第三届中国 R 语言会议

自由的统计语言

主办：

中国人民大学应用统计科学研究中心

中国人民大学统计学院

协办：

统计之都网站

2010 年 6 月 14 日

目　　录

目录 · · · · · i
R语言简介 · · · · · 1
中国人民大学应用统计科学研究中心简介 · · · · · 2
中国人民大学统计学院简介 · · · · · 3
统计之都简介 · · · · · 4
第三届中国 R 语言会议日程 · · · · · 5
中国人民大学地图 · · · · · 6

演讲摘要 7
傻瓜软件是怎样炼成的：用 gWidgets 包创建图形用户界面 · · · · · 9
谢益辉
R 在时间序列中的应用–TAR · · · · · 10
胡一睿
R 与文本挖掘 · · · · · 11
李舰
R与Python在昼夜节律分析中的应用 · · · · · 12
杨仁东
非参数回归的 R 语言实现 · · · · · 13
陈堰平
加速R，矢量化运算与并行计算 · · · · · 14
稳国柱
R 与因子分析的新发展 · · · · · 15
祝迎春

R/BioConductor 在斑马鱼心脏再生领域的应用 **16**

甄一松

Nlme 包及数学建模中的 R **17**

江麒

R 语言简介*

R 是一个有着统计分析功能及强大作图功能的软件系统，由新西兰奥克兰大学统计系的 Ross Ihaka 和 Robert Gentleman 共同创立。 R 语言可以看作是由 AT&T 贝尔实验室所创的 S 语言发展出的一种方言。因此， R 即是一种软件也可以说是一种语言。

R 是在 GNU 协议 General Public Licence 下免费发行的，它的开发及维护现在则由 R 开发核心小组 *R Development Core Team*具体负责，这个团队的成员大部分来自大学机构（统计及相关院系），包括牛津大学、华盛顿大学、威斯康星大学、爱荷华大学、奥克兰大学等。除了这些作者之外，R 还拥有一大批贡献者（来自哈佛大学、加州大学洛杉矶分校、麻省理工大学等），他们为R 编写代码、修正程序缺陷和撰写文档。

R 内含了许多实用的统计分析及作图函数。作图函数能将产生的图片展示在一个独立的窗口中，并能将之保存为各种形式的文件（jpg，png，bmp，ps，pdf，emf，pictex，xfig；具体形式取决于操作系统）。统计分析的结果也能被直接显示出来，一些中间结果（如 P–值，回归系数，残差等）既可保存到专门的文件中，也可以直接用作进一步的分析。

R 的功能很大程度上是通过程序包（Package）来实现的，迄今为止，R 中的程序包已经是数以千计，各种统计前沿理论方法的相应计算机程序都会在短时间内以软件包的形式得以实现，这种速度是其它统计软件无法比拟的。

由于 R 强大的功能和它在统计理论及应用上的优势，我们希望 R 能够在国内的各领域有更多的发展，而本次会议正是秉承这一目标而召开的。

*内容主要选自 *R for beginners, Chinese Edition 2.0*及谢益辉《现代统计图形》

中国人民大学应用统计科学研究中心简介

中国人民大学应用统计科学研究中心前身是成立于1988年的统计科学研究所。十几年来，中心积极培育中青年学术骨干，逐渐发展并形成了经济与社会统计、统计调查与数据分析和风险管理与精算三个各具特色的研究方向。几年来，中心建设的重点研究平台是：1）统计理论与建模方法和应用研究；2）满意度统计理论、方法和应用研究；3）国际竞争力理论方法及其应用研究；4）数据挖掘技术中的统计理论、方法与应用研究；5）改进我国政府统计数据质量及其抽样调查制度的理论方法研究；6）统计在社会科学中的应用研究；7）风险管理与保险精算应用研究；8）六西格玛管理应用研究；9）环境经济核算理论方法和应用研究。此外，中心本着创建和发展面向实际应用的研究中心的宗旨，创建了：竞争力与评价研究中心；数据挖掘中心；六西格玛质量管理研究中心；保险精算中心；统计资讯研究中心等子机构，在突出应用主题的研究中心下，本着联系实际和服务实际的思想，创建了面向实际应用的网站，建立新型的学术交流、知识普及和与用户零距离连接的模式。随着我国经济体制的进一步改革，中心积极适应市场经济的需要，面向全国开放，加强国际学术交流与合作，推动重大应用统计项目的研究。

