R语言在经济学中的应用

南开大学周恩来政府管理学院 吕小康

2017-07-13

R简介

R是一个免费自由且跨平台通用的统计计算与绘图软件。

- 它有Windows、Mac、Linux等版本，均可免费下载使用。

从R主页中选择download R链接可下载到对应操作系统的R安装程序。

- 打开链接后的网页会提示选择相应的CRAN镜像站。目前全球有超过一百个CRAN镜像站，用户可选择就近下载。

R与STATA等统计软件的区别

R为开源免费的软件，其他基本为商业付费软件。

- 如果你有钱，可以只选贵的、不选对的；但如果你没钱……

R是一种脚本语言，强调英文命令操作。

- R的学习比较费时、对汉字编码不友好，但掌握之后的自由性更强

R在数据可视化上的表现更佳，选择更丰富。

- R的统计绘图是它最有标志性的功能，可以制作达到出版的各种图形

R在经济学中的综合应用

R及与之相关的配套开源软件（如RStudio）已构成一个丰富的数据分析网络生态，具有同类软件很难同时满足的多种可能性。

用于课程教学

用于数据获取与预处理

用于数据的计量分析

用于数据可视化

用于撰写学术报告

R 已可方便导入各类型的数据

利用Hadley等人开发的诸多R包，已可方便导入各种类型的数据，这为R成为一种“兼容并蓄”的统计分析软件奠定了重要基础。

readxl包: xls or xlsx

```
library(readxl)
read_excel("file.xlsx")
```

rvest包：在线文本

data.table包：导入大数据文件（> 100 G）

R 已可方便导入各类型的数据

Haven包： SAS, SPSS, STATA

```
# SAS
read_sas("mtcars.sas7bdat")
write_sas(mtcars, "mtcars.sas7bdat")

# SPSS
read_sav("mtcars.sav")
write_sav(mtcars, "mtcars.sav")

# Stata
read_dta("mtcars.dta")
write_dta(mtcars, "mtcars.dta")
```

利用R，几乎可以分析类型任何类型的数据，而避免在各类统计软件之间相互转化和跟踪。

作为课堂教学的辅助软件

可以作为两门经济学基础课程的教学辅助软件

- 《概率论与数理统计》
- 《计量经济学》

我本人在清华大学出版社2017年出版的《R语言统计学基础》，内容差不多覆盖经济学类入门概率论与数理统计的教学要求，全程使用R作为分析和绘图软件。

概率论与数理统计的课程教学

这里以抽样分布的教学设计为例进行说明。

抽样分布（sampling distribution）的基础知识

标准定义

- 抽样分布是样本统计量的分布。这显然精确而“无用”，即不能帮助人真正理解什么是抽样分布。
- 实质是重复抽样的假想前提下形成的一个统计推论框架，它在现实中是不一定存在的。

抽样分布的具体解释

- 抽样分布是对同一总体，做出相同样本容量的、重复(无限) 多次的简单随机抽样取遍样本统计量的所有可能值后所体现出来的取值规律性。
- 对这一规律性，采用概率论的基本知识加以描述，即可概括为某一**分布**（distribution），也即 $F(x) = P(X \leq x)$
- 如果能够找到 $F(x)$ 的精确数学表达形式，后续的统计推论即可基于这一概念框架而得到概率意义上的精准推导。

抽样分布的教学难点

理论框架的“非现实性”

- 现实中的研究通常只能有一次抽样，不可能对同一总体进行反复抽样从而得到关于样本统计量的精确分布的直观感知

数据并非总是随机抽样获得的

- 通过随机化实验，以及通过普通的观测收集的数据，也需要进行推论统计。但此时很难直接套用基于“重复抽样”获得的抽样分布理论。

某些统计量的抽样分布数学推导较为困难

- 况且，能够找到精确数学形式的分布总是少见的。很多统计量本身就是很难找到精确分布，然而推论总是要做下去……

建立经验感知的方式：模拟

抽样分布的建立需要一对互相联系的概念：总体（population）与样本（sample）。

不妨以这样的思路进行教学:

建立经验感知的方式：模拟

抽样分布的建立需要一对互相联系的概念：总体（population）与样本（sample）。

不妨以这样的思路进行教学:

- 先从假想的理论分布（如正态分布、二项分布、指数分布等）总体进行重复抽样，模拟某一简单样本统计量的分布，并与数学推导的结果进行对比解释；

建立经验感知的方式：模拟

抽样分布的建立需要一对互相联系的概念：总体（population）与样本（sample）。

不妨以这样的思路进行教学:

- 先从假想的理论分布（如正态分布、二项分布、指数分布等）总体进行重复抽样，模拟某一简单样本统计量的分布，并与数学推导的结果进行对比解释；
- 到假想的理论分布总体进行重复抽样，模拟某一很难或无法从数学推导获得精确分布的样本统计量的抽样分布；

建立经验感知的方式：模拟

抽样分布的建立需要一对互相联系的概念：总体（population）与样本（sample）。

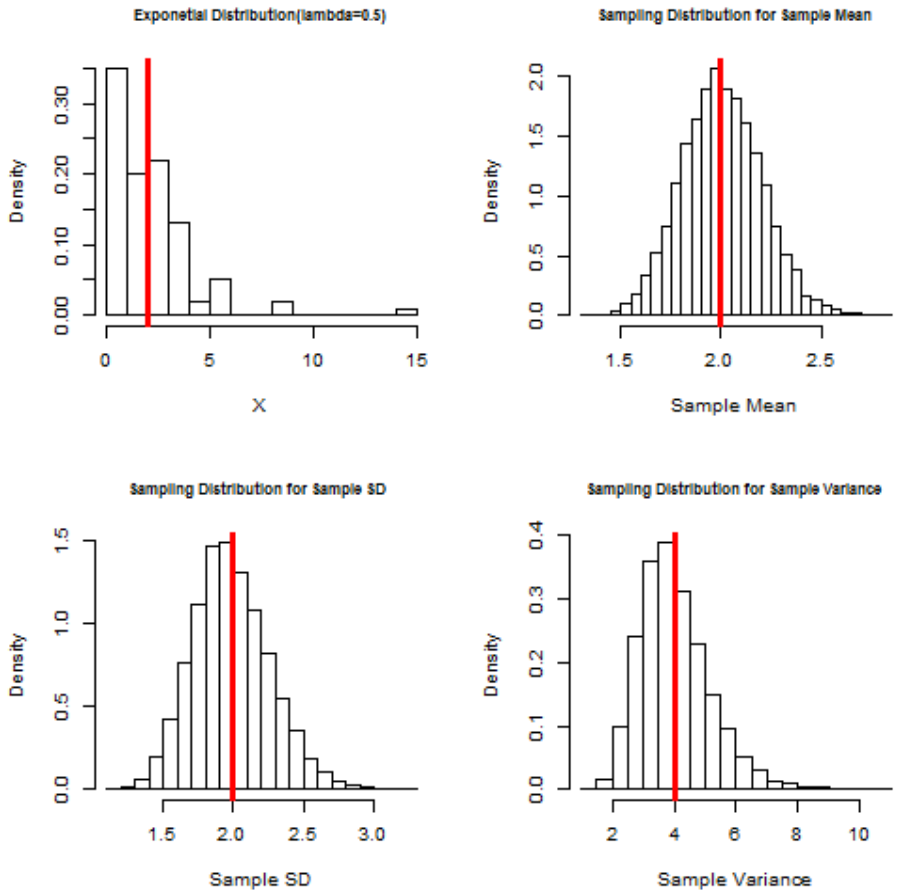
不妨以这样的思路进行教学:

