21世纪统计学规划教材

Statistics: An Introduction Using R

统计学导论
——基于R语言

李勇　金蛟　编著

北京大学出版社
PEKING UNIVERSITY PRESS

21世纪统计学规划教材

Statistics: An Introduction Using R

统计学导论
——基于R语言

李勇　金蛟　编著

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目 (CIP) 数据

统计学导论：基于 R 语言 / 李勇，金蛟编著. — 北京：北京大学出版社，2016.9
(21 世纪统计学规划教材)
ISBN 978-7-301-27472-9

Ⅰ.①统… Ⅱ.①李… ②金… Ⅲ.①统计学—高等学校—教材 Ⅳ.①C8

中国版本图书馆 CIP 数据核字 (2016) 第 205405 号

书　　名	统计学导论——基于 R 语言 TONGJIXUE DAOLUN
著作责任者	李勇　金蛟　编著
责任编辑	曾琬婷
标准书号	ISBN 978-7-301-27472-9
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http://www.pup.cn　新浪微博：@ 北京大学出版社
电子信箱	zpup@pup.cn
电　　话	邮购部 62752015　发行部 62750672　编辑部 62767347
印 刷 者	北京大学印刷厂
经 销 者	新华书店
	787 毫米 × 980 毫米　16 开本　15.75 印张　328 千字 2016 年 9 月第 1 版　2016 年 9 月第 1 次印刷
定　　价	36.00 元

“21世纪统计学规划教材”
编委会

内 容 简 介

本书主要介绍统计学的基本思想、原理和方法，使读者对统计学及统计学的思维方式有一个整体的了解. 本书主要内容包括：统计学的发展和应用领域、概率理论、数据收集的概念和方法、对数据总体信息的描述、常用的参数估计和假设检验方法. 书中注重以概率理论解释常见统计方法的原理，并通过计算机模拟帮助读者理解统计思想和原理，以避免把统计学片面地理解为简单的加减乘除计算公式，进而增强学生运用统计思想和方法提出问题、分析问题和解决问题的能力.

本书适合作为高等院校本科生学习统计学知识的入门教材.

前　言

统计学是通过收集数据和分析数据来认识未知现象的一门科学, 它在政府管理、工业、农业、林业、商业、教育、军事、自然科学和社会科学等领域有广泛的应用. 在大数据时代, 统计学的基本思想和方法成为人们日常学习、工作和生活的必备素养.

为了提高北京师范大学本科生的统计学素养水平, 笔者于 2014 年底提出了在文科和理科分别开设 "统计学导论" 通识课的设想, 得到北京师范大学教务管理部门以及文、理科各学院的支持, 并于 2015—2016 学年分别为文科和理科讲授该通识课.

笔者以北京高等教育精品教材《统计学导论》为基础, 结合国家级精品课程、国家级精品资源共享课和国家级精品视频公开课 "统计学导论" 建设过程中的教学实践经验, 通过 R 语言中的函数替代繁杂的数学推导, 降低对于学生先修课要求, 完成了适合文科学生使用的讲义《统计学导论 B》, 并用于 2015 年秋季学期文科学生的 "统计学导论" 通识课的教学实践.

笔者在讲义《统计学导论 B》基础上, 结合文科 "统计学导论" 通识课主讲教师和听课学生的意见, 完成了本教材. 本教材起点低, 适合具有高中数学知识背景的读者作为统计学入门读物, 也可以作为普通高等院校的文、理科各专业的统计学入门教材.

本教材的目的是: 介绍统计学的基本思想、原理和方法, 使读者对统计学及统计学的思维方式有一个整体的了解, 以便他们能够以统计学的视角看待日常学习、工作和生活中的问题, 并能在统计思想指导下利用 R 软件解决简单的统计学问题; 当读者以后遇见专业领域的复杂统计问题时, 能够想到与统计学家合作, 各取所长高效率地解决问题. 为此, 在教学过程中, 主讲教师应将统计学研究流程框架应用到各个知识和方法的讲授过程中, 讲清这些知识和方法的来龙去脉, 使学生认识问题背景分析的重要性, 培养学生统计创新的能力和跨学科合作的能力.

