

概率统计 B

第零章 序言

原著：陈家鼎、刘婉如、汪仁官
制作：李东风，邓明华

2025 春季学期

本章目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计
- 4 统计学与人工智能
- 5 教学安排
- 6 R 简介

本节目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计
- 4 统计学与人工智能
- 5 教学安排
- 6 R 简介

教学目的

- 掌握概率论和数理统计的基本数学知识。
- 训练用概率论和数理统计方法对实际问题进行数学建模的能力。
- 学会解决常见的统计分析问题。

课程信息

- 周学时 3 (单周三 5-6, 每周五 1-2), 理教 203;
- 助教信息
 - ① 虞家伟 (2201110030 at pku.edu.cn)
 - ② 张平 (zhangping0114 at pku.edu.cn)
 - ③ 英乃文 (ying0126v at pku.edu.cn)
- 考评: 暂定平时作业: 20%, 期中考试: 40%, 期末考试: 40%.
- 期末考试在 6 月 20 日周五上午, 期中考试大约在第 13 次课。
- 两周一次作业, 周三 12: 00 之前交作业。在教学网提交电子版, 第三周开始交作业。

本节目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计
- 4 统计学与人工智能
- 5 教学安排
- 6 R 简介

什么是概率统计

- 概率统计是研究自然界中随机现象统计规律的数学方法，又称数理统计方法；
- 概率统计是应用概率的理论来研究大量随机现象的规律性；

<https://baike.baidu.com/item/概率统计/1486966>

什么是统计学

- Wiki: Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, presentation, and organization of data.
- 统计是一门研究数据的收集、分析、解释、呈现和组织的数学分支；

<https://en.wikipedia.org/wiki/Statistics>

统计学的产生与发展 (I)

- 统计学起源于收集数据的活动；比如人口普查或者抽样调查；英国学者葛朗特在 1662 年发表的著作《关于死亡公报的自然和政治观察》，标志着这门学科的诞生。
- 数理统计学的另一个重要源头来自天文和测地学中的误差分析问题。伽利略、拉普拉斯分布、高斯分布；

<https://baike.baidu.com/item/概率论与数理统计/36319>

统计学的产生与发展 (II)

- 19 世纪后期高尔顿、K · 皮尔逊等发展了统计相关和回归理论；
- 20 世纪上半叶费歇尔，K · 皮尔逊建立起统一的理论框架；奈曼与 E · 皮尔逊发展了假设检验理论；我们概率统计系的前辈许宝騄先生在奈曼-皮尔逊理论上也有很重要的贡献；瓦尔德发展了统计决策理论，建立了序贯分析理论；

概率论的发展历史 (I)

- 概率的概念起源于中世纪以来的欧洲流行的用骰子赌博；
- 分赌本问题：甲、乙二人赌博，各出赌注 30 元，共 60 元，每局甲、乙胜的机会均等，都是 $\frac{1}{2}$ 。约定：谁先胜满 3 局则他赢得全部赌注 60 元，现已赌完 3 局，甲 2 胜 1 负，而因故中断赌情，问这 60 元赌注该如何分给 2 人？
- 帕斯卡和费尔马建立了“数学期望”的概念，解决了“分赌本问题”；1657 年，惠更斯专著《论掷骰子游戏中的计算》计算古典概率；

<https://baike.baidu.com/item/概率论与数理统计/36319>

概率论的发展历史 (II)

- 贝努利家族: 雅可布·贝努利发现并证明了证明了大数定律; 1713 年其著作《猜度术》出版;
- 法国数学家拉普拉斯将古典概率论向近代概率论进行推进, 明确给出了概率的古典定义, 证明了“棣莫弗—拉普拉斯定理”, 1812 年出版了他的著作《分析的概率理论》;
- 1906 年, 俄国数学家马尔科夫提出了“马尔科夫链”的数学模型。1934 年, 前苏联数学家辛钦又提出一种在时间中均匀进行着的平稳过程理论。

<https://baike.baidu.com/item/概率论与数理统计/36319>

概率论的发展历史 (III)

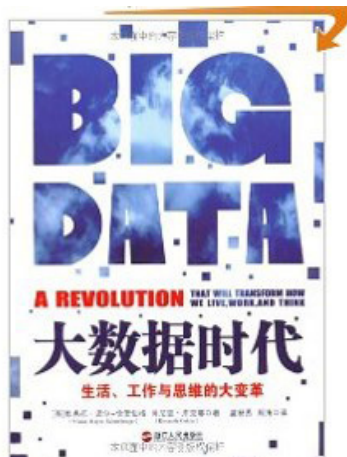
- 20 世纪初完成的勒贝格测度与积分理论及随后发展的抽象测度和积分理论，为概率公理体系的建立奠定了基础。在这种背景下柯尔莫哥洛夫 1933 年在他的《概率论基础》一书中首次给出了概率的测度论式定义和一套严密的公理体系。他的公理化方法成为现代概率论的基础，使概率论成为严谨的数学分支。
- 1946 年, S.M. 乌拉姆和 J. 冯·诺伊曼提出了蒙特卡罗方法。他们是在参与美国的“曼哈顿计划”提出了这一算法。蒙特卡罗方法是一种建立在概率论与数理统计基础上的计算方法，在多个学科的研究中起着非常重要的作用；

