统 计 应 用 研 究 丛 书

现代数据分析技术

沈学桢 主编

立信会计出版社

·统计应用研究丛书·

现代数据分析技术

XIANDAI SHUJU FENXI JISHU

沈学桢 主编

立信会计出版社

图书在版编目(CIP)数据

现代数据分析技术/沈学桢主编. —上海:立信会计出版社,2005.1
ISBN 7-5429-1376-X

Ⅰ.现... Ⅱ.沈... Ⅲ.统计分析 Ⅳ.C813

中国版本图书馆 CIP 数据核字(2005)第 003744 号

出版发行	立信会计出版社
经　销	各地新华书店
电　话	(021)64695050×215
	(021)64391885(传真)
	(021)64388409
地　址	上海市中山西路 2230 号
邮　编	200235
网　址	*www.lixinaph.com*
E-mail	*lxaph@sh163.net*
E-mail	*lxzbs@sh163.net*(总编室)

印　刷	上海申松立信印刷厂
开　本	890×1240 毫米　1/32
印　张	7.375
插　页	2
字　数	194 千字
版　次	2005 年 1 月第 1 版
印　次	2005 年 1 月第 1 次
印　数	3 000
书　号	ISBN 7-5429-1376-X/F・1246
定　价	14.00 元

如有印订差错　请与本社联系

前　言

对每位统计专业课的任教教师而言，拥有一本既能引发学生兴趣，又能很好贯彻教学要求的教材，恐怕是一种普遍的热切愿望，怀着对这种愿望的追求，我和教研室的同仁们，在经过了若干年认真的资料准备、反复的课堂实践之后，终于完成了《现代数据分析技术》一书的编写工作。

在本教材编写之前，我们在学习、研究国内外同类和相关教材的过程中，结合本身的教学需要及学生的实际情况，逐渐确立了本书的指导思想：为那些已初步掌握统计概念、基本方法的本科学生，提供进一步数据分析的思路、方法及途径；同时研究国内外教材的实践也提示我们：一本受到教师、学生欢迎的教材，必定是系统性、科学性、适用性结合得恰到好处的一种教学和学习指南。因此我们全体参编人员，全力以赴地在本书中追求以下特色：

1. 教学内容的系统性。我们在汲取国内外多种同类及相关教材精华的基础上，本教材在内容的取舍、繁简的安排上，既注意与学生已掌握的统计基础知识的衔接，又顾及后续课程及学生将来职业发展的需要，做出了比较合理的安排。比如在基本统计分析方法这一章中，我们注意既精练且较全面地回顾了描述性统计方法在今天数据分析中的功能作用、应用条件，同时也考虑安排了相当篇幅来介绍现代统计分析潮流中涌现的探索性数据分析的新思路、新理念等，使学生对数据分析方法的整个发展过程有了一定的了解及理解。

2. 研究方法的科学性。近几十年来，由于计算机、网络等高新技术的迅速发展，使得一些计算工作量大、计算过程复杂的数据处理工作得

以顺利、方便地展开。本教材顺应这种潮流，通过介绍统计分析通用软件 SPSS 的使用，将数据分析中的定性分析思想与计算机技术在定量分析方面的优势相结合，使学生对运用现代化分析手段对数据进行有效、高速处理的理念及过程，有一定的感性、理性的体会及认识。

3. 课堂教学的适用性。长期的教学实践使我们深切感受到，一本满足“方便教与学”的教材一定具有较广的适用性。因此我们在设计教材体例，安排统计理论背景介绍与计算机软件操作程序的比例方面，充分考虑到教材内容的深度、广度与教学资源、教学设施、允许课时的匹配。尤其在案例的选编时，尽量选择涉及社会学、经济学、管理学、心理学众多领域中短小精悍的案例，以便突出、方便教师在有限的教学时间内从容讲解、演示案例，方便学生由浅入深，逐步接受统计思想、统计技术的特点。

