自由的统计语言

谢益辉 主编

2008 年 12 月 15 日

目　　录

历史篇 **5**

基础篇 **11**

应用篇 **15**

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 1–3 页

第一届中国 R 语言会议简介

谢益辉[1,*]

[1] 中国人民大学统计学院
北京市海淀区中关村大街 59 号中国人民大学明德主楼 1037，北京 100872

R 是一门用于统计计算和作图的语言，它最初由新西兰奥克兰大学统计系的 Robert Gentleman 和 Ross Ihaka 合作编写，它与 S 语言非常相似，加上它是自由软件，因此又被称为 GNU S。自 1997 年开始，R 语言开始由一个核心团队开发，团队成员来自世界各地的大学和研究机构。随着 R 语言在国际上的飞速发展以及广泛应用，国内也有部分研究者开始学习使用R。R 官方机构于 2004、2006、2007、2008 年在奥地利、美国和德国分别举办了 useR! 会议，而且近两年内还没有在欧洲及美国之外的地区举办官方会议的计划，因此我们决定在北京举办国内首次 R 语言会议，期望对国内 R 用户提供学习和研究上的帮助，同时也为精通 R 语言的专家提供集中交流的机会。

1 会议背景

2008 年 10 月 5 日下午 2 时，我在 COS 论坛上看到 S-Plus & R 讨论区的文章数目为 9336 篇，然后打开 R 作了如下计算：

```
> (x = difftime("2008-10-05", "2006-05-19"))
Time difference of 870 days
> 9336/as.integer(x)
[1] 10.73103
```

COS 全称为 “Capital of Statistics”，中文意即 “统计之都”，网址 `http://cos.name`；论坛创办于 2006 年 5 月 19 日，创办宗旨是为国内广大统计学相关领域的学者、研究人员和应用者提供无偿统计技术支持。刚才 R 告诉我，网站已经成立 870 天，在 S-Plus & R 讨论区平均每天大约有 11 篇新的帖子诞生，这在国内 R 语言领域算得上是非常活跃的地带。事实上，中国国内目前为止专门提供 R 语言讨论的网站据我所知也不过两三家[1]。

如果说平均每天 11 篇新帖子这个数据并不那么令人振奋，那么我们不妨再仔细查探一下每天的原始数据。图 1 展示了 COS 论坛 S-Plus & R 讨论区文章数目的时间序列，上图

*电子邮件：xie@yihui.name；主页：http://www.yihui.name
[1] 另有一家 http://rbbs.biosino.org/Rbbs/

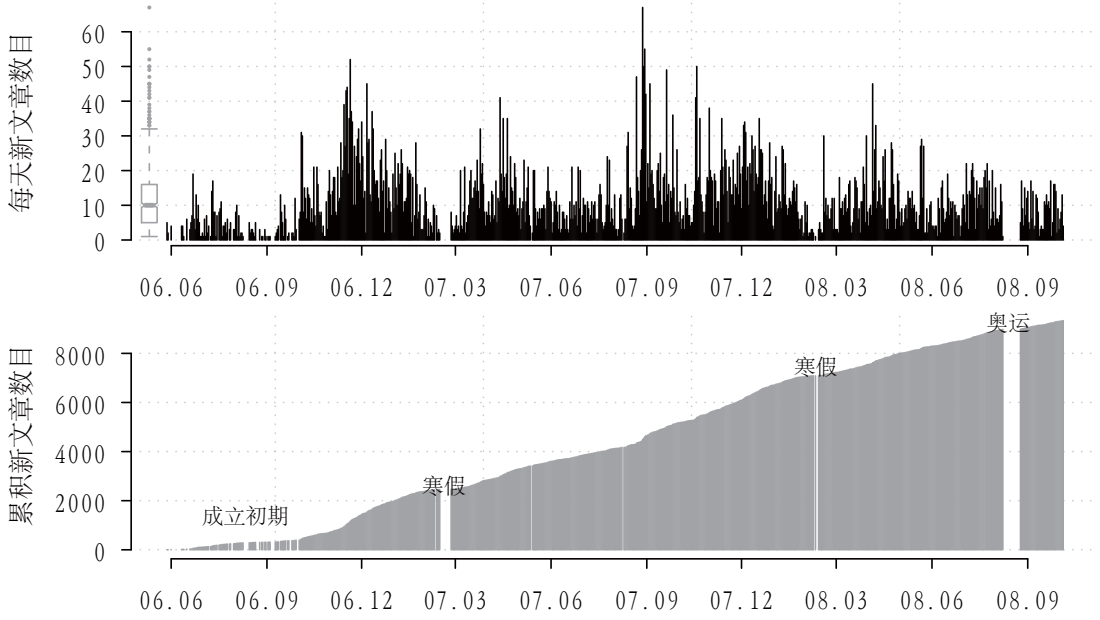

图 1: COS 论坛 S-Plus & R 版块文章数目时间序列图

为每天帖子数量，下图为累积数量。可以看出，该讨论区在 06 年的 12 月、07 年的 4 月、9 月、12 月以及 08 年 4 月每天新帖数量甚至能够超过 40 篇。总的说来，图 1 告诉我们的信息是：

