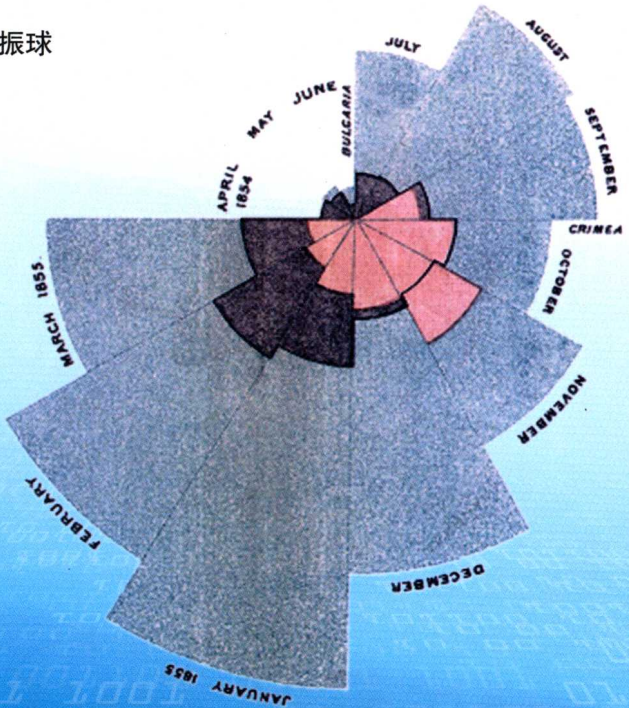
- 深入浅出，丝丝入扣，快速掌握 R 语言数据可视化的方法与技巧
- 玩转 R 语言，开心做科研

R 语言与医学统计图形

Medical Statistics Graphics with R

主　编　张铁军　陈兴栋　刘振球



网络增值服务
ONLINE SERVICES

人民卫生出版社

语言

与医学统计图形

Medical Statistics Graphics with R

主　审　何　纳

主　编　张铁军　陈兴栋　刘振球

编　者（以姓氏笔画为序）

方　圆（复旦大学）

艾自胜（同济大学）

吕　明（山东大学齐鲁医院）

刘振球（复旦大学）

严　琼（复旦大学）

杜　雨（东北财经大学）

谷鸿秋（国家神经系统疾病临床医学研究中心）

张　杰（香港理工大学）

张铁军（复旦大学）

陈兴栋（复旦大学）

姚应水（皖南医学院）

索　晨（复旦大学）

徐　萍（中国科学院上海分院）

秘　书　方绮雯（复旦大学）

人民卫生出版社

图书在版编目(CIP)数据

R语言与医学统计图形 / 张铁军，陈兴栋，刘振球主编. —北京：人民卫生出版社，2017

ISBN 978-7-117-25728-2

Ⅰ. ①R… Ⅱ. ①张… ②陈… ③刘… Ⅲ. ①程序语言－应用－医学统计 Ⅳ. ①R195.1-39

中国版本图书馆CIP数据核字(2017)第314285号

R语言与医学统计图形

主　　编：张铁军　陈兴栋　刘振球

出版发行：人民卫生出版社（中继线 010-59780011）

地　　址：北京市朝阳区潘家园南里19号

邮　　编：100021

E - mail：pmph @ pmph.com

购书热线：010-59787592　010-59787584　010-65264830

印　　刷：北京铭成印刷有限公司

经　　销：新华书店

开　　本：710×1000　1/16　　印张：18

字　　数：333千字

版　　次：2018年2月第1版　2018年2月第1版第1次印刷

标准书号：ISBN 978-7-117-25728-2/R·25729

定　　价：89.00元

打击盗版举报电话：010-59787491　E-mail：WQ @ pmph.com

（凡属印装质量问题请与本社市场营销中心联系退换）

序　言

几千年的人类文明历程中，随着科学技术的不断发展，人类认知世界的方式也在不断改变。举例来说，我们的古人认为天圆地方，这是他们用眼睛观察的结果，当然肯定也经过了大脑的思考。虽然从今天看来，这是一个错误的结论，但是我们并不能因此否定“观察”这一认知世界的最直接方式。近代以来，随着现代科学的快速发展，人类观察到了以前肉眼不可见的物质，比如细胞、遥远的天体、遗传物质等。这些发现均是通过“观察”而获得的，它们的出现大大地推动了人类文明的进程，让人类认知世界的方式由表及里，由定性到定量，由单纯现象发展到一定的规律，这就正如马克思教导我们要认识物质的本质一样。

