学校编码：10384　分类号__密级__
学号：15420141151963　UDC___

厦门大学

硕 士 学 位 论 文

持续期及寿命模型的事件概率区间研究

A Study of Bounds on Event Probabilities for Duration and Lifetime Models

陈晓雪

指导教师姓名：沈雁 副教授
专　业 名 称：数量经济学
论文提交日期：2017 年 4 月
论文答辩时间：2017 年 4 月
学位授予日期：2017 年 6 月

答辩委员会主席：__
评阅人：__

2017 年 4 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下，独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外，该学位论文为（　　）课题（组）的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

目录

Contents

摘要

目前，有关于寿命期、生存持续期等方面数据的统计研究在很多领域都是一个很重要的课题，比如，生物医学、工程学以及社会科学。在近几十年中，持续期的研究在经济研究中也是一个逐渐快速增长的领域。目前关于持续期在经济领域的研究主要基于从工业管理以及生物科学领域发展出来的统计方法，并且关注点主要在于对持续期数据的分析以及模型设定上，然而在对未来的持续期以及生命周期的推断预测却较少地被研究。

为了能够对持续期以及生命周期有更加精准与更加系统的推断与预测，从方法论上出发，本论文关注对持续期和寿命模型建立事件概率的置信区间（confidence bounds on event probabilities, 即 CBEP），我们的目标是面对不同分布不同类型的数据，都能够使用一种统一化的方法来构建事件发生概率的置信区间，从而使我们放松对数据的要求。在文章中，我们将观测值进行了一种“正态化”转换，然后以此为出发点，不断进行多种尝试，并通过蒙特卡洛模拟试验，最后保留了三种效果较好的方法。通过文中提出的这几种方法与蒙特卡洛模拟结果，我们考察了这几种 CBEP 下的有限样本表现，综合总结表现的优点与不足，我们可以得到以下结论：（1）无论观测值是来自于何种分布，建立的 CBEP 大体上表现良好；（2）针对于某些特殊分布，如对数正态分布或 Weibull 分布，它们具有特定的置信区间的构建方法，而本文所建立的 CBEP 与之相比表现也比较优良。最后本文通过美国 1968-1976 年的罢工数据进行了实证检验，进一步验证了三种方法的良好表现。

关键词：Box-Cox 转换；持续期模型；寿命模型；生存概率

Abstract

Nowadays, the statistical analysis of what are variously referred to as lifetime or survival time data is an important topic in many areas, including the biomedical, engineering, and social sciences. In recent decades, the study of duration in the field of economic research is gradually becoming a rapidly growing area. Current researches on duration in the economic area are mainly based on statistical methods in industrial management and the developments in the field of biological sciences, and focus mainly on the analysis of duration data and the model specification. However, it has been studied little about the future duration and the prediction of life cycle.

In order to build a more accurate system to make inferences about the duration data and to make some predictions of life cycle, we start our paper from the methodological view. Our paper concerns the problems of constructing confidence bounds on event probabilities (CBEP) for duration and lifetime models. We want to propose a unified approach to construct the confidence bounds whatever the distributions of the data are, so that we can relax assumptions of the distributions. In the paper, we deal with the data through a “normalizing” transformation of the observations, and did a lot of try, which we have checked by Monte Carlo simulations. As a result, we selected three methods that perform much better. And then we investigate the finite sample performances of the proposed CBEPs. We find (i) it generally performs well no matter what distribution that the observations come from, and (ii) it is comparable with the CBEPs that are specifically designed for a particular distribution such as lognormal and Weibull. Finally, we use the strike data of America from 1968 to1976 to further test our methods and the results performed pretty well.

Keywords：Box-Cox transformation; Duration model; Lifetime model; Surviving probability