本节目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计**
- 4 统计学与人工智能
- 5 教学安排
- 6 R 简介

大数据时代

- 大数据时代 — 生活、工作与思维的大变革 (Big Data, A Revolution that will transform how we live, work and Think)



大数据实例

- 大数据变革公共卫生：Google 根据网络搜索预测流感发生，预测与官方数据相关性高达 97%.
- 大数据变革商业：FareCast 通过已有价格记录预测机票价格，预测准确度达到 70%. 为使用者平均每张机票节省 50 美元；
- 对冲基金通过分析 Twitter 数据信息预测股票表现；
- Amazon 和 Netflix 根据用户网络查询进行产品推荐。

流感预测 (I)

● Google flu Trends, Nature 文章

nature

Vol 457 | 19 February 2009 | doi:10.1038/nature07634

LETTERS

Detecting influenza epidemics using search engine query data

Jeremy Ginsberg¹, Matthew H. Mohebbi¹, Rajan S. Patel¹, Lynnette Brammer², Mark S. Smolinski¹ & Larry Brilliant¹

Table 1 | Topics found in search queries which were found to be most correlated with CDC ILI data

Search query topic	Top 45 queries		Next 55 queries	
	n	Weighted	n	Weighted
Influenza complication	11	18.15	5	3.40
Cold/flu remedy	8	5.05	6	5.03
General influenza symptoms	5	2.60	1	0.07
Term for influenza	4	3.74	6	0.30
Specific influenza symptom	4	2.54	6	3.74
Symptoms of an influenza complication	4	2.21	2	0.92
Antibiotic medication	3	6.23	3	3.17
General influenza remedies	2	0.18	1	0.32
Symptoms of a related disease	2	1.66	2	0.77
Antiviral medication	1	0.39	1	0.74
Related disease	1	6.66	3	3.77
Unrelated to influenza	0	0.00	19	28.37
Total	45	49.40	55	50.60

The top 45 queries were used in our final model; the next 55 queries are presented for comparison purposes. The number of queries in each topic is indicated, as well as query-volume-weighted counts, reflecting the relative frequency of queries in each topic.

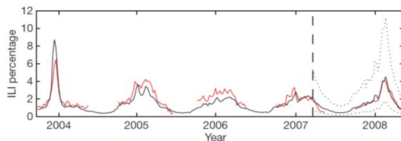


Figure 2 | A comparison of model estimates for the mid-Atlantic region (black) against CDC-reported ILI percentages (red), including points over which the model was fit and validated. A correlation of 0.85 was obtained over 128 points from this region to which the model was fit, whereas a correlation of 0.96 was obtained over 42 validation points. Dotted lines indicate 95% prediction intervals. The region comprises New York, New Jersey and Pennsylvania.

流感预测 (II)

● Google flu Trends 的改进, PNAS 文章



Accurate estimation of influenza epidemics using Google search data via ARGO

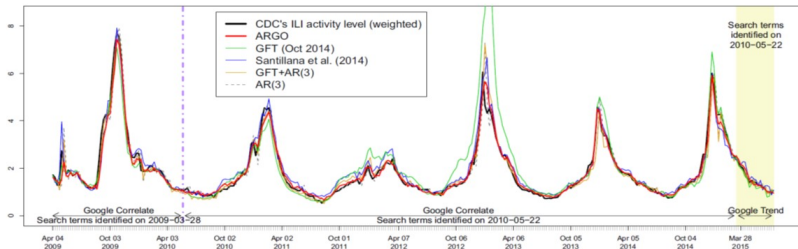
Shihao Yang^a, Mauricio Santillana^{b,c,1}, and S. C. Kou^{a,1}

^aDepartment of Statistics, Harvard University, Cambridge, MA 02138; ^bSchool of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and ^cComputational Health Informatics Program, Boston Children's Hospital, Boston, MA 02115

Edited by Wing Hung Wong, Stanford University, Stanford, CA, and approved September 30, 2015 (received for review August 6, 2015)

Accurate real-time tracking of influenza outbreaks helps public health officials make timely and meaningful decisions that could save lives. We propose an influenza tracking model, ARGO (AutoRegression with Google search data), that uses publicly available online search

CDC's ILI reports have a delay of 1–3wk due to the time for processing and aggregating clinical information. This time lag is far from optimal for decision-making purposes. To alleviate this information gap, multiple methods combining climate, demo-



