应用统计学系列教材 Texts in Applied Statistics

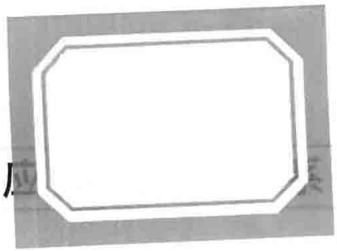
非参数统计（第2版）

Non-parametric Statistics (Second Edition)

王 星 褚挺进 编著

Wang Xing Chu Tingjin

清华大学出版社



应用统计系列教材 Texts in Applied Statistics

非参数统计（第2版）

Non-parametric Statistics (Second Edition)

王 星 褚挺进 编著

Wang Xing Chu Tingjin

清华大学出版社

北 京

内 容 简 介

本书是非参数统计教材，内容从经典非参数统计推断到现代前沿，包括基本概念、单一样本的推断问题、两独立样本数据的位置和尺度推断、多组数据位置推断、分类数据的关联分析、秩相关和分位数回归、非参数密度估计、一元非参数回归和数据挖掘与机器学习共计9章. 本书配有大量与社会、经济、金融、生物等专业相关的例题和习题，还配置了一些实验或案例. 方便结合R软件进行探索、研究.

本书可以作为高等院校统计、经济、金融、管理专业的本科生课程的教材，也可以作为其他相关专业研究生的教材和教学参考书，另外，对广大从事与统计相关工作的实际工作者也极具参考价值.

图书在版编目(CIP)数据

非参数统计/王星，褚挺进编著. --2版. --北京：清华大学出版社，2014
应用统计学系列教材
ISBN 978-7-302-37156-4

Ⅰ.①非… Ⅱ.①王… ②褚… Ⅲ.①非参数统计－高等学校－教材 Ⅳ.①O212.7

中国版本图书馆CIP数据核字(2014)第148335号

责任编辑：刘　颖
封面设计：常雪影
责任校对：王淑云
责任印制：王静怡

出版发行：清华大学出版社
网　址：http://www.tup.com.cn，http://www.wqbook.com
地　址：北京清华大学学研大厦A座　　邮　编：100084
社总机：010-62770175　　邮　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质　量　反　馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 刷 者：三河市君旺印务有限公司
装 订 者：三河市新茂装订有限公司
经　　销：全国新华书店
开　　本：170mm×230mm　　**印　张**：23.5　　**字　数**：458千字
（附光盘1张）
版　　次：2009年3月第1版　2014年9月第2版　**印　次**：2014年9月第1次印刷
印　　数：1～3000
定　　价：46.00元

产品编号：060604-01

第 2 版前言

习惯于用数据思考和决策的人都清楚, 和二三十年前相比, 现在的数据分析面临着更大的挑战. 在咨询领域, 数据误解、噪声数据、快速成像所产生的危害呈指数增长. 研究显示, 今天大数据分析所涉及的数据所呈现出的复杂特征并没有和几十年前小规模数据的特征有多大区别. 此外, 数据分析工具和封装的程序越来越容易获得, 令人兴奋的可视化技术越来越吸引年轻人的目光, 越来越技术化的数据分析孤立于通过观察并依循数据特点而进行的分析之外. 这些现象都表明我们的学生在尊重数据特点做出正确分析决定的能力方面训练不足.

经过五年多的等待, 《非参数统计》第二版终于面世了, 我很欣慰, 因为这次出版适逢大数据时代, 算作是我和我的团队献给我一直深爱的数据分析事业的一份礼物吧!

《非参数统计》第一版获得许多读者和同行青睐, 第二版在保留第一版全部优点和特色基础上, 作了许多优化、改进和创新. 这些优化、改进和创新包括:

(1) 内容进行了全面更新, 勘误了每一章, 扩充了 U 统计量理论, 添加了新的非参数回归内容.

