“十二五”国家重点出版物出版规划项目

现代统计学系列丛书

统计学：从概念到数据分析（第二版）

吴喜之　刘超

Statistics: from Concepts to Data Analysis

高等教育出版社

五”国家重点出版物出版规划项目
现代统计学系列丛书

统计学：
从概念到数据分析（第二版）

Statistics:
from Concepts to Data Analysis

吴喜之　刘超

ngjixue:cong Gainian dao Shuju Fenxi

高等教育出版社·北京

内容提要

本书主要介绍了概率基础、统计的基本概念、描述性统计、估计、假设检验、回归与分类等内容，同时介绍了一些经典的数据挖掘方法以及如何用R软件来实现相应的计算目标。

本书着重直观讨论，尽量少用公式，避免数学推导，强调统计学的基本内容及应用，使读者能够完整、准确地理解统计学的概念，学会利用R软件进行数据分析。

本书主要是为非统计学专业的学生和读者编写，读者不需要任何概率统计基础知识。

图书在版编目（CIP）数据

统计学：从概念到数据分析 / 吴喜之，刘超编著
. -- 2版. -- 北京：高等教育出版社，2016.4
（现代统计学系列丛书）
ISBN 978-7-04-044972-3

Ⅰ. ①统… Ⅱ. ①吴… ②刘… 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035156号

策划编辑	张晓丽	责任编辑	杨 帆	封面设计	赵 阳	版式设计	于 婕
插图绘制	杜晓丹	责任校对	张小镝	责任印制	朱学忠		

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街4号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印 刷	高教社(天津)印务有限公司		http://www.hepmall.com
开 本	787mm×960mm 1/16		http://www.hepmall.cn
印 张	14.75	版 次	2008年6月第1版
字 数	260千字		2016年4月第2版
购书热线	010-58581118	印 次	2016年4月第1次印刷
咨询电话	400-810-0598	定 价	26.40元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换

物料号 44972-00

现代统计学系列丛书编委会

(按姓氏笔画排序)

总　序

统计学是一门收集、整理和分析数据的科学和艺术。这里的“数据”通指“信息的载体”，涵盖了大千世界中的文本、图像、视频、时空数据、基因数据等。统计学是一个独立的学科，在历史上曾隶属于数学，但统计学与数学有着本质的区别，因此统计学教育有其自身的特点和要求，这些特点表现为：(1) 统计学研究的是随机现象，而数学研究的是确定性的规律；(2) 统计学是一门应用性很强的学科，许多概念和原理来自于实际的需要，不是数理逻辑的产物；(3) 数据在统计学中扮演了重要的角色。目前，统计学已被列为一级学科。

在过去的 30 年中，随着生命科学、信息科学、物质科学、资源环境、认知科学、工程技术、经济金融和人文科学等众多学科的发展，产生了许多新的统计学分支，如风险管理、数据挖掘、基因芯片分析等。此外，计算机及其有关软件在统计教育和应用中扮演了越来越重要的角色，它们提供了越来越多的图形表达和分析的方法，使得许多原来教科书中重要的内容，现在已变得无足轻重。统计教育必须要改革才能适应高速发展的形势。

大学的统计教育可分为两大类，一类是非统计学专业的课程，另一类是统计学专业的教学设计。非统计学专业的学生学习统计的目的是为了应用，在大学阶段，课程不多，主要是学习基础的统计概念和方法，学会使用统计软件，培养其解决实际问题的能力。统计学专业的课程设置十分重要，应向国际靠拢，对教师队伍的要求也较高。虽然这两类学生的教育有很多共同点，但在课程设置中必须加以区分。

我国的统计教育在过去受苏联的影响很深，把统计学作为数学的一个分支，在内容上偏理论，少应用，过于强调概率论在统计中的作用。统计学是一门应用性很强的学科，应从实际问题、从数据出发，通过统计的工具来揭示数据内部的规律。用“建模”的思路来教统计，使学生能更加容易理解统计的概念和方法，知道如何将实际问题抽象为统计模型，反过来又指导实践。对非统计学专业的学生，要强调统计的应用。学生要能熟练地使用至少一个统计软件包。对于统计学专业的学生，要培养学生对实际问题的建模能力。有些实际问题可直接应用现有的统计方法来解决，如问卷调查的统计分析。有些问题在初次接触时并不像一个统计问题，必须有坚实的统计基础和对实际问题的洞察力，才

