厦门大学研究生方法论公共课教材

米红　张文璋　编著

实用现代统计分析方法及SPSS应用

专门适合非统计专业研究生及相关人士

融实用性、方法性、技术性、先进性、可操作性为一体

当代中国出版社

厦门大学研究生方法论公共课教材

实用现代统计分析方法及SPSS应用

米红　张文璋　编著

当代中国出版社

图书在版编目（CIP）数据

实用现代统计分析方法及 SPSS 应用/米红，张文璋编著.—北京：当代中国出版社，2004

ISBN 7-80170-295-6

Ⅰ.实… Ⅱ.①米…②张… Ⅲ.统计分析—软件包，SPSS Ⅳ.C819

中国版本图书馆 CIP 数据核字（2004）第 021401 号

责任编辑 刘海燕
装帧设计 吴家凯
出版发行 当代中国出版社
地 址 北京地安门西大街旌勇里 8 号
邮政编码 100009
发 行 部 电话（010）66572157
印 刷 北京地质印刷厂
开 本 880×1230 毫米 1/32
印 张 11.75 印张 342 千字
版 次 2004 年 5 月第 1 版
印 次 2004 年 5 月第 1 次印刷
定 价 26.00 元

前　言

21世纪，随着我国改革开放的进一步深入和社会经济的全面发展，我们面临着严峻的挑战和难得的发展机遇。在逐步完善的社会主义市场经济大潮中，要成为一名弄潮儿，必须掌握定量分析方法，精通定量分析与定性分析相结合的技能。实用现代统计分析方法已广泛应用于人文社会科学、管理科学，因此文科学生需要有一本既比较全面又针对性强的关于统计分析方法以及计算机统计软件使用方法的实用教材。目前国内市面上出售的文科用统计分析方面的教材远远不能满足这一要求，而理工科用的统计分析教材虽然其统计分析方法比较全面，但是没有很好地、有针对性地解决这些学科中许多复杂的实际问题，而文科学生又无法接受高深的数学知识。《实用现代统计分析方法与SPSS应用》正是适应了这一需要，为非统计专业的文科学生而写的。

本书的特点是实用性较强，内容广泛，并有所侧重。强调对实用统计分析方法基本思想的理解和应用，培养使用统计软件SPSS的能力，把统计分析方法与SPSS紧密结合在一起。对各种统计方法的原理只进行通俗的、描述性的说明，不作严格的数学推导。除了第九章和第十章需用到一些简单的矩阵代数外，其余各章所用的数学知识都很简单。

在本书的写作过程中，始终得到厦门大学研究生院副院长陈甬军、曲晓辉给予各方面的鼓励和支持。本书是作者在为厦门大学1997～2002级研究生教学的基础上总结而成的，在编写过程中，吕谭华、杨绮、顾慧慧等同学为本书的练习题提供了许多宝贵资料。厦门大学高

等教育研究所2002级研究生周仲高和郭书君同学为本书的案例及统稿提供了较多帮助，当代中国出版社刘海燕为本书的出版给予了大力支持，在此一一表示感谢。

由于编者的学识和水平有限，书中错误之处及不妥之处在所难免，恳请读者批评指正。

米红　张文璋

2004年4月1日于厦门大学

目录

第 1 章

概　论

第一节　市场经济呼唤统计学

许多人简单地认为统计（Statistics）就是收集数字，这仅仅是统计学的原始意义。现代统计学已远远超出了这个范围，发展成为广泛应用于经营管理、社会科学、自然科学等领域的科学方法。它是研究客观事物数量特征和数量关系的方法论学科，能够告诉人们如何打开几扇窗口去探索一个未知的世界，教会人们怎样用一种新的方式来思考问题，因此统计学是一门很实用的学科。

统计作为一种强有力的定量分析方法，在社会经济、政治、生活等领域得到了广泛的应用，并起着日益重要的作用。大至国家的宏观决策，小至企事业单位的微观管理，都离不开统计的应用。现代市场经济对统计信息的需求急剧增加，同时也对统计理论与方法提出了更高的要求。

