纯粹数学与应用数学专著 第38号

数论方法在统计中的应用

方开泰 王元 著

科学出版社

纯粹数学与应用数学专著 第33号

数论方法在统计中的应用

方开泰 王 元 著

科 学 出 版 社

(京)新登字092号

内 容 简 介

本书描述了一系列应用统计问题，这些问题都可以用数论方法来解决．阐述了计算各种统计分布的概率与矩的数值方法，均匀设计方法．介绍了SNTO程序用来寻找有界闭区域上一个连续函数整体极大与极大值点．给出了一些寻找许多著名多元分布的代表点的方法，及数论方法与投影追踪法的结合，统计推断的一些新方法．附录中给出维数18之内 *glp* 集合的生成矢量，均匀设计表，及一些理论结果的证明．

本书英文版由英国 Chapman and Hall 出版．

本书可供大学统计系学生、研究生、教师阅读，也可供科研工作者、应用统计工作者参考．

Number-theoretic Methods in Statistics
Chapman and Hall, 1994

纯粹数学与应用数学专著　第33号
数论方法在统计中的应用
方开泰　王元　著
责任编辑　毕　颖
科学出版社出版
北京东黄城根北街16号
邮政编码：100717
中国科学院印刷厂印刷
新华书店北京发行所发行　各地新华书店经售
*
1996年1月第一版　开本：850×1168 1/32
1996年1月第一次印刷　印张：10 5/8 插页：2
印数：1–1000　字数：273000

ISBN 7-03-004886-5/O·822

定价：25.00元

序 言

这本书包含了作者过去 17 年的某些研究工作. 我们的工作开始于 1976 年, 在那时, 第一位作者正在研究合金钢标准化问题, 其中需计算很多 5 维正态分布的概率, 因通常的近似计算多重积分的方法都显得无效, 于是他询问第二位作者关于数值积分的数论方法问题. 这样一来, 他的问题被数论方法加以解决了. 数论方法 (NTM) 的实质是在 s 维单位立方体上找到一个点集, 它是均匀散布的, 这个集合可以用来代替蒙特卡罗方法中的随机数. 这两位作者随后就意识到这个想法还可以用于一些其他问题, 例如, 试验设计中的问题.

1978 年, 中国导弹研究部门向作者提出一个试验设计问题, 其中共有六个因子, 每个因子至少要考虑 12 个水平. 因实验耗费很昂贵, 他们要求设计出不超过 50 个实验的有效试验设计. 这就引导了我们研究如何利用数论方法来安排试验的问题, 于是我们建议了一个所谓“均匀设计”的试验设计方法. 将均匀设计用于这个问题, 每个因子有 31 个水平, 只安排了 31 次实验即得到了满意的结果. 此后, 均匀设计逐渐在中国的农业、纺织工业、手表工业、自然科学与军事科学研究等得到普及与应用.

1978 年, 另一个中国工业部门要求我们解决一个典型的几何概率问题. 一方面很难得到这个问题的解析解, 另一方面, 如果用蒙特卡罗方法进行模拟, 则精度又欠佳, 所以我们又用数论方法来尝试, 其结果与蒙特卡罗方法得到的结果相比较, 机器时间仅为 1/100 而精度却提高了一位数字. 受到这些成果的鼓舞, 我们决定在北京中国科学院中组织一个“数论在统计中的应用”的研究小组. 一些年轻研究生魏刚、袁克

海与张金廷、副研究员方碧琪、教授朱尧辰是我们小组的成员，他们的一些工作也包含在本书中. 1991 年，我们得到香港政府基金的资助，叶扶德、朱力行、吴启宏博士等参加了研究并获得了许多新的结果.

本书将阐述数论方法的思想及其在统计中的应用. 因为我们主要着眼于应用，所以一些有用的结果的证明仅仅放在本书的附录 B 中. 关于有关的数论结果，读者可以去看华罗庚与王元〔Hua and Wang (1981)〕的专著. 本书只假定读者有微积分的背景及大学统计课的水平.

