上海财经大学证券期货学院丛书

Data Analysis Method

数据分析方法

施锡铨　范正绮　编著

上海财经大学出版社

上海财经大学证券期货学院丛书

数据分析方法

施锡铨　范正绮　编著

上海财经大学出版社

数据分析方法

SHU JU FEN XI FANG FA

施锡铨　范正绮　编著

责任编辑　江　玉
封面设计　吕　莱

出　版　上海财经大学出版社
（上海市中山北一路 369 号　邮编 200083）
发　行　新华书店上海发行所
印　刷　上海市印刷七厂一分厂
装　订　上海市印刷七厂一分厂
开　本　850×1168mm　1/32
印　张　17.25（插页 2）
字　数　464 千字
版　次　1997 年 10 月第 1 版　1997 年 10 月第 1 次印刷
印　数　1—3000
书　号　ISBN 7—81049—159—8/F・122
定　价　34.50 元

作者简介

施锡铨 1944年7月出生于浙江湖洲。1967年毕业于复旦大学数学系，1982年获华东师范大学概率统计专业理学硕士学位。现为上海财经大学教授。长期从事应用数理统计研究工作，其研究方向与主要兴趣为抽样与再抽样理论、鞍点逼近及小样本理论、统计质量管理、统计计算、经济统计等方面。多次发表论文于《The Annals of statistics》、《Quality Engineering》、《科学通报》、《数学学报》等。与系统所冯士雍教授合作专著《抽样调查——理论、方法与实践》，主编教材《抽样调查的理论与方法》，主译《统计学》。

范正绮 1946年3月出生，湖北黄陂人。1967年毕业于北京地质学院。先后任职于武汉水利电力学院科研所、复旦大学管理学院计算机室。现任上海财经大学计算中心副研究员。早期从事工程地质与岩石力学研究及其实验数据分析，目前对经济建模、统计预测及计算机统计模拟很感兴趣。在《应用概率统计》、《岩石力学与工程力学》、《统计研究》、《会计研究》、《上海金融》等杂志发表15篇论文。并参与《抽样调查的理论与方法》编写及《统计学》一书的翻译工作。

总　序

龚浩成

我国的证券市场从1984年发行第一张股票算起，至今已有12年了；从第一家证券交易所成立算起，至今也将近6年。这期间，我国的证券市场发展迅速。仅以上海流通市场为例，在1984年至1985年间，年交易量仅20万元；而1991年上海证券交易所成立后的第一年里，交易量就达到了63亿元（以单向统计计算），增长了3万多倍。1994年上海证券交易所成交量达25 000亿元，比1991年又增长了400倍。至1995年，上海证券交易所交易金额达55153亿元，又比1994年翻了一番多。可以想见，随着我国市场经济体制的不断确立和推进，证券市场仍将以较快的速度向前发展。

证券市场的快速发展对证券业人才的培养提出了新要求。与其他相关行业比较，证券业从业人员总体知识水平不算低，其中也不乏大学本科学历以上的从业人员。但也应该看到，我国的证券市场是在近几年内发展起来的新生事物，从业人员大多是从其他行业转入的，即使具有本科以上学历的人员，相当一部分人的知识结构也与证券业所需的知识结构存在很大差距。所以，加快培养大批合格的证券期货专门人才，已成为我国证券期货市场健康发展的关键因素。有鉴于此，上海证券期货交易所与上海财经大学合作创办了我国第一家证券期货学院，旨在为我国证券市场的发展提供教学、科研和培训等方面的配套服务。目前，学院正在本科教育、研究生教育和社会从业人员培训三个方面进行着卓有成效的工作。

学校以教学、科研为主，教材是教学的基础，而科研则是教师提高教学质量的保证。学院成立后的第一件事就是抓教材建设，并

鼓励教师多出科研成果。为此，学院根据教学科研的需要，及时组织教师撰写有关的专著、译著及专业教材，继今年年初已在上海三联书店出版发行了《证券投资学》、《基金管理学》、《商品投资学》、《金融期货与期权》、《社会保障经济理论》等五本专著、专业教材后，目前学院又有一批专著、译著、专业教材已进入出版、撰写阶段。