- 先从假想的理论分布（如正态分布、二项分布、指数分布等）总体进行重复抽样，模拟某一简单样本统计量的分布，并与数学推导的结果进行对比解释；
- 到假想的理论分布总体进行重复抽样，模拟某一很难或无法从数学推导获得精确分布的样本统计量的抽样分布；
- 再到从实际的、不满足特定分布的总体进行重复抽样，模拟对应的样本统计量的分布，验证数学推导的结果是否能够应用于现实，并理解数学推导的局限性与模拟的自由性

样本均值等常见统计量的抽样分布示例

已知总体 $X \sim Exp(0.5)$，即服从某指数分布，注意该分布本身是右偏的，且是一无限总体。现从中抽取样本量为100的样本，重复10000次，每次计算如下样本统计量，再绘制这些样本统计量的的直方图，即可在一定程度上展示该些样本统计量的抽样分布形状，以便形成直观的感知。

- 样本均值（sample mean）
- 样本标准差（samlpe standard deviation）
- 样本方差（sample variance）
- 其他需要的样本统计量



左上图为给定总体的概率密度图，右上图为样本均值的抽样分布示意图；左下图为样本标准差的抽样分布示意图，右下图为样本方差的抽样分布示意图。红色虚线为虚线表示总体均值、 总体标准差和总体方差所在位置。

生成样本均值抽样分布的代码

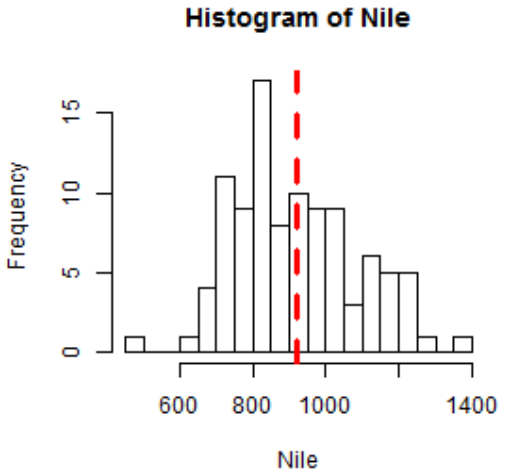
```r
sample_mean <- numeric(10000)
for (i in 1:10000) {
  set.seed(i)
  samples <- rexp(100, 0.5)
  sample_mean[i] <- mean(samples)
}
hist(
  sample_mean,
  prob = T,
  main = "Sampling Distribution for Sample Mean",
  xlab = "Sample Mean",
  breaks = 30,
  cex.main = 0.8
)
abline(v = 2, col = "red", lwd = 3)
```

其余可交由学生思考复制。

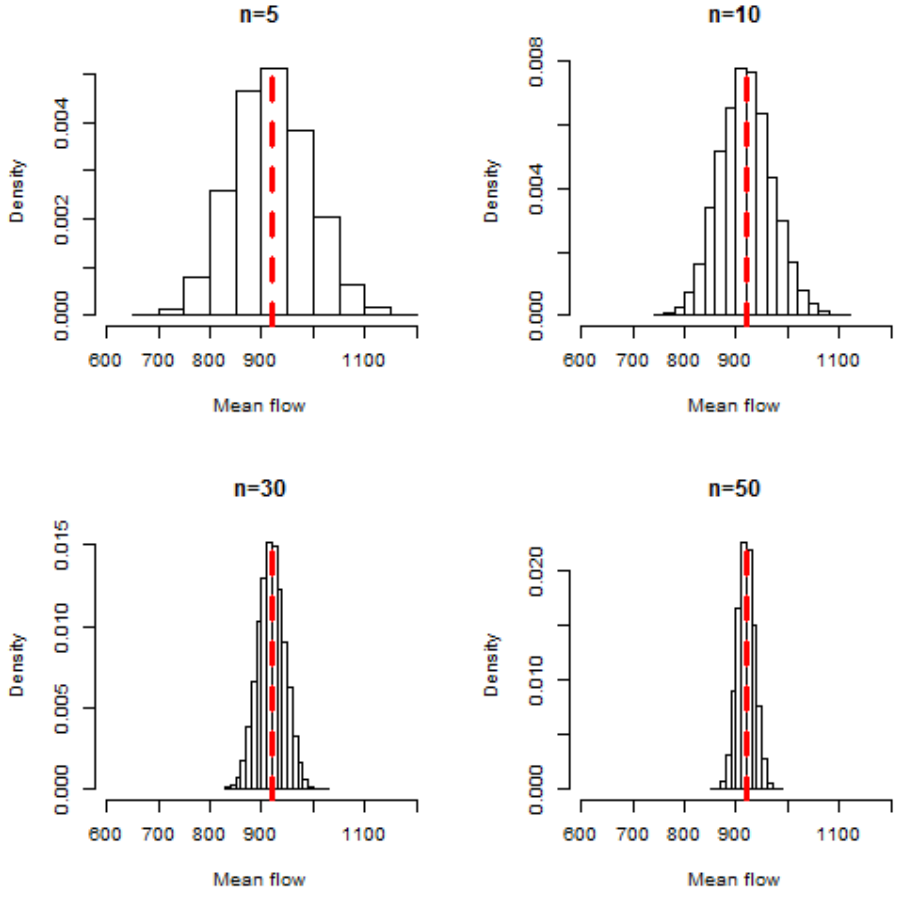
真实的观测数据：Nile 流量

Nile数据是R自带的数据，记录了尼罗河在埃及阿斯量(Ashwan) 地 区1871-1970 年这100年间的年流量值。这一数据服从什么特定的精确分布吗？--不清楚。

```
hist(Nile, breaks = 30)
abline(v = mean(Nile), col = "red", lty = 2, lwd = 3)
```



