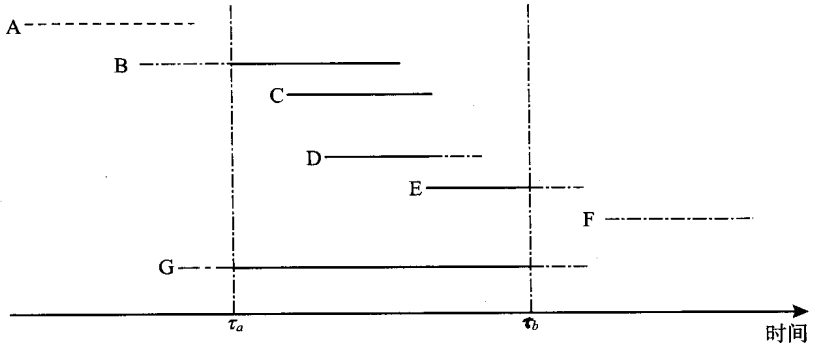

图 1－3 删失类型