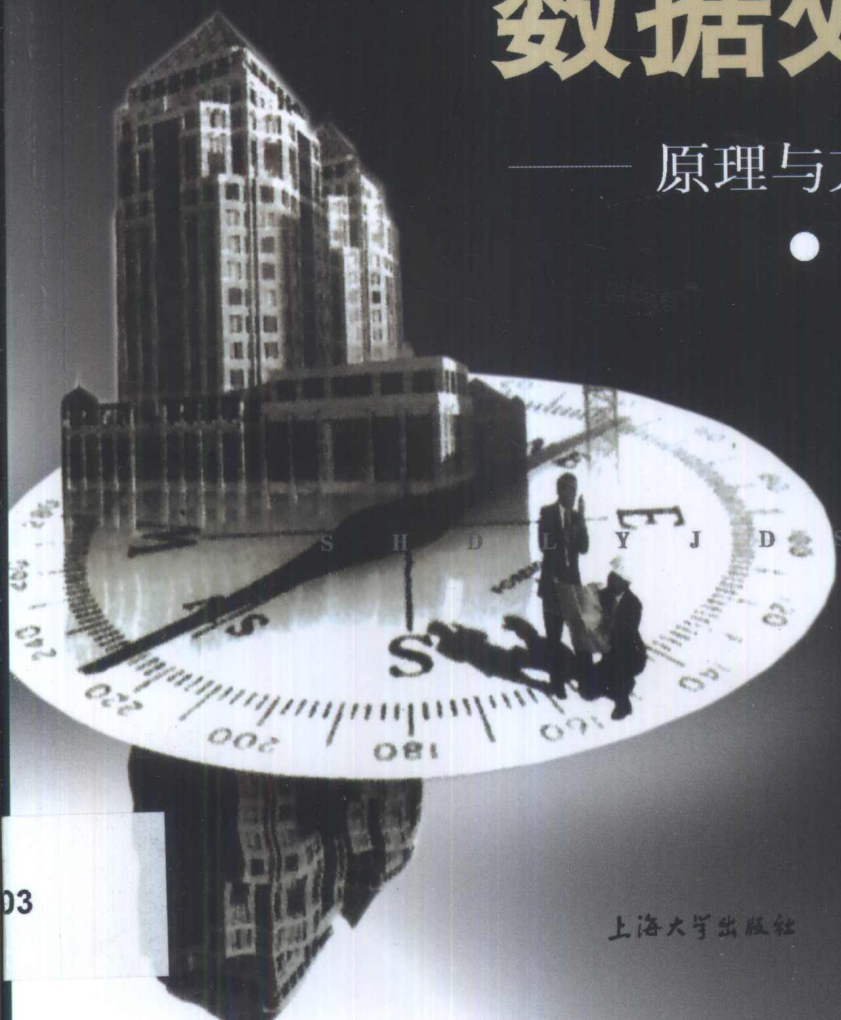
社会学与社会发展丛

社会定量研究的数据处理

——原理与方法

翁定军／编著



上海大学出版社

社会学与社会发展丛书
重点学科建设系列成果

社会定量研究的数据处理

——原理与方法

翁定军　编著

上海大学出版社
·上海·

图书在版编目(CIP)数据

社会定量研究的数据处理:原理与方法 / 翁定军编著. —上海:上海大学出版社,2002.10
(社会学与社会发展丛书)
ISBN 7-81058-481-2

Ⅰ.社... Ⅱ.翁... Ⅲ.定量社会学-数据处理
Ⅳ.C91-03

中国版本图书馆 CIP 数据核字(2002)第 077843 号

上海大学出版社出版发行
(上海市延长路 149 号 邮政编码:200072)
上大印刷厂印刷 各地新华书店经销
开本:890×1240 1/32 印张:9.5 字数:261 千字
2002 年 10 月第 1 版 2002 年 10 月第 1 次印刷
印数 1~2 100
定价:18.00 元

