暨南大学经济管理实验中心实验教材

数据统计分析及R语言编程（第二版）

Data Statistics Analysis and R Language Program

王斌会　编著

北京大学出版社 PEKING UNIVERSITY PRESS
暨南大学出版社 JINAN UNIVERSITY PRESS

经济管理国家实验教学示范中心
经济管理省级实验教学示范中心 共同资助

暨南大学经济管理实验中心实验教材

数据统计分析及R语言编程（第二版）

Data Statistics Analysis and R Language Program

王斌会 编著

北京大学出版社 PEKING UNIVERSITY PRESS
中国·北京

暨南大學出版社 JINAN UNIVERSITY PRESS
中国·广州

图书在版编目（CIP）数据

数据统计分析及 R 语言编程 / 王斌会编著 . —2 版 . —广州：暨南大学出版社，2017. 6
ISBN 978 - 7 - 5668 - 2100 - 3

Ⅰ. ①数… Ⅱ. ①王… Ⅲ. ①统计数据—统计分析②程序语言—程序设计 Ⅳ. ①O212. 1
②TP312

中国版本图书馆 CIP 数据核字（2017）第 093292 号

数据统计分析及 R 语言编程（第二版）
SHUJU TONGJI FENXI JI R YUYAN BIANCHENG（DIERBAN）
编著者：王斌会

出 版 人：徐义雄
责任编辑：曾鑫华
责任校对：邓丽藤
责任印制：汤慧君　周一丹

出版发行：暨南大学出版社（510630）
电　　话：总编室（8620）85221601
　　营销部（8620）85225284　85228291　85228292（邮购）
传　　真：（8620）85221583（办公室）　85223774（营销部）
网　　址：http：//www. jnupress. com　http：//press. jnu. edu. cn
排　　版：广州市天河星辰文化发展部照排中心
印　　刷：广东广州日报传媒股份有限公司印务分公司
开　　本：787mm × 1092mm　1/16
印　　张：15
字　　数：365 千
版　　次：2014 年 8 月第 1 版　2017 年 6 月第 2 版
印　　次：2017 年 6 月第 2 次
印　　数：3001—6000 册
定　　价：39. 00 元

第二版前言

统计学是研究不确定性现象数量规律性的方法论科学，在众多的专业、学科领域中都起着重要的作用，具有很强的应用性，是进行科学研究的一项重要工具，在自然科学、社会科学和经济管理等领域已得到越来越广泛的应用。随着计算机的普及和统计软件的广泛使用，了解和运用它的人迅速增加。作为数据处理非常有用的方法，它在各个领域都卓有成效。

众所周知，数据的统计分析是以概率统计为基础，应用统计学的基本原理和方法，结合计算机对实际资料和信息进行收集、整理和分析的一门科学。因此，它的原理较为抽象，对学生的数学基础要求也较高，教学中存在着大量的数学公式、数学符号、矩阵运算和统计计算，必须借助于现代化的计算工具。

R 语言是属于 GNU 系统的一个自由、免费、源代码开放的软件，是一个用于统计计算和统计制图的优秀工具。在目前保护知识产权的大环境下，开发和利用 R 语言将对我国的统计事业有非常重大的现实意义。

本书是关于 R 语言的一本入门教材，由于主要针对初学者，将重点放在了对 R 语言工作原理的解释上。R 语言涉及广泛，因此对于初学者来讲，了解和掌握一些基本概念及原理是很有必要的。读者在打下扎实的基础后，进行更深入的学习将会变得轻松许多。本着深入浅出的宗旨，本书配有大量图表，使用尽可能通俗的语言，使读者容易理解而又不失细节。

本书的特色是：

（1）原理、方法、算法和实例分析相结合：鉴于目前计算机统计分析软件已是统计分析应用中不可缺少的工具，本书特别强调各种统计分析的 R 语言算法实现，使得给出的计算方法更有实用价值。