但若以它为“总体”，再从中进行简单随机抽样，然后观察某些特定统计量（如样本均值、样本方差、样本中位数）的分布，仍可获得经验感知。



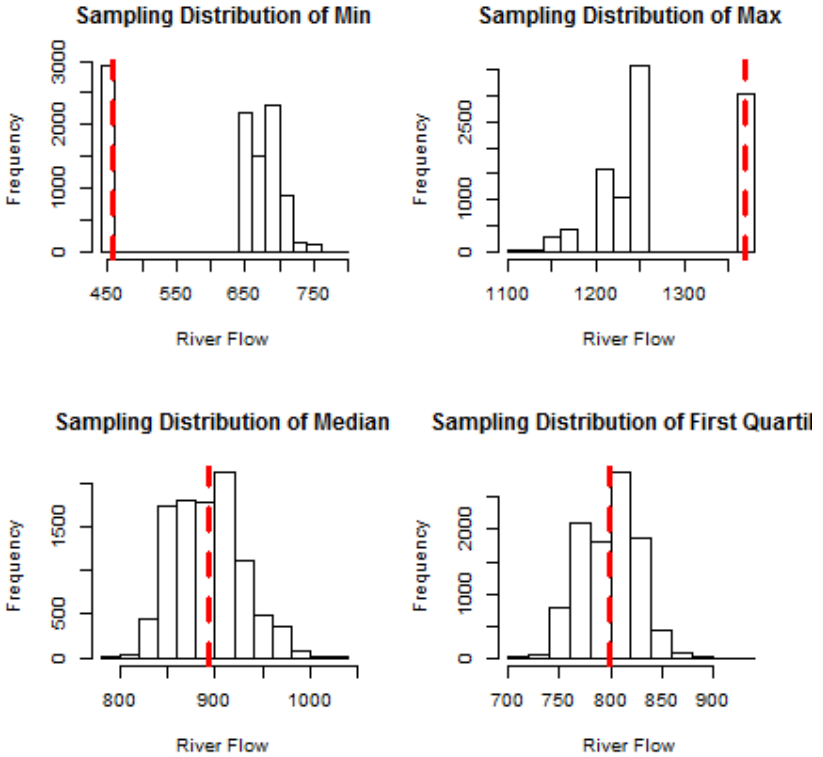
红色虚线表示“总体”均值所在的位置。这是不是就是中心极限定理（Central Limit Theorem）告诉我们的道理呢？

部分代码示例

```
x <-
  data.frame(
  a = numeric(10000),
  b = numeric(10000),
  c = numeric(10000),
  d = numeric(10000)
  )
  par(mfrow = c(2, 2))
  for (i in 1:10000)
  {
  set.seed(i)
  samples <- sample(Nile, 5)
  x$a[i] <- mean(samples)
  }

  hist(
  x$a,
  xlab = "Mean flow",
  main = "n=5",
  probability = T,
  xlim = c(600, 1200)
  )
  abline(
  v = mean(Nile),
  lwd = 3,
  lty = 2,
  col = "red"
  )
```

Nile数据，其他样本统计量的抽样分布



样本最小值、样本最大值、样本中位数、样本第一四分位数的抽样分布示意图（样本容量为30）。红色虚线分别表示“总体”最小值、最大值、中位数和第一四分位数所在的位置。

关于代码教学的建议

教师只需在课堂讲解说明并提供一个图形的原始代码，即可要求学生仿照此代码，自行绘制其他图形，并作为作业进行布置。如此可加深学生对抽样分布形成过程的直观认识。

关于代码教学的建议

教师只需在课堂讲解说明并提供一个图形的原始代码，即可要求学生仿照此代码，自行绘制其他图形，并作为作业进行布置。如此可加深学生对抽样分布形成过程的直观认识。

实际上关于抽样分布还有一些基于`RStudio`的`shiny`平台搭建的动态呈现模式，例如Nicole Radziwill就制作了相关的简单结果展示。这可以作为课堂教学的参考。

如果教师本人精力允许，可以带领学生自行制作相关网页。这样收获更大。

关于代码教学的建议

教师只需在课堂讲解说明并提供一个图形的原始代码，即可要求学生仿照此代码，自行绘制其他图形，并作为作业进行布置。如此可加深学生对抽样分布形成过程的直观认识。

实际上关于抽样分布还有一些基于`RStudio`的`shiny`平台搭建的动态呈现模式，例如Nicole Radziwill就制作了相关的简单结果展示。这可以作为课堂教学的参考。

如果教师本人精力允许，可以带领学生自行制作相关网页。这样收获更大。

实际上，有了重复抽样情形下的抽样分布及模拟，教师还可利用`R`进行自助分布（Bootstrap Distribution）、随机化分布（Randomization Distribution）等理论分布的模拟，如此可将推论的背景框架推广至其他观测数据或实验数据的情形。

计量经济学的课程教学

这里使用一个经常在计量经济学中使用到的数据（Affairs）进行示例。这是美国 *Psychology Today* 杂志于1969年采集的关于婚外情的数据。该数据经常用于广义线性模型的示例。

```
if(!require(AER)) install.packages("AER")
data("Affairs")
head(Affairs)
```

```
##    affairs gender age yearsmarried children religiousness education
## 4        0   male  37        10.00       no             3        18
## 5        0 female  27         4.00       no             4        14
## 11       0 female  32        15.00      yes             1        12
## 16       0   male  57        15.00      yes             5        18
## 23       0   male  22         0.75       no             2        17
## 29       0 female  32         1.50       no             2        17
##    occupation rating
## 4           7      4
## 5           6      4
## 11          1      4
## 16          6      5
## 23          6      3
## 29          5      5
```

OLS回归

```
fm_ols <- lm(affairs ~ age + yearsmarried + religiousness + occupati
data = Affairs)
summary(fm_ols)
```

```
##
## Call:
## lm(formula = affairs ~ age + yearsmarried + religiousness + occupation +
##     rating, data = Affairs)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.0382 -1.7076 -0.7780  0.2086 12.8134
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.60816    0.79660   7.040 5.31e-12 ***
## age           -0.05035    0.02211  -2.278   0.0231 *
## yearsmarried   0.16185    0.03690   4.387 1.36e-05 ***
## religiousness -0.47632    0.11131  -4.279 2.18e-05 ***
## occupation     0.10601    0.07110   1.491   0.1365
## rating        -0.71224    0.11829  -6.021 3.03e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.087 on 595 degrees of freedom
## Multiple R-squared:  0.1314,    Adjusted R-squared:  0.1241
## F-statistic:    18 on 5 and 595 DF,  p-value: < 2.2e-16
```

OLS 回归

查看模型拟合值

```
fit <- fitted(fm_ols)
head(fit)
```

```
##           4           5          11          16          23          29
##  1.8279300  0.7779612  3.2055393 -0.1406526  2.1685688  0.2559938
```

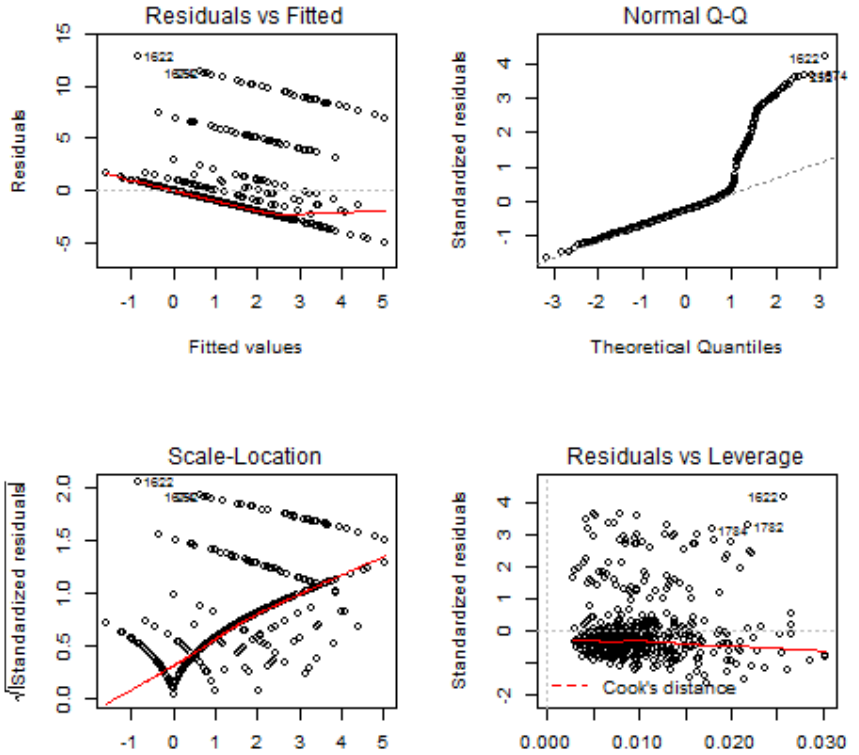
查看模型残差

```
re <- residuals(fm_ols)
head(re)
```

```
##           4           5          11          16          23          29
## -1.8279300 -0.7779612 -3.2055393  0.1406526 -2.1685688 -0.2559938
```