总　序

邓伟志

自邓小平提出社会学在中国“需要赶快补课”以后，上海大学文学院的前身——复旦大学分校便于1980年3月在国内率先设置了社会学系，并且同时创办了全国第一家社会学刊物——《社会》，随后又成立了社会学研究所以及若干个研究中心。20年来，社会学系的教师、社会学所的研究人员以及《社会》杂志的编辑，三方面共同结合，呕心沥血，编写了不少教材。有的教材为多所学校采用，有的教材很受学生欢迎。教材是教学的蓝图；教材是学生成长的阶梯；教材是激发学生学习热情的酵母。现在中国许多省市、世界上许多国家都有上海大学社会学系的校友，有的校友已成为国家栋梁之才。学生今日之高素质不能不认为同昨日之好教材有一定的正比例关系。

可是，随着光阴的流逝，再好的教材也有过时的地方。现在已经到了20世纪的最后一年的最后几个月。手再伸得长一点，我们已经可以与21世纪握手了。时间老人已经拿起大锤，准备去敲响21世纪的大钟了。

在21世纪，中国的社会主义建设的步伐将有长足的进展。在21世纪，中国的市场经济将更加繁荣，更加规范。在21世纪，多种所有制之间的构成将出现新的格局。经济的这些变化必然要作用于社会。何况经济本身也是社会的一个方面呐！

在21世纪，中国的社会结构将会变得更加匀称，更加合理；中国的社会变迁将更加迅速，更加有序。这些都需要我们社会学界用世纪眼光去观察，去审视，都需要我们在社会学教材中加以反映和提炼。社会

学只有聚焦时代、反映时代的义务，没有漠视时代、脱离时代的权利。时代在前进，实践在发展，我们编写教材的工作不敢有一丝一毫的松懈。

在21世纪，全球化的进程将大跨度地向前推进。社会学的学说、学派将层出不穷。包括那些在20世纪里未能引人注目的某些社会学观点，在21世纪说不定会在课堂上成为主课。据联合国教科文组织统计，社会学的分支学科约为110余个。可是，据中国学者的不完全统计，中国已有140余个社会学的分支学科。所有这些，恰是我们上海大学文学院社会学系同仁学习和借鉴的对象，恰是我们编写新教材的重要参考资料。

时代在呼唤新教材。实践在鞭策我们不断开拓进取，编出新教材。在编写过程中，大家兢兢业业，精益求精，边编写，边讲课，边听取意见，边修改。至于说是否做到了人们常说的"全新"，我们不敢妄言，似乎也无此奢望。我们只是在求新，未必做到全新。渴望同行多多给予指点，因为我们还要继续编写，继续出版，继续奋斗。

教材的编写只有起点，没有终点。

2000年6月10日

于上海大学A楼601室

前　言

无论从哪个角度讲，社会统计都称不上是一门独立的学科。它像其他应用性统计——教育统计、心理统计、经济统计等等一样，依附于数学分支学科——数理统计。任何新的统计技术、统计理论的出现和发展，全都仰仗数理统计，社会统计只不过是数理统计的理论和技术在社会研究领域中的应用。社会统计不仅缺乏自身发展的能力，还缺乏自身发展的动因，它不具有自身提出研究问题的可能性，它所解决的全部问题都是由社会调查这门方法学科提出的。它是运用数理统计的原理分析社会调查研究的结果。社会统计与其说是一门学科，不如说是社会定量分析的一种手段，是处理社会调查数据的一种技术。

称其为技术，并不意味着它在社会研究中的地位无足轻重。任何事物总有质和量两个方面，社会现象也不例外。以往的社会调查研究往往只注重定性分析，忽视定量分析。然而，定性与定量相结合的研究方法已经成为目前社会调查中的一种趋势，定量分析在社会研究中表现出了无可否认的优越性，人们利用统计技术建立各种数学模型，揭示数据后面隐藏的各种社会现象之间的关系、规律和发展趋势。定量分析以及定量分析的手段——社会统计技术，在社会研究的方法中已经取得了毋庸置疑的合法地位。

随着计算机技术的发展，各种统计软件的出现和普及，那些过去由于运算复杂而难以应用的统计技术在今天已经变得非常容易，社会统计在分析和处理社会调查数据中的作用和地位也显得愈益重要。可以说，离开了社会统计这门技术，对于那些数量浩大的调查数据根本无法

