格致方法·定量研究系列　吴晓刚　主编

分析复杂调查数据(第二版)

[美] 李殷凷（Eun Sul Lee）
罗纳德·N.福索佛（Ronald N. Forthofer）著
张卓妮 译

- 革新研究理念
- 丰富研究工具
- 最权威、最前沿的定量研究方法指南

23

格致出版社　上海人民出版社

格致方法·定量研究系列　吴晓刚　主编

分析复杂调查数据

[美] 李殷凰(Eun Sul Lee)
罗纳德·N. 福索佛(Ronald N.Forthofer) 著
张卓妮 译

SAGE Publications ,Inc.

格致出版社　上海人民出版社

图书在版编目(CIP)数据

分析复杂调查数据:第2版/(美)李殷鸎(Lee, E. S.),(美)福索佛(Forthofer, R. N.)著;张卓妮译. —上海:格致出版社:上海人民出版社,2012
(格致方法·定量研究系列)
ISBN 978-7-5432-2121-5

Ⅰ. ①分… Ⅱ. ①李… ②福… ③张… Ⅲ. ①抽样调查统计-研究 Ⅳ. ①C811

中国版本图书馆CIP数据核字(2012)第129136号

责任编辑　王亚丽

格致方法·定量研究系列
分析复杂调查数据(第二版)
[美]李殷鸎　罗纳德·N. 福索佛 著
张卓妮 译

出　版　世纪出版集团 www.ewen.cc
格致出版社 www.hibooks.cn
上海人民出版社
(200001 上海福建中路193号24层)

格致出版
编辑部热线 021-63914988
市场部热线 021-63914081

发　行　世纪出版集团发行中心
印　刷　浙江临安曙光印务有限公司
开　本　920×1168毫米 1/32
印　张　5.25
字　数　103,000
版　次　2012年7月第1版
印　次　2012年7月第1次印刷
ISBN 978-7-5432-2121-5/C·73
定　价　15.00元

出版说明

由香港科技大学社会科学部吴晓刚教授主编的“格致方法·定量研究系列”丛书，精选了世界著名的 SAGE 出版社定量社会科学研究丛书中的 35 种，翻译成中文，集结成八册，于 2011 年出版。这八册书分别是：《线性回归分析基础》、《高级回归分析》、《广义线性模型》、《纵贯数据分析》、《因果关系模型》、《社会科学中的数理基础及应用》、《数据分析方法五种》和《列表数据分析》。这套丛书自出版以来，受到广大读者特别是年轻一代社会科学工作者的欢迎，他们针对丛书的内容和翻译都提出了很多中肯的建议。我们对此表示衷心的感谢。

基于读者的热烈反馈，同时也为了向广大读者提供更多的方便和选择，我们将该丛书以单行本的形式再次出版发行。在此过程中，主编和译者对已出版的书做了必要的修订和校正，还新增加了两个品种。此外，曾东林、许多多、范新光、李忠路协助主编参加了校订。今后我们将继续与 SAGE 出版社合作，陆续推出新的品种。我们希望本丛书单行本的出版能为推动国内社会科学定量研究的教学和研究作出一点贡献。

总 序

往事如烟，光阴如梭。转眼间，出国已然十年有余。1996年赴美留学，最初选择的主攻方向是比较历史社会学，研究的兴趣是中国的制度变迁问题。以我以前在国内所受的学术训练，基本是看不上定量研究的。一方面，我们倾向于研究大问题，不喜欢纠缠于细枝末节。国内一位老师的话给我的印象很深，大致是说：如果你看到一堵墙就要倒了，还用得着纠缠于那堵墙的倾斜角度究竟是几度吗？所以，很多研究都是大而化之，只要说得通即可。另一方面，国内（十年前）的统计教学，总的来说与社会研究中的实际问题是相脱节的。结果是，很多原先对定量研究感兴趣的学生在学完统计之后，依旧无从下手，逐渐失去了对定量研究的兴趣。

我所就读的美国加州大学洛杉矶分校社会学系，在定量研究方面有着系统的博士训练课程。不论研究兴趣是定量还是定性的，所有的研究生第一年的头两个学期必须修两门中级统计课，最后一个学期的系列课程则是简单介绍线性回归以外的其他统计方法，是选修课。希望进一步学习定量研

