统计计算与图形

黄湘云

2018-09-19 11:53:55 CST

目录

欢迎

R 软件主要用于统计计算和统计绘图，因其提供了完整的绘图系统，实现了大量的统计方法，而且相比于其他统计软件，具有免费和更新快的特点，在当今数据时代浪潮下，占有一席之地。Markdown 是一个文本标记语言。欢迎来到 R 语言的世界，写点关于发展历史等容易吸引人的东西，让读者有继续看下去的欲望。

结构

介绍书籍写作工具，写作风格设定，结构说明，R 语言介绍[1]

(a) 历史　　(b) 现在

图 1: R 语言

历史

历史数据分析

[1]https://www.r-project.org/about.html

```
# 获取数据
pdb <- tools::CRAN_package_db()
pdb <- pdb[,c("Package","Published")]
# pdb <- readRDS(file = "data/pdb.RDS")
library(ggplot2)
ggplot(pdb[,c("Package","Published")], aes( as.Date(Published) ) ) +
  geom_bar(color = 'springgreen4') +
  geom_line(data = data.frame( date = as.Date( c("2011-01-01","2012-10-20") ),
                               count = c(130,148)), aes(x = date , y = count),
            arrow = arrow(angle = 15, length = unit(0.15, "inches"))) +
  annotate("text", x = as.Date("2010-11-01"), y = 128, label = "(2012-10-29,148)") +
  scale_x_date(date_breaks = "1 year",date_labels = "%Y") +
  labs(x = "Published Date" ,y = "Count" ) +
  theme_minimal()
```