- 两年多以来，S-Plus & R 讨论区的发帖数量维持在较高的水平（均值为 10.7，中位数为 10），但从图 1 上图可知发帖数量波动性很大，左侧的箱线图中大量的离群点也反映出这个特征
- 累积发帖数量几乎呈直线持续上升，说明国内的 R 用户需求比较稳定，一直对 R 语言的应用存在诸多疑惑

管窥中，我们看到了局部的需求高峰和不稳定性以及全局的稳定性，第一届中国 R 语言会议因此在中国人民大学统计学院和应用统计科学研究中心的支持下拉开序幕。

2 会议目标

国外 R 语言的应用已经非常普遍，它开源、免费、统计功能强大的特点吸引了众多学术单位和科研机构甚至公司企业，相比之下它在国内的应用则显得冷冷清清。本次会议设立了三大目标：

1. 聚集各行各业的 R 语言用户，展示 R 在统计应用中的多种灵活功能，扩大参会者的知识面

2. 训练出一批 R 语言先锋队，推进 R 在国内的应用步伐
3. 编写 R 语言学习材料，展示 R 语言的统计功能

我们希望本次会议的参会者能够领会到 R 语言中体现出来的“开放、共享、互助”的精神，在自我学习的同时也将它介绍到更广大的统计相关研究者中去。本论文集即为第二目标的实现之一，除此之外，我们还希望能够建立一支具备扎实技术功底的 R 队伍，在未来的时间里以翻译英文材料、编写小文档等形式为大家扫清学习障碍。

3 会议成果

本次会议取得了如下成果：

1. 两天会议时间里举行演讲报告 20 余场
2. 收到论文 20 余篇，编撰为论文集《自由的统计语言》
3. 建立了 R 帮助文件翻译团队，COS 维基（`http://cos.name/wiki`）提供了便利的合作翻译平台

4 未来发展

第二届？……第 N 届？……

致谢

感谢中国人民大学应用统计科学研究中心提供会议经费支持，感谢广大 R 爱好者参加本次会议，感谢会议服务的同学们。

历史篇

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 7–7 页

A Technical Introduction to the History of R

Kurt Hornik[1,*]

[1] Department für Statistik und Mathematik, Wirtschaftsuniversität Wien
Augasse 2-6, A-1090 Wien, Austria

Abstract This paper gives a technical introduction to the history of R.

Keywords R language; History

*Email: Kurt.Hornik@wu-wien.ac.at; Homepage: http://www.wu-wien.ac.at/cstat/hornik/

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 8–8 页

R 的那些事儿

丁国徽[1,*]　　李虹[2]

[1,2] 中国科学院上海生命科学研究院

摘要　本文对 R 的历史，体系结构，以及当前中国现状进行一个简要的概述。

关键词　R 语言；历史；体系；现状

A Leisure Look at the R Language

Guohui Ding[1,*]　　Hong Li[2]

[1,2] Shanghai Institutes For Biological Sciences

Abstract　This paper briefly describes the history and structure of R as well as the current situation of R in China.

Keywords　R language; History; Structure; Current situation

*电子邮件：ghding@gmail.com
*Email: ghding@gmail.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 9–10 页

R 语言的历史背景、发展历程及现状

谢益辉[1,*]　　郑冰[2]

[1] 中国人民大学统计学院
北京市海淀区中关村大街 59 号中国人民大学明德主楼 1037，北京 100872
[2] 杭州电子科技大学财经学院
浙江省杭州市下沙高教园区杭州电子科技大学财经学院统计学专业，杭州 310018

摘要　R 是一门用于统计计算和作图的语言，受 S 语言影响发展而来。R 语言最初由新西兰奥克兰大学统计系的 Robert Gentleman 和 Ross Ihaka 合作编写。自 1997 年开始，R 语言开始由一个核心团队开发，团队成员来自世界各地的大学和研究机构。迄今为止，R 源代码已经经历了近 70 次主要更新，功能也在不断完善、增强中，主要统计功能包括线性模型/广义线性模型、非线性回归模型、时间序列分析、经典的参数/非参数检验、聚类和光滑方法等。R 语言具有免费、开源及统计模块齐全的特征，已被国外大量学术和科研机构采用，其应用范围涵盖了数据挖掘、机器学习、计量经济学、实证金融学、统计遗传学、自然语言处理、心理计量学和空间统计学诸多领域。相比之下，R 语言在国内的发展显得比较落后。因此本文对 R 语言的历史背景、发展历程及现状分别作出介绍，以期引起相关学者和研究人员的注意，并推动 R 语言在中国的广泛应用。

关键词　GNU；R 语言；S 语言；历史；现状

The History of R Language and Current Developments

Yihui Xie[1,*]　　Bing Zheng[2]

[1] School of Statistics, Renmin University of China
Room 1037, Mingde Main Building, Renmin University of China, Beijing, 100872, China
[2] College of Business, Hangzhou Dianzi University
College of Business, Hangzhou Dianzi University, Hangzhou City, Zhejiang, 310018, China

Abstract　R is a language and environment for statistical computing and graphics; it was initially written by Robert Gentleman and Ross Ihaka. R is often considered as the GNU S, from which much code has been written into R without modifications. Since mid-1997 there has been a core group with write access to the R source. The R distribution contains functionality for a large number of statistical procedures. Among these are: linear and generalized linear models, nonlinear regression models, time series

*电子邮件：xie@yihui.name；主页：http://www.yihui.name
*Email: xie@yihui.name; Homepage: http://www.yihui.name

analysis, classical parametric and nonparametric tests, clustering and smoothing. Now R has been applied to a wide range of areas including data mining, machine learning, econometrics, chemometrics, computational physics, econometrics, empirical finance, statistical genetics, natural language processing, psychometrics, and spatial statistics, etc. Compared with the rapid overseas developments and applications of R, the current situation of R in China still needs a lot of efforts in terms of the popularization of its powerful statistical functionalities, therefore we provide an introduction to the history and developments of R to arouse the attention of more people in academic and scientific research fields.