此为试读，需要完整PDF请访问：www.ertongbook.com

究方法的可以在第二年修读另外一个三学期的系列课程,其中头两门课叫“调查数据分析”,第三门叫“研究设计”。除此以外,还有如“定类数据分析”、“人口学方法与技术”、“事件史分析”、“多层线性模型”等专门课程供学生选修。该学校的统计系、心理系、教育系、经济系也有一批蜚声国际的学者,提供不同的、更加专业化的课程供学生选修。2001 年完成博士学业之后,我又受安德鲁·梅隆基金会资助,在世界定量社会科学研究的重镇密歇根大学从事两年的博士后研究,其间旁听谢宇教授为博士生讲授的统计课程,并参与该校社会研究院(Institute for Social Research)定量社会研究方法项目的一些讨论会,受益良多。

2003 年,我赴港工作,在香港科技大学社会科学部,教授研究生的两门核心定量方法课程。香港科技大学社会科学部自创建以来,非常重视社会科学研究方法论的训练。我开设的第一门课“社会科学里的统计学”(Statistics for Social Science)为所有研究型硕士生和博士生的必修课,而第二门课“社会科学中的定量分析”为博士生的必修课(事实上,大部分硕士生在修完第一门课后都会继续选修第二门课)。我在讲授这两门课的时候,根据社会科学研究生的数理基础比较薄弱的特点,尽量避免复杂的数学公式推导,而用具体的例子,结合语言和图形,帮助学生理解统计的基本概念和模型。课程的重点放在如何应用定量分析模型研究社会实际问题上,即社会研究者主要为定量统计方法的“消费者”而非“生产者”。作为“消费者”,学完这些课程后,我们一方面能够读懂、欣赏和评价别人在同行评议的刊物上发表的定量研究的文章;另一方面,也能在自己的研究中运用这些成熟的

方法论技术。

上述两门课的内容，尽管在线性回归模型的内容上有少量重复，但各有侧重。“社会科学里的统计学”(Statistics for Social Science)从介绍最基本的社会研究方法论和统计学原理开始，到多元线性回归模型结束，内容涵盖了描述性统计的基本方法、统计推论的原理、假设检验、列联表分析、方差和协方差分析、简单线性回归模型、多元线性回归模型，以及线性回归模型的假设和模型诊断。“社会科学中的定量分析”则介绍在经典线性回归模型的假设不成立的情况下的一些模型和方法，将重点放在因变量为定类数据的分析模型上，包括两分类的 logistic 回归模型、多分类 logistic 回归模型、定序 logistic 回归模型、条件 logistic 回归模型、多维列联表的对数线性和对数乘积模型、有关删节数据的模型、纵贯数据的分析模型，包括追踪研究和事件史的分析方法。这些模型在社会科学研究中有着更加广泛的应用。

修读过这些课程的香港科技大学的研究生，一直鼓励和支持我将两门课的讲稿结集出版，并帮助我将原来的英文课程讲稿译成了中文。但是，由于种种原因，这两本书拖了四年多还没有完成。世界著名的出版社 SAGE 的“定量社会科学研究”丛书闻名遐迩，每本书都写得通俗易懂。中山大学马骏教授向格致出版社何元龙社长推荐了这套书，当格致出版社向我提出从这套丛书中精选一批翻译，以飨中文读者时，我非常支持这个想法，因为这从某种程度上弥补了我的教科书未能出版的遗憾。

翻译是一件吃力不讨好的事。不但要有对中英文两种

语言的精准把握能力,还要有对实质内容有较深的理解能力,而这套丛书涵盖的又恰恰是社会科学中技术性非常强的内容,只有语言能力是远远不能胜任的。在短短的一年时间里,我们组织了来自中国内地及港台地区的二十几位研究生参与了这项工程,他们目前大部分是香港科技大学的硕士和博士研究生,受过严格的社会科学统计方法的训练,也有来自美国等地对定量研究感兴趣的博士研究生。他们是:

香港科技大学社会科学部博士研究生蒋勤、李骏、盛智明、叶华、张卓妮、郑冰岛,硕士研究生贺光烨、李兰、林毓玲、肖东亮、辛济云、於嘉、余珊珊,应用社会经济研究中心研究员李俊秀;香港大学教育学院博士研究生洪岩璧;北京大学社会学系博士研究生李丁、赵亮员;中国人民大学人口学系讲师巫锡炜;中国台湾"中央"研究院社会学所助理研究员林宗弘;南京师范大学心理学系副教授陈陈;美国北卡罗来纳大学教堂山分校社会学系博士候选人姜念涛;美国加州大学洛杉矶分校社会学系博士研究生宋曦。