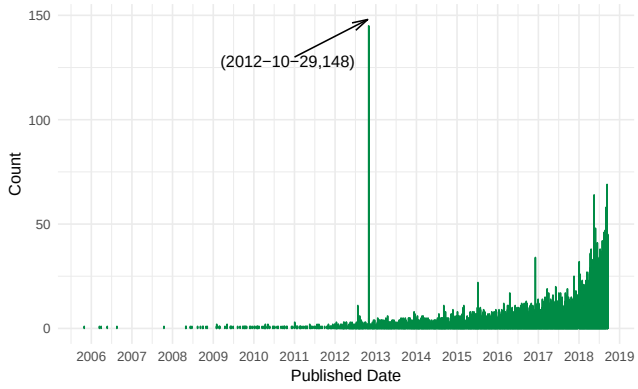


图 2: 有趣的是在 2012 年 10 月 29 日更新的 R 包多达 148 个

R 语言进入到各方各面

如何重复本书，写作工具，如何提交 PR，参与写作

益辉维护的 R Markdown 生态及其它 R 包，如图 4

[2]2018 年 8 月 9 日

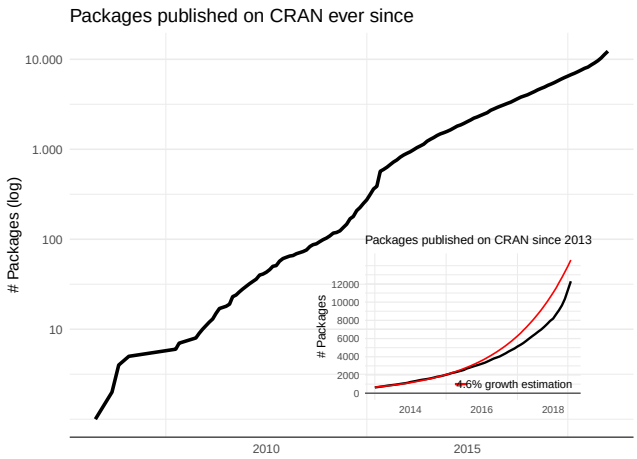


图 3: R 包发布量变化趋势

package	published	dl_last_month	stars	forks	last_commit	depends_count	watchers
animation	2017-03-30	8332	130	53	0.8	1	18
blogdown	2018-07-15	4601	768	172	0.3		78
bookdown	2018-02-18	7054	1061	479	0.1		86
DT	2018-01-30	52185	291	105	0.1		57
evaluate	2018-07-17	260513	62	23	0.8	1	7
formatR	2017-04-25	40403	141	39	8.2	1	13
fun	2011-08-12	618	27	29			7
highr	2018-06-09	207813	27	9	2	1	5
knitr	2018-02-20	255314	1613	701	0	1	131
markdown	2017-04-20	218789	52	65	11.6	1	29
mime	2016-07-07	255885	16	5			3
MSG	2016-02-13	129	19	5			7
printr	2017-05-19	861	91	24	14.9		6
Rd2roxygen	2018-08-02	196	19	9			3
rmarkdown	2018-06-11	196734	1218	559	0.1	1	126
rticles	2018-07-06	7552					
servr	2018-05-30	4564	164	21	1.2	1	19
testit	2018-06-14	4611	22	9	0.4		5
tinytex	2018-07-07	172343	217	21	0		9
tufte	2018-07-15	4002	153	52	0.8		21
xaringan	2018-07-10	1253	511	91	1		55
xfun	2018-07-06	170484	18	5	0.1		5

图 4: 益辉在 Github 上维护的 R 包[2]

后记

这本书是在 RStudio 内用 R Markdown (Xie et al., 2018) 写的，Git 控制版本，bookdown (Xie, 2016) 组织章节，knitr (Xie, 2015) 调用各类编程语言或解释或编译代码块，将执行结果返回到 R Markdown 文件，Pandoc 再将其转化为 Markdown 和 HTML 文档，进一步转化为 PDF 格式文档则需要 TinyTeX 发行版[3]。

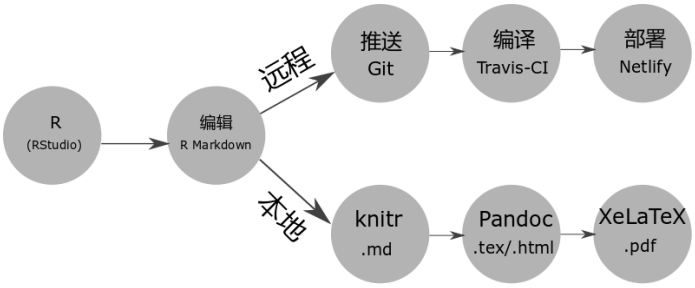


图 5: 工作流程图[4]

这个网站是通过 Travis-CI 把编译结果（即 _book 目录）推送到 Netlify 实现部署。在 Travis-CI 和 Netlify 都与 Github 绑定的情况下，源代码一发生改变就会触发编译，编译成功就会自动部署，这个过程即持续集成和连续部署，你正在阅读的版本是 2018-09-19 在 Travis 上构建的。

说明

Alegreya 罗马体显示正文，AlegreyaSans 等线体显示数字，sourcecodepro 等宽体显示代码，Alegreya 字体源文件见 https://github.com/huertatipografica/ 和样式见 https://huertatipografica.com/en，我们可以通过安装 LaTeX 包 alegreya 获得。R 包名称在文中以粗体显示，代码块输出用 #> 表示，以区分普通的代码注释 #

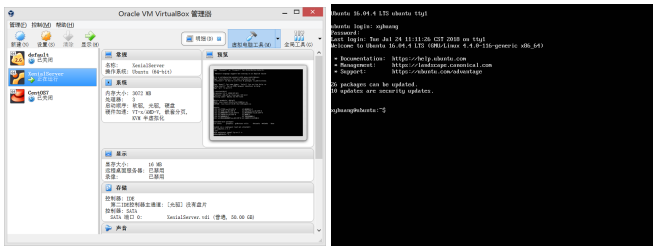
绘图使用的中文字体是思源宋体和思源黑体，由 **showtext** 包安装和调用，**tikzDevice** 和 **fontcm** 处理其中的数学公式，**xkcd** 设置漫画手写体风格。