（2）解决统计软件用于统计学教学和科研中存在的问题：国内目前缺乏适合开展统计分析教学科研的统计分析软件，如 SAS、SPSS、S－PLUS 等统计软件，由于没有版权，需要昂贵的购买费用，更新很慢，并且需要大量的维护费用，许多内容与教科书设置不完全一致，财经管类学生和研究人员使用较为困难。

（3）提供了一些用于统计分析的 R 语言程序，特别是统计模拟方面的内容，并及时加入现代统计的一些新方法。本书中的所有结果、图形和算法都是由 R 语言给出的。

（4）研究如何将统计软件的数据处理与统计教学相结合，形成一套完整的教学与科研相结合的统计过程。在教学与科研一体化的功能上，在数据编辑、统计分析、统计设计、统计绘图和统计帮助上充分体现多媒体教学的特点。

本书的最大特点在于从数据处理的角度来讲解统计分析，而不是从统计分析的角度来介绍数据处理。也就是说，本书在数据收集与处理上采用了一套比较方便的流程，即用一组数据贯穿于整个数据统计分析过程，这样可使读者不必花很多时间去了解各种数据的特性，并寻找合适的统计方法来进行数据分析。

本书的内容安排吸收了国内外有关统计分析教材的特点，在章节的安排上遵循由浅入深、由简到繁的原则，对统计量和分布进行了较为详细的介绍，增加了许多探索性统计分析的内容和一些统计推断的内容，同时附加了一些数据结构和矩阵运算的概念。书中的主要内容是笔者在暨南大学多年从事统计计算教学的研究成果的基础上编写而成的，还包括笔者多年从事统计计算教学的心得体会。

2006 年初，笔者在日本访问期间，同志社大学的金明哲教授告诉笔者，即使在知识产权保护相当完善的发达国家，许多大学也在广泛采用 R 语言进行统计分析和教学，不仅因为它是免费的，还因为它是实时更新的（大约每三个月更新一次），更重要的是，它不断吸收最先进的统计技术。所以金教授建议笔者在国内开展 R 语言方面的研究，并积极鼓励笔者撰写 R 语言指导书，介绍 R 语言的特色和优势，于是促成了《R 语言统计分析软件教程》《多元统计分析及 R 语言建模》及本书的出版。

本书是国内第一本用 R 语言从数据处理角度编写的统计分析教程，这次修订主要扩展了三个方面的内容：

（1）对全书进行了适当的扩充和调整，每章增加了相应的练习题。

（2）优化了部分章节的代码和操作，公开了本书自编函数的源代码，使读者可以深入理解 R 语言函数的编程技巧，也使读者可以在不向作者索求开发包的情况下使用本书，并用这些函数建立自己的开发包。

