



普通高等教育“十一五”国家级规划教材

全国统计教材编审委员会“十一五”规划教材



非参数统计

第二版

★ 吴喜之 编著



中国统计出版社
China Statistics Press



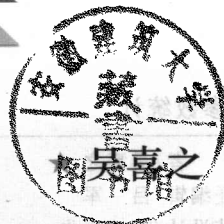
普通高等教育“十一五”国家级规划教材

全国统计教材编审委员会“十一五”规划教材



非参数统计

第二版



吴嘉之 编著

 中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

非参数统计/吴喜之编著. —2版.

—北京:中国统计出版社,2006.9

ISBN 7-5037-4997-0

I. 非…

II. 吴…

III. 非参数统计—高等学校—教材

IV. 0212.7

中国版本图书馆 CIP 数据核字(2006)第 104407 号

非参数统计

作 者/吴喜之

责任编辑/吕 军

装帧设计/艺编广告

出版发行/中国统计出版社

通信地址/北京市西城区月坛南街 57 号 邮政编码/100826

办公地址/北京市丰台区西三环南路甲 6 号

电 话/邮购(010)63376907 书店(010)68783172

印 刷/河北天普润印刷厂

经 销/新华书店

开 本/787×1092mm 1/18

字 数/210 千字

印 张/14.625

印 数/1—5000 册

版 别/2006 年 10 月第 2 版

版 次/2006 年 10 月第 1 次印刷

书 号/ISBN 7-5037-4997-0/O·59

定 价/29.00 元

中国统计版图书,版权所有。侵权必究。

中国统计版图书,如有印装错误,本社发行部负责调换。

出版说明

“十一五”时期是继续深化教育改革,加强素质教育,努力建设有利于创新型科技人才生长的教育培训体系的关键时期。为了更好地培育统计创新型科技人才,适应统计教育培训的新形势,全国统计教材编审委员会制定了《“十一五”全国统计教材建设规划》(以下简称规划)。规划坚持“以人为本”的科学发展观,坚持统计教育与实践相结合,坚持统计教育同国际接轨,坚持培养创新型的统计人才的指导思想,编写符合国民经济发展和统计事业发展需要的统计教材。

这批教材是在深入分析统计教育形势和统计教材建设发展状况,总结多年来统计教材建设经验的基础上,本着以建设本科统计教材为主的方针,积极探索研究生层次的统计教材,力争使规划统计教材的编写做到层次分明,有针对性和实用性。建设精品教材,是编委会自成立以来孜孜以求的目标。考虑到统计教材建设的实际情况,“十一五”期间,本科教材主要以修订为主,对以往规划统计教材中使用面广,得到广大教师和学生普遍认可的教材组织了修订。修订后的教材,淘汰了过时的内容和例子,增加了计算机操作和大量的案例,编写手法也做了一定的调整,在实用性、可操作性等方面有了较大的改进。

近年来,我国现代化建设快速发展,高等教育规模持续扩大,尤其是研究生教育规模的扩大,使得高等学校研究生统计教学工作面临着许多新情况、新问题,任务艰巨。因此,必须坚持科学发展观,在规模持续发展的同时,把提

高研究生统计教学质量放在突出的位置,培养全面发展的创新型的统计人才。教材是统计教学的载体,建设高质量的研究生层次的统计教材是统计教育发展的需要。因此,编委会在“十一五”期间对研究生的统计基础课教材做了些有益的探索。根据《规划》的要求,这批教材主要采取招标和邀请的方式组织有关院校的专家、学者编写。

值得特别提出的是,在这批教材中,有《非参数统计》、《概率论与数理统计》、《经济计量学教程》、《医学统计》、《应用时间序列分析》、《多元统计分析》、《统计学》7种教材入选国家教育部组织编写的“普通高等教育‘十一五’国家级规划教材”,更加充实和完善了“十一五”期间统计教材的建设。

为了便于教学和学习,这批教材里面包含了与之相配套的《学习指导与习题》,使得这批教材在编辑出版上形成了比较完整的体系。我们相信,这批教材的出版和发行,对于推动我国统计教育改革,加快我国统计教材体系和教材内容更新、改造的步伐,打造精品教材,都将起到积极的促进作用。