查看用于模型诊断的相关图示

```
opar <- par(no.readonly = T)
par(mfrow = c(2, 2))
plot(fm_ols)
```



关于回归假设诊断的一个常见范例

线性模型最大的优点可能在于数学形式上的简单性。但这仍依赖于许多基本假定（assumptions）。对这些假定进行检验，是经济学教学和实证研究环节中不可忽视的内容。

通过一些“极端”的教学示例，可培养学生检验模型基本假定的良好习惯。

Frank Anscombe(1918-2001) 是20世纪著名的英国统计学家， 他于1973年发表了一篇具有深远影响的文章，讨论图形在统计检验中的作用。其所构造的一组数据经常被作为演示数据。

Anscombe 四重奏(Anscombe's Quartet)

请观察以下数据。

```
anscombe
```

```
##    x1 x2 x3 x4    y1   y2    y3    y4
## 1  10 10 10  8  8.04 9.14  7.46  6.58
## 2   8  8  8  8  6.95 8.14  6.77  5.76
## 3  13 13 13  8  7.58 8.74 12.74  7.71
## 4   9  9  9  8  8.81 8.77  7.11  8.84
## 5  11 11 11  8  8.33 9.26  7.81  8.47
## 6  14 14 14  8  9.96 8.10  8.84  7.04
## 7   6  6  6  8  7.24 6.13  6.08  5.25
## 8   4  4  4 19  4.26 3.10  5.39 12.50
## 9  12 12 12  8 10.84 9.13  8.15  5.56
## 10  7  7  7  8  4.82 7.26  6.42  7.91
## 11  5  5  5  8  5.68 4.74  5.73  6.89
```

x1, x2, x3这三列完全相同，x4 与前三例不同；y1, y2, y3, y4各不相同。

Anscombe的“四个”回归方程

```
fit1 <- lm(y1 ~ x1, data = anscombe)
coefficients(fit1)
```

```
## (Intercept)          x1
##   3.0000909   0.5000909
```

```
fit2 <- lm(y2 ~ x2, data = anscombe)
coefficients(fit2)
```

```
## (Intercept)          x2
##    3.000909    0.500000
```

```
fit3 <- lm(y3 ~ x3, data = anscombe)
coefficients(fit3)
```

```
## (Intercept)          x3
##   3.0024545   0.4997273
```

```
fit4 <- lm(y4 ~ x4, data = anscombe)
coefficients(fit4)
```

```
## (Intercept)          x4
##   3.0017273   0.4999091
```

Anscombe的“四个”线性相关系数

```
attach(anscombe)
cor(x1, y1)
```

```
## [1] 0.8164205
```

```
cor(x2, y2)
```

```
## [1] 0.8162365
```

```
cor(x3, y3)
```

```
## [1] 0.8162867
```

```
cor(x4, y4)
```

```
## [1] 0.8165214
```

```
detach(anscombe)
```

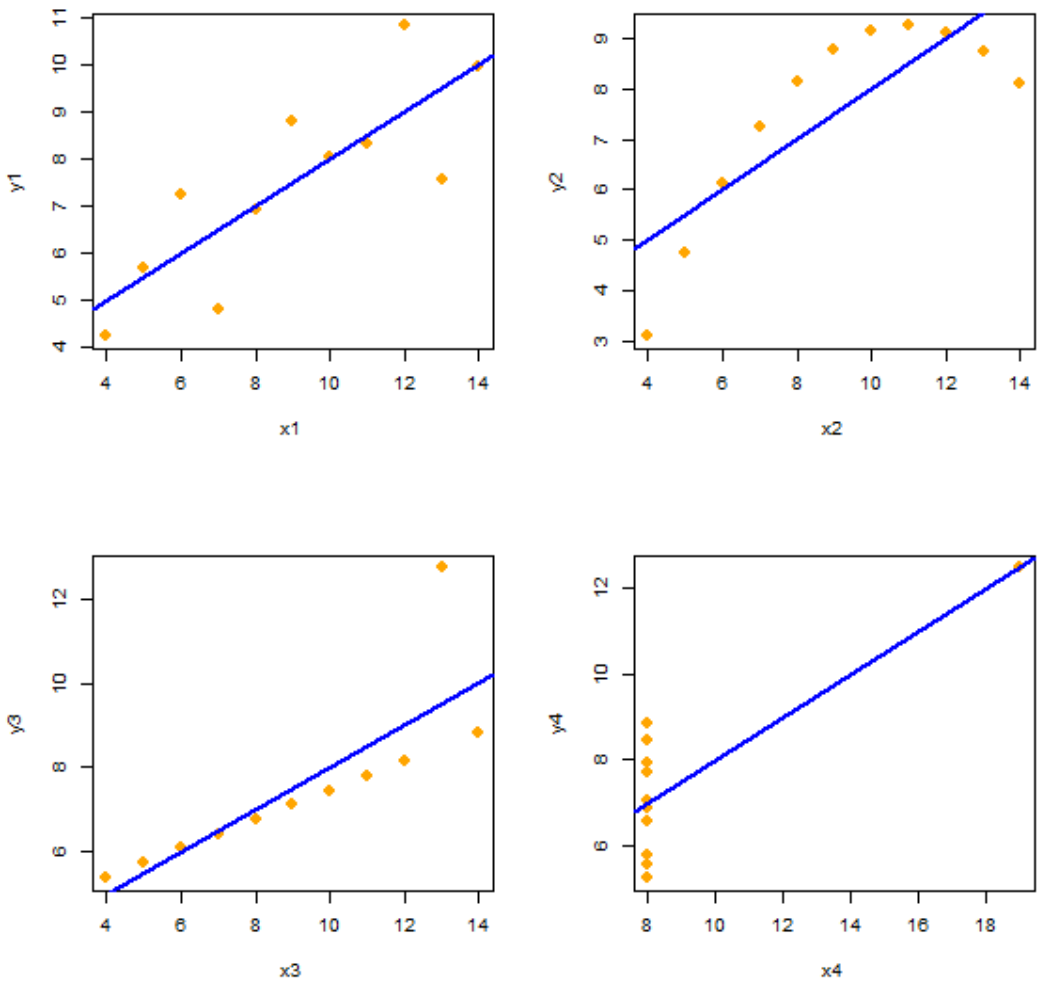
相同的回归系数，相同的相关系数

从近似角度看，这几乎可以统一为一个回归方程：

$$\hat{y} = 0.5x + 3$$

从近似的角度看，它们还拥有相同的相关系数：0.816。

但是，如果绘制各自的散点图……


y1
x1
y2
x2
y3
x3
y4
x4

结论：统计数字会骗人！

结论：统计数字会骗人！

不能迷信回归系数

结论：统计数字会骗人！

不能迷信回归系数

结合可视化进行模型检验有其优势

结论：统计数字会骗人！

不能迷信回归系数

结合可视化进行模型检验有其优势

结合R来做可视化较为便利

相关代码

```
attach(anscombe)
opar <- par(no.readonly = T)
par(mfrow = c(2, 2))
plot(x1, y1, col = "orange", pch = 20, cex = 2)
abline(lm(y1~x1), col = "blue", lwd = 2)
plot(x2, y2, col = "orange", pch = 20, cex = 2)
abline(lm(y2~x2), col = "blue", lwd = 2)
plot(x3, y3, col = "orange", pch = 20, cex = 2)
abline(lm(y3~x3), col = "blue", lwd = 2)
plot(x4, y4, col = "orange", pch = 20, cex = 2)
abline(lm(y4~x4), col = "blue", lwd = 2)
par(opar)
detach(anscombe)
```

