21世纪
人口学系列教材

社会统计分析与数据处理技术
——STATA软件的应用

杨菊华/著

中国人民大学出版社

21世纪人口学系列教材

社会统计分析与数据处理技术

——STATA软件的应用

杨菊华 著

中国人民大学出版社

图书在版编目（CIP）数据

社会统计分析与数据处理技术——STATA 软件的应用/杨菊华著.
北京：中国人民大学出版社，2008
（21 世纪人口学系列教材）
ISBN 978-7-300-08997-3

Ⅰ. 社…
Ⅱ. 杨…
Ⅲ. 社会统计-统计分析-应用软件，STATA-高等学校-教材
Ⅳ. C91-03

中国版本图书馆 CIP 数据核字（2008）第 020967 号

21 世纪人口学系列教材
社会统计分析与数据处理技术——STATA 软件的应用
杨菊华 著

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010-62511242（总编室）	010-62511398（质管部）	
	010-82501766（邮购部）	010-62514148（门市部）	
	010-62515195（发行公司）	010-62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京东君印刷有限公司		
规　　格	185 mm×230 mm　16 开本	**版　　次**	2008 年 3 月第 1 版
印　　张	29	**印　　次**	2008 年 3 月第 1 次印刷
字　　数	546 000	**定　　价**	49.80 元（本书附赠光盘）

总 序

中国是世界第一人口大国。无论是在宏观上，还是在微观上，中国社会经济的发展都与人口的状况和问题紧密相关。在中国追求现代化的历史过程中，人口问题始终是我们长期面临的重大问题。人口问题的复杂性世所公认，但中国人口问题可能更复杂。一个多世纪以来，人口问题几乎成了所有经济学家、社会学家和政治家最为关注的热门话题之一。尽管如此之多的人士研究人口问题，但是近百年来，在对中国人口问题的认识上，却经历了反复的曲折，积累了无数的教训。

对一个事物的正确认识，需要有科学的理论和科学的方法。中国人口问题如此重大，如此复杂，任何人要想深入了解、分析和研究人口问题，必须要掌握基本的人口学理论与方法。20世纪70年代中期，中国开始了具有历史意义的现代人口学科的创建过程。人口学教材建设在一开始就被列为学科创建的最重要工作之一。到20世纪80年代中期，中国的大学里已经系统地建立了完整的人口学学士、硕士和博士的培训课程，出版了一系列与国际前沿接轨的人口学教材，如《人口统计学》、《人口经济学》、《世界人口》、《人口理论教程》、《人口思想史》、《人口社会学》、《人口规划与预测》、《人口地理学》、《现代人口分析技术》、《人口普查的分析方法》等，并培养了一大批人口学的专门人才。从那时到现在，国际上人口科学的发展日新月异，新方法、新理论、新流派层出不穷。与此同时，中国人

口的研究也迅速向深度和广度推进。收集到的人口数据堆积如山，人口数量、结构和质量，人口与经济，人口与社会，人口与环境资源等各种问题几乎全部纳入了研究视野。但从20世纪90年代初期以来，人口学教材的编写高潮却逐渐回落。反映国际最新进展并与中国人口实践紧密结合的新教材，无论是在种类上还是在水平上，都远远不能满足现实的需要。

中国人民大学人口学系（所）是国内最早开设人口学学位教育的机构，曾经培养过上千名人口学专业的本科生、硕士生和博士生，20世纪80年代初编写的《人口统计学》、《世界人口》、《人口理论教程》等著作曾在人口学界产生了重大而深远的影响。进入21世纪后，为了满足人口学教育的需要，在教育部和中国人民大学出版社的大力支持下，中国人民大学人口学系（所）决定牵头组织编写一套能够反映国际国内人口学最新进展的《21世纪人口学系列教材》。本系列教材将涵盖人口学主要分支学科的内容，为各个大学和培训机构开设人口学公共课和专业课提供科学、准确的教学资料。

尽管本套教材是在长期教学实践基础上编写的，编写过程中我们也力争逻辑严谨、内容充实、形式活泼，但疏漏在所难免，望广大读者指正，以便再版时修改。

翟振武

内容摘要

事物之间彼此联系，这些联系有其内在机制，并且以多样形式而存在。社会科学研究就是要发现这些形式，并解释这些机制。系统的研究理论、科学的研究方法、合理的研究步骤有助于更好地发现关联事物之间的显性和潜在规律。西方社会科学研究积累了大量的理论和方法，而在中国，由于研究对象的复杂性及量化水平偏低，社会科学研究至今依然停留在规范研究领域，侧重于理论思辨和逻辑推理，以事实为依据的实证研究才刚刚起步。