进行深入的分析。也同样可以说，离开了统计软件，那些复杂的统计技术根本无法应用。统计软件与统计技术已经融为一体。

有了统计软件，人们再也不用作手工运算了，再也不用去记忆那些复杂的统计公式了。鼠标轻轻一点，结果便在眼前。一切都变得那么容易，那么方便。然而，统计软件毕竟只是一种计算工具，无非是计算得更快更准。它不能代替人们对统计原理的理解，不会告诉人们应该使用哪种统计方法，更不会告诉人们如何解释、分析统计结果。

显然，统计软件的掌握离不开统计原理的掌握。确切地讲，离开了统计原理，不可能掌握统计软件。掌握了统计原理，却可以轻易地掌握统计软件。主次关系不能颠倒。

本书取名为《社会定量研究的数据处理——原理与方法》，实际上是把社会统计的原理方法与统计软件结合了起来，针对的是社会调查数据的处理方法和分析方法。重点在于阐述统计原理、统计方法和统计过程，强调说明每种方法的主要含义和适用范围。每种方法均配备了两种形式的例题，一种是传统的手工计算方法的例题。虽然在实际应用中已不可能再采用手工计算的方法了，它却能够展示统计方法的全过程，有助于说明和理解其中的统计原理。离开了统计过程，对统计方法和统计原理的理解将是不完整的。另一种是以统计软件为计算工具代替手工运算的例题，这有助于从实际应用的角度掌握统计方法。

本书采用的统计软件为 SPSS for Windows 10.0 版本。从思维形式讲，软件操作的学习和掌握是一种动作思维，离开了实际操作，不可能掌握一门软件的使用方法。本书对于 SPSS 的操作介绍只是引导性的，非常简略，重点在于从统计原理解释和分析 SPSS 的输出结果。

社会统计的原理存在于数理统计之中，其间涉及到了复杂的数学推演。本书力图能以通俗的方式表达其中的统计原理和思想，尽量回避那些过于复杂的数学演算和证明，以方便数学基础不太强的读者学习。然而，统计的基础毕竟是概率，不可能完全绕过某些高等数学和概率方面的知识，有些统计公式的数学表达式本身就极为复杂。其实，表达式的复杂并不一定意味着原理的深奥难懂，有时是复杂的数学符号带来了抽象性，也似乎带来了难度，稍加用心，其原理并不难以理解。

值本书出版之际，作者特别要表示对胡申生老师的衷心感谢。正是在胡申生老师的支持下，作者的社会统计专业知识得以提高；他也为本书的出版作出了很大努力。此外，也对陈友放老师给予的帮助表示谢意，本书附录中的各项概率数值表均由陈友放老师提供。

限于作者水平，书中难免会有各种不足，真诚欢迎各位同仁指正。

翁定军

2002 年 6 月

目　录

第一章
数据的初步整理

在社会定量研究中，我们会收集到许多数据，这些数据能够对我们的研究提供很有用的信息，帮助我们认识社会现象的内在规律，了解社会现象之间的关系，预测今后的发展趋势，等等。因此，这些数据是非常宝贵的。但是，这些数据所提供的信息，并不是一目了然的，必须对它们进行认真的、科学的整理和分析后，通过统计检验，才能取得可靠的结论。本章承接社会调查研究方法中的第二阶段，主要对数据进行初步整理的常用方法作一些介绍。为帮助我们认识和整理数据，先介绍数据的特点与分类。

第一节　数据的特点与分类

一、数据的特点

要整理社会调查中的大量数据，首先要了解数据有一些什么特点。数据的第一个特点是离散性，调查研究中的数据或观测数据都是以一个个分散的数字形式出现的。离散性表示数据在数轴上的变化是不连续的、间断的，数目总是有限的。第二个特点是波动性，波动性又称变异性。观测数据总是在一定的空间和时间范围内不断变化的，很难收集到完全相同的数据。即使有相同数据，也是很少量的。数据波动性的原因在于调查研究中存在着的各种误差。误差按其性质可分为三类。