Keywords GNU; R language; S language; History; Development

基础篇

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 13–13 页

R Frequently Asked Questions

刘思喆 (sunbjt<at>gmail.com)

April 3, 2008

本文档内容来源多样，既有来自于 R 官方文档，也有来自于互联网的非正式捐赠文档；还有若干来自于 COS 论坛的讨论问题。

文档的最初目的是为初学者提供一个快速认识 R 软件的平台，所列问题不可能详尽，如果想系统学习 R 软件，建议大家精读一本关于 R 的原版书籍。

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 14–14 页

R 在植物生态生理学中的应用 —— 从数据管理到图表制作

代占武[1,*]　　魏太云[2,†]

[1] 波尔多法国农科院，葡萄生态生理与功能基因研究组，33883
[2] 中南大学数学科学与计算技术学院，410075

摘要　R 是一个开源软件，功能强大，越来越受到各个领域研究者的青睐，其在植物生态生理学中的应用在国际上日渐普遍，但尚有待国内植物生态生理学者的发现和运用。本文以一个试验为例，简要介绍了 R 的数据输入输出、方差分析、多重比较、图表制作、线性相关性分析及其线性方程的拟合等操作方法，注重实用和可重复性。其中图的制作部分，给出了植物生态生理论文发表过程中多图组合及不同图形 (柱状图和折线图) 的代码。它们可以作为模板，稍加改动便可轻松运用于研究者个人研究论文的发表过程中。

关键词　R；植物生态生理；数据管理；图表制作

Using R in Plant Ecophysiology — from data management to result presentation

Zhanwu Dai[1]　　Taiyun Wei[2]

[1] UMR-1287 EGFV, Bordeaux INRA, 33883 Bordeaux, France
[2] School of Mathematical Science and Computing Technology, Central South University, 410075

Abstract　R, an open source software with powerful potential in data analysis, has been extensively used in many scientific research areas. In plant ecophysiology, more and more researchers are adopting R for their data management, however, its usage needs further discovery for researchers in China. Here, taking an experiment dataset as an example and with special emphases on practicability and reproducibility, we briefly showed how to use R to input and output experimental data, to perform analysis of variance and multiple comparisons, to form tables, to draw figures, and to conduct liner regression. In the section of graphics presentation, scripts were provided for combining multiple figures and choosing different types of figures (including barplot and line-point figures with error bars), which are commonly used in publication of experimental results in plant ecophysiology. These scripts can be readily adopted in reader's research and publication with minor modification.

Keywords　R; plant physiology; data analysis; R graphics

*电子邮件：zhwdai@yahoo.com
†电子邮件：weitaiyun@gmail.com

应用篇

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 17–17 页

Data Mining with R

John Maindonald [1,*] Yihui Xie[2,†]

[1] Centre for Mathematics and Its Applications, Mathematical Sciences Institute
Building 27 (John Dedman Mathematical Sciences Building)
Australian National University, Canberra, ACT 0200, Australia
[2] School of Statistics, Renmin University of China
Room 1037, Mingde Main Building, Renmin University of China, Beijing, 100872, China

Abstract The name "Data Mining", commonly used to describe a style of data analysis that makes a virtue of exploratory approaches, emerged from the computer science community. In a recent book, Ken Berk describes it as a "muscular" version of EDA (Exploratory Data Analysis). Statistical Learning and Machine learning draw from similar streams of ideas, and have similarly strong connections into computer science, but may pay more attention to the literature and traditions of probability theory and of theoretical statistics. "Analytics", focusing on applications in business and commerce, is another name that has come into wide use in recent years.

This talk will offer a statistician's view of these different names for data analysis, with their differences in style, concepts, terminology and notation. It will comment on the challenges and innovations that they have fostered. It will comment on common deficiencies in the frameworks of understanding and theory, arising in part from limited attention to insights from the statistical tradition. It will comment on key ideas. Finally, it will comment on what R offers to these diverse communities, in specific analysis tools, as a unifying framework for development of new abilities, and as a means of access to a wide range of methodologies.

