# 可重复的科学研究与 Sweave 的应用

谢益辉

中国人民大学统计学院

2009年3月3日

# 目录

# 你曾在 Word 中遇到过多少麻烦？

- 你总出于自己的审美观考虑格式：页边距、加粗、行距、列表……
- 老板（编辑、客户……）要求一级标题用二号字、正文小四、……
- 数学公式太难看（不知道如何让它们在行内垂直居中对齐）
- 不知如何生成目录、双页页眉，或目录和页眉太难看
- 阅读起来觉得满满一页密密麻麻都是字，太累
- 图表不知如何环绕文字以使得文字不要留出大块空白（字多了加图，图多了加字）
- 不知如何引用图表编号以及参考文献编号，或者页编号、章节编号
- 不知如何断词（英文）使得页面边界不齐
- 一边写一边考虑格式，思路不能集中在写作上面

# 你曾在 Word 中遇到过多少麻烦？（续）

- 终于有一天辛辛苦苦排好了版，再过了一个月，排版要求又变了！
- 还有更多令人抓狂的事情
  - 作为期刊编辑，收来的文章格式千奇百怪
  - 作为出版社，考虑把 A4 纸的书改成 B5 纸（然后发现图表乱了套、图变得不清晰）
  - 作为 Office 用户，2003 版的 Word 打不开 2007 的文档（现已有转换补丁）
  - ……
- 我们真的需要“所见即所得”吗？我们能忍受“所见即所得”的诱惑吗？

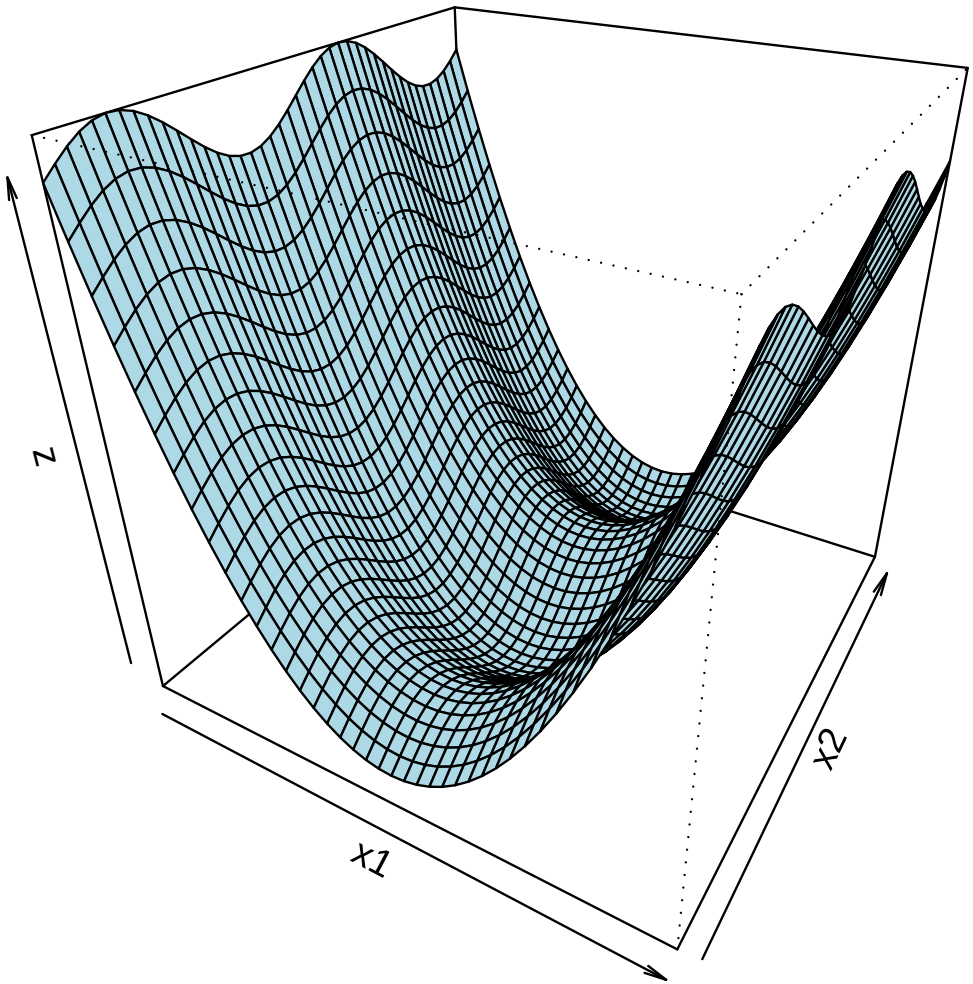
# 什么是“美”

