全国统计教材编审委员会“十二五”规划教材

统计学：从数据到结论

第四版

吴喜之 编著

中国统计出版社
China Statistics Press

全国统计教材编审委员会“十二五”规划教材

统计学：从数据到结论

第四版

吴喜之 编著

中国统计出版社
China Statistics Press

图书在版编目(CIP)数据

统计学：从数据到结论 / 吴喜之编著. —4 版. — 北京 ：中国统计出版社，2013.3

全国统计教材编审委员会“十二五”规划教材

ISBN 978—7—5037—6789—0

Ⅰ. ①统… Ⅱ. ①吴… Ⅲ. ①统计学 Ⅳ. ①C8

中国版本图书馆 CIP 数据核字(2013)第 044840 号

统计学：从数据到结论

作　　者/吴喜之
责任编辑/梁　超
封面设计/上智博文
出版发行/中国统计出版社
通信地址/北京市丰台区西三环南路甲 6 号　　邮政编码/100073
电　　话/邮购(010)63376909　书店(010)68783171
网　　址/http://csp.stats.gov.cn/
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/16
字　　数/267 千字
印　　张/15
印　　数/1—5000 册
版　　别/2013 年 3 月第 4 版
版　　次/2013 年 3 月第 1 次印刷
定　　价/30.00 元

全国统计教材编审委员会

出版说明

“十二五”时期，是我国全面实施素质教育，全面提高高等教育质量，深化教育体制改革，推动教育事业科学发展，提高教育现代化水平的时期。“十二五”伊始，统计学迎来了历史性的重大变革和飞跃。2011年2月，在国务院学位委员会第28次会议通过的新的《学位授予和人才培养学科目录(2011)》(以下简称“学科目录”)中，统计学从数学和经济学中独立出来，成为一级学科。这一变革和飞跃将对中国统计教育事业产生巨大而深远的影响，中国统计教育事业将在“十二五”时期发生积极变化。

正是在这一背景下，全国统计教材编审委员会制定了《“十二五”全国统计教材建设规划》(以下简称“规划”)。根据“学科目录”在统计学下设有数理统计学，社会经济统计学，生物卫生统计学，金融统计、风险管理与精算学，应用统计5个二级学科的构架，“规划”对“十二五”全国统计规划教材建设作了全面部署，具有以下特点：

第一，打破以往统计规划教材出版学科单一的格局。全面发展数理统计学，社会经济统计学，生物卫生统计学，金融统计、风险管理与精算学，应用统计5个二级学科规划教材的出版，使“十二五”全国统计规划教材涵盖5个二级学科，形成学科全面并平衡发展的出版局面。

第二，打破以往统计规划教材出版层次单一的格局。在编写出版好各学科本科生教材的基础上，对研究生教材出版进行深入研究，出版一批高水平高层次的研究生教材，为我国研究生教育、尤其是应用统计研究生教育提供教学服务。同时，积极重视统计专科教材出版，联合各专科院校，组织编写和出版适应统计专科教学和学习的优秀教材。

第三，打破以往统计规划教材出版品种单一的格局。鼓励内容创新，联系统计实践，具有教学内容和教学方法特色的、各高校自编的相同内容选题的精品教材出版，促进统计教学向创新性、创造性和多样性

发展。

第四，重视非统计专业的统计教材出版。探讨对非统计专业学生的统计教学问题，为非统计专业学生组织编写和出版概念准确、叙述简练、深入浅出、表达方式活泼、练习题贴近社会生活的统计教材，使统计思想和统计理念深入非统计专业学生，以达到统计教学的最大效果。

第五，重视配合教师教学使用的电子课件和辅助学生学习使用的电子产品的配套出版，促进高校统计教学电子化建设，以期最后能形成系统，提高统计教育现代化水平。

第六，重视对已经出版的统计规划教材的培育和提高，本着去粗存精、去旧加新、与时俱进的原则，继续优化已经出版的统计教材的内容和写作，强化配套课件和习题解答，使它们成为精品，最后锤炼成为经典。