与国内外 "统计学导论" 的教科书相比, 本教材注重以概率理论解释常见统计方法的原理, 并通过计算机模拟帮助读者理解概率统计思想和原理, 以避免读者把统计学片面地理解为简单的加减乘除计算公式. 本教材将著名的 R 软件融合在各章节之中, 一方面, 使学生能够通过计算机来模拟简单的随机模型, 进而比较不同的统计方法的特性; 另一方面, 还可以使学生借助于计算机和 R 软件将所学的统计思想、原理和方法应用于解决实际问题.

本教材由五章组成, 第一章介绍统计学的基本概念、研究流程和思维模式, 使读者对于统计学及其研究流程和特点有一个概括的认识; 第二章介绍必要的概率论知识, 为理解统计学基本方法和原理奠定理论基础; 第三章介绍收集数据的方法, 使读者理解收集数据方法的

原理; 第四章介绍数据的描述性统计分析方法, 为进一步统计建模奠定基础; 第五章介绍参数估计效果的衡量指标和基本的统计学分析方法, 使读者理解点估计、假设检验和线性回归模型的基本原理.

借此机会, 笔者感谢张淑梅教授、陈梦根教授、杜勇宏副教授和李慧博士对于本教材提出的宝贵修改意见; 感谢黄文贤博士、张娟博士、徐祥灿硕士、徐帅帅硕士、张雪晴硕士、崔星宇硕士对于 R 语言程序代码和文字的校对工作; 感谢北京大学出版社曾琬婷编辑对本教材的精心校对和编辑. 由于时间仓促, 书中恐有不少谬误, 恳请读者指正, 以便修改.

李勇　金蛟

2016 年 5 月于北京师范大学

目　　录

第一章 绪 论

在自然界和人类社会中, 存在着大量的未知现象需要探索. 例如, 人们想要读懂遗传天书 —— 基因序列 (由 A, C, G, T 构成); 执政者想要知道国家经济运行是否正常; 火车客运计划者希望知道下一个春运高峰的客流量分布; 生产管理者想要知道生产线是否在正常工作; 保险公司想要知道各种灾害的分布情况; 药厂想要知道新研制的药品是否更有效; 人们想要知道什么样的饮食习惯更有利于健康, 吸烟与患肺癌之间的关系如何, 某减肥产品是否像其广告声称的那样有效率为 75%, 明天是否下雨; 等等.

§1.1 未知现象的认识过程与统计学

在客观世界中, 需要认识的现象无穷无尽. 人类对未知现象的探索经历了神化、定性分析和定量研究的阶段.

案例 1.1 未来的天气预报问题.

(1) 神化: 古时人们不能解释天气的变化, 把它归结为神的支配. 在此观点下, 人们采用祭祀神的方式求风调雨顺, 但是大多不如愿.

(2) 谚语: 关于气压、湿度、云、雨、冰雹等气象要素定性观测的经验总结, 如 "天上钩钩云, 地下雨淋淋" "朝霞不出门, 晚霞行千里" "东虹日头, 西虹雨" 等. 诸葛亮成功将谚语用于大雾预报, 完成 "草船借箭" 的壮举.

(3) 当地定量观测预报: 用当地天气测量数据预报未来天气. 从 17 世纪起, 科学仪器 (如气压表) 的发明使得人们可定量测量天气状态, 并将这些数据用于当地天气预报.

(4) 台网定量观测预报: 用全球气象观测网数据预报未来天气. 电报的发明使得人们能够远距离交换信息, 开始使用气象观测网数据做天气预报. 计算机硬件、互联网和数据分析技术的发展, 使得现代数值天气预报的精确度越来越高, 并可以通过技术手段干预天气, 如人工降雨、除冰雹等.

1.1.1 与天气预报案例相关的概念

未知现象 (不确定现象) 在特定条件下, 不能预知结果的现象.

必然现象 在特定条件下, 能预知结果的现象.

