应用R软件和Epicalc程序包分析流行病学数据

Analysis of Epidemiological Data Using R and Epicalc

主 编 Virasakdi Chongsuvivatwong
主 译 蔡 乐
中文总审校 姜润生 刘 苹 吴锡南

人民卫生出版社

应用R软件和Epicalc程序包分析流行病学数据

Analysis of Epidemiological Data Using R and Epicalc

主　编　**Virasakdi Chongsuvivatwong**

主　译　蔡　乐

中文总审校

姜润生　　刘　苹　　吴锡南

译　者（以姓氏笔画为序）

许传志　　孙艳春　　李伟明

李晓梅　　何利平　　张晓磐

罗家洪　　赵科颖　　蔡　乐

（主译　中文总审校　译者单位：昆明医学院）

人民卫生出版社

图书在版编目（CIP）数据

应用R软件和Epicalc程序包分析流行病学数据/蔡乐主译
—北京：人民卫生出版社，2008.7
ISBN 978-7-117-10296-4

Ⅰ.应… Ⅱ.蔡… Ⅲ.流行病学—应用软件 Ⅳ.R18-39

中国版本图书馆CIP数据核字（2008）第082315号

应用R软件和Epicalc程序包分析流行病学数据

主　　译：蔡　乐
出版发行：人民卫生出版社（中继线010-67616688）
地　　址：北京市丰台区方庄芳群园3区3号楼
邮　　编：100078
网　　址：http://www.pmph.com
E-mail：pmph@pmph.com
购书热线：010-67605754　010-65264830
印　　刷：北京汇林印务有限公司
经　　销：新华书店
开　　本：787×1092　1/16　　印张：16
字　　数：379千字
版　　次：2008年7月第1版　2008年7月第1版第1次印刷
标准书号：ISBN 978-7-117-10296-4/R·10297
定　　价：31.00元

前　言

数据分析是流行病学研究中非常重要的一项工作。随着计算工具能力的稳步提高，流行病学研究也因计算机的不断推广而得到增强。现在有很多商业化的统计软件包已在世界范围内被流行病学家广泛应用。对发达国家而言，购买软件的费用不是一个大问题，但对发展中国家来说，购买软件的费用则显得过高。因此，很多研究者不得不依赖于使用盗版软件。

可用的免费软件包为数不多。例如，EpiInfo 是一个免费软件，主要用于数据录入和进行简单的数据分析，但它不适用于纵向研究资料的数据处理；其回归分析模块不能处理重复测量数据和多水平模型资料，而且作图能力有限。高级数据分析家发现其局限性太多。

R 软件是一个相对较新和免费的统计软件，正日益受到人们的关注。在世界各地著名统计专家的支持和开发下，它几乎能满足所有流行病学数据分析的需求，但与类似的统计软件相比（如 Stata），该软件相对难学和难用。因此，本书的目的是试图构建一个桥梁，使得来自发展中国家的研究者容易学习 R 软件，同时，也推进该软件在发展中国家的应用。

我在流行病学领域已经有 20 多年的研究经验，尤其热衷于数据分析的教学工作。在软件资源共享精神的激励下，我花费了大量的心血来开发和使用 R 软件。我用了 3 年的时间开发和增加 R 软件的新模块，使得新研究者乐于使用 R 软件，20 章以上的讲稿内容和练习都配备了相应的数据集，使得学习者能够自学该软件。

在世界卫生组织和泰国研究基金的资助下，我已经在很多发展中国家包括泰国、缅甸、朝鲜和马尔代夫组织召开了 R 软件的学术研讨会，并受到了这些国家的欢迎。由于有了这些经验，我希望该软件能够得到流行病学家的支持，尤其是鼓励那些没有能力购买昂贵商业统计软件包的人们使用。

Epicalc 是一个包括 40 多个功能的模块。在正确安装 R 软件和 Epicalc 后，只要运行 R 软件，Epicalc 随时可以使用。R 软件由一系列经世界著名统计学家研制的模块组成，能够覆盖较宽范围内的数据分析。