广义线性模型

广义线性模型（Generalized Linear Models）的一般形式：

$$f(\mu_Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$$

其中

- $f(\mu_Y)$表示响应变量的条件均值的某种函数（称为连接函数，link function）。
- 此时对 Y 不再有服从正态分布的要求，而可以服从任何指数分布族中的某一分布。
- 设定好连接函数与分布类型后，就可以利用极大似然法通过多次迭代推导出各参数值。

常用的广义线性模型

- Probit/Logistic 回归模型
- Poisson 回归模型
- Negative Binomial 回归模型
- Zero Inflation 回归模型
- Tobit 回归模型
- ……

这些都可通过R的相关函数方便求得。

广义线性模型

Probit 回归

```
fm_probit <- glm(I(affairs > 0) ~ age + yearsmarried + religiousness
data = Affairs, family = binomial(link = "probit"))
summary(fm_probit)
```

```
##
## Call:
## glm(formula = I(affairs > 0) ~ age + yearsmarried + religiousness +
##     occupation + rating, family = binomial(link = "probit"),
##     data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6143  -0.7678  -0.5841  -0.2368   2.4615
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     0.97667    0.36537   2.673 0.007516 **
## age            -0.02202    0.01032  -2.134 0.032821 *
## yearsmarried    0.05990    0.01712   3.499 0.000468 ***
## religiousness -0.18365    0.05172  -3.551 0.000384 ***
## occupation      0.03751    0.03285   1.142 0.253399
```

Probit 回归

查看模型拟合值

```
fit <- fitted(fm_probit)
head(fit)
```

```
##           4           5          11          16          23          29
## 0.26779939 0.16359028 0.47303362 0.07518241 0.33574947 0.11842761
```

Logistic/Logit 回归

```
fm_logit <- glm(I(affairs > 0) ~ age + yearsmarried + religiousness +
data = Affairs, family = binomial(link = "logit"))
summary(fm_logit)
```

```
##
## Call:
## glm(formula = I(affairs > 0) ~ age + yearsmarried + religiousness +
##     occupation + rating, family = binomial(link = "logit"), data = Affairs
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6633  -0.7500  -0.5750  -0.2691   2.4189
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.74904    0.62802   2.785 0.005352 **
## age            -0.04009    0.01785  -2.245 0.024748 *
## yearsmarried    0.10543    0.02952   3.572 0.000354 ***
## religiousness -0.32332    0.08950  -3.613 0.000303 ***
## occupation      0.07250    0.05677   1.277 0.201565
## rating         -0.46842    0.08928  -5.247 1.55e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 675.38  on 600  degrees of freedom
```

Poisson 回归

```
fm_pois <- glm(affairs ~ age + yearsmarried + religiousness + occupa
data = Affairs, family = poisson)
summary(fm_pois)
```

```
##
## Call:
## glm(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating, family = poisson, data = Affairs)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.5968  -1.5728  -1.1627  -0.7067   8.3473
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.533905   0.196924  12.867  < 2e-16 ***
## age           -0.032255   0.005851  -5.512 3.54e-08 ***
## yearsmarried   0.115698   0.009908  11.677  < 2e-16 ***
## religiousness -0.354037   0.030892 -11.460  < 2e-16 ***
## occupation     0.079828   0.019449   4.105 4.05e-05 ***
## rating        -0.409443   0.027381 -14.953  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 2925.5  on 600  degrees of freedom
```

Negative Binomial 回归

```
if(!require(MASS)) install.packages("MASS")
fm_nb <- glm.nb(affairs ~ age + yearsmarried + religiousness + occup
data = Affairs)
summary(fm_nb)
```

```
##
## Call:
## glm.nb(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating, data = Affairs, init.theta = 0.142555597,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.1139  -0.8067  -0.6943  -0.4533   2.4548
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.189666   0.727326   3.011  0.00261 **
## age           -0.002624   0.020312  -0.129  0.89722
## yearsmarried   0.084819   0.034205   2.480  0.01315 *
## religiousness -0.422227   0.104081  -4.057 4.98e-05 ***
## occupation     0.060443   0.066262   0.912  0.36167
## rating        -0.431331   0.107449  -4.014 5.96e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.1426) family taken to be 1)
```

Zero Inflation 回归

```
if(!require(pscl)) install.packages("pscl")
fm_zero <- zeroinfl(affairs ~ age + yearsmarried + religiousness + o
yearsmarried + religiousness + occupation + rating, data = Affairs)
summary(fm_zero)
```

```
##
## Call:
## zeroinfl(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating | age + yearsmarried + religiousness + occupation
##     rating, data = Affairs)
##
## Pearson residuals:
##     Min      1Q  Median      3Q     Max
## -1.4643 -0.5190 -0.3827 -0.2444 14.3993
##
## Count model coefficients (poisson with log link):
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)    2.201940   0.210148  10.478  < 2e-16 ***
## age           -0.007238   0.006343  -1.141    0.254
## yearsmarried   0.049224   0.010990   4.479 7.50e-06 ***
## religiousness -0.131668   0.031154  -4.226 2.37e-05 ***
## occupation     0.016029   0.020101   0.797    0.425
## rating        -0.118672   0.028693  -4.136 3.53e-05 ***
##
## Zero-inflation model coefficients (binomial with logit link):
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -1.74075    0.62879  -2.768 0.005633 **
```

Tobit 回归

```
library(AER)
fm_tobit <- tobit(affairs ~ age + yearsmarried + religiousness + occ
data = Affairs)
summary(fm_tobit)
```

```
##
## Call:
## tobit(formula = affairs ~ age + yearsmarried + religiousness +
##     occupation + rating, data = Affairs)
##
## Observations:
##          Total  Left-censored     Uncensored Right-censored
##            601            451            150              0
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    8.17420    2.74145   2.982  0.00287 **
## age           -0.17933    0.07909  -2.267  0.02337 *
## yearsmarried   0.55414    0.13452   4.119 3.80e-05 ***
## religiousness -1.68622    0.40375  -4.176 2.96e-05 ***
## occupation     0.32605    0.25442   1.282  0.20001
## rating        -2.28497    0.40783  -5.603 2.11e-08 ***
## Log(scale)     2.10986    0.06710  31.444  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 8.247
```

