Chapter 1: Introduction to R

Shujia Wong

Contents

Purpose of the course

- Introduction to basic statistical concepts
  - Descriptive data analysis
  - Random variables and probability distributions
  - Statistical inference
- Build and analyze basic statistical models
  - Regression models
  - Model diagnostics
- Introduction in using R
  - Understanding and programming correct code
  - Make correct interpretation from output

1 An overview of R

What is R?

- R is a language and environment for statistical computing and graphics
- was created in the early 1990s by Ross Ihaka and Robert Gentleman, at the University of Auckland
- is based upon the S language that was developed at Bell Laboratories

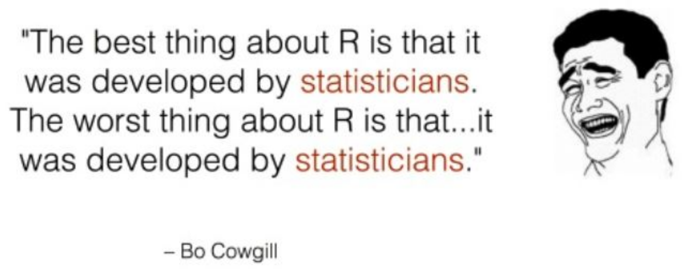
Why use R?

- Widely used in statistics and applied sciences
  - data scientists,The New York Times, Google,Facebook,Twitter ...
- Reliable free open source and cross platforms
- Versatile: Python, Matlab, MySQL, Perl, JAVA, C++, Fortran
- Extensible: new methods become available on a weekly basis (7000+ packages)
- Flexible: unlike other programs (e.g., SAS and SPSS)
- High quality visualization and graphics tools
- High level language with many built-in-functions (allows quick programming)

Disadvantages:

- Steep learning curve (frequent use helps at lot)

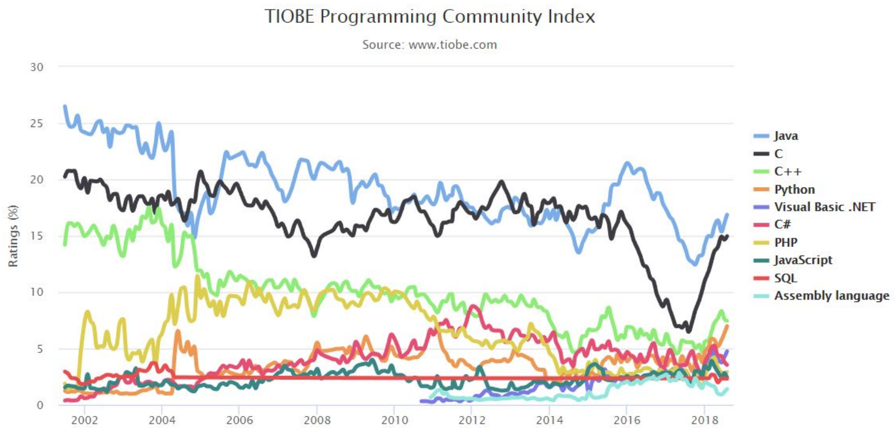
Pros and Cons



"The best thing about R is that it was developed by statisticians. The worst thing about R is that...it was developed by statisticians."

– Bo Cowgill

R or Python?



TIOBE Programming Community Index

Source: www.tiobe.com

How to get R and RStudio

- Download and install R: You must do this first.
  - http://CRAN.R-project.org/
- Download and install RStudio: Powerful integrated development environment (IDE) for R
  - https://www.rstudio.com/

R Environments

- Prompt: >
- Current working direction: getwd()
- Change working direction: setwd(“F:/AdvStat”) (or setwd(“F:\\AdvStat”))
- R is *case sensitive*, help(lm) and Help(lm) are different !
- Word processors are not recommended

Help on examples

from commandline

- help.start(),library()
- ls(package:stats), library(help="stats")
- ?t.test,?sum, ?solve
- methods(plot) # plotting functions
- example(boxplot) #examples
- demo(lm.glm) #demonstrations of generalized linear model
- mean.default #study code of function

