STATISTICS

21世纪统计学系列教材

非参数统计

王 星 编著

中国人民大学出版社

21世纪统计学系列教材

非参数统计

王　星　编著

中国人民大学出版社

图书在版编目（CIP）数据

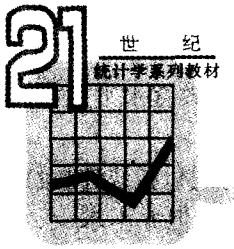
非参数统计/王星编著.
北京：中国人民大学出版社，2005
（21世纪统计学系列教材）
ISBN 7-300-06269-5

Ⅰ.非…
Ⅱ.王…
Ⅲ.非参数统计-高等学校-教材
Ⅳ.0212.7

中国版本图书馆CIP数据核字（2005）第006901号

21世纪统计学系列教材
非参数统计
王　星　编著

出版发行	中国人民大学出版社		
社　　址	北京中关村大街31号	**邮政编码**	100080
电　　话	010－62511242（总编室）	010－62511239（出版部）	
	010－82501766（邮购部）	010－62514148（门市部）	
	010－62515195（发行公司）	010－62515275（盗版举报）	
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com（人大教研网）		
经　　销	新华书店		
印　　刷	北京东方圣雅印刷有限公司		
开　　本	787×965毫米 1/16	**版　　次**	2005年1月第1版
印　　张	20.25	**印　　次**	2005年1月第1次印刷
字　　数	365 000	**定　　价**	30.00元（含光盘）



总 序

改革开放以来，高等统计教育有了很大的发展。随着课程设置的不断调整，有不少教材出版，同时也翻译引进了一些国外优秀教材。作为培养我国统计专门人才的摇篮，中国人民大学统计学系自 1952 年创建以来，走过了风风雨雨，一直坚持着理论与应用相结合的办学方向，培养能够理论联系实际、解决实际问题的高层次人才。随着新知识经济和网络时代的到来，我们在教学科研的实践中，深切地感受到，无论是自然科学领域、社会科学领域的研究，还是国家宏观管理和企业生产经营管理，甚至人们的日常生活中，信息需求量日益增多，信息处理技术更加复杂，作为信息技术支柱的统计方法，越来越广泛地应用于各个领域。

面对新的形势，我们一直在思索，课程设置、教材选择、教学方式等怎样才能使学生适应社会经济发展的客观需要。在反复酝酿、不断尝试的基础上，我们决定与统计学界的同仁，共同编写、出版一套面向 21 世纪的统计学系列教材。

这套系列教材聘请了中科院院士、中国科技大学陈希孺教授，上海财经大学数量经济研究院张尧庭教授，中国科学院数学与系统科学研究所冯士雍研究员等作为编委。他们长期任中国人民大学的兼职教授，一直关心、支持着统计学系的学科建设和应用统计的发展。中国人民大学应用统计科学研究中心 2000 年已成为国家级研究基地，这些专家是首批专职或兼职研究人员。这一开放性研究基地

的运作，将有利于提升我国应用统计科学研究的水平，也必将进一步促进高等统计教育的发展。

这套教材是我们奉献给新世纪的，希望它能促进应用统计教育水平的提高。这套教材力求体现以下特点：

第一，在教材选择上，主要面向经济类统计学专业。选材既包括统计教材也包括风险管理与精算方面的教材。尽管名为统计学系列教材，但并不求大、求全，而是力求精选。对于目前已有的内容较为成熟、适合教学需要、公认的较好的教材，并未列入本次出版计划。

第二，每部教材的内容和写作，注意广泛吸收国内外优秀教材的成果。教材力求简明易懂、内容系统和实用，注重对统计方法思想的阐述，并结合大量实际数据和实例说明统计方法的特点及应用条件。

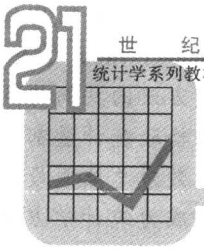
第三，强调与计算机的结合。为着力提高学生运用统计方法分析解决问题的能力，教材所涉及的统计计算，要求运用目前已有的统计软件。根据教材内容，选择使用 SAS、SPSS、TSP、STATISTICA、EViews、MINITAB、Excel 等。