流行病学的发展，也是一样的。1758 年 Lind 用水手多吃水果和蔬菜来预防坏血病，后来证实维生素 C 能预防坏血病。1767 年 Bakor 报告饮苹果酒者经常发生腹绞痛，后来才发现腹绞痛系铅酒壶的铅溶入酒中引起。18 世纪 Pott 报告扫烟囱的小孩易得阴囊癌，后来证实系煤烟灰中含多环芳烃（PAH）引起的。1854 年 Snow 发现伦敦宽街饮用污染井水居民腹泻发病高，不饮者低，控制后很少发病。30 年后 Koch 在水中分离到霍乱弧菌。20 世纪 80 年代后计算机技术迅速发展，我在多伦多大学访问期间，花了不少时间学习使用计算机分析肝癌危险因素、解决宫颈癌合适的普查方案、乳腺癌与营养关系等课题。

计算机开始应用的年代还需要自己编写复杂的程序，后来有许多商业软件如 SPSS、SAS 等，只要付费即可应用。随着时代的前进，出现了无需付费的 R 软件，它有一套完整的数据处理系统，它还有优秀的统计制图功能。简而言之，R 简洁而强大。

这本书系统介绍了 R 的绘图功能，目的就是让科研工作者拥有良好的科研构图能力，它将支持你的学习和科研。

俞顺章

2017 年 12 月 20 日

前　言

一图胜千言！这是对图形最好的概括。一张优质的统计图形，不仅能够准确、生动的展示出我们的结果，更能给人一种数据之美、图形之美的享受。同时，在科研论文中添加一张精美的统计图形，对于文章整体的质量也是一种提升。

那么，如何利用数据绘制出一张准确而精美的统计图形呢？相信有很多软件可以做到。但是，毫无疑问，从实现的简便性和实现的效果上来说，R 语言是这方面的翘楚。

随着数据科学和计算机科学的不断发展，R 语言近十年也在蓬勃发展，它在统计建模、机器学习、数据挖掘、生物信息等方面表现十分抢眼。同时，R 语言的绘图功能也一直为人所称道。无论是基础绘图系统还是 *ggplot2* 绘图系统，均可以利用少量的代码绘制出精美的统计图形。因此，近几年，R 语言的身影在科研论文中，包括 *Lancet*、*JAMA* 之类的顶级医学期刊论文中，屡见不鲜。这种现象提示我们，一篇优质的科研论文，除了研究设计、数据分析、论文撰写这些重要环节外，结果的呈现方式也非常重要。这也就是我们撰写这本书的初衷——让更多优质的统计图形出现在我们的科研论文中。

本书从结构上来说，主要分为六大篇，共十四章，其中，前三篇主要介绍静态图形的绘制，包括基础绘图包、*ggplot2*、地图等方面；第四篇介绍了 R 语言几种主要的动态交互绘图系统，比如 *plotly*、*recharts*、*leaflet* 等；第五篇介绍了一些另辟蹊径的图形和医学科研中独有的统计图形，比如生存曲线和 meta 分析森林图；第六篇介绍了 R 语言中统计表格的制作方法。与此同时，笔者在行文的过程中，穿插了很多实际工作中可能遇到的问题，比如图形颜色的选取、高质量图形的保存与导出、中文字体的选择等等。在代码的编排上，除了代码中不同类型文本之间（比如数字和字符串）有颜色区分外，笔者也做了大量的代码注释，以帮助读者更好的理解。此外，文中使用的所有数据，要么来自于 R 语言的内置数据集，要么是笔者利用函数构建的随机数据

集，因此读者都可以轻松获取相应的数据进行代码练习，但是有一点需要声明，文中所有随机生成的数据集，比如肿瘤发病率数据集，均是为了图形的展示，并无任何实际意义。本着“授人以鱼不如授人以渔”的原则，书中大量的代码和图形均是从简单入手，其目的是为了使读者更好的掌握 R 语言的绘图技巧，而一些复杂的图或者“成品图”在书中鲜有出现。绘图是一种创意的考验，但是创意是建立在良好的基础之上的。由于篇幅有限，书中部分图形被封装到每个章节的二维码中，读者可以扫码看图。