例如, 今天不能预知明天是什么样的天气, 因此明天的天气是未知现象; 又如, 在标准大

气压下, 将纯净的水加热到 100 °C 必然沸腾, 这是一个必然现象.

定义 1.1.1 **统计学**是通过收集数据和分析数据来认识未知现象的一门科学.

在此定义中包含了: 统计学的研究对象为 "未知现象" ; 研究的途径为 "收集数据和分析数据", 其中 "收集数据" 是指要科学地收集数据, "分析数据" 是指要科学地分析数据. 这里数据不仅包含定量数据, 也包含属性数据.

《大英百科全书》中指出: "统计学是一种收集数据、分析数据, 并根据数据进行推断的艺术和科学." 其内涵与上述定义相同.

1.1.2 统计学研究流程

如今, 人们主要是利用统计学原理来认识未知现象. 天气预报案例的研究流程如图 1.1 所示. 这种认识过程是通过如下三个步骤的循环逐步实现的:

(1) 明确研究问题, 通过观察或实验获取必要的观测资料;

(2) 通过分析所得资料推断未知结果;

(3) 通过实践检验推断结果, 寻找下一步研究的方向.

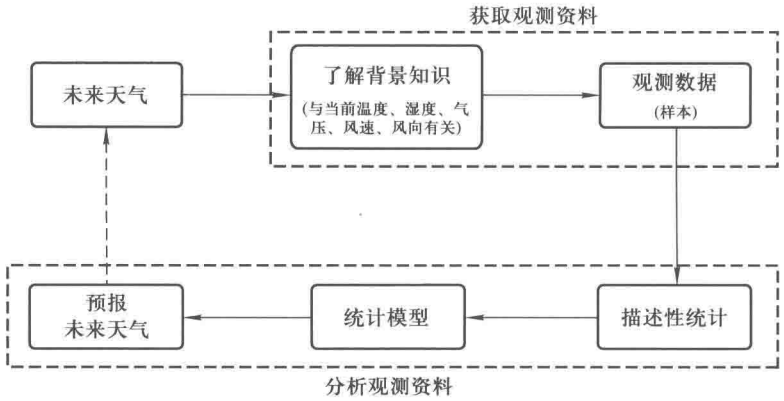

图 1.1 未来天气变化的研究流程图

这里第三步尤为重要, 如果不用实践检验而欣赏已有研究结果, 就是迷信. 例如, 在未来天气预报案例 1.1 中, 坚信 "未来天气由神支配" 是迷信; 坚信 "谚语预报" 是迷信; 坚信 "当地定量观测预报" 也是迷信; 坚信 "目前的天气预报方法" 同样是迷信. 迷信是阻碍未知现象研究的绊脚石.

一般地, 统计学认识未知现象的研究流程如图 1.2 所示. 统计学通过不断重复这一过程, 逐步认识未知现象. 例如, 在案例 1.1 中的谚语预报、当地定量观测预报、台网定量观测预报等研究, 都是沿着流程图进行工作的. 前后两研究流程区别在于背景知识的不同, 后一流程

的背景知识中, 包含了前一研究流程的研究结果, 起点更高; 而后一研究流程的研究将得到关于天气预报问题的新认识.

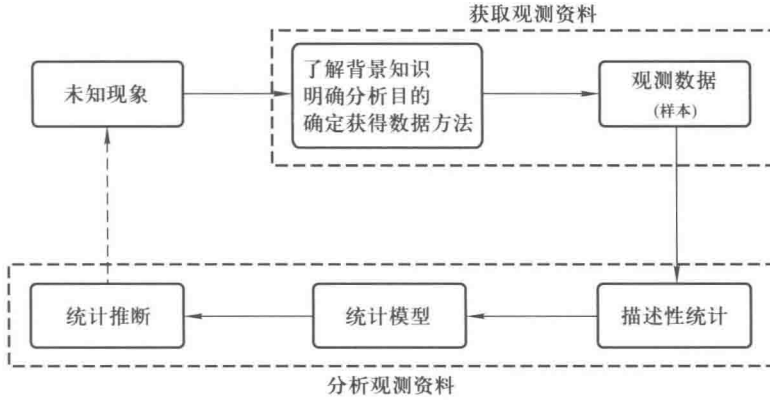

图 1.2 统计学研究流程图

例 1.1.1 了解地震的孕育过程, 是人类梦寐以求的愿望. 为达此愿, 应该收集哪些数据, 如何收集这些数据?

