

# Data Analysis Using R: Introduction

2016 Spring Term

罗智超 (ROKIA.ORG)

## 1 联系方式

- 姓名: 罗智超 (Rokia.org)
- Email&QQ: [1814347@qq.com](mailto:1814347@qq.com)
- QQ: 1814347
- 课程 QQ 群:167472549
- 课件 : [http://rokia.org/?page\\_id=303](http://rokia.org/?page_id=303)
- Git : <http://www.github.com/zhichaolu/DataAnalysis/>
- 课程相关资料 : <http://pan.baidu.com/s/1qWv6NY8> 提取码 : lt3l

## 2 注意事项

- 为兴趣、热情而不是为考试而学习
- 将手机关闭或者调整成静音状态
- 尽量坐在前排，如果你想学习

### 3 课程将给你什么？

- 一起度过一学期**痛并快乐**的学习时光
- 增强你的简历
- 增强你的动手处理和分析数据的能力
- 掌握一门将来可能赖以生存的技能
- 掌握常用的统计（数据科学）模型和方法
- 以上的收获将基于你的坚持与付出

注：本学期的课件相比其它学期加强了作业练习的要求，加强以真实项目训练为导向，提升学生动手能力和独立解决问题能力，而不是单纯的知识点学习。

## 4 课程教材

- R In Action, Data Analysis and Graphics with R by Robert I. Kabacoff
- The Art of R Programming by Norman Matloff
- ggplot2: Elegant Graphics for Data Analysis by Hadley Wickham
- 统计建模与 R 软件 by 薛毅陈立萍

## 5 辅助阅读材料

- Statistics: From Data to Evidence by Xizhi WU
- Statistics: With the application of R by Xizhi WU
- Data Manipulation With R by Phil Spector
- An Introduction to R by Bill Venables & David Smith
- R for Programmers by Norman Matloff
- The Lady Tasting Tea show Statistics Revolutionized Science in the Twentieth Century

## 6 前修知识

- 统计学原理
- 计算机一、二级
- 热爱编程

## 7 推荐网站

- <http://www.rokia.org/>
- <http://www.r-project.org/>
- <http://www.rstudio.com/>
- <http://library.xmu.edu.cn/>
- <http://www.pinggu.org/>
- <http://COS.NAME/>
- <http://Coursera.org/>
- <http://www.jstatsoft.org/>
- <http://www.r-bloggers.com/>
- <http://www.stackoverflow.com/>
- <http://51qiangda.com/>
- <http://ctex.org/>
- <http://www.kaggle.com/>
- <http://www.kdnuggets.com/>
- <http://en.savefrom.net/>

## 8 推荐软件

- Adobe Acrobat Professional
- CNKI E-Learning
- CTEX
- GIT
- RStudio/R
- UltraEdit

## 9 推荐电脑

MacBook Pro

## 10 数据分析的重要性

- 几乎所有科学研究都要涉及数据分析
- 几乎所有文章发表都需要涉及数据分析
- 熟练掌握数据分析技能会使你在工作学习中成为“香馍馍”
- 养成数据分析思维



11 数据分析如烹小鲜

Step	Data Analysis	Cooking
1	Data, Software, Model	Raw Material, Kitchenware, cookbook
2	Define the research question	Decide which dish to eat
3	Collect Data	Shopping
4	Data Cleaning	Cleaning
5	Data Preparation	Chopping
6	Modeling with theory and data	Cooking with cookbook
7	Hypothesis Test	Taste
8	Report	Decoration

数据分析与烹饪最主要的差别是煮完菜要洗碗  
如果对烹饪不感兴趣的童鞋可以欣赏这部电影，包括主题曲



图 1: 《陪安东尼度过漫长岁月》



图 2: 小野二郎

## 12 数据分析的流程

- Define the question
- Define the ideal data set
- Determine what data you can access
- Obtain the data
- Clean the data
- Exploratory data analysis
- Statistical prediction/modeling
- Interpret results
- Challenge results
- Synthesize/write up results
- Create reproducible code