中心现有专兼职研究人员 29 人，学术委员会委员 19 人，其中既有统计科学领域国内外著名的学术带头人，如中科院院士严加安教授、陈木法教授、彭实戈教授；又有一批全国知名学者和业务骨干，如袁卫教授、吴喜之教授、耿直教授、赵彦云教授和原国家统计局局长谢伏瞻研究员等。中心研究队伍强大的教育背景、研究成果和学术声誉将使本中心成为全国一流并具有国际声誉和影响的开放式应用统计研究机构。

中国人民大学统计学院简介

中国人民大学统计学科始建于 1950 年，两年后成立统计学系，是新中国经济学科中最早设立的统计学系， 2003 年 7 月，成立中国人民大学统计学院。多年来，本学科一直强调统计理论和统计应用的结合，不断拓宽统计教学和研究领域，成为统计学全国重点学科。教育部人文社会科学重点研究基地“应用统计科学研究中心”也设在统计学院。学院拥有统计学和风险管理与精算学两个博士点，统计学、概率论与数理统计、风险管理与精算学、流行病与卫生统计学四个硕士点，应用经济学下设统计学博士后流动站。

统计学院现有教师 33 人，其中教授 14 人，副教授 11 人，博士生导师 13 人。国内兼职教授 11 名，海外客座教授 10 人。 50 多年来，共培养不同层次人才 5000 多人。 2008 年 9 月，在校学生总人数为 523 人，其中本科生 305 人，硕士生 142 人，博士生 76 人，大多数毕业生在金融、保险、证券、基金、信息等领域从事数据采集和分析工作。

统计之都简介

“统计之都”（Capital of Statistics，简称 COS）网站成立于 2006 年 5 月，其主旨为传播统计学知识并将其应用于实际领域。纵观现今国内统计学理论和应用的发展，一方面我们不难发现统计学在应用领域的巨大潜力——现代管理、咨询、商业、经济、金融、医药、生物等等，无不需要数据的力量，而另一方面我们也不得不承认，国内统计学的应用很大程度上受理论的制约——无论是应用界的人们对统计学基础理论知识的欠缺，还是学术界所研究的理论对应用领域问题的轻视。

“统计之都”网站便是基于这样的认识而创建的。我们希望，统计理论研究者能充分关注应用问题，而统计应用者也能正确把握统计学基本知识，将统计学这门应用学科真正的潜力开发出来。

“统计之都”为非赢利性质网站，但大力欢迎所有商界和研究领域的朋友与我们在实际应用问题上合作。我们的口号是：

中国统计学门户网站，免费统计学服务平台

我们怀着“十年磨一剑”的决心，要将“统计之都”创建成中国的统计学门户网站；我们抱着“己欲立而立人、己欲达而达人”的信条，要将“统计之都”以免费统计学服务平台的形式坚持办下去。我们希望“统计之都”在专业知识体系上有真正的王者风范，在面对用户需求时却又以谦恭的态度为大家服务。

第三届中国 R 语言会议日程

	培训	时间
6 月 14 日		
上午（地点：明德法学楼 0101）		
	欢迎辞	9:00~9:15
	刘思喆：R 基础培训 1	9:15~10:20
休息		
	刘思喆：R 基础培训 2	10:30~11:30
午餐（我们将引导您至就餐地点）		
下午		
	谢益辉：R 作图培训1	14:00~15:30
休息		
	谢益辉：R 作图培训2	15:40~17:00
培训结束		

主持人	演讲	时间
6 月 15 日		
上午（地点：明德法学楼 0101）		
范建宁	欢迎辞	9:00~9:15
	谢益辉：傻瓜软件是怎样炼成的：用 gWidgets 包创建图形用户界面	9:15~9:40
	胡一睿：R 在时间序列中的应用-TAR	9:40~10:10
全体合影（明德法学楼前台阶上集合）、休息		
熊熹	李舰：R 与文本挖掘	10:30~11:00
	杨仁东：R 与 Python 在昼夜节律分析中的应用	11:00~11:30
午餐（北区食堂二层）		
下午（地点：明德法学楼 0101）		
谢漫锜	陈堰平：非参数回归的 R 语言实现	14:00~14:30
	稳国柱：加速 R，矢量化运算与并行运算	14:30~15:00
	祝迎春：R 与因子分析的新发展	15:00~15:40
休息		
关菁菁	甄一松：R/BioConductor 在斑马鱼心脏再生领域的应用	16:10~16:40
	江麒：Nlme 包及数学建模中的 R	16:40~17:20
综合讨论		
会议结束		

