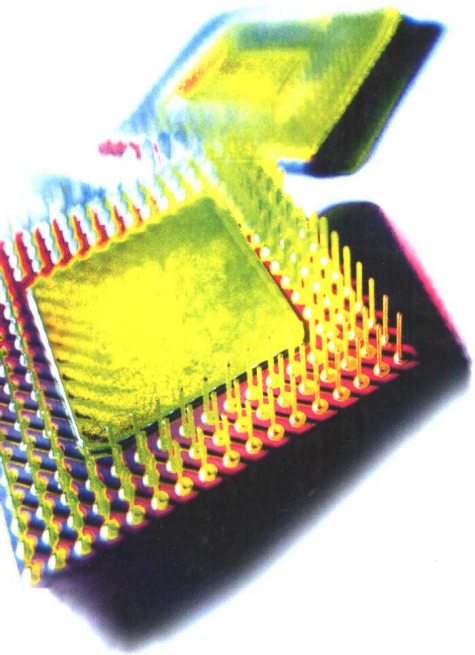
数据分析方法

SHUJU FENXI FANGFA

SHUJU FENXI FANGFA

SHUJU FENXI FANGFA



董麓 编著

东北财经大学出版社
DUFEP Dongbei University of Finance & Economics Press

数据分析方法

董 麓 编著

东北财经大学出版社

图书在版编目（CIP）数据

数据分析方法/董麓编著 .—大连：东北财经大学出版社，2001.11

ISBN 7-81044-972-9

Ⅰ.数… Ⅱ.董… Ⅲ.数据-分析方法

Ⅳ.0212

中国版本图书馆 CIP 数据核字（2001）第 080246 号

东北财经大学出版社出版

（大连市黑石礁尖山街 217 号 邮政编码 116025）

总 编 室：（0411）4710523

营 销 部：（0411）4710525

网 址：http://www.dufep.com.cn

读者信箱：dufep@mail.dlptt.ln.cn

东北财经大学印刷厂印刷 东北财经大学出版社发行

开本：880 毫米×1230 毫米 1/32 字数：159 千字 印张：5 1/2

印数：1—3 000 册

2001 年 11 月第 1 版 2001 年 11 月第 1 次印刷

责任编辑：孙 平 责任校对：孙 萍

封面设计：张智波 版式设计：丁文杰

定价：20.00 元

内容提要

社会经济领域的研究离不开数据分析。统计分析方法和计算机技术是现代数据分析的基础，本书从这两个角度出发，介绍了从数据采集到数据分析的实际工作过程和方法。主要内容包括数据、数据组织和数据管理的基本知识，使用 SAS 软件进行探索性数据分析、统计分析和计量分析的基本方法和具体案例。

本书可以作为大专院校财经类专业的学生学习统计学和计量经济学的实验教材，也可以作为从事社会经济领域研究工作者的工具书，希望学习 SAS 软件程序设计的读者也可以从中受益。

前　言

在社会经济研究领域，数据分析越来越受到重视。数据分析离不开统计方法和计算机技术的应用，随着计算机科学的迅猛发展，基于计算机技术与统计方法相结合而产生的现代数据分析技术越来越展示出它无穷的魅力。但是，国内介绍这方面内容的书尚不多见，已有的书籍，或者是偏重于统计理论和方法，或者是偏重于计算机应用。从统计理论和方法的角度来讲述，使许多统计理论知识不足的读者很难理解，常常望而却步。从计算机应用的角度来讲述，又使得具备统计理论知识但计算机知识薄弱的读者难以灵活地运用计算机技术来分析问题。因此需要一本能够指导读者快速掌握基于统计分析和计算机技术相结合的数据处理方法和循序渐进的书籍。

在财经类院校中，学生在学习统计学和计量经济学这两门课程时往往感到困难。一些复杂的数学计算，使他们望而却步，进而失去了继续学习的兴趣。然而，到四年级作毕业设计，以及毕业走上工作岗位以后，很多人都深深体会到这些知识的重要性和自身知识的欠缺。针对这些情况，作者在教学中努力探索有益于激发学生学习兴趣，较好掌握这两类课程的教学方法。经过多年的探索和实践后发现，在教学过程中结合计算机软件处理实际问题，可以大大激发学生的学习兴趣。以往教师担心学生使用计算机软件后会忽视对具体数学处理过程的学习，实践表明，当学生使用计算机完成作业以后，反而激发了他们探求数学方法和原理的欲望。学习过这门课程的学生，毕业论文中进行定量分析的比例显著提高，很多学生使用的数学方法和计算机技术甚至超出了教师讲授的范围。然而，目前国内关于这方面的适用教材还比较少。