我们在第一章中描述了一系列应用统计问题，它们可以作为“数论在统计中的应用”的典型问题背景. 所有这些问题都可以用数论方法来满意地加以解决. 此外，我们将叙述数论方法的含义及在 s 维立方体、球面、球体与单纯形上寻找均匀散布点集的方法. 在第二章中，我们将阐述计算各种统计分布的概率与矩的数值方法，这些分布中也包括信息数据所要求的在一个球与单纯形上的统计分布. 在第三章，我们介绍一个所谓的 SNTO 程序用来寻找有界闭区域上一个连续函数的整体极大与极大值点. SNTO 可以用来处理多峰函数的最优化问题，而且不要求函数有连续微商. 我们还将给出 SNTO 在极大似然估计、似然比统计量、非线性回归、稳健回归等上的许多应用. 一个连续多元分布常常要求一个有限点集作为它的代表，而这个集合则要求包括分布较多的统计性质. F-偏差与均方差被用来作为点集的代表性的度量. 用这两个度量，我们在第四章中给出一些寻找许多著名多元分布的代表点的方法，其中包括椭球对称分布，多元 l_1- 范数对称分布及多元 Liouville 分布类等. 在这章中，我们还给出数论方法在统计模拟问题上的应用，所以一些困难的几何概率问题可以用这种办法来解决. 在第五章，我们介绍均匀设计方法的思想. 这个方法也可以用于带混料试验，其中因子是互依的. 在最后一章，我们给出数论方法与投影追踪法的结合，然后给出

统计推断的一些新方法，例如多元正交性的检验、平均矢量的稳健估计等. 一些结果是首次在本书发表的. 为了帮助读者了解数论方法的理论及其应用，在本书附录中，我们给出维数 18 之内 *glp* 集合的生成矢量、均匀设计表，及一些理论结果的证明.

我们衷心感谢以下机构，他们对我们提供了经费支持，或电脑与其他设施的支持：中国国家自然科学基金会；中国科学院数学研究所及应用数学研究所；香港 UPGC 基金，香港大学 Sino-British 交换学者基金，香港中文大学伟伦基金与联合基金，香港浸会学院科学基金，香港政府基金，纽约大学(石溪) CEEC Glorious Sun 学者基金，及美国公共健康中心的 DA01070 基金. 本书的一些材料，作者已在 1990 年第一届亚洲数学大会， 1990 年国际泛华统计学会召开的第一届统计最近发展及 1992 年多元分析及其应用国际会议上的邀请报告中发表过；也在 North Carolina 大学 (Chapel Hill) 、 Florida 大学、 California 大学 (Los Angeles) 、香港中文大学及香港大学等报告过. 在此我们对很多参加者的宝贵意见致以感谢. 对于 Stanford 大学的 T.W.Anderson 与 I. Olkin 教授， California 大学 (Los Angeles) 的 P.M. Bentler 教授，Toronto 大学的 P.A.S. Fer 教授， Georgia 技术学院的 Y.L. Tong 教授与 Maryland 大学的 Grace Yang 教授的鼓励与帮助，我们致以衷心感谢. 作者由衷感谢香港浸会学院副校长曾宪博教授与理学院院长吴清辉教授的支持与鼓励. 吴启宏博士对本书给予了很多宝贵的帮助. 秦宇红女士承担了中文版的排版工作，在此深表谢意.

方开泰

香港浸会学院数学系

北京，中国科学院应用数学研究所　100080

王　元

北京，中国科学院数学研究所　100080

目　录

第一章　数论方法导引

§ 1.1　统计问题

统计在 20 世纪有了巨大的发展，许多有效的统计方法成为各门科学研究的强有力工具．当今自然科学、社会科学、工业与商业中不断涌现出新问题，这些问题有下面的特性：(a) 多元性，(b) 非正态性，(c) 非线性，(d) 有异常数据与 (e) 有丢失数据．这些复杂性需要我们发展更为强有力的统计方法，从而需要愈来愈多的纯粹数学的工具与结果．近 15 年来，微分几何与群论在统计中有广泛的应用并获得丰硕的成果，例如 Efron (1975), Amari (1985, 1987), Eaton (1989) 与 Diaconis (1989). 本书旨在介绍所谓的 **数论方法** (NTM), 它可以用来解决一系列统计问题，其中有些问题是人们长期以来试图用其他方法，但并未获得满意解答的． NTM 或称 **伪蒙特卡罗方法** 是数论与近似分析交叉的产物．就像许多交叉学科一样，有其自身的魔力与吸引人之处．数论方法的广泛应用导源于数值积分，并成功地应用于与之有关的问题如插值法，积分方程与微分方程的数值解法等． Korobov (1963), 华罗庚与王元 (1961, 1963, 1981) 与 Niederreiter (1978b, 1988, 1992) 给出了这一方法详细的文献．