随机误差：又称偶然误差。是由于研究中的一些偶然的却又不易控制的因素所引起的误差。在社会调查中，随机误差不仅包括调查者主观和客观上的不可控制的偶然因素引起的观测误差，也包括调查对象的一些偶然的不可控制的因素造成的误差，以及调查者与调查对象交互作用而导致的一些误差。抽样调查中的抽样误差也属于随机误差，它是由于样本范围与总体范围的不同而产生的误差。随机误差产生的原因是复杂的，其存在却是绝对的。随着观测次数的增加，随机误差的变化会表现出一定的规律性和特点，它总是围绕着被观测对象的真正值上下波动。随着试验次数的增加，随机误差的波动范围逐渐趋小，随机误差逐渐消除。

系统误差：是观测过程中服从某种确定性规律的误差，由确定性因素引起。其结果是数据往往偏向一端。在社会研究中，这种误差主要是由于研究者的偏见或被研究者的偏见造成的。比如，调查家庭收入，人们的回答往往低于实际值；中青年女性遇到年龄问题时，也可能倾向于把自己的年龄说得低一些；在家庭关系或夫妻关系、同事关系、上下级关系之类的问题中，人们的回答又往往会高于实际融洽程度。在抽样调查中，违背随机抽样的原则也会导致系统误差。系统误差不会随试验次数的增加而消除，但是，通过科学地设计问卷和严格按照随机抽样原则抽取样本，可以在一定程度上减少甚至消除系统误差。

过失误差：一般指在数据收集过程中由于人为的过失而造成的误差，如调查过程中听错、测错、传错、记错或整理中抄错等。在数据的统计整理过程中，必须鉴别、舍弃这些含有过失误差的数据，否则会严重影响计算结果的准确程度，得出不正确的结论。

在一组原始数据中，随机误差、系统误差和过失误差总是错综复杂地存在着，因而造成观测数据的波动性和变异性。在统计学中，称这些具有波动性的数据为变量。按一般习惯用英语字母 x，y 等表示，把一组数据记为 $x_1, x_2, \cdots, x_n$ 或 $y_1, y_2, \cdots, y_n$ 等。在不会引起混淆的地方，往往省略 $x_1, x_2, \cdots, x_n$ 中的下标，这样，x，y 既用以表示变量，也用以表示数据。

数据的第三个特点是规律性，观测数据虽然总是波动的，但这个波

动并不是杂乱无章的，而是在一定的范围内，呈现出一定的规律性。这个规律性虽然难以直观辨出，但是，对数据进行分析整理、统计检验后就显得很清楚。也正是由于数据具有规律性这一特点，统计检验才有必要和可能。

二、数据的分类

变量是概念的具体化，它以具体的测量数据表现出来。因此，数据的分类也就是变量的分类。根据不同的标准，可以对数据作不同的分类。

数据按其来源和由什么方法观测得到，可以分为计数数据和测量数据两类。

计数数据是指计算个数的数据，它具有独立单位，以事物的不同特性归属为标准。此类数据不具有数值的意义，只是一种分类符号，只能对其个数“计数”，而无法对其“数值”进行运算。比如性别、民族、职业等都属于计数数据，相当于后面的定类、定序数据。计数数据一般都取整数形式，除特殊情况外，不取小数和分数形式。

测量数据，一般是指借助于一定的测量工具或测量标准得到的数据，像身高、体重、温度等是用不同的测量工具得到的数据，爱好评价、智商等是根据一定的测量标准得到的数据。

按数据是否连续来划分，可分为连续型数据和离散型数据。上面所讲的计数数据大都属于离散变量，又称不连续变量，其数字形式是取整数，两个单位之间不适于再分成细小数目，因为比单位小的数目在观测中是不存在的，比如人，只能说一个人，两个人。两个单位之间是独立而间断的。

用测量工具得到的数据，一般都是连续型变量，它的单位可以划分为细微的数目，细微的程度可以达到不能看见而只能想象的程度，比如，长度可以分为千米、米、分米、厘米、毫米、微米等，即两个单位之间存在无限个数据。