写这本书的目的就是提供一本适合大专院校财经类专业的学生学习统计学和计量经济学的实验教材，以及为从事社会经济领域研究工

作的读者提供一本使用计算机进行数据分析的工具书。本书将数据分析的全部过程——从数据采集到数据的组织、管理、加工，以及统计分析作为一个完整的内容，通过大量的实例，向读者介绍如何利用计算机技术和统计分析软件来协调、有效地进行数据分析工作。书中特别注意交叉、融合运用计算机技术与统计方法，将两者有机地结合在一起，尽可能在数据分析的过程中发挥各自的所长。对于这两门学科中具有共同点，却由于表述不同而长期割裂的内容，在把目标聚焦在数据分析过程的基础上进行了沟通。书中对很多重要的技术和方法的背景都做了说明，以便于读者深入学习。

本书选择著名的 SAS 软件进行案例分析，所有程序都可以在 SAS6.12 系统下直接运行。全书共分 4 章。第 1 章从统计学和计算机应用两个角度介绍了数据的基本概念和基础知识。包括数据收集的方法，计算机辅助数据收集方法对传统技术的变革；数据的分类，定量与定性资料，时间序列的数据，统计上的数据分类与计算机应用软件要求对数据分类的区别和联系；数据的组织方法，收集的数据如何组织在一起，数据组织的基本单位与结构，如何从众多的数据中获得满足要求的子集，数据编码的方法与意义；数据的管理，数据的存储方式与结构，数据的兼容性与数据的保密，以及数据管理的技巧。第 2 章介绍使用 SAS 应用软件建立、编辑、存储和管理数据文件的具体方法，SAS 程序工作的环境，SAS 程序的语法和结构。修饰数据的方法，包括一些非常重要的思想和技术如 SQL（数据库结构化查询语言）等。第 3 章介绍数据的探索性分析方法，主要包括两个方面的内容。数据的合法性检验，介绍了进行数据合法性检验的统计方法，如交叉分析，查找异常数据的方法，以及对样本数据进行随机性检验的方法。还介绍了利用程序调试手段，查询错误数据位置的方法，如断点的设置与程序的追踪技术等。探索数据规律和数据之间的关系的方法，介绍了如何使用图形来展示数据之间的关系，每种方法都提供了 SAS 程序设计的范例。第 4 章举例介绍使用 SAS 程序进行统计分析的经济计量分析的方法，包括假设检验，方差分析，相关与回归分析，多元统计分析，定性资料分析和时间序列分析等。本书虽然面向社会经济领域的问题研究，但对从事其他领域的研究，希望使用统计方法

和计算机技术进行数据分析的读者也具有一定的参考价值。

感谢我的导师，天津财经学院的肖红叶教授在本书的写作和出版过程中给予的指导和帮助。天津财经学院的周恒彤教授在我学习统计方法的过程中给予了很多有益的指导。我的同学沈鸣博士、周国富博士、于忠义博士在本书的写作过程中也给予了很多帮助，在此一并致谢。

由于水平有限，错误在所难免，欢迎广大读者批评指正。

董藩 2001.9

于天津财经学院

目 录

第 1 章　数据的基本知识和计算机管理

伴随科学和技术的迅猛发展，人类的经济和社会生活以及其他活动都发生了巨大的变化。大规模的专业化生产、商业、金融、保险及其他各种社会和经济活动都以追求最大经济效率为目标。所从事的活动规模越大，就越需要对发展和计划进行周密的安排。而计划的设计、实施，以及对未来成效的评估，不论工业、商业或政府活动，都不可缺少地要以客观数据资料（data）[①] 作为依据。各种分析、决策性资料都需要有健全的依据，因此就应该用科学的方法、客观的态度进行收集、分析、加工，使之成为信息（information），从而具有使用价值。本章介绍有关数据收集、分类、组织、存储与管理的基础知识。

