第2版

社会学教材教参方法系列

SM

分类数据分析的统计方法

Statistical Methods for Categorical Data Analysis(Second Edition)

〔美〕丹尼尔·A.鲍威斯（Daniel A. Powers）
谢　宇（Yu Xie）/著

任　强　巫锡炜　穆　峥　赖　庆/译

社会科学文献出版社 SOCIAL SCIENCES ACADEMIC PRESS(CHINA)

第2版

社会学教材教参方法系列

SM

分类数据分析的统计方法

Statistical Methods for Categorical Data Analysis(Second Edition)

〔美〕丹尼尔·A. 鲍威斯 (Daniel A. Powers)
谢　宇 (Yu Xie) /著

任　强　巫锡炜　穆　峥　赖　庆/译

社会科学文献出版社
SOCIAL SCIENCES ACADEMIC PRESS (CHINA)

社会学教材教参方法系列

分类数据分析的统计方法（第2版）

著　　者／〔美〕丹尼尔·A. 鲍威斯　谢　宇
译　　者／任　强　巫锡炜　穆　峥　赖　庆

出 版 人／谢寿光
总 编 辑／邹东涛
出 版 者／社会科学文献出版社
地　　址／北京市西城区北三环中路甲29号院3号楼华龙大厦
邮政编码／100029
网　　址／http：//www. ssap. com. cn
网站支持／（010）59367077
责任部门／社会科学图书事业部（010）59367156
电子信箱／shekebu@ ssap. cn
项目经理／杨桂凤
责任编辑／杨桂凤
责任校对／张锦文
责任印制／郭　妍　岳　阳　吴　波

总 经 销／社会科学文献出版社发行部
（010）59367080　59367097
经　　销／各地书店
读者服务／市场部（010）59367028
排　　版／北京中文天地文化艺术有限公司
印　　刷／三河市世纪兴源印刷有限公司

开　　本／787mm×1092mm　1/16
印　　张／21.75
字　　数／383千字
版　　次／2009年7月第1版
印　　次／2009年7月第1次印刷

书　　号／ISBN 978-7-5097-0866-8
著作权合同登记号／图字01-2009-2487号
定　　价／39.00元

本书如有破损、缺页、装订错误，
请与本社市场部联系更换

图书在版编目（CIP）数据

分类数据分析的统计方法（第2版）/〔美〕鲍威斯（Powers，D. A.），谢宇著. —北京：社会科学文献出版社，2009.7
（社会学教材教参方法系列）
ISBN 978-7-5097-0866-8

Ⅰ.分… Ⅱ.①鲍…②谢… Ⅲ.数据-统计分析 Ⅳ.C813

中国版本图书馆CIP数据核字（2009）第096673号

中文版序

《分类数据分析的统计方法》（第2版）的中文版终于和读者见面了，我感到非常高兴。

《分类数据分析的统计方法》是我和Daniel Powers合著的，也是我的第一本书。第一版于2000年由美国的学术出版社（Academic Press）出版，第二版于2008年由英国的翡翠出版社（Emerald Group）出版。很荣幸的是，我们能在2009年英文第2版刚刚出版后不久就见到由社会科学文献出版社出版发行的中文版。

《分类数据分析的统计方法》是为社会科学——特别是社会学——做定量研究的学者和学生专门写作的教材和参考书。本书介绍、探讨了许多社会科学定量研究中实际碰到的统计方法问题。这些方法是社会科学定量研究人员都应该掌握的基本功，也是我对自己的所有学生都要求其学会的。可惜的是，一些国内的学者还认为本书包括的内容“太复杂”了。他们应该知道，社会现象本身要更复杂得多。再复杂的统计方法都是建立在我们对更复杂的社会现象做大量简化的基础之上的。虽然统计方法最终不可能让我们完美地了解社会现象，但不同的统计方法可以更好地适用于不同的社会科学研究应用之中。换句话说，统计方法虽然不能给我们十全十美的答案，但适当的统计方法相比不适当的统计方法会给我们更可信、更有学术意义的答案。所以，一个社会科学定量研究做得好的学者应该掌握各种不同的统计方法，才能做到对症下药。我希望本书中文版的出版有利于提高国内社会科学定量研究的水平。