本教材以经济类、工商管理类专业本科生为主要对象，亦可作为统计专业专业课、选修课的教材或参考教材。另外，由于介绍统计方法时基本不涉及数学原理的推导、求证，所以还方便具有初等数学基础的读者作为自学参考。

本书各章编写的分工如下：第一、第二、第三、第六章，由沈学桢编写；第四章，由徐静编写；第五章，由唐庆银编写；第七章，由卫志红编写；第八章，由彭江燕编写。沈学桢负责全书的总纂和定稿工作。编辑蔡莉萍女士在本书的编写和出版过程中，提出了许多建设性的意见，在此一并表示感谢。由于我们的水平有限，书中的不当之处难免，恳请读者批评指正，以共同推进数据分析方法教学的研究。

沈学桢

2005.1

目　　录

第一章 总 论

第一节 数据分析概述

一、数据分析意义

统计作为一门研究“数据的收集、整理、分析”的科学，其目的是为了通过分类与测度取得的信息来理解和认识现时世界。近几十年来，随着计算机及其他高新技术的发展及介入，数据收集的速度、广度均有了长足的发展，并由此引发了对数据分析技术的更新、扩充以及对分析结果的效度、信度进行评估的迫切要求。

传统的数据分析技术，建立在经典理论及众多假设的基础上，对数据分布的形态、结构等，有着严格的要求，近似在进行一种数据实验，符合这种实验要求条件的数据往往不多；而近几十年来统计学界所讨论的数据分析技术，其本意更在乎于完全“让数据说话”，即用一种比较宽松、灵活的方式，对数据不作任何分布上、结构上的限制，甚至对建模数据的要求也更为宽松，为的是使数据更符合生活中的实际情况，研究的结论更客观，研究的结果更具有实际意义。

尽管数据分析技术作为统计学中的一个新的研究领域、研究方向，其基本理论还有待于完善，其正式定义还在统计学家们的讨论之中，但其在统计研究中所表现出来的最大限度地将复杂理论转化为应用工作者可以接受的方法的努力以及尊重实际数据的追求，预示着数据分析技术广阔的发展空间及应用前景。

二、数据分析工作过程

在实际生活中，往往需要结合一个具体的社会经济问题，来开展统

计分析工作。一般而论，统计分析过程需经历以下几个阶段：

（一）理解问题、提出问题阶段

对一个具有一定工作经验、一定统计、业务背景知识的分析人员来说，善于将某个经济问题转化为统计问题，是搞好统计分析的基础。我们在工作中所面临的大量有待解决的问题当中，有相当一部分是可以转化为统计问题或用统计方法来处理的，由于这些问题的表现方式千姿百态，形成的原因错综复杂，解决问题的线索不像教科书中的习题那样清晰、一目了然，而更需要我们用统计思维的方式来思考问题的实质，并将其用统计的语言来表达，用统计的方法来解决。这阶段主要工作包括：考虑数据的实际背景，数据分析要达到的目的，如何定义变量，可能适用的统计方法及有否可利用的物理定律、经验公式、先验信息、历史资料等。

（二）有效收集统计数据阶段

要获得有意义的结论，数据的质量及数量十分重要。常用的数据收集方式有：

1. 实验设计，注意解决数据的随机性问题。

2. 抽样调查，注意解决问卷的设计、样本的代表性及无应答及假应答引起的误差。

3. 利用历史数据及研究资料，注意解决资料的适用性及调整问题。

往往可以结合使用多种收集数据的方法，并注意其与以后的分析方法的匹配。

（三）数据分析阶段

此阶段工作又可分成两步：

1. 考察数据，主要检查数据质量，概述数据特征。包括：了解数据背景，观察数据有否错误、异常、缺失，校正数据中错误；用探索性方法计算某些统计量及配以合适的图形，反映数据的位置特征、离散特征；对需要建立新变量的数据进行适当数据变换，使数据更符合特定分析方法的假设。