我的写作环境是 VBox + Ubuntu 16.04.4 Server + PuTTY + Xming，如图 6 所示。

书籍生成过程中，R 进程和 Pandoc 版本信息如下：

[3] https://yihui.name/tinytex/

[4] Inkscape 绘制



(a) VBox 虚拟机 (b) Ubuntu Server 系统

(c) PuTTY SSH 登陆客户端 (d) 显示服务器 X Windows System

图 6: 远程办公四剑客

```
xfun::session_info()
#> R version 3.5.0 (2017-01-27)
#> Platform: x86_64-pc-linux-gnu (64-bit)
#> Running under: Ubuntu 14.04.5 LTS
#>
#> Locale:
#>   LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
#>   LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
#>   LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
#>   LC_PAPER=en_US.UTF-8       LC_NAME=C
#>   LC_ADDRESS=C               LC_TELEPHONE=C
#>   LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
#>
#> Package version:
#>   assertthat_0.2.0   backports_1.1.2    base64enc_0.1.3
#>   bookdown_0.7       cli_1.0.0          codetools_0.2-15
#>   colorspace_1.3-2   compiler_3.5.0     crayon_1.3.4
#>   curl_3.2           digest_0.6.17      evaluate_0.11
#>   fansi_0.3.0        ggplot2_3.0.0      glue_1.3.0
```

```
#>   graphics_3.5.0     grDevices_3.5.0    grid_3.5.0
#>   gtable_0.2.0       highr_0.7          htmltools_0.3.6
#>   jsonlite_1.5       knitr_1.20         labeling_0.3
#>   lattice_0.20.35    lazyeval_0.2.1     magrittr_1.5
#>   markdown_0.8       MASS_7.3.49        Matrix_1.2.14
#>   methods_3.5.0      mgcv_1.8.23        mime_0.5
#>   munsell_0.5.0      nlme_3.1.137       pillar_1.3.0
#>   plyr_1.8.4         R6_2.2.2           RColorBrewer_1.1.2
#>   Rcpp_0.12.18       reshape2_1.4.3     rlang_0.2.2
#>   rmarkdown_1.10     rprojroot_1.3-2    scales_1.0.0
#>   stats_3.5.0        stringi_1.2.4      stringr_1.3.1
#>   tibble_1.4.2       tinytex_0.8        tools_3.5.0
#>   utf8_1.1.4         utils_3.5.0        viridisLite_0.3.0
#>   withr_2.1.2        xfun_0.3           yaml_2.2.0
rmarkdown::pandoc_version()
#> [1] '2.3'
```

授权

第一部分

统计图形

介绍

介绍这部分讲什么内容，如何展开

第一章　基础图形

数据可视化是一种重要的数据分析手段，Claus O. Wilke 目前正在写《Fundamentals of Data Visualization》，并且给出了在线网络版[1]。

1996 年 R 语言横空出世 (Ihaka and Gentleman, 1996) 带数学符号的注释 (Murrell and Ihaka, 2000)