什么是大数据

- 数据没有办法在可容忍的时间下使用常规软件方法完成存储、管理和处理任务
 - ① 大量 (Volume): 存储, 计算量
 - ② 多样 (Variety): 多来源, 多格式
 - ③ 快速 (Velocity): 增长速度, 处理速度要求
 - ④ 价值 (Value)

张文涛. 阿里巴巴研究中心: 大数据时代

<http://wenku.baidu.com/view/8d259ac40c22590102029d29.html>

数据来源

- 互联网企业：微博、视频网站、电子商务网站；
- 物联网、移动设备、终端中的商品、个人位置、传感器采集的数据；
- 联通、移动、电信等通信和互联网运营商；
- 天文望远镜拍摄的图像、视频数据、气象学里面的卫星云图数据；
- 生物测量数据 (高通量测序、图像)；
- 医学健康数据。

张文涛. 阿里巴巴研究中心：大数据时代

<http://wenku.baidu.com/view/8d259ac40c22590102029d29.html>

大数据分析技术

- 数据处理：自然语言处理技术
- 统计和分析：A/B test; top N 排行榜；地域占比；文本情感分析
- 数据挖掘：关联规则分析；分类；聚类
- 模型预测：预测模型；机器学习；建模仿真

张文涛. 阿里巴巴研究中心：大数据时代

<http://wenku.baidu.com/view/8d259ac40c22590102029d29.html>.

大数据技术

- 数据采集：ETL 工具 (extract、transform、load)
- 数据存取：关系数据库；NoSQL；SQL 等
- 基础架构支持：云存储；分布式文件系统等
- 计算结果展现：云计算；标签云；关系图等

张文涛. 阿里巴巴研究中心：大数据时代

<http://wenku.baidu.com/view/8d259ac40c22590102029d29.html>

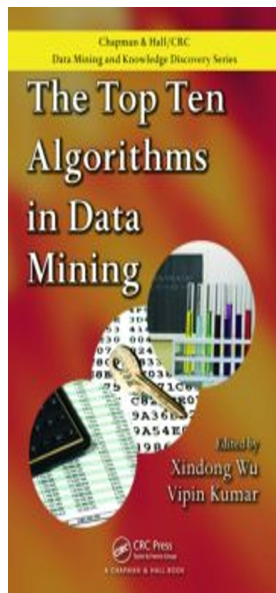
大数据方向

- 云计算是基础设施架构;
- 大数据是灵魂资产;
- 分析、挖掘是手段;
- 实现与预测是最终目标。

张文涛. 阿里巴巴研究中心: 大数据时代

<http://wenku.baidu.com/view/8d259ac40c22590102029d29.html>

数据挖掘 Top 10 算法



Knowl Inf Syst (2008) 14:1–37
DOI 10.1007/s10115-007-0114-2

SURVEY PAPER

Top 10 algorithms in data mining

Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang · Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu · Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg

Received: 9 July 2007 / Revised: 28 September 2007 / Accepted: 8 October 2007
Published online: 4 December 2007
© Springer-Verlag London Limited 2007

Abstract This paper presents the top 10 data mining algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006: C4.5, k -Means, SVM, Apriori, EM, PageRank, AdaBoost, k NN, Naive Bayes, and CART. These top 10 algorithms are among the most influential data mining algorithms in the research community. With each algorithm, we provide a description of the algorithm, discuss the impact of the algorithm, and review current and further research on the algorithm. These 10 algorithms cover classification,

数据挖掘 Top 10 算法

- C4.5 (Decision Tree);
- k-Means (Clustering);
- SVM (Classification, machine learning).
- Apriori (Frequent pattern finding, association rule mining);
- EM (Maximum likelihood estimation with missing data);
- PageRank (Ranking, Markov Chain);
- AdaBoost (Boosting classification accuracy);
- kNN (Classification);
- Naive Bayes (Classification);
- CART (Decision Tree, Classification and Regression Tree).

本节目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计
- 4 统计学与人工智能**
- 5 教学安排
- 6 R 简介

AlphaGO 系列

- AlphaGO: 2016 年 1 月 28 日 Nature, 2016 年 3 月 AlphaGO 对李世石 4:1 胜出.
- AlphaGOZero: 2017 年 10 月 19 日发表 Nature 文章, AlphaGO Zero 对战 AlphaGO 的战绩是 100:0.
- AlphaZero: 2018 年 12 月 7 日,《科学》杂志以封面文章发表了文章介绍了 AI 算法 AlphaZero. 同一个算法, 在三种棋上成为顶级高手.