R 软件运行环境能够同时处理很多数据集。用户既可以通过复制其路径，也可以引用数据集的名字作为前缀来使用数据集中的变量。R 软件在处理数据集方面的优势也是

其在数据处理方面的缺陷。当创建一个变量或修改一个已经存在的变量时，如果没有数据集的名字作为前缀，这个新变量将独立于它的父数据集。如果加入前缀，则原始数据虽然改变了，却没有复制进搜索路径中。细心的用户需要移去搜索路径中的数据，重新复制新数据进去。R 软件在这方面显得很“笨拙”。如果复制太多的数据到搜索路径中而没有清除，将会导致系统因过度装载而中断运行，或者会导致分析员混淆变量应该位于的路径。

Epicalc 提供了一种概念性的解决方法，当分析员工作于某个数据集时，只需采用几个命令就能解决问题。在 Epicalc 中，用户没有必要去专门识别某个数据集，并能非常有效地避免过度装载搜索路径。除了很容易清空内存以外，Epicalc 可以对在其他软件中（如 SPSS 或 Stata）已经加了标签的变量，或者在 Epicalc 自身中加了标签的变量进行调适，使变量含义变得容易理解。

R 软件有非常强大的作图功能，任何时候对一个变量的特征进行概括总结时，Epicalc 都能自动绘制非常好的特征分布图。这个变量用另一个分类变量进行分割时，图形也能很容易地自动生成。自动作图功能也能应用于单向和双向制表；变量和分类变量标签值的描述也能通过描述性图形很好地体现出来。

其他加入 Epicalc 中的流行病学模块，包括样本含量的计算、1∶n（n 可以变动）配对作表、Kappa 统计量，以及从一张表或 Logistic 回归模型的分析结果制作 ROC 曲线。R 软件有一些高级的回归模型功能。例如多项 Logistic 回归、有序 Logistic 回归、生存分析和多水平模型。应用 Epicalc 可以产生美观的 OR 值表和 95%可信区间，只要进行微小的修改就可以应用到论文中。

虽然 Epicalc 在 R 软件中的运行与常规方法不同，但是安装 Epicalc 对 R 软件中已有的功能或新功能没有任何影响。Epicalc 的功能仅仅是增加了 R 软件数据分析的有效性，使 R 软件的运行更容易。本书主要介绍 R 软件的使用并侧重于 Epicalc 的学习，读者应该具有一些基础的计算机应用知识。有了 R 软件、Epicalc 以及提供的数据集，用户应该能够通过每课的学习，了解数据管理的概念、相关的统计学原理、数据分析和强大的作图功能。

本书前 4 章介绍 R 软件的概念和一些重要的基本元素，比如标量、向量、矩阵、数组和数据框架的简单处理方法。第 5 章介绍一些简单的数据分析。第 6 章介绍日期和时间变量，并在第 7 章通过一些数据集来得到完全的诠释，描述性统计量和行列表伴随自动生成的图形，使得重要的结果能得到更全面的展示。第 8 章通过行列表来观察暴发，对各种类型的风险评估比如风险比率，可以用数字和图形来显示。第 9 章数据集的分析得到进一步扩展，处理不同水平间的联系或 OR 值，并对如何分层作表，计算 Mantel-Haenzsel 的 OR 值，以及 OR 值的同质性检验进行了详细的解释，同时附以图形说明。结合图形，混杂这个概念能够得以更好的理解。

深入剖析 R 软件以前，读者在第 10 章里可以对数据清除和标准数据操作进行充分的练习，并通过一些简单的循环语句以增加数据管理的有效性。随后也可以从这本书中学习如何应用这些循环语句来绘制精彩的图形。

第 11 章介绍散点图，简单线性回归和方差分析。第 12 章通过分层散点图的制作来增强混杂的概念以及连续性结果变量的交互作用。第 13 章讨论曲线模型。第 14 章是线性模型到广义线性模型。

第 15 章介绍二分类变量的 Logistic 回归，并与第 9 章中学到的分层作表进行比较。第 16 章讨论配对病例对照的概念，并介绍 1∶1 和 1∶n 配对表格，最后将介绍条件 Logistic 回归分析。第 17 章采用病例对照研究介绍了多分类 Logistic 回归分析。第 18 章介绍有序 Logistic 回归分析。