结合统计意义和探索性数据分析介绍各种常见统计图形

plot 函数对象

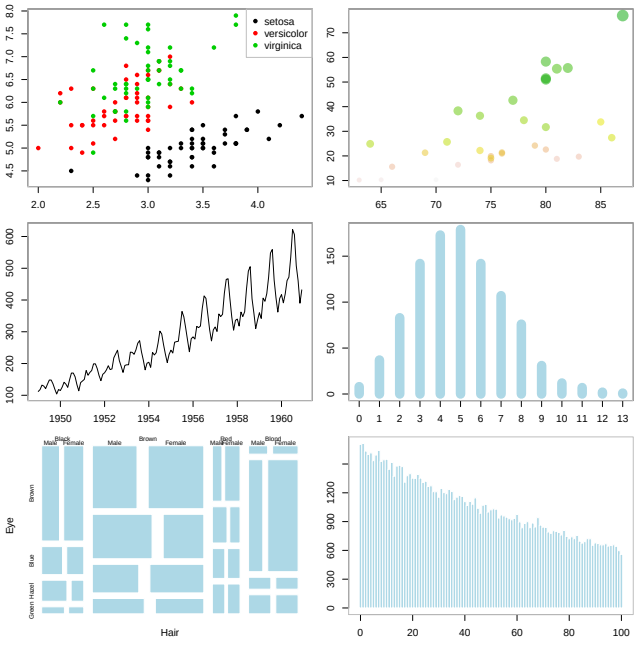


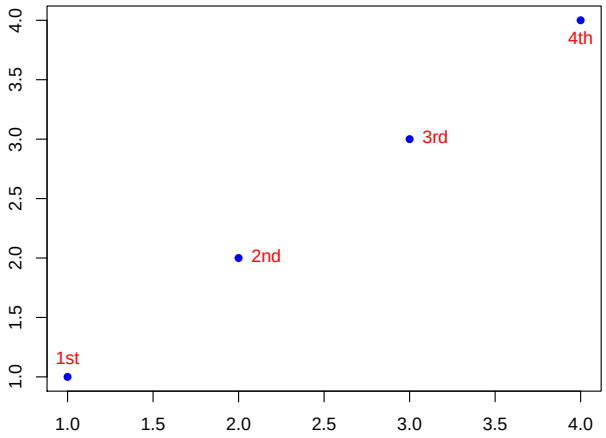
图 1.1: plot 函数对象

在介绍各种统计图形之前，先介绍几个绘图函数 `plot` 和 `text` 还有 `par` 参数

[1]https://serialmentor.com/dataviz/

设置，作为最简单的开始，尽量依次介绍其中的每个参数的含义并附上图形对比。

```
x <- 1:4
y <- x
plot(x, y, ann = F, col = "blue", pch = 16)
text(x, y,
  labels = c("1st", "2nd", "3rd", "4th"),
  col = "red", pos = c(3, 4, 4, 1), offset = 0.6
)
```



其中 labels，pos 都是向量化的参数

1.1　散点图

散点图（点图），抖动图（箱线图），一维的二维的

高亮某些点，按类别绘散点图[2]

```
data("iris")
pch <- rep(16, length(iris$Petal.Length))
pch[which(iris$Petal.Length < 1.4)] <- 17
```

[2]https://stackoverflow.com/questions/51804892/use-vectorized-plotting-arguments-when-plotting-by-factor-in-r

```
stripchart(Petal.Length ~ Species,
  data = iris,
  vertical = TRUE, method = "jitter",
  pch = pch
)
```

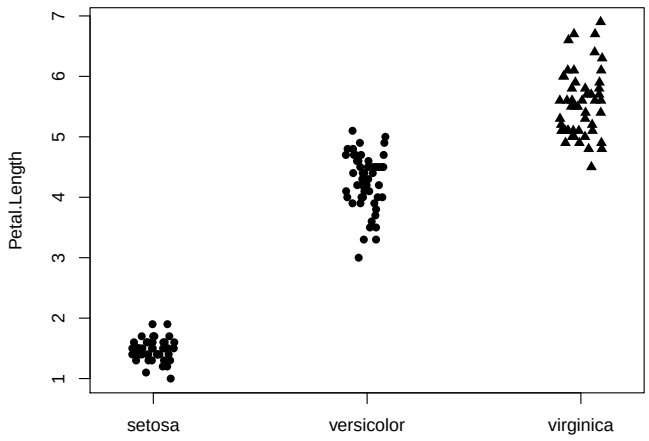


图 1.2: 错误的散点图画法

```
methods(stripchart)
getAnywhere(stripchart.default)
getAnywhere(xy.coords)
```

pch 没有向量化实际只是取了前三个值 16 16 17 对应于 Species 的三类

高亮某些点关键是高亮的分界点是由区分意义的

```
data("iris")
stripchart(Petal.Length ~ Species,
  data = iris, subset = Petal.Length > 1.4,
  vertical = TRUE, method = "jitter", ylim = c(1, 7),
  pch = 16
)
stripchart(Petal.Length ~ Species,
  data = iris, subset = Petal.Length < 1.4,
```

```
  vertical = TRUE, method = "jitter", add = TRUE,
  pch = 17
)
```