北
中关村大街
双榆树
东校门
西校门
中国人民大学附属中学
人民大学北路
苏州桥方向
东风5楼
东风4楼
东风6楼
东风7楼
文化大厦
网球场
体育馆
求是园
红1楼
红2楼
红3楼
东区食堂
青年公寓
公共教学1楼
东风2楼
留学生食堂
东风1楼
公共教学2楼
留学生2楼
留学生3楼
人文楼
公共教学3楼
游泳馆
小足球场
运动场
世纪馆
培训楼
育贤楼
品园1楼
品园2楼
品园3楼
品园4楼
品园5楼
西区食堂
聚英楼
明德楼
学生事务楼
北区食堂
知行1楼
知行2楼
知行3楼
知行4楼
知行5楼
静园22楼
校医院
一勺池
百家廊
孔子像
图书馆
林园
静园
品园

演讲摘要

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 9 页

傻瓜软件是怎样炼成的：用 gWidgets 包创建图形用户界面

谢益辉[1,*]

[1] 爱荷华州立大学统计系
3211 Snedecor Hall, Ames, IA 50010

摘要 R 默认的命令行用户界面 (CLI) 对一些用户来说在一定程度上加大了学习和使用的难度，相比之下，图形用户界面 (GUI) 则显得更为有好。在 R 中已经存在若干种创建 GUI 的可能性，例如 R 自带的 tcltk 包，著名的 Rcmdr 包就是基于 tcltk 包的一个应用。本文要介绍的是 gWidgets 包，它相比起其它创建 GUI 的 R 包来说，在使用上极大减轻了用户设计界面的负担，而且包内各个函数的使用也具有很大的相似性，进一步，它也将 R 的数据结构（如数据框）与图形界面元素（如数据表）很好地融合起来，为熟悉 R 基础的用户提供了形象的开发环境。本文以一个基于 gWidgets 的 R 包 formatR 的开发为例，介绍 gWidgets 包在创建窗口、按钮等图形元素及其相应事件中的应用。最后，我们也简要介绍一下编写 R 包的方法。

关键词 图形用户界面；gWidgets；formatR

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

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 10 页

R 在时间序列中的应用–TAR

胡一睿[1,*]

[1] 北京师范大学
北京师范大学 707 信箱，100875

摘要 通过 R 演示数据预处理、点值图，选取合适的门限值建立回归模型，最终通过 R 得到预测趋势及拟合的图形。

关键词 TAR；点值图；回归；预测

*电子邮件：Doudou@mail.bnu.edu.cn

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 11 页

R 与文本挖掘

李舰[1,*]

[1] 上海源略数据服务有限公司

摘要 概括地介绍了文本挖掘的技术和应用以及一个利用 R 做运算引擎的文本挖掘系统的实现，并就文本挖掘在系统层面的发展进行了探讨。

关键词 文本挖掘；中文分词；大规模矩阵运算；云计算

*电子邮件：lijian.pku@gmail.com；主页：http://www.leejian.name

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 12 页

R 与 Python 在昼夜节律分析中的应用

杨仁东[1,*]

[1] 中国农业大学生物学院
中国农业大学西区生命科学研究中心 2063 房间，北京 100193

摘要 R 是一款开源的跨平台的数值统计和数值图形化展现工具。R 拥有自己的脚本语言和大量的统计、图形库。R 能够很方便的完成大多数统计计算的任务。Python 是一种通用的脚本语言，可以方便完成像连接数据库、文本处理、文件操作等任务，并可被作为胶水语言和其他语言结合起来使用。本文介绍了一种新的算法 (ARSER) 采用 Python 调用 R 程序来实现，并将其应用于生物时间序列的周期性的识别。通过对模拟与真实的实验数据的计算，证明了 ARSER 算法有效的解决了短时间序列的周期性识别问题。

关键词 R；Python；时间序列分析；周期识别；昼夜节律

*电子邮件：cauyrd@gmail.com

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 13 页

非参数回归的 R 语言实现

陈堰平[1,*]