2. 数据分析，主要根据分析要求与目的，构造合适的模型来反映问

题。包括:按原则建立一个尽可能简单的模型,对模型的参数作出估计;通过残差检查模型的拟合程度,检查模型的有关变量的假设是否得到满足,根据要求修改模型。

(四)出具报告、表述结论阶段

将统计分析的结果使用简明、准确的语言作出报告。主要包括:分析所用背景材料的来源;分析过程中使用的具体方法,其数学依据及推导过程,分析者的结论、意见及建议。报告的主体部分文字要简洁、清晰、合理,符号要加以说明;文中的图表数值要准确,逻辑关系要清楚。

三、统计分析流程图示

统计分析流程如图 1-1 所示。

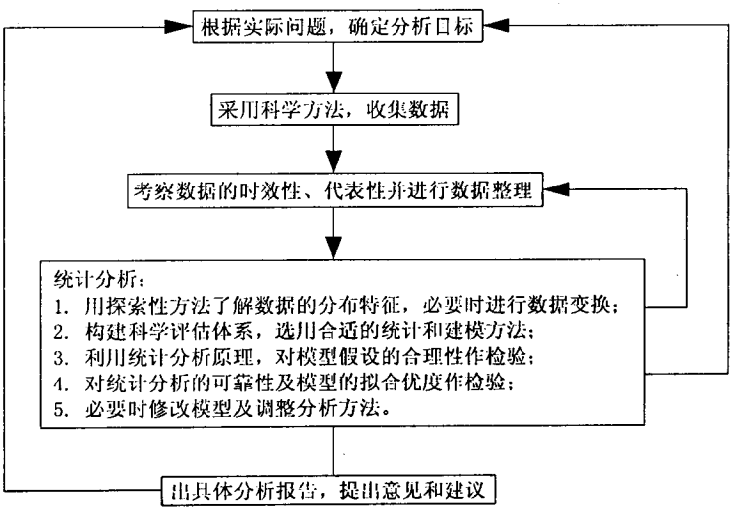

图 1-1

第二节 统计测量的基本概念

一、统计测量的意义

统计测量是用一套符号系统去描述某个被观察对象某种属性的过

程。它具有客观性、数量性、科学性等特点。

二、统计测量的基本概念

（一）变异

变异（Variety）指现象属性之间的差异，是统计学得以存在的前提。统计学亦可称是一门研究变异的科学。例如，人之间若无性别差异，就没有“性别”的概念，也就没有了研究性别的需要。统计学的存在意义表明：统计就是专门研究社会现象、自然现象在同一或不同时间、地点、条件下现象的差异的。

（二）变数

变数（Variable）指现象间某一属性差异的量化表现。它有两部分组成：即属性（即变量名）和数值（即变量值）。变数可以表示变量某一属性的量化程度，其数值需用特定的计量尺度来测定，如3个学生的身高分别为1.70米，1.66米，1.56米。

（三）自变量和因变量

对两个或两个以上变量间的关系进行分析时，我们将两个变量中的原因变量称为自变量（Independent Variable），即不受其他因素影响而发生变化在前的变量；将两个变量中的结果变量称为因变量（Dependent Variable），即受自变量变化影响而跟着发生变化的变量。

（四）离散变量和连续变量

离散变量（Discrete Variable）是指变量数值的个数是有限的，数值与数值之间无法找到更小单位的数值，如汽车的产量、家庭子女数等；连续变量（Continuous Variable）是指变量的取值个数是无限的，它们可能取值的范围能连续充满某一区间，例如身高、体重、产值等。