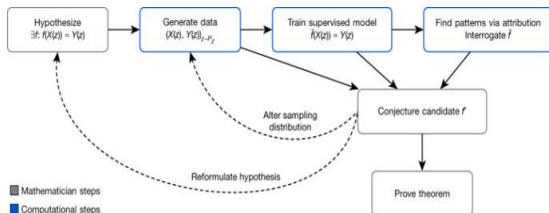


AlphaFold

- 2018 年 11 月 2 日，在第 13 届全球蛋白质结构预测竞赛（CASP）上，AlphaFold 获得了预测 43 种蛋白中的 25 种蛋白结构的最高分，在 98 名参赛者中排名第一。
- 2020 年 1 月 15 日，DeepMind 关于 AlphaFold 模型与代码通过了同行评审发表在《Nature》，并且模型和代码已经开源。
- 2020 年 12 月 1 日，AlphaFold2 在第 14 届 CASP 比赛中无可争议地夺冠，其预测接近了实验检测的精度；
- 2021 年 7 月 15 日，期待已久的 AlphaFold2 方法的配套论文在《Nature》发表，相关代码公布于世。一周后 AlphaFold2 再登《Nature》，预测了人类 98.5% 的蛋白质，预测数据免费开放。
- AlphaFold2: 2024 Nobel Prize for Chemistry for "revolutionary" work on proteins. John Jumper, Demis Hassabis, David Baker

AI 辅助数学定理证明

- 2021 年 12 月 1 日,《Nature》发文“Advancing Mathematics by Guiding Human Intuition with AI”, 论文的一作是来自 DeepMind 的 Alex Davies 博士。
- 论文采用一种机器学习模型, 来发现数学对象之间的潜在模式和关联, 用归因技术加以辅助理解, 并利用这些观察进一步指导直觉思维和提出猜想的过程。



<https://www.nature.com/articles/s41586-021-04086-x>

《新一代人工智能发展规划》

- 2017 年 7 月 20 日，国务院发布《新一代人工智能规划》(国发【2017】35 号)
 - a. 第一步，到 2020 年人工智能总体技术和应用与世界先进水平同步，人工智能产业成为新的重要经济增长点，人工智能技术应用成为改善民生的新途径，有力支撑进入创新型国家行列和实现全面建成小康社会的奋斗目标
 - b. 第二步，到 2025 年人工智能基础理论实现重大突破，部分技术与应用达到世界领先水平，人工智能成为带动我国产业升级和经济转型的主要动力，智能社会建设取得积极进展。
 - c. 第三步，到 2030 年人工智能理论、技术与应用总体达到世界领先水平，成为世界主要人工智能创新中心，智能经济、智能社会取得明显成效，为跻身创新型国家前列和经济强国奠定重要基础。

诺奖得主 Thomas J. Sargent 观点

- 2018 年 8 月 11 日诺贝尔经济学奖获得者 Thomas J. Sargent 在由厚益控股和《财经》杂志联合主办主题为「共享全球智慧引领未来科技」的世界科技创新论坛上表示：人工智能其实就是统计学，只不过用了一个很华丽的辞藻，其实就是统计学。好多的公式都非常老，但是所有的人工智能利用的都是统计学来解决问题。



<https://www.zhihu.com/question/289782814>

华为创始人任正非观点

- 2019 年 1 月 20 日，华为创始人兼 CEO 任正非 1 月 20 日在接受央视《面对面》采访时，谈及基础研究和基础学科的重要性时说，人工智能就是统计学，计算机与统计学就是人工智能。你说我们要进入大数据时代，大数据时代做啥？统计。说明我们国家在数学上重视不够，第二个在数学中的统计学重视不够。



<https://v.qq.com/x/cover/6gc6hfhj8u4wuchb/j0029cbs3er.html>

本节目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计
- 4 统计学与人工智能
- 5 教学安排**
- 6 R 简介

概率论的内容

- 随机事件与概率；
- 随机变量及其概率分布；
- 随机变量的数字特征；
- 随机向量其概率分布, 大数定律及中心极限定理。

数理统计的内容

- 统计估值;
- 假设检验;
- 回归分析方法.

教学要求

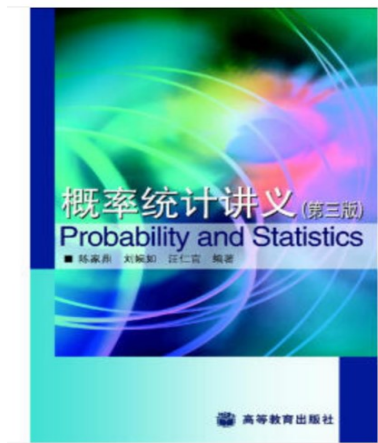
- 认真预习；
- 完成作业；
- 自己学习一种统计数据分析软件，建议学习 R。