“十二五”期间，编审委员会将本着“重质量，求创新，出精品，育经典”的宗旨，组织我国统计教育界专家学者，编写和编辑出版好本轮教材。本轮教材出版后，将能够形成学科齐全、层次分明、品种多样、配套系统的高质量立体式结构，使我国统计规划教材建设再上新台阶，这将对推动我国统计教育和统计教材改革，推动我国统计教育事业科学发展，提高我国统计教育现代化水平产生积极意义。

让教师的教学和学生的学习事半功倍，并使学生在毕业之后能够学以致用的统计教材，是本轮教材的追求。编审委员会将努力使本轮教材好教、好学、好用，尽力使它们在内容上和形式上都向国外先进统计教材看齐。限于水平和经验，在教材的编写和编辑出版过程中仍会有不足，恳请广大师生和社会读者提出批评和建议，我们将虚心接受，并诚挚感谢！

全国统计教材编审委员会

2012 年 7 月

再版说明

这本书已经有了近十年的历史，现在将要出第四版。前面三版已经作为参考书或教科书在许多学校使用。各个学校的师生对本书提出许多宝贵的意见，并且指出了很多错误和不妥之处。读者的支持和鼓励，对本书各版的诞生起着关键的作用。第四版在许多地方对前面几版进行了修改和增减。

免费的自由编程的R软件在国际上已经成为统计教学和科研的主要软件，本书第四版全部采用R软件来描述计算过程，彻底放弃了使用商业软件。R软件非常强大，凡是国际上出现的新方法，都会很快地上传到R的网站上，在发达国家，不能想象一个统计教师或者统计研究生不会熟练使用R。从R的功能和使用者的人数来说，它已经远远超过所有昂贵的商业软件。R软件的绝大部分程序包的代码都是公开的，透明是防止腐败的最好方式。此外，由于R在中国的普及越来越广泛，网上关于R的互动和帮助的环境也已经形成，中国学生和实际工作者完全可以赶上国际统计界使用R的主流(虽然已经至少落后了10年)。

在强大的免费R软件不断普及的情况下，对于缺乏经费的中国教育系统以及并非富裕的学校师生来说，教学中继续通过昂贵的商业软件来讲授统计变得越来越缺乏吸引力。用商业软件教学的一个客观效果是鼓励非法盗版行为。由于避免了对商业软件菜单的点击鼠标的繁琐而又冗长的细节叙述，整本书都显得简洁明了，节省了大量的篇幅(第四版比前一版减少了一百多页)。课文中所有计算过程都附有可以实现的R语句，在每章最后仅仅对R语句做些汇总或说明。

虽然R软件是编程语言，但由于其简单易懂，任何从来没有使用过R的人都可以毫不费力地通过复制和粘贴书上的代码重新实现书上的所有例题。书后附录中的R代码练习更可以帮助读者尽快地掌握R语言。

许多人，比如各层管理人员，并不一定都进行第一线的实际数据计算，但为了理解手中关于本单位及有关方面信息的意义，为了更好地进行明白的决策，他们必须理解各种统计推断结果的意义。对这些人，不一定要求能够使用软件，更不需要理解数学推导，但他们必须明白各种统计概念和方法以及输出结果的意义，明白那些数据分析人员在做什么。相信本书对他们肯定会有所裨益。

在内容方面，本版专门添加了有广泛应用前景的机器学习的回归和分类方法，并且把这些内容及经典的回归和判别分析等归到一章。此外，把多元分析的除判别分析之外的其他内容合并到一章之中。这一版还取消了非参数检验一章，把其中一些常用的非参数检验加入到假设检验的一章中。

作为教科书，本书内容对于每周两学时的课程似乎太多。我觉得，什么讲或者什么不讲应该根据学生的需要由老师自己安排。实际上，对于任何课程，最好是由任课教师来决定讲哪些内容以及如何讲。因为他们最了解他们所面对的学生。教科书编者的思维方式不见得和老师的一致，而老师最好按照自己的理解来讲述。一个好的教科书，应该给教师以较大的余地和自由。

笔者希望读者在阅读本书时能够以理解统计方法的含义为主，学会处理数据，提高学习和应用能力。**在任何国家及任何制度下都能够生存和发展的知识和能力，就是科学，是人们在生命的历程中应该获得的。**

希望读者继续对本书予以宝贵的支持和批评指正。

吴喜之

2012 年 10 月

第一版前言

什么在本书中等待着你们去发现，去探讨，去欣赏呢？当然不是数学公式和定理定义的堆砌，也不是和枯燥的公文报表相关的政府工作的培训。这是一门充满了哲学韵味的认识世界的学问。