限于水平和经验,这批教材的编审、出版工作还会有缺点和不足,诚恳欢迎教材的使用单位、广大教师和同学们提出批评和建议。

全国统计教材编审委员会

2006年6月

前言

本书的目的是用简明的语言,不多的数学工具并通过大量例子来尽可能直观地介绍非参数统计的基本方法。它可以作为统计学专业本科一学期(2学时)的应用非参数统计课程的教材,也可以作为实际工作者自学或查阅的参考书。所需要的预备知识为统计学教程中的最基本的内容。读者只要知道总体和样本,随机变量及分布,统计量,检验和估计的基本概念等即可以看懂本书。虽然计算机并不是学会本书内容所必需的,但是不能想象,一个不会用计算机的统计工作者如何在实践中生存。

本书在引进每一个方法时,都通过数据例子来说明该方法的意义和使用过程。所有例题的计算和绘图都是由笔者完成的。笔者还核算了每一章后面的所有习题。由于这些习题都只涉及基本概念和方法,相信读者完全可以独立完成。由于本书的基本原理和方法广泛适用于许多不同的领域,这里的例子和习题尽量取自不同的领域和学科,以扩展读者的思路。

本书的第一章引言部分包含以下几类内容:(1)对统计和非参数统计以及计算机软件应用的一般论述;(2)对一些初等统计内容,特别是对本书常遇见的问题作了回顾;这些问题有一般的检验与置信区间问题,特别的 χ^2 检验问题,S探索性数据分析问题等;(3)初等统计不一定有的问题,比如渐进相对效率(ARE)和局部最优势(LMP)检验,顺序统计量,秩,线性秩统计量和线性记分问题。这里的(1)和(2)可以根据使用者的情况酌情处理,最好先浏览

一下,而在需要时再读。第(3)部分内容在书中多次涉及;但由于仅与理论推导和对方法的评价有关,可作为有兴趣的人的参考。

从第二章到第七章依次序为关于位置的单样本,两样本和多样本模型,尺度问题,相关与回归问题以及分布及一些 χ^2 检验问题。这些一般都可以讲;但是如果时间安排不开,可以对正态记分部分,仅作举一反三的式的介绍。这并不是它不常用,而是因为其思想仅仅是别的统计量的推广。最后两章为非参数密度估计和回归与稳健统计简介。这两部分中每一部分都可以构成数倍于本书厚度的专著。它们在统计中占有重要的地位,这里的内容仅打算让读者作一初步了解。

本书在编写过程中始终得到国家统计局教育中心的关心和帮助。苏州大学的汪仁官教授极其认真地审阅了全书,并提出了宝贵的意见;自然,所有的意见都是非常合理的而且均被采纳了;这使我回忆起36年前敬爱的汪老师为我们仔细批改数学分析作业的感人情景。本书的大部分内容和例子曾在中国人民大学讲过,在此也对积极参与课堂教学的统计96级同学一并表示感谢。

编者水平有限;欢迎各方面能对本书的错误和不当之处予以批评指正。

吴喜之

1999年11月20日

第二版说明

本书的第一版发行以来,在许多学校被选为“非参数统计”课程的教科书或参考书。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。没有他们的支持和鼓励,本书的第二版不可能面世。

和第一版相比较,第二版对许多内容完全重新写过,还进行了一些调整,同时加强了对概念和方法的解释,使得该书更加容易理解。第二版还对例题和习题都做了一些修订和增减,并且都在光盘中给出了数据。此外,还增加了一些内容,特别是关于如何通过编程来理解方法,以及用软件来实现数值计算的内容。本书在课文中关于计算方法的叙述中主要使用了免费的,功能强大的,需要自己动手写程序的 R 软件;力图清楚地用 R 语句来描述计算的细节。这也是一些“黑匣子”式的傻瓜软件所无法比拟的。R 软件是使用 S 语言来编程的(和 S-plus 的编程语言一样);在其问世的不到 10 年的时间,已经成为国外统计研究生的首选软件。它有强大的网上支持系统。多数最新的统计计算方法,在进入商业软件之前,就已经以 R 语言的形式在 R 网站上免费提供了。使用本书的师生最好也使用 R 语言。掌握 R 软件对其他统计方向的学习和研究都会有很大的帮助,甚至会有一种到了自由天地的感觉。