统计软件

- SAS (商业软件);
- SPSS (社科方面使用广泛);
- Stat;
- S-plus;
- R。

教材

- 陈家鼎、刘婉如、汪仁官：概率统计讲义 (第三版), 高等教育出版社, 2004 (<https://item.jd.com/10337785.html>)。



参考书

- 陈家鼎，郑忠国，概率与统计，北京大学出版社，2007 年；
概率与统计(第二版) 概率分册，概率与统计(第二版) 统计分册，北京大学出版社，2016 年；
- 陈家鼎、孙山泽、李东风、刘力平，数理统计学讲义(第三版)，高等教育出版社，2015 年；
- 何书元，概率论与数理统计(第二版)，高等教育出版社，2013 年；
- 何书元《概率论》，北京大学出版社，2005 年。

参考书

- Sheldon M. Ross, 概率论基础教程 (A First Course in Probability), 郑忠国、詹从赞翻译, 人民邮电出版社, 2006 年;
- 李贤平, 概率论基础 (第三版), 高等教育出版社, 2010 年; 李贤平, 陈子毅, 概率论基础学习指导书, 高等教育出版社, 2011 年。
- 陈希孺, 概率论与数理统计, 中国科学技术大学出版社, 2009 年;
- 茆诗松, 程依明, 濮晓龙, 概率论与数理统计教程, 2004 年; 概率论与数理统计教程 (习题与解答), 2005 年; 概率论与数理统计教程 (第二版), 高等教育出版社, 2011 年;
- 盛骤, 谢式千, 潘承毅, 概率论与数理统计 (第四版), 高等教育出版社, 2010 年。

本节目录

- 1 课程信息
- 2 概率统计简介
- 3 大数据呼唤统计
- 4 统计学与人工智能
- 5 教学安排
- 6 R 简介

声明

- 部分内容从丁国徽“R 导论”改编 (<http://www.r-user.com/R-guide/>)
- 部分内容来自李东风“S 语言入门” (<http://www.math.pku.edu.cn/teachers/lidf/docs/statsoft/html/s/index.html>)

内容提纲

- R 历史;
- 什么是 R?
- R 统计分析;
- R 编程。

R 开发者

- R 于上世纪 90 年代
- 开发者: Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand.



Ross Ihaka



Robert Gentleman

R 的前身—S 语言

- S 于 1976 由贝尔实验室 (AT&T) 开发。
- 后来发展成为商业软件- “S-plus”, AT&TLucent to Insightful Coporation.

Version 1	1976-1980	Honeywell GCOS, Fortran-based
Version 2	1980-1988	Unix; Macros, Interface Language
	1981-1986	QPE (Quantitative Programming Environment)
	1984-	General Outside Licensing, books
Version 3	1988-1998	C-based; S functions & objects
	1991; rev. 1992	Statistical Models; informal Classes & Methods
	1993-	Exclusive license to StatSci (later MathSoft) for S-Plus.
Version 4	1998-	Formal class-method model; Connections; Large Objects
	1999-	Interfaces to Java, Corba?

R Project

- 网站: www.r-project.org
- R core group, 1997.
- CRAN (The comprehensive R archive network)



[\[Home\]](#)

Download

[CRAN](#)

R Project

[About R](#)

[Logo](#)

[Contributors](#)

[What's New?](#)

[Reporting Bugs](#)

[Conferences](#)

[Search](#)

[Get Involved: Mailing Lists](#)

[Developer Pages](#)

[R Blog](#)

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To [download R](#), please choose your preferred [CRAN mirror](#).

If you have questions about R like how to download and install the software, or what the license terms are, please read our [answers to frequently asked questions](#) before you send an email.