面对 21 世纪，我国的人文社会科学肩负着时代的重托。社会发展问题、经济可持续发展问题、国际竞争力问题、金融风险问题、保险精算问题、人口与社会保障问题、环境保护问题等等，这些都迫切地等待着我们去深入地研究。要解决这些问题，时代要求我们必须抛开偏见，正确理解与批判地吸收建立在发达商品经济基础上的外来文化，加强数学方法、统计学方法的学习，提高我们的定性分析与定量分析相结合的能力。这样，中国才会在新的世纪里大步赶上世界发达国家。

第二节　统计学的研究对象及其学科分类

一、统计学的研究对象

1992年11月，国家技术监督局正式批准统计学为一级学科，国家标准局颁布的学科分类标准已将统计学列为一级学科，1998年教育部进行的专业调整也将统计学归入理学类一级学科。建设一级学科统计学的构想反映了统计学学科建设的内在要求，使它逐渐符合国际统计学发展的大趋势。所谓一级学科统计学，是指研究搜集和分析数据、研究客观事物数量特征和数量关系的方法论科学。一级学科统计学首先是一门方法论，它是研究客观现象（包括自然现象和社会现象）数量特征和数量关系、具有明确对象的方法论科学。统计方法论的性质，是指它作为一门认识方法论科学，为人们提供一套从不确定的现象中探索现象规律性的理论和方法。这里作为统计学研究对象具体体现的“数据”，是指进行各种统计（指统计工作）、计算、科学研究或技术设计等所依据的数值。

统计数据所具有的不同特点，使得统计学百花园色彩纷呈，各具特色。数据中的实验数据主要来自自然技术现象，如对产品配方检验得到的数据等等，这类数据大多在可控条件下通过物理测量取得，这类数据的搜集和整理工作并不复杂，研究的重点在于数据分析。另一类是观察数据，它主要来自社会经济现象，如国内生产总值（GDP）数据、某年度的货币购买力数据等等。由于社会经济现象的复杂性，尤其是不能通过一定条件下的物理或化学实验进行研究，致使观察数据的搜集往往十分困难，统计学不仅要研究观察数据的整理、分析技术，而且要花很大力气研究观察数据的调查搜集技术。正因为实验数据和观察数据有不同特点，所以以实验数据作为研究对象的自然技术统计学，如生物统计学、统计力学等等；以观察数据作为研究对象的社会经济统计学，如农业统计学、工业统计学等等，就表现出很不相同的特点。社会经济统计学利用统计指标、统计分组方法，不厌其详地研究数据搜集的技术，研究资料来源、指标口径和计算方法，至于

数据整理、尤其是数据分析的技术，则由于社会经济各专门统计的共同特点，出于简化篇幅的考虑，一般安排在社会经济统计学原理中做统一研究。自然技术统计学的生物统计学等等，与社会经济统计学的农业、工业统计学则恰恰相反，它的研究重点往往放在对数据所做的各种分析上，至于数据搜集、整理的技术，则考虑到自然技术各专门统计所具有的共同特点，一般在自然技术统计学原理的数理统计学中作简要讨论（之所以作简要讨论，是因为实验数据的搜集和整理远比观察数据的搜集整理简单）。从上面的分析中不难看出，自然技术统计学和社会经济统计学本没有不可逾越的鸿沟，两者只是由于研究对象所具有的不同特点，才产生了不同的理论体系和学科特色。建设一级学科统计学的构想，兼容自然技术统计学与社会经济统计学，反映了统计学发展的内在要求，对促进自然技术统计学和社会经济统计学各自的发展，都具有重要的意义。

二、统计学的学科分类[①]

统计学作为一门研究客观事物数量特征和数量关系的方法论科学，其内容构成错综复杂，既有层次性，又有交叉性，所以对其学科的分类迄今未得到合理的解决。较为流行的划分是把统计学分为社会经济统计学和数理统计学，或者分为描述统计与推断统计。这些分类都无法完全包括现代意义上的统计学内容，是不全面的。与一级统计学相对应，我们把统计学分为理论统计学、应用统计学与其他统计学等（如图 1－1 所示）。