第一章 绪论

一、研究背景与探究意义

目前，对关于寿命期、生存期或持续期等方面数据的统计研究在很多领域都是一个很重要的课题，比如，生物医学、工程学以及社会科学。那么，首先要明确的一个问题就是：怎样理解定义寿命、持续期模型？简单地，我们从实际角度举例来讲，在社会科学领域中，对于一些工业大国，罢工可能是一种劳动者们表达自己要求的一种手段，因此罢工是一个非常常见的正常现象，而此时可能为了预期某国政府所颁布的某项新政策体制对未来发展的影响，我们则可能需要关心未来预计发生罢工的持续时间，那么就需要这样一组数据：该国家在某一个时间段内每一次罢工从开始一直到结束的持续时间长度，而这种测量时间长度的数据就是所说的持续期数据，很容易理解，顾名思义，即是持续时间长度的数据；另一方面，例如在医学领域，由于研究需要了解一种绝症类疾病对人类剩余寿命的影响，那么可能需要收集的数据则为已患病患者的从被确诊到最后死亡的这个剩余寿命长度，也可能是为了某项医学研究，观察试验小白鼠从出生到死亡的寿命长度，这些就是所谓的寿命数据。从本质上来看，持续期、寿命、以及在某些文章中提到的失效期等都属于对同一个事物的不同角度的描述，例如寿命数据也可以理解为，从患者患病开始，一直到死亡所持续的时间，那么它同样也是持续期数据，所以可以说是不同视角下所产生的不同的名称，归根结底可以看作是一种类型的数据，也就是我们所关心的事情的从开始到结束的这个时间长度。事实上，从数学的角度来讲，持续期及寿命数据可以简单地理解为“非负随机变量”，从这个方面来看，持续期与寿命模型就变成了非常熟悉、非常常见的在概率论与数理统计中所学的模型分布。对寿命分布方法论的研究应用可以从工业制造产品的持续期研究一直扩展到对于人类疾病及其治疗等方面的研究，而之所以要将持续期与寿命模型单列出来进行专门的讨论，则是因为其数据本身的特性所带来的丰富的现实意义。在社会研究中，持续期的研究可以包括婚姻持续期、政策影响力持续期；在工程研究中可以包括接纳新技术所需时间、公司寿命、产品持续期，在医学研究中，可以包括病情潜伏期等等。更具体地，例如，在生物医学上，通

过对持续期与寿命数据分析判断，可以预测一种疾病对人类的危害程度，并同样可以检验一种试验层面的治疗手段的有效程度，从而考虑是否可以在临床中广泛的应用；在工程学中，通过抽样检验预测某些零部件与制造产品的有效时间，从而对其失效期有合理的预测，及时制止其失效而造成的可能损失；更广泛地，在社会科学中，我们可以将持续期与寿命模型分析广泛的用于探究各种政策体制与行为策略所能产生的影响。

在发展初期，处理持续期及寿命模型数据的方法都是相对老旧的，大约在 1970 年以后，关于持续期及寿命模型研究的方法论才得到了快速的发展，从而进一步扩展了其研究应用领域；在 1980 年左右，寿命数据分析计算机软件包的推广使用使得对寿命持续期模型的研究分析进入了一个新的时代。而在近三十年中，持续期在经济研究领域中的研究运用才逐渐呈现出了快速发展的趋势。例如，Kiefer（1988）认为，失业持续时间长短比失业率是一个更加有意义的研究指标，故其将经济持续期中的失业持续期通过使用风险函数来进行了分析，提供了详尽的研究，其中，所谓风险函数，是生存函数的反向表述，即某个体（事项）已存活（持续）至某一时刻，在这一时刻即时死亡（停止）概率。实际上，持续期在经济领域也涉及很多方面的应用，除了上文所说的失业持续期，还可以包括宏观经济中的罢工持续期以及金融交易持续期等。从本质上来讲，由于持续期及寿命模型在经济领域发展起步相对较晚，目前关于持续期在经济领域中的研究主要是基于在工业管理以及生物科学领域发展出来的统计方法与研究思路，并且关注点主要在于对持续期寿命数据的模型设定上，同时对于数据的分析研究也十分依赖于所设定的模型分布类型。另外，从概率的角度上出发，在对未来的持续期以及生命周期推断预测的研究方面也相对较少。

二、研究内容

本篇论文想要探究的主要问题就是对持续期及寿命模型的事件发生概率建立一个统一化的方法体系来构造置信区间。其根本出发点就是欲立足于从概率角度出发对未来持续期以及生命周期的推断预测这个总体目标上。这里有两点需要作如下解释：首先，所谓持续期与寿命模型的事件发生概率，指的是，已知某个

体或事件的一系列持续期寿命数据，想要进一步了解的是，在未来的某一时刻，该个体仍然存活的概率有多大，或是该事件仍然持续的概率是多少。在实际应用中，举例来讲，我们可以去预测某一次罢工时间会超过一个所设定的临界值的概率大小，从而去估量未来罢工深入的程度，以及可能会对该产业造成影响的大小；第二点，本篇论文想要建立的是一个统一化的方法体系，在已有的对持续期寿命模型研究中，更多的是利用该历史数据进行分析并将其进行模型的设定与拟合，再进一步去对未来进行预测，而本文希望做到的则是从方法论的角度出发，可以放松对于数据的模型分布假设，即该数据无论属于何种分布的数据类型，都可以运用文章中所建立的统一化的方法体系得到一个较为理想的事件发生概率的置信区间，对于分布假设具有一定的稳健性。