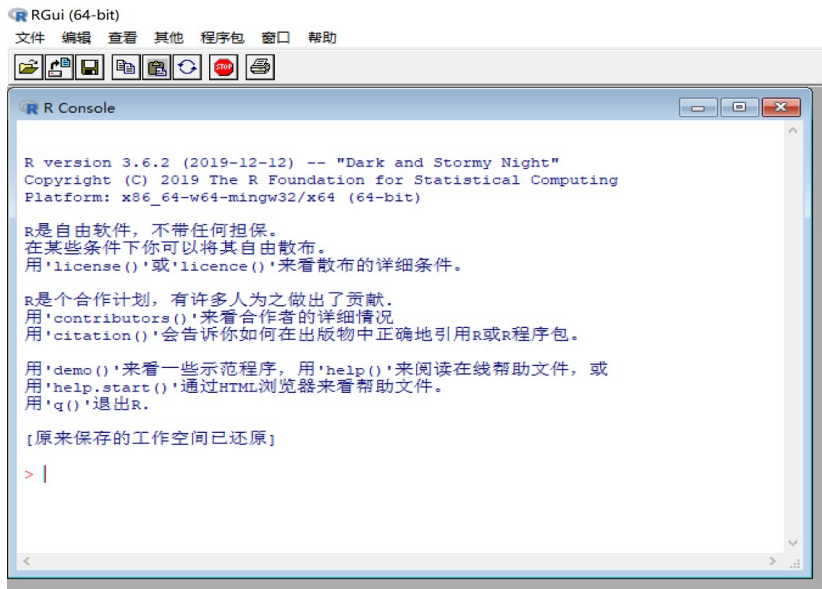
News

- [R version 4.0.3 \(Bunny-Wunnies Freak Out\)](#) has been released on 2020-10-10.
- Thanks to the organisers of useR! 2020 for a successful online conference. Recorded tutorials and talks from the conference are available on the [R Consortium YouTube channel](#).
- [R version 3.6.3 \(Holding the Windsock\)](#) was released on 2020-02-29.
- You can support the R Foundation with a renewable subscription as a [supporting member](#)

R 特点

- 免费自由软件。
- 丰富的图形，图形可以直接输出到 Latex.
- 可编程，自己开发。
- 多种软件包 (External Packages)，软件包多由著名的统计学家开发。许多新的统计分析方法都可以从 External Packages 中找到。

R 运行界面



R 数据类型和数据结构

- 基本的数据对象是“向量”。
- 向量元素可以有多种类型
 - ① 数值型 (numeric)
 - ② 字符型 (Character)
 - ③ 逻辑型 (Logic)
- 组合数据结构: Array, Matrix, List.
- 缺失值: NA, is.na().

R 运算符

Operators					
Arithmetic		Comparison		Logical	
+	addition	<	lesser than	! x	logical NOT
-	subtraction	>	greater than	x & y	logical AND
*	multiplication	<=	lesser than or equal to	x && y	id.
/	division	>=	greater than or equal to	x y	logical OR
^	power	==	equal	x y	id.
%%	modulo	!=	different	xor(x, y)	exclusive OR
%/%	integer division				

¹⁰The following characters are also operators for R: \$, [, [[, :, ?, <-,

向量

- 向量属性: `length()`, `mode()`.
- 赋值

```
1 > students <- c("张三","李四","王五")
2 > students
3 [1] "张三" "李四" "王五"
4 > x<-c(1:10)
5 > x
6 [1] 1 2 3 4 5 6 7 8 9 10
7 >
```

- 运算
 - ① 简单运算: `sin()`, `sqrt()`, `tan()`, `cos()` 等。
 - ② 简单统计: `max()`, `min()`, `mean()`, `var()` 等。

矩阵和数组

- 命令 `array(data = NA, dim = length(data), dimnames = NULL)`
- 命令 `matrix(data = NA, nrow = 1, ncol = 1, byrow = FALSE, dimnames = NULL)`
- 运算：简单四则运算，矩阵乘法等。
- 缺省顺序是“列优先”方式。

矩阵排列顺序

```
1 > x<-c(1:20)
2 > dim(x)<-c(4,5)
3 > x
4 [,1] [,2] [,3] [,4] [,5]
5 [1,]    1    5    9   13   17
6 [2,]    2    6   10   14   18
7 [3,]    3    7   11   15   19
8 [4,]    4    8   12   16   20
9 > y<-matrix(x,nrow=4,ncol=5,byrow=TRUE)
10 > y
11 [,1] [,2] [,3] [,4] [,5]
12 [1,]    1    2    3    4    5
13 [2,]    6    7    8    9   10
14 [3,]   11   12   13   14   15
15 [4,]   16   17   18   19   20
16 >
```

矩阵和数组操作

- 矩阵乘法: `%*%`
- 矩阵除法: 需要借助于方程求解, QR 分解等操作。 `solve(A,b)`, `qr()`, `eigen()`.
- 矩阵合并: `cbind()`, `rbind()`
 - ① `cbind()` 作为列向量合并
 - ② `rbind()` 作为行向量合并
- 一个非常有用的函数 `apply(x,margin, fun)`: 对 `x` 的某维 (`margin`) 进行 `fun` 函数操作。

列表(List)

- List 是一种特殊的向量，向量中元素类型可以不一样。一般的向量必须是同一类型。
- 下面的两个式子是不一样的：

```
1 > x<-list("张三",86,TRUE)
2 > y<-c("张三",86,TRUE)
3 > x[2]
4 [[1]]
5 [1] 86
6      ^^ I
7 > y[2]
8 [1] "86"
9 >
```

因子 (Factor)

- 处理统计中的名义变量或者有序变量。
- 命令 `factor(x, levels = sort(unique(x), na.last = TRUE), labels, exclude = NA, ordered = FALSE)`
- 对因子可以采用 `table()` 进行频数统计。

```
1 > x<-c("男","男","女","女","男")
2 > y<-factor(x)
3 > y
4 [1] 男 男 女 女 男
5 Levels: 男 女
6 > table(y)
7 y
8 男 女
9 3  2
10 >
```


数据框架 (Data Frame)

- 数据框架是一种特殊的列表，不同的列可以有不同的数据类型。
- 数据框的每列是变量，行是观测。每列可以给出名字，该名字可以作为变量名引用。这种数据结构方便于统计建模。

```
1 >d<-data.frame(name=c("李明","张建","王聪"),
2   + age=c(30,35,28),height=c(180,162,175))
3 > d
4 name age height
5 1 李明 30      180
6 2 张建 35      162
7 3 王聪 28      175
8 >
```

- 数据框架和变量名的连接 attach(), detach().