1.1　数据的收集

数据收集是数据分析的基础性工作，能否正确、成功地收集到所需要的数据，将关系到整个研究工作的成败。本节介绍收集数据的方法和步骤。

1.1.1　数据收集方法简介

收集数据的方法很多，往往因为研究的问题不同而不同，有时也受到人力和财力状况的制约。经常使用的方法有调查法、观察法、实验法和现有资料查询以及网络查询等，图 1—1 列出了这些方法以及

① 数据资料在统计书籍中常称为资料，在计算机书籍中常称为数据、数据集、数据文件等。

它们的适用范围。

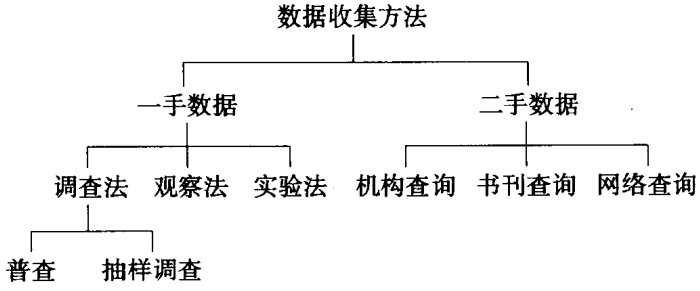


图 1—1　数据收集方法示意图

一手数据是指直接收集的资料，即原始数据。二手数据是指利用已经存在的资料。在实际收集数据时，并不局限于仅采用一种方法，也可以同时使用几种方法。

调查法包括普查和抽样调查，是社会和经济研究中最常用的资料收集方法。普查适用于必须收集每个单位资料的统计调查。通常在收集区域性或全国性的资料时，有时要对全部资料一一调查，这时就需要采用普查的方法，常见的例子如人口普查。如果普查的范围很大，花费的人力和财力就会很多。

抽样调查的数据来源于总体中的一小部分。通常把研究的对象如个人、家庭、田块、工厂、商店称为抽样单位或单位，要研究的全部对象称为总体。抽样调查就是从总体中抽取一小部分单位，这些被抽出的单位构成了总体的一个样本，然后用抽出的样本来推断总体。和普查相比，抽样调查具有节省时间、节约金钱、有较大的作业弹性和调查范围，以及可以对误差进行控制等优点。因此，凡是能够采用抽样调查收集数据的研究，都应该采用抽样调查。抽样调查的应用范围非常广泛，例如，调查商品市场、金融市场情况、农产品产量、土地使用状况、劳动力的多少与失业人数、工业生产、人民健康水平与家庭收支等。

观察法是调查人员直接或利用仪器在现场观察调查对象的活动，通过观察结果来收集资料的方法。由于观察法在实行时不让被调查者察觉，所以这种方法最适用于任何人都可以接触的数据，或是可以直

接以观察获得数据的情况。例如，车站人流统计、交通流量、货架上的价格标识等。

实验法是研究者在研究领域内，为发现一个特定过程或系统的某些现象或规律，而设计的一系列试验。例如，验证某种药物疗效的试验，设计者选择一批试验对象，把他们随机地分为两组。一组试验对象服用要检验的药物，另一组服用安慰剂，药物和安慰剂在外观上没有区别。跟踪观察并统计试验对象的治愈情况，然后对两组数据进行统计、对比分析，从而得出药物是否有效的结论。近些年试验设计发展很快，在很多领域被广泛应用。

很多研究工作可以利用已经收集并整理好的数据，即二手数据。这些数据可以向有关的研究部门、公司和政府机构查询或者购买。一些公开出版的图书、报刊也提供可供分析的数据。通过计算机网络也可以收集数据，目前已经有很多专门提供各种数据的信息公司，通过网络提供数据的有偿服务。使用二手数据能节省大量的时间和金钱，但必须认真考察数据是否可靠，是否适用于所做的研究工作。

1.1.2 计算机辅助数据收集

计算机技术的发展不仅使得数据收集工作更加方便、准确和有效，而且在很多方面带来了革命性的变化。下面我们通过一个例子来说明计算机技术的影响，图 1—2 是使用计算机辅助访问调查的示意图。