尽管 NTM 与 **蒙特卡罗方法** 有密切关系，但在过去，只有少数统计学家研究数论方法及其在统计中的应用． NTM 在统计中的第一个应用自然是计算多元分布的概率与矩 (例如方开泰与吴传义 (1979), 张尧庭与方开泰 (1982)). 方开泰 (1980) 及王元与方开泰 (1981), 首先将 NTM 的想法用于试验设计．他们提出了一个新的设计，即均匀设计，这或许是数值积分之后， NTM 在统

计中的第一个应用. 自此以后，均匀设计在中国的许多领域均获得了应用与成果. Shaw (1988) 给出 NTM 用于贝叶斯 (Bayes) 统计的详细讨论，特别是关于后验分布的数值计算. 最近王元，方开泰，袁克海与 Bentler (王元与方开泰 (1990a, 1990b, 1992), 方开泰与王元 (1991) 及方开泰，袁克海与 Bentler (1990, 1992)) 系统地研究了 NTM 在统计中的应用. 这一方向吸引了愈来愈多的统计学家及从事应用统计学的人，他们的经验表明 NTM 是一个强有力的工具.

读者可能有这样的问题： 什么是数论方法？它是怎样用于统计中的各种问题的？我们将在本节介绍本书中所要讨论的一些统计问题及解决这些问题的主要想法，并阐明所有这些问题均可以归结为一个关键问题：如何在 s 维单位立方体 C^s 上找出一个均匀散布的点集？

注意这里所提的点集“均匀散布”的大意是指该点集在 Weyl (1916) 意义下的偏差较小，而不是统计意义下的均匀分布集合.

1.1.1 多元分布的概率与矩的计算

假定 s 维随机矢量 $\boldsymbol{x}$ 有一个连续**概率密度函数** (p.d.f.) $p(\boldsymbol{x})$, 其落在区域 D 上的概率为

$$p = \int_D p(\boldsymbol{x})\, d\boldsymbol{x}. \tag{1.1.1}$$

通常 D 为一个 s 维矩形 $[\boldsymbol{a},\boldsymbol{b}] = [a_1, b_1] \times \cdots \times [a_s, b_s]$, 所以

$$p = \int_{a_1}^{b_1} \cdots \int_{a_s}^{b_s} p(x_1, \cdots, x_s)\, dx_1 \cdots dx_s. \tag{1.1.2}$$

这一积分可以化为标准形式

$$I(f) = \int_{C^s} f(\boldsymbol{x})\, d\boldsymbol{x} = \int_0^1 \cdots \int_0^1 f(x_1, \cdots, x_s)\, dx_1 \cdots dx_s. \tag{1.1.3}$$

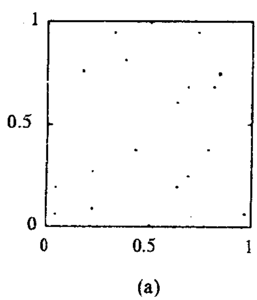
此处 $f(\boldsymbol{x})$ 为单位立方体 C^s 上的连续函数. 如果得不到 I 的解

析表达式，则我们通常用蒙特卡罗方法中的 **样本均值法** 去近似计算 $I=I(f)$, 即

$$I \approx \frac{1}{n}\sum_{k=1}^{n} f(\boldsymbol{y}_k), \tag{1.1.4}$$

此处 $\{\boldsymbol{y}_k\}$ 为 C^s 上均匀分布的一个随机样本，即 $\boldsymbol{y}_1,\cdots,\boldsymbol{y}_n$ 独立同分布 (i.i.d.) 并遵从在 C^s 上均匀分布 $U(C^s)$. 蒙特卡罗方法的效率不高，即只有当 n 很大时，由 (1.1.4) 才能得到好的逼近. 第二章将指出，当 $I(f^2)<\infty$ 时，在概率意义下，蒙特卡罗方法的平均收敛速度为 $O(1/\sqrt{n})$, 而在任何情况下亦不低于 $O(\sqrt{\ln(\ln(n))/n})$.