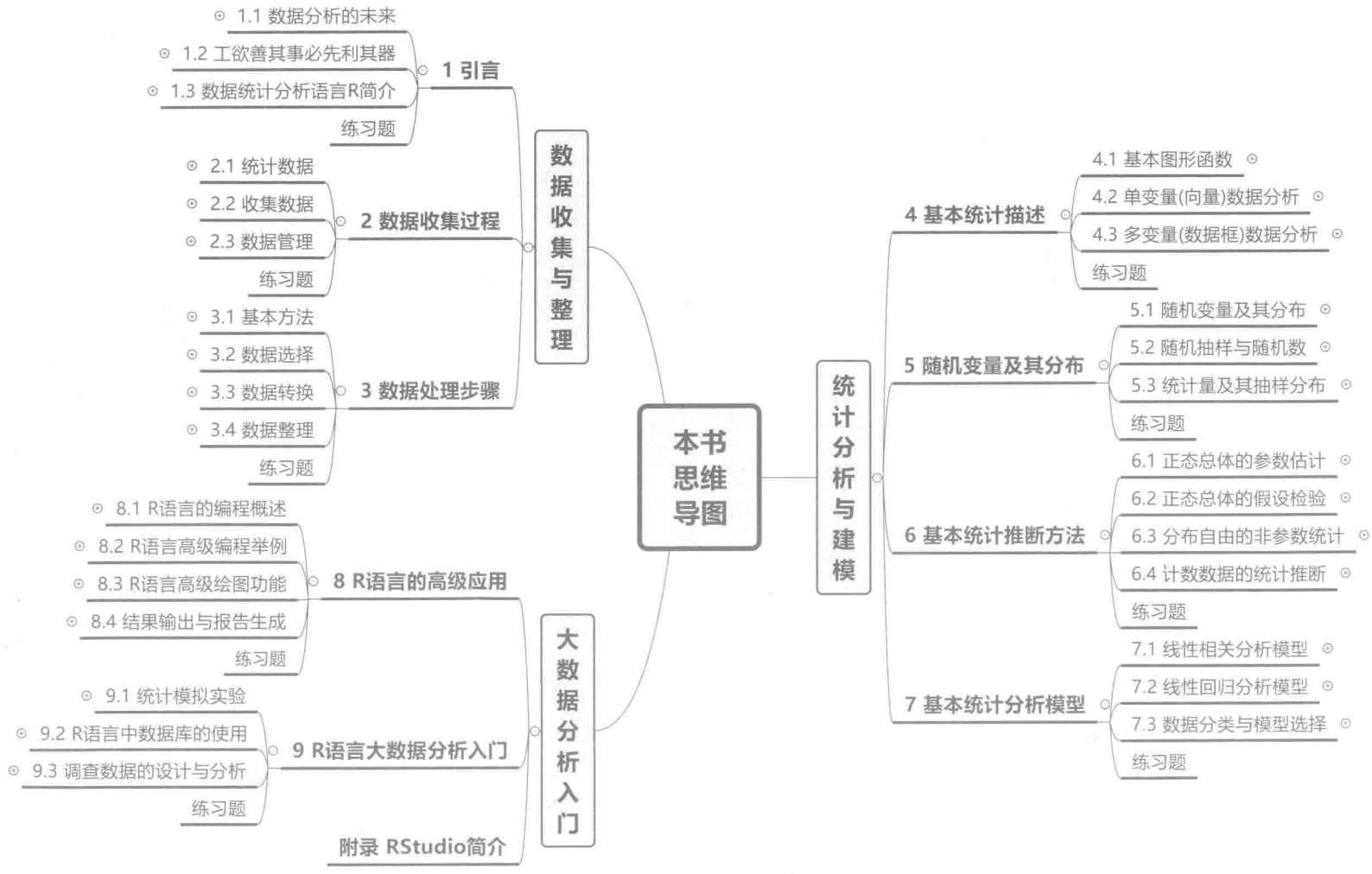
（3）建立了本书的 R 语言学习博客（Rstat. leanote. com），书中的数据、代码、例子、习题都可直接在网上下载使用。

本书的完成得到了暨南大学统计学系尹居良、侯雅文、谢贤芬老师，广东金融学院汪志宏、何志锋老师，广东财经大学王志坚、李雄英老师等的帮助；暨南大学统计学系研究生颜斌、徐锋、洪嘉灏、瞿尚薇、张佳萍、邓文、蒋鸽、史立新、刘弥然、赵子然、谢杰等人为本书的出版提供了一些资料和信息，在此深表谢意！