本书的特点是着重于对有关统计方法的理解，而不是对这些统计方法的理论

证明。为了方便读者，我们另外通过互联网提供了用不同统计软件（包括 aML、GAUSS、LEM、LIMDEP、R、SAS、STATA、TDA、WinBUGS 和 OpenBUGS）编写的例题程序。网址是：http：//www. powers - xie. com；或通过我的个人主页 http：//www - personal. umich. edu/ ~ yuxie/链接；或通过 Daniel Powers 的个人主页 http：//webspace. utexas. edu/dpowers/www 链接。

本书的最初来源是我和 Daniel Powers 的教学讲义。我们也用本书作为教材教过许多学生，学生对本书有过很多好的建议。我们再次感谢他们。

翻译本书的主要负责人是北京大学的任强老师，他已和我认识多年。他在第1版出版不久就想翻译本书。当时我认为时机还没有成熟。2008 年我们出第2版时，他正好在我这里做访问学者，给翻译本书提供了很好的机会。另外三位学生巫锡炜、穆铮、赖庆也积极参加了翻译工作，做了重要的贡献。在此，我向他们四位表示感谢。

最后，我想感谢社会科学文献出版社的谢寿光社长和杨桂凤编辑。如果没有他们的积极支持和辛苦工作，就没有本书的中文版。多谢了！

谢　宇

2009 年 6 月 1 日于安娜堡

前　言

在本书中，我们试图对社会科学研究中的分类数据分析以及应用的方法和模型做一个全面的介绍。本书主要面向研究生和社会科学应用研究的学者，同时也可以作为参考工具书。

一个区别于其他教科书主题的特点是我们有明确的目标，即整合转换方法（transformational approach）和潜在变量方法（latent variable approach），它们是处理分类数据分析的两种完全不同但相互补充的方法。在人口学和生物统计领域，处理分类数据分析的统计或转换方法是研究者最为常见的方法，而潜在变量方法则被经济学家经常使用。第1章将会讨论这两种方法。

我们假定读者已经具备初步的知识，例如，掌握了应用回归课程的知识，但不需要高级数理统计知识。尽管一些技术细节不可避免，但是我们借助了大量实例来帮助大家理解本书。一些读者可能会略过书中的技术部分，但这样也不会失去本书的很多精华。

为了充分利用互联网技术，我们为本书设置了网站（http://webspace.utexas.edu/dpowers/www/）。[①] 网站包含了书中用不同软件处理所讨论例子的数据集和程序编码，这些软件包括 GLIM（Numberical Algorithms Group Ltd.，1986），LIMDEP（Greene，2007），SAS（SAS Institute，2004），Stata（Stata，2007），TDA（Rohwer & Pötter，2000），以及 R（R Development Core Team，2006）。为

① 此主页被链接在 YuXie.com 和 Powers-Xie.com 上。

了描述估计的细节和介绍几个标准统计软件包不能估计模型的特殊程序，网站提供了一些 GLIM 的宏命令和 GAUSS（Aptech Systems，1997）与 R 子程序，如 aML（Lillard & Panis，2003）。当获得新的程序时，我们会继续更新网站内容。

◎ 第2版新增内容

我们已经更新了每一章的内容，并新增了一章关于二分类变量的多层模型（第5章）。第5章详细介绍了边际最大似然估计和现代贝叶斯估计方法（Bayesian estimation methods）。我们也针对纵贯数据分析的 Rasch 模型和随机系数模型进行了讨论，重新组织了事件史模型这一章（第6章），扩展了离散时间模型和 Cox 回归模型。对次序因变量模型（第7章）和名义因变量模型（第8章）这两章也进行了更新。

◎ 本教材在分类数据模型课程中的使用

本书适合于为期一个学期的分类数据建模课程。第1章和第2章是一般性介绍与课程基础。我们的观点是，无论数据类型如何，回归类建模方法都是一个合适的分析方法。第3章介绍并详细讨论了针对二分类数据的回归模型。第4章深入讲解了分析列联表的模型。第5章讨论了针对二分类数据的多层/分层模型。第6章介绍了事件史分析技术。第7章和第8章回顾了针对次序和非次序分类因变量的模型。这部分内容与第4章的列联表方法和第3章介绍的潜在变量分析框架是有关联的。