三、统计数据的种类

统计测量的结果即为表明现象品质、数量特征的数据。数据是描述现象测量结果的数值。

数据的种类与统计处理方法的选择密切相关。数据按来源、表现形式、性质，可以有多种分类形式。

（一）点计数据和度量数据

1. 点计数据是指通过计数得到的数据。例如，企业数、员工数、设备台数。

2. 度量数据是指使用一定工具按一定标准测量所得到的数据。例如，用量表测得的学生的智商，用秤测得的产品的重量等。

（二）定性数据和定量数据

1. 定性数据又称品质数据，是指反映现象某一属性、特征的数据。定性数据用来表明现象的属性特征，其表现形式通常为类别。这类数据是由定类尺度、定序尺度计量所得到的（见下节）。对定性数据，合适的统计处理方法往往是计算各组的频数或频率。

2. 定量数据又称数量数据，是指反映现象数量多少、特征的数据。定量数据用来表明现象的量化程度，其表现形式通常为数值。这类数据是由定距尺度、定比尺度计量所得到的（见下节）。对定量数据，往往有很多种合适的统计处理方法。

四、统计测量的尺度

在对定性数据、定量数据进行统计分析之前，必须首先解决对定性、定量变量的量化测定问题。变量量化的测定直接关系到统计方法的选择。统计学中将这种变量的量化测定称为变量的测量等级或测量尺度。美国统计学家、社会学家史蒂文斯（S. S. Stevens）1968 年按照变量的性质和数学运算的特点，将变量的测量尺度分成以下四种：

（一）定类尺度

定类尺度（Nominal Measurement）也称名义测定，是最粗略、等级最低的测定层次。它是用特定的编码来表示事物按品质标志进行的分组。例如，人口统计中的按性别分组，市场调查中的按地区分组等。为了便于计算机识别，可以将男性的编码定为 1，女性的编码定为 0，或者反过来定义。这里的编码没有任何量上的意义，除对每类变量可以计数，即计算每类变量发生的频数、频率外，这些编码所代表的各个类别之间关系对等，不能区分大小，也不能进行任何数学运算。

定类测定必须遵循两个重要的方法原则：①互斥原则，即每一个单位只能归入某一类中，不能同时归入两类。②穷尽原则，即所有被研究的单位都可以归入恰当的类中，没有一个单位无从归属。例如，人口调查中，人按性别分类，要么归入男性，要么归入女性，不可能同时归入两类，也不可能无所归属。

（二）定序尺度

定序尺度(Ordinal Measurement)也称序次测定，是一种不仅可以对事物进行分类，而且可以按某种标准对事物进行排序，进而根据排序确定类别大小、优劣的测定。例如，企业可以根据产品的质量，将一等品、二等品、三等品用编码 1,2,3 或 3,2,1 来表示。这里编码排序已有特定意义，编码值的增大或缩小，表明产品质量优劣程度的变动方向。

定序测定比定类测定要精确，但不能测定出各类别之间的间距大小，所以测定结果只能比较大小，确定优劣的方向，不能对编码值进行任何加、减、乘、除等数学运算。

（三）定距尺度

定距尺度(Interval Measurement)也称间距测定，是比定序测定量化程度高一层次的测定。它不仅可以对事物进行分类、排序，而且可以测定不同类别之间的间距大小。例如，可将某产品的市场占有率按从高到低的顺序排队，形成 40%，30%，20%，10%，5%的序列。这里的定距尺度的取值已不是类别编码，而是具有计量单位的实际值。定距尺度的取值可以进行加、减运算。如上例，若假设甲地市场占有率为 40%，乙地市场占有率为 20%，运用减法可以得到两地的市场占有率的差额为 20%，这里的差额是有实际意义的。但是定距尺度的取值不能进行乘、除运算，这是因为在具体问题中，很难准确定义这一测定层次中变量取 0 值时的准确含义。如上例，我们可以说某产品甲地的市场占有率比乙地高 20%，却不能说甲地的市场比乙地大 2 倍，就好像不能说 40℃比 20℃暖和 2 倍一样。由此可见，定距尺度不仅可以显示类别的差异，反映差异的绝对量，而且在测定过程中由于具有确切的计量单