为了适应分析实际数据的各种需要,本书还在每一节(除了少数介绍性章节之外)的最后加入了使用 R、SPSS 或 SAS 等统计软件分析有关数据的程序语句和

各种选项的说明。这里要指出的是,编者尽量使本书提供的 R 程序是直接根据公式或定义写成的;这里的 R 程序没有按照专业化编写软件所通常遵循的高效率和漂亮输出的原则;这是因为那将使得显示基本公式和概念的语句淹没在为了形式和效率而加入的大量其他语句之中,而使得有关程序难以读懂。希望本书在编程上起着-一个抛砖引玉的作用,鼓励读者编出更加高效、更加漂亮的程序。

本书选择的与内容有关的 SPSS 软件选项和 SAS 软件语句(或选项)的原则是容易理解和掌握;当然,由于编者知识有限,对于有些方法,没有找到(因此也无法提供)合适及方便的 SPSS 或 SAS 方法;希望读者提出建议,使得再版时予以弥补。

由于使用软件比查表更加方便和可靠,有人说,你自己都不查表,为什么要教学生去查表呢?的确,编者除了在最初等的统计课教学过程中曾经涉及到少数统计表之外,从来都是使用软件。“己所不欲,勿施于人”,本书原本不想再提供任何统计分布表,但为了部分没有计算条件的读者易于理解,还是提供了少数最常用的表格,以备不时之需。希望有条件的读者尽量使用计算机,而不去查表。实际上,如果没有计算机的支持,很难对有一定规模的数据在任何统计方向进行较深入的分析。

一些读者提出,本书内容对于每周两学时的课程似乎太多。我觉得,对于初学者或者实际应用部门的人来说,可以略去不讲的章节包括:§ 1.6、§ 1.7(正态记号部分)、§ 3.3、§ 4.3、§ 5.2、第六章、第十章及第十一章。此外,第一章主要是用于介绍、回顾或参考的;可以有选择地在需要时讲、也可以完全不讲;这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。

本书所附的光盘中包含例题和习题的各种形式的
数据。

希望读者继续对本书予以宝贵的支持和批评指正。

吴喜之

2006年4月

第一章 引言

- 1.1 统计的实践 1
- 1.2 关于非参数统计 3
- 1.3 假设检验及置信区间的回顾 4
- 1.4 χ^2 检验简单回顾 9
 - 1.4.1 基于随机化模型的 χ^2 检验 10
 - 1.4.2 关于离散分布的列联表 χ^2 检验 11
- 1.5 熟悉手中的数据和数据变换 12
- 1.6 渐过相对效率(ARE);局部最优势(LMP)检验 14
- 1.7 顺序统计量,秩,线性秩统计量及正态记分 16
- 1.8 计算机统计软件的应用 20
- 1.9 习题 23

第二章 总体比例的检测和置信区间

- 2.1 小总体情况-超几何分布 27
- 2.2 大总体情况-二项分布及其大样本正态近似 31
- 2.3 习题 38

第三章 单样本问题

- 3.1 广义符号检验和有关的置信区间 39
 - 3.1.1 广义符号检验:对分位点进行的检验 41
 - 3.1.2 基于符号检验的中位数及分位点的置信区间 45
- 3.2 Wilcoxon 符号秩检验,点估计和区间估计 49
 - 3.2.1 Wilcoxon 符号秩检验 49
 - 3.2.2 基于 Wilcoxon 符号秩检验的点估计和置信区间 55
- 3.3 正态记分检验 58
- 3.4 Cox-Stuart 趋势检验 61