能从中发掘出统计模型。要培养学生的这种能力及统计思想 (统计思想是统计文化的一部分, 是用统计学的逻辑思考问题) 。教师在授课中要结合较多的应用例子, 要求学生做案例研究, 鼓励学生参加建模比赛, 参加企业的实际项目。

为满足我国统计教育发展的需要, 我们计划编写一套面向高校本科生、特别是一般院校, 适用于统计学专业和非统计学专业的系列教材。系列教材的编写宗旨是: 突出教学内容的现代化, 重视统计思想的介绍, 适应现代统计教育的特点及时代发展的新要求; 以统计软件为支撑, 注重统计知识的应用; 内容简明扼要, 生动活泼, 通俗易懂。编写原则为: (1) 从数据出发, 不是从假设、定理出发; (2) 从归纳出发, 不是从演绎出发; (3) 强调案例分析; (4) 重统计思想的阐述, 弱化数学证明的推导。系列教材分为两个方向, 一个面对统计学专业, 另一个面对非统计学专业和应用统计工作者。

系列教材是适应形势的要求, 由高等教育出版社邀请专家组成“现代统计学系列丛书编委会”负责选题、审稿, 由高等教育出版社出版。

以上是我们编写这套教材的背景和理念, 希望得到读者的支持, 特别是高校领导和教学一线教师的支持。我们希望使用这套教材的师生和读者多提宝贵意见, 使教材不断完善。

现代统计学系列丛书编委会

扫描二维码, 获取更多丛书信息

第二版前言

自本教材第一版出版以来，大数据的概念逐步深入人心，数据分析方法被越来越多的人接受和使用。我们认为：**任何统计研究和应用都应该是问题驱动 (problem driven) 或者是数据驱动 (data driven)，而不应该是模型驱动 (model driven)。统计的发展是基于应用的，没有应用背景的数学式的统计没有存在的必要。缺乏应用背景的统计模型再漂亮也没有任何实际意义。**

因此，我们在第一版教材的基础上对《统计学：从概念到数据分析》进行了修订。其变化主要体现在以下几个方面：

第一，本书对部分章节进行了调整和删改。具体来说，将第四章变量的分布的抽样分布等内容单独设为一章，从而将第一版的共九章修订为共十章；将第一版的第二章变量和数据的概率与随机变量一节与第四章变量的分布相关内容合并，使得内容更加紧凑和流畅；在经典回归与分类这一章增加了对于回归利用交叉验证的例子一节，使得读者对算法建模的模型评价方法有所了解；在现代回归与分类这一章删除了最近邻方法和人工神经网络，增加了 boosting 回归等相应方法和交叉验证。

第二，本书加强了统计软件 R 的使用。将第一版中作为脚注的 R 程序放到每道例题的求解中，并且删除了第一版中的 SPSS 和 Excel 操作。此外，删除了各章的软件的使用，因为全书已经贯穿着统计软件的使用。本书例题与习题中涉及的相关数据可在 http://smss.buaa.edu.cn/szdw/ jxls/lc/index.htm 下载。

第三，考虑到教材的广泛使用，本书在每一章增加了延伸阅读，并且更新了部分数据，使用与老百姓生活紧密相关的例子，突出统计方法的实际应用，让数据分析更具有时代气息，增强教材的可阅读性。

特别感谢广大读者和授课教师，是他们的鼓励、建议帮助我们不断完善教材。

感谢高等教育出版社的相关编辑，没有他们的高效率的辛勤劳动，新版教材也不可能这样快和读者见面。

由于作者水平有限，书中错误和疏漏在所难免，敬请读者指正。

吴喜之　刘超

2015 年 10 月

第一版前言

统计是一个对所有领域都有用的工具。一个本科生，无论其主修方向是什么，如果能够掌握一些统计数据分析方法，无疑会受益匪浅。本教材主要是为非统计专业的学生和读者编写。强调统计学最基本的内容及应用，使读者不但能够逻辑完整地准确理解统计学的概念，而且能学会如何通过计算机统计软件进行数据分析。本书包括了基本概念、描述性统计、估计、假设检验、回归与分类等统计学内容。特别要提出的是，本书还介绍了如何用计算机实现在数据挖掘中用于回归与分类的常用方法。