(2) 可读性、易读性进一步提高. 为了做到这一点, 我们对每一个章节的每一个句子, 都经过了字斟句酌、反复推敲, 尽可能使用短句子, 同时, 继续邀请优秀的本科生参与试读教材, 充分听取他们的意见, 力争使第二版的内容更加生动、深入浅出和言简意赅.

(3) 调整结构体系, 将原来的第一章 R 基础调整至附录, 原来的十章依次分九章排列 —— 为每一章添加了一个实验或案例, 强调了结合问题背景根据复杂数据分布特点进行数据分析和信息解读的培养思想. 这些实验和案例可以激发学生的学习兴趣, 也为教师提供了丰富生动的教学内容.

在编写和修订的过程中, 对我支持最多的是我的家人和我的团队. 特别感谢我的助教王聪同学协助整理了大部分案例和勘误表, 许泳铎同学调整了部分实验 R 程序, 尤其是褚挺进老师加盟了我的教学团队, 协助修订了第 8 章和第 9 章, 最后, 还要感谢清华大学出版社编辑负责的编辑校对工作.

王 星

2014 年 6 月 10 日于中国人民大学应用统计中心 & 统计学院

第 1 版前言

统计是一个面向问题解决的、系统收集数据和基于数据做出回答的过程, 其本质是通过在随机现象中寻找分布规律回答现实问题的科学过程. 实际问题的复杂性和人类认知的局限性, 造成反映实际问题的数据在问题表示的充分性、代表性和分布的单一性等方面, 与传统的统计应用要求不相匹配, 于是催生了对数据分布假定宽松的非参数统计的兴起与发展. 尤其是最近 20 年来, 随着信息技术和网络技术的快速发展, 基于大量数据计算探索数据分布特点的数据分析方法层出不穷, 成为非参数统计发展的新主题, 代表着统计学未来的方向. 非参数统计自然成为连接统计学、信息学和计算机科学等交叉研究的桥梁, 共同推动数据分析和信息利用整体地向前发展.

本书是一本专门讲授非参数统计理论和方法的教科书. 内容主要分为两个部分: 传统的非参数统计推断和现代非参数统计方法. 传统的非参数推断内容由单一样本、两样本及多样本非参数统计估计和假设检验、分类数据的关联分析方法、定量数据的相关和回归等内容构成; 现代非参数统计方法部分包含非参数密度估计、非参数回归和数据挖掘与机器学习技术等内容.

本书的主要特色是结合 R 软件讲解非参数统计方法的原理和应用, 我们的宗旨是塑造有独立专业思考能力, 对所学知识有比较地选择, 并能够使用恰当方法解决实际问题的统计专业人才. 据此, 我们在课程设计中, 专门设计了学生在接受知识的过程中对知识的运用和鉴别能力的训练. 本书大部分例题都给出 R 源程序解法示例, 各种理论条件的检验、讨论、分析和比较, 鼓励学生针对数据的特点, 独立编写数据分析程序. 为加强与 R 的结合, 书中图形大部分由 R 生成, 我们广泛收集了很多领域数据分析实例和应用编写成本书的例题和习题, 以扩展学生的应用领域, 提高学生解决实际问题的能力.

本书可作为统计、经济、管理、生物等宏、微观专业领域本科三四年级以上学生以及相关研究人员学习非参数统计方法的教材, 也可以用作统计研究或从事数据分析的方法的参考书. 本书的先修课程只需具备初等统计学基础. 对统计基础略感陌生的读者, 可以阅读第 2 章相关内容作为补充. 本书的内容可以安排在一学期 54 课时内完成, 建议安排 10 课时左右用于学生上机实践. 本书备有丰富的习题, 兼有理论推导、方法应用和上机实践题目.