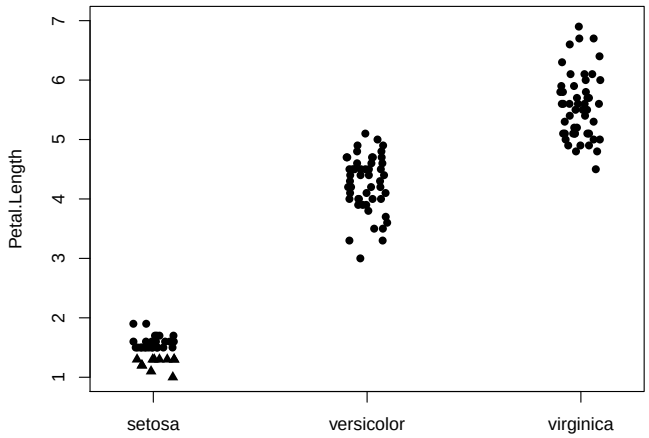


图 1.3: 正确的散点图

1.2　折线图

应用：时序图

1.3　条形图

条形图

应用：自协方差、自相关、偏自相关、协相关和协方差图，

1.4　直方图

```
with(data = faithful, {
  hist(eruptions, seq(1.6, 5.2, 0.2), prob = TRUE)
```

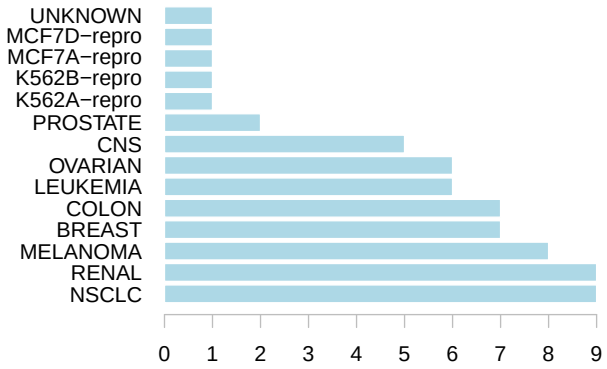
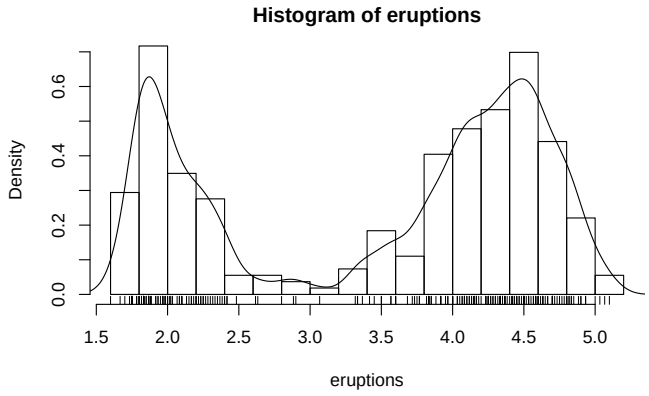


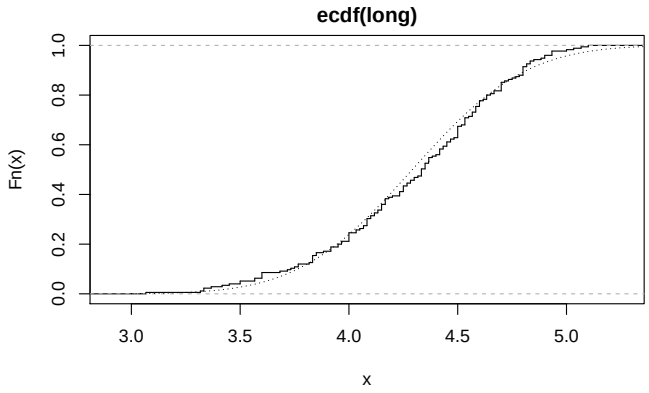
图 1.4: 条形图

```
  lines(density(eruptions, bw = 0.1))
  rug(eruptions) # show the actual data points
})
```



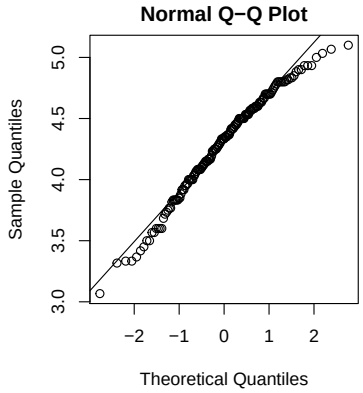
1.5　经验累积分布图

```
with(data = faithful, {
  long <- eruptions[eruptions > 3]
  plot(ecdf(long), do.points = FALSE, verticals = TRUE)
  x <- seq(3, 5.4, 0.01)
  lines(x, pnorm(x, mean = mean(long), sd = sqrt(var(long))), lty = 3)
})
```