3.5	关于随机性的游程检验	64
3.6	习题	68
第四章 两样本位置问题		
4.1	两样本和多样本的 Brown-Mood 中位数检验	73
4.2	Wilcoxon(Mann-Whitney)秩和检验及有关置信区间	78
4.2.1	Wilcoxon(Mann-Whitney)秩和检验	78
4.2.2	M_X-M_Y 的点估计和区间估计	84
4.3	正态记分检验	86
4.4	成对数据的检验	88
4.5	习题	91
第五章 多样本数据模型		
5.1	Kruskal-Wallis 秩和检验	95
5.2	正态记分检验	100
5.3	Jonckheere-Terpstra 检验	102
5.4	区组设计数据分析回顾	104
5.5	完全区组设计:Friedman 秩和检验	107
5.6	Kendall 协同系数检验	110
5.7	完全区组设计:关于二元响应的 Cochran 检验	112
5.8	完全区组设计:Page 检验	114
5.9	不完全区组设计:Durbin 检验	115
5.10	习题	118
第六章 尺度检验		
6.1	两独立样本的 Siegel-Tukey 方差检验	122
6.2	两样本尺度参数的 Mood 检验	125
6.3	两样本及多样本尺度参数的 Ansari-Bradley 检验	128
6.4	两样本及多样本尺度参数的 Fligner-Killeen	

检验	131
6.5 两样本尺度的平方秩检验	133
6.6 多样本尺度的平方秩检验	135
6.7 习题	137
第七章 相关和回归	
7.1 Spearman 秩相关检验	141
7.2 Kendall τ 相关检验	143
7.3 Theil 非参数回归和几种稳健回归	148
7.4 习题	155
第八章 分布检验和拟合优度 χ^2 检验	
8.1 Kolmogorov-Smirnov 单样本检验及一些正态性检验	159
8.1.1 Kolmogorov-Smirnov 单样本分布检验	159
8.1.2 关于正态分布的一些其他检验和相应的 R 程序	163
8.2 Kolmogorov-Smirnov 两样本分布检验	166
8.3 Pearson χ^2 拟合优度检验	167
8.4 习题	170
第九章 列联表	
9.1 二维列联表的齐性和独立性的 χ^2 检验	172
9.2 低维列联表的 Fisher 精确检验	176
9.3 对数线性模型与高维列联表的独立性检验简介	179
9.3.1 处理三维表的对数线性模型	179
9.3.2 假设检验和模型的选择	181
9.4 习题	184
第十章 非参数密度估计和非参数回归简介	
10.1 非参数密度估计	187

目 录

10.1.1	一元密度估计	188
10.1.2	多元密度估计	191
10.2	非参数回归	193
第十一章 稳健统计方法简介		
附表		
参考文献		

第一章

引言

§ 1.1 统计的实践

虽然“统计学(statistics)”的定义在当今世界的百科全书和统计教科书中于文字、侧重点或描述方式有所出入,但就其所包含的总体内容和应用领域来说则差不多.比如“大英百科全书(The New Encyclopaedia Britannica)”(1993)在一开始说:“统计学是一门收集数据,分析数据,并根据数据进行推断的艺术和科学.最初与政府收集的数据有关,现在包括了范围广泛的方法和理论.”该书随后还列举了主要应用领域并详尽介绍了统计学的各方面的内容.以Kotz和Johnson(1983)主编的13卷“统计科学百科全书(Encyclopedia of Statistical Sciences)”是迄今最完整的关于统计的具有权威性的百科全书.它说“统计学”这个术语表示“涉及收集、展示和分析数据的普遍方法和原理的领域.”它还列举了四十多个运用统计的领域,它们包括: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,历史研究,人类遗传学,水文学,工业,法律,语言学,文学,劳动力计划,管理科学,市场营销学,医学诊断,气象学,军事科学,核材料安全管理,眼科学,制药学,物理学,政治学,心理学,心理物理学,质量控制,宗教研究,社会学,调查抽样,分类学,博彩和气象改善.

统计在每一个应用领域都有它自己的目标 and 特点, 有的还有自己的名字, 如生物统计, 统计质量控制, 空间统计, 政府统计等等. 各个应用统计领域既有个性又有共性. 多数普遍应用的统计方法最初是为某一个应用领域而发展的, 然后为其它领域所利用. 这些统计方法和原理逐渐形成统计学的基础.