为什么蒙特卡罗方法的效率这样低？关键在于 $\{\boldsymbol{y}_k\}$ 在 C^s 上的散布不是很均匀的. 例如图 1.1 (a) 表示用蒙特卡罗方法在 $U(C^2)$ 中得到的一个大小为 17 的样本点图，而图 1.1 (b) 则为由 NTM 得到的 17 个点的点图. 易见后者在 C^2 上的散布较前者均匀. "均匀散布" 的精确定义将在 1.2 节中给出.



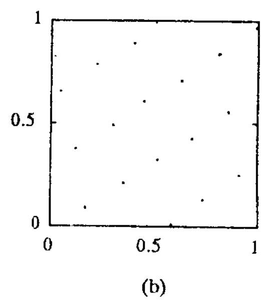


图 1.1 由蒙特卡罗方法与 NTM 生成点集的位置

命 $\mathcal{P}=\{\boldsymbol{c}_1,\cdots,\boldsymbol{c}_n\}$ 为由 NTM 得到的 C^s 上均匀散布的一个点集. 由于它们比由随机数得到的点 $\{\boldsymbol{y}_k\}$ 散布得更均匀，所以可以用它们来代替 (1.1.4) 中的 $\{\boldsymbol{y}_k\}$, 于是得到

$$I \approx I(f,n) = \frac{1}{n}\sum_{k=1}^{n} f(\boldsymbol{c}_k), \tag{1.1.5}$$

第二章中，我们将证明存在 $\mathcal{P}$ 使 $|I(f,n)-I| = O(n^{-1}(\log n)^{s-1})$. 这比由蒙特卡罗方法构造的 (1.1.4) 在概率意义下的误差阶 $O(n^{-\frac{1}{2}})$ 好多了.

显然我们可以应用 NTM 的数值积分方法来计算多元分布的函数，例如矩，后验概率与某些贝叶斯统计中的有关矩.

用 NTM 来近似计算 C^s 上多重积分的关键问题是选取一个 C^s 上均匀散布的点集. 为此目的，数学家们建议了不少有效的 NTM, 例如 Korobov (1959a, 1963), 华罗庚与王元 (1960, 1964, 1965), Hlawka (1962) 与 Halton (1960), 我们将在 1.3 节中介绍这些方法.

假定随机向量 $\boldsymbol{x}$ 在 R^s 的一个区域 D 上分布. D 可以是 s 维球面，球体或单纯形. 求 D 上的概率与矩的问题归结为下面形状的积分的数值计算问题

$$I(f,D) = \int_D f(\boldsymbol{x})\,dv, \tag{1.1.6}$$

此处 dv 表示 D 的体积元素. 例如当 D 为 s 维单位球面时，则研究 D 上的分布是 **方向数据分析** 的统计基础 (Mardia (1972), Watson (1983)). 若 D 为单纯形

$$\begin{aligned} T_s = \{(x_1, \cdots, x_s):\ & x_i \ge 0, \\ & i = 1, \cdots, s,\ x_1 + \cdots + x_s = 1\}, \end{aligned} \tag{1.1.7}$$

则来自 D 上一个分布的一个观察值就称为一个 **配方**. Aitchison (1986) 详尽地给出了 **配方数据** 的统计分析.

王元与方开泰 (1990a) 建议了下面的方法来近似计算 $I(f, D)$: **指标函数法**, **变换法** 与 **直接法**. 类似于蒙特卡罗方法的 **成败** (hit-miss) **法**, 我们可以发展所谓 **NT-成败法** (§4.7). 这些方法的关键为找一个在 C^s 或 D 上均匀散布点集. 我们将在第二章与第四章作详细讨论.