本书写作过程中, 得到众多老师的支持与鼓励. 感谢吴喜之先生多年来在非参数统计前沿和方法论上的引领和指导, 感谢袁卫、金勇进、易丹辉、张波、赵明德、

谢邦昌和郁彬等教授在学科发展动态上的启迪与建议, 感谢赵彦云、高敏雪等教授的支持与鼓励, 感谢朱建旭同学参与了第 10 章的部分编写, 协助整理了部分文献、图表和习题, 感谢孙兆楠、赵博元、詹瑾、李扬、王旭、王爱玲、伍燕然以及研究生讨论班的各位学生对部分内容进行的相关讨论, 感谢责任编辑王海燕和赵从棉, 正是凭借着她们对本书出版计划的坚定而耐心的支持, 才有本书的问世.

感谢我的恩师、朋友、学生和家人与我相伴的岁月!

王　星

中国人民大学统计学院

E-mail:wangxingscy@gmail.com

目　录

第1章　基本概念

1.1　非参数统计概念与产生

1. 非参数统计的概念

回顾数理统计基础知识可知, 分布是回答不确定性问题的基本统计工具, 对数据的分布做出推断是统计推断的根本任务. 典型的统计推断过程是从假定分布族开始的, 从数据到结论通常由 5 个步骤组成: 分布族假定, 抽样, 统计量和抽样分布, 推估和检验, 评价模型. 假定分布族是对实际问题的数学描述, 它是统计推断的基础. 比如, 研究某类商品的市场占有率, 假定在平均的意义之下, 每个消费者是否占有待研究商品来自两点分布 $B(1,p), 0<p<1$; 在研究保险公司的索赔请求数时, 可能假定索赔请求数来自 Poisson 分布 $\mathcal{P}(\lambda), 0<\lambda<\infty$(当然还可能有其他类型的分布假定); 在研究肥料对农作物产量的影响效果时, 假定平均意义之下, 每测量单元 (可能是) 产量服从正态分布 $N(\mu+x\beta,\sigma^2)$, 其中 x 是肥料的用量. 数据样本被视为从分布族的某个参数族抽取出来的总体的代表, 未知的仅仅是总体分布具体的参数值, 这样推断问题就转化为分布族的若干个未知参数的估计问题, 用样本对这些参数做出估计或进行假设检验, 从而得知数据背后的分布, 这类推断方法称为**参数方法**.

然而在许多实际问题中, 要对数据的分布做出具体的假定常常需要很多背景知识, 特别是在探索性的问题研究中, 人们往往对总体的信息知之甚少, 很难对总体的分布形式和统计模型做出相对比较明确的假定. 甚至在有些情况下, 能够对问题尝试数学描述本身就是问题的核心. 比如在人为控制因素不多的大部分经济和社会问题中, 数据的分布形态和数据之间的关系常常是不能任意假定的, 最多只能对总体的分布做出类似于连续型分布或者关于某点对称等一般性的假定. 这种不假定总体分布的具体形式, 尽量从数据 (或样本) 本身获得所需要的信息, 通过估计而获得分布的结构, 并逐步建立对事物的数学描述和统计模型的方法称为**非参数方法**.

问题 1.1(见光盘数据 chap1student.txt)　我们想比较两组学生的成绩是否存在差异, 传统的方法如 t 检验可以帮助我们分析问题. 但是应用 t 检验的一个基本前提是两组学生的成绩服从正态分布, 应用附录 A 中介绍的探索性数据分析方法绘制两组数据的分布, 如图 1.1 所示, 很难看出数据的分布是对称的. 这样, 应用 t 检验会有怎样的问题? 我们将在第 2 章回答这个问题.

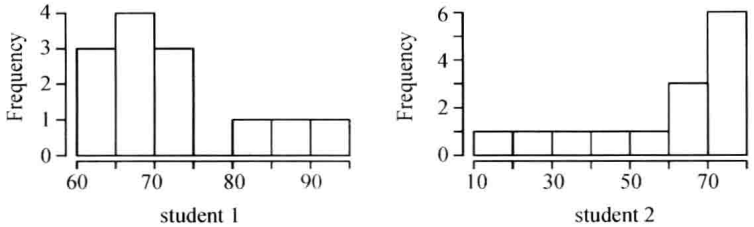


图 1.1　两组学生成绩的直方图

问题 1.2(见光盘数据 iqeq.txt)　我们希望比较两组被试的 IQ 成绩和 EQ 成绩之间是否存在着相关性, 传统的方法如 Pearson 相关系数检验可以帮助我们分析问题. 应用附录中介绍的探索性数据分析方法绘制两组数据的散点图, 如图 1.2 所示, 很难看出数据之间是否存在相关性. 这样的数据分布, 应用 Pearson 检验能测量出真实的关系吗? 我们将在第 5 章回答这个问题.