## 13 数据分析的三项基本技能

- 数据操作能力
- 统计编程及绘图能力
- 统计理论及业务理解能力

## 14 数据分析的问题类型

- 描述 ( Descriptive )
- 探索 ( Exploratory )
- 推断 ( Inferential )
- 预测 ( Predict )
- 因果 ( Casual )
- 机理 ( Mechanistic )

## 15 数据的来源

- 普查
- 抽样调查
- 试验设计

## 16 数据载体

- Tab-delimited text
- Comma-separated text
- Excel file
- JSON File
- HTML/XML file
- Database



## 17 统计学 VS 数据分析 VS 数据科学

- 统计学

收集、分析、展示和解释数据的科学。( 吴喜之 )

- 数理统计学

数理统计学是数学的一个分支。它的任务是研究怎样用有效的方法去收集和使用带随机性影响的数据。( 陈希孺 )

- 数据分析

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析, 提取有用信息和形成结论而对数据加以详细研究和概括总结的过程 ( 百度百科 )

## 18 数据科学家

数据科学家是指能采用科学方法、运用数据挖掘工具对复杂多量的数字、符号、文字、网址、音频或视频等信息进行数字化重现与认识，并能寻找新的数据洞察的工程师或专家（不同于统计学家或分析师）。一个优秀的数据科学家需要具备的素质有：懂数据采集、懂数学算法、懂数学软件、懂数据分析、懂预测分析、懂市场应用、懂决策分析等。（百度百科）

一个数据科学家不能只擅长数字（这种人被称为统计学家或分析师），也要能够理解业务：什么样的数据或结果才是有参考性的；能够找到新的数据集并为其创造新产品；然后能够让 CEO 们理解这一切。这是一个艰巨的任务，这个世界上这类人是很少的。作为顶尖的数据科学家，不要求他们对环境做出什么积极的改变，但是需要他们尝试做一些真正先进的东西，帮助大家更好的解决业务上的问题。

## 19 数据科学家的六种能力：

1. 对数据的提取与综合能力；
2. 统计分析能力；
3. 数据洞察与信息挖掘能力；
4. 开发软件能力；
5. 网络编程能力；
6. 数据的可视化表示能力。

## 20 数据科学家涉及学科：

1. 计算机科学：数据获取、数据解析、数据存放、和数据安全
2. 数理统计学：数据分析、数据过滤、数据挖掘、和数据优化
3. 图形设计学：显示数据结果，比如将数据表达成三维图形，以便更好地理解 and 利用
4. 人机交互学：在用户和数据之间建立有机联系，使得人对数据的使用更方便