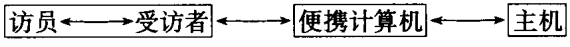


图 1—2 计算机辅助调查示意图

从图 1—2 中可以看出，访员和受访者通过和主机联网的便携式计算机在主机的监控下进行访谈。访员和受访者可以随时和在主机旁边工作的专家进行信息交流，专家可以实时地监控调查的全部过程，指导访员的行为，随时对已经收集的数据进行加工和分析。这种新型的调查方式是对传统方式的重大变革，它将传统的静态访问调查方式发展成为动态访问调查方式。计算机辅助访问调查还从根本上改变了数据从收集、整理、录入、分析分步进行的传统工作模式，将它们融为一体，从而减少了中间环节，大大提高了调查数据的质量，加快了

数据向信息转化的速度，提高了信息的使用价值。

计算机技术的发展还为数据的保存、组织、管理、共享、复用、分析等方面提供了更多、更有效的方法。

1.1.3 数据收集的步骤

数据收集应该有计划按步骤地进行，数据收集的基本步骤包括：

- 发现问题，明确调查目的。发现问题是数据收集的第一个阶段。这一阶段的工作是找出要研究的问题的实际处境和所期望的状态之间的差别，即通过对现实和期望的比较，找出它们之间的差距，然后对差距进行评价以便找出存在的问题。获得期望值的方法通常是利用历史数据进行推测，或者根据目前制定的计划，也可以根据竞争对手或用户需求情况来确定。
- 确定数据收集方法。收集数据之前要先确定采用何种资料，是一手资料还是二手资料，如何获得所需要的数据。然后确定数据的格式，以便于数据的收集和整理，在确定格式时应该尽量采用标准的格式。
- 确定调查方案。定义研究的总体，确定样本的大小、样本的性质，如何分配样本单位。在大规模调查前可以先作小规模调查，以便找出方案存在的缺点。通常在这一步还要编制整个调查的资金预算。
- 收集数据。按制定的计划和方法收集数据。从事数据收集的人员，应该具备一定的专业知识和技能，通常应该在收集数据之前对收集人员进行培训。
- 数据整理。对收集的数据进行整理，将不符合要求的资料剔除，统一计量单位和书写格式，并制作相应的表格。

1.2 数据的分类

数据的分类方法很多，不同的领域和学科往往有各自的分类方法。即使在同一个领域，由于研究问题的角度不同，也会产生不同的分类方法。下面介绍几种在社会经济研究中常见的数据类型：

1.2.1 按数据与时间的关系划分

从数据和时间关系的角度进行观察，可以把数据划分为三种类型：

- 研究的问题是在一个固定时间，或者一段时间内的统计对象的数量特征和各对象之间的关系。例如，人口、产量、消费者需求与爱好、就业状况等。这类资料会随着时间改变，所以要逐次收集。
- 研究的问题是统计对象及各对象之间的关系随时间变化的情况或趋势。这类资料一般是统计对象的变化及对象之间的关系呈时间数列，要由定时收集的资料求出，利用研究结果可以预测未来。例如，分析社会经济因素与人民生活状况间的关系，预测对未来的需求。
- 研究的问题是对时间相对稳定的对象。常见的如地理、地质资料，气候、土壤类型，石油储量等。

1.2.2 从描述和度量事物的角度划分

从对事物的描述和度量的角度来观察，可以把数据分为定量的和定性的两类。定量数据又可以分为：

- 计量的。例如，人的身高、体重、血压等，气象上的温度、相对湿度等等。这一类资料的特点是：原则上它的取值可以是在某一区间内的任一实数，通常称这类资料是连续的，或考察的指标是连续的。它的统计分析是与具有密度的连续随机变量的分布有关。
- 计数的。例如，一个平方米内某一种害虫的个数、一个居民区内拥有的电视机的台数、一个单位内职工的总人数等等。这一类资料的特点是：它们的取值范围是整数，大部分还只在非负整数范围内取值，通常称这一类资料是计数的，或考察的指标是计数的。它的统计分析是与离散的随机变量的分布有关。

这两类资料原则上是可以分清的，但实际上有时也有困难。例如，一个人的年龄，按理可以认为是连续的，然而实际上只能按年或月或日计算，是计数的。从这里可以看到，有时为了方便，连续的资料是可以离散化的。