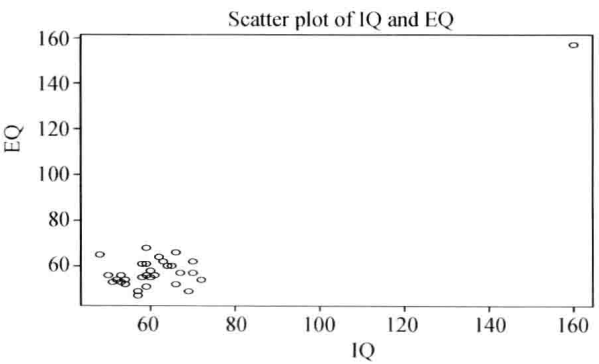


图 1.2　两组学生 IQ 成绩和 EQ 成绩相关散点图

问题 1.3　我们希望从光顾超市的用户购买清单数据中分析出哪些物品可能会被客户同时购买, 传统列联表分析能够给我们提供一些思路, 但是当物品数量很大的时候, 传统方法很难出现有效的结果. 我们将在第 5 章回答如何解决类似的问题.

以上这些问题, 并不总是能够在参数统计的框架结构中找到对应的方法, 数据驱动的方法会带领数据分析的实践者突破传统的框架, 思考如何对数据进行合理的运用. 总而言之, 非参数统计学是统计学的一个分支. 相对于参数统计而言, 非参数统计有以下几个突出的特点.

(1) 非参数统计方法对总体的假定相对较少, 效率高, 结果一般有较好的稳定性, 即不会由于总体分布与数据之间不一致导致发生大的结论性错误. 在经典的统计框架中, 正态分布一直是最引人瞩目的, 可以描述许多相对而言更为确定的问题.

比如：自动生产链处于稳定状态下的产品的质量. 然而, 正态分布并不是神话, 用于探索性问题时并不总是合适的, 随意对数据做出假定可能方便了计算和解释, 但可能产生错误的判断. 在某些推断问题中, 当数据不能支持显著性的结论, 常常表现为模型没有通过检验, 一些分析人士往往将原因归为信息量太少. 样本量不足可能是结论不显著的一个原因, 然而追加样本量在很多行业中代价是巨大的. 另外一个可能的解决方法是尝试更为宽松的模型假设, 即换用更有效的方法取代一味地增加样本量, 在节约成本和降低资源环境代价的条件下, 有效率地解决问题.

(2) 非参数统计可以处理所有类型的数据, 有广泛的适用性. 我们知道, 统计数据按照数据类型可以分为两大类：定性数据 (包括类别数据和顺序数据) 和定量数据 (包括等距数据和比例数据). 拿检验来说, 一般而言, 参数统计主要针对定量数据, 原因是理论上容易得到比较好的结果, 然而实践中, 我们所收集到的数据常常不符合参数统计模型的假定. 比如：数据只有顺序, 没有大小, 这时很多流行的参数模型无能为力, 尝试非参数方法是自然的. 即便对于定量数据而言, 也常常出现数据测量误差问题、不同分布数据混合问题, 此时传统的统计推断未必适用于噪声密集的数据环境, 如果将这些数据转化为顺序数据, 有可能弱化颗粒噪声的影响, 尝试用非参数方法分析, 甚至可能获得理想的结果.