Package libraries

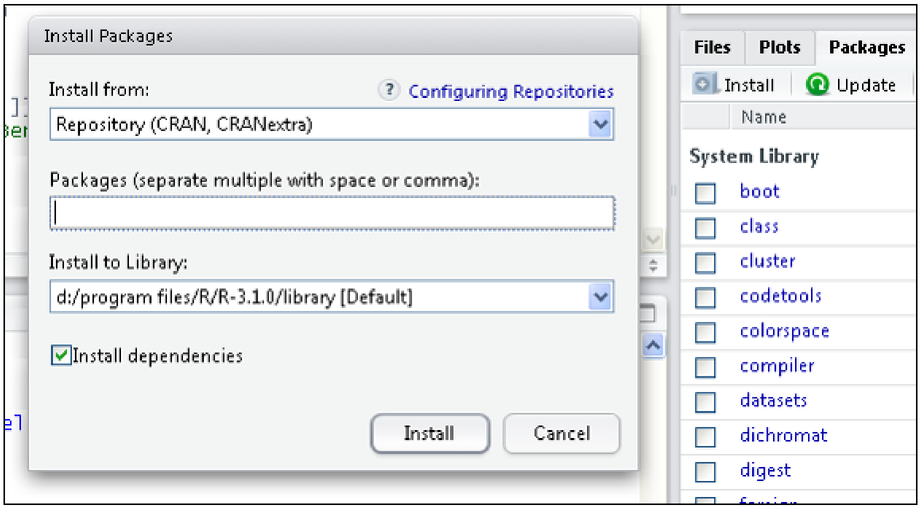
- What is a package?
  - A collections of R *functions*, *data*, and *compiled code* in a well-defined format
  - Packages extend the functionality of R: most of the value to R comes from the 7000+ packages out there
- Where do the packages come from?
  - Most packages are distributed “centrally” via CRAN (**C**omprehensive **R** **A**rchive **N**etwork)
  - There are lots of mirrors of CRAN.

* CTEX.ORG : http://ftp.ctex.org/mirrors/CRAN/
* Beijing Jiaotong University: http://mirror.bjtu.edu.cn/cran
* University of Science and Technology of China: http://mirrors.ustc.edu.cn/CRAN/
* Xiamen University: http://mirrors.xmu.edu.cn/CRAN/
* etc

Installing Packages

- R: install.packages("PASWR2")

- RStudio:



Install Packages

Install from: Repository (CRAN, CRANextra)

Packages (separate multiple with space or comma):

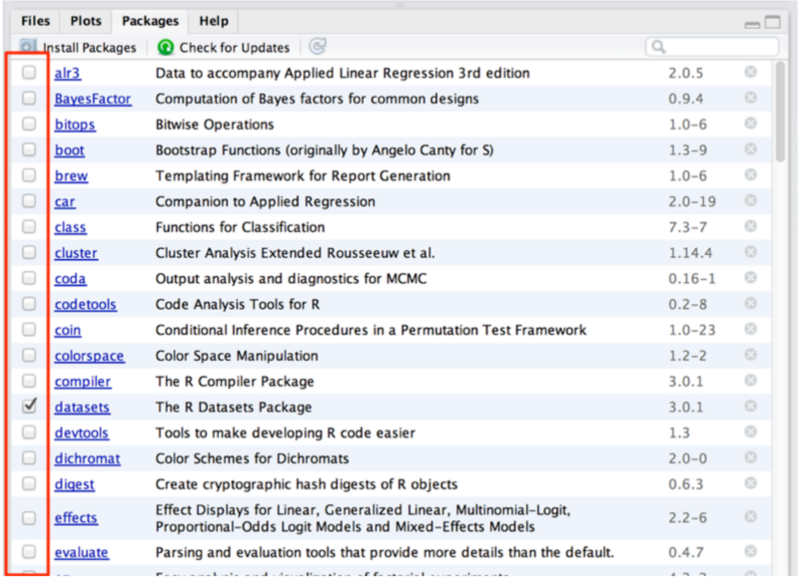
Install to Library: d:/program files/R/R-3.1.0/library [Default]

☑ Install dependencies

Loading Packages

- R: Loading package “datasets”: library("datasets")