感谢中国人民大学出版社的同志们，他们怀着发展我国应用统计科学的热情和提高统计教育水平的愿望，经过反复论证，使这套教材得以出版。感谢参与教材编写的同行专家、统计学系的教师。愿大家的辛勤劳动能够结出丰硕的果实。我们期待着与统计学界的同仁，共同创造应用统计辉煌的明天。

易丹辉

2000 年 8 月

于中国人民大学



前 言

随着信息技术的发展，无论是试验数据还是观测数据都能轻松地从各行各业产生，亟待处理分析以产生应有的价值。由此产生的统计问题也在规模和复杂性上激增。一些复杂的随机环境，如生物医学、经济社会、航天太空等领域，由于无法对产生数据的总体进行过分细致的假定，而导致庞大而复杂的参数模型的出现，比如有的金融模型即便有成千上万个参数来描述，也未必能尽如人意。

现实世界中，哪里有不确定性，哪里就有统计。随着科技的进步和测量仪器越来越先进，我们发现自然界和人类社会的不确定性不仅没有消失，反而一直在增加，增加的数量远远超出人类为获得仅有的一点认识所付出的努力和探索。比如有的金融模型甚至需要成千上万个参数。与方法不同的是，观察数据却在飞快地增长，并试图占据数据库的每个角落。因而有人形容目前的数据分析市场呈现出一派“数据繁荣，知识匮乏”的尴尬境况。而“知识匮乏”表达了人们对与数据和问题相适应的研究方法的强烈的渴求。在现实的驱使下，整个数据分析方法都朝向模型假定越来越宽松的方向发展，这也许就是非参数统计自20世纪40年代形成后的60年间能广泛应用，并获得蓬勃发展的最好理由。

像所有统计学方法一样，非参数统计方法也分为两大部分，一为总体参数之间的差异性比较，内容主要涉及单一总体、两总体以及多总体位置或尺度的非参

数估计和假设检验。这些方法基本形成于20世纪40年代到60年代，它们构成了传统非参数统计的基本内容，这个时期的代表人物有Wilcoxon和Pitman等人。最近几年，随着新问题的不断涌现，新的方法也发展很快。这些方法的共同特点是思想鲜明、原理简单、计算简便、易于应用。大部分估计量或检验统计量都可能通过中心极限定理，建立较完整的估计和检验理论。由于非参数对总体假设较参数统计更加宽泛，加之有更宽的适用性，从而也使非参数统计成为统计学方法论体系中不可或缺的重要分支。传统的非参数统计的第二部分内容是不同变量之间统计关系的研究，特别是分类（定性）数据的研究，比如：以χ^2列联分析为代表的定性变量的独立性研究。进入70年代以后，随着计算机技术的发展，无论在描述数据的分布上，还是在刻画数据之间的关系方面，与传统相比，非参数技术都获得了质的飞跃，比如非参数密度估计和非参数回归等。这些方法不仅对数据来源的总体几乎没有要求，而且允许对数据进行大量的计算，从而获得更为稳健及精确的结果，其中的主要代表人物有Silverman，J.Fan等。

为兼顾传统与现代，本书也自然分为两个部分，第2章至第7章主要介绍传统的非参数统计方法，其中第2章是非参数统计的基本概念和学习非参数统计所必备的一些知识，第3、4、5章分别介绍传统的单一总体、两总体以及多总体非参数统计估计和假设检验的内容。接下来的第6章主要关注定性数据的关联分析方法，而第7章则是定量数据之间的相关关系和回归系数的非参数推断方面的内容。第8章和第9章是现代非参数统计的一些基本概念和方法，其中第8章介绍非参数密度估计，第9章则是非参数回归的基本原理和应用。

中国人民大学统计学院在最近几年的非参数统计课程教学实践经验基础上，逐步摸索出一套运用S-Plus或R学习非参数统计的教学方法。在我们的课堂上，学生可以通过学习非参数统计提高用程序设计语言表达统计分析过程的能力。在几年的教学实践中，我们也发现部分学生在学习之初对S-Plus或R有一定的畏难情绪。与菜单式的软件相比，学习S-Plus或R的确需要花费更多的时间，特别是又要配合非参数统计的各种方法来设计和实践、体会方法的各种应用技巧。然而，与可编程软件相结合教授统计方法的模式，是国外知名统计院校的普遍做法，因为单纯将统计方法当做一个工具，通常并不能真正解决分析建模问题。而在学习方法的过程中，加强学生的实践和操作，将提高学生对统计方法的综合运用能力。另外，考虑到目前统计书籍市场，用中文介绍S-Plus或R语言的书籍非常稀缺，因此，作者在第1章中，以S-Plus为例，介绍S语言的实践基础，这些内容完全适用于R语言。实践表明，阅读完第1章内容并结合部分习题上机实习，基本能够掌握S-Plus和R软件的使用技巧。此外，本书大部分例题都给出