Keywords R; Data Mining; Statistical Learning; Machine learning; Analytics

[*] Email: john.maindonald@anu.edu.au; Homepage: http://wwwmaths.anu.edu.au/~johnm/
[†] Email: xie@yihui.name; Homepage: http://www.yihui.name

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 18–18 页

数据挖掘过程在 R 环境下的应用

刘思喆[1,*]

[1] 北京中彩在线科技有限责任公司
北京宣武区广安门大街 48 号、100054

摘要 数据挖掘是指“从数据中提取出隐含的特别的，过去未知的，潜在有价值的信息”[5] 的一门学科。它可能涉及到数据仓储 (数据库)、机器学习和人工智能、统计学、高水平计算、并行计算、数据可视化等多项技术。而 R 作为一种高度灵活的统计分析、绘图软件，不但拥有庞大的、持续更新的包 (包括贝叶斯推断、分类方法、计量经济学、生态学、金融学、遗传学、机器学习、稳健统计、空间统计、生存分析、时间序列等多个方面) 来扩展其数据分析方法，而且还有着便捷的接口来调用数据库，这些都为 R 作为数据挖掘平台提供了基本技术保障。本文从实际应用角度介绍了 R 语言在数据挖掘方面的优势，并举例说明了 R 在数据挖掘流程中的各项技术。

关键词 数据挖掘；R；数据可视化

The Application of Data Minging with R

Sizhe Liu[1,*]

[1] China Lottery Online Ltd.co
No.48 Guang'anmen avenue, Xuanwu District, Beijing , 100054

Abstract Data Mining is a discipline which discovery unknown , hidden knowlege of interest. The application of data mining may involves statistics, Data Warehouse, Machine Learning, artificial intelligence, visualisation, Parallel and high-performance computing. R, as a language and environment for statistical computing and graphics, provides features and capabilities for data mining which includes a large collection of packages(covering Bayesian, Cluster, Econometrics, Environmetrics, Finance, Genetics, Machine Learning, Robust, Spatial, Survival, Time Series ...etc.) extending its statistical and graphical techniques, as well as interfaces connecting to databases. This article introduces advantages of R in data mining applications and techniques used in the process of data mining.

Keywords Data Ming; R; Visualisation

*电子邮件：sunbjt@gmail.com
*Email: sunbjt@gmail.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 19–20 页

Bioconductor项目简介及其在生物信息学中的应用

陈钢[1,*]　　代利坚[1,†]　　王建新[1,‡]

[1]湖南省长沙市岳麓区麓山南路中南大学主校区计算机楼302，邮编：410083

摘要

Bioconductor是一个基于R语言的开源软件项目，致力于生物数据的分析和处理。海量生物数据的分析和处理是推动当前生物信息学和系统生物学迅速发展的主要动力之一。新的实验手段和计算方法层出不穷，导致不同数据之间的整合、不同方法之间的衔接、方法和数据的集成逐渐成为一个在现实研究工作非常棘手的问题。在Bioconductor社区卓有成效的努力下，大量的数据库和相关算法被集成到Bioconductor这个统一的框架中，加上R语言在统计分析上的卓越性能，有效地解决了相关科研工作中的很多现实问题，极大地方便了相关领域的研究工作。

本文以人类蛋白质相互作用网络的GO语义相似性分析为例，通过与Java语言的对比，展示了R语言以及Bioconductor的强大威力和方便快捷。

如有需要，可以向作者发邮件索要本文涉及的所有代码。

关键词　Bioconductor；生物信息学；系统生物学；蛋白质相互作用；

Introduction to Bioconductor and its application in Bioinformatics

Gang Chen[1,*]　　Lijian Dai[1,†]　　Jianxin Wang[1,‡]

[1]Room 302, Computer Building, Central South University, Yuelu district, Changsha, Hunan

Abstract

Bioconductor is an R-based open source and open development software project to provide tools for the analysis and comprehension of genomic data. The analysis and comprehension of huge biological data is an important power for the rapid development of bioinformatics and systems biology. As the rapid development of biological experimental technologies and computational methods, the integration of various experimental data and different computational methods has became a serious challenge for related research work. The statues has been changed by Bioconductor project. Lots of databases and

*通讯作者；电子邮件：chen.gang1983@gmail.com；主页：http://netlab.csu.edu.cn
†电子邮件:dailijian@gmail.com
‡电子邮件：jxwang@mail.csu.edu.cn
*Corresponding Author; Email: chen.gang1983@gmail.com; Homepage: http://netlab.csu.edu.cn
†Email: dailijian@gmail.com
‡Email: jxwang@mail.csu.edu.cn

related algorithms have been added to this united framework. Incorporating the power of R statistic language, the problems mentioned above have been solved.

The analysis of GO semantic similarity of protein interactions network has been taken as an example to prove the power of Bioconductor in this paper. Compare to JAVA, R and Bioconductor is much more convenient for Bioinformatics.

Keywords Bioconductor; Bioinformatics; Systems Biology; Protein Interactions

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 21–21 页

R 软件与最优化

魏太云*

中南大学数学科学与计算技术学院
湖南省长沙市中南大学铁道学院 11 舍 306 长沙410075

摘要 最优化是一类重要问题，在科学的各个领域中被广泛用到。免费而又强大的开源软件 R 可以方便地解决各类最优化问题。本文就线性规划、整数规划、目标规划、非线性规划、图与网络分析在 R 中的求解方法进行了初步探究。

关键词 R 软件，最优化，线性规划，整数规划，目标规划，非线性规划，图与网络分析，TSP

R Software And Optimization

Taiyun Wei*

School of Mathematical Science and Computing Technology, Central South University
Changsha,410075

Abstract Optimization is a class of important issues , it has been widely used in the various fields of science . Free and powerful open source software R can easily solve various types of optimization problem. This paper discusses the basic usage of R and linear programming(LP), integer programming(IP), goal programming(GP), non-linear programming(NLP) and Programming involving graphs or networks(GRAPH).