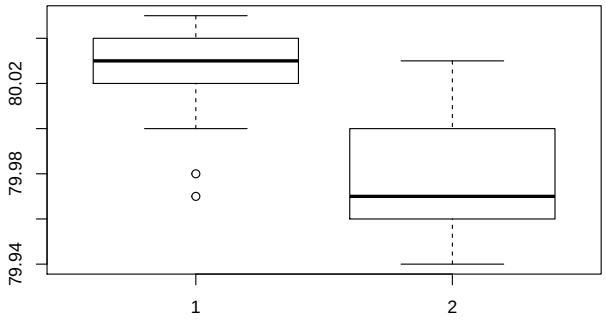
1.6　QQ正态分布图

```
with(data = faithful, {
  long <- eruptions[eruptions > 3]
  par(pty = "s") # arrange for a square figure region
  qqnorm(long)
  qqline(long)
})
```



1.7　箱线图

```
A <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04, 79.97,
       80.05, 80.03, 80.02, 80, 80.02)
B <- c(80.02, 79.94, 79.98, 79.97, 79.97, 80.03, 79.95, 79.97)
boxplot(A, B)
```




```
with(data = iris, {
  op <- par(mfrow = c(2, 2), mar = c(4, 4, 2, .5))
  plot(Sepal.Length ~ Species)
  plot(Sepal.Width ~ Species)
  plot(Petal.Length ~ Species)
  plot(Petal.Width ~ Species)
  par(op)
  mtext("Edgar Anderson's Iris Data", side = 3, line = 4)
})
```

1.8　等高线图

contour

1.9　透视图

persp 多元分布函数图像

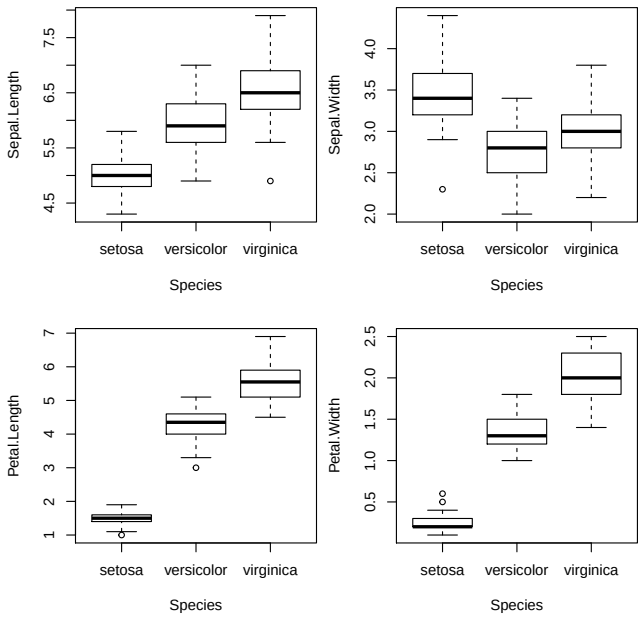


图 1.5: 安德森的鸢尾花数据

1.10　热图

image

应用：heatmap, raster 图像

1.11　树图

dendrogram

层次聚类/分类/回归树

1.12　图形参数

`par()` 图形版面设置

1.13　数学注释

数学符号注释，图1.6 自定义坐标轴 (Murrell and Ihaka, 2000)。

```
# 自定义坐标轴
plot(c(1, 1e6), c(-pi, pi), type = "n",
     axes = FALSE, ann = FALSE, log = "x")
axis(1, at = c(1, 1e2, 1e4, 1e6),
     labels = expression(1, 10^2, 10^4, 10^6))
axis(2, at = c(-pi, -pi / 2, 0, pi / 2, pi),
     labels = expression(-pi, -pi / 2, 0, pi / 2, pi))
text(1e3, 0, expression(italic("Customized Axes")))
box()
```