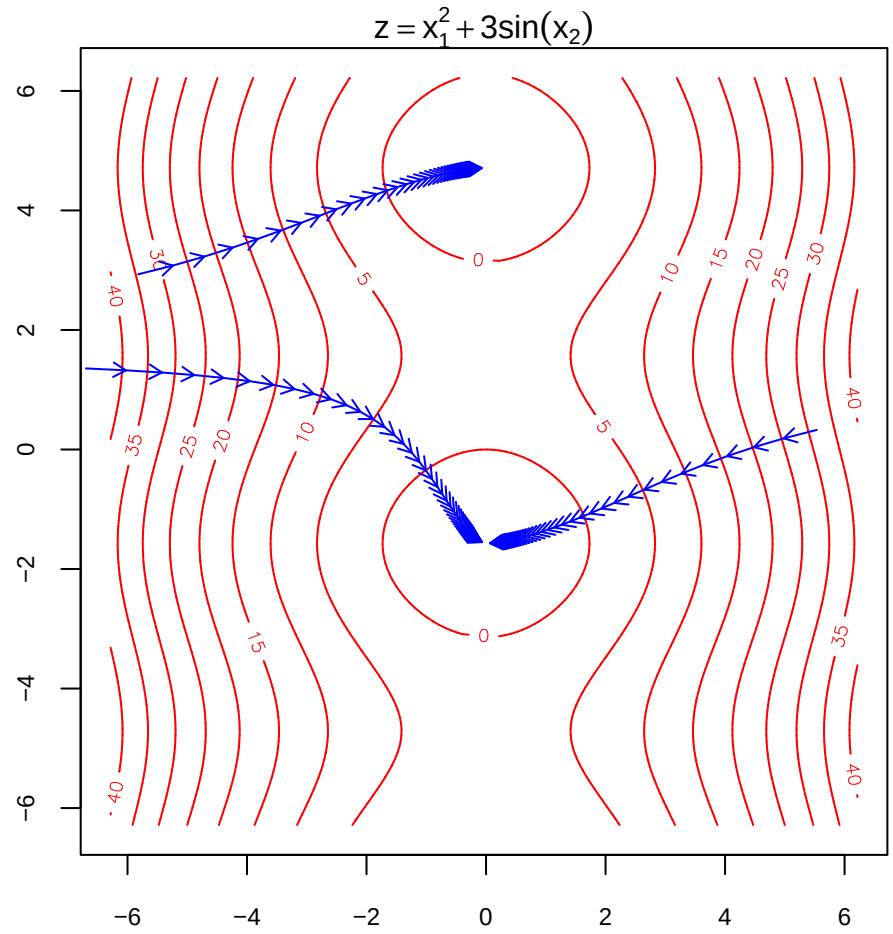
与公式编辑器相比:

$$
\begin{aligned}
SS_{\text{total}} &= SS_{\text{error}} + SS_{\text{treatments}} \\
\sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X})^2 &= \sum_{i=1}^{r}\sum_{j=1}^{n_i}(X_{ij}-\bar{X}_i)^2 + \sum_{i=1}^{r}(\bar{X}_i-\bar{X})^2 \\
F &= \frac{SS_{\text{treatments}}/(r-1)}{SS_{\text{error}}/(n-r)} \sim F(r-1, n-r)
\end{aligned}
$$

# 什么是“美”（续）

与 Word 中的位图相比:



# 什么是“美”（续）

让我们来看看:

- 几篇论文
- 一本书
- 几本论文集

# 一些花絮

- 神奇人物Knuth（为了书籍出版而写了TeX系统、一个bug赏2.56美元）
- 为什么页面那么窄（还是为你的眼睛多操操心吧）
- 一个艰巨的任务：请你去国外统计系教授的网站上找出一篇Word文档

# 安装LaTeX

## 编译程序

- 中文LaTeX: http://www.ctex.org
- 核心程序: MiKTeX
- 还有很多版本

## 编辑器

- 任何文本编辑器均可
- CTeX自带WinEdt作为编辑器
- 记事本、Tinn-R等

# LaTeX文档结构

```
\documentclass{article}
\begin{document}
Small is beautiful.
\end{document}
```

# LaTeX文档结构（续）

```
% 声明文档类型：论文（书籍、报告、幻灯片）
\documentclass[a4paper,11pt]{article}
% 标题、作者
\author{H.~Partl}
\title{Minimalism}
% 开始文档
\begin{document}
% 生成标题
\maketitle
% 插入目录
\tableofcontents
% 开始一节
\section{Start}
Well, and here begins my lovely article.
% 又一节
\section{End}
\ldots{} and here it ends.
% 结束文档
\end{document}
```

# LaTeX学习材料

- 勤看LShort-cn（CTeX自带）
- 宏包的帮助
- FAQ
- Google

# 贝尔实验室的S语言

- Fortran程序对于统计分析来说过于低层
- 统计数据分析过程复杂，模式化的程序难以适应分析需要
- 统计图形是（探索型）数据分析的重要输出
- S语言的主要作者John Chambers获得了ACM的软件系统奖

# 奥克兰大学的 R 语言

- 作者Ross Ihaka和Robert Gentlemen（首字母都是R）
- 基于Scheme语言，恰逢S语言的发布
- 改进 S 语言
- 作者都对统计计算感兴趣