◎ 致谢

在本书写作的各个阶段，我们从下列学者的鼓励以及与他们的联系中获益匪浅：Paul Allison，Mark Becker，John Fox，Richard Gonzalez，Leo Goodman，David Grusky，Robert Hauser，Michael Hout，Kenneth Land，Scott Long，Charles Manski，Robert Mare，Bill Mason，Susan Murphy，Trond Peterson，Thomas Pullum，Adrian Raftery，Steve Raudenbush，Arthur Sakamoto，Herbert Smith，Michael Sobel，Chris Winship，Raymond Wong，Larry Wu 和 Kazuo Yamaguchi。此外，我们对许多学习这门统计课程的研究生表示感谢，是他们激励我们写这本书的。

资助丹尼尔·A. 鲍威斯的奥斯汀得克萨斯大学的主任基金、资助谢宇的国家自然科学基金的青年学者基金和密歇根大学基金对本书的研究提供了部分资助。

我们也要感谢外部评审对早期初稿提出的宝贵意见，以及 Pam Bennett，John Fox，Kimberly Goyette 和 James Raymo 对书稿最后版本的仔细校对和在第 1 版中对实例的编程工作。感谢 Meichu D. Chen 和许多研究生，他们指出了第 1 版中的一些错误。特别感谢 Cathy（Hui）Liu 对第 2 版新内容的仔细阅读。我们也要感谢 Cindy Glovinsky 卓越的编辑工作。我们将对书中仍然存在的错误负责。

最后，我们感谢学术出版社（Academic Press）和 Elsevier 的编辑 J. Scott Bentley 提出这个项目，并努力使第 1 版面世，同时促使我们完成第 2 版的编写工作。感谢 EmeraldInsight 的 Rachel Brown 女士对出版第 2 版的帮助，也要感谢 Macmillan 对编排此书的帮助。

丹尼尔·A. 鲍威斯

谢 宇

目录
CONTENTS

第1章

绪　论

1.1　为什么需要分类数据分析?

生育、结婚、入学、就业、职业、迁移、离婚和死亡的共同之处是什么？答案在于它们都是社会科学研究中经常碰到的分类变量（categorical variables）。实际上，社会科学研究中的绝大多数观测结果都是以分类变量的形式加以测量的。如果你是一位从事实际研究工作的社会科学家，你可能在很多具体的研究工作中遇到过分类变量（即使你从未用过任何专门的统计方法来处理分类变量，这也是千真万确的）。如果你目前是在读研究生，打算将来成为一名社会科学家，你可能还没有但很快就将遇到分类变量的问题。请注意，就你在目前为止的人生中是否遇到过分类变量的表述本身就是一种分类测量!

在过去大约25年中，针对分类数据分析的统计方法和技术得到了快速发展。很大程度上得益于商业软件的普及和计算的简化，近年来，它们在实际研究中的应用已经变得越来越普遍。因为其中一些资料相当前沿而且散见于多个学科，因此，我们认为需要有本书来专门对该主题做系统讨论。本书的目的是为应用社会科学家使用合适的分类数据分析工具提供帮助。在本章中，我们将首先定义何谓分类变量，然后介绍我们针对此主题的处理思路。

1.1.1　分类变量界定

在我们看来，分类变量指的是那些只用有限个取值或类别加以测量的变量。这一定义将分类变量与连续变量区分开来。从原理上讲，连续变量可以被认为具

有无限多个取值。

尽管分类变量的这一定义非常清楚，但是其在实际工作中的应用仍然显得非常含糊不清。社会科学家持续关注的许多变量明显都是分类的。这些变量包括：种族、性别、迁移状态、婚姻状态、就业、出生和死亡。然而，一些概念上的连续变量有时候被处理成连续的，而其他时候则被处理成分类的。当某个连续变量被处理成一个分类变量时，这被称作连续变量的类别化（categorization）或离散化（discretization）。在实际研究工作中，常常必须进行类别化，因为某个连续变量的实质含义（substantive meaning）或实际测量（actual measurement）就是分类的。年龄就是一个很好的例子。尽管在概念上是连续的，但是，出于实际和应用的考虑，年龄在实际研究中常常被处理成分类的。实质上，基于某些研究目的，年龄会被作为质性状态（qualitative states）的代理，定性地表示个体在某一关键点所出现的状态转换。个体在法律和社会状态上的变化首先出现在其转入成人期的时候，随后出现在其退出劳动力市场的时候。出于应用的原因，年龄通常被表达成单岁或5岁年龄组。①