- RStudio: package “datasets” is loaded



□	alr3	Data to accompany Applied Linear Regression 3rd edition	2.0.5
□	BayesFactor	Computation of Bayes factors for common designs	0.9.4
□	bitops	Bitwise Operations	1.0-6
□	boot	Bootstrap Functions (originally by Angelo Canty for S)	1.3-9
□	brew	Templating Framework for Report Generation	1.0-6
□	car	Companion to Applied Regression	2.0-19
□	class	Functions for Classification	7.3-7
□	cluster	Cluster Analysis Extended Rousseeuw et al.	1.14.4
□	coda	Output analysis and diagnostics for MCMC	0.16-1
□	codetools	Code Analysis Tools for R	0.2-8
□	coin	Conditional Inference Procedures in a Permutation Test Framework	1.0-23
□	colorspace	Color Space Manipulation	1.2-2
□	compiler	The R Compiler Package	3.0.1
☑	datasets	The R Datasets Package	3.0.1
□	devtools	Tools to make developing R code easier	1.3
□	dichromat	Color Schemes for Dichromats	2.0-0
□	digest	Create cryptographic hash digests of R objects	0.6.3
□	effects	Effect Displays for Linear, Generalized Linear, Multinomial-Logit, Proportional-Odds Logit Models and Mixed-Effects Models	2.2-6
□	evaluate	Parsing and evaluation tools that provide more details than the default.	0.4.7

Installing and Loading Packages

- *Installed*: install.packages("...") means...
  - That the package files are stored on your computer
  - Your version of R is able to load the package
- *Loaded*: library("...") means...
  - That R has opened the package files, and “knows” what they contain
  - You can use the functions/data stored in the package
- The upshot of this
  - A package must be *installed* before you can load it
  - A package must be *loaded* before you can use it
- Separating install from load avoids inconsistency
  - Install everything you might want to use sometime
  - Load only those things you need to use now!

Finally, What packages should you install?

After you installed the latest version of R and RStudio, you should install the following packages:

1. tidyverse: a set of core packages.

   # Install from CRAN

   install.packages("tidyverse")

   # Or the development version from GitHub

   install.packages("devtools")

   devtools::install_github("hadley/tidyverse")

2. PASWR2: codes, datatsets and functions from the text book.

   install.packages("PASWR2")

3. bookdown: a package to facilitate writing books and long-form articles/reports with R Markdown.

   - knitr - Easy dynamic report generation in R.
   - rmarkdown - Dynamic documents for R.
   - xfun
   - tinytex - A lightweight and easy-to-maintain LaTeX distribution
   - yaml

Core tidyverse packages

library(tidyverse) will load the core tidyverse packages:

ggplot2 for data visualisation.

dplyr for data manipulation.

tidyr for data tidying.

readr for data import.

purrr for functional programming.

tibble for tibbles, a modern re-imagining of data frames.

stringr for strings.

forcats for factors.

2 Using R

Frequently used operators

Operator	Description
<-	Assign
+	Sum
-	Difference
*	Multiplication
/	Division
^	Exponent
%%	Mod
%*%	Dot product
%/%	Integer division
%in%	Subset

Operator	Description
|	Or
&	And
<	Less
>	Greater
<=	Less or =
>=	Greater or =
!	Not
!=	Not equal
==	Is equal

2.1 Scientific calculating

R as a calculator

```
> 3^2+sqrt(4);factorial(5);log(10)

## [1] 11
## [1] 120
## [1] 2.302585

> exp(2); pi; sin(pi/3)

## [1] 7.389056
## [1] 3.141593
## [1] 0.8660254

> print("Hello world")

## [1] "Hello world"

> date()

## [1] "Tue Sep 18 12:02:39 2018"
```

R as a number generator

Sequence: seq(from, to, by=)

```
> x<-(1:12)
> x

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12

> seq(12)

##  [1]  1  2  3  4  5  6  7  8  9 10 11 12

> seq(4, 6, 0.25)

## [1] 4.00 4.25 4.50 4.75 5.00 5.25 5.50 5.75 6.00
```

R as a number generator

Repetition: rep(x, times, …)

```
> rep(10, 3)

## [1] 10 10 10

> rep(c(1:3), 3)

## [1] 1 2 3 1 2 3 1 2 3

> rep(c(3.14, 2.71), 3)

## [1] 3.14 2.71 3.14 2.71 3.14 2.71
```