在数据获取、预处理和可视化中的应用

- tidyverse系列数据处理包
  - dplyr: 数据操纵
  - tidyr: 数据操纵
  - stringr: 文本数据操纵
  - rvest: 在线抓取文本
  - ……
- 可视化系列数据处理包
  - ggplot2
  - ggtheme
  - ggvis
  - shiny
  - wordcloud2
  - ……

数据处理示例1：一手问卷调查数据

我们项目组目前正在编制《中国医患社会心态调查问卷》，问卷已经基本完成编制并已进行预测试。对初测数据的统计分析工作正在进行。初测问卷使用问卷星填答，要求被调查者使用自身手机或在访问员的手机上完成填答。数据示例见Excel文件。

以下命令可简单地统计被试的地理位置分布。

```
library(readxl)
library(stringr)
library(tidyverse)
PDSurveyBasic <- read_excel("PDSurveyBasic.xlsx")
ip.location <- str_extract(PDSurveyBasic$ip, "(?<=\\().*(?=\\))") %>%
  str_split("-", n = 2, simplify = TRUE) %>%
  as_tibble %>%
  transmute(province = .[[1]], city = .[[2]]) %>%
  group_by(province) %>%
  summarise(n=n()) %>%
  arrange(desc(n))
```

地理位置信息分布结果

```
## # A tibble: 27 x 2
##    province     n
##       <chr> <int>
##  1     天津    97
##  2     云南    71
##  3     辽宁    63
##  4     新疆    57
##  5     河南    47
##  6     山东    44
##  7     北京    36
##  8     四川    34
##  9     重庆    34
## 10     山西    29
## # ... with 17 more rows
```

数据获取与处理示例2

政府工作报告抓取与分析

传统社会科学的量化分析以对数字数据（numeric data）的量化分析为主，对文本数据（text data）的分析较少。这主要是受研究工具的局限所致。

R及Python等开源软件的出现，很大程度上改变这种现状，使得文本分析成为当下社会科学研究的一大潮流。

中国政府网提供了自1954年以来所有的政府工作报告全文。这里以中国政府工作报告（2017）为例做一简单的R语言示例（该示例得益于雪晴数据网陈堰平老师的讲座）。

政府工作报告的抓取与简单分析

2017政府工作报告

```
if (!require(rvest)) install.packages('rvest')
if (!require(wordcloud2)) install.packages('wordcloud2')
if (!require(jiebaR)) install.packages('jiebaR')
if (!require(stringr)) install.packages('stringr')
url2017 <-
  "http://www.gov.cn/premier/2017-03/16/content_5177940.htm"
report2017 <- read_html(url2017)
text2017 <- report2017 %>%
  html_nodes("p") %>%
  html_text() %>%
  paste(collapse = "")
writeLines(text2017, "report2017.txt")
library(jiebaR)
cutter <- worker(
  bylines = T,
  user = "./UsrWords.txt",
  stop_word = "./stopWords.txt",
  output = "report2017output.txt"
)
report_seg_file <- cutter["./report2017.txt"]
report_segged <-
  readLines("./report2017output.txt", encoding = "UTF-8")
report <- as.list(report_segged)
doc.list <- strsplit(as.character(report), split = " ")
term.table <- table(unlist(doc.list))
term.table <- sort(term.table, decreasing = TRUE)
```

政府工作报告的抓取与简单分析

```
head(vocabDF, 10)
```

```
##    Var1 Freq
## 1  发展 6125
## 2  改革 3332
## 3  推进 3185
## 4  建设 2646
## 5  经济 2548
## 6  推动 2058
## 7  加快 1960
## 8  政府 1960
## 9  创新 1764
## 10 企业 1715
```

政府工作报告的抓取与简单分析

```
library(wordcloud2)
wordcloud2(vocabDF, color = "random-light", backgroundColor = "grey")
```

如何通过循环来遍历所有年份政府工作报告的链接，留待大家作为思考题。

提示如下：

```
url <- "http://www.gov.cn/guowuyuan/baogao.htm"
reports <- read_html(url)
links <- reports %>%
  html_nodes(".history_report a") %>%
  html_attr("href") %>%
  str_trim()
head(links)
```

```
## [1] "http://www.gov.cn/premier/2017-03/16/content_5177940.htm"
## [2] "http://www.gov.cn/premier/2016-03/17/content_5054901.htm"
## [3] "http://www.gov.cn/guowuyuan/2015-03/16/content_2835101.htm"
## [4] "http://www.gov.cn/guowuyuan/2014-03/14/content_2638989.htm"
## [5] "http://www.gov.cn/premier/2013-03/19/content_2357136.htm"
## [6] "http://www.gov.cn/premier/2012-03/15/content_2067314.htm"
```

数据获取、处理与可视化

经济学研究的常用数据、世界银行数据可使用两个R包获取：

- WDI
- wbstats

一个复制Hans Rosling的Gapminder软件的动态交互式气泡图

- Hans Rosling的TED演讲，中文翻译版
- R中的复制

ggplot系列图形

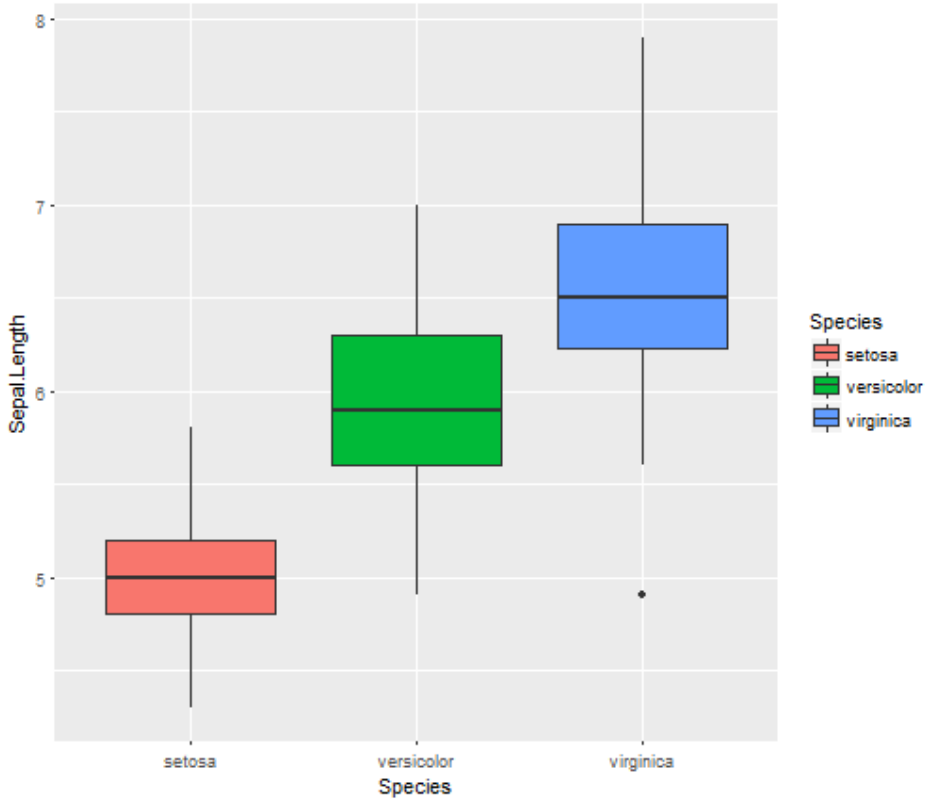
利用ggplot2及ggthemes、ggsci等包，可便捷产生符合特定杂志风格的图形。