定性资料又可以分为：

- 有序的。有些资料，既不能计量，也不能计数。例如，这一块布的颜色比那一块深，但无法量化；又如，评定毛料的手感程度，感觉这一种比那一种丰满；评定酒或茶的好坏时，只能评出一个顺序，而无法量化。这一类指标和资料，我们称它为有序的。
- 名义的。有些资料不是计量的、计数的，也不是有序的，它仅仅是一个名义值。例如，给各种不同的颜色赋以代号，给不同的书籍赋以代码等等。这些赋值只起一个名义作用，它的值的顺序和大小并无统计意义。这一类资料，我们称它为名义的。①

1.2.3 数据在计算机中的分类

用计算机处理数据时，也需要说明数据的类型。这是为了使应用软件能有效地识别、存储数据和确定对数据的运算。计算机中常见的数据类型有整数、实数、双精度数②、字符串、布尔、日期、图形等。定量数据在计算机中一般定义为整数、实数或双精度数。定性资料一般定义为字符串，也可以用整数表示有序的资料。下面是一个定义计算机数据类型的例子。

［例 1—1］某公司生产打印机（p）和计算机（c）两种设备。由公司职员负责在四个地区销售。表 1—1 列出了销售人员的姓名（name）、销售金额（sale）、销售地区（region）和销售的设备类型（type）。

姓名、销售地区、设备类型是名义的，所以这三个变量应该定义为字符串类型。销售金额是计量的，所以应该定义为实数类型。有的读者可能认为这里的销售金额都是整数，所以应该把它定义为整数类型。但是多数计算机应用软件的整数是指在一定范围内的整数，常见的如在 - 32767 至 32767 之间。如果数据超出了应用软件整数类型规定的范围，就应该改用实数处理。有些应用软件（如 MICROSOFT EX-

① 参阅张尧庭：《定性资料的统计分析》，桂林，广西师范大学出版社，1991。
② 双精度数也是实数，但表示的数据的有效位数更多。有关计算机数据类型的内容请参阅专业书籍。

CEL）提供了短整数、长整数两种整数类型，短整数的范围一般在－32767至32767之间，长整数的表示范围则很大。定义数据类型时并不是选择精度越高，或数据范围越大越好。因为复杂的数据结构占用的存储空间大于简单的存储结构，例如，存储一个整数使用两个存储单元，而存储一个实数可能需要四个或更多的存储单元。数据结构越复杂，占用计算机的存储空间越多，程序处理所需要的时间越长。

表1—1　　**销售情况表**

name	sale	region	type
stafer	9664	east	p
young	22969	east	p
stride	27253	east	p
topin	86432	east	c
spark	99210	east	c
vetter	38928	west	c
curci	21531	west	p
marco	79345	west	c
greco	18523	west	p
ryan	32915	west	p
tomas	42109	west	p
thalman	94320	south	c
moore	25718	south	p
allen	64700	south	c
steiam	27634	south	p
farlow	32719	north	p
smith	38712	north	p
wilson	97214	north	c

1.3 数据的组织

合理地组织数据，便于数据的收集、整理和分析。使用计算机存储和分析数据时，必须首先按照使用的计算机应用软件的要求，对数据进行组织。本节介绍数据组织的基本概念和在计算机中组织数据的基本原理和方法。

1.3.1 观测单位与数据项

观测单位是收集和记录数据的基本单位。研究学生的学习情况，观测单位是每个学生；研究家庭的收入、支出问题，观测单位是每个家庭。抽样调查时观测单位往往就是抽样单位。一般每个单位都包含若干个数据项或称为变量、字段，例如，研究家庭收入和支出问题时我们需要了解每个家庭的收入、人口、开支情况、受教育程度等等，收入、人口、开支情况和受教育程度都是数据项。数据项是记录数据的最小单位，统计上称这些数据项为指标，收集一个单位的全部指标的数据称为一次观测。观测单位和数据项的概念可以通过表 1—2 来说明。表 1—2 记录的是学生所在专业的情况：

表 1—2　　学 生 情 况

学　号	姓　名	专　业	班　级
34200	苗　晶	信　息	2001
35012	王　敏	统　计	2002
35228	丁　辉	信　息	2001
35618	王　健	统　计	9906
36120	刘　芳	金　融	2006
36336	陈绍瑞	金　融	2006

表 1—2 中的每一行为一个观测单位，每个单位包含四个数据项，分别是学号、姓名、专业、班级。

1.3.2 数据在计算机中的组织

使用计算机组织数据时，逻辑上把一个单位的数据组织在一起称为一个记录，每个记录包含相同的数据项，全部记录的集合称为一个