R as a probability calculator

- Binomial probability $P(X=k)=C_n^k p^k(1-p)^{n-k}$: dbinom(k, n, p)

```
> dbinom(2, 5, 0.60)

## [1] 0.2304
```

- Probability of $P(|Z|<1.96)$ with $Z \sim N(0,1)$

```
> pnorm(1.96,0,1)-pnorm(-1.96,0,1)

## [1] 0.9500042
```

- Poisson probability$P(X=k|\lambda)=\frac{\lambda^k}{k!}e^{-\lambda}$: dpois(k, l)

```
> dpois(2,1)

## [1] 0.1839397
```

R as a sampler

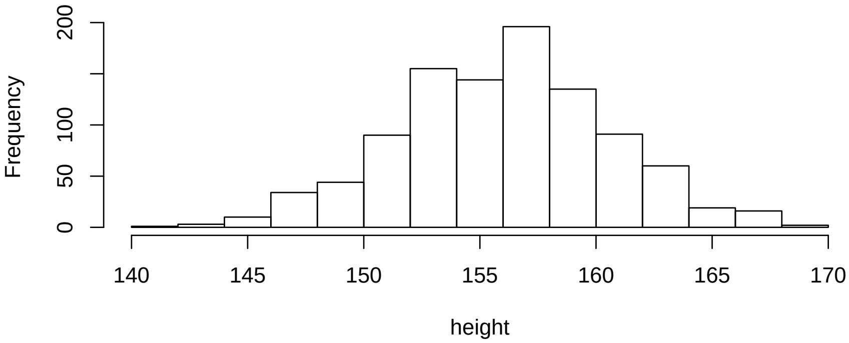
- Darw 5 people from a 50 group

```
> sample(1:50,5);sample(1:50,5);sample(1:50,5,replace=T)

## [1] 28 42  7 30 16
## [1] 22  2  8 14  7
## [1] 41  5 19 44 16

> set.seed(1234);sample(1:50,5)

## [1]  6 31 30 48 40

> set.seed(1234);sample(1:50,5)

## [1]  6 31 30 48 40
```

R as a simulator

Average height of female students in a class is 156 cm, with standard deviation being 4.6 cm. If we randomly take 1000 female students from this population, what is the distribution of height?

```
> height<-rnorm(1000,mean=156,sd=4.6)
> hist(height,main="")
```



Vectors

```
> x <- 1:5
> length(x)

## [1] 5

> sum(x)

## [1] 15

> x1 <- seq(1, 5, by = 1)
> x2 <- seq(1, 5, length = 5)
> x1==x2

## [1] TRUE TRUE TRUE TRUE TRUE

> x <- letters[1:5]
> x == c("b")

## [1] FALSE  TRUE FALSE FALSE FALSE
```

Matrix Operations

Let a be a scalar, $\boldsymbol{A}$ and $\boldsymbol{B}$ be two real matrices

```
> a<-2
> A<-matrix(c(1,2,3,4),nrow=2,ncol=2) # Real matrix
> A
```

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    2    4

> B<-matrix(c(1,2,2,7),nrow=2,ncol=2) # Symmetric real matrix.
> B

##      [,1] [,2]
## [1,]    1    2
## [2,]    2    7
```

Matrix Multiplication

- Dot product: $\boldsymbol{AB}$

```
> A%*%B

##      [,1] [,2]
## [1,]    7   23
## [2,]   10   32
```

- Cross product: $\boldsymbol{A}^T\boldsymbol{B}$

```
> crossprod(A,B) # t(A)%*%B

##      [,1] [,2]
## [1,]    5   16
## [2,]   11   34
```

Matrix Multiplication

- Entry-wise multiplication:

```
> A*B

##      [,1] [,2]
## [1,]    1    6
## [2,]    4   28
```

- Entry-wise division:

```
> A/B

##      [,1]      [,2]
## [1,]    1 1.5000000
## [2,]    1 0.5714286
```

Soving a linear system

- Matrix inversion: A^{-1}

```
> solve(A)

