高等院校统计专业规划教材

非参数统计

Non-parameter Statistics

吴喜之 编

中国统计出版社

高等院校统计专业规划教材

非 参 数 统 计

吴喜之 主编

中国统计出版社

(京) 新登字 041 号

图书在版编目 (CIP) 数据

非参数统计 / 吴喜之主编.
- 北京：中国统计出版社，1999.12
高等院校统计学专业规划教材
ISBN 7-5037-2929-5

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 (1999) 第 10214 号

责任编辑 / 张美华
责任校对 / 刘开颜
封面设计 / 张建民
出版发行 / 中国统计出版社
通信地址 / 北京市三里河月坛南街 75 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 / (010) 63459084、63266600-22500 (发行部)
印　　刷 / 科伦克三莱印务 (北京) 有限公司
经　　销 / 新华书店
开　　本 / 850×1168mm　1/32
字　　数 / 170 千字
印　　张 / 6.75
印　　数 / 5000 册
版　　别 / 1999 年 12 月第 1 版
版　　次 / 1999 年 12 月第 1 次印刷
书　　号 / ISBN 7-5037-2929-5 / O. 33
定　　价 / 14.00 元

出版说明

“九五”期间是我国社会主义市场经济体制逐步完善和发展的重要时期，一方面，随着高等教育体制改革和统计改革的深入发展，对统计教育模式和统计人才培养目标都提出新的要求，另一方面，科学技术的飞速发展也促使统计技术发生了重大变革，新理论、新方法和新技术不断涌现并被应用于统计实践。为了适应这种新形势的需要，全国统计教材编审委员会制定了《1996—2000年全国统计教材建设规划》，根据《规划》的要求，编委会采取招标的方式组织全国有关院校的专家、学者编写了这批统计学专业“规划教材”。

这批教材力求以邓小平理论为指导，在总结“八五”期间规划统计教材建设经验的基础上，认真贯彻以下原则：①理论紧密联系实际的原则；②解放思想、转变观念、大胆探索、努力创新的原则；③正确处理继承与发展关系的原则。通过不懈努力，把这批教材建设成为质量高、适应性强、面向21世纪的新教材。

相信通过这批教材的出版、发行，对推动我国统计教育改革和加快更新、改造我国统计教材体系、教材内容的步伐将起到积极的促进作用，同时对我国统计教材建设也将起到较好的示范、导向作用。

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之处，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

1999年3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，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

目　　录

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, 和气象改善.

统计在每一个应用领域都有它自己的目标和特点, 有的还有

自己的名字,如生物统计,统计质量控制,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想,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一两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. 其很大一部分内容是和正态理论相关的. 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，或者是对某些参数值进行各种检验，比如检验正态分布的均值是否相等或等于零等等. 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或者甚至灾难性的结论. 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，就是在对于总体的任何知识都没有的情况下,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方法.然而, 在总体的分布族已知的情况下, 不需要任何先验知识就成为它的缺点;因为它没有充分利用已知的关于总体分布的信息,所做出的结论就不如参数方法得到的精确.

在不知总体分布的情况下如何利用数据所包含的信息呢? 一组数据的最基本的信息就是次序. 如果可以把数据点按大小次序排队, 每一个具体数目都有它的在整个数据中(从最小的数起)的位置或次序, 称为该数据的 秩(rank).数据有多少个观察值, 就有多少个秩. 在一定的假定下, 这些秩和它们的统计量的分布是求得出来的, 而且和原来的总体分布无关. 这样就可以进行所需要的统计推断了. 这是本书所讲的非参数统计的 基本思想.

注意, 非参数统计的名字中的"非参数(nonparametric)"意味着其方法不涉及描述总体分布的有关参数; 它被称为和分布无关(distribution-free), 是因为其推断方法 和总体分布无关;不应理解为与所有分布(例如有关秩的分布)无关.