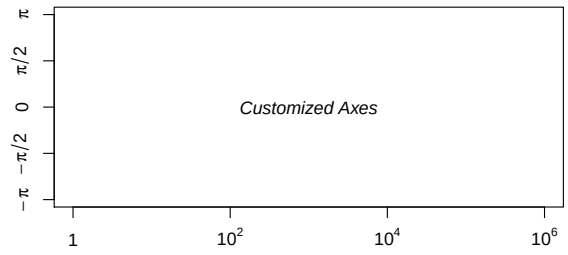


图 1.6: Creating Customized Axes With Suitable Annotation

1.14　旋转坐标抽标签

旋转坐标抽标签的例子来自手册《R FAQ》的第 7 章第 27 个问题 (Hornik, 2017)，在基础图形中，旋转坐标轴标签需要 `text()` 而不是 `mtext()`，因为后者不支持 `par("srt")`

```
## Increase bottom margin to make room for rotated labels
par(mar = c(5, 4, .5, 2) + 0.1)
## Create plot with no x axis and no x axis label
plot(1 : 8, xaxt = "n",  xlab = "")
```

```
## Set up x axis with tick marks alone
axis(1, labels = FALSE)
## Create some text labels
labels <- paste("Label", 1:8, sep = " ")
## Plot x axis labels at default tick marks
text(1:8, par("usr")[3] - 0.5, srt = 45, adj = 1,
     labels = labels, xpd = TRUE)
## Plot x axis label at line 6 (of 7)
mtext(side = 1, text = "X Axis Label", line = 4)
```

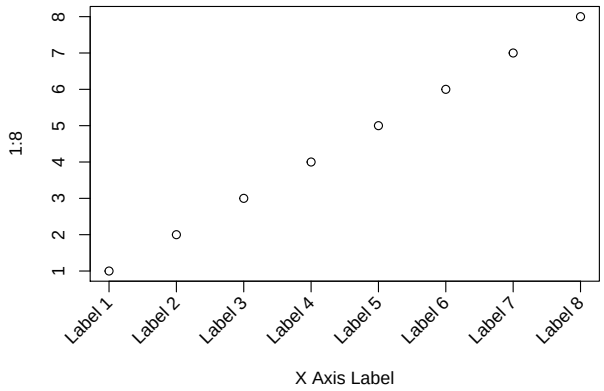


图 1.7: 旋转坐标轴标签

`srt = 45` 表示文本旋转角度，`xpd = TRUE` 允许文本越出绘图区域，`adj = 1` to place the right end of text at the tick marks；You can adjust the value of the 0.5 offset as required to move the axis labels up or down relative to the x axis.

1.15　双纵轴

How to plot multiple time series plots in a grid, where each plot has two y axes? https://stackoverflow.com/questions/52082801

```
library(ggplot2)
dd <- data.frame(
  x = 1:11, y = c(rnorm(22),rpois(22,5)),
  id = gl(2, 11, labels = paste0("ser", 1:2)),
  panel = gl(2, 22, labels = c("norm","pois"))
```

```
)
# 设置左手刻度为 norm 右手刻度为 pois
ggplot(dd, aes(x, y, col = panel)) +
  geom_line() +
  facet_wrap(~id) +
  scale_y_continuous(sec.axis = sec_axis(~.))
```

图 1.8: 双纵轴时间序列

第二章　高级图形

2010 年 Hadley 基于图形语法开发了 ggplot 包，随后一直保持维护和开发，在 2015 年进行重构，推出 ggplot2，经过几年的发展，现已进入软件开发的稳定阶段。

图 2.1: Hadley Wickham 的帅照

介绍 grid 绘图系统，特别是 ggplot2 包

图形语法，图层，点线，映射，统计

2.1　ggplot2

图形语法

2.1.1　生态

```
grep('^(gg)', .packages(T), value = TRUE)
#> [1] "ggplot2"
```

2.2　动态图形

一类由静态图形合成为 gif, avi, svg 等，一类面向特定输出设备的动态图形，基于 JavaScript 库的交互图形，基于 OpenGL/ggobi/asymptote 的真三维图形