##      [,1] [,2]
## [1,]   -2  1.5
## [2,]    1 -0.5
```

- Matrix division: $\boldsymbol{A}^{-1}\boldsymbol{B}$

```
> solve(A,B) #Identical to: solve(A)%*%B

##      [,1] [,2]
## [1,]    1  6.5
## [2,]    0 -1.5

> solve(A)%*%B

##      [,1] [,2]
## [1,]    1  6.5
## [2,]    0 -1.5
```

Logical operators

```
> 5>4;5>=4;5==4;5!=4

## [1] TRUE
## [1] TRUE
## [1] FALSE
## [1] TRUE

> 5 >= 4 & 4 <= 5 # AND

## [1] TRUE

> 5 >= 4 | 4 == 5 # OR

## [1] TRUE
```

Functions: standard normal

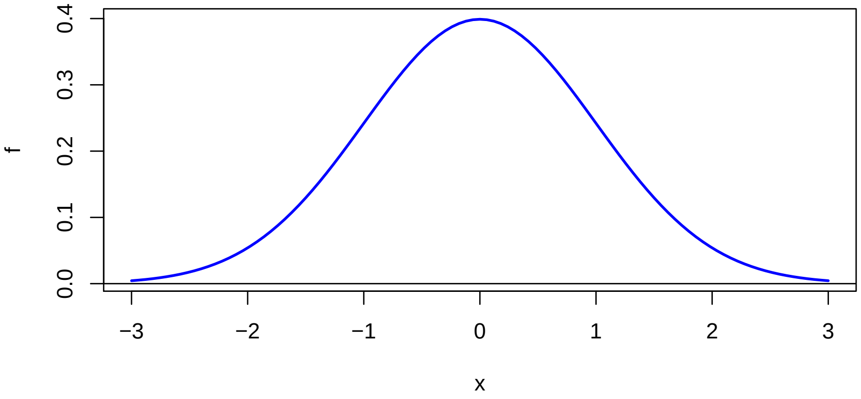
$$f(x) = 1/\sqrt{2\pi}\exp(-x^2/2)$$

```
> f <- function(x){exp(-x^2/2)/sqrt(2*pi)}
> f(1)

## [1] 0.2419707
```

```
> f(1) == dnorm(1) # built-in-function

## [1] TRUE

> class(f)

## [1] "function"

> integrate(f,-2,2)

## 0.9544997 with absolute error < 1.8e-11
```

Normal density curve

```
> f<-function(x){exp(-x^2/2)/sqrt(2*pi)}
> plot(f,xlim=c(-3,3),col="blue",lwd = 2)
> abline(h=0)
```



2.2 Data Analysis

Descriptive Statistics

```
> x<-0:9
> x

##  [1] 0 1 2 3 4 5 6 7 8 9

> mean(x);median(x);mode(x)

## [1] 4.5
## [1] 4.5
## [1] "numeric"
```

```
> var(x);sd(x)

## [1] 9.166667
## [1] 3.02765
```

Simply: summary()

```
> x<-rnorm(20,1,2);x

##  [1]  2.04450953 -2.98110658  0.29468007  4.06866388 -0.78040515
##  [6]  0.36584193  2.22403890 -0.09796437  3.12963293 -2.62382662
## [11]  3.61654682 -0.07417713 -1.18704919  1.17546953  0.91070953
## [16]  0.67540988  0.27711350  1.51194329  1.44317011  5.04198771

> summary(x)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -2.98111 -0.08012  0.79306  0.95176  2.08939  5.04199
```

2.3 Statistical Modeling

Linear model

```
> x<-0:19; y<-x+rnorm(20,0,1);out<-lm(y~x)
> summary(out)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.42281 -0.57087 -0.09943  0.73880  1.84445
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.66904    0.40359   1.658    0.115
## x            0.92433    0.03632  25.452 1.45e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9365 on 18 degrees of freedom
## Multiple R-squared:  0.973,Adjusted R-squared:  0.9715
## F-statistic: 647.8 on 1 and 18 DF,  p-value: 1.449e-15
```

Linear model

```
> library(xtable)
> xtable(out)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6690	0.4036	1.66	0.1147
x	0.9243	0.0363	25.45	0.0000