第 19 章介绍 Poisson 回归和不均匀分布的 Extra-Poisson 回归，包括应用负二项误差分布对结果进行建模。第 20 章讨论多水平模型和纵向数据分析。对随访时间的队列研究，第 21 章将介绍生存分析，第 22 章讨论 Cox 比例风险模型。

第 23 章介绍日常工作中的样本含量计算，专业分析者须掌握的文件技术将在第 24 章进行阐述。第 25 章介绍处理大型数据集的策略，第 26 章介绍处理有关态度方面的数据，最后一章介绍如何为撰写论文准备表格。

每一章节会给出一些参考文献，大多数章节附一些练习题，其解答将在本书末尾给出。

Virasakdi Chongsuvivatwong, M.D., Ph.D
泰国宋卡王子大学
（蔡　乐　译）

目　录

1 第1章 开始使用R软件

本章主要讨论第一次使用R软件时的一些基本操作，包括安装、数据录入、文件输出、命令文件的创建以及其他文档操作。注意：本书是在Windows操作系统的基础上写的。

安装

R软件是在GNU公共使用权限术语下使用的，即R软件是免费软件。R软件和Epicalc程序包的最新版本以及相关文件可以从CRAN（the Comprehensive R Archive Network）下载，该网址为：http://cran.r-project.org/。世界范围内有很多R的镜像储存地址，使用者可以选取距离最近的地址进行下载。

R的安装文件大约28MB。只要双击该文件并根据安装提示就可以进行安装。安装R后，R的快捷图标应该出现在桌面上。右击该图标可以修改R的启动属性，把缺省的开始文件夹替换为自己的工作文件夹，否则输入和输出文件将保存在程序文件夹下，这不是一种好的工作方式。用户可以根据不同的研究课题为工作文件夹创建多个快捷图标。

与本书有关的文件均储存在“C:\RWorkplace”文件夹下。图标的属性“Start in”的文本应设为“C:\RWorkplace”（不要键入双引号，它们在本书中代表某个对象或技术性名称）。

R检测电脑操作系统所使用的语言，并在菜单和对话框中用相应的语言显示。例如，如果使用中文Windows XP操作系统，则菜单和对话框都会以中文显示。由于本书是用英文写的，因此建议把语言设为英文，这样计算机上显示的所有信息就会与本书的一样。在R图标的属性选项“Shortcut”的“Target”中增加“Language=en”，注意在“Language”前加一个空格。

因此，R2.4.0版本的图标文本框为：“C:\Program Files\R\R-2.4.0\bin\Rgui.exe” Language= en，见图1。

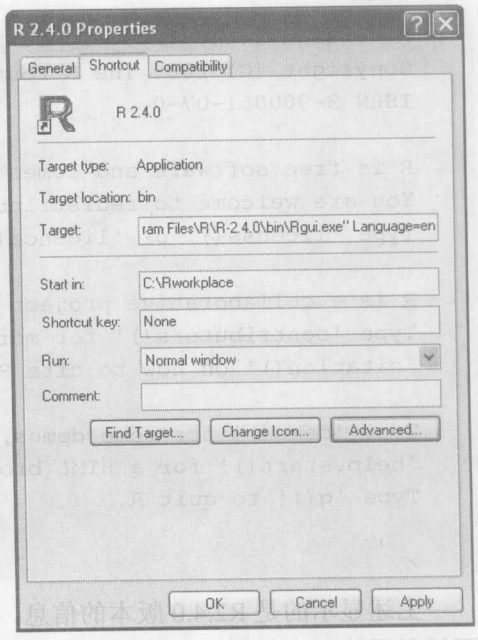

图1

为了有效地使用本书，必须安装一个专门的文本编辑器如 Crimson Editor 或 Tinn-R。此外，Epicalc 程序包也需要安装和加载。

Crimson 编辑器

这个软件的安装与其他软件的常规安装方法一样，只要运行 setup.exe 文件并按照提示安装即可。

Crimson 编辑器能辅助用户使用 R。对很多计算机软件比如 C++，PHP 以及 HTML 文件，Crimson 编辑器有很强大的命令文件编辑功能。它能显示行号并能匹配开括号和闭括号。这些特性很重要，因为在 R 的命令中会经常用到。