S-Plus 程序，解释并分析输出结果，鼓励学生自己编写非参数程序。为方便初学者，本书从第 3 章至第 8 章每章习题之后还给出当前一章的部分 S-Plus 源程序代码示例。

本书适用于统计、经济、管理、生物等宏观、微观专业领域本科三、四年级以上学生以及相关研究人员学习非参数统计的教材，也可以用做统计研究或从事数据分析的参考书。本书的先修课程只需具备初等统计学基础。对统计基础略感陌生的读者，可以浏览第 2 章相关内容作为补充。本书的内容可以安排在一学期 54 课时内完成，建议安排 10 课时左右用于学生集中上机实践。本书备有丰富的习题，兼有理论推导、方法应用和程序设计题目，我们希望学生通过足够数量的习题练习掌握非参数统计方法，并选择合适的环境，正确使用这些方法。为便于读者学习，本书制作了幻灯片，其中收录了书中部分例题数据和每章的教学大纲。

值得一提的是，我在人大统计学院的老前辈、非参数统计的倡导者吴喜之先生，在本书的写作过程中，给予我无尽的鼓励和支持，对本书提出诸多宝贵建议。另外，李舰同学提供了部分 S-Plus 和 R 函数程序，罗玉波同学帮助解决了部分图形排版技术问题，并帮助制作幻灯片，王军侠同志在书稿录入和排版方面做了大量工作，出版社的编辑纠正了很多疏误。在此一并向这些同志表示诚挚的谢意！

由于作者水平及时间所限，书中错误在所难免，恳请同行专家和广大读者给予批评指正。

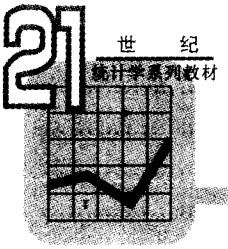
通信地址：中国人民大学统计学院　王星（收）　邮编：100872

E-mail：wangxing@ruc.edu.cn

王　星

中国人民大学统计学院

2004 年 11 月



目 录

第1章

S-Plus 基础

1.1 S-Plus 环境

S语言和S的扩展 S-Plus，是由 AT&T Bell 实验室于 20 世纪 70 年代末 80 年代初研制开发的，其中 Rick Becker 和 John Chambers 是主要创始人之一。现在流行的版本是 S-Plus 6.0，而就在本书写作过程中，S-Plus 6.2 正在试验室阶段。在西方发达国家，S-Plus 被作为统计专业软件，是学习统计方法和从事统计研究的本科高年级以上学生和统计研究人员必备的统计计算工具。像大部分传统的统计软件一样，S-Plus 环境也提供了现代的菜单和人机交互的操作方式，用户可以通过点击鼠标快捷地选择模型分析数据。然而，菜单的处理数据环境并非 S-Plus 最初设计的本意，也不是 S-Plus 最突出的特点，它的点选模式是最近几年为满足商业应用才开发出来的。S-PLus 的主要应用特点如下：

1.S-Plus 拥有强大的面向对象的开发环境，其中所有的函数、数据及模型都可被视为对象，用户可以灵活自如地应用 S 语言编写定制从对象操作、数据处理、尝试模型到评估结果的一个数据整体处理方案。

2. 作为标准的统计语言，S 拥有几乎所有特别是最新的统计计算函数，有些更新的统计方法还可以从网上下载。不仅如此，在 S-Plus 中，用户可以随处自定义各种函数，延伸基本的分析方法。

3. 作为面向对象的语言，S 集数据的定义、插入、修改和函数计算等功能于一体，语言风格统一，可以独立完成数据分析生命周期的全部活动。

4. S-Plus 提供了非常丰富的 2D 和 3D 图形库，不仅可以帮助数据分析员分析数据，而且对已生成的图形，用户可以用修改其中每个细节，调整图形的属性满足报表报送需求。另外，S-Plus 图形还可以非常方便地输出到 Latex 等正式出版的文章编辑器中，从而生成高品质的科技文章。