在当前已有的研究中，存在着一些建立一个合理的事件概率置信区间的方法，但是这些方法都有一个限制，就是他们都是基于一些特定的分布类型，比如，lognormal 模型、Weibull 模型。同时，也有一些模型假设下，并没有一个可提供的方法来构建其事件概率置信区间。在这种情况下，就面临着一些实际的问题：第一，很多情况下对于给出的数据，究竟应该使用哪一种寿命分布模型我们经常是不明确的；第二，即使我们选择了一个正确的模型，这个模型的事件概率置信区间也有可能是不可获得的、不合适的、又可能是在实际情况下太复杂而难以实现的；第三，对于标准的非参数方法或者无分布方法可能并不能够给出合理的事件概率置信区间，因为这些方法通常要求一个不切实际的大样本才能得到一个合理的表现。所以，对于构建一个能够适用于任何的持续期以及寿命模型分布类型的统一化的事件概率置信区间计算方法是本文所迫切希望得到的，这样一来，在各实证领域研究中就能够减轻分布假设对分析的影响。

综合来看，本论文的任务则是，在给定一组已知的历史寿命数据或持续期数据后，我们想要了解在未来的某一时刻，某一事件仍然持续（存活）的概率有多大，虽然研究者可能并不清楚该数据来自于何种寿命模型，但也并不需要通过模型拟合来推断出该数据来自于何种模型。一旦本文的方法体系得以在理论上得到较好的验证，将对未来持续期以及生命周期的推断预测具有重要的实际意义。

那么，要进行何种尝试才能够放松数据对于模型分布假设的依赖？方法体系建立出来以后，又要怎样去检验其表现是否良好，以及对于不同分布的稳健性又

如何？在实际应用中，是否能够得到良好的体现？这些就是本篇文章将要解决的主要问题。

三、研究思路与方法

本篇文章的现实目标是对未来持续期与生命周期进行预测及推断，着眼点为事件发生概率，而分析的入手点则是事件发生概率的置信区间。在数理统计中，预测事件发生概率相当于一个点估计，是一个具体的数值，但是这个数值本身并不能反映出这个点估计的精度如何，所以，考虑将置信区间纳入文章的研究范围，置信区间的估计解决这个问题：通常情况下将会设置一定的置信系数，置信系数可以反映区间估计的可靠程度例如，置信系数设置为 90%，则说明真值出现在所估计的区间中的概率至少为 90%。故通常情况下，在数理统计中都是在一定的置信系数下，去寻找尽可能短的置信区间。

下面，开始考虑问题：当获得一组持续期寿命数据时，在并不清楚它服从何种分布类型的情况下，该如何去处理这组数据来构建其事件发生概率的置信区间。

这里将考虑使用一种常见的数据转换方法对数据进行一定的变换，即 Box-Cox 幂转换方法。Box、Cox（1964）提出了基于极大似然法的幂转换模型，致力于将非负连续观测值转换成为正态或接近正态。本文则希望通过使用 Box-Cox 转换对持续期寿命数据进行加工处理，能够使其成为正态数据或近似正态数据。Box-Cox 转换帮助我们将不同类型的数据都可以转换为正态数据，也就使得不同的数据类型都能够得到了统一，具有了相同的起点，这也是本文欲建立统一化的方法体系所迫切希望看到的，故可以认为，Box-Cox 转换是整篇文章的基础与重点。

在得到转换数据后，本文所要解决的问题就得到了相应的简化，即如何对一组经过转化而得到的正态数据构造事件发生概率的置信区间。从理论推导层面上出发，本篇文章进行了很多构建方法的理论尝试，并通过使用 Monte Carlo 模拟的方法随机模拟产生几种分布类型下的数据进行模拟检验，来考察文中所尝试提出的事件概率置信区间构建方法的有限样本表现，初步挑选出了表现相对平稳且合理的三种方法，并进行进一步的比较分析。最后，文章将利用 Lancaster（1972）

以及 Kennan（1985）研究的罢工持续期数据来对所使用的方法进行进一步的实证检验分析。

在生成模拟数据、处理与分析模拟数据以及实证数据中，所用到的统计软件主要包括 matlab、R、以及 EXECL，另外，由于数据量的庞大，模拟试验中的 matlab 运算所借助的平台为厦门大学经济学院提供的高性能计算机群 HPC (High Performance Computing)。