Keywords R package，Optimization，LP，IP，GP，NLP，GARPH，TSP

*电子邮件：weitaiyun@gmail.com；主页：http://hi.baidu.com/cloud_wei
*Email: weitaiyun@gmail.com; Homepage: http://hi.baidu.com/cloud_wei

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 22–22 页

贝叶斯统计与 R

丁鹏*

北京大学数学科学学院概率统计系
北京大学42楼607室,邮编：100871

摘要 贝叶斯统计的思想可以上溯到贝叶斯，远远早于频率学派；但是贝叶斯统计在相当长的时间内并未被大家接受。即使接受了贝叶斯的思想，在实际中贝叶斯方法受到计算能力的极大限制，限制了贝叶斯统计的发展。但是借助现代的统计计算方法，如Gibbs抽样和Markov Chain Monte Carlo（MCMC），贝叶斯统计成了一种具有可操作性的统计方法。本文简介统计软件 R 中是如何进行贝叶斯统计推断的，大部分统计方法都用 R 中的包MCMCpack，涉及的模型有单参数、多参数的推断问题和广义线性模型。

关键词 贝叶斯统计；MCMC；MCMCpack; R

Bayesian Statistics and R

Peng Ding*

Department of Probability and Statistics
School of Mathematical Sciences, Peking University
Room 607,Building 42,Peking university, 100871

Abstract Bayesian Statistics can be traced back to Bayes, much earlier than frequentist. However, Bayesian Statistics did not win widely acceptance in a long period. Even though someone may accept Bayesian philosophy, it is still difficult to carry out a real problem due to the limitation of computation ability. But with the help of modern statistical computing methods, such as Gibbs Sampling and Markov Chain Monte Carlo(MCMC),Bayesian has become a feasible statistical method. In this paper, we will introduce how to do Bayesian Inference in the software R , especially the package MCMCpack in R . Single parameter, multi-parameter problems and generalized linear models(GLM) are mentioned in this paper.

Keywords Bayesian Statistics, MCMC, MCMCpack; R

*电子邮件：dingyunyiqiu@163.com
*Email: dingyunyiqiu@163.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 23–23 页

R 在遗传统计学中的应用

侯丽平[1,*]　　顾东风[1,†]

[1] 中国医学科学院&北京协和医学院 阜外心血管病医院 循证医学部及群体遗传研究室
北京市西城区北礼士路167号，北京 100037

摘要　人类的大多数疾病是环境和遗传因素共同作用的结果。遗传统计在疾病的遗传易感因素研究中起着至关重要的作用。R 提供多个遗传统计相关的扩展包，可以用来作为遗传研究的数据分析工具。本文对遗传统计数据分析中数据整理、多态位点基本信息的获取、Hardy-Weinberg平衡检验、连锁不平衡的计算、关联研究常用的分析方法及家系图的绘制进行简要介绍。

关键词　R 语言；遗传统计学

Applied Genetic Statistics Using R

Liping Hou[1,*]　　Dongfeng Gu[1,†]

[1] Department of Evidence Based Medicine & Division of Population Genetics and Prevention
Cardiovascular Institute and Fu Wai Hospital
Chinese Academy of Medical Sciences and Peking Union Medical College
No. 167 Beilishi Road, Beijing, 100037

Abstract　Most diseases are the results of the interaction of multiple genetic and environmental factors. Genetic statistics is very important in unraveling the contributions of environmental and genetic factors to health differences. The CRAN contains a lot of contributed extension packages about genetic statistics, which can be used for data analysis for genetic study. In this paper, we provide an introduction to data analysis in genetic statistics, including data cleaning, getting summary information of genetic variation, Hardy-Weinberg equilibrium test, computing of pairwise linkage disequilibrium, statistical methods for association study, and genograms drawing.

Keywords　R Language; Genetic statistics

[*]电子邮件：houliping@gmail.com
[†]通讯作者；电子邮件：gudongfeng@vip.sina.com
[*]Email: houliping@gmail.com
[†]Email：gudongfeng@vip.sina.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 24–24 页

R 中使用sampling和survey包进行抽样调查和分析

刘重杰*

南京大学金陵学院
江苏省南京市浦口区学府路8号南京大学浦口校区图南505，南京210089

摘要 我们在研究工作中常常会采用抽样调查的方法来收集信息，并对总体进行推断。各种概率抽样方法基于科学的理论，通过严格的设计和步骤来获显著的优点。在抽样调查领域已有很多专业统计软件，比如SPSS、SAS、Stata等均实现了这方面的功能。R 作为统计领域的专业工具，诞生之日起使用的范围就迅速扩展。本文使用 R 中已有的**sampling**和**survey**包来实现各种抽样调查方法及其对应的数据分析方法。本文仅作为一个简单介绍，只涉及了无放回简单随机抽样、有放回简单随机抽样、分层抽样、整群抽样、系统抽样、多阶段抽样中的二阶段抽样方法，这些抽样方法实现的细节可以在相应帮助文档中找到。同时需要指出的是，在这两个包中中还有实现了其他一些抽样方法；在 R 中，还有其他的包实现了更多的抽样方法。

关键词 抽样调查；简单随机抽样；分层抽样；整群抽样；系统抽样；多阶段抽样

Using sampling and survey Packages to Survey Sampling and Survey Data Analysis in R

Chonjie Liu*

Jinling College, Nanjing Univ.
Room 505, Siyuan Library, Nanjing Univ., Nanjing City, Jiangsu, 210089, P.R.C.