江山代有才人出，各领风骚数百年。R 语言很小巧（安装包仅有 70M），但是强大，R 语言很庞大（至今已有超过 13 000 个扩展包），但是优雅。在本书中，希望你能跟随笔者的脚步一起感受 R 语言的魅力，体会数据之美，图形之美。

本书的编写得到了国家重点研发计划、国家自然科学基金、教育部博士点基金、上海市自然科学基金的资助。我们特别邀请了我国著名流行病学家俞顺章教授与何纳教授对本书进行审阅，两位老师丰富的学识和严谨的科学态度为本书增色不少。感谢安徽医科大学段禹同学和复旦大学左佳鹭、袁黄波以及蔡宁同学在文字编排上的协助；感谢山东大学张文超同学在本书封面设计上的帮助；同时感谢“医学方”微信公众平台的支持。本书编撰的同时也收到了大量的网友反馈，他们所提出的意见和建议帮助我们进一步完善了本书的内容，在此一并致谢！

我们求学于复旦、成长于复旦，衷心感谢复旦大学对于本书的大力支持；感谢每一位编委的辛劳付出；感谢 R 语言道路上的先驱与前辈，是你们的智慧成就了 R 语言的今天，才让我们能够站在巨人的肩膀上继续前行。

由于笔者水平有限，对于 R 语言理解有限，书中难免有疏漏错误之处，恳请各位专家、老师、前辈、同学批评指正（邮件可发送至 zhenqiuliu@outlook.com），谢谢！