关于每一位译者的学术背景,书中相关部分都有简单的介绍。尽管每本书因本身内容和译者的行文风格有所差异,校对也未免挂一漏万,术语的标准译法方面还有很大的改进空间,但所有的参与者都做了最大的努力,在繁忙的学习和研究之余,在不到一年的时间内,完成了三十五本书、超过百万字的翻译任务。李骏、叶华、张卓妮、贺光烨、宋曦、於嘉、郑冰岛和林宗弘除了承担自己的翻译任务之外,还在初稿校对方面付出了大量的劳动。香港科技大学霍英东南沙研究院的工作人员曾东林,协助我通读了全稿,在此

我也致以诚挚的谢意。有些作者，如香港科技大学黄善国教授、美国约翰·霍普金斯大学郝令昕教授，也参与了审校工作。

我们希望本丛书的出版，能为建设国内社会科学定量研究的扎实学风作出一点贡献。

吴晓刚

于香港九龙清水湾

序

当乔治·盖洛普(George Gallup)正确预测到富兰克林·罗斯福(Franklin D. Rooselvelt)成为1936年美国总统选举赢家时,公众民意调查进入了科学抽样的时代,那时使用的方法是配额抽样,一种代表目标总体的非概率抽样方法。但是,同样的方法却错误地预测托马斯·杜威(Thomas Dewey)将成为1948年的总统,但实际上哈利·杜鲁门(Harry S. Truman)才是最终的赢家。这种方法之所以失败,是因为配额抽样是非概率的,而且盖洛普的配额抽样框基于1940年的人口普查,忽略了二次大战期间的城市移民。

今天的调查抽样方法和早些时候相比已有了很大进步,现在我们依赖于复杂的概率抽样设计。一个关键的特征就是分层:目标总体被分成一系列不同阶层的子总体,阶层内的样本量由抽样者控制且通常与该阶层的人口规模成比例。另一个特征是整群和多阶段抽样:先抽取组群——即在不同阶层选到的一系列不同等级的集群,到最后才在这些"最后阶层"的集群内部选取个体成员。比如,美国综合社会调查使用的就是分层多阶段集群抽样设计。Kalton(1983)为调查

抽样提供了很好的入门介绍。

当调查设计具有这个复杂的性质时,数据的统计分析就不再是一个简单的运行回归(或任何其他的模型)分析的事情了。现在的调查全都伴随着抽样权重以支持正确的统计推断,大部分关于统计分析的教材,通过假定简单随机抽样而没有处理抽样权重的问题,而这个被忽略的抽样权重可能对统计推断产生重要的影响。在最近的二三十年间,数据分析的统计方法也已取得了巨大进展。这些很可能就是Michael S. lewis-Beck选择为《分析复杂调查数据》(*Analyzing Complex Survey Data*)出第二版的原因。

Lee和Forthofer这本书的第二版为我们在调查抽样设计和调查数据分析的联合方面提供了最新的情况。作者在本书开头先回顾了调查抽样设计常见的类型,并通过解释什么是抽样权重及其如何产生和调整,揭开了抽样权重的神秘面纱。然后他们详细讨论了方差估计以及考虑抽样权重后的复杂截面调查数据的初级和多变量分析等主要问题。他们重点关注了基于设计的方法,这种方法直接在分析中涉及抽样设计(虽然他们也讨论了基于模型的视角,该视角在某些分析中能扩充基于设计的方法),他们还以流行的软件为例阐释了这种方法的使用。读者将会在以样本为基础作出统计推断的实践中发现本书的巨大益处。

廖福挺

此为试读,需要完整PDF请访问: www.ertongbook.com

目 录

第1章

概　论

调查分析似乎通常是在所有的样本观察值都以同等的概率被独立选中的情况下实行的。如果在数据收集中使用的是简单随机抽样(SRS),这种分析就是对的,但实际上样本选择比 SRS 复杂得多。某些样本观察值可能比其他观察值以更高的概率被选中,某些之所以被包括在样本中,是因为他们属于某个特定组别的成员(比如家庭),而不是被独立选取的。我们可以在调查数据的分析中简单地忽略这些与 SRS 相悖的事实吗?使用调查数据分析统计书中的标准技术是否合适?或者是否有特别的方法和计算机程序更适合复杂调查数据的分析?这些问题将在随后的章节中进行论述。