数据输入

- 从键盘输入 `scan()`
- 从文件读入: `read.table()`; `read.csv()`; `read.dta()`;
- 通过表单进行数据编辑:
 - ① `data.entry()`
 - ② `edit()`
 - ③ `fix()`

数据存储

- 每个 R 对象都可以存储和加载。

```
1 > save(x, file = "x.Rdata")  
2 > load("x.Rdata")
```

- 格式化输出

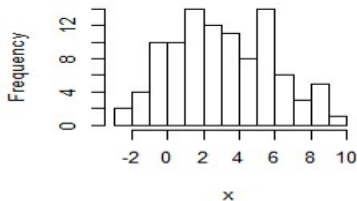
```
1 > write.table(x, file = "x.txt", sep = "\t")
```

描述性统计量

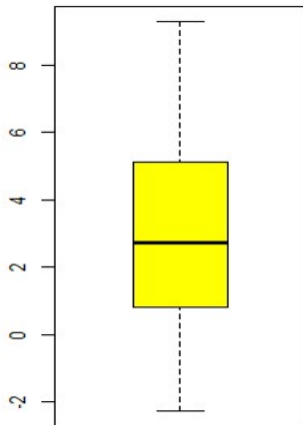
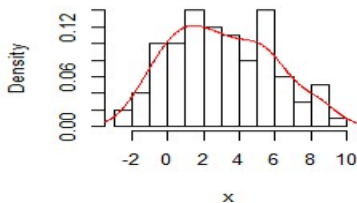
```
1 > x<-rnorm(100,3,3)
2 > mean(x)
3 [1] 3.010877
4 > sd(x)
5 [1] 2.803744
6 > summary(x)
7 Min. 1st Qu. Median Mean 3rd Qu. Max.
8 -2.2626 0.8385 2.7122 3.0109 5.1098 9.2760
9 > layout(matrix(c(1,2,3,3),2,2))
10 > hist(x, breaks=12, main=Histogram)
11 > hist(x, freq=FALSE, breaks=12, main=Desity Plot)
12 > lines(density(x), col=red)
13 > boxplot(x, col=yellow)
14 >
```

描述性统计量

Histogram



Density Plot



常用概率分布

law	function
Gaussian (normal)	<code>rnorm(n, mean=0, sd=1)</code>
exponential	<code>rexp(n, rate=1)</code>
gamma	<code>rgamma(n, shape, scale=1)</code>
Poisson	<code>rpois(n, lambda)</code>
Weibull	<code>rweibull(n, shape, scale=1)</code>
Cauchy	<code>rcauchy(n, location=0, scale=1)</code>
beta	<code>rbeta(n, shape1, shape2)</code>
'Student' (t)	<code>rt(n, df)</code>
Fisher-Snedecor (F)	<code>rf(n, df1, df2)</code>
Pearson (χ^2)	<code>rchisq(n, df)</code>
binomial	<code>rbinom(n, size, prob)</code>
geometric	<code>rgeom(n, prob)</code>
hypergeometric	<code>rhyper(nn, m, n, k)</code>
logistic	<code>rlogis(n, location=0, scale=1)</code>
lognormal	<code>rlnorm(n, meanlog=0, sdlog=1)</code>
negative binomial	<code>rnbinom(n, size, prob)</code>
uniform	<code>runif(n, min=0, max=1)</code>
Wilcoxon's statistics	<code>rwilcox(nn, m, n), rsignrank(nn, n)</code>

概率函数前缀

- 密度函数: dxxx
- 累计分布函数: pxxx
- 分位数函数: qxxx
- 模拟随机变量: rxxx

统计推断

- `qxxx ()`
- `t.test()`
- `wilcox.test(stats)`
- `kruskal.test(stats)`
- `var.test();`
- `shapiro.test();`
- `qqnorm();`
- `anova();lm();`