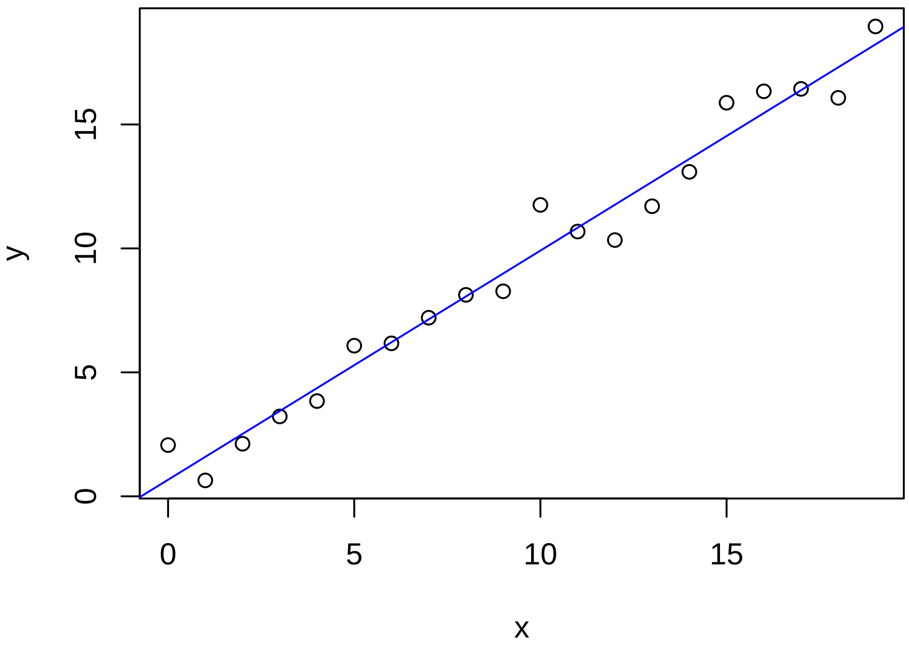
```
> xtable(anova(out))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	568.17	568.17	647.81	0.0000
Residuals	18	15.79	0.88		

2.4 Data Visualization

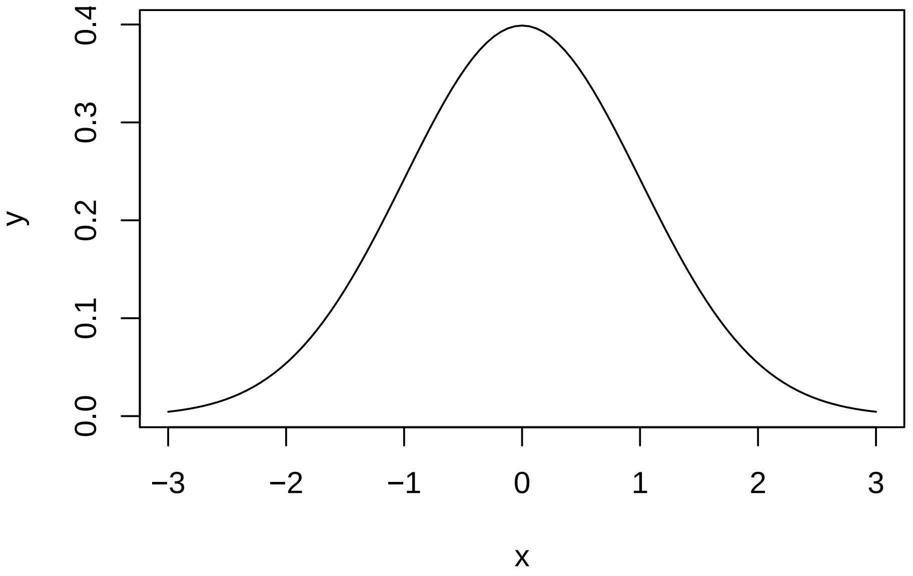
Data Plot

```
> plot(x,y)
> abline(lm(y~x),col='blue')
```



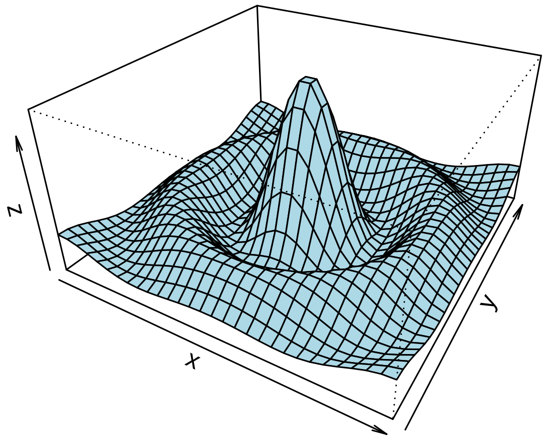
Density Curve of Standard Normal