统计作为一门科学, 随着其应用的发展和深入, 涉及大量的数据及复杂的模型; 因而也需要先进的计算机和越来越多的数学. 事实表明, 数学和计算机的大量运用加速了统计学的发展, 也更新了统计学的面貌. 当前, 统计是计算机数值计算的最重要的用户. 今天的统计学如果没有计算机是不可想象的.

统计应用的广泛性既造就了一批为各个具体应用领域服务的, 并懂得该领域内容的统计学家, 同时也造就了一些相对独立于某一两项具体应用, 从事于研究具有普遍性的统计方法或原理的统计学家. 后者所研究的内容有时也被称为“数理统计.” 他们对目前广泛应用的大量的统计模型有着重要的贡献. 然而这些似乎“脱离”某一两个具体应用领域的表面现象以及他们所使用的复杂的数学工具, 使得有些人认为统计(或数理统计)就是数学或数学的一个分支. 实际上, 也的确有许多人把统计学当成数学来研究. 这些自然要引起一些争论. 这没有关系, 在数学和许多其它科学领域之间都不可能划出明确的界限.

从思维方式来说统计和数学在研究目标和思想方法上是有差异的, 数学是以公理系统为基础, 以演绎为基本思想方法的逻辑体系. 它属于少数可以和世界具体事物无关的自成体系的学科. 数学可以完全脱离实际而存在. 而其它科学均是以实际事物为研究对象的. 统计是为各个领域服务的, 它以归纳为其基本思维方式, 而且归纳和演绎并用. 统计是仅有的系统地研究推断的科学(Efron, 1990). 统计学科也仅有在实际应用中才能得到发展和提高. 如果没有应用, 统计没有存在的必要(Box, 1990). 多年来, 统计作为一个学科之所以如此硕果累累, 就是因为它有一个比数学还要广阔的思维基础(Shafer, 1990).

虽然大多数现代统计方法是由统计学家根据实际问题发展出来的,但是绝大多数的统计应用是由那些没有专业统计背景的实际工作者来实施的.新的统计问题一般也是由他们提出的.世界上有无数涉及统计的问题需要解决,问题在于是否有人知道这是统计问题.不能要求每一个统计学家都了解某一实际领域的细节.也不能要求实际工作者精通统计的所有方法.理想的情况是:实际工作者有统计的基本常识;当他们遇到问题时,能够识别该问题是否涉及统计.如果是统计问题而他们又解决不了,他们可以找到愿意了解该领域实践的统计学家去寻求支援.由此,应在实际工作者中尽可能地普及统计的基本知识,而统计学家应该对至少一个或几个实际领域有较深刻的了解.

§ 1.2 关于非参数统计

在初等统计学中,最基本的概念是总体,样本,随机变量,分布,估计和假设检验等.其很大一部分内容和正态理论相关的.在那里,总体的分布形式或分布族往往是给定的或者是假定了的.所不知道的仅仅是一些参数的值或他们的范围.于是,人们的任务就是对一些参数,比如均值和方差(或标准差),进行点估计或区间估计.或者对一些参数进行各种检验;比如检验正态分布的均值是否相等或等于某特定值等等;也有对于拟合好坏进行的各种检验.最常见的检验为对正态总体的 t 检验, F -检验, χ^2 和最大似然比检验等.

然而,在实际生活中,那种对总体的分布的假定并不是能随便作出的.有时,数据并不是来自所假定分布的总体;或者,数据根本不是来自一个总体;还有可能,数据因为种种原因被严重污染.这样,上一段所说的在假定总体分布的情况下进行推断的做法就可能产生错误的、甚至灾难性的结论.于是,人们希望在不假定总体分布的情况下,尽量从数据本身来获得所需要的信息.这就是非参数统计的宗旨.因为非参数统计方法不利用关于总体分布的知识,所以,即使在对于产生数据的总体信息极端缺乏的情况下,它也能很容易而又很可靠地获得结论.这时,非参数方法往往优于参数