(3) 非参数思想容易理解, 计算容易. 作为统计学的分支, 其统计思想非常深刻, 很多原理与参数统计思想平行, 容易发展生成算法. 特别是伴随计算机技术的发展, 最近的非参数统计更强调运用大量计算求解问题, 这些问题很容易通过编写程序求解, 计算结果也更容易解释. 非参数统计方法在小样本的时候, 可能涉及更多不常见的统计表, 过去会对一些非专业的使用者造成不便. 如今很多统计软件, 如 R 中都已提供现成的函数供人们计算和使用, 一些统计量的精确分布或近似分布都可以轻松地从软件中更为精确地得到, 取代纸质编制的粗糙且不精确的表.

当然, 非参数方法也有一些弱点, 如果人们对总体有充分的了解且足以确定其分布类型, 非参数方法就不如参数方法具有更强的针对性, 有效性可能会差一些. 所以非参数统计并非要取代参数统计, 它作为参数统计的一个有力的补充, 符合人类认识问题、解决问题的认知过程.

2. 学习非参数统计的意义

统计学研究的是从数据到结论的数据研究方法, 数据分析工作常常不是一个单纯方法的简单应用, 而是一个对数据内部规律认识的从无到有、从局部到整体的判断、推断和下结论的过程. 在这个过程中, 常常需要数据分析的综合思考, 如数据的收集、选择、分布推断等, 这个过程对于培养学生解决问题的能力非常必要.

大部分非参数统计方法的基本思想与参数统计思想平行, 很容易参照参数统计的内容进行比对, 可以增强学生对方法比较和选择的思考和训练, 扩充学生的知

识体系. 非参数统计可以补充学生在统计计算方面的训练, 有利于培养信息创新型人才.

统计学专业在西方著名大学中的主要招生对象是本科高年级学生或研究生, 这说明统计专业主要培养学生解决问题的能力, 特别是数据的处理、比较和分析能力, 能力体现在深度和广度两个层面, 很难想像一名仅会一两种方法的学生能够比较客观、恰当地运用数据分析工具协助实际部门解决各种复杂难题, 非参数统计则在思想的层面上扩展学生的专业方法选择能力, 在数据问题中提高学生动手解决问题的能力.

3. 非参数统计的历史

一般认为, 非参数统计概念的形成主要归功于 20 世纪 40—50 年代化学家 F.Wilcoxon 等人的工作. Wilcoxon 于 1945 年提出两样本秩和检验, 1947 年 Mann 和 Whitney 将结果推广到两组样本量不等的一般情况. 继 F.Wilcoxon 之后的 50—60 年代, 多元位置参数的估计和检验理论相继建立起来, 这些理论极大地丰富了试验设计不同情况下的数据分析方法, 小样本检验和异常数据诊断方面得到了成功的应用. Pitman 于 1948 年回答了非参数统计方法相对于参数方法来说的相对效率的问题, 1956 年, J.L.Hodges 和 E.L.Lehmann 则发现了一个令人惊讶的结果, 与正态模型中 t 检验相比较, 秩检验能经受住有效性的较小损失. 而特别对于厚尾分布所产生的数据, 秩检验可能更为有效. 第一本论述非参数应用的书也是在这个时代于 1956 年由 S.Siegel 撰写, 有人记载从 1956—1972 年, 该书被引用了 1824 次. 这也说明非参数统计在 20 世纪 60—70 年代的发展相当活跃.

20 世纪 60 年代, Hodges 和 Lehmann 从秩检验统计量出发, 导出了若干估计量和置信区间. 这些方法为后来非参数方法成功应用于试验设计数据开启了一道大门. 之后, 非参数统计的应用和研究获得巨大的发展, 其中较有代表性的是 60 年代中后期; Cox 和 Ferguson 则最早将非参数方法应用于生存分析.