由于笔者知识和水平有限，书中难免有错误和不足之处，恳请读者批评指正！

王斌会

2017 年 4 月于暨南园


本书思维导图
数据收集与整理
1 引言
1.1 数据分析的未来
1.2 工欲善其事必先利其器
1.3 数据统计分析语言R简介
练习题
2 数据收集过程
2.1 统计数据
2.2 收集数据
2.3 数据管理
练习题
3 数据处理步骤
3.1 基本方法
3.2 数据选择
3.3 数据转换
3.4 数据整理
练习题
大数据分析入门
8 R语言的高级应用
8.1 R语言的编程概述
8.2 R语言高级编程举例
8.3 R语言高级绘图功能
8.4 结果输出与报告生成
练习题
9 R语言大数据分析入门
9.1 统计模拟实验
9.2 R语言中数据库的使用
9.3 调查数据的设计与分析
练习题
附录 RStudio简介
统计分析与建模
4 基本统计描述
4.1 基本图形函数
4.2 单变量(向量)数据分析
4.3 多变量(数据框)数据分析
练习题
5 随机变量及其分布
5.1 随机变量及其分布
5.2 随机抽样与随机数
5.3 统计量及其抽样分布
练习题
6 基本统计推断方法
6.1 正态总体的参数估计
6.2 正态总体的假设检验
6.3 分布自由的非参数统计
6.4 计数数据的统计推断
练习题
7 基本统计分析模型
7.1 线性相关分析模型
7.2 线性回归分析模型
7.3 数据分类与模型选择
练习题

目　录

1 引 言

1.1 数据分析的未来

市场上流行一个观点：数据越便宜，数据分析技术越昂贵。目前数据在中国很难获取，大家都把数据当资源来卖。国外就不一样，国外开放很多数据，因为国外的人认为，数据里面的信息才是资源。他们把数据源放开，若有本事就从里面寻找信息吧。所以，国外分析数据的人才的薪酬很高。

将来，中国的数据提供商必定会转型，会做咨询和分析，而不是单纯地卖数据。他们不卖数据了，数据分析师就开始值钱了。相信这一天很快就会到来！

1.1.1 趋势预测

下面我们通过一个例子来说明数据分析的未来趋势[①]。

① 谢益辉．数据科学家的崛起．（2012－11－25）．http：//cos. name/2012/11/the－rise－of－data－scientists/.

2012 年美国总统大选是奥巴马的胜利，但实际上也是统计学家的胜利。奥巴马当选之夜，我看见推特上有一条消息被疯狂转载：

NATE SILVER ELECTED 44TH PRESIDENT OF UNITED STATES

当然这是一句玩笑话，但 Nate Silver 是谁？他号称“竞选预测之神谕”：2008 年的总统大选他预测对了最终结果，而且美国 50 州的投票结果他预测对了 49 个；2012 年的总统大选他又预测对了，并且是 50 州全对。Silver 是一名统计学家，毕业于芝加哥大学，随后在毕马威会计师事务所“度过了令自己后悔的四年时间”（不喜欢那里的工作），后来转向预测棒球选手的成绩，再后来转向政治方面的数据分析和预测。总统大选的预测是一件噪声很大的工作，各家有各家的预测和分析，各种突发事件可能会导致某位候选人的支持率在短期内大幅波动。Silver 的工作就像机器学习中的“集成学习”（他自己的描述是“贝叶斯统计”，用自己的先验信息和数据得到后验信息），集合众多民意调查结果，根据自己的经验判断去平衡它们。

我想说的不是这个预测本身，而是我所感觉到的统计学家的变化。换成时髦的词就叫数据科学家。他们和具体的行业紧密相连，有扎实的统计基础，也有丰富的行业经验。不仅如此，大家都会玩编程、做数据可视化。看看 Silver 在纽约时报网站的博客就有感觉了。数据科学家正在“入侵”一些我们以前不能想象的行业，例如总统大选。除了 Silver 和其他一大批统计学家做预测之外，奥巴马还有一个数据分析部门，利用各种预测建模和数据挖掘手段来提高奥巴马连任总统的概率，例如他们有一则招聘广告就提到了 R、MySQL、Python 等工具。

1.1.2 数据科学家

你如果搜索一下“数据科学家”，就会看到有关它的各种光鲜的描述。很多光鲜的东西都是“坑”，当然这不是绝对的。媒体报道容易流于表面，这没什么奇怪的，数据科学家应该是一类综合人才，并不应该只懂一门技术的好手，例如纯统计。对统计学家来说，贝叶斯谁不会！半夜三点把你叫醒你都能三秒内背出贝叶斯定理，但让你把贝叶斯统计用到总统大选上，可能就没多少人做得了这件事情了。