这是一本从入门到具体数值分析与计算的课本。我们着重直观解释统计的基本概念，尽量少用公式，避免引入只有专业统计人员才需要了解的数学推导及定理证明。

学习本书不需要读者事先学过概率论或微积分，因此，完全可以放在大学本科任何一个学期讲授。对于学过概率论或概率论与数理统计课程的读者，完全可以跳过第二章的第二节及全部第四章的内容，并且在授课时主要强调数据的描述、数值计算及结果的解释。

我们提倡启发式教学，以理解为主，杜绝死记硬背。为此，书中以“思考一下”的形式给出很多启发性的问题、注解和补充，供老师和学生讨论和思考。这些问题也可以作为习题。而在习题部分，强调数值计算和分析。希望读者用计算机来实现所有的数值计算题。我们不希望在统计课程中存在了至少五十年的“手工计算 → 查表 → 手工计算”的前计算机时代的模式再延续下去。

本书每一章后面都介绍了如何用 R 、SPSS 、SAS 等软件来实现相应的计算目标。不仅如此，书中对每一个计算结果的获得以及每一个重要图形的绘制，都在脚注中说明了如何用 R 软件来实现。希望这种方式对教和学都有帮助。

可能有人担心，没有了手工计算、死记硬背的术语定义或数学推导就不易考试了。教学不是为了考试。教材也不能为迎合考试而编写。其实，考试的方式是多种多样的。开卷或闭卷、用计算机或不用计算机都可以考查学生的能力。一种行之有效的课堂闭卷考试方式为选择题。这些题目包括关于基本概念的是非题、对给定具体应用的统计方法选择题、计算机输出结果的解释题（也是选择）等。为有助于拉开学生成绩间的距离，题目量可以很大，而做题时间

要短,部分人能够答完就行了。

本教材的内容在中国人民大学的非统计专业本科及研究生的统计学课程中得到使用。希望读者能够提出宝贵的意见。

中国人民大学 统计学院
吴喜之

目　录

第一章

引 言

1.1 什么是科学方法?

我们天天都在使用“科学”这个词语, 但是, 有多少人认真考虑过科学的真正含义呢?

人们对世界的认识来源于他们所获得的信息 (或数据). 而在总结这些信息时人们头脑中会形成一些模型 (也称假说或理论). 这些模型会指导人们做进一步的探索, 直到遇到这些模型无法解释的现象. 这时, 人们会改进这些模型, 或者干脆建立新的模型使得新模型不仅可以解释旧模型可以解释的现象, 而且还能解释旧模型无法解释的现象. 这就是科学的方法. 而只有用科学方法进行的探索才叫**科学**. 下面举两个人们熟知的例子.

- **天文学:**公元 2 世纪, 托勒玫致力于传播宇宙地心说, 这一思想影响了 1 300 多年. 地心说可以对当时条件下的一些天文观测提供解释. 1543 年, 在哥白尼的《天体运行论》一书中阐明了日心说, 把托勒玫的理论大大改进了. 随后, 开普勒发现行星运动原理. 伽利略开始将望远镜用于天文观测. 牛顿又建立了运动和万有引力定律. 在新的观测的基础上, 赖特在 1750 年提出宇宙是由众多星系构成的看法. 18 世纪末, 赫歇尔首先用望远镜进行了巡天观测, 奠定了现代恒星天文学的基础.

- **从牛顿到爱因斯坦:** 牛顿发现了运动定律和万有引力定律, 这些定律在当时可以解释相当一部分观测的现象. 然而, 后来在亚原子尺度上, 以及在行星观测中出现了一些用牛顿的惯性定律或万有引力定律无法解释的现象. 这就导致了爱因斯坦狭义和广义相对论的产生. 相对论是建立在光速在真空中不变的假定前提之下的. 如果人们观测到光速在真空中可变, 则又会对相对论进行修正.