Abstract Researchers often use sample survey methodology to obtain information about a large population by selecting and measuring a sample from that population. Due to variability among items, researchers apply scientific probability-based designs to select the sample. In the area, there are many professional statistical softwares, such as SPSS, SAS, Stata, etc. R is a language and environment for statistical computing and graphics, which is used more and more widely. This article introduces the **sampling** and **survey** packages for survey sampling and describes how you can use their functions to analyze survey data. This is only a simple introduction to this field, so only including simple random sampling with replacement, simple random sampling without replacement, strata sampling, cluster sampling, and two-stage sampling methods, details could be found in help documents of these packages. And special sampling methods or its corresponding analysis methods are not give here, but which have implemented in other packages, and more and more sampling methods have been added into R .

Keywords survey sampling; simple random sampling; strata sampling; cluster sampling; systematic sampling; multistage sampling

*电子邮件：pkuabel@gmail.com
*Email: pkuabel@gmail.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 25–25 页

R 在水文模拟中的应用初探

王化儒[1,*]　　索安宁[2]　　梁玉莲[1]　　国庆喜[1]

[1] 东北林业大学林学院
哈尔滨市香坊区和兴路26号, 哈尔滨150040
[2] 国家海洋环境监测中心
大连沙河口区凌河街42号，大连116023

摘要　本研究利用 R 中的 TOPMODEL 包和 RHydro 包在流域水文表面分析的基础上对其水文过程进行模拟，设置是否进行凹陷点填充处理来考察其对地形指数等水文响应单元的影响，并对流域水文表面分析和水文过程模拟结果进行了展示。凹陷点填充对地形指数较高的地区产生较大的影响，改变了提取河网的连续性，导致分离出的流域也发生了较大的改变。研究表明 TOPMODEL 模型的模拟结果与实测数据能够较好的拟合，Nash-Sutcliffe 效率达到0.828，说明TOPMODEL模型能够准确地刻画流域的主要水文过程。可见，R 在水文模拟领域具有巨大的潜力，必将越来越受到该领域相关研究人员的欢迎。

关键词　水文模型；R；TOPMODEL；地形指数

Application of R on Hydrological Modeling

Huaru Wang[1,*]　　Qingxi Guo[1]　　Yulian Liang[1]　　Anning Suo[2]

[1] College of Forestry ,Northeast Forestry University
26 Hexing Road,Xiangfang District,Harbin,150040,China
[2] National Oceanic Environment Monitoring Center
42 Linghe Street,Shahekou District,Dalian,116023,China

Abstract　Using the packages of TOPMODEL and RHydro, we simulated the runoff process based on the hydrological surface analysis. The impact of the sink-filling treatment on the classification of hydrological response unit such as topographical index was examined and the results of hydrological surface analysis and the runoff simulation were also exhibited in R with several packages. The sink-filling treatment affected the area with larger topographical index greatly, and it changed the continuity of river net and leaded to the great change of the watershed separation. Our simulated results showed that the simulation of TOPMODEL fitted nearly perfectly with observation and the Nash-Sutcliffe efficiency was 0.8278993. The main hydrological process could be exactly described with TOPMODEL. Obviously, R has great potential in hydrological modeling and will gain more and more welcome from the modeler in the related research fields.

Keywords　Runoff model;R;TOPMODEL;Topological index

*通讯作者；电子邮件：huaru.wang@gmail.com
*Email: huaru.wang@gmail.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 26–26 页

分位数回归模型在 R 环境下的实现

左辰[1,*]　　潘岚锋[1,2,†]

[1] 中国人民大学统计学院
北京市海淀区中关村大街59号品园二楼104室、100872
[2] 中国人民大学统计学院
北京市海淀区中关村大街59号品园二楼108室、100872

摘要　传统回归模型的实际应用中往往会遇到与 Gauss-Markov 条件不符的情况，从而使模型的稳健性欠佳。分位回归模型是20世纪80年代的一种新的回归模型，该模型放宽了 Guass-Markov 假设条件，从而提高模型的稳定性。Roger Koenker 发布了 R 中专门用于处理分位回归模型的软件包 **quantreg**。本文借助该软件包展示了分位回归模型的基本形式，参数的渐进分布性质，参数估计的几种方法，以及对“位置漂移模型”和“位置尺度漂移模型”的检验方法。作为总结，文章最后给出了一个实例展示R中处理分位回归模型的整个过程。

关键词　分位数水平；渐近分布;参数估计;假设检验

The Application of Quantile Regression in R

Zuo Chen[1,*]　　Pan Lanfeng[1,2,†]

[1]School of Statistics ,Renming University of China
Room104, Pingyuan Buildings, Zhongguancun Street No.59, 100872
[1]School of Statistics ,Renming University of China
Room108, Pingyuan Buildings, Zhongguancun Street No.59, 100872

Abstract　Traditional linear regression model is based upon the strong Guass-Markov condition, which in practice is often violated, thus threatens the robustness of linear model. The quantile regression model, developed since the 1980s, has relatively weak conditions and thus enhances the model stability. Roger Koenker has developed the package of quantreg specially to deal with quantile regression models. In this essay we utilizes the software to display the basic model definition, asymptotic parameter distribution, parameter estimation methods, and three types of tests of location model and location-shift model. An example is presented at the end of the article to display the whole process of statistical analysis with this model.