5. 与 S-Plus 类似的有新近开发的 R，R 最早由 Auckland 大学统计系的 Robert Gentleman 和 Ross Ihaka 于 1995 年开始研制开发，从 1997 年起免费公开发布。R 在计算功能上比 S-Plus 有更高的效率，但由于推出时间较短，因而不如 S-Plus 普及。S-Plus 和 R 二者在语法和功能实现上区别不大，因而学习 S-Plus 的用户可以轻松转到 R，反之亦然。

总之，S 使数据分析员能更好地思考和生成数据分析的方法，它的弹性和可扩展性，为灵活分析数据和学习统计方法提供了一个良好的实验平台。值得注意的是，S-Plus 可以在 Windows 和 Unix 两种不同的操作系统中运行，因而有更广的商业应用。但考虑到本书主要针对学生作为教材之用，所以本章所讲全部针对 Windows 下的 S-Plus 的语法和操作。

1.1.1 用户界面

在 Windows 操作系统中启动 S-Plus，可以见到如图 1—1 所示的图形界面。虽然不同的 S-Plus 版本界面略有不同，但主要的内容还是类似的，其中一般都会包括六项内容：Object Explorer（对象浏览器），Commands Window（命令窗口），Graph Sheets（图形操作表），Script Window（草稿窗口），Menus（菜单）和 Toolbars（工具栏）。后面两项功能与 Windows 相应的功能类似，这里不再赘述，其中前四部分是 S-Plus 特有的界面，其主要功能如下。

1. Object Explorer（对象浏览器）。对象浏览器的目的是快速浏览系统中可供操作的对象，包括数据、函数和图形。它的操作方式类似于 Windows 资源管理器，在左侧用树形结构列出不同对象类型，右边显示每一类对象的具体内容。双击对象可以编辑对象。

2. Commands Window（命令窗口）。命令窗口是 S-Plus 的核心部分。所有可以通过 S-Plus 菜单执行的命令都可以在命令窗口通过输入提交命令的方式来执

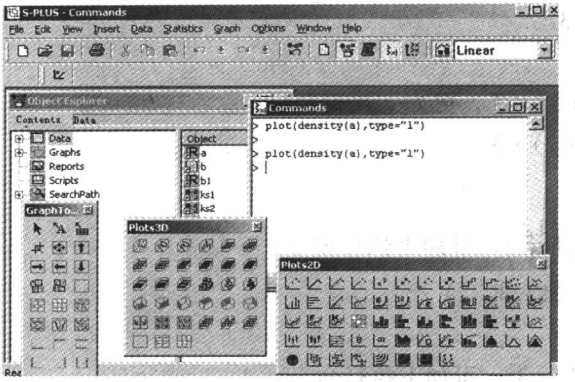


图 1—1　S-Plus 6.0 启动界面

行，反之不行。也就是说，S-Plus 中大部分的命令只能通过函数的提交方式得以执行，菜单中只显示出 S-Plus 中强大功能的很少一部分。正因为如此，S-Plus 才被誉为可编程的个性化的软件。用户可以在其中创建自己的函数，书写自己的数据处理流程，延伸数据分析方法。

3.Graph Sheets（图形操作表）。图形操作表的作用是编辑图形，用户只要点击图形上所感兴趣的任何部分就可以修改这些内容，比如图标、颜色、坐标轴、标题等都可以重新定义和更新。Insert 菜单和 Format 菜单可以辅助完成这些任务。Insert 菜单可以在图形操作表中加入各种图形组件，Format 菜单用于移动现存图形中组件的位置。

4.Script Window（草稿窗口）。草稿窗口由上下两部分构成，上面的面板用于起草程序代码，下面的面板用于程序的说明和备注。在草稿窗口中，也可以像在命令窗口中一样执行程序，所不同的是在 Script 中，可以执行一批命令。因而当一批命令全部执行完成之后，除非中间设置屏幕打印命令（输出图形也算屏幕打印），用户只能看到最后一行的计算结果，不利于检查程序错误。在命令行中，一次只执行一行命令，因而实现了完全的人机交互，但不利于及时保存程序。因而程序员应交互利用两个窗口编写程序。