位，差异的程度可得以较精确地反映。

（四）定比尺度

定比尺度（Ratio Measurement）也称比率测定，是最高等级的测定层次。测定的取值为实际值。它不仅可以根据现象的性质对变量进行分组、排序，而且可以对数值进行加、减、乘、除任何数学运算。定比测定与定距测定的最大区别在于：定距尺度中没有绝对“零”点，“零”值在某一数列中表示一个相对位置、相对水平，不是“没有”、“不存在”的意思。例如，某一地区的温度为0℃，它表示一种温度水平，而不是“没有”温度；而定比尺度有绝对“零”点，“零”值表示“没有”、“不存在”。例如，一个企业某月的利润为“零”，则表明这个企业某月没有利润。因此对具有相同对比基数的两值进行测定时，加、减、乘、除均有实际意义。例如，甲企业的利润为66万元，乙企业的利润为22万元，在定距测定时，我们可以说两者的利润相差44万元，在定比测定时，我们可以说甲企业的利润是乙企业的3倍。

统计测定这四个层次的关系是：对事物的测定或计量可以由低级到高级逐次递进。测定层次越高，测定结果所包含的信息量越大，这是由不同测定尺度的计量功能的大小不同所造成（见下表）的。在实际统计分析中，根据具体问题的需要，选择合适的测定层次来既满足规定的精度要求又不使信息有太大的浪费，是十分必要的。

四种测定尺度数学特征的比较如表1-1所示。

表1-1

测定尺度数学特征比较表

测量尺度	数学关系			
	=or=	>or<	+or−	×or÷
定类尺度	√			
定序尺度	√	√		
定距尺度	√	√	√	
定距尺度	√	√	√	√

第三节 SPSS 概 述

SPSS统计分析软件是其全称“Statistics Package for the Social Science”的英文缩写。随着SPSS产品服务领域扩大和服务深度的增加，SPSS公司于2000年正式将英文全称改为“Statistics Product and Service Solution”，意为统计产品与服务的解决方案，标志着SPSS的战略方向的重大调整。作为与SAS、MINITA等齐名的世界著名统计分析软件，近几十年来，SPSS软件在包括我国在内的世界各国的社会科学及自然科学的众多领域，得到了广泛的接受及运用。在我国，由于SPSS软件其强大的数据管理、分析功能，直观易学的操作特点，灵活、丰富、友好的界面，在经济学、医学、心理学、金融、教育、交通、体育等各个领域，发挥了巨大的作用。在我们统计学的教学过程中，深感SPSS软件在处理大数据集时的上述优越性。所以本书以SPSS for Windows11.0版本作为讲解统计分析的重要工具，除了介绍SPSS软件的基本操作外，主要是结合案例进行分析。

一、SPSS for Windows运行方式简介

SPSS软件为用户提供了三种基本运行方式，分别为：完全窗口菜单方式、程序运行方式、混合运行方式。

（一）完全窗口菜单方式

完全窗口菜单方式是在SPSS运行中，通过点击主菜单或子菜单的各功能项以及通过对话框对话来完成所有统计分析过程的一种方式。用户可以根据自己需要、根据选项上相应的解释说明，通过菜单操作来完成一系列的统计分析任务。完全窗口菜单方式的操作通常在数据编辑窗和输出观察窗内进行，用户不需要具备计算机编程知识，只要掌握一定的windows操作原理，略知统计基本原理，就可以在短时期内，掌握这种方便、直观的运行方式。本书主要介绍这种方式。

（二）程序运行方式

程序运行方式是用户根据自己的需要，在syntax窗口直接运行手

工编写好的有关 SPSS 命令程序的一种运行方式。对有特殊需要又熟悉 SPSS 语言编写程序的用户而言，是一种行之有效的方式。

（三）混合运行方式

混合运行方式是以上两种运行方式的结合。首先按菜单运行方式输入数据，利用对话框选择分析功能及参数，然后，不马上提交系统执行，而是通过 Paste 按钮，将选择的过程及参数转换成 SPSS 命令，在 syntax 窗口提交运行。混合运行方式一般适用熟悉统计分析过程的统计专业人员。