1.1.2 最优化与统计

最优化与统计之间有着密切的关联，统计中的许多问题，例如最大似然估计，回归分析中的参数估计，最优试验设计与最优量化等，均可以化为最优化问题.

假定 D 为 R^s 中的一个有界闭区域，$f(\boldsymbol{x})$ 为 D 上的一个连续函数. 今往寻求一个点 $\boldsymbol{x}^* \in D$ 使

$$M = f(\boldsymbol{x}^*) = \max\nolimits_{\boldsymbol{x}\in D} f(\boldsymbol{x}). \tag{1.1.8}$$

现在我们来讨论统计中出现的一些最优化问题.

(a) 最大似然估计 (MLE)

假定 $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$ 为来自总体 p.d.f. $g(\boldsymbol{x}, \boldsymbol{\phi})$ 的一个样本，此处参数 $\boldsymbol{\phi} \in \Theta \subset R^s$. 其中 Θ 称为**参数空间**. 则 $\boldsymbol{\phi}$ 的 MLE $\hat{\boldsymbol{\phi}}$ 为使**似然函数**

$$L(\boldsymbol{\phi}) = \prod_{i=1}^{N} g(\boldsymbol{x}_i, \boldsymbol{\phi}) \tag{1.1.9}$$

在 $\hat{\boldsymbol{\phi}}$ 达到其极大，即

$$L(\hat{\boldsymbol{\phi}}) = \max_{\boldsymbol{\phi}\in\Theta} L(\boldsymbol{\phi}). \tag{1.1.10}$$

当总体为正态分布时，$\hat{\boldsymbol{\phi}}$ 有解析表达式. 若总体分布为非正态时，一般说 $\hat{\boldsymbol{\phi}}$ 没有解析表达式，例如，威布尔分布，贝塔分布或哥西分布，这时 $\hat{\boldsymbol{\phi}}$ 只能由数值方法求得. 因似然函数一般都是非单峰的，所以由传统的方法很难求出 $\boldsymbol{\phi}$ 的 MLE.

(b) **似然比统计量**

考虑总体有 p.d.f. $g(\boldsymbol{x}, \boldsymbol{\phi})$ 的参数假设检验

$$H_0: \boldsymbol{\phi} \in \omega \subset \Theta, \quad H_1: \boldsymbol{\phi} \notin \omega, \quad \boldsymbol{\phi} \in \Theta, \tag{1.1.11}$$

其**似然比准则**为

$$\lambda = \frac{\sup_{\boldsymbol{\phi}\in\omega} L(\boldsymbol{\phi})}{\sup_{\boldsymbol{\phi}\in\Theta} L(\boldsymbol{\phi})}, \tag{1.1.12}$$

其中表达式的分子与分母均涉及到最优化问题.

(c) 稳健回归与非线性回归

考虑回归模型

$$EY = h(\boldsymbol{x},\boldsymbol{\phi}), \qquad \boldsymbol{\phi} \in \Theta \subset R^s, \tag{1.1.13}$$

此处 $\boldsymbol{\phi}$ 为回归参数矢量或回归系数矢量. 回归分析的主要任务之一是由观察值 $\{Y_i,\boldsymbol{x}_i, i=1,\cdots,N\}$ 来估计 $\boldsymbol{\phi}$. 命

$$Q(\boldsymbol{\phi}) = \sum_{i=1}^{N}(Y_i - h(\boldsymbol{x}_i,\boldsymbol{\phi}))^2. \tag{1.1.14}$$

则广泛使用的 **最小二乘估计** $\hat{\boldsymbol{\phi}}$ 满足

$$Q(\hat{\boldsymbol{\phi}}) = \min_{\boldsymbol{\phi}\in\Theta} Q(\boldsymbol{\phi}). \tag{1.1.15}$$

若 h 为 $\boldsymbol{\phi}$ 的线性函数，则 $\hat{\boldsymbol{\phi}}$ 有解析表达式. 否则我们只能用非线性最优化数值方法以求出最小二乘估计 $\hat{\boldsymbol{\phi}}$ (Ratkowsky (1983), Nash 与 Walker-Smith (1987)).