回归分析

```
> x<-rnorm(20,2,9)
> y<-rnorm(20,x,4)
> xylm<-lm(y~x)
> plot(x,y)
> abline(xylm)
> summary(xylm)
Call:
lm(formula = y ~ x)

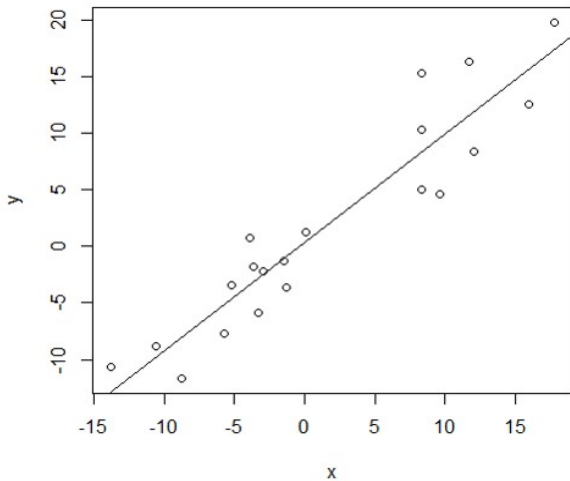
Residuals:
Min       1Q   Median       3Q      Max
-4.9760  -3.0802   0.5903   2.0069   7.0204

Coefficients:
(Intercept)  0.33964      0.76276      0.445      0.661
x            0.95941      0.08429     11.382  1.18e-09 ***

-----
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.359 on 18 degrees of freedom
Multiple R-squared:  0.878, Adjusted R-squared:  0.8712
F-statistic: 129.5 on 1 and 18 DF,  p-value: 1.177e-09
```

回归分析



R 图形功能

- 高级图形函数

- ① `plot()`, `boxplot()`, `hist()`, $\dots$.

- 低级图形函数

- ① `line()`, `point()`, `text()`, `legend()`, `title()`, `axis()`,

- ② `abline()`, $\dots$.

- 交互式图形函数

- ① `locator()`, `identify()`.

控制语句

- `if(cond) expr`
- `if(cond) cons.expr else alt.expr`
- `for(var in seq) expr`
- `while(cond) expr`
- `repeat expr`
- `break`
- `next`

循环语句 (I)

- 和一般程序语言类似, 采用 for 语句
 - ① `for(i in 1:n)`
 - ② `for(i in c(1:10))`
- 也可用 while 循环语句, 然后跳出循环
 - ① `break` 终止循环, 从循环中跳出
 - ② `next` 终止本次循环, 跳到下一步循环
- 应当避免明显的循环语句 (解释性语言), 而使用系统自带的向量处理函数代替。如 `apply()`, `tapply()`, `sapply()` 等。

循环语句 (II)

- ”Bad” 循环

```
1  >## multiply two vectors
2  >for(i in 1:n) {
3    >   d[i] <- a[i] * b[i]
4  >}
5  >##compute the inner product
6  >s <- 0
7  >for (i in 1:n){
8    >   s <- s + d[i]
9  >}
10 >
```

- ”good” 循环

```
1  >s<-sum(a*b)
```

自定义函数

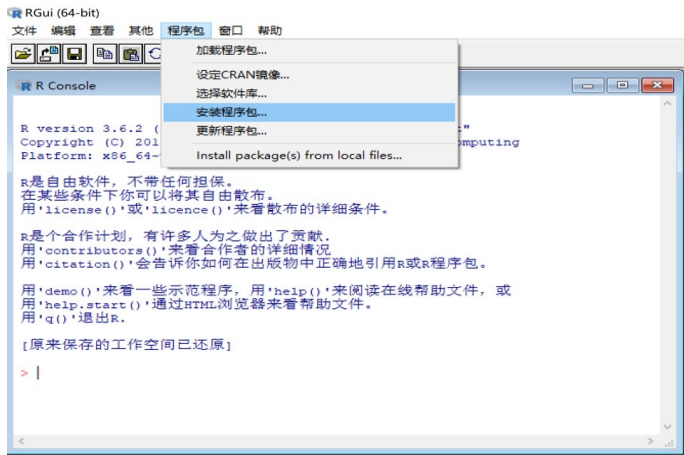
- 函数定义方法

```
1  funcname <- function(x, na.rm = FALSE)
2  {
3    ... lots of code ...
4    ## a return value
5  }
```

- 自定义函数调用：用 funcname() 调用函数。

调入软件包

- library() 调入软件包，软件包库函数；
- detach(package: xxx) 卸载软件包；
- 可以在 R 主菜单下载入软件包。



R 网上资源

- A Course (The ppt is showed with R Development Core Group)
<http://faculty.washington.edu/tlumley/Rcourse/>
- A Paper (citing R in a publication): Ihaka R. & Gentleman R. 1996. R: a language for data analysis and graphics. Journal of Computational and Graphical Statistics 5: 299–314.
- Several Books
 - ① Using R for Data Analysis and Graphics—An Introduction. J.H. Maindonald.
 - ② An Introduction to R. The R Development Core Team.
 - ③ simpleR –Using R for Introductory Statistics. John Verzani
 - ④ R for Beginners. Emmanuel Paradis
 - ⑤ The R Reference Manual Base Package. The R Development Core Team