从上面的例子可以看出科学的一些特点. **科学可以定义为对关于宇宙的所有方面的知识的认真的、系统的、合乎逻辑的研究.** 这些知识则是由考察最好

的可利用的证据得到的, 并且这些知识总是应该在发现更有力的证据时随时予以纠正和改进. **科学也可以定义为任何知识系统, 这些知识涉及物理世界及其可经受无偏见观测和系统实验的现象.**

科学方法是目前已知的筛去谎言和错觉的最好方式, 对其步骤可做如下大致的描述:

(1) 观测宇宙的某些方面.

(2) 发明或提出可以解释这些观测的假说或假设, 它必须和观测结果相容.

(3) 利用该假说进行预测.

(4) 用实验来检验这些预测, 或者做进一步观测并根据结果修正假说.

(5) 重复第 (3),(4) 步直到在理论和实验或观测中没有发现矛盾为止.

任何假说, 如果能够说明很多现象, 也可称为理论. 但任何理论都不能达到绝对的真理. 所有的科学理论都应该是**可证伪的 (falsifiable**[1]**)**, 这意味着应该存在某种实验或可能的发现证明该理论有问题. 看不见摸不着的神的存在是无法证伪的, 因此宗教不是科学, 而是信仰. 目前基于不能重复观测或重复实验的现象而产生的许多说法, 都不是科学, 最多是信仰. 没有证伪, 科学是不会发展的. 从上面天文学和物理学的例子可以看出, 科学的理论是在否定中发展的. 每当发现目前理论所不能解释的实验结果或观测, 就产生对理论进行改进或更新的动机或需要.

注意, 一个科学理论即使被发现有局限性, 也不是不能应用. 比如, 在一般条件下, 牛顿定律还是适用的. 现在谁都知道地球大概是个球体. 但在发现地球是球形之前, "地平说" 可以解释很多现象. 即使是今天, 在盖普通房子之前进行测量时, 也不必考虑地球的曲率.

科学是靠证据说话的. 一个理论适用与否是靠实验或观测, 靠辩论是不行的. 古希腊的伟大哲学家亚里士多德用各种理由辩论说男人和女人的牙齿数目不同. 他是好的辩才, 但不是好的科学研究人员. 基于含糊不清或者不适当的前提的逻辑推理是没有多大意义的. 科学研究必须是毫无偏见的. 科学的结论应该独立于研究人员的文化背景、社会背景、种族、习惯、宗教和政治信仰等因素.

当然, 也存在制造假的研究结果的现象. 但是, 除非造假者的结论没有多大意义, 否则造假总是会被人发现的. 最有名的造假案例是 1989 年美国犹他大学的彭斯和英国南安普敦大学的弗莱什曼冷核聚变以及韩国科学家黄禹锡克隆胚胎干细胞的例子. 也有科学家犯错误的案例, 其中最突出的是在伦琴发现 X 射

[1]这个词的英文定义为: falsifiable. *adj*: capable of being tested (verified or falsified) by experiment or observation [syn: {confirmable}, {verifiable}].

线之后, 法国教授布朗洛宣称发现 N 射线. 但无人能够重复布朗洛的实验, 人们还证明他的观测有误.

此外, 权力、宗教和意识形态也会对科学造成严重干扰. 拥护哥白尼的天体运行论的布鲁诺被罗马教廷以“异端分子和异端分子的老师”的罪名, 于 1600 年 2 月 17 日被烧死在罗马鲜花广场. 伽利略由于收集、分析了日心说的证据, 于 1633 年被罗马天主教廷判决软禁, 在软禁中度过余生, 结果使得地中海地区的科学传统完全停止了. 在 20 世纪 30 到 60 年代, 苏联的全苏列宁农业科学院院长李森科出于政治与其他方面的考虑, 把得到实验支持的孟德尔和摩尔根遗传学斥为资产阶级的异端邪说, 并在斯大林的支持下对苏联的研究基因的学者实行迫害. 李森科事件是政治权威取代科学权威裁决科学争论的典型案例. 这件事也对中国遗传学界产生了恶劣影响.