```
> x<-seq(-3,3,length=101)
> y<-dnorm(x) # assign standard normal values to y
> plot(x,y,type='l')  # 'l' stands for line
```



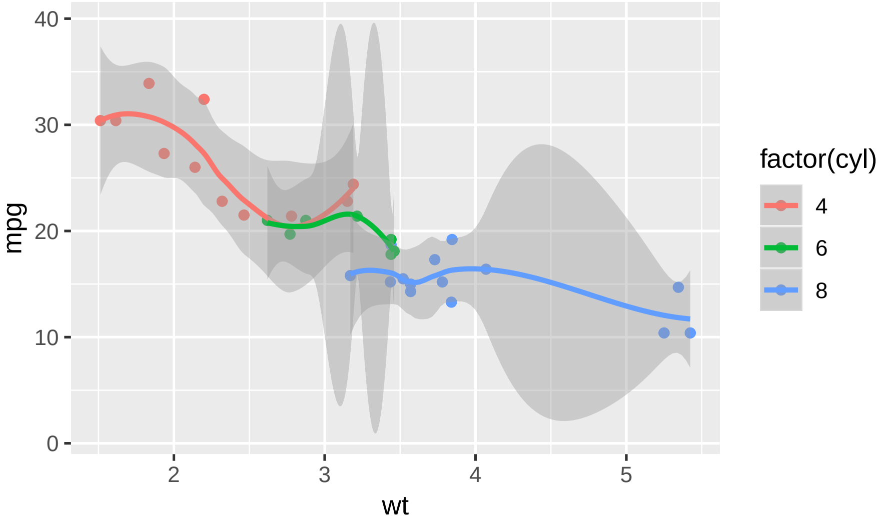
3D surface curve

```
> x<-seq(-10,10,length=30)
> y<-x
> f<-function(x,y){
+   r<-sqrt(x^2 + y^2)
+   10*sin(r)/r }
> z<-outer(x,y,f)
> z[is.na(z)]<-1
> persp(x,y,z,theta=30,phi=30,expand=0.5,col="lightblue")
```



ggplot2