解 要了解未来地震发生的时间、地点和震级变化规律, 当然应该收集与这三个量有关的数据: 地震目录、地形变和地下水位等. 由于地震是一种自然现象, 只能通过观测的方式来收集有关的数据. ■

例 1.1.2 一个人在市场上买了一篮表面上看起来十分新鲜的水果, 回到家详细检查后才发现篮中有许多烂水果, 这里的分析结果为什么和实际情况不相符合?

解 这是因为, 获取观测资料的方法有问题, 这些观测资料不能很好地代表篮中的所有水果, 而在分析观测数据的过程中没有注意到这个问题, 使用的分析方法不得当, 使得分析结果与实际不符. ■

统计学常识

(1) 要用科学的方法收集与所研究未知现象有关的数据;

(2) 需要根据数据的背景特点 (获取过程、数据质量、数据分布特点) 调整分析方法, 以取得更好的分析结果.

例 1.1.3 摇奖机是否公平是亿万彩民关心的问题. 请设计一个检验摇奖机是否公平摇奖的方案 (流程).

解题思路 (1) 分析问题背景, 提出问题量化指标: 要想考查一台摇奖机是否公平摇奖, 首先要了解公平的含义, 确定用什么样的量来刻画公平;

(2) 如何获取数据: 确定量化指标和什么量有关系以及获取这些量的观测数据的方法;

(3) 如何从收集到的数据中提取量化指标信息, 并给出问题的解答.

解 这里 "公平" 是指摇出每个号码的概率相等. 根据频率稳定于概率的思想, 可以按如下方案检验摇奖机的公平性:

(1) 将带号码的球装入摇奖机后, 开动摇奖机摇出一个球, 记录下该球的号码.

(2) 重复第一步, 考查摇出的各个号码的频率的稳定性. 当所有号码的频率都比较稳定之后, 再进行下一步.

(3) 考查各个号码的频率的接近程度, 以此来判断这台摇奖机的公平性. ■

1.1.3 随机现象

案例 1.2 投掷一枚质地均匀的硬币, 结果不是正面向上就是反面向上. 随着投掷次数的增加, 却呈现出如下规律: 正面向上的比例接近于 0.5 (见表 1-1).

表 1-1 投掷硬币的实验结果

实验者	抛硬币次数	出现正面的次数	出现正面的频率
蒲丰	4040	2048	0.5069
德莫根	4092	2048	0.5005
费勒	10000	4979	0.4979
皮尔逊	12000	6019	0.5016
皮尔逊	24000	12012	0.5005
罗曼诺夫斯基	80640	39699	0.4932

定义 1.1.2 像这类在特定条件下不能事先预知结果且各个结果都具有频率稳定性的现象称为**随机现象**.

这里的结果包括现象所有可能出现的结果. 如所考查的现象为降雨量 w, 则 $w = 0$ 和 $w \in [0, 0.5]$ 均为此现象可能出现的结果. 结果的频率稳定性: 该结果出现的次数与观测总数之比随实验次数的增加而趋于稳定. 这里随机现象与一般书中的定义不同, 读者可仔细体会.

例 1.1.4 假定我们关心北京市一月份的平均气温, 试讨论它与随机现象间的关系.

解 "北京市" 和 "一月份" 的特定条件下, 不能预知这个平均气温, 它是一个不确定现象. 考虑到温室效应气体的作用, 它不是一个随机现象.

考虑到温室效应气体增加的因素, 北京市一月份的平均气温是由确定因素和随机因素共同决定的复合现象. ■

§1.2 描述未知现象的理想模型与现实模型

未知现象是众多因素作用的结果, 这些因素可分为两类: 确定因素和随机因素.