理论统计学包括各种统计基础理论，又可以分为描述统计学和推断统计学。描述统计学指以总体全面资料或非随机性局部资料为基础的统计理论与方法体系，包括统计总体论（有关总体、指标和分组等理论）、统计设计、统计调查、统计整理、统计指数、动态分析理论、统计平衡理论、统计数据库等等，不同于仅研究如何整理和概括大量数据的“描述统计学”。推断统计学指依据随机样本推断总体特征的

① 杨灿：《统计学基本问题研究》，《统计研究》，1993 年第 3 期；黄良文、黄沂木：《大学科统计刍议》，《统计研究》，1995 年。

理论与方法体系，也就是数理统计学，它又可以分为理论数理统计学和应用数理统计学。理论数理统计学侧重于统计方法的数理基础，包括概率论、经典统计理论、贝叶斯理论、统计判决理论等。应用数理统计学（现代意义上的数理统计学）则侧重于统计方法的应用形式，包括抽样技术、试验设计、相关分析、方差分析、多重应答分析、多元统计分析、序贯分析、线性统计模型、时间序列分析、非参数统计等。这里的描述统计学与推断统计学并无“普通统计学”与“高级统计学”之分，实际上，推断统计学的某些内容是非常初等的，而描述统计学中的某些方法（如统计指数理论）却具有相当的理论深度和复杂性。

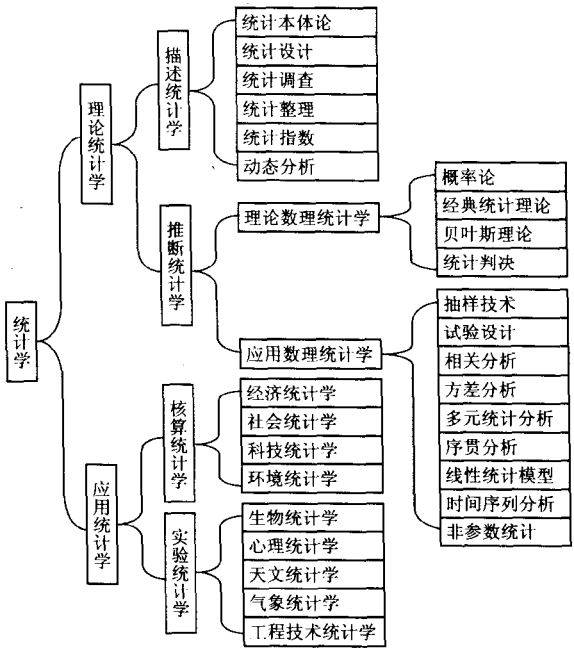

图 1-1　统计学分类

应用统计学只涉及某一特定现象领域的统计研究，又可以分为核算统计学和实验统计学。核算统计学是通过核算手段研究社会现象及

其过程的数量特征或统计规律性的理论与方法体系，包括经济统计学、社会统计学、科技统计学、环境统计学等等。而实验统计学是运用实验手段研究自然现象本身及其过程的数量特征或统计规律性的理论与方法体系，包括统计物理学、生物统计学、天文统计学、气象统计学、心理统计学、农业试验统计学、工程技术统计学等。

除了理论统计学和应用统计学外，还有统计史学、统计法制学、比较统计学等其他统计学科，以及经济计量学、保险精算学、运筹学、信息论等边缘学科。

从统计学的学科分类可以看出，统计学的内容是十分丰富的，其研究和应用的领域非常广泛。本书主要是为非统计专业的学生和统计工作者提供一本关于实用统计分析方法的读物，所以，主要包括了应用数理统计的一些内容。本书强调统计分析方法的基本思想和应用条件，培养用计算机进行统计计算的能力，并希望通过案例分析提高学生的解决实际问题的能力。

第三节　实用统计分析方法概述

一、变量（Variable）的分类

要进行统计分析，离不开统计数据。在搜索数据之前，必须首先了解数据的种类。数据涉及到变量的取值，通常用变量的取值来描述数据。变量可按多种方法分类，这些分类有助于选择适当的统计分析方法作进一步的分析与研究。下面按三种方法对变量进行分类：按间隙分类、按作用分类和按测量尺度分类。