今天典型的社会调查反映了统计理论和关于社会现象的知识两者间的结合,过去 70 年间从许多不同的调查中获得的经验塑造了它的发展。社会调查的进行是为了满足用以讨论社会、政治和公共卫生议题的信息需要。为了满足这种信息需要,在政府内部和外部都设立了调查机构。但是,在早期提供这些信息的尝试中,调查小组们最关心的是实地调查中的操作问题,如抽样框的建立,员工培训/监督,以及成本的降低等,理论上的抽样议题获得的只是次级重视(Ste-

phan, 1948)。随着这些实际问题得到解决,现代抽样实践方法的发展远远超越了 SRS。复杂抽样设计已经走在了前面,随之而来的是一系列分析问题。

因为早期调查通常需要的只是描述性统计,所以很少有人对分析问题感兴趣。更近一些时期,社会和政策科学家对分析研究的需要已经增加,那些并不参与数据收集过程的研究人员,正利用可用的社会调查数据对各种不同的现实议题进行分析。这个传统上被称为二手资料分析(Kendall & Lazarsfeld, 1950)。研究者通常并没有对复杂抽样设计的发展给予应有的关注,并假设这些设计对将要使用的分析程序没有什么影响。

二手资料分析中统计技术使用的增加以及近年来对数线性模型、logistic 回归和其他多变量技术的使用(Aldrich & Nelson, 1984; Goodman, 1972; Swafford, 1980),并没有把设计和分析更紧密地结合起来。这些技术断定数据收集使用的是简单有放回随机抽样(SRSWR),但这个假设在应用了观测单位的分层和整群方法以及不等概率选择法的社会调查中几乎无法得到满足。因此,利用 SRSWR 假设的社会调查分析可能导致有偏差和误导性的结果。比如,Kiecolt 和 Nathan(1985)在他们关于二手资料分析的书中承认了这个问题,但是他们几乎没有就如何把抽样权重和其他设计特征融入分析之中提供什么建议。最近一篇关于公共健康和流行病学的文献研究表明:对基于设计的调查分析方法的使用正在逐渐增加,但依然处于较低水平(Levy & Stolte, 2000)。

任何对抽样进行了限制且那些限制超越了 SRSWR 所做限制的调查在设计上都很复杂,并需要特别的分析性考虑。

本书回顾了由复杂抽样调查引起的分析问题,为分析策略提供了入门介绍,并利用某些可用的软件进行了实例阐释。我们讨论的重点在于使用抽样权重以校正不同的代表性,以及抽样设计对抽样方差估计的影响,我们也对权重的产生和调整程序做了一些讨论。但许多其他重要的处理非抽样误差和缺失数据的议题并没有在本书中得到充分说明。

本书介绍的最基本的方法是分析复杂调查数据的传统方法。这种方法现在被称为基于设计的(或基于随机化的)分析。另一种不同的分析复杂调查数据的方法,是所谓的基于模型的分析。正如统计学的其他领域,近些年来基于模型的统计推断在调查数据分析中已引起了更多的注意。这些构造模型的方法在调查数据分析的不同步骤——定义参数、定义估计值和估计方差——中得以介绍;但是,并没有普遍被接受的模型选择或使某个特定模型有效化的规则。

然而,对基于模型的方法的理解对调查数据分析者扩充基于设计的方法来说非常重要。在某些情况下,这两种方法会产生相同的结果,但在其他情况下结果却不同。基于模型的方法可能在描述性数据中没什么作用,但在推断性分析中却是有用的。我们将在适当的地方介绍基于模型的视角,并进一步提供与这些问题的处理方式相关的参考资料。恰当地实施基于模型的分析需要对一般统计模型知识的掌握,而且还需要从调查统计员那里得到某些参考信息。此书中与这种不同的方法或相关题目有关的章节用星号(*)标出。

自本书第一版出版以来,对复杂调查数据进行分析的软件情况已有相当大的改进。方便用户的程序现在很容易就可获得,许多常用的统计方法现在也已合并成不同的程序

此为试读,需要完整PDF请访问: www.ertongbook.com