[1] 中国人民大学统计学院

摘要 非参数回归因其形式灵活、限制条件少以及拟合效果好的特点，近年来受到越来越多的关注。非参数回归的估计方法有 Nadaraya-Watson 核估计、Müller 估计、局部多项式估计以及样条估计等方法。这些方法都涉及到大量的计算，并且应用到具体的非参数回归模型中时形式也需要做相应调整。R 语言的强大计算功能以及编程的灵活性，使得非参数回归模型的估计方法很容易用 R 语言实现。本文主要介绍了与非参数回归相关的几个 R 语言包：sm、np、splines、KernSmooth 等，并介绍了应用局部多项式方法、样条方法估计非参数分位数回归模型时的 R 语言实现过程。

关键词 R 语言；局部多项式估计；B 样条估计；分位回归

*电子邮件：yanping.chen@cos.name

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 14 页

加速 R，矢量化运算与并行计算

稳国柱[1,*]

[1] 豆瓣 www.douban.com

摘要 R 是一门面向科学计算特别是统计计算的语言，跟 MATLAB 类似，其循环结构的实现效率很差。但因为它们都包含了矢量化运算的支持，可以弥补循环效率不足的问题。事实上，大部分循环结构的代码都可以转换成矢量化运算的形式。利用矢量化的思想设计算法，不但可以充分利用硬件及现代科学计算工具包所提供的高性能计算支持，还可以很方便地实现并行/分布计算，使得算法随着数据规模的增加而具有可扩展性。本主题还会演示几个实例来阐述这些原理。

关键词 高性能计算；矢量化运算；并行/分布计算

*电子邮件：guozhuwen@gmail.com；主页：http://www.wentrue.net/blog/

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 15 页

R 与因子分析的新发展

祝迎春[1,*]

[1] 定谊科技

摘要 回顾传统因子分析技术以及指出常见错误用法，展示作者收集整理最新的（已经由 R 实现）与因子分析相关的程序包和个人编写的代码。系统讲述了完整的因子分析使用流程、方法与应用。

关键词 因子分析；多水平；贝叶斯；非线性；平行分析；独立因子分析

*电子邮件：ereree@126.com

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 16 页

R/BioConductor 在斑马鱼心脏再生领域的应用

甄一松[1,*]

[1] 中国医学科学院，阜外心血管病医院
教育部基因与临床重点实验室，北京 100037

摘要 通过对既往文献斑马鱼心脏再生芯片的数据分析，我们希望找到与再生过程非常密切的基因，用于后续的生物功能学研究。我们对数据采用常规的质控分析步骤，采用 Limma 进行差异表达基因的分析，对数据的聚类采用 Cluster3.0。我们得到的结果与发表文献相似，说明我们的分析方法是准确的，为我们今后的工作奠定了基础。

关键词 斑马鱼；生物芯片；心脏再生

*电子邮件：zhenyisong@cardiosignal.org

第三届中国 R 语言会议手册，北京， 2010 年 6 月 14 日 – 15 日
主办：中国人民大学应用统计科学研究中心、中国人民大学统计学院；第 17 页

Nlme 包及数学建模中的 R

江麒[1,*]

[1] 中国人民大学统计学院
北京市海淀区中关村大街 59 号中国人民大学东风七楼 203，100872

摘要 本文演讲主要包括两部分。前一部分介绍了作者在参加 2010 年北美数学建模比赛中解决 Problem B 的建模成果。作者针对连环杀人事件(serial murder)的数据，建立了贝叶斯概率模型，并且利用 Gibbs 抽样和 Metropolis Hasting 算法，对于模型参数进行后验推断，从而很好地解释了凶手的犯罪行为。第二部分，介绍 R 中 nlme 包的使用。近年来，在诸多领域（包括纵向数据分析、教育评价、动物生理学分析等），混合效应模型被广泛地用在分层数据(Hierarchical data)的处理中。本文主要介绍了基于 R 中 nlme 包的分层数据分析，比如如何判定数据适用于混合效应模型，如何确定随机效应项，以及如何加入协变量来解释随机效应。

关键词 MCM；Gibbs Sampling；M-H algorithm；Mixed Effect Model；NLME

*电子邮件：jiangqi8908@gmail.com