近年来 **稳健回归** 引起愈来愈多的注意，(1.1.14) 中函数 $Q(\boldsymbol{\phi})$ 常常被换成一个更为 **稳健的函数**

$$Q_1(\boldsymbol{\phi}) = \sum_{i=1}^{N}|Y_i - h(\boldsymbol{x}_i,\boldsymbol{\phi})|, \tag{1.1.16}$$

或更一般的

$$Q_2(\boldsymbol{\phi}) = \sum_{i=1}^{N}u(Y_i, h(\boldsymbol{x}_i,\boldsymbol{\phi})). \tag{1.1.17}$$

此处 $u(x,y) > 0$ 为满足某些性质的 x,y 的函数 (Huber (1972), Hampel 等 (1986)). 由于在 $Q_1(\boldsymbol{\phi})$ 的表达式中出现了绝对值运算，所以当用牛顿型方法来求 $\boldsymbol{\phi}$ 的 **稳健估计** $\hat{\boldsymbol{\phi}}^*$ 使

$$Q_1(\hat{\boldsymbol{\phi}}^*) = \min_{\boldsymbol{\phi}\in\Theta} Q_1(\boldsymbol{\phi}). \tag{1.1.18}$$

成立时，将会遇到困难. 通常参数空间 Θ 是 R^s 或矩形 $[\boldsymbol{a},\boldsymbol{b}]$, 此处 $\boldsymbol{a},\boldsymbol{b}\in R^s$. 有时 $\Theta\subset R_+^s$, 其中

$$R_+^s=\{(x_1,\cdots,x_s):x_i\geq 0,\ i=1,\cdots,s\}. \tag{1.1.19}$$

在这种情况下，无论是 $\boldsymbol{\phi}$ 的最小二乘估计或稳健估计都没有解析解. 好几个学者 (Boot (1964), Waterman (1974), 方开泰，王东谦和吴国富 (1982) 与方开泰与贺曙东 (1985)) 建议了不同的方法来处理这个问题.

所有上述的统计问题都是最优化问题 (1.1.8) 的特例. 在优化理论中有许多数值方法可以用于这个问题，例如牛顿 - 高斯方法，单纯形方法，截尾牛顿方法与共轭梯度方法. 对于这些方法，函数 f 需要假定为单峰，否则仅可得到一个局部极值. 但在统计中出现的函数往往不是单峰的. 进而言之，通常的方法还需假定 f 有一阶或二阶微商. 若在 $f(\boldsymbol{x})$ 的表达式中含有 “max”, “min” 与 $|x|$, 则在有些点上微分不存在. 最近涌现出许多寻求整体极值的方法 (Horst 与 Tuy (1990)). 另一方面，在过去的 20 年中蒙特卡罗最优化方法也很活跃 (Rubinstein (1986)). 但是我们可以将 NTM 用于最优化并用于统计. 这对于其他方法来说多了一个竞争对手.

下面的计算程序很简单：命 $\mathcal{P}_n=\{\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n\}$ 为 D 上的一个均匀散布点集. 命

$$M_n=\max_{1\leq k\leq n}f(\boldsymbol{x}_k)=f(\boldsymbol{x}_{k_n}) \tag{1.1.20}$$

为 f 在 $\mathcal{P}_n$ 上的最大值， $\boldsymbol{x}_{k_n}$ 表示 $\mathcal{P}_n$ 中的一个点它使 f 达到在 $\mathcal{P}_n$ 中的最大值. 显然当 $n\to\infty$ 时， $M_n\to M$, 所以当 n 充分大时有 $M_n\cong M$ 及 $\boldsymbol{x}_{k_n}\cong\boldsymbol{x}^*$.

使用最优化中的 NTM 有许多有利之处. 首先只需要假定函数 f 是连续的而不需要假定函数是单峰与可微的，所以这个方法特别适合于寻求有多个局部极大的函数的最大值，其次 NTM 只要计算函数 f 在 $\mathcal{P}_n$ 上的值，而不需要求 f 的微商值，所以