2.2.1　gganimate

与此类似的还有 mapmate

2.2.2　ggobi

2.2.3　OpenGL

2.2.4　JavaScript

第三章　出版级图形

要达到出版级的专业图形，我们需要从配色、字体等方面再下功夫

3.1　配色

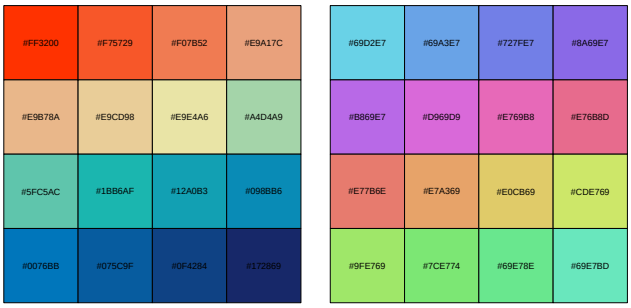
配色是快速提升图片美感的捷径，主要介绍一下基础调色板，**colorspace**，**RcolorBrewer** 和 **viridis** 三个用于配色的 R 包。除了上述要求外，还应在字体和图形质量上考虑

调制两个色板

```
n <- 16
# Colors from https://github.com/johannesbjork/LaCroixColoR
color_pal <- c("#FF3200", "#E9A17C", "#E9E4A6", "#1BB6AF", "#0076BB", "#172869")
color_pals_1 <- (grDevices::colorRampPalette(color_pal))(n)
scales::show_col(colours = color_pals_1)
# colors in colortools from http://www.gastonsanchez.com/
fish_pal <- c("#69D2E7", "#6993E7", "#7E69E7", "#BD69E7",
              "#E769D2", "#E76993", "#E77E69", "#E7BD69",
              "#D2E769", "#93E769", "#69E77E", "#69E7BD")
color_pals_2 <- (grDevices::colorRampPalette(fish_pal))(n)
scales::show_col(colours = color_pals_2)
```

3.2　字体

xkcd 等字体

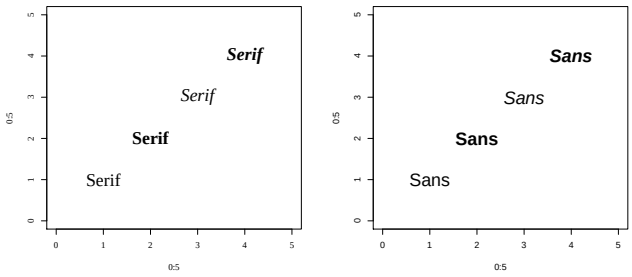


(a) 渐变　(b) Hue-Saturation-Value (HSV) color model

图 3.1: 调色板

```
font_add_google("Alegreya Sans", "aleg")

font_add_google("Source Code Pro", "sourcecodepro")

font_add_google("Source Sans Pro", "sourcesanspro")

font_add_google("Source Serif Pro", "sourceserifpro")

font_add_google("Roboto", "roboto")
# Ubuntu Ubuntu Mono Ubuntu Condensed
```

```
# library(showtext,quietly = TRUE)
# showtext.auto()
# pdf("google-fonts.pdf")
# font_add_google("Alegreya Sans", "aleg")
par(family = "serif")
plot(0:5,0:5, type="n")
text(1:4, 1:4, "Serif", font=1:4, cex = 2)
par(family = "sans")
plot(0:5,0:5, type="n")
text(1:4, 1:4, "Sans", font=1:4, cex = 2)
# dev.off()
```



(a) Alegreya Sans 字体　(b) Sans 字体

图 3.2: 两种字体

3.3　保存

图形设备，抗锯齿，颜色模式 LAB，RGB，CMYK 适应展示设备需要

```
embedFonts(file = "google-fonts-ex.pdf",outfile = "google-fonts-ex-embed.pdf")
```

另外一个方法就是使用 `cairo_pdf` 而不是 `pdf`

参考文献

Hornik, K. (2017). R FAQ.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Murrell, P. and Ihaka, R. (2000). An approach to providing mathematical annotation in plots. *Journal of Computational and Graphical Statistics*, 9(3):582–599.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2016). *bookdown: Authoring Books and Technical Documents with R Markdown*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138700109.

Xie, Y., Allaire, J., and Grolemund, G. (2018). *R Markdown: The Definitive Guide*. Chapman and Hall/CRC, Boca Raton, Florida. ISBN 978-1138359338.

索引