确定因素 能够事先确定的因素.

随机因素 不能事先确定, 但具有频率稳定性的因素.

在投掷硬币的例子 (案例 1.2) 中, "硬币质地均匀" 是确定因素; 在北京平均气温的例子 (例 1.1.4) 中, "北京" 和 "一月份" 是确定因素. 在天气预报的例子 (案例 1.1) 中, 未来天气与当前温度 x_1, 湿度 x_2, 气压 x_3, 风速 x_4, 风向 x_5 等确定因素有关, 还与其他的随机因素 ε 有关, 如图 1.3 所示. 如果已经知道所有的这些因素, 未来的天气就完全定下来了, 即未来天气可以用如下的统计学模型来描述:

$$y = f(x_1, x_2, x_3, x_4, x_5, \cdots, x_m) + \varepsilon, \tag{1.1}$$

其中 f 描述了所有确定因素对于未来天气的影响方式, 在数学上称之为函数, 而 ε 是影响未来天气的随机因素 (在给定确定因素的条件下). 统计学中我们的目的就是研究影响未来天气的所有确定因素是什么, 随机因素的变化规律是什么.

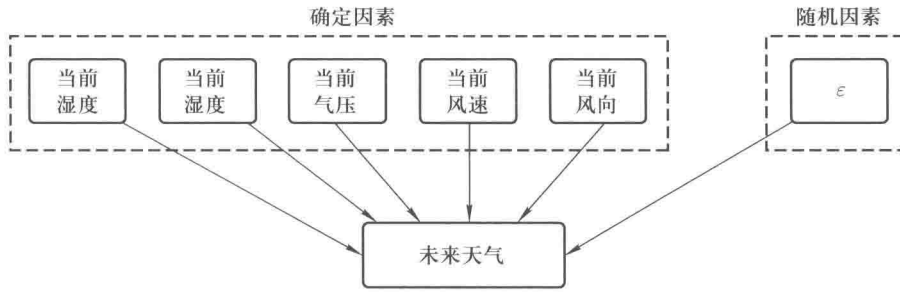

图 1.3 未来天气变化规律

虽然对于未来天气的研究已经获得丰硕的成果, 但这只是万里长征的第一步, 远未达到认识该现象的程度. 我们甚至还不能确定影响未来天气的确定因素到底有多少个, 随机因素的变化规律是什么, 更谈不上 f 的结构表达式. 当前关于未来天气的研究结果都是基于已知的影响天气的确定因素所建立的近似模型

$$y = \hat{f}(x_1, x_2, x_3, x_4, x_5, \cdots, x_k) + \eta, \tag{1.2}$$

其中 k 是 m 的近似, $\hat{f}$ 是 f 的近似, 而

$$\eta = y - \hat{f}(x_1, x_2, x_3, x_4, x_5, \cdots, x_k)$$

是 ε 的近似. 随着研究过程的推进, 模型 (1.2) 的近似程度不断提高, 最终会完全认识未来天气, 即得到模型 (1.1).

在研究的过程中, 模型 (1.1) 是追求的目标, 是我们的理想, 因此把这类描述未知现象的模型称为**理想模型**, 而把研究过程中所建立的那些形如 (1.2) 式的模型称为**现实模型**或**统计模型**, 简称为**模型**.

我们现在认为案例 1.1 中的 "神话" 研究阶段是迷信, 因为未来的天气不是由神来支配的. 同样, 在未知现象研究过程中, 认为现实模型无懈可击也是**迷信**, 因为它仅是理想模型的近似. 只有破除迷信, 才能逐步认识未知现象, 并在模型改进研究过程中推动统计学的发展.

按照统计学研究流程 (图 1.2), 在一个研究循环过程中, 我们是依据已经获得的研究背景知识, 通过收集和分析数据来构建更好的现实模型, 以更加深刻地认识未知现象. 统计学就是通过不断地建立新的现实模型来逐步认识未知现象的.