数据科学家的概念是近几年在美国提出的，在中国发展如何，我们拭目以待。

下面是最近关于数据科学家的一则新闻，供大家参考[①]。

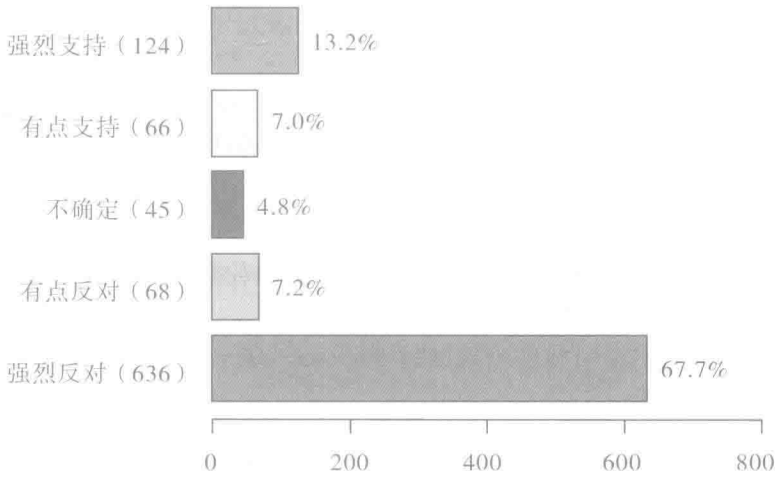
最近的 KDnuggets 民意测验显示出民众对特朗普移民禁令的强烈反对。测验的问题只有一个：你支持特朗普的移民禁令（13769 号行政令，“阻止外国恐怖分子进入美国的国家保护计划”或称为“禁穆令”）吗？

KDnuggets 组织的这次针对 1 000 名分析专业人士和数据科学家的民意测验，结果显示：约 75% 的全球受访者和约 77% 的美国受访者反对特朗普的移民禁令。这一民意测验反映出了很强的两极化特征，有着强烈观点的一方都占了显著多数。

KDnuggets 读者调查显示：全世界有近 75% 的数据科学家反对特朗普的移民禁令，但中国居然有 60% 的数据科学家表示支持！这实在让人难以琢磨。

① http://www.kdnuggets.com/2017/02/poll-data-scientists-oppose-trump-immigration-ban.html.

投票结果如下图所示（R 代码）[①]，总体显示出民众对禁令的强烈反对。综合“强烈反对”和“有点反对”的结论来看，约有75%的投票反对禁令，有20.2%表示支持，只有4.8%的人表示“不确定”。因为大部分问卷对象是数据科学家，所以该投票结果充分证明：全球的数据科学家们强烈反对特朗普的移民禁令。



我们也注意到结果中透露出的明显的特征。对移民禁令“强烈反对”的人数是“有点反对”人数的9倍多；“强烈支持”的人数也几乎是“有点支持”人数的2倍。

特朗普禁令遭到了全球范围内的反对，但令人讶异的是，各个地区仍有人会支持禁令，特别是在亚洲（主要是在中国、日本和印度）和非洲/中东（在以色列人的带动下），其支持度较高。

根据调查显示，下列国家中，一些国家（有5个以上受访者的国家）有80%及以上的数据科学家反对禁令；还有些国家则站在对立一方，这些国家（有5个以上受访者的国家）有30%及以上的数据科学家支持禁令。

反对的国家	韩国（100%）、丹麦（100%）、匈牙利（100%）、爱尔兰（100%）、瑞典（100%）、荷兰（92%）、土耳其（88%）、西班牙（86%）、德国（84%）、墨西哥（83%）、法国（80%）、巴西（80%）、澳大利亚（80%）
支持的国家	以色列（83%）、中国（60%）、日本（43%）、比利时（33%）、意大利（33%）、印度尼西亚（33%）、俄罗斯（33%）、印度（30%）