不知读者们是否意识到，统计已经渗入到人们的社会、生活、工作等各个领域。每天新闻媒介报道的各个方面都离不开各种统计数据和各种分析与预测。人们可能对于这些统计内容觉得习以为常，也可能会有一些好奇或神秘感。由于国情不同，统计的地位与人们对统计的看法也不同。在发达国家，一般民众觉得统计学和数学类似，是一门高不可攀但极易找到满意工作的学问。在中国，又有一些人认为统计就是处理政府报表的职业。但自从中国向世界开放之后，越来越明确的一点是，没有什么学科或领域能够真正离开统计。

以应用为目标学习统计，究竟是为什么？是为了流利地背诵一大堆定义、概念和抽象的名词和术语吗？是为了学习如何进行推导和证明一些复杂的定理和公式吗？这些问题不仅学生会思考，更重要的是统计教师要思考。本书的目的是希望读者在学习之后，能够知道实际中哪些是统计问题，最好能够自己解决一部分统计问题，即使不能解决也知道能够在哪里查到答案和向谁请教。知识固然重要，更重要的是通过学习获得解决和处理问题的能力。

学习并不总是一个令人生畏或至少成为某种负担的过程。人们学会走路、说话、骑车、下棋、打球等大都是在一种乐趣中进行的。为什么涉及日常生活的每一个方面的统计就不能和看侦探小说那么引人入胜呢？其实任何一门科学，都有其趣味性，而只有把科学研究当成游戏的人才会真正成为大师。这门课并不想使读者都成为统计学家，而仅仅想让读者如同学会使用电脑、手机、学会辩论、上网或讨价还价那样愉快地认识或理解在人生中无法躲开的统计。

本书由浅入深地把统计最基本和最有用的部分在这么一本不厚的教科书中完整地介绍给读者，而且让读者可以边学习，边着手用统计软件处理数据。篇幅大、语言罗嗦的教材对读者是个负担，不但浪费了资源，也抓不住要领。因此，作者力图惜墨如金，既节省篇幅，又要把该解释的全部说清。希望读者慢慢咀嚼，不必图快。

很少有一本统计教材包括像本书那么多的统计内容。我觉得，这些内容本来并不深奥，只是其貌似复杂的数学工具把它搞成阳春白雪，再加上强调数学推导的教学方式，使得统计显得高不可攀。本教材要还这些统计应用以其本来面目。使得统计变成人人都能够基本上理解和掌握的有用工具。多数使用计算机的人都不是计算机专业的，多数开汽车的都不会修汽车，但这对他们毫无妨碍。难道不会推导或背诵与统计有关的数学公式就不能应用统计这个工具了吗？

本书每一章的主要部分是用日常语言来引进和解释一些概念，如果可能，就通过例子来说明。如果不涉及应用，这部分就足够了。在本书例题的分析中，同时提供简洁明了的软件代码，可以使读者一边看书，一边自己计算，这会给多数想要自己动手分析数据的读者以方便。每章后面的小结中还展示了与概念及计算有关的一些数学公式以及软件的说明，使那些精力充沛的读者能更深刻地理解内容。这种安排使得本教材能够适用于各种不同水平、不同要求的读者群体。

本教材不仅可供没有学过概率论和数理统计的非统计专业的本科生和研究生使用，也可以供统计专业的本科生作为理解统计本来含义的教材使用（以代替不能满足需要的“描述统计学”等类课程），它还可以为各领域的广大实际工作者作为应用各种统计方法的参考书。为了读者可以使用各种软件来进行分析，本书所涉及的所有电子版数据都为文本格式。

软件方面，本书则采用免费的自由软件 R[①]。经验表明，在学习统计内容的时候学习软件比上专门的软件操作课更有效。R 软件既采用了最简单的编程语言，又拥有最丰富的统计资源。一个大学本科生通常可以