21 Data Science Venn Diagram

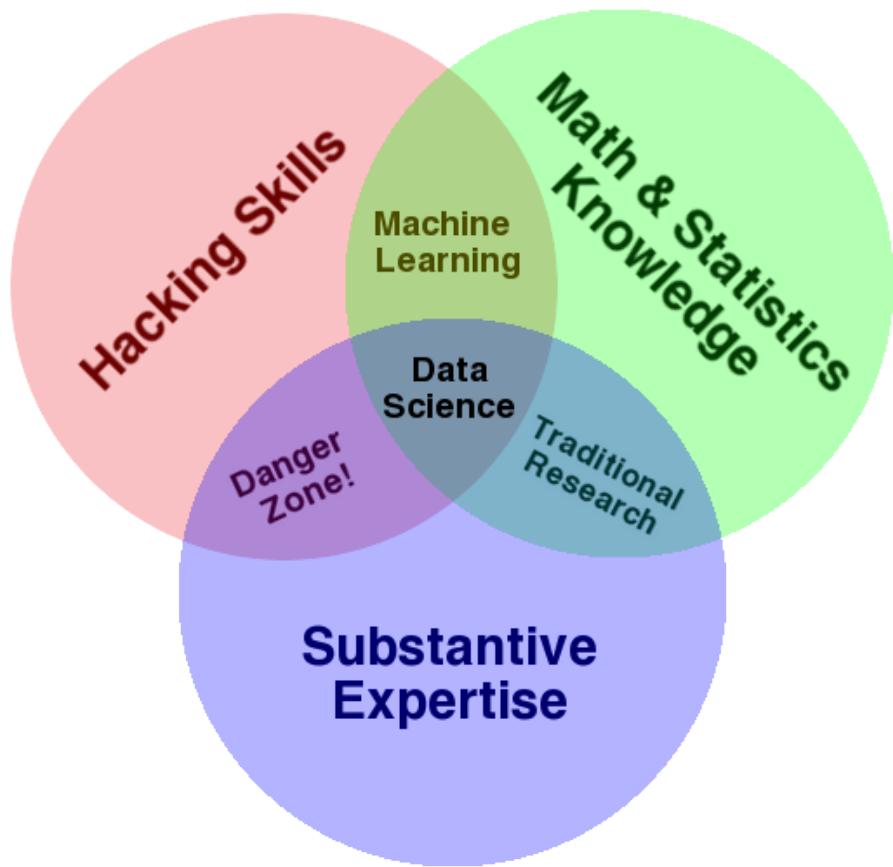


图 3: DataScienceVennDiagram

## 22 真实数据

- “脏”数据
  - 格式不规范
  - 缺失值
  - 错误数据
  - 异常值
- 格式不规范

23 Dirty Data

	A	B	C	D	E	F	G	H	I
1	Qualifications by Year Level and Gender								
2					National				
3					Year 11		Year 12		Year 13
4	Qualificat	Gender							
5									
6	National Certificate of Educational Achievement								
7	NCEA (Level 1)								
8		Male			5,929		6,427		5,170
9		Female			0		60		38
10	NCEA (Level 2)								
11		Male			194		5,395		5,027
12		Female			0		58		38
13	NCEA (Level 3)								
14		Male			2		128		3,276
15		Female			0		0		36
16									

图 4: DirtyData1

图 5: DirtyData2



## 24 辛普森悖论

- 辛普森悖论 ( Simpson's Paradox ) 亦有人译为辛普森诡论, 为英国统计学家 E.H. 辛普森 ( E.H.Simpson ) 于 1951 年提出的悖论, 即在某个条件下的两组数据, 分别讨论时都会满足某种性质, 可是一旦合并考虑, 却可能导致相反的结论。
- 辛普森悖论就像是欲比赛 100 场篮球以总胜率评价好坏, 于是有人专找高手挑战 20 场而胜 1 场, 另外 80 场找平手挑战而胜 40 场, 结果胜率 41%, 另一人则专挑高手挑战 80 场而胜 8 场, 而剩下 20 场平手打个全胜, 结果胜率为 28%, 比 41% 小很多, 但仔细观察挑战对象, 后者明显较有实力。简单的将分组资料相加汇总, 是不一定能反映真实情况的。