# R 语言现状

- 开源、免费、灵活、统计方法模型繁多
- 19 位核心成员
- 超过 1500 个程序附加包
- 论文引用次数呈指数增长
- 邮件列表中的邮件不计其数
- 跨地域、跨行业的协作

# 主页导航

http://www.r-project.org

- 关于
- 下载镜像（中国香港有一个镜像网站）
- R 组织
- 文档（官方文档、用户贡献文档、卡片）
- 其它

# 其它网络资源

- COS 论坛R版块: http://cos.name/bbs/thread.php?fid=15
- 入门示例: http://www.statmethods.net/
- 小技巧和小提示: http://onertipaday.blogspot.com/

# 相关书籍

- Peter Dalgaard, *Introductory Statistics with R*（初等统计）
- Brian S. Everitt and Torsten Hothorn, *A Handbook of Statistical Analyses Using R*（涵盖较多统计模型，理论部分少，实例多）
- Venables and Ripley, *Modern Applied Statistics with S* (MASS)（经典，注重理论和统计计算细节）
- Paul Murrell, *R Graphics*（详细解释R图形）

# 我怎样学习R

- 一天看两次，一次看半天
- 学习要点：数据结构+积累经验

# 程序编辑器

任意文本编辑器都可以

- Tinn-R
- R 自身的编辑器
- 记事本
- Emacs/ESS
- WinEdt/R-WinEdt
- Kate

# 什么是科学

## John W. Tukey 的观点

1. intellectual content（包含知识成分）
2. organization into an understandable form（有可理解的形式）
3. reliance upon the test of experience as the ultimate standard of validity（经得起实践的检验）

# 统计的道德比技术重要

## Vardeman & Morris 的观点[a]

[a]Vardeman, S.B. and Morris, M.D. (2003) Statistics and ethics: Some advice for young statisticians. *The American Statistician* 57, 21–26.

*[A]t its core statistics is not about cleverness and technique, but rather about honesty. Its real contribution to society is primarily moral, not technical. It is about doing the right thing when interpreting empirical information. Statisticians are not the world's best computer scientists, mathematicians, or scientific subject matter specialists. We are (potentially, at least) the best at the principled collection, summarization, and analysis of data. Our subject provides a framework for dealing transparently and consistently with empirical information from all fields; means of seeing and portraying what is true; ways of avoiding being fooled by both the ill intent (or ignorance) of others and our own incorrect predispositions[...]*

# 可重复的数据分析

## Charlie Geyer的观点[a]

[a] http://www.stat.umn.edu/~charlie/Sweave/

- Research should be reproducible. Anything in a scientific paper should be reproducible by the reader.
- Whatever may have been the case in low tech days, this ideal has long gone. Much scientific research in recent years is too complicated and the published details to scanty for anyone to reproduce it.
- The lack of detail is not entirely the author's fault. Journals have severe "page pressure" and no room for full explanations.
- For many years, the only hope of reproducibility is old-fashioned person-to-person contact. Write the authors, ask for data, code, whatever. Some authors help, some don't. If the authors are not cooperative, tough.

# 可重复的数据分析（续）

## Charlie Geyer的观点

- Even cooperative authors may be unable to help. If too much time has gone by and their archiving was not systematic enough and if their software was unportable, there may be no way to recreate the analysis.
- Fortunately, the internet comes to the rescue. No "page pressure" there!
- Nowadays, many scientific papers also point to "supplementary materials" on the internet, either at the journal's or the author's web site. It doesn't matter so long as the material is permanently available. Data, computer programs, whatever should be there.

# 统计计算与分析的可重复性

## 模拟与优化算法

- 模拟涉及到随机数，如何保证别人的结果和你一样？
- 随机数种子！
- 计算机的随机数是怎样产生的？[a]
- 优化算法一定是稳定的吗？（某教授与结构方程模型；John Fox的sem包）

[a] http://www.yihui.name/en/post/41.htm

## 数据分析

- 让“可重复性”变得可操作的关键在于数据源（国内的普遍情况）
- 更重要的是分析过程！

# 一个写报告的经历

- 主要任务：分省市汇总、统计、评价
- 数据源：Foxpro数据库
- 工作
  - 创建数据库链接
  - 读取数据
  - 在地区变量（省市）中循环输出CSV文件（数据表）和PNG文件（图形）
  - 写入HTML文件
  - 打开HTML文件，全选、复制、粘贴到Word中

# LaTeX与R的结合

- Friedrich Leisch: Sweave User Manual, 2008 URL: `http://www.stat.uni-muenchen.de/~leisch/Sweave`
- 为什么能结合？（LaTeX文档是纯文本文件，易操作）
- 细节实现：搜索文档中的R代码块、执行之、替换或添加到原位置生成LaTeX文档、运行LaTeX生成PDF、PS等文档
- 示例

# 谢谢大家!

- 主页: http://www.yihui.name
- 邮件: sprintf("%s@%s", "xie", "yihui.name")