最后，通过一系列的模拟与实证研究，可以检验以下结论是否能够得到实现：第一，无论观测数据来自于何种分布类型，文章中所建立的事件概率置信区间大体上表现良好；第二，针对于某些特殊分布，如对数正态分布或 Weibull 分布，它们具有专门针对该分布所构建时间事件发生概率置信区间的方法，而文章中提出的方法体系与为其专门设计的置信区间之间是可以比较的。

四、研究框架

本文共分为六章。

第一章为绪论，概括介绍了研究背景与研究意义、研究内容、研究思路与方法、研究框架以及创新和不足。

第二章为文献综述，回顾已有研究的成就与不足。

第三章讨论了三种建立事件概率置信区间的理论研究方法。

第四章通过蒙特卡洛模拟试验对第三章中提出的三种置信区间构建方法与已有的方法进行比较探讨。

第五章是将所用方法进行实证分析。

第六章为结论，总结了全文的主要工作与贡献，同时对未来研究做出了展望。

五、创新与不足

（一）创新之处

本文的创新之处主要包括以下两个方面：

1. 研究角度的创新

基于目前国内外已有的参考文献，关于对持续期和寿命模型的事件概率进行的研究很少。而本文则从这个新的角度出发，着眼于对事件发生概率的预测区间估计。同时，已有的文献中，基本关注点都在于对分布模型的构建拟合上，但本文所做的尝试则是放松对模型分布的假设，构造一个基于所有分布模型数据都可使用的一个统一化的方法体系，这样在未来的实证研究中，使得对于数据处理可以得到相应的简化，而不需要进行复杂的模型拟合工作。

2. 研究方法的创新

参考近几年学者的研究，借鉴 Box-Cox 幂转换的数据转换方法来进行数据的转换，为文章能够放松分布假设这个目标奠定了良好的基础。同时，在获得了转换的数据后，进行了不同的尝试，最后保留了三种表现较好的方法，包括正态近似、非中心t分布近似以及 bootstrap（重抽样）方法。

（二）不足之处

由于本人对数理统计学科的学习程度与对统计软件的灵活应用能力有限，故仍存在以下不足之处，还有很大需要学习进步的空间，在以后的学习研究中希望能够得到进一步的完善：

1. 应用分析较少

在目前已有的持续期寿命模型领域的研究，特别是在国内，可以发现，在经济领域的研究始终慢于工程学医学等领域，故本文进行理论分析的目标是为了更好的将文中所提出的方法在未来能够应用于经济领域的实证研究当中，为经济现象与发展趋势做出解释与预判，为经济领域研究做出一定的贡献。但是本篇文章大部分篇幅都是进行了理论分析和模拟检验，实证分析部分的偏重较少。由于理论方法是实证分析的基础，而理论方法的产生要经过大胆假设、寻找理论支持、模拟检验验证这重重程序，一旦中途产生问题或模拟结果不理想，则需要从头开始层层检查修改，需要投入更多的时间和精力成本、故本篇文章的侧重点主要是在理论推导与模拟检验上，有理由相信：理论的基石一旦坚固，实证的万丈高楼也将更加坚实地平地而起。故将方法更加充分地运用于实际的应用并研究现实意义可能要在后续时间慢慢进行扩充探究。

2. 数据处理

在对所提出的方法进行蒙特卡洛模拟检验时，为了构建完整的研究体系，对于样本量、参数、置信水平、检验概率点都对不同的取值进行了检验，在层层循环下，就有了庞大的数据量。另外在方法中还尝试使用了 bootstrap 重抽样，即在有限样本下进行大次数的重复抽样，也会导致数据量特别大，在老师帮助下使用了 HPC 运行，然而数据的产生也要耗费近一个月的时间，且每一次对方法的改进甚至推翻重改都是对时间成本的一个挑战，而在这个限制下，则需要控制模拟的次数，从而可能会产生这样的结果：模拟的数据并不能够达到最理想的状态。反思整个代码运行过程，有一部分的原因可能是所编写的 matlab 程序的优化不够，导致某些代码部分有浪费时间的嫌疑，这一方面可能还需要不断的改进优化。另一方面，整篇文章中所研究的数据都是基于完整数据，但是，有一个很重要的问题：截尾数据在持续期寿命模型分析中也是一个非常重要且普遍的数据类型，其中所谓截尾数据是指，在实验终止之时，某事件仍在持续，即该持续期寿命数据只得到了上界，未获得下界。但是在本文并没有将截尾数据纳入到考虑范围，希望在未来的研究中，能够延续本文对完整数据的研究，将事件发生概率置信区间的构建方法扩展运用到对截尾数据的分析中去。