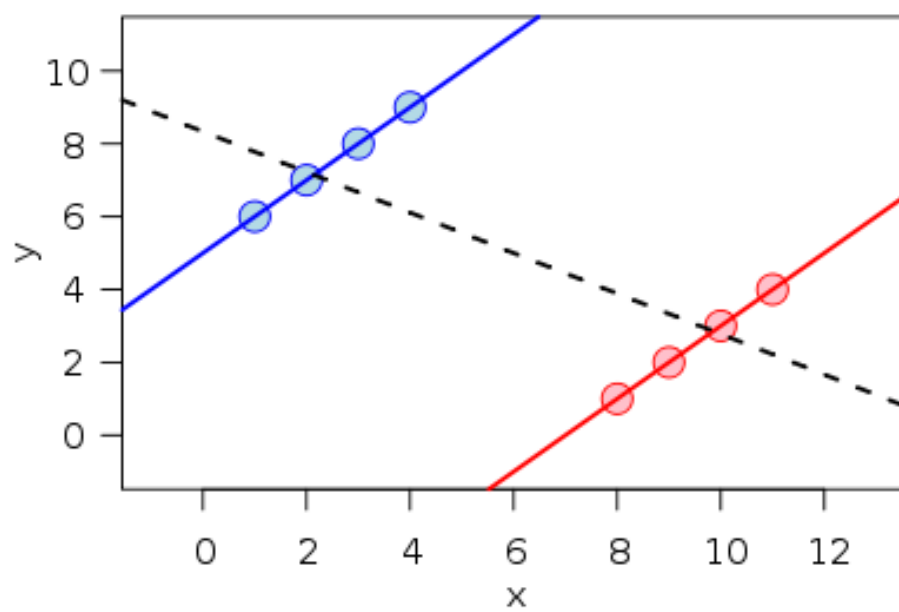


图 6: Simpson's Paradox Graphic

## 25 数据可视化

数据可视化在数据分析过程中起到非常核心的作用，并且随着交互式数据可视化工具的发展，其作用将更加明显。

## 26 为什么是 useR?

- S 语言的理念，用它的发明者 John Chambers 的话说就是 “to turn ideas into software, quickly and faithfully
- John Chambers 是这样对 R 语言定义的
  - An interface to computational procedures of many kinds;
  - Interactive, hands-on in real time;
  - Functional in its model of programming;
  - Object-oriented, “everything is an object”;
  - Modular, built from standardized pieces;
  - Collaborative, a world-wide, open-source effort.

## 27 关于 R 的前世今生的几篇文章

- 刘思~~囡~~, [R You Ready?——大数据时代下优雅、卓越的统计分析及绘图环境](#)
- 谢益辉, 郑冰 (2008). [R 语言的历史背景、发展历程和现状](#). 1st China R Conference

## 28 SAS 和 R 的比较分析

- 谢益辉在统计之都的[这篇文章](#)后面的评论记录了 SAS 和 R 用户的一场口水战
- 胡江堂有两篇[Think SAS\(二\)](#)[Think SAS\(二\)](#)值得读读

- 太灵活、选择太多
- 需要一定编程基础
- 真正陡峭的是后面的统计知识基本功（因为它的诞生本身就是用于统计研究）

- 热爱编写代码
- 多动手写代码
- 多看别人代码

- R 帮助文档
- [stackoverflow.com](http://stackoverflow.com)
- [stats.stackexchange.com/](http://stats.stackexchange.com/)
- [google/baidu](http://google/baidu)

### 32.1 Anscombe Data

##	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
```

思考：这样的数据表达方式是否有效

### 33 Anscombe Data Mean

```
##      x1      x2      x3      x4      y1      y2      y3      y4
## 9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909

## [1] 8.65250 7.45250 10.47125 8.56625 9.35875 10.49250 6.33750
## [8] 7.03125 9.71000 6.92625 5.75500

##      x1      x2      x3      x4      y1      y2      y3      y4
## 9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
```

### 34 Anscombe Data Variance

```
##      x1      x2      x3      x4      y1      y2      y3
## 11.000000 11.000000 11.000000 11.000000 4.127269 4.127629 4.122620
##      y4
## 4.123249
```

### 35 Anscombe Data Plot

[AnscombeDataPlot]:(image/anscombeplot.png)

#### 35.1 Monty Hall

三门问题 ( Monty Hall Problem ) 亦称为蒙提霍尔问题、蒙特霍问题或蒙提霍尔悖论。大致出自美国的电视游戏节目 Let's Make a Deal。问题名字来自该节目的主持人蒙提·霍尔 ( Monty Hall )。参赛者会看见三扇关闭了

的门，其中一扇的后面有一辆汽车，选中后面有车的那扇门可赢得该汽车，另外两扇门后面则各藏有一只山羊。当参赛者选定了一扇门，但未去开启它的时候，节目主持人开启剩下两扇门的其中一扇，露出其中一只山羊。主持人其后会问参赛者要不要换另一扇仍然关上的门。问题是：换另一扇门会否增加参赛者赢得汽车的机会率？