① S=c('强烈反对', '有点反对', '不确定', '有点支持', '强烈支持'); D=c(636,68,45,66,124);B=barplot(D,xlim=c(0,800),names.arg=paste(S,'(',D,')'),horiz = T,las=1,col=2:6);text(D,B,labels = paste(round(D/sum(D)*100, 1),"%"), pos=4)。

1.2 工欲善其事必先利其器

1.2.1 四大分析利器简介

要想成为一个优秀的数据分析师，就先必须掌握四大分析利器。

一、数据管理工具

1. 电子表格软件

这方面最为突出的有微软 Microsoft office 的 Excel，金山 WPS Office 的电子表格也是不错的选择。

2. 数据库管理软件

如常用的 Oracle、SQL Server、MySQL 等属于专门的数据库系统，本书不做介绍。

二、报告撰写工具

这方面最常用的文字编辑软件当属微软 Microsoft office 的 Word，而金山 WPS Office 的 Writer 也是不错的选择。

三、结果展示工具

这方面最为好用的当属微软 Microsoft office 的 PowerPoint，金山 WPS Office 的 Presentation 也是不错的演示工具。

四、数据分析工具

这方面的软件比较多，如常用的 SAS、SPSS 和 Matlab 等，还有后起之秀如 Stata 和 R 语言。这其中除了 R 语言外，其他皆为收费软件，而且价格不菲。R 语言不仅免费还开源，是一个跨平台系统。

前三类工具作为常用的办公软件，大多数人都会使用，也不是本书的重点，在此不做详细介绍，读者可到各自官方网站上了解。

1.2.2 四大分析利器的比较

一、办公软件比较

办公软件	安装文件大小	优势	不足
Microsoft office	约 1 000M	程序庞大，功能越来越强大	安装文件庞大，购买费用较高，升级频繁，兼容性较差
WPS Office	约 60M	小巧，免费，模板丰富，符合国人的使用习惯	功能有待加强，缺少数据库和绘图模块

二、统计分析软件比较

统计分析软件	安装文件大小	优势	不足
SAS	约 2 000M	程序庞大，功能越来越强大，可解决大数据问题	安装文件庞大，购买费用较高，升级频繁，兼容性较差，命令较多，编程困难
SPSS	约 800M	界面友好，操作方便	购买费用较高，升级麻烦，功能有待加强
R 语言	约 60M	小巧，免费，开源，附加包丰富	需要编程，入门较为困难

1.2.3 数据分析工具的选择

通过上面的分析，作为一个数据分析师，笔者认为可按下面的思路来选择数据分析工具。

一、首选 WPS + R

如果仅仅是做一般的数据统计分析，数据量不是特别大（十万级以下），而且要求系统免费、开源、跨平台，那么首选的数据统计分析软件组合应该是 WPS + R。

二、次选 Excel + R

如果你的数据量较小（65 536 行×256 列），使用的是 Windows 平台的话，考虑 Microsoft office（低于 2007 版）在国内的流行程度，也可考虑用 Excel + R，但 Excel 收费。

三、不差钱选 Access + SAS

如果你的数据量很大（十万级以上），使用的是 Windows 平台的话，一般用户可用 Microsoft office 的 Excel（高于 2007 版）+ SPSS（收费），企业用户可用 Microsoft office 的 Access（高于 2007 版）+ SAS（费用较高）。

四、专业选 Oracle + R

如果你的数据量是百万甚至是千万级，那么一般要使用专业的数据库软件进行分析，如 MS Sqlserver、Oracle 和 MySQL，本书暂不做介绍。

WPS Office 的电子表格与 Microsoft office 的 Excel 相互兼容，并有一致的操作界面，符合国人的使用习惯，WPS 的电子表格的缺点是不包含 Excel 的基本的数据分析模块。

综上所述，常规的数据分析师，特别是高校的教师和学生进行教学和科研时，选用 ET + R就可以了。如果你的电脑已经装有 Microsoft office，那么用 Excel + R 将是最好的组合。