张铁军　陈兴栋　刘振球

2017 年 6 月 20 日于西苑 8 号楼

目　录

第四篇 动态交互绘图系统

第五篇 绘图番外篇

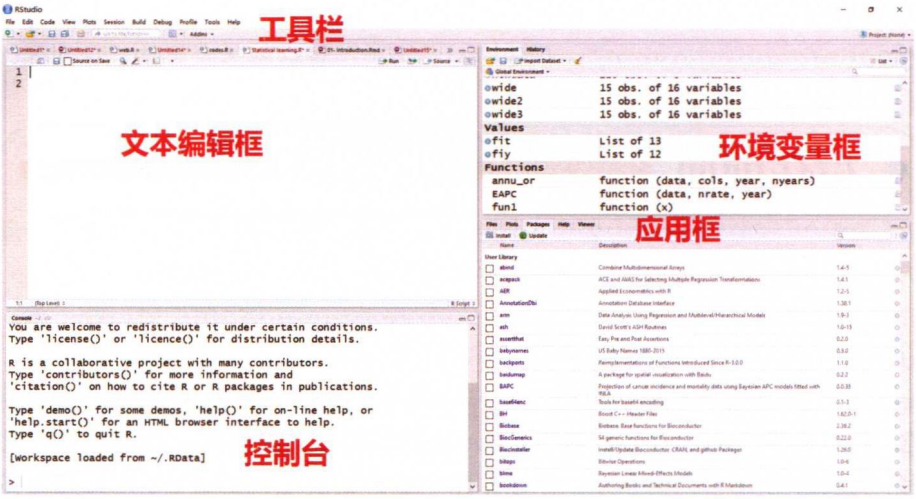
第六篇 统计表格的制作

写在前面

R 语言是近十年来快速崛起的一门以数据分析为特色的计算机编程语言，对于还未接触过它的读者来说，一切都是陌生而新鲜的。开篇的这段文字，仅仅针对 R 语言的新手，而对于有一定 R 语言基础的读者，比如知道如何下载安装 R 语言，就完全可以跳过这一段。R 语言是完全免费的开源软件，大家可以在其官方网站上（https://cran.r-project.org/），根据自己的计算机操作系统，选择相应的版本进行下载。截至 2017 年 6 月 30 日，R 语言的最新版本为 3.4.1。R 语言的安装全部是点击式操作，因此不涉及复杂的编译或配置过程，故在此不详述。安装完毕后，Windows 系统用户可以在桌面上发现两个 R 语言的快捷方式图标，其中一个是基于 64 位系统，另外一个是基于 32 位系统，两者并不冲突，因此可以不予理会。双击其中一个图标，即可进入 R 语言的图形用户界面（R GUI）。该界面相对简陋，对于非计算机专业的用户来说极不适应，因此在这里给大家推荐另外一款与 R 语言配套的集成开发环境（IDE）——Rstudio。Rstudio 是目前最流行的 R 语言 IDE，其界面友好，操作方便，针对普通用户的版本同样是免费的。大家可以在其官网下载最新版本（https://www.rstudio.com/）。Rstudio 的安装同样简单方便，大家无需将 R 语言与 Rstudio 安装在同一目录下，只需要注意一点，即先安装 R 语言，再安装 Rstudio 即可。

安装工作完毕之后，大家可以直接打开 Rstudio 进行操作，此时 R 语言是无需同时打开的。Rstudio 的基本界面如下图所示。

其中，文本编辑框的功能类似于一个普通的文本编辑器，主要用于代码的编写，此处编写的代码不会自动运行，需要选中相应的代码，然后点击文本编辑框右上角的绿色“Run”按钮，或者使用“Control + R”（Windows），“Commond + R”（Mac OS）的键盘组合运行代码。控制台是 R 语言代码输入和结果输出的地方，在此处键入代码，摁下“enter”键即可即时得到相应的结果。从文本编辑框中运行的代码，其代码和结果也会在控制台显示出来。右上角的环境变量框展示的是在不同操作环境中的变量和数据（默认是全局环



Rstudio 基本操作界面

境，即 global environment）。右下角是 Rstudio 特有的应用框，从左至右依次是“Files”（文件展示窗口），“Plots”（图形展示窗口），“Packages”（包展示窗口），“Help”（帮助文档窗口），“Viewer”（视图窗口）。

R 语言是函数式编程的计算机语言，在实际应用中，我们会使用大量的已经封装好的函数。比如求解一组数据的算数平均数，不需要逐个相加求和再除以数据的个数，而只需要调用 mean() 函数（注意，在接下来的行文中，凡是 R 语言函数，均会带上小括号，以便同普通文本区分）。函数就像一台机器，如果想要得到输出，就必须要有输入才行。在函数内部，存在若干参数，比如 mean() 函数中的 x，这些参数就像是机器的控制按钮，选择的参数不同，得到的结果也会不同，当然，这其中部分参数是“必需参数”，即该参数不能忽略，一旦忽略则会报错。另外一些参数称为“非必需参数”，只有在执行特定需求时才会使用它们，比如 mean() 函数中的 na.rm 参数，该参数针对的是原始数据中存在缺失值的情形，如果原数据没有缺失值，则该参数可以忽略。还有一类参数称为“缺省参数”，也称作“默认参数”，通俗来讲，即该参数具有一个“天生”的值，如果不对其进行修改，那么每次默认都使用该值。“非必需参数”和“缺省参数”界限并非十分明显，但是“必需参数”是每一次都必须要接受相应值的。

R 语言中还有一个显著的特点，就是包（package）的存在。包就像 R 语言的“弹药库”，不同的包其功能不全相同或者完全不同（因为研究领域不同）。在 R 语言安装好之后，大家会发现已经有大约 20 几个包存在，比如 *MASS*, *utils* 等，我们可以把这些包称为“系统包”（system library），它们无需

再次下载，就像战士随身配备的手枪和匕首，只需使用 library() 函数进行加载就可调用（部分包，比如 *utils*，甚至无需加载就可以直接使用）。而对于其他包，当我们需要使用时，第一步是下载它。如果该包已经在 CRAN 上发布，那么我们可以使用 install.packages() 函数下载，比如下载 *ggplot2*，可以使用如下代码：

```
install.packages('ggplot2')# 包的名称需要用引号包括
```