20世纪80年代以来，西方现代社会统计技术开始传入，定量分析方法得到推广。过去20年来，包括SPSS和SAS在内的一些统计软件在中国也逐步得到应用。其中SPSS传入中国较早，简便易学，故国内使用者众多；SAS功能十分强大，但学习应用相对比较困难，应用范围较小。这些软件凭借先入优势，占据了一定的市场，影响其他功能更强大、使用更方便的软件的进入，在西方社会科学分析中常用的Stata软件就是一例。虽然已经开始有了许多使用者，但是配套的教材却很稀缺。

鉴于以上情况，作者编写了《社会统计分析与数据处理技术——STATA软件的应用》一书，本书以Stata的第9版为例，详细介绍该软件在社会科学研究领域的应用。本书采取循序渐进的原则，由浅入深，由易到难。在遵循国外相关教材

的体例的基础上，也考虑到国内学习者目前定量研究的数据处理能力，重点放在数据的处理上。与一般的 Stata 的使用手册不同，本书除介绍 Stata 的一些具体使用方法外，在一些章节还介绍相关的统计原理、数据处理的思路和缘由、研究方法等。

本书共分为四部分。第一部分介绍社会科学研究领域的定量分析过程与方法，作为社会统计、数据分析的理论指导。考虑到许多研究者对 Stata 软件还比较陌生，本书前面章节系统介绍了 Stata 的一些基本特征以及数据的读入与描述。

第二部分为数据处理，包括变量的生成、处理，以及数据分组、合并、转换两部分。这是该书的重点，也是本书最具特色之处，许多有关统计软件应用的书籍均未涉及。而该技能也是目前定量研究者最为缺乏的。

第三部分为描述性分析，包括变量描述和图形的制作。前者是指获得变量的频数分布、均值等统计量，列联表的生成及解释等等。后者介绍 Stata 制图功能。与其他软件相比，Stata 的制图功能十分强大。

第四部分是推断性统计，包括参数估计、假定检验、方差分析、一元和多元线性回归（OLS）和二元非线性（logistic）回归等内容。

为了方便读者，本书附有光盘一张，内容包括：

（1）每章正文使用的数据及数据处理程序。

（2）每章复习思考题使用的数据及答案程序。

（3）1989—2006 年“中国健康与营养调查”（China Health and Nutrition Survey，简称 CHNS）部分公开使用数据，可供读者练习使用。CHNS 数据由美国北卡罗来纳大学和中国预防科学医学院联合采集。该调查旨在探讨中国社会和经济转型与计划生育政策对人们健康和营养的影响。关于该数据的详细情况和具体内容，请参考 http：//www. cpc. unc. edu/projects/china。

读完本书后，读者不仅可以掌握一定的社会统计分析原理和数据处理的技能，而且还可以掌握 Stata 软件的基本应用方法。当然，统计分析只是整个研究过程中的一个步骤，其具体实施还依赖于理论的指导。

本书的适用范围和读者对象包括社会学科的学生以及 Stata 统计软件的使用者。在国内社会学和人口学定量分析方法需求不断发展的情况下，希望本书能为专业教学和普及培训提供一本通用的教材，对国内定量研究方法的深入和发展起到推动作用。

21世纪人口学系列教材

目录

第一部分　定量分析研究过程与方法

第一章　社会科学研究方法与过程 …… 3
1.1　理论的作用 …… 5
1.2　提出问题 …… 6
1.3　生成假定 …… 9
1.4　收集数据 …… 13
1.5　分析数据 …… 15
1.6　检验假定 …… 21
1.7　陈述分析结果、避免误用分析结果 …… 22

第二部分　数据处理技术

第二章　Stata 入门 …… 27
2.1　特点与功能 …… 27
2.2　版本 …… 30

2.3 运行方式 …… 30
2.4 界面 …… 31
2.5 帮助系统 …… 39
2.6 语法和命令 …… 46
2.7 本书的体例 …… 49

第三章 Stata 数据的读入与熟悉 …… 52
3.1 log（记录）文件 …… 52
3.2 数据的记忆（存储）空间 …… 57
3.3 数据的读入 …… 60
3.4 数据的保存 …… 66
3.5 数据的类型与压缩 …… 67
3.6 数据库的描述 …… 69