Crimson 的安装和设置将在第 24 章介绍。

Tinn-R

Tinn-R 可能是与 R 软件联合使用的最好的文本编辑器，它是为 R 的命令文件专门设计的。除了突出显示 R 代码的语法外，Tinn-R 可以与 R 交互使用一些特殊的菜单和工具按钮。这意味着 Tinn-R 可以高亮度显示命令的某些部分，并通过单击按钮直接发送到 R 中运行。Tinn-R 可以从 www.sciviews.org/Tinn-R 下载。

开始使用 R 软件

当修改 R 图标的启动属性以后，双击桌面上 R 的图标即可。R 启动后屏幕将会显示以下信息：

```
R version 2.4.0 (2006-10-03)
Copyright (C) 2006 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for a HTML browser interface to help.
Type 'q()' to quit R.

>
```

上述显示的是 R2.4.0 版本的信息，该版本于 2006 月 10 月 3 日发布。第二段信息简要解释了该版本的授权和版权问题，第三段信息是关于 R 软件的贡献者以及如何在文献

中引用 R 的问题，第四段信息是为第一次使用 R 的用户提供一些建议使用的命令。

在本书中，R 命令以“>”符号开始，与在 R 控制台窗口中显示的类似。用户不用键入这个符号，只需键入命令即可。R 命令和输出行均以 Courier New 字体显示，而注释文本以 Times New Roman 字体显示。Epicalc 命令以斜体字显示，而标准的 R 命令以正常字体显示。

第一步是退出 R。单击窗口最右上角的叉号按钮或者直接键入如下 R 命令：

```
> q()
```

一个对话框将出现在屏幕询问“Save workspace image？”并会出现三个选择：“Yes”，“No”and“Cancel”。如果选择“Yes”，两个新文件将会创建在工作目录下。任何先前键入的 R 命令都会保存在一个名为“.Rhistory”的文件中，而现在使用的工作空间将保存在名为“.Rdata”的文件中。注意这两个文件名都没有前缀。在下一次的计算中，当 R 在这个文件夹启动后，上次保存的工作环境和使用过的命令将自动恢复出来。继续在这种方式下使用 R（退出和保存未命名的工作空间）将导致这两个文件变得越来越大。通常人们喜欢每次更新启动 R，因此当屏幕询问“Save workspace image？”时建议选择“No”，或者也可以键入：

```
> q("no")
```

这样退出 R 时就不会保存工作空间并防止上述对话框的出现。

也可以通过键入如下命令在退出 R 前保存工作空间：

```
> save.image("C:/RWorkplace/myFile.RData")
```

“my File”是文件名，当你退出 R 时应该选择回答“No”。

R 的程序库和程序包

R 被定义为包含许多传统和高级统计方法的运行环境，称为函数。一些方法被构建进 R 的基础环境中，但大多数方法是用程序包来补充。一个程序包是一些函数、数据集和文件的简单组合。R 的程序库是程序包的集合，通常单独集中在计算机的一个文件夹中。

R 有 25 个程序包（称为“标准”或“推荐”的程序包），更多的程序包可以通过 CRAN 的网站获得。当执行 R 时只有 7 个程序包被装载到内存中。为了知道是哪些程序包被装载进了内存中，你可以键入：

```
> search()
[1] ".GlobalEnv"        "package:methods"   "package:stats"
[4] "package:graphics"  "package:grDevices" "package:utils"
[7] "package:datasets"  "Autoloads"         "package:base"
```

上述显示的是存放在 R 搜索路径中的程序包。当 R 被用于执行某项工作时，它将从搜索路径中寻找相应的目标进行操作。首先，它将从一个全球环境‘.GlobalEnv’内查看，这总是第一个搜索位置。如果 R 不能找到它需要的目标，它将会从第二个搜索路径“package:methods”去找，如果找不到再找下一个，依次类推。任何属于上述装载的软件包中的函数在 R 中都可以找到。