（一）按间隙（Gaps）划分

根据一个变量紧挨着的两个观测值之间是否有空隙（缺口）划分，可以把变量分为两类：离散型变量（Discrete variable）和连续型变量（Continuous variable）。如果一个变量的观测值之间有空隙，该变量称为离散型变量，否则称为连续型变量，如图 1－2（A）所示。更准确地说，当一个变量的任意两个可能取值之间没有其他取值时，该变量是离散的；当一个变量的任意两个可能取值之间还有其他可能取

值时，该变量是连续的。例如，性别（设男性取值为 0，女性取值为 1）、企业数目、分组情况（设 A 组取值为 1，B 组取值为 2 等）等为离散型变量；身高、体重、血压、GDP 等为连续型变量。

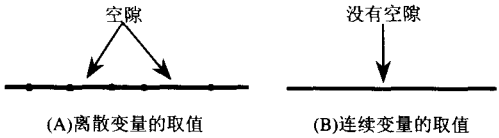

图 1-2　离散型变量与连续型变量

需要指出的是，由于分析的需要，离散型变量经常作为连续型变量处理。而连续型变量也可以作为离散型变量处理，如可以把“血压”变量分为“低”、“中”、“高”三组变为离散型变量。

（二）按作用划分

根据一个变量在分析时的作用，可以把变量分为因变量（Dependent variable）或自变量（Independent variable）。如果一个变量由其他变量来描述，该变量称为因变量或反应变量（Response variable）；如果一个变量与其他变量一起用于描述因变量，该变量称为自变量或预测变量（Predictor variable）。例如，在分析家庭收入、性别等因素对消费支出的影响时，收入变量和性别变量是自变量，消费支出变量是因变量。

一个变量是因变量还是自变量，与统计分析的目的有关。同一个变量在某种分析中作为因变量，而在其他分析中可能作为自变量。

（三）根据测量尺度划分

根据变量测量精度不同，可把变量由低到高分为四种尺度：定类变量、定序变量、定距变量和定比变量。

1. 定类变量

定类变量又称为名义（Nominal）变量。这是一种测量精确度最低、最粗略的基于“质”因素的变量，它的取值只代表观测对象的不同类别，例如“性别”变量、“职业”变量等都是定类变量。定类变量的取值称为定类数据或名义数据。定类数据的共同特点是用不多的

名称来加以表达，并由被研究变量每一组出现的次数及其总计数所组成，这种数据是枚举性的，即由计数——而得。惟一适合于定类数据的数学关系是“等价关系”。因而，在定类数据中，同一组内各单位是等价的，同时若更换各不同组的符号并不会改变数据原有的基本信息。因此，最常用来综合定类数据的统计量是频数、比率或百分比等。

2. 定序变量

定序变量又称为有序（Ordinal）变量、顺序变量，它的取值的大小能够表示观测对象的某种顺序关系（等级、方位或大小等），也是基于“质”因素的变量。例如，“最高学历”变量的取值是：1 - 小学及以下、2 - 初中、3 - 高中、中专、技校、4 - 大学专科、5 - 大学本科、6 - 研究生以上。由小到大的取值能够代表学历由低到高。定序变量的取值称为定序数据或有序数据。适合于定序数据的数学关系是“大于（>）”和“小于（<）”关系。在定序数据中，同一组内各单位是等价的，相邻组之间的单位是不等价的，它们存在“大于”或“小于”的关系。而且，并进行保序变换（或称单调变换），则不改变数据原有的基本信息即等级顺序。最适合用于综合定序数据取值的集中趋势的统计量是中位数。

3. 定距变量

定距变量又称为间隔（Interval）变量，它的取值之间可以比较大小，可以用加减法计算出差异的大小。例如，“年龄”变量，其取值60与20相比，表示60岁比20岁大，并且可以计算出大40岁（60 - 20）。定距变量的取值称为定距数据或间隔数据。定距数据是一些真实的数值，具有公共的、不变的测定单位，可以进行加减乘除运算。定距数据的基本特点是两个相同间隔的数值的差异相等，例如，年龄的60岁与50岁之差等于40岁与30岁之差。对于定距数据，不仅可以规定“等价关系”以及“大于关系”和“小于关系”，而且也可以规定任意两个相同间隔的比值或差值。如果将每个数值分别乘以一个正的常数再加上一个常数，即进行正线性变换，并不影响定距数据原有的基本信息。因此，常用的统计量如均值、标准差、相关系数等都可直接用于定距数据。