1.3 假设检验的回顾

假定潜在的观察值(是随机变量)为$X_1,\ldots,X_n$. 一般用小写$x_1,\ldots,x_n$来记它的一个实现(是具体数值). (注意, 当不会发生误解时, 为了叙述方便, 我们并不一定总是强调观察值及其实现的区别.) 比方说, 要用这些观察值来对均值或某位置参数θ进行推断. 当人们觉得该数据可能成为θ大于某值θ_0的证据时, 就要进行假设检验了.在这种情况下, 零假设H_0要用人们希望拒绝的$\theta=\theta_0$, 而备择假设H_1用数据所可能支持的$\theta>\theta_0$. 这是个单边检验问题.类似地, 还有另外一个方向的单边检验问题(H_1:$\theta<\theta_0$) 和双边检验问题(H_1:$\theta\neq\theta_0$).

有了假设之后, 就要寻求和检验的目的有关的检验统计量 $T=$

$T(X_1,\ldots,X_n)$.它是观察值的函数. 观察值本身是随机变量,因而,作为观察值函数的检验统计量也是随机变量. 但是, 当已经获得观察值的具体数目$x_1,\ldots,x_n$之后, 就可以得到T的一个数值实现 $t=T(x_1,\ldots,x_n)$.我们必需能够(通过或者查表,或者用计算器或统计计算软件)计算在零假设下T落入和该值有关的某区间的精确概率或近似概率. 由于检验方法是由检验统计量 T 决定的, 通常也称 T 为检验. 在上面的检验位置参数的问题中, 如果 T 大就说明θ大;具体的说,要计算概率$P(T>t)$; 它称为$p-$值. 如果$p-$值很小,这说明这里的观察值的实现在零假设下是小概率事件. 如果拒绝零假设的话,犯第一类错误(H_0正确时拒绝它) 的概率也很小 (等于$p-$值). 这时的$p-$值可以作为显著性水平. 统计计算软件的输出通常就有$p-$值.

反之,如果$p-$值很大,则拒绝零假设可能犯错误的概率也大,因此不能拒绝.但是, 不能拒绝也可能犯第二类错误(H_1正确时不拒绝H_0).第二类错误无法用零假设下的概率来解释, 一定要考查H_1下的概率.在H_1正确时拒绝H_0的概率称为势(power). 势和检验统计量的选择很有关系. 势依赖于许多因素; 其中包括显著性水平,参数的真值,样本大小及检验统计量的选择. 一般来说,利用信息越多的检验统计量, 势越大. 在其它条件一样的情况下,势越大, 则该检验越有效.例如, 对于同样大的样本, 本书介绍的的符号检验就不如 Wilcoxon 符号秩检验势大, 因为后者利用了更多的信息. 但如果符号检验运用比其它检验更大的样本, 则它可以比其它检验有更强的势.

在位置参数的检验中, 人们(包括本书)往往把单边假设检验中的零假设写成不等式, 诸如 $H_0:\theta\leq\theta_0\Leftrightarrow H_1:\theta>\theta_0$ 而不是 $H_0:\theta=\theta_0\Leftrightarrow H_1:\theta>\theta_0$. 这是因为如果能够拒绝 $H_0:\theta=\theta_0$, 则更能拒绝 $H_0:\theta\leq\theta_0$ 的缘故. 关于$p-$值的检验统计量的概率计算是在$\theta\leq\theta_0$的范围内最不容易拒绝的θ_0点处计算的.

许多传统的问题事先给定一个显著性水平α, 这时,就要拿它

和$p-$值比较. 如果$p-$值小, 就可以拒绝零假设了;否则不能拒绝.

在不能拒绝零假设时, 要避免"在水平α时, 接受零假设"之类的说法. 在拒绝零假设时, 要认识到可能犯第一类错误的概率α. 而在提及"接受零假设"时, 一定要涉及(在备择假设正确时)犯第二类错误的概率. 然而,在实践中, 犯第二类错误的概率多不易得到; 这时,说"接受零假设"就容易产生误导. 实际上,不能拒绝零假设的原因很多, 可能是证据不足(比如样本太少),也可能是检验效率低, 换一个更有效的检验之后就可以拒绝了, 当然也可能零假设本身就是对的.

在数学上, 可以说零假设和备择假设是对称的, 接受和拒绝也是对称的.但是在统计实践中, 除了最简单的情况, 相对于简单的零假设, 备择假设往往不是简单的一点; 而检验的目的主要是拒绝. 这和在其它科学领域一样.一个科学的假设或模型(比如牛顿定律)仅仅是暂时地被接受,直到新的证据把它拒绝并代之以改进的新假设或理论(比如爱因斯坦相对论). 没有任何新的假设或模型能够证明是完全正确的, 它们之所以能被接受, 是因为能解释以前的模型不能解释的想象. 它们将存在到被新的模型取代为止.科学, 特别是物理学, 是在不断的否定或拒绝中发展的.