的确，社会科学研究中的常用方法有意地将可能的响应限定为有限的可能取值，在这一意义上，它们都是粗略的。正因为如此，我们在前面就已经指出，如果不是全部的话，社会科学中的绝大多数观测结果都是分类的。

那么，相对于连续变量而言，哪些变量在经验研究中应当被认为是分类的呢？答案取决于许多因素，其中的两个是它们在理论模型中的实质含义和测量精确度。将某个变量处理成分类变量的一个必要条件就是它的取值反复出现并在样本中达到一定的比例。② 正如随后将要介绍的那样，区别连续变量和分类变量，对响应变量（response variables）来讲比对解释变量（explanatory variables）要重要得多。

1.1.2 因变量和自变量

因变量（也称响应变量、结果变量或内生变量）表示某项研究中一个被解

① 受教育水平是另一个例子。如果不加以类别化，就不能揭示“低于12年教育”、“高中文凭”、“大学学历”或“研究生学历”之间的实质区别。一些类别提供了教育分布上重要节点的一个简明表达。

② 注意，连续变量也可能被删截，这意味着得到一个超过某一特定门槛或转折点的取值的概率为零。当某一连续变量被删截时，未被删截的部分仍然是连续的，而被删截的部分则像一个分类变量。

释的总体特征。自变量（也称解释变量、先决变量或外生变量）是被用来解释因变量发生变异的那些变量。具体来讲，所关注的特征是因变量（或其转换形式）以某一自变量或一组自变量的取值作为条件的总体均值（population mean）。正是在这一意义上，我们认为在回归类统计模型（regression-type statistical models）中，因变量取决于自变量、被自变量所解释或者是自变量的函数。

采用“回归类统计模型”这一表述，意味着对因变量的期望值或其他特征进行预测的模型是自变量的一个回归函数。从原理上讲，尽管我们可以设计出能最好地对因变量或其转换形式的任何总体参数（例如中位值）进行预测的模型，但在实践中，我们通常使用回归这一术语来表明是要预测条件均值（conditional means）。当回归函数为自变量的线性组合时，我们就得到了所谓的线性回归。它们被广泛应用于连续因变量。

1.1.3 分类因变量

尽管分类变量和连续变量具有许多共同的属性，但是我们希望在这里突出它们之间的一些差异。作为因变量，分类变量和连续变量之间的差别需要特别加以重视。相比而言，当它们在回归类统计模型中被用作自变量时，差别相对不那么重要。我们所采用的回归类统计模型的定义包括针对方差和协方差分析的统计模型，这些模型可以通过因变量对一组虚拟变量进行回归的形式加以表达，在协方差分析的情况下，还包括其他的连续协变量。因此，在回归类模型中纳入分类变量作为解释变量并不存在任何特别的难处，因为它主要涉及建立与自变量不同类别相对应的虚拟变量，所有已知回归模型的性质都可以直接推广到方差和协方差分析模型。正如本书随后将要介绍的，当我们将分类变量作为因变量处理时，情况彻底改变了，因为线性回归的许多知识都无法简单地加以应用。简而言之，分类数据分析（即涉及分类因变量的分析）需要专门的统计方法。

尽管在回归类模型中将分类变量作为自变量分析的方法已经成为标准统计知识基础的一部分，而且现在还成了社会科学中获取最高学位所必须掌握的内容，但是分类因变量的分析方法远不为人们所熟悉。大部分有关分类数据分析方法的基础性研究只是近年来才发展起来的。本书的目标是系统介绍与分类数据分析相关的几个重要主题，以便将这些知识整合到社会科学研究中来。

与连续变量的分析方法不同，分类数据的分析方法特别注意因变量的测量类型。分析某一类分类因变量的方法可能并不适合分析另一类分类变量。

1.1.4 测量类型

当某个变量被用作因变量时，测量类型在确定恰当的分析方法方面起着关键的作用。考虑到三个方面的差别，我们针对四种测量类型提出了一个分类模式（typology）。[①] 我们首先区分定量和定性测量之间的差别。二者之间的差别在于：定量测量严格地用数值来标示变量的实质含义；而定性测量的数值不具有实质含义，有时只是作为区分某些相互排斥的、具有唯一性的特征（或属性）的分类。定性变量属于分类变量。