常用ggplot系列可视化包

- ggplot2
- ggthemes
- ggsci
- ggcorrplot
- ……

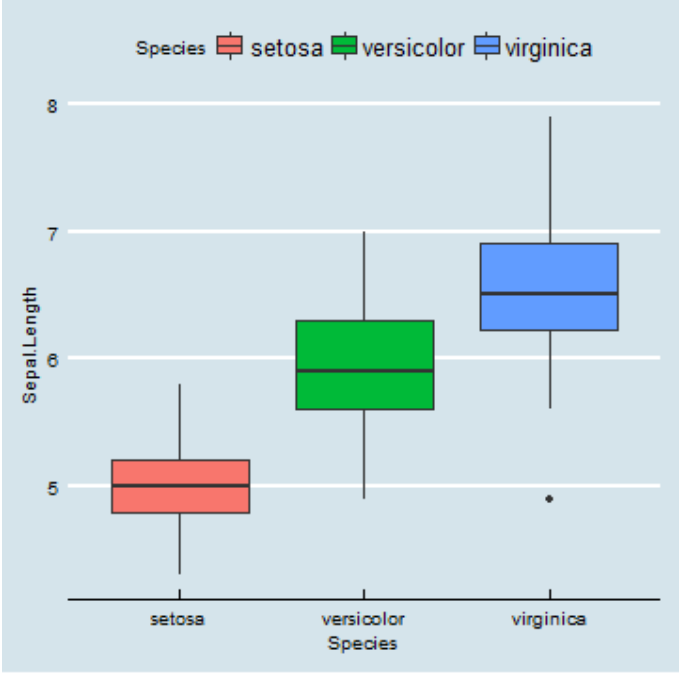
ggplot2 原始风格

```
library(ggplot2)
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species))
```



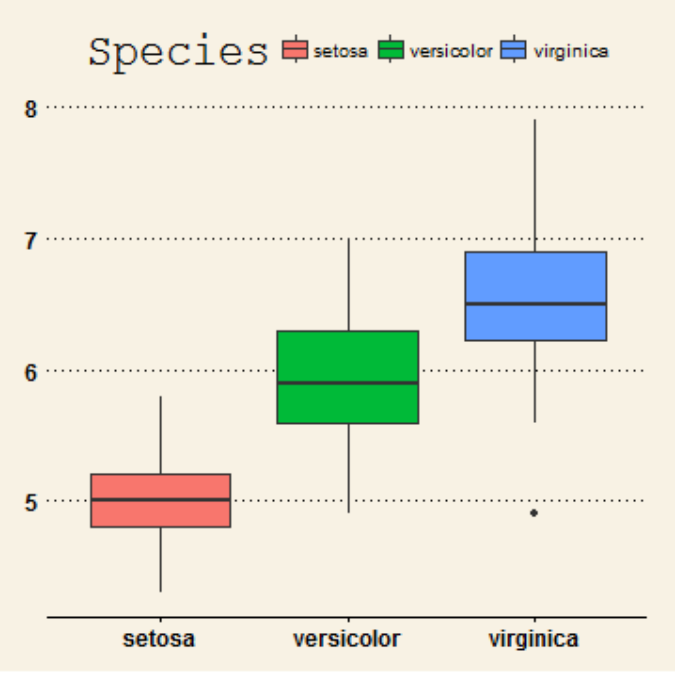
The Economist 风格图形

```
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
  ggthemes::theme_economist()
```



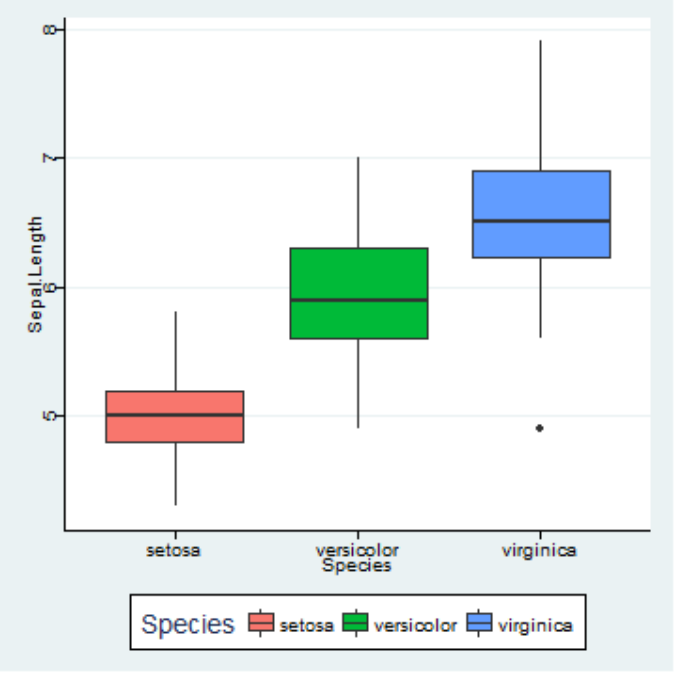
The Wallstreet Journal 风格图形

```
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
  ggthemes::theme_wsj()
```



Stata风格图形

```
ggplot(iris) +
  geom_boxplot(aes(x = Species, y = Sepal.Length, fill = Species)) +
  ggthemes::theme_stata()
```

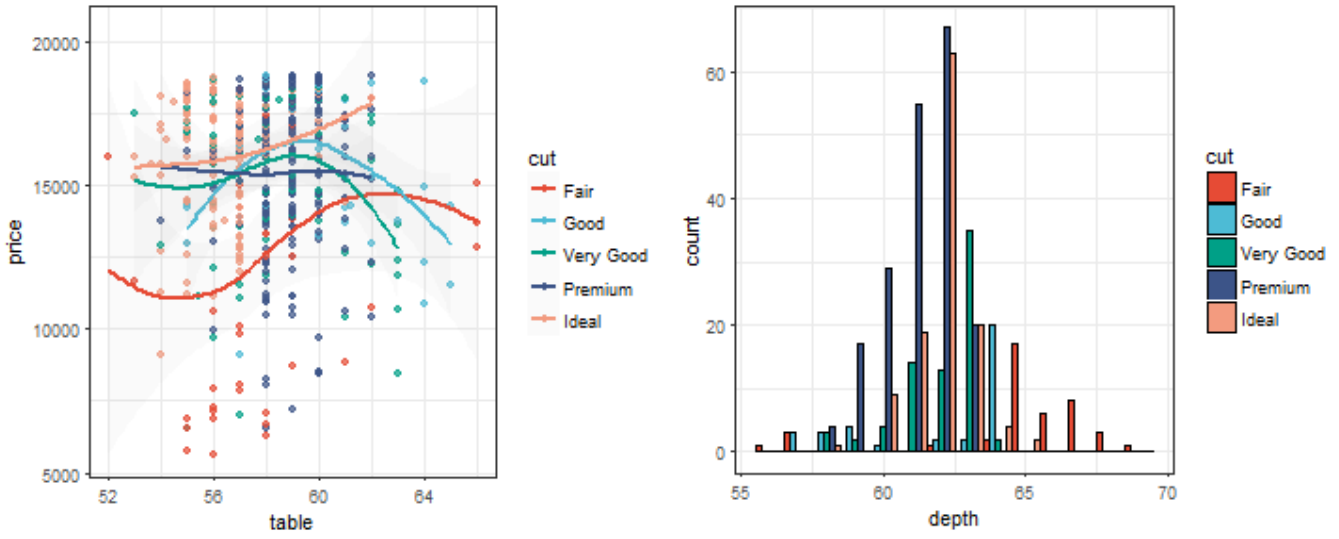


Nature 风格

```
library("ggsci")
library("ggplot2")
library("gridExtra")
data("diamonds")
p1 = ggplot(subset(diamonds, carat >= 2.2),
       aes(x = table, y = price, colour = cut)) +
  geom_point(alpha = 0.7) +
  geom_smooth(method = "loess", alpha = 0.05, size = 1, span = 1) +
  theme_bw()
p2 = ggplot(subset(diamonds, carat > 2.2 & depth > 55 & depth < 70),
       aes(x = depth, fill = cut)) +
  geom_histogram(colour = "black", binwidth = 1, position = "dodge")
  theme_bw()
```