就单变量位置参数来说,置信区间一般来说是和双边检验有联系. 比如我们有均值μ的估计量$\hat{\mu}$; 它作为检验统计量去检验H_0:$\mu=\mu_0$ 对备择检验H_1:$\mu\neq\mu_0$. 这时, 如果显著性水平为α,则存在临界值C_α 使得在零假设下不拒绝的概率为 $P(|\hat{\mu}-\mu_0|<C_\alpha)=1-\alpha$. 这就导出了 μ的$100(1-\alpha)\%$ 置信区间的公式 $(\hat{\mu}-C_\alpha, \hat{\mu}+C_\alpha)$.虽然这里没有检验问题(没有给出具体的某个$\mu$的值), 但是可以认为, 如果$\mu_0$在此区间中, 则对于水平$\alpha$不能拒绝零假设.在置信区间的两个端点是随机的(是样本的函数)这个意义上, 人们说"该区间包括μ的概率是$1-\alpha$." 无论知道与否,总体均值μ是确定的数, 并非随机变量; 因此 "μ 落在该区间的概率是$1-\alpha$" 的说法是不合适的, 可能会使人认为μ 是随机的.当置信区间由实际数据计算出来之

后，它就成为一个固定的区间；它或者包含μ，或者不包含，没有任何概率可言. 当然，在贝叶斯统计中，作为随机变量的参数和置信区间的概念是和这里不同的.

关于连续性修正的注. 在实践中，当用连续分布去近似离散分布时，常常要用连续性修正(continuity corrections). 应用中最常用于近似的连续分布是正态分布.通常对于一个离散分布的点的概率$P(X=x)$用连续（正态）分布的相应的区间的概率$P(x-\frac{1}{2}\leq X\leq x+\frac{1}{2})$来近似. 对于离散分布的$P(X\leq x)$用连续分布的$P(X\leq x+\frac{1}{2})$来近似.它把离散集中的一个点$x$变换成一个单位区间$[x-\frac{1}{2}\leq X\leq x+\frac{1}{2}]$;相应的离散点的概率就变换成连续分布密度函数曲线下面在一个单位区间上的面积. 这种对x加或减$\frac{1}{2}$的调整就称为连续性修正.

在近似时最经常利用连续性修正的离散分布为二项分布，超几何分布和 Poisson 分布. 用来近似的正态分布都有和它们同样的均值和方差.例如对于二项分布$Bin(n,p)$随机变量X，概率$P(X\leq x)$由

$$\Phi\left\{\frac{x+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right\}$$

来近似. 对于超几何分布

$$P(X=x)=\frac{\binom{S}{x}\binom{N-S}{n-x}}{\binom{N}{n}}$$

概率$P(X\leq x)$由

$$\Phi\left\{\frac{x+\frac{1}{2}-np}{\sqrt{np(1-p)(N-n)/(N-1)}}\right\}$$

来近似，这里$p=S/N$.

另一个常用的连续性修正为Yate连续性修正，它是用χ^2分布来近似2×2列联表的精确概率. 对于一个2×2列联表

a	b	$a+b$
c	d	$c+d$
$a+c$	$b+d$	$n=a+b+c+d$

来说，进行度量的统计量为

$$\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)};$$

而Yate的修正度量为

$$\frac{n(|ad-bc|-n/2)^2}{(a+b)(c+d)(a+c)(b+d)},$$

它有近似的自由度为1的χ^2分布.

对各种离散分布的连续性修正，都有一些另外的建议，以适应于不同的情况和假定. 到底用不用，用哪一种，应按照具体情况而定.

1.4 χ^2检验简单回顾

在本书中许多检验统计量渐近地具有χ^2分布. 它们一般都被称为χ^2检验. 它们有共性，也有特性. 下面就和本书有关的一些χ^2检验作一介绍.这些内容大部分都可以在标准的初等统计书中找到.

χ^2检验一直应用于极其广阔的应用领域. 所有这些应用有一个共同点,这就是拥有足够大的样本使得在零假设下通过多元的