Keywords　regression quantile; asymptotic behavior;parameter estimation;hypothesis test

*通讯作者；电子邮件：john_zuo@ruc.edu.cn
†电子邮件: idoget@hotmail.com
*Email: john_zuo@ruc.edu.cn
†Email: idoget@hotmail.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 27–27 页

基于 R 软件的统计模拟

奚潭[1]　　詹鹏[2]

[1,2] 南京财经大学统计系，统计2006级

摘要　针对R软件统计模拟，本文首先从统计模拟概念、统计模拟方法、统计模拟的一般步骤三方面对统计模拟进行了概述，其次梳理了 R 软件的模拟功能，最后使用R语言成功模拟了赶火车问题和大数定律、中心极限定理。

关键词　统计模拟；R语言；大数定律；中心极限定理

Statistical simulation based on R

Tan Xi[1]　　Peng Zhan[2]

[1,2] Nanjing University of Finance And Economics

Abstract　In order to explain statistic simulations used R, concepts, methods as well as general steps of statistic simulations were summarized firstly; then R's functions in simulation were made clear of. This paper were concluded with two cases, that is the simulation of catching the train as well as that of large numbers, law central limit theorem, all of which were successfully achieved by using R.

Keywords　statistical simulation; R; large numbers law; central limit theorem

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 28–28 页

R 语言在南京市大学生幸福感统计分析上的应用

詹鹏[1]　　奚潭[2]

[1,2]南京财经大学统计系，统计2006级

摘要　在南京市大学生幸福感调研所获得的大量数据基础上，本文基于 R 软件的强大统计功能，运用 R 语言的常用统计方法 (描述性分析、方差分析、相关性分析、回归分析等) 分析南京市大学生幸福指数状况及其影响因素，并结合国内外学者对幸福指数的研究结论，得到南京市大学生幸福指数的总体状况，以及学习、家庭、朋友等因素的重要性。而合理地处理和分析了调查数据。特别地，R 语言在方差分析上的灵活性充分体现了R语言精简、操作方便等优点。

关键词　R语言；常用统计方法；幸福指数

A Statistical analysis of well-being sence of the college students in Nanjing based on R

Peng Zhan[1]　　Tan Xi[2]

[1,2]Nanjing University of Finance And Economics

Abstract　In this paper, general statistic methods, such as descriptive analysis, analysis of variance, correlation analysis, regression analysis and etc, were used to analyze self well-beings and its influential factors of undergraduates in Nanjing. All analysis were based on R's mighty functions in statistics and scores of datum, which were procured by investigating self well-beings of undergraduates in Nanjing. Then some conclusions related to self well-beings, which were done by scholars from both homeland and abroad, were referred to gain the general situations and importance of factors like study, family and friends. In that way, investigated datum were dealt with and analyzed appropriately. Specially, the flexibility of R in analysis of variance has proved some advantages of R, like contracted, very easy to be used, and etc.

Keywords　R language; General methods of statistics; self well-being

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 29–29 页

统计动画程序包 animation 介绍及其在教学和数据分析中的应用

谢益辉[1,*]

[1] 中国人民大学统计学院
北京市海淀区中关村大街 59 号中国人民大学明德主楼 1037，北京 100872

摘要 统计学理论的飞速发展直接带来的一个问题就是学习负担的加重：我们需要以更快的速度掌握各种现代统计方法，以适应复杂问题和数据的统计分析需要。传统的统计学教科书模式在现代统计教育中缺乏灵活性的劣势随着当今计算机技术和统计软件的进步逐渐凸显，这使得我们开始借助这些技术寻求对教科书的补充方式。本文基于 R 函数包 **animation** (Xie, 2008) 从动画的角度介绍了统计学与动画的密切联系，并给出了具体的教学演示以及数据分析示例。

关键词 统计学；动画；教学；统计计算；数据分析

animation: An R Package for Statistical Animations with Applications in Teaching and Data Analysis

Yihui Xie[1,*] Xiaoyue Cheng[1,†]

[1] School of Statistics, Renmin University of China
Room 1037, Mingde Main Building, Renmin University of China, Beijing, 100872, China

Abstract The rapid developments of statistical theories has led to the burdens of learning – we have to master a lot of modern methods to be adapted to the need for complicated statistical problems and data analysis. Traditional textbooks have obvious disadvantages in statistical teaching as they are rather inflexible, especially in this era of advanced computer technologies as well as powerful statistical software packages such as R. Based on the R package **animation** (Xie, 2008), we will introduce the applications of animations in both teaching and data analysis.