思考一下

1. 那些关于耳朵识字、透过封闭的瓶子取物等报道的事件是科学的吗?
2. 举出一些科学、信仰和魔术的实例. 它们之间有什么区别?
3. 有人说他们见到了不明飞行物, 这能够证明它们存在吗?
4. 谈谈你对电视上对一些问题进行辩论的看法.
5. 举例说明利用科学方法得到的结论应该是可重复的.
6. 举例并讨论如果信仰和偏见混入了大前提, 哪怕逻辑过程再正确, 也不可能得到科学的结论.
7. 为了某种目的, 一些人伪造研究结果, 这些最终会被发现吗? 举例说明.
8. 有人说科学也是信仰, 所信仰的是“宇宙存在规律”. 请大家讨论这个“信仰”是真的宗教式的信仰, 还是可以用证据来说明的.

1.2 什么是统计学?

人们总是在现实世界中收集各种证据, 试图找出一些规律或模型来描述所研究的对象. 物理学、化学、生物学、地理学、天文学都是这样做的, 统计学也一样. 不同的是, 那些自然科学本身的规律是比较确定的. 叙述简洁的牛顿定律就是一个很好的例子. 此外, 在一定条件下化学反应中的结果也是完全确定的, 许多天体的运动轨道也是基本可以确定的. 而世界上还有许多事物是无法用确定性的理论来描述的. 比如, 一个企业家去年增加投资可能利润有所增加, 而今年增加投资就可能赔本. 再如, 保险公司希望减少汽车保险中的风险,

这就需要找出具有哪些特征的人群具有高风险. 这些问题绝对不能从逻辑推理来得到解决方案, 必须通过分析相关的数据才能总结出规律.

我们继续以汽车保险为例. 国外保险业数据表明年轻人、开红色车的人、开跑车的人均容易出车祸. 从物理学角度, 鲜明的红色应该更醒目, 性能好的跑车应该更安全, 眼明手快的年轻人更不易出事. 但保险公司的规律是统计学家从事故数据总结出来的. 现在, 心理学家可以从这一类人有炫耀倾向的心理因素来解释. 统计学仅仅是从数据中找出规律而已. 但是在统计学家找出什么类型的人易出事故的规律之前, 心理学家往往不会往这方面去想.

因此, 在无法用自然科学的定律来解释的情况下, 在研究许多说不清楚原因的现象时, 统计学可以通过研究数据来找出规律甚至答案. 统计学进行推断的基础就是数据. 因此可以像不列颠百科全书那样, 把统计学定义为: **统计学是收集、分析、展示和解释数据的科学**[1]. 这个定义是国际上普遍接受的. 这里所说的数据就是科学中的事实和证据. 这意味着, 统计学是一门科学, 其方法是科学的方法. 在大数据时代, 数据不仅限于数字, 它也可能是图像或者是文字. 实际上, 任何信息都可以称为数据. 所以, **统计学是数据的科学和艺术**.

延伸阅读: 统计学的英文名字 statistics

统计学的英文名字 statistics 大约在十八世纪中叶由德国学者阿亨瓦尔 (Gottfried Achenwall) 所创造, 是由状态 status 和德文的政治算术联合推导得出的, 第一次由约翰·辛克莱 (John Sinclair) 所使用, 于 1797 年出现在不列颠百科全书中. 统计一词的本意是国势学, 即关于一个国家基本情况的调查. 这些情况可以是描述性的事实, 也可以是数据, 其中最重要的一部分, 就是人口情况. 18 世纪, 又有人把这种基于实情的条目放到了表格中, 按行和列组织起来, 这样更容易对比, 因此促进了数据的使用. 因为较早地重视数据、使用数据, 德意志民族也一直以“精准”著称. 基本上和德国同期, 英国人也开始探索统计在政治和社会生活中的作用. 1662 年, 约翰·格朗特 (John Graunt) 调查了伦敦的人口死亡情况, 分析了新生男孩和女孩的比例, 发展了现在保险公司所用的那种类型的死亡率表. 随后发表了他第一本也是唯一一本手稿《对死亡的自然和政治观察》, 该书被后世公认为是统计学的开山之作. 中文的“统计”一词来源于日本. 在日本明治维新时期, 日本向西方国家学习, 成立了国家统计局. 1872 年进行了第一次全国人口普查, 1882 年出版了第一本统计年鉴. 1903 年, 中国学者翻译了日本人横山雅南所著的《统计讲义录》一书, 把“统计”一词引入中国.

[1] Encyclopedia Britannica, 2006.