1.1.2　算术运算

一、算术运算

S-Plus 默认的命令提示符是“>”，如果用户在命令窗口下看到这个“>”，就可以在其后进行运算。

1. 计算 7×3，可如下执行命令：

```
> 7 * 3
> 21
```

2. 计算$(7+2)\times 3$，可如下执行命令：

```
> (7+2) * 3
> 42
```

3. 计算$\log_2\left(\frac{12}{3}\right)$，可如下执行命令：

```
> log (12/3, 2)
> 2
```

4. 计算幂2×3^2，可如下执行命令：

```
> 2 * (3^2)
> 18
```

常用的初等函数如下：sin()，cos()，tan()，exp()；log(N,a)表示$\log_a N$，查看内存中的对象用objects()。而要了解S-Plus中的所有可处理的对象，用library()命令。

二、赋值

给变量赋值用“<-，或”，两个字符串，比如将3赋给变量x，修改变量y的取值为4，可以使用命令：

```
>x<-3
```

或

```
> y_1+x
```

屏幕打印变量x如下：

```
> x
> 3
```

或

```
> print(x*y)
> 12
```

1.2 向量的基本操作

一、连接命令

前一节，我们介绍的运算都是对单个标量数据的操作。然而，统计处理中最

常见的是对向量整体的操作和变换。比如，超市里某一品牌或某一类商品价格打8折，这时需要对这个品牌或类的所有商品价格进行乘0.8的运算，等等。如果事先将这些价格输入一个向量，再对这个向量的每一个元素执行同样的操作，就可能提高运算效率。在S中，用于创建任意大小的数据向量的命令是c（catenation的缩写）语句，用户只需将组成向量的每个元素列出，并用c组合起来即可。基本运算如下：

【例1.1】

```
> A.brand< - c(15,27,89)
> A.brand
  [1]  15  27  89
```

对整个向量处理的语法与标量的处理完全一致，如：

```
> A.brand * 0.8
  [1]  12.0  21.6  71.2
```

上面的程序将向量中每个元素乘0.8，其他运算完全类似。生成非数值类型的向量也可以使用c语句，如下所示：

```
> title.text< - c(" This" " is" " for" " sale")
> title.text
  [1]  " This" " is" " for" " sale"
```

定义向量之后，可以在对象浏览器（Object Explorer）中看到新定义向量的名称，双击该名称可以在一个电子表格中对数据的各部分进行修改和编辑操作，完全类似于Office Excel的操作，这不是我们关注的重点。实际上，对向量的操作并不需要将数据集完全打开。此处我们重点介绍在命令窗口下如何对向量操作，这些操作主要包括查找数据、插入数据、更新数据、删除数据、向量与向量的合并、拆分向量等。在S-Plus中，这操作符具有统一性的特点，这为用户学习带来极大的方便。

1. 向量 a 中第 i 位置的元素表示

向量 a 中第 i 位置的元素表示为：$a[i]$，比如：

【例1.1续】

```
> A.brand[1]
  [1]  15
```

```
> A.brand[length(A.brand)]
  [1]  89
```

如果输入的位置超出向量的长度，则 S-Plus 输出 NA。NA 表示数据缺失，如下所示：

```
> A.brand[6]
  [1]  NA
```

提取向量 a 的第 $i_1, i_2, \cdots, i_k$ 位置上的元素的语法为：$a\{c(i_1, i_2, \cdots, i_k)\}$。

也可以用两步法定义一个新的向量表示要查找的位置，然后再提取这些数据，例如：

```
> position. a< - c (1, 3, 5)
> subset. a< - A. brnd [position.a]
  [1]   15  89  NA
```

2. 在向量中插入新的数据

在 A.brand 向量末尾添加两个新产品价格的书写方法是：

【例 1.1 续】

```
> add.a.brnad< - c(A.brand,189,240)
> add.A.brand
  [1]  15  27  89  189  240
```

在向量的开头插入新数据的书写方法如下：

```
> addbegin.A.brand< - c(12,13,A.brand)
  [1]  12  13  15  27  89
```

思考题 1.1　如何在向量的第 i 个位置后插入数据？

3. 向量与向量的合并将 A.brand 和 B.brand 两个向量合并为一个新向量的书写方法是：

```
> B.brand< - c (35, 40, 58)
> AB.brand< - c (A.brand, B.brand)
  [1]   15  27  89  35  46  58
```

向量中元素的个数，称为向量的长度，查看向量中元素的个数，可以使用 length 命令，如下所示：