在社会定量研究中，最常用的一种分类方法是按变量的测量层次划分，根据测量层次由低到高依次分为定类变量、定序变量、定距变量

和定比变量。

定类变量又称定类数据或类别数据，它是事物性质的划分，实质上只是一种分类体系。比如，性别、婚姻状况、出生地等，反映的都是事物的某一类性质，从变量的角度讲，都属于定类变量。若用"1"表示男性，"2"表示女性，此时的1和2只是表明类别的不同，是一种分类符号，不反映事物本身的数量状况，不具有数值上的意义。用数学语言讲，定类变量只能用等于或不等于表示，既不能用大于或小于来比较大小，更不能用加、减、乘、除对其作数学运算。

定序变量也称定序数据或等级数据。定序变量的数据具有某种逻辑顺序，但没有相等的单位，也没有绝对的零点。如学历、工作效率、等级评定、赞成程度、喜爱程度等，它们具有高低、大小、强弱的差异，我们据此可以对它们进行比较，但是不能说明它们的具体差异量，或者说，定序变量具有等于或不等于、大于或小于的关系，却不能进行加、减、乘、除的运算。

定距变量也称定距数据或等距数据。定距变量的数据是具有相等单位却没有绝对零点的数据。由于具有相等单位，就引入了数量变化的概念，因此，定距变量不仅能将变量区分类别和顺序，还可以确定变量之间的数量差别和间隔距离，或者说，定距变量不仅具有等于或不等于、大于或小于的关系，还能进行加减运算。智力测验中的智商和气温测量中的温度都属于典型的定距变量，我们不仅能比较智商高低或温度高低，而且能通过加减运算具体说明高低相差多少。定距变量开始真正显示了事物在数量方面的差异。但是，定距变量没有绝对的零点，因而不能进行乘除运算。以温度为例，比如，今天温度为6℃，昨天是3℃，我们可以说今天比昨天高3℃(6－3＝3)，却不能说今天温度是昨天温度的两倍(6/3＝2)，若昨天温度是零下3℃，更易看出定距变量不能进行乘除运算。温度计上的零摄氏度只是一个相对的概念，是对水开始结冰的临界点的规定，并不表示"没有"温度。

定比变量也称定比数据或比例数据。定比变量除具有上述三种变量的全部性质之外，还具有实际意义的绝对零点，所以，它的测量数据既可加减也可乘除，如收入、体重、年龄、出生率等都属于定比变量。拿

收入来说，如甲的收入为 1 800 元，乙的收入为 1 200 元，可以说甲比乙收入高 600 元，也可以说甲的收入是乙的 1.5 倍。收入为零是一个绝对零点，确实表示没有收入。

定距变量和定比变量在实际运用中有时很难区分。为简单起见，人们一般将这两类变量合并为一类，对定距定比不再加以区分。这样，四类变量减少为三类变量。SPSS 统计软件即采用三类划分法。本书也作三类划分。

此外，还有两点需要做一些说明：

(1) 高层次的变量包含了低层次变量的全部性质。因此，高层次的变量可以当作低层次的变量来处理。以年龄为例，年龄可以看作是定距变量(定距定比不再区分)，假如对年龄进行分组，12 岁以下称为童年，13～16 岁称为少年，17～28 岁称为青年……这样，年龄便成为定序变量。而低层次的变量却不能转化为高层次的变量。

(2) 变量类型的确定直接关系到统计方法的选用。对于不同的变量(数据)类型，应采用不同的统计方法。

在定量研究中，结论的正确性不仅依赖统计方法的正确性，更依赖数据的正确性，数据的正确性是结论正确性的前提。以定量方式研究社会现象，遇到的困难主要来自两个方面。一是数据的真实性，即数据是否真实地反映了客观实在。真实性问题不是依靠测量技术的完善能够得到完全解决的问题。另一困难是对社会现象的量化问题。对于社会现象往往缺乏科学的严谨的量化指标，其中的一个表现就是在社会调查中遇到的大量变量都是定类变量和定序变量，定类变量和定序变量的量化程度是很低的，从本质上讲，它们不是“数”，只是类别符号或等级的顺序，适用的统计方法相对也是比较少的。然而，这是一个技术性问题。随着测量技术的不断完善和统计方法的不断发展，这个问题会不断地得到逐步解决。

第二节　统计表与统计图

当社会调查研究的第二阶段调查阶段结束以后，我们会收集到大