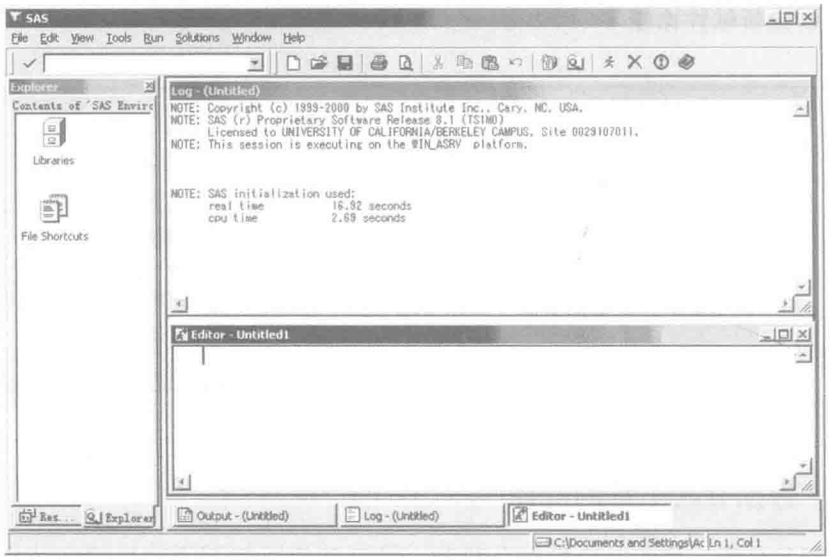
由于 Excel 已成为电子表格（ET）类软件的事实标准，因此我们下面在说法上将不区分电子表格和 Excel。

1.2.4 常用的数据分析软件

一、专业的数据分析软件 SAS

（1）优点：系统权威，内容全面，是数据处理和统计分析的标准软件。

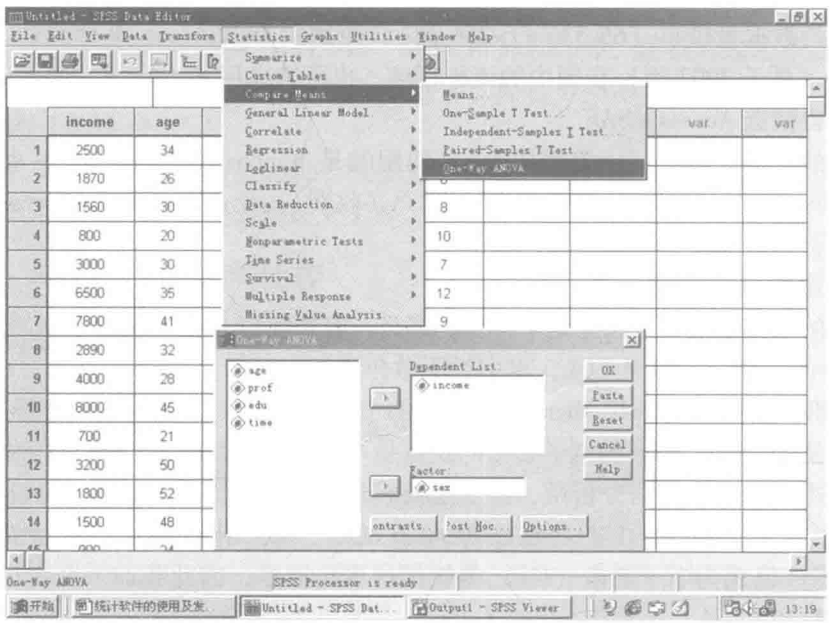
（2）缺点：系统庞大，编程复杂，购买费用较高。



二、方便的数据分析软件 SPSS

（1）优点：操作方便，使用简单，是非统计人员的首选。

（2）缺点：内容不全，编程麻烦，购买费用较高。



三、强大的数值分析软件 Matlab

（1）优点：编程方便，矩阵运算能力强大，是数值计算和图像处理的首选。

（2）缺点：统计方法不多，需一定的编程经验，购买费用较高。