第四章 变量的生成与处理 …… 87
4.1 变量的测量水平 …… 87
4.2 新变量的生成、规则及注意事项 …… 89
4.3 利用系统变量（下划线变量）生成变量 …… 92
4.4 生成字符型变量 …… 100
4.5 生成数值型变量 …… 101
4.6 生成分组变量 …… 105
4.7 生成虚拟变量 …… 107
4.8 egen 命令 …… 112
4.9 日期变量 …… 119
4.10 变量类型的转换 …… 121
4.11 给数据、变量和变量的属性贴标签 …… 124
4.12 重新命名变量 …… 129

第五章 数据的合并、转换与集合 …… 136
5.1 数据合并中的几个主要概念 …… 137
5.2 纵向合并——增加样本量 …… 139
5.3 横向合并——增加变量 …… 146
5.4 数据的转换（reshape） …… 171
5.5 数据的分组（group） …… 187

5.6 数据的集合（collapse 和 contract） …… 188

第三部分 描述性统计分析

第六章 数据的描述 …… 201
6.1 频数分布 …… 202
6.2 条件频数分布 …… 210
6.3 频数分布的常见错误分析及解决方法 …… 214
6.4 变量的中央趋势和离散趋势 …… 214
6.5 描述数值型数据统计量的其他方法 …… 219
6.6 列联表 …… 227

第七章 图形的制作与数据的描述 …… 243
7.1 散点图 …… 245
7.2 线图 …… 258
7.3 条形图 …… 265
7.4 直方图 …… 274
7.5 圆形图（饼图） …… 282
7.6 箱线图 …… 289
7.7 矩阵图 …… 292
7.8 图形的保存、编辑与合并 …… 294

第四部分 推断性统计分析

第八章 参数估计、假定检验与方差分析 …… 307
8.1 参数估计 …… 307
8.2 假定检验 …… 319
8.3 方差分析 …… 330

第九章 线性回归 …… 356
9.1 变量间的相关关系 …… 357
9.2 线性回归概述 …… 366
9.3 一元线性回归 …… 376
9.4 多元线性回归 …… 381

第十章　Logistic（非线性）回归 …… 412
10.1　基本原理 …… 413
10.2　回归模型 …… 413
10.3　模型的估计方法 …… 414
10.4　模型的应用与分析结果的解释 …… 415

部分复习思考题参考答案与提示 …… 421

参考文献 …… 448

后记 …… 450

第一部分

定量分析研究过程与方法

第一部分

常微分方程的定性与分析法

第一章

社会科学研究方法与过程

社会科学研究探究事情的真相，关注事物之间的联系及其原因。联系以多样形式而存在。它可以发生在同一个体的不同特征之间（如性别与收入），不同个体之间（如男性和女性收入的差异），个体与群体之间（如不同社会阶层的个体收入），个体与社会制度之间（如职场的性别隔离和收入不均）。事物之间的联系不是无序的，而是有其内在机制，科学研究就是要寻找并解释这种（些）机制。适当的研究理论，科学的研究方法，合理的研究步骤有助于有效地发掘事物之间关联的显性和潜在机制。然而，20 世纪 80 年代以来，随着西方现代社会统计技术和方法的传入，在定量分析方法得到推广与普及的同时，统计分析方法的滥用、误用和不规范现象也普遍存在（《人口研究》，2002）；理论、方法和包括定量、定性数据在内的经验材料分析脱节的情形也很严重。造成这些现象的原因是多方面的。既有对理论把握的欠缺，也有对研究方法掌握的不足，还有对研究过程和方法本身的不了解，等等。

本章旨在勾勒出社会科学（尤其是社会学和人口学）研究领域定量研究的过程与方法，探讨在此过程中需要遵循的行为规范，寻求整合研究理论与方法的最佳途径。我们知道，科学研究方法主要可以划分为两大类：演绎法（deduction）和归纳法（induction）。前者从一般到个别，即从（1）逻辑或理论上预测的模式到

(2) 观察检验预期的模式是否确实存在；换言之，演绎法是从“为什么”推延到“是否”。归纳推理从个别到一般，从一系列特定的观察中，发现一种模式，在一定程度上代表所有给定时间的秩序。这里介绍的过程和方法遵循从理论到实践的演绎规则。需要指出的是，两种途径的结合可以帮助人们寻求对事物更有力、更完整的理解（巴比，邱泽奇译，2005）。

研究过程是社会科学家在探求事物原理和关联时遵从的一系列行为；研究方法是研究过程实施的手段。一方面，方法寓于研究过程之中；另一方面，方法指导研究过程和步骤的实施。通过这些行为和方法，研究者最终得以回答研究问题。从演绎法的角度出发，图 1—1 描述了定量研究的 5 个基本步骤和过程：提出问题，生成假定，收集数据，分析数据，检验假定（Frankfort-Nachmias and Leon-Guerrero，2003）。这些步骤密切联系，相互制约，有时甚至重合，构成科学研究的完美形式，也是深化认识的最佳途径。