进入 20 世纪 70—80 年代, 非参数统计获得了蓬勃的发展, 特别是 Efron 于 1979 年提出 Bootstrap 方法之后, 使得非参数方法借助计算机技术和大量计算获得更稳健的估计和预测, 因而在应用领域取得了长足的进展. 而以 P.J.Huber 以及 F.Hampel 为代表的统计学家从计算技术的实现角度, 为衡量估计量的稳定性提出了新准则. 20 世纪 90 年代有关非参数统计的研究和应用主要集中在非参数回归和非参数密度估计领域, 其中较有代表性的人物是 Silverman 和 J.Q.Fan.

20 世纪 90 年代后, 算法建模思想发展飞快, 成为非参数统计的新宠儿. Vapnik(1974) 等从学习的角度规范了预测模型选择的方法框架. Brieman L(1984, 2001) 和 Donoho(1994) 等为数据驱动的探索性构造模型敞开了大门, 大规模计算和自动化分析的需要将非参数统计引入机器学习领域, 如自适应模型选择、决策树、组合决策树 (Bagging, 装袋法;Boosting, 助推法)、随机森林以及关联分析等.

1.2 假设检验回顾

经典统计方法回答问题的基本逻辑是考察样本数据是否支持我们对总体的某种猜测. 这些没有被数据验证的猜测就是假设, 求证的过程就是假设检验. 比如研究问题:

(a) 新引进的生产过程是否优于旧过程?

(b) 几种不同的肥料哪一种更有效?

(c) 大学生的就业率与城市失业率之间是否存在关系?

从传统统计的观点来看, 这些问题都可以视为对不同分布总体的选择问题. 选择的依据首先可以从描述统计的图形和基本统计量上观察出一些现象, 但是关于分布间差异的粗略判断, 并没有揭示这种差异是本质的还是随机偶然因素造成的, 真实的情况只取其中之一, 而样本的表现则在两者之间. 检验就是希望能够找到一个合理地做出可靠性判断的临界点, 从而产生判别的条件和准则.

基本的假设检验原理是从两个互为对立的命题, 即假设开始的: 零假设和备择假设. 对这两个互为对立的假设而言, 事先还要假设分布族和数据, 比如: 假设分布族是正态的, 那么对总体的选择就可以简化为对参数的选择, 就像我们在大部分教科书中所看到的那样, 假设一般都以参数的形式出现, 这是参数统计的普遍特征. 比如: 在上述问题 (a) 中, 可以如此书写假设 $H_0 : \theta = \theta_0$, 称为零假设; 而新过程优于旧过程的猜测则描述成另一个假设 $H_1 : \theta > \theta_0$, 称为备择假设. 当然, 这里给出的是一个常规的单边检验问题. 类似地, 如果我们的猜测是另一个方向的或无倾向性, 则有单边检验问题 ($H_1 : \theta < \theta_0$) 或双边检验问题 ($H_1 : \theta \neq \theta_0$).

假设检验的基本原理是, 先假定零假设成立, 样本被视为通过合理设计所获得的总体的代表. 一旦总体分布确定, 那么抽样分布也就确定了, 从而理论上样本应该体现总体的特点, 统计量的值应该位于抽样分布的中心位置附近, 不能距离中心位置太远. 这显然是零假设成立的一个几乎必然的结果, 就像在理想环境下投一枚均匀硬币 100 次, 正面和负面出现的次数应近乎相等, 因为这是在均匀假设前提下的几乎必然的抽样结果. 反之, 如果真实情况是预先未知的, 我们需要通过实验推测真实情况. 比如, 在少量的实验中, 我们发现了正面和负面出现的次数之间出现了差异, 甚至较大的差异, 一种推翻均匀假设的想法油然而生. 用逆否命题进行推断是假设检验的本质. 当然, 差距的大与小需要测量, 如果样本量的值偏离抽样分布的中心位置过远, 则从小概率原理很难发生的统计观点出发, 认为有很大的把握怀疑这个离群的样本点 (outlier) 是从假定总体中取得的, 几乎必然地认为这些样本与备择命题更匹配, 从而拒绝数据对零假设的支持, 接受数据对备择假设的支持. “过远”是一个统计的概念, 我们用显著性来衡量. 几乎必然的含义是, 虽然拒绝