```
> library(ggplot2)
> qplot(wt, mpg, data=mtcars, color=factor(cyl), geom=c("point", "smooth"))
```



3 Web resources for R

Textbook Supports
http://alanarnholt.github.io/PASWR2E-Book/

- Overview of the Book
- Supplementary Materials
- R-Scripts
- Errata

R Homepage
https://www.r-project.org/

- List of CRAN mirror sites
- Manuals
- FAQs
- Site seach
- Mailing lists
- Links

CRAN - Comprehensive R Archive Network

http://cran.fhcrc.org/

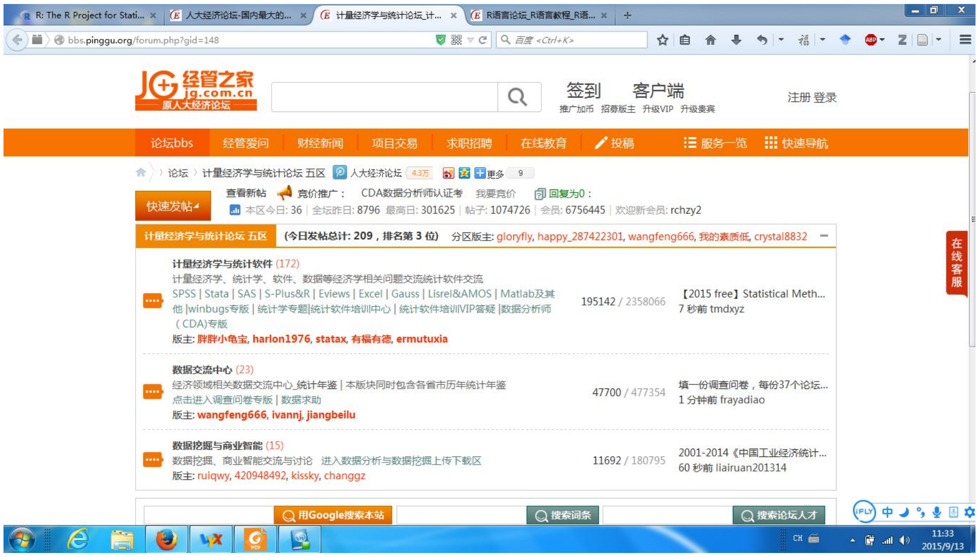
- CRAN Mirrors
  - About 90 sites worldwide
  - About 20 sites in US
- R Binaries
- R Packages
  - 7000+ packages
- R Sources
- Task Views

CRAN Task Views

Organizes the 7000+ R packages by application

- Bayesian
- Econometrics
- Finance
- Time Series
- Meta Analysis
- Optimization
- Machine Learning
- Spatial
- etc

An Excellent Forum (in Chinese)



论坛 › 计量经济学与统计论坛 五区 人大经济论坛 4.3万 更多 9

查看新帖 竞价推广： CDA数据分析师认证考 我要竞价 回复为0：

本区今日: 36 | 全坛昨日: 8796 最高日: 301625 | 帖子: 1074726 | 会员: 6756445 | 欢迎新会员: rchzy2

计量经济学与统计论坛 五区 (今日发帖总计: 209，排名第 3 位) 分区版主: gloryfly, happy_287422301, wangfeng666, 我的素质低, crystal8832

计量经济学与统计软件 (172) 计量经济学、统计学、软件、数据等经济学相关问题交流统计软件交流 SPSS | Stata | SAS | S-Plus&R | Eviews | Excel | Gauss | Lisrel&AMOS | Matlab及其他 |winbugs专版 | 统计学专题|统计软件培训中心 | 统计软件培训VIP答疑 |数据分析师（CDA)专版 版主: 胖胖小龟宝, harlon1976, statax, 有福有德, ermutuxia	195142 / 2358066	【2015 free】Statistical Meth... 7 秒前 tmdxyz
数据交流中心 (23) 经济领域相关数据交流中心_统计年鉴 | 本版块同时包含各省市历年统计年鉴 点击进入调查问卷专版 | 数据求助 版主: wangfeng666, ivannj, jiangbeilu	47700 / 477354	填一份调查问卷，每份37个论坛... 1 分钟前 frayadiao
数据挖掘与商业智能 (15) 数据挖掘、商业智能交流与讨论 进入数据分析与数据挖掘上传下载区 版主: ruiqwy, 420948492, kissky, changgz	11692 / 180795	2001-2014《中国工业经济统计... 60 秒前 liairuan201314

- http://bbs.pinggu.org/

Quick R

http://www.statmethods.net
Site maintained by Robert Kabacoff, author of R in Action

- Introductory R Lessons
- R Interface
- Data Input
- Data Management
- Basic Statistics
- Advanced Statistics
- Basic Graphs
- Advanced Graphs

Other useful R sites

- **Stackoverflow**: the primary resource for help with R
  - http://stackoverflow.com/
- **R Bloggers**: Aggregation of about 450 R blogs
  - http://www.r-bloggers.com

- **R Graph**: Gallery Examples of many possible R graphs
  - http://addictedtor.free.fr/graphiques
- **Google**: everything!

Course Materials

- Course Materials: Go to
  - https://github.com/Andrewsky123/Advanced-Statistics-With-R
  - then click clone or download and then Download.zip to your computer.
- Tutorial for writing scientific documents with R Markdown https://github.com/Andrewsky123/R-Markdown-Notes

References

Casella, G. and Berger, R. (2002). *Statistical Inference.* Wadsworth Group, Duxbury, 2nd edition.

Cohen, Y. and Cohen, J. Y. (2008). *Statistics and data with R: an applied approach through examples.* Wiley, Chichester, U.K.

Kabacoff, R. (2015). *R in action: data analysis and graphics with R.* Manning, Shelter Island, NY, 2nd edition.

Ugarte, M. D., Militino, A. F., and Arnholt, A. T. (2016). *Probability and Statistics with R (Text book).* CRC Press, Boca Raton, FL, 2nd edition.