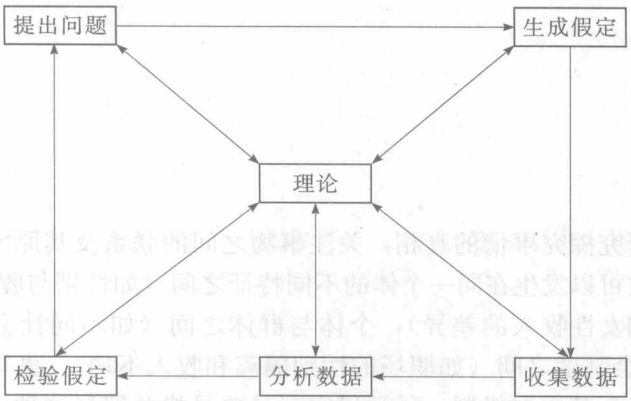

图 1—1　社会科学研究流程

注：双向箭头表示相互影响，单向箭头表示先后顺序。

特别值得关注的是居中的理论。如其所示，每一个研究步骤都与理论互动，既被理论影响，也检验理论的适应性。概而言之，理论处理的是科学的逻辑层面，问题的提出和假定的生成都是在理论的指导下进行的；收集和分析数据则是观察层面，而假定检验则是比较预期的逻辑和实际观察之间的吻合度（巴比，邱泽奇译，2005：12）。当然，由于人力、财力、物力的局限或其他人为和非人为因素，并不是所有研究都必须经历这些步骤。数据收集步骤往往被忽略，有时我们只能使用、分析他人收集的二手资料。

下面，我们主要以性别、教育、职业与收入的关系为主线，辅之以其他研究实例，通过对具体问题的剖析，首先阐述理论在科学研究中的重要性，然后系统

介绍每个研究步骤及主要注意事项。通过对这些问题的分析与探讨，我们希望能够引起社会科学研究领域的学者对定量研究方法的过程及规范性的广泛关注和重视。

1.1 理论的作用

理论是指人们由实践概括出来的关于自然和社会现象的系统结论，也是对两个或多个可观察的属性关系的解释。理论尝试着在观察到的现象和概念之间建立一种联系，解释为什么一种现象以一种特殊的方式与另一种现象相关联。因此，理论关心的是是什么（what is）和为什么（why），而不是应该（should be）如何。理论不是价值判断，而是对事物之间联系的系统的、言之成理的客观描述、总结和解释。由于种种条件的制约，社会科学领域的理论是不可以被证实的，而只可能被证伪（郭志刚等，2003）。

社会现象之间的联系是复杂多样的，只有在科学理论的指导下，在科学方法的帮助下，才能更好地找寻、把握事物之间的关联机制，二者相辅相成。然而，理论与方法脱节的现象十分严重。一方面，我们常常看到，一些定量分析忽视理论，流于数字的游戏。另一方面，我们也不难发现，不少理论研究缺乏实证经验的支持，成为理论空谈。因此，如图 1—1 所示，在科学研究中，理论始终处于核心地位，与每个研究步骤互动，指导每一步骤的实施，同时也被研究过程和分析结果检验、修正。这一描述再现了理论与方法之间的依存关系；二者如同车之两轮、鸟之双翼，是科学研究中不可或缺的成分。

理论因其解释的对象和构建的不同可以分为三个层次：宏观理论（grand theory，total theory，paradigm）、中观理论（middle range theory）和微观理论。

理论可以是抽象、宏观的。社会科学家往往尝试构建一个完整的理论体系，以便更为精确地解释事物原理和社会生活的一些共性特征，并最终揭示社会客体的本质和社会发展规律。这样的理论着眼于大的社会结构或体制。马克思的资本理论、辩证唯物主义和历史唯物主义理论，韦伯的官僚体制效率理论，帕森斯的功能理论，等等，都是人们耳熟能详的宏观理论的典型代表。

然而，宏观理论的提出往往需要数代人的努力和经验的积累。作为常人的我们，即便穷毕生精力，也难以在宏观理论方面有所建树。同时，由于宏观理论难以具体量化和检验，故在社会科学研究中，宏观理论与经验研究的脱节现象比较严重（李培林等，2005）。正是关注到了这一点，默顿（Merton，1968）提出了介于宏观范式和微观分析框架之间的“中观理论”。