下载好之后，如果需要使用它，则必须使用 library() 函数对其进行加载。

第一篇

R语言基础绘图系统

第一章

基础绘图包之高级绘图函数

R 语言有强大的绘图系统，数据可视化一直是 R 语言独步江湖的强大杀招。与 R 语言美轮美奂，妙不可言的绘图相比，其他常用的统计软件，比如 SPSS 和 SAS，在统计绘图方面则显得相形见绌。Python 作为当下非常流行的一种通用计算机编程语言，在统计绘图方面也在向 R 语言学习。

R 语言至今已拥有超过 13 000 个扩展包（packages），大部分能在 CRAN（comprehensive R archive network，https://cran.r-project.org/）上找到，小部分还未正式发布，托管在 Github 上或者 RForge（https://r-forge.r-project.org/）上。在这些扩展包中，专门用作数据可视化的包不下 100 个，而各种形形色色的包，多少都会自带一些绘图函数，以满足自身“特异性”的数据可视化的需要，比如用于量化金融的 *quantmod* 包，就有专门用于绘制股票走向的蜡烛图函数。

在众多绘图包中，*ggplot2* 是当之无愧的“武林盟主”！曾有人断言，*ggplot2* 的出现，将统计绘图提升了一个档次，到了一个全新的境界。而依托 *ggplot2* 开发的绘图包也超过 20 个，且个个都能独当一面，成为各领域数据可视化的翘楚！比如香港大学余光创博士开发的 *ggtree* 包，专门用来绘制各式各样的进化树。

随着 R 语言的不断发展与普及，各种优秀的绘图包如雨后春笋般不断涌现，但是，任凭风起云涌，沧海桑田，R 语言基础绘图包的“江湖地位”一直不可动摇。基础绘图包并不是单独指某一个包，而是由几个 R 包联合起来的一个“联盟”，比如 *graphics*、*grDevices* 等。这些包在我们第一次装载 R 语言时就会被自动下载，因此无需手动下载，而且在初次使用的时候，也无需使用 library() 函数对其进行加载。曾几何时，这个联盟也是独领风骚，好不风光！时至今日，面对众多俊俏后生，基础绘图包更像是一个看惯了世事纷争，看淡了潮起潮落的老者，老成持重的守着自己的一方土地。

第一节 par() 函数详解

一、par() 函数简介

为什么一开始就介绍 par() 函数，而不是基础绘图包中的绘图函数？笔者认为基础绘图包之所以能够很好的完成绝大多数的数据可视化任务，最主要的原因就是其强大而深厚的“内功”——par() 函数！

par 是英文单词 parameters 的简写，即参数，简而言之，这是一个专门用来设置绘图参数的函数！何为参数？参数是一个函数的肉体，没有参数，函数体则没有相应的输入值，函数也就无从运算；参数同样是一个函数的灵魂，没有参数，函数会一成不变，无法做到灵活运用。我们在实际应用中，通过改变参数的传入值，得到不同的结果。更多关于参数的内容，在接下来的章节中，大家会大量的接触到，所以，其真正的含义，大家会逐渐明晰。

我们的第一步任务就是要了解和熟悉 par() 函数的参数，请首先打开 R Gui 或者 Rstudio，在控制台键入：

```
help("par")
#所有符号必须英文状态！更简单的写法是直接在par前面添加一个问号即可，对于一些非系统包中的函数或其他名词，在进行搜索时，可以在其前面添加两个问号，比如"??linear model"。
```

Rstudio 会弹出一个帮助窗口（如果是 R Gui，则会弹出一个网页窗口），一般在电脑屏幕的右下角。par() 函数参数众多，功能不一，但是笔者认为其中能经常使用到的，或者说对我们的统计图形能够起到本质改变作用的也就 20 个左右。若能把这 20 个参数熟练掌握也就可以了，其余的呢，在遇到不同的数据可视化需求时，再来现学也不迟。另外，par() 函数以及接下来介绍的其他函数，只要属于基础绘图包，在开启 R 之后，是无需加载扩展包的，直接可以使用。

二、par() 函数参数介绍

par() 函数目前大概有 60 多个参数，但是此处只选取其中大约 20 个较为常用的参数进行详解。其余的参数，请读者朋友自行翻阅帮助文档。

1. adj 英文单词 adjustment 的简写，用于调整图中字符的相对位置，属于微调。取值：adj = c(x, y)，表示字符边界矩形框的左下角相对坐标点（x，