在定量变量这一类别中，进一步区分连续变量和离散变量往往非常有用。连续变量也被称为定距变量（interval variables），可以取任意实数值。例如，收入和社会经济地位等变量在其可能的取值范围内通常被作为连续变量处理。离散变量只能取整数值（integer values）且往往表示事件计数（event counts）。例如，每个家庭的孩子数、某一青少年的违法行为次数、某一路口每年的交通事故数等都是离散变量的例子。根据前面的定义，离散（但定量的）变量也属于分类变量。

定性测量可以进一步区分出次序（ordinal）测量和名义（nominal）测量两种。次序测量产生有次序关系的定性变量（ordered qualitative variables）或次序变量（ordinal variables）。对某个包含次序关系的定性变量，通常的做法是采用数值来标示排序信息（ordering information）。但是，与次序变量各类别相对应的数值只反映某一特定属性上的排序，因而相邻数值之间的距离并不相同。对枪支管制的态度（坚决支持、支持、中立、反对和坚决反对）、职业技能水平（高级、中级、低级和无技能）和受教育水平的分类（小学、中学、大学和研究生）等都是次序变量的例子。

名义测量产生无次序关系的定性变量（unordered qualitative variables），往往被称作名义变量（nominal variables）。名义变量的类别之间并没有内在的次序，也没有数值距离。种族和民族（白人、黑人、西班牙裔和其他）、性别（男性和女性）以及婚姻状况（未婚、已婚、离婚和丧偶）等都是无次序关系的定性变量的例子。但是，值得注意的是，次序变量和名义变量之间的区别并不总是那么清晰，其中的区别在许多情况下取决于所研究的问题。同一个变量对某些研究者而言可能是次序变量，而对其他研究者而言则可能是名义变量。

---

① 有关的历史背景，请参见 Duncan（1984）的重要著作 *Notes on Social Measurement*。

为了进一步解释最后一点，下面我们以职业为例加以说明。不同的职业往往使用开放式问题加以测量，然后采用三位数字码人工将被调查者的回答编码成一个分类体系，但是这些数字编码并不表示具有实质意义的量。考虑到可能的职业类别非常多（对于现代社会而言，编码体系中通常至少包含几百类），这就期望，也的确有必要通过数据简化来减少某一职业测量的详细程度。一种数据简化的方法是将详细职业编码合并成大的职业类别，并将其看作次序测量或者名义测量（Duncan，1979；Hauser，1978）；另一种方法是从某一社会经济指数（socioeconomic index，SEI）的维度将职业尺度化（scale）（Duncan，1961），从而变成一个定距变量。最近，Hauser 和 Warren（1997）对 Duncan 的方法提出了挑战，建议最好将职业尺度化成职业收入和职业教育两个独立的维度，从而对职业社会经济地位加以测量。Hauser 和 Warren 的研究阐明了当名义测量被尺度化为定距测量时考虑多个维度的重要性。

图 1－1 概括了我们针对四种测量类型提出的类型学框架。根据该类型学划分，分类变量有 3 个类别：离散的、次序的和名义的，我们将在本书中对其加以讨论。只有当变量的可能取值等于或超过 3 个时，分类变量 3 个类别之间的区别才有意义。当变量的可能取值为 2 个时，我们就有一种被称作二分类变量的特例。二分类变量可以是离散的、次序的或名义的，这取决于研究者的解释。例如，如果研究者的兴趣是研究中国一孩政策的遵从情况，那么因变量就是夫妻生育过的孩子数是否超过一个。为了简便起见，假设在某一特定样本中，一名妇女至少有一个孩子且不超过两个。我们对 y 进行编码。如果某妇女有一个孩子，则 $y=0$；如果她有两个孩子，则 $y=1$。在这种情况下，因变量可以被解释成离散的（孩子数减去 1）、次序的（一孩或一孩以上）或名义的（遵从相对于不遵从）。幸运的是，研究者可以采用相同的统计方法来处理这三种情况——只是对结果的实质含义的理解会随着解释的不同而有所不同。

图 1－1 四种测量的分类模式