① 第一版主要使用 SPSS 和部分地应用 Excel，第二版之后加了 SAS 和 R 的应用，第三版则以 R 软件为主，第四版则全部用 R 软件。

在一天内学会R的基本计算，在一周内学会统计基本课程的计算。

在前计算机时代，几乎所有的统计教科书都给出了各种与分布有关的表格。但随着计算机的普及，所有统计软件（无论是商业的还是免费的）都给出了和各种分布有关的各种函数，把人们从繁琐而又不精确的查表中解放出来。目前很多国外的统计教科书都不再提供既占用篇幅又比较粗糙的分布表。本书不准备提供任何和分布有关的表格。本书第四章会介绍如何使用软件来进行与概率分布有关的计算。

这个教材的全部内容曾作为非统计专业硕士和博士的课程分别在北京大学光华管理学院及中国人民大学讲授过，受到普遍欢迎。实践证明，这本书的大部分内容完全能够轻轻松松地在一个学期（每周三个学时）中全部讲完。一些热心而又好奇的非统计背景的人士也曾读过本教材的全部内容，没有任何理解上的问题。当然，根据不同的教学对象和需要，有些章节可以完全不讲或少讲。

本书前面的章节，是对统计基本概念的介绍。而后面的部分则是更有针对性的一些统计模型和方法。一般传统统计学的课程包括前六章，或最多前七章的内容，而第八章属于多元统计分析的课程内容，第九章一般属于时间序列课程包含的内容，第十章简单介绍了生存分析，第十一章对指数进行了必要的介绍。目前大多数流行的统计应用都已包含在本教材内。

本书的编写是在国家统计局教育中心的建议和鼓励下产生，并得到其大力支持。本书还受到北京大学、中国人民大学以及各兄弟院校老师和学生的鼓励和帮助。中国统计出版社一直关心着本书的写作和出版。特别要指出的是敬爱的汪仁官老师又一次为我所写的统计教材进行了非常认真的审校，使我重新感受到做学生的幸福，中国统计界的老前辈茆诗松老师也热心地对本书提出了许多宝贵而又中肯的建议。他们的审校和建议使本书避免了许多错误和不妥之处。没有这些支持和帮助，本书是不可能面世的。谨在此对所有各方面表示衷心的感谢。

吴喜之

2003年6月

目 录

第一章　一些基本概念

1.1　统计是什么?

你想过下面的问题吗?

1. 当你买了一台电脑时, 被告知三年内可以免费保修. 那么, 厂家凭什么这样说? 说多了, 厂家会损失, 说少了, 会失去竞争力, 也是损失. 到底这个保修期是怎样决定的呢?
2. 在同一年级中, 同样统计学的课程可能由一些不同教师讲授. 教师讲课方式当然不一样, 考试题目也不一定相同. 那么如何比较不同班级的统计学成绩呢?
3. 大学或企业的排名是一个非常敏感的问题. 不同的机构得出不同的结果, 各自都说自己是客观、公正和有道理的. 到底如何理解这些不同的结果呢?
4. 任何公司和个人都有一个信用问题. 如果他们在试图得到贷款时并没有不还贷的不良记录, 如何根据其背景资料来判断其信用等级呢?
5. 我国东部和西部的概念是一个比较笼统的概念. 如何能够根据某些标准或需要, 选择一些指标来把各省, 或各市县甚至村进行分类呢?
6. 疾病传播时, 如何能够通过被感染者入院前后的各种经历得到一个疾病传染方式的模型呢?
7. 如何通过问卷调查来得到性别、年龄、职业、收入等各种因素与公众对某项事物(比如商品或政策)的态度的关系呢?
8. 一个从来没有研究过红楼梦的统计学家如何根据比较写作习惯得出红楼梦从哪一段开始就不是曹雪芹的手笔了呢?
9. 如何才能够客观地得到某个电视节目的收视率, 以确定插播的广告价格是否合理呢?
10. 如何根据税务部门过去的税收记录来预测下一年的税收收入, 供政府部门制定预算时参考?
11. 如何根据某地区的寿命记录来确定人寿保险的既有竞争力, 又有利可图的定价?

其实, 这些都是统计应用的例子. 这样的例子太多了, 无法一一列举. 因为统计学可以应用于几乎所有的领域, 包括精算、农业、动物学、人类学、考古学、审计学、晶体学、人口统计学、牙医学、生态学、经济计量学、教育学、选举预测和策划、工程、流行病学、金融、水产渔业研究、遗传学、地理学、地质学、