这一系列教材和专著的出版，对系统总结证券市场发展过程中的理论和实践，促进证券投资学科体系的建设，加速培养成熟的证券投资与经营管理人才，具有重要的理论意义和现实意义。证券市场在西方已有上百年的历史，而在我国则是一个新兴的蓬勃发展的市场，它在快速发展过程中必然暴露出种种不成熟的弱点，系统的理论指导就成为迫切需要的营养。我相信这套丛书的出版必将对我国证券市场的健康发展起到积极的促进作用。只要每一个市场参与者经过认真的学习与培训，把握了市场的运作规律，在法规的约束下冷静、理智地进行投资，我国的证券市场必然进一步得到规范的发展。

前　言

写下《数据分析方法》这个名字，未免涌起一阵感慨，记得读书期间崇尚的是函数论、抽象代数这类抽象理论，当时在同学之间还戏称计算机专业为“打洞”专业——因为他们需要在输入带上打洞。星移斗转，时代的确不同了！现在我们再也离不开计算机，我们的研究工作中也有大量的数据需要计算机一起处理，例如本书的另一位作者在很长时期内经常性地处理岩石力学实验数据、水文地质的有关数据以及灌浆等工程数据；我们还处理过可靠性方面的大量数据，包括电视机的可靠性试验、地铁客流模拟中的乘客流量等随机数据、市场调查的数据处理以及证券市场的有关数据等等。社会主义市场经济不断发展，不断繁荣，使我们“陷入了数据的包围之中”。研究生们，尤其是攻读博士学位的研究生，无论是财政专业，还是税收专业，无论是经济理论专业，还是实用的会计专业，对数据处理都正在产生浓厚的兴趣，他们发现国内外学术刊物之间的差异之一正是在于国外刊物强调对数据的科学处理以及用数据分析来揭示内在规律。即使是证券期货学院的本科生们，也越来越明白数据处理的重要性，他们对MINITAB、TSP、SAS等统计软件表示出浓厚的兴趣，也曾经提出开一些管理统计软件及数据分析课程的愿望。我们感觉到了在财经院校普及数据分析工作方面自己应尽的责任，这也成了我们写作《数据分析方法》的动力。经过努力，作为上海财经大学证券期货学院拟议中的一套丛书之一，《数据分析方法》即将展现在读者面前。

应当承认，数据分析是一件相当困难的工作，对摆在人们面前的一大堆“乱七八糟”的数据从何处着手，又怎样引导到正确的轨道，从而揭示一定的规律，这件工作不仅仅需要耐心，更需要进取心，要主

动去设想并收集尽可能多的信息，了解数据来源以及明确的分析目的，综合在一起考虑才有可能找到较理想的分析工具，建立起较令人满意的模型。

数据分析问题不存在固定的解决方法。没有一个人敢于断言什么样的数据可以用什么样的方法解决，做到分门别类，一清二楚。可以这样认为，敢于这样说的人也许并不是个数据分析工作者。在本书中我们只能尽可能地将各种方法一一叙述并偶尔作一些建设性建议，希望读者在理解各种方法的特点之后灵活地处理摆在你面前的数据。美国 Wisconsin-Madison 大学统计系有一门课程，教授通过计算机交给学生许多数据，学生的任务在于利用一切已学到的知识刻划这些数据可能的规律，教授根据学生使用工具的合理性、所得结论的逻辑性等给出评价与分数。我想，这才真正地体现出数据分析工作的特色。

数据分析的基本方法和工具实在太多，要在一本书中全部集其大成是不可能的。我们原计划从基本统计手段、多元分析方法、时间序列分析及经济建模四个方面阐述，后算下来篇幅实在太多，这可能不适合作为教材，因此决定拆成两部分写作，本书即为计划中的第一部分，基本统计方法当然包括在内，考虑到证券期货学院的需要，又收进了简单的时间序列分析部分。写作本书的一个基本思路是，只叙述方法的本身操作以及有关这种方法的思想，尽可能地不涉及数学证明等内容。每种方法都配有一定例题，在每章后面基本上附有有关的 SAS 指令（第八～十章的 SAS 指令均附在第十章后面），每章后都配有一定量的习题。

施锡铨负责本书第一章、第三至十章的内容，范正绮撰写了第二章与第十一章。各章中最后一节的SAS有关指令均由范正绮负责编程。

最后要强调一点，数据分析工作也离不开好的统计软件。美国 SAS 软件公司上海办事处对我们的教学与科研工作给予了极大的关心与支持，特别是张少华、李斌等朋友们关心尤甚。在此向他们表

示深深的谢意!