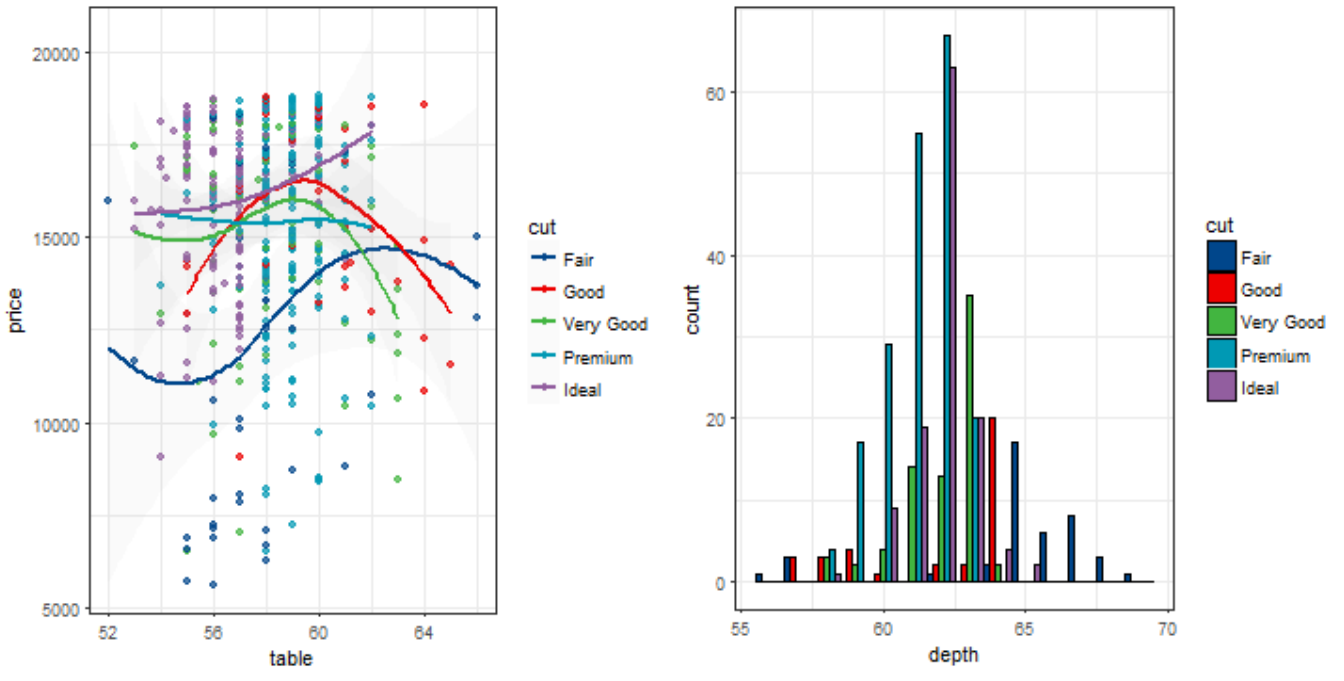
```
p1_npg = p1 + scale_color_npg()
p2_npg = p2 + scale_fill_npg()
grid.arrange(p1_npg, p2_npg, ncol = 2)
```

Nature 风格


price
table
20000
15000
10000
5000
52
56
60
64
cut
Fair
Good
Very Good
Premium
Ideal
count
depth
60
40
20
0
55
60
65
70
cut
Fair
Good
Very Good
Premium
Ideal

Lancet 风格

```
p1_lancet = p1 + scale_color_lancet()
p2_lancet = p2 + scale_fill_lancet()
grid.arrange(p1_lancet, p2_lancet, ncol = 2)
```



更多的R可视化图例

- RStudio图库
- ggplot2图库
- ggthemes示例
- ggsci示例

用于撰写学术报告

- rmarkdown: html 格式报告
- xaringan: html 格式幻灯片
- rticles: AER 等经济学类顶级刊物LaTeX模板
- stargazer: 生成LaTeX表格

结论与建议

- R 对于经济学教学与研究来说，是一个“无价”而高效的工具。
  - 无价的两层含义
    - 本身是免费的
    - 作用是巨大的

结论与建议

- R 对于经济学教学与研究来说，是一个“无价”而高效的工具。
  - 无价的两层含义
    - 本身是免费的
    - 作用是巨大的
- 通过RStudio 等IDE（集成开发环境），R已形成一个良好的数据分析生态
  - 可导入SAS、STATA、SPSS等常见格式的数据
  - 可用来生成动态化、交互式报告
  - 几乎可直接用来撰写论文与书籍
  - 有强大的社区支持
- R 自身的学习周期较长，不易上手
  - 这是包括Python等开源软件存在的一种通行问题
  - 可能需要进一步加强基础课程建设来加以解决

结论与建议

- 尽早让学生接触数据分析的基本流程
    - 在没有概率论支撑的情况下就可引入数据分析的基本理念，培养面向数据的分析思维
- 尽早、全面地引入计算工具，深度参与统计教学
    - 对于科研型院校，可适当强调程序思维和编程操作的基础性地位
    - 要将相关的数据分析的计算机技术当成基础数学三大模块（微积分、线性代数、概率论与数理统计）并列的模块加以重视
- 应更加突出可视化在数据分析中的作用
- 建立统一的、开放的、可编辑的数据展示与教学安全平台，对于节约教师的精力有很大的作用(GitHub可以作为一个有效的平台)。

常用资源

- 计量经济学中的常用 R 包索引：https://cran.r-project.org/web/views/Econometrics.html
- 用R做计量分析网站：https://econometricswithr.wordpress.com/
- Using R for Introductory Econometrics(Wooldridge 计量经济学导论配套R语言网站): http://www.urfie.net/
- bookdown官方网站：https://bookdown.org/home/
- *R for Data Science* 在线版本：http://r4ds.had.co.nz/

谢谢观看！

吕小康

xkdog@126.com

南开大学周恩来政府管理学院

本幻灯片由谢益辉的 R 包 **xaringan** 生成。原始文档可从以下链接下载：

https://github.com/xkdog/StatsUsingR

简略版可从以下网址在线观看（图片未能正确显示）：

https://github.com/xkdog/StatsUsingR/blob/master/R4Eco201707.Rmd

http://rpubs.com/xkdog/r4eco2017