4. 定比变量

定比变量又称为比率（Ratio）变量，它与定距变量意义相近。细微差别在于定距变量中的“0”值只表示某一取值，不表示“没有”。例如，人的身高就是一个定比变量，如果身高值为“0”米，则表示这个人不存在。而定比变量的“0”值表示“没有”。而在测定温度的摄氏表中，0℃并不表示没有温度，因为还有在零点以下的温度。定比变量的取值称为定比数据或比率数据。定比数据也同样可进行算术运算和线性变换等。通常对定距变量和定比变量不需再加以区别，两者统称为定距变量或间隔变量。

一般地，定类变量和定序变量用于描述定性数据，属于定性变量；而定距变量和定比变量用于描述定量数据，属于定量变量。

同其他分类标准一样，一个变量在不同分析中可当作不同尺度的变量。例如，“年龄”在某些分析中（如回归分析）当作定距变量，而在另外一些分析中（如方差分析）可通过分组作为定类变量处理。

另外，较高尺度的变量包含了较低尺度变量的性质。定序变量包含了定类变量的所有特征，定距变量同时包含了定序变量和定类变量的特征。这种性质允许在分析数据时把一些较高尺度变量作为较低尺度变量处理。例如，定距变量可当作定类变量或定序变量看待，而定序变量可作为定序变量分析。

以上通过三种不同方法对变量进行分类。这些分类是可以重叠的。一个变量可能是离散型变量、自变量、定类变量（如“最高学历”），也可能是连续型变量、因变量、定距变量（如“血压”）。按间隙分类和按测量尺度分类的重叠如图 1－3 所示。

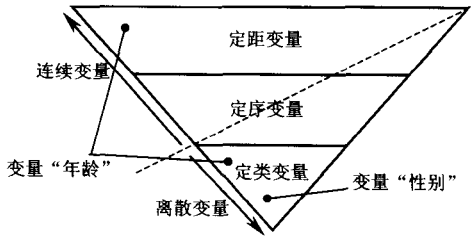

图 1－3　变量分类的重叠

从图 1－3 中可以看出，定类变量必须是离散变量，而定距变量和定序变量可以是离散变量或连续变量；连续变量必须是定序变量或定距变量。例如，变量“性别”是离散变量又是定类变量；变量“年龄”可当作定距变量、连续变量，也可以作为定类变量、离散变量。而自变量与因变量是根据分析目的而不是按变量本身性质来划分的，所以图 1－3 中没有包括这种分类。

二、统计分析方法的分类与选择

对数据进行统计分析时，选择正确的分析方法是非常重要的。选择统计分析方法时，必须考虑许多因素，主要有：(1) 统计分析的目的；(2) 所用变量的特征；(3) 对变量所作的假定；(4) 数据的收集方法（即抽样过程）。选择统计分析方法时一般考虑前两个因素就足够了。

(一) 根据统计分析目的不同进行分类

统计分析方法根据统计分析目的的不同，可以分成四大类：相关分析方法、结构简化方法、分类分析方法、预测决策方法[①]。

(二) 根据变量特征的不同进行分类

根据变量的分类不同分类方法，把变量分为因变量、自变量以及定量变量、定性变量，可把统计分析方法一一进行归类（如表 1－1 所示），这是正确选择统计分析方法的一种有效方法。

表 1－1　统计分析方法分类表

变量类型		统计分析方法	统计分析目的
因变量	自变量		
定量	定量	回归分析（或线性模型）、相关分析	描述一个或多个自变量与一个因变量之间的因果依存关系，或变量之间的相关关系。
定量	定性	T 检验、方差分析	描述一个连续型因变量与一个或多个定类自变量之间的关系。
定量	定性、定量	协方差分析（或线性模型）	描述在控制了一个或多个连续型自变量的影响下一个连续因变量与一个或多个定类自变量之间的关系。

① 详见何晓群编著：《现代统计分析方法与应用》，中国人民大学出版社，1998 年。