数据分析是件困难的工作,我们写作这方面的书也算是遇到了一件困难的任务。本书可能有不少谬误之处,恳望读者对本书的错误不吝指教,以使我们获益匪浅,先在这里致谢了!

施锡铨

1997 年 7 月于上海财经大学

FOREWORD

With the socialist market economics having been developing at high speed in China, lots of problem with complexity must be studied and solved, such as market potential forecasting, company management, economic modeling, stock market technical analysis, etc. In research work, pertinent data are often incomplete and initially seldom in a form that specifically supports decision making. The use of statistical methods to transform relevant data into usable summaries is data analysis.

This book is intended to introduce data analysis methods to readers. Maybe almost every reader wants to know which statistical procedure is suitable for which data set . Unfortunately we can't do so in this book. Because of the complexity and randomness, for a given set of data, perhaps there are several statistical procedures to deal with. Sometimes the results are the same, sometimes are different. Which is correct or which is the best? Understanding for data, knowing information about the background of the data and experience is very important. In this book, we only show the readers various statistical procedures, show them how statistics can be used to analyze data. We hope the readers can understand statistical data analysis methods by studying thd examples which we carefully collected in this book.

We emphasized statistical idea in this book. The discussion is not calculus-based. We almost ignored all mathematical proofs because it isn't the first. But we hope that the readers have had

fundamental training in probability and statistics. If one has had a first course in statistics, he will read this book without difficulty。

In original, there are four areas in our plan: statistical procedures (included regression, data comparisons, etc.) multivariate statistical analysis, time series analysis and economical modeling. The content is too much to include in one book, so we divided the content into two parts. This book introduces statistical methods and time series analysis and forecasting methods.

As well known, to do data analysis, computers and a useful computer statistical software package are necessary. In chapter 2, we introduce the SAS system. This work was supported by SAS Institute Ltd. (Shanghai Representative Office). We are thankful to them for their consistent support in our research work. In our experience, SAS system is a useful tool in data analysis. In this book, we will introduce relevant SAS syntax and commands at the end of almost every chapter, and show the readers how to use SAS system to realize the procedures described in this chapter.

Chapter 3 offers statistical description about data set. Data comparisons (include hypothesis test), regression, nonparametric and variance analysis are in chapter 4—7. In chapter 8—10, we discussed various time series, included the time series with trend or no trend, and we deal with seasonal series. Box-Jenkins modeling procedure is in chapter 11.

We thank Shanghai University of Finance and Economics, Stock and Future Institute for their supporting.

目　　录

第一章　数据分析引论

人类对变幻无穷的大千世界充满着好奇心，千方百计地运用自己的智慧及各种有效手段试图探索和理解现实世界，其中必不可少且极有成效的手段之一是数据分析。其实，作为一门工具，数据分析的最终目的就是使人们进一步认识世界。一般教科书中很少对数据分析下确切的定义。顾名思义，数据分析主要注重于数据的收集、整理及分析，这使人们很自然地联想起“统计学”。众所周知，统计学正是被定义为“收集、分析和解释数据的一门科学”。于是这两者可能常常被混为一谈。事实上，数据分析的主要工具之一的确是统计学，然而它们之间尚存在着本质上的差异。

统计推断的实施必然地需要某些相当严格的条件，在统计学中，似乎一切都在控制之中，数据的特性可以被确定。譬如，一批数据是独立的或在一定程度上相依的，误差项服从同一正态分布等等；变量之间的关系可以得到证实。统计学家希望这些“实验室”产品与遇到的数据分析问题有一定的关系，或者说所遇到的问题可以简化为统计理论已经研究过的某些东西。

数据分析则从对立的方向来处理问题，只有极少场合，数据的分布是已知的，且更经常地样本分布，与统计学中所研究的任何分布形式都不类同。没有任何东西处于分析人员的控制之中，也没有什么假设。然而这样一些并非符合统计学中严格要求的数据却包含了有价值的信息，可以提供我们去学习、研究和作出决策。处理这类数据通常使用的是数据分析技巧，从而数据分析技巧是工具。

著名的统计学家 J. Neyman 曾经这样地概括数据分析问题：“在实际工作与相应的理论之间决不可能存在严格的一致，然而我们生