例 1.2.1 *考查模型 $y = \sin x + e$ 所产生数据的特点.*

解 (1) 在已知 x 和 e 的条件下, y 的值被完全确定, y 是一个必然现象. 此时, y 是 (x, e) 的二元函数, 图 1.4 给出了该函数的图像.

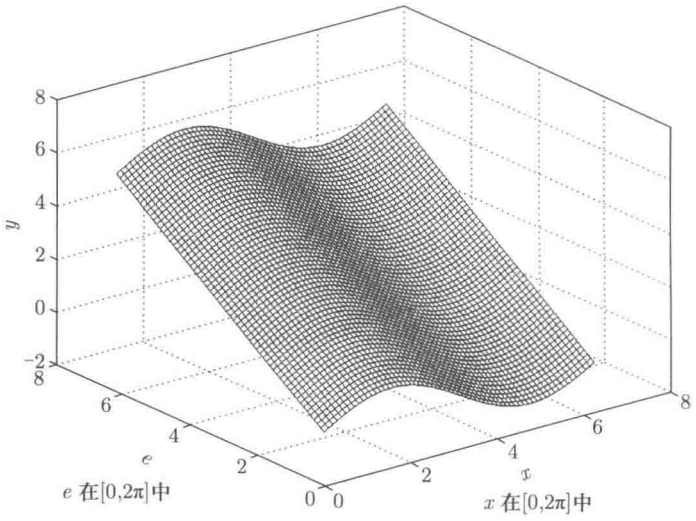

图 1.4 $y = \sin x + e$ 完全由 x 和 e 确定

例如, 当 $x = 0, e = 0.5$ 时, 可以用如下的 R 程序代码模拟 y 的观测值:

$$\texttt{x = 0; e = 0.5; y = sin(x) + e;}$$

模拟的结果永远是 0.5.

(2) 在仅知 x 的条件下, 不能预知 y, 因此 y 是不确定现象. 下面在仅知 $x = 0$ 的情况下,

讨论 y 的观测值变化情况.

① 当 e 是以 0.5 的可能性取 1, 以 0.5 的可能性取 0 的时候, y 是一个随机现象. 可以用如下的 R 程序代码模拟 y 的 16 次观测结果:

```
n=16;x=0;
e=sample(0:1,n,replace=T);
y=sin(x)+e;
y
```

上述程序代码运行后, 模拟的 y 在 RStudio 或 R 软件的控制台窗口显示如下:

```
[1] 1 1 0 1 0 1 1 1 1 0 0 1 0 1 1 0
```

注意, 将如上述程序代码再运行一次, 得到的结果多半不同, 其原因是 e 为未知的随机因素.

② 当 e 是以某种规律依次出现时, 如当

$$e = (-1)^{[\log_2(n)]}, \quad n = 1, 2, \cdots \tag{1.3}$$

时, y 为不确定现象, 且不具备频率的稳定性. 这里方括号表示取整运算 (R 语言中用函数 floor 完成取整运算任务), $\log_2$ 为以 2 为底的对数函数 (R 语言中用函数 log2 完成 $\log_2$ 的计算任务). 可以用如下的 R 程序代码模拟 y 的前 16 次观测值:

```
n=16;x=pi/2;
e=(-1)^floor(log2(1:n));
y=sin(x)+e;
y
```

运行上述程序代码后, 模拟的 y 在 RStudio 或 R 软件的控制台窗口显示如下:

```
[1] 2 0 0 2 2 2 2 0 0 0 0 0 0 0 0 2
```

注意, 将上述程序代码再运行一次, 会得到相同的结果, 其原因是未知的确定因素 e 的内在变化规律由 (1.3) 式所决定. ■

此例说明, 除了随机现象和必然现象之外, 还存在既不是随机的也不是必然的现象. 也就是说, 存在着不能事先确定且不具备频率稳定性的现象, 其原因是存在某种与该现象有关的确定因素.

在现实研究过程中, 无法知道与未知现象对应的理想模型, 只能不断地创建近似效果更好的 (现实) 模型. 研究过程中所建的模型只是理想模型的近似, 因此我们认为: