陈希孺 方兆本
李国英 陶 波/著
中国科学技术大学出版社

非参数统计

陈希孺文集

陈希孺 方兆本
李国英 陶 波/著
中国科学技术大学出版社

非参数统计

陈希孺文集

内 容 简 介

非参数统计是数理统计学中一个体系博大、理论精深且富有实用价值的分支，本专著对非参数统计的理论和方法进行了系统的论述，内容上有一定的广度和深度，经典全面，反映了本学科的现代面貌，语言表达具有简洁、朴实的特点，适合于高等学校概率和统计专业的本科生、研究生与老师，数理统计研究工作者以及具有相当数学水平的实用统计工作者阅读.

图书在版编目(CIP)数据

非参数统计/陈希孺，方兆本，李国英，陶波著. —合肥：中国科学技术大学出版社，2012.4

(陈希孺文集)

ISBN 978-7-312-02283-8

Ⅰ. 非… Ⅱ. ①陈…②方…③李…④陶… Ⅲ. 非参数统计—高等学校—教材 Ⅳ. O212.7

中国版本图书馆 CIP 数据核字(2011)第 227889 号

出版 中国科学技术大学出版社
安徽省合肥市金寨路 96 号，邮编：230026
网址：http://press.ustc.edu.cn

印刷 合肥晓星印刷有限责任公司

发行 中国科学技术大学出版社

经销 全国新华书店

开本 710 mm×960 mm 1/16

印张 20.25

插页 1

字数 370 千

版次 2012 年 4 月第 1 版

印次 2012 年 4 月第 1 次印刷

定价 38.00 元

总　　序

陈希孺先生是我国杰出的数理统计学家和教育家，1934年2月出生于湖南望城，1956年毕业于武汉大学数学系，先后在中国科学院数学研究所、中国科学技术大学数学系和中国科学院研究生院工作，1980年晋升为教授，1997年当选为中国科学院院士，并先后当选为国际统计学会(ISI)的会员和国际数理统计学会(IMS)的会士．陈先生的毕生精力都贡献给了我国的科学事业和教育事业，取得了令人瞩目的成就，做出了若干具有国际影响的重要工作，这些基本上反映在他颇丰的著述中：出版专著和教科书14部，统计学科普读物3部．在陈先生的诸多著作中，教科书占据重要的位置，一直被广泛用作本科生和研究生的基础课教材，在青年教师和研究人员中也拥有众多读者，影响了我国统计学界几代人．

陈希孺先生多年来一直参与概率统计界的学术领导工作，尤其致力于人才培养和统计队伍的建设．在经过“文革”十年的停顿，我国统计队伍十分衰微的情况下，他多次主办全国性的统计讲习班，带领、培养和联系了一批人投入研究工作，这对于我国数理统计队伍的振兴和壮大起到了重要作用．陈先生在中国科学技术大学数学系任教长达26年之久，在教书育人和学科建设等方面做出了重要贡献．中国科学技术大学概率统计学科及其博士点能有今天

这样的发展，毋庸置疑，是与陈先生奠基性的工作以及一贯的悉心指导和关怀完全分不开的.

陈希孺先生是我十分敬重的一位数学家. 1983 年我国首批授予博士学位的 18 人中，就有 3 位出自他的门下，一时传为佳话. 令人扼腕浩叹的是，陈先生已于 3 年前过早地离开了我们. 我坚信，陈先生在逆境中奋发求学的坚强意志，敦厚的为人品格，严谨的治学态度，奖掖后学的高尚风范，连同他的大量著作，将会成为激励我们前行的一笔非常宝贵的精神财富.

这次出版的《陈希孺文集》，是在陈希孺先生的夫人朱锡纯先生授权下，由中国科学技术大学出版社编辑出版的. 该文集收集了陈先生在各个时期已出版的著述和部分遗稿，迄今最为全面地反映了陈先生一生的科研和教学成果，一定会对学术界和教育界具有重要的参考价值. 值得一提的是，中国科学技术大学出版社决定以该文集出版的经营收益设立“陈希孺统计学奖”，我想，这应该可以看做我们全体中国科学技术大学师生员工对为学校的发展做出贡献的老一辈科学家和教育家的一种敬仰和感念吧！

中国科学技术大学校长
中 国 科 学 院 院 士

2008 年初冬于中国科学技术大学

前　言

非参数统计是数理统计学的一个分支，它形成于20世纪40年代，而在第二次世界大战以后得到迅速发展，至今已成长为一个体系博大、理论精深且富有实用价值的分支，受到数理统计学者和应用工作者的重视.

本书的目的是对非参数统计的理论和方法作一比较严格而系统的论述.本书属于专著性质，我们的主观愿望是在取材上能有一定的广度和深度，以反映本学科的现代面貌；另一方面，我们也希望本书的读者面能尽量广一些，因而在预备知识的要求方面以及在使用的数学工具的深度方面，不能不加以一定的限制，并舍弃一些过于专业的材料.因此，本书也带有“非参数统计引论”的性质.

本书的读者对象是高等学校概率和统计专业的本科生、研究生与教师，数理统计研究工作者以及具有相当数学水平的实用统计工作者.本书绝大部分内容所需要的预备知识是大学数学系的微积分（或比较扎实的非数学系的高等数学）及少量的矩阵知识，以及相当于复旦大学数学系主编的《概率论与数理统计》一书中的概率与统计知识，个别地方用到一点测度论和分析概率论的知识.

本书是根据我们以往在学习、讲授和研究这个分支过程中所积累的一些笔记资料并参考了国外近年出版的一些专著编写而成的.1982年夏初稿完成，曾在四川大学的

一个讲习班上试用过，在此基础上，我们又分头收集材料，做了整理和加工，最后由陈希孺执笔定稿，以利于在表述方式、行文和符号等方面达到统一.本书是集体工作的成果，但在最后定稿中出现的问题，由执笔者负责.由于作者们的水平所限，书中不妥之处在所难免，希望同行学者、专家及其他读者不吝指正.

作　者

目　　次

引　言

本书是一本非参数统计的专著.开宗明义第一件事,就是解释什么是非参数统计和非参数统计方法.不过,这个问题不是三言两语能说清楚的.关于本书在取材上的考虑、各章内容的安排及所用的符号也需作说明,故写了这篇引言.其中有的内容,须具备一些非参数统计的基础知识才能理解,不过初次接触时,浏览一遍就可以了.

先谈第一个问题.

顾名思义,"非参数统计"是"参数统计"的对立面.所谓**参数统计**,粗略地说,就是在一般教程中常见的一套基于正态假定的统计方法,如 χ^2 检验、t 检验和 F 检验,正态线性回归,狭义的多元分析等.Fraser 的著作[68] 有一个副标题,意谓非参数方法是当正态假定被一般假定取代时的统计方法.这种说法作为一个初步的解释是可以的,但作为非参数统计的正式定义则不可以.因为正态分布族固然是最重要的参数分布族,但不是唯一的重要参数分布族.

把以上的说法加以引申,并适当抽象化,就可以提出下面的说法,以作为划分参数统计问题与非参数统计问题的分界线:如果在一个统计问题中,所假定的总体分布族的数学形式已知,而只包含有限个(通常为数很少)未知的实参数,则这个统计问题是参数性的;否则,就是非参数性的.诸如可靠性统计中估计负指数分布和 Weibuil 分布参数的问题,产品抽样验收中对二项分布参数的检验问题等,都是参数统计问题.而确定连续分布的容忍限,一般的两样本检验问题(检验两组样本所来自的总体有相同分布),对称分布(形式未知)的对称中心的估计及形状未知的回归函数的估计问题等等,则都是非参数统计问题.常见的拟合优度检验问题也是非参数性的,因为在此问题中,理论分布是一个有待检验的假设而并非假定.

这个划分的准则是合理且明确易行的,在统计文献中多采取这个说法.本书作者也同意这一观点,但认为还不可完全拘泥于文字.例如,在 Gauss-Markov 线性模型$\{\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}, \mathrm{E}(\boldsymbol{e}) = 0, \mathrm{Cov}(\boldsymbol{e}) = \sigma^2 \boldsymbol{I}\}$中,并未对模型中的随机误差向量 $\boldsymbol{e}$ 的分布的具体形式作什么假定,故按上述准则,这模型的统计问题应是非参数性的,但一般并不这样看:在非参数统计著作中,都不把用最小二乘法处理这个模型的理论和方法包括进去.这是因为,在这个模型中,人们感兴趣的是回归系数向量 $\boldsymbol{\beta}$,它只涉及有限个实参数,且 $\boldsymbol{X\beta}$ 是一个简单的线性形式.又如极值统计,根据极值统计方法中对底分布(参看 1.2.3 小节)并无特定要求一节,可将其归入非参数统计范围,但是,极值分布只有三种简单的参数族,因此,把它列入参数统计范围,也言之成理.我们的结论是:在多数情况下,参数问题和非参数问题的划分是明确的;在少数情况下,则可因人的看法及侧重点的不同而异,这还要

考虑到习惯.

除了统计问题有参数与非参数之分以外,在文献中,也常把一定的统计方法划为"参数性"的或"非参数性"的.然而,要对此定出合理的划分标准,则问题更复杂了.大体上说,可以把非参数统计方法理解为"处理非参数统计问题的统计方法".但取之作为正式定义,仍觉有所不足.因为有些重要的统计方法在参数和非参数统计问题中都有用,矩法就是一个例子.又如最小二乘估计法,在形式上看更适合用于参数问题.但如把 Gauss-Markov 模型视为一种非参数模型,这个方法的主要应用则在非参数统计中.

本书作者倾向于采取一种较有伸缩性的观点:一种统计方法,如果主要用于非参数性的统计问题(包括其属性可有争议,但习惯上认为是非参数性的统计问题),则这个方法称为是非参数性的;否则是参数性的.按照这种观点,一个非参数统计方法也可以用于典型的参数统计问题中(反之亦然),但不是它的主要用武之地.一个典型的例子是秩方法(第 4 章),它是非参数统计的主要方法.但秩方法不仅可以而且确实也用于一些典型的参数统计问题中.置换检验(第 5 章)按其方法的性质及 Fisher 引进这种方法的思想,是非参数性的统计方法.又如,用次序统计量去构造连续分布的容忍区间,是一种非参数方法,但它也可用于正态分布的情况.不过对后者而言,人们一般使用形如 $\bar{x} \pm cs$ 的区间.又如,由于考虑到 Gauss-Markov 模型在习惯上并不视为非参数性的,故最小二乘估计法就不宜看作是一种非参数方法.

有的统计方法,在参数和非参数问题中都有重要应用,不能勉强将其划入某一类.这时,只好就具体情况而言:该方法适于参数应用还是非参数应用.例如矩估计法,在参数估计中是一个重要方法,但若只假定总体分布有均值,而其他一无所知,则除了用样本均值去估计总体均值外,别无其他良法,这种性质的应用并不限于直接用样本矩估计总体矩,例如,概率密度函数的核估计法(见 6.1 节),可以认为是一种导源于矩估计思想的方法.所有这些都可以说是"矩方法的非参数性应用".又如,二项分布的概率 p 的检验,无疑是一种参数统计方法,但它在非参数性的对比试验模型中也有重要应用(即符号检验,见 3.1 节),这可以说是二项分布参数检验法的非参数应用.

下面谈一下文献中关于这个问题的一些值得注意的观点:

Walsh 在文献[172]中把一个方法的非参数性归结为在应用上的广泛性.按照这种观点,二项分布参数的检验法是一种非参数方法,因为有许多问题最后都归结到这个检验问题.本书作者认为,应用的广泛性确实是非参数方法的一个特点(因为它对模型的要求少,应用自然就广了),但以此作为划分标准则不恰当,

因为，一种方法在应用上是否“广泛”，其看法因人而异，尤其重要的是：划分的标准应当着重于方法的性质，而不在于应用上是否广泛.

在另一个极端，Kendall 等[98] 主张，只能讲统计假设的“参数”或“非参数”性，而不能将这些形容词用于检验、统计量等.换句话说，问题有参数与非参数之分，而统计方法则否.他们认为，将一个检验称为“参数”或“非参数”这种提法，已造成了不少混乱.按照这种观点，下面这个两样本问题：$\{X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 分别是抽自总体分布 F 和 G 的随机样本，要检验假设 $F\equiv G\}$ 可称为非参数性的.而一个具体的检验方法，例如 Смирнов 检验（见 4.6 节）则不能冠之以“参数性”或“非参数性”.本书作者认为：这种看法似乎过于绝对化.因为大多数统计方法的应用有其重点所在.据此对方法的性质（参数或非参数）进行划分，不仅可能而且是有益的.

Conovor 在文献[49]中提出如下的看法：一个统计方法，如果适合下述条件之一，就可以称为**非参数方法**：① 它可用于属性数据；② 它可用于只分别大小次序而不计具体数值的数据；③ 它可用于通常的数据（有具体数值的），其所来自的分布族不能用有限个实参数刻画.我们认为：Conovor 这个提法明确指出了非参数统计方法的主要应用所在.但以此作为定义则似乎欠妥.比方说：最大似然估计与似然比检验可用于属性数据，但这些方法是参数性的.按 Conovor 的观点，最小二乘估计法是非参数方法，因为这方法主要用于 Gauss-Markov 模型，而后者的分布族不能用有限个实参数刻画.因此，我们认为 Conovor 的定义有些失之过宽.

美国统计学家常使用一个被称为“分布无关”（distribution-free）的概念，并理解为“非参数性”的同义语.所谓“分布无关”，一般是指下述情况：设样本 X 的分布 F 属于分布族 $\mathscr{F}$，而 $\mathscr{F}_0\subset\mathscr{F}$.设 $T=T(X)$ 为一统计量，若当 $F\in\mathscr{F}_0$ 时，T 的分布不依赖于 F，则称 T **相对于** $\mathscr{F}_0$ **为“分布无关”的**，例如设 $X=(X_1,\cdots,X_n)\sim N(\mu,\sigma^2)$，$\mathscr{F}=\{N(\mu,\sigma^2):-\infty<\mu<\infty,\sigma^2>\infty\}$，指定 μ_0，而作 t 统计量 $T=\sqrt{n}(\bar{X}-\mu_0)/s$，则对 $\mathscr{F}_0=\{N(\mu_0,\sigma^2):\sigma^2>0\}$ 而言，T 是分布无关的.因此，“分布无关”这个概念并不专用于非参数性的分布族，对参数族也很重要和常见.这个概念主要用于假设检验，$\mathscr{F}_0$ 是原假设，如果检验统计量 T 相对于原假设 $\mathscr{F}_0$ 为分布无关，就可以作出基于 T 的相似检验.一个显著的例子是秩统计量.在一些非参数检验问题中，往往在原假设成立之下，样本 $X_1,\cdots,X_n$ 为独立同分布且分布连续.这时秩统计量相对于原假设下的分布族为“分布无关”的.秩统计量的重要性正在于此.“分布无关”与“非参数”这两个概念不能混同（前者是指统计量的某种性质；后者是指统计问题和方法的性质）.不过，“分布无

关”这个概念更多的是用于非参数分布族之中，且在非参数方法中使用的统计量往往有“分布无关”性.因此，“分布无关”这个概念确实从一个重要侧面刻画了“非参数性”的含义.

关于将统计问题和统计方法划分为“参数”或“非参数”的标准问题，以上讲了作者自己的看法及文献中的一些提法.但我们并不是要把某一种意见(包括我们自己的意见)作为结论性的观点推荐给读者，而只是想提供一些材料，供读者思考这个问题时参考.

现在转到第二个问题.

由于对非参数统计和非参数方法的含义有不同的理解，故对于非参数统计中应当包含哪些题材的问题，看法也就不尽一致了.本书在取材问题上的指导思想，除了考虑到对非参数统计的理解及材料在理论和应用上的价值外，也还在一定程度上考虑到习惯与叙述上的方便，而不拘泥于一种想法.例如有关拟合优度的 χ^2 检验和列联表的内容，在性质上可归入非参数统计的范围，但在习惯上则多数是放到其他著作中论述.所以在本书中不包括这一内容.次序统计量的有些内容虽然明显地属于参数统计的范围，但为了照顾到系统性与叙述上的方便，也将其收进本书中，又如第 6 章中的稳健统计，有一种观点认为这个题材不属于非参数统计，作者也基本上同意这个观点，但考虑到稳健性与非参数的概念有密切联系，并且这方面的材料较新，而在中文文献中介绍得很少，因此将这一题材也收进本书.下面简介各章的内容.

第 1 章是次序统计量.把一组样本 $X_1,\cdots,X_n$ 按大小次序排列为 $X_{(1)}\leqslant\cdots\leqslant X_{(n)}$，后者(或其一部分)就称为**次序统计量**.这一章叙述了这种统计量的分布和渐近分布及其在统计问题上的重要应用.渐近理沦方面包含 3 个内容：样本分位数的渐近正态性；极值($X_{(1)}$ 和 $X_{(n)}$)的渐近分布；次序统计量的线性组合 $\sum_{i=1}^{n}C_{ni}X_{(i)}$ 的渐近正态性.这些结果有很强的理论和实际意义.

第 2 章是 $\bar{U}$ 统计量.它是 Hoeffding 在 1948 年引进的一类统计量(定义见 2.1 节)，在一些非参数性的估计和检验问题中有用.这一章介绍了这种统计量的定义、方差的计算与渐近正态性定理，列举了一些例子并说明其应用.同时介绍了关于其极限理论的若干深入结果，但由于它们的统计意义较小，因此不予详细论证.

第 3、4 章是基于秩的统计方法.这是非参数统计的中心内容.因此在本书中也给予了较大的篇幅.所谓秩，就是指各样本在其大小比较中所占的位次：若将样本 $X_1,\cdots,X_n$ 按大小排列为 $X_{(1)}<\cdots<X_{(n)}$，而 $X_i=X_{(R_i)}$，则称 X_i 的秩

为 R_i，而 $(R_1,\cdots,R_n)$ 称为**秩统计量**. 第 3 章主要讨论线性秩统计量的极限分布. 对于同分布的情况做了严格的论证；对不同分布和结存在的情况则只介绍有关结果(因为严格的证明过于繁复). 第 4 章可分为两部分：前两节是讨论秩检验在两种准则(渐近相对效率与局部最优) 下的优良性. 总的结论是：秩检验与传统的参数检验相比，处于有利的地位. 本章后四节讨论秩方法在各种统计问题中的应用 —— 估计问题、方差分析模型、独立性检验、Колмогоров 与 Смирнов 检验等. 应当指出：Колмогоров 检验并不是秩方法，因其与 Смирнов 检验(它是一个秩检验) 的密切联系而放在这里.

第 5 章的主要内容是置换检验，又称随机化检验，是 Fisher1935 年在其名著[66] 中引进的，它基于 Fisher 所提出的试验设计三原则之一 —— 随机化原则. 置换检验有重要的理论意义，因为它把方差分析的理论置于更现实的基础上. 本章通过较多的例子详细分析了条件检验与置换检验的基本思想，对其大样本理论(包括大样本极限分布及渐近效率) 作了严格的处理.

第 6 章是关于稳健统计. 这也是比较新的题材. 本章所选择的材料，主要是关于稳健性的基本概念以及较富于实际意义的理论和结果. 这方面的材料可认为是比较定型的.

本书所用的符号，大多数在统计文献中是标准的，例如用 E,Var,Cov(或 COV) 记随机变量的均值、方差与协方差(协方差阵). 用

$$X_n \xrightarrow{P} X, X_n \longrightarrow X, \text{a.s.}(\text{或} \lim_{n\to\infty} X_n = X, \text{a.s.})$$

分别表示：当 $n \to \infty$ 时，变量 X_n 依概率或以概率 1 收敛于变量 X，若 X_n 的分布为 F_n，而 F 为一分布函数，则以 $F_n \xrightarrow{\mathscr{L}} F$ 或 $X_n \xrightarrow{\mathscr{L}} F$ 表示 F_n 或 X_n 依分布收敛于 F；以缩写记号“iid.”(independently-identically distributed) 表示“独立同分布”. 若某一总体的总体分布为 F，而 $X_1,\cdots,X_n$ 为抽自该总体的独立随机样本，则常记为 $X_1,\cdots,X_n \sim F$. 有时也用记号“$X \sim F$”表示“变量 X 有分布函数 F”；$X \overset{\mathrm{d}}{=} Y$ 表示 X 与 Y 的分布相同；又 I_A 或 $I_A(x)$ 表示集 A 的指示函数，即在 A 上的值为 1 而在 A 外为 0.

本书也涉及少量的矩阵向量运算，按一般的表示法：当将 X 视为向量时，总是看作列向量，向量或矩阵 $\boldsymbol{A}$ 的转置记为 $\boldsymbol{A}'$，故 $\boldsymbol{X}'$ 表示行向量. 向量 $\boldsymbol{a}$ 与 $\boldsymbol{b}$ 的内积记为 $\boldsymbol{a}'\boldsymbol{b}$. 向量 $\boldsymbol{a}$ 的长度常记为 $\|\boldsymbol{a}\|$. 在任何特定意义下，两点 x 与 y 的距离也常记以 $\|x-y\|$.

第 1 章

次序统计量