二、SPSS for Windows 的两个基本窗口

（一）SPSS 的数据编辑窗口

启动 SPSS for Windows 后便激活了数据编辑窗，数据编辑窗如图 1-2 所示。

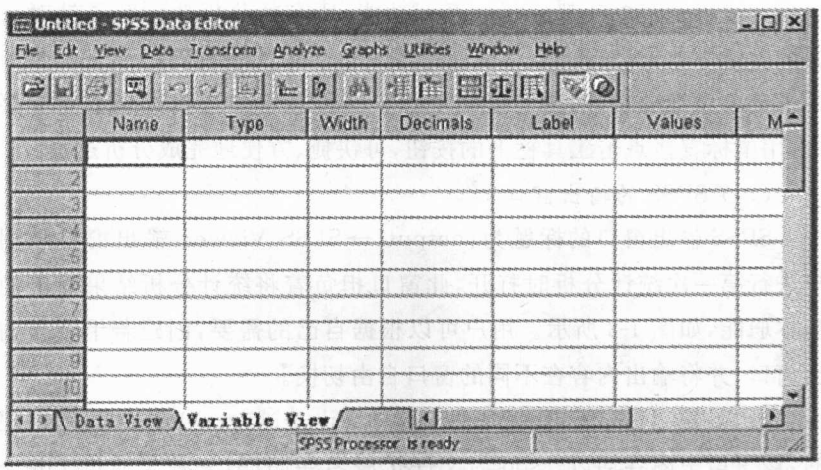

图 1-2　SPSS 数据编辑窗

SPSS 数据编辑窗的窗口标题为：data editor，它是 SPSS for Windows 的主程序窗口，此窗口担负着对 SPSS 数据文件的输入、修改、管理等一系列基本操作。SPSS 数据文件均以扩展名为 .sav 的形式存储。

SPSS 数据编辑窗口由四部分组成：窗口主菜单、工具栏、数据编辑

区、系统状态显示区。

窗口主菜单由 10 个菜单项组成，它们的名称及基本功能如下：

File：SPSS 文件操作菜单

Edit：SPSS 文件编辑菜单

View：SPSS 窗口状态设置菜单

Data：SPSS 数据文件建立与编辑菜单

Transform：SPSS 数据转换菜单

Analyze：SPSS 统计分析菜单

Graphs：SPSS 统计图表菜单

Utilities：SPSS 实用程序菜单

Window：SPSS 各窗口切换菜单

Help：SPSS 帮助菜单

上述每个菜单项都包括一系列功能，所有的分析功能都是针对数据编辑窗中数据进行的。可以使用各菜单项中的各种功能或进一步展开子菜单中细分功能，进行分析工作。操作过程中当找到合适功能菜单，用鼠标直接点击工具栏上的按钮，可快捷、方便地完成分析过程。

（二）SPSS 的输出窗口

SPSS 输出窗口的标题为：output1→SPSS Viewer，输出窗口随用户进行第一次统计分析时打开，此窗口担负着将统计分析结果输出的基本职能，如图 1-3 所示。用户可以根据自己的需要，创建若干个新输出窗口，并将输出内容在不同的窗口自由切换。

SPSS 输出窗口有两种：一种是通过 File→New→Output 路径创建的，该窗口中内容均以 .spo 为文件扩展名存储；另一种是通过 File→New→draft view 创建的，该窗口中内容均以 .rtf 为文件扩展名存储，它的本文内容可以直接被 word 等其他软件直接读取及编辑。

SPSS 输出窗口由四部分组成：窗口主菜单、工具栏、分析结果显示区、系统状态显示区。

窗口主菜单由 10 个菜单项组成，它们是：File，Edit，View，