```
MontyHall<-function(Dselect,Dchange){
  Dcar<-sample(1:3,1)
  #print(Dcar)
  if (Dcar==Dselect & Dchange==0) return(1)
  else if (Dcar!=Dselect & Dchange==0) return(0)
  else if (Dcar==Dselect & Dchange==1) return(0)
  else return(1)
}

MontySim<-function(n,Dchange){
  win<-0
  for(i in 1:n){
    Dselect<-sample(1:3,1)
    win<-win+MontyHall(Dselect,Dchange)
  }
  pwin<-win/n
  return(pwin)
}

MontySim(10000,0)
MontySim(10000,1)
```

事实上不换门的话，赢得汽车的几率是  $1/3$ 。换门的话，赢得汽车的几率是  $2/3$ 。因为，当你从三扇门中选了门 1 后，这扇门后面有奖的几率是  $1/3$ ，另两扇门是  $2/3$ 。但接下来主持人给了你一个线索。如果奖品在门 2 后，主持人将会打开门 3；如果奖品在门 3 后，他会打开门 2。所以如果你改选的话，只要奖品在门 2 或门 3 后你就会赢，两种情况你都会赢！但是如果你不改选，只有当奖品在门 1 后你才会赢。

## 36 课后练习

- 评估下自己的 R 语言水平[测验](#)
- 登陆 QQ 群
- 安装课程要求的软件
- 下载课件及相关教材

## 37 每周“大牛”

- Sir R.A.Fisher ( 1890 ~ 1962 ), 全名 Ronald Aylmer Fisher, 生于伦敦, 卒于 Adelaide ( 澳洲 )。英国统计与遗传学家, 现代统计科学的奠基人之一, 并对达尔文进化论作了基础澄清的工作。
- Fisher 以天文学学士毕业于剑桥大学, 也因对天文观测误差的分析, 使他开始探讨统计的问题。毕业後几年, 他曾到加拿大务农, 工作于投资公司, 也当过私立学校的老师。并在 1915, 1918 发表两篇重要文章, 前者探讨相关系数的分布; 後者证明遗传上的连续变异, 可用许多遵守孟德尔律的基因变异所叠加来解释。他一生在统计和生物的研究兴趣与才华, 已经清楚地表现出来。
- 1919 年他拒绝在 K. Pearson 下工作, 任职于 Rothamsted 农业实验场。他负责的主要工作是植物播殖实验的设计, 希望透过尽量少的时间、成本与工作量, 得到尽量多的有用资讯; 另外是要整理该实验场 60 年来累积的实验资料。Fisher 在这里发展他的变异数分析理论, 研究假说测试, 并且提出实验设计的随机化原则, 使得科学试验可以同时多参数之检测, 并减少样本偏差。



- 他在 1925 所著《研究工作者的统计方法》影响力超过半世纪，遍及全世界。而他在 Rothamsted 的工作结晶，同时也表现在为达尔文演化论澄清迷雾的巨著《天择的遗传理论》(1930) 中，说明孟德尔的遗传定律与达尔文的理论并不像当时部份学者认为的互相矛盾，而是相辅相成的。并且认为演化的驱动力主要来自选择的因素远重于突变的因素。这本著作将统计分析的方法带入演化论的研究。为解释现代生物学的核心理论打下坚实的基础。也因这本著作，Fisher 1933 年获得伦敦大学的职位，从事 RH 血型的研究。
- 1943 至 1957 年他回剑桥大学任教,1952 年受封爵士,1956 年出版《统计方法与科学推演》，最後三年，则在澳洲为国协科技研究组织 (CSTRO) 工作，并卒於任上。