Keywords animation; teaching; data analysis; statistical computation; R language

*通讯作者；电子邮件：xie@yihui.name；主页：http://www.yihui.name
*Email: xie@yihui.name; Homepage: http://www.yihui.name
†Email: chengxiaoyue@gmail.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 30–30 页

随机微分方程入门 — 基于 R 语言的模拟与推断

陈堰平[1,*]

[1] 中国人民大学统计学院
北京市海淀区中关村大街 59 号中国人民大学，北京 100872

摘要 本文用通俗的方法介绍了与随机微分方程（以下简称 SDE）有关的基本概念与数值实现，包括布朗运动、Ito 积分的基本思想与数值模拟，SDE 的基本概念、数值解以及参数估计的数值方法。本文并不追求数学上的严格与完美，只是一次向不具备高等概率论和随机过程等方面知识的读者介绍 SDE 的尝试。文中有来自金融和生态学的实例，例子程序全部是由 R 语言编写，有的用到了 **sde** 包。

关键词 随机微分方程；模拟；推断；R 语言

*电子邮件：ypchen_cos@yahoo.com.cn

高级篇

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 33–33 页

基于 R 后台的网页应用
或基于网页可定制界面的 R 应用

李晓煦[1,2,*]　　李崇亮[2,3]　　吴锋[2,4,†]

[1]香港中文大学教育心理系
香港，沙田，香港中文大学，何添楼104
[2]北京大学深圳研究生院人文社科学院
广东，深圳，西丽大学城，518055
[3]北京大学心理系
北京，海淀区，100871
[4]北京大学社会学系
北京，海淀区，100871

摘要　介绍、示范若干层面基于 R 后台的网页应用或基于网页可定制界面的 R 应用，讨论社会科学领域的学术网页如何实现经典范例与软件互动界面相得益彰的融合。

关键词　Rweb；R extension for MediaWiki；RwebFriend for Wordpress；行为科学与心理学统计

Web Powered by R or R Powered by Web

Xiaoxu LI[1,2,*]　　Chongliang LI[2,3]　　Feng WU[2,4,†]

[1]Dept. of Educational Psychology, Chinese Univ. of H. K.
104, Ho Tim Bldg, Chinese Univ. of H.K., Hong Kong
[2]School of Humanities and Social Sciences, Shenzhen Graduate School of Peking Univ.,
Xili Univ. Town, Shenzhen, Guangdong, 518055
[3]Dept. of Psychology, Peking Univ.
Beijing, 100871
[4]Dept. of Sociology, Peking Univ.
Beijing, 100871

Abstract　Introduce web applications powered by R or R interface powered by web applications. Discuss how to integrate the social sciences web contents and relevant interactive web-based software interfaces.

Keywords　Rweb; R extension for MediaWiki; RwebFriend for Wordpress; Behaviorial and Psychological Statistics

[*]电子邮件：lixiaoxu@gmail.com；主页：http://lixiaoxu.lxxm.com
[†]主页：http://qixianglu.cn
[*]Email: lixiaoxu@gmail.com; Homepage: http://lixiaoxu.lxxm.com
[†]Homepage: http://qixianglu.cn

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 34–34 页

利用 R (D) COM Server 实现 R 与MS Office的整合

李舰[1,*]

摘要 R (D) COM Server 是Thomas Baier 和Erich Neuwirth开发的基于DCOM通信的组件工具，它提供了COM接口用来实现客户端与R的连接。该文重点介绍了这个工具与MS Office的整合，并给出了一个实际开发中的例子。

关键词 COM；R接口；MS Office

Integrating MS Office with R Using R (D) COM Server

Li Jian[1,*]

Abstract R (D) COM Server is a component based on DCOM, which is developed by Thomas Baier and Erich Neuwirth. It used to connect a client application with R. This paper tells something for the integration of R and MS Office, and gives an example.

Keywords COM ; R interface ; MS Office

*电子邮件：lijian.pku@gmail.com；主页：http://www.lijian-homepage.com
*Email: lijian.pku@gmail.com; Homepage: http://www.lijian-homepage.com

第一届中国 R 语言会议论文集，北京，2008 年 12 月 13 日 – 14 日
主办：中国人民大学统计学院、应用统计科学研究中心；第 35–35 页

无处不在的数据及 R 的探索方法

谢益辉[1,*]

[1] 中国人民大学统计学院
北京市海淀区中关村大街 59 号中国人民大学明德主楼 1037，北京 100872

摘要 统计是一门与实际生活紧密相连的学科，其数据来源也是多种多样的，然而实际问题中的数据原始形式一般不会是统计所需的规整的表格形式，因此我们有必要了解更多的收集和整理数据的工具、方法。本文主要以网页数据为例，对数据的获取和表达作出初步的探讨。

关键词 数据；网页；连接；正则表达式

Explore Irregular Data with R

Yihui Xie[1,*]

[1] School of Statistics, Renmin University of China
Room 1037, Mingde Main Building, Renmin University of China, Beijing, 100872, China

Abstract Statistics is a discipline closely related to our real life, and there are also a large variety of data sources which are usually unlikely to meet the common needs for statistical analysis, therefore we need methods and tools to gather such data and tidy them up. This paper will mainly discuss on how to fetch and deal with irregular data them using R in a few examples.

Keywords Data; Web; Connection; Regular Expression

[*] 电子邮件：xie@yihui.name；主页：http://www.yihui.name
[*] Email: xie@yihui.name; Homepage: http://www.yihui.name