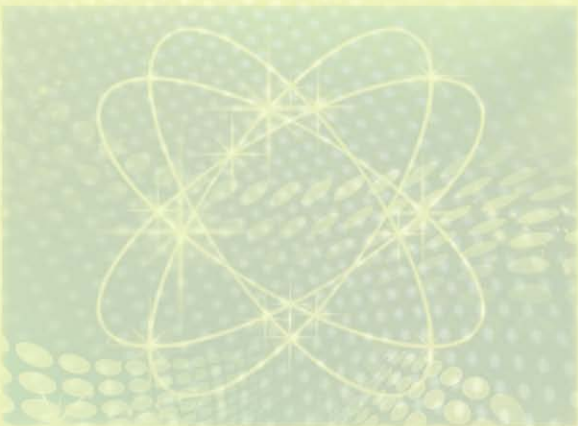
抽样方法与统计试验

邓国华 著



江西人民出版社

图书在版编目(CIP) 数据

抽样方法与统计试验 / 邓国华著. — 南昌 : 江西人民出版社, 2014.1
ISBN 978-7-210-06352-0

Ⅰ. ①抽… Ⅱ. ①邓… Ⅲ. ①抽样调查统计 Ⅳ. ①C811

中国版本图书馆 CIP 数据核字(2013) 第 312246 号

抽样方法与统计试验
邓国华 著
责任编辑: 徐 旻
出版: 江西人民出版社
发行: 各地新华书店
地址: 江西省南昌市三经路 47 号附 1 号
编辑部电话: 0791-86898965
发行部电话: 0791-86898801
邮编: 330006
网址: www.jxpph.com
E-mail: gjzx999@126.com
2014 年 7 月第 1 版 2014 年 7 月第 1 次印刷
开本: 787 毫米×1092 毫米 1/16
印张: 14.75
字数: 220 千
ISBN 978-7-210-06352-0
赣版权登字—01—2014—260

定价: 30.00 元
承印厂:
赣人版图书凡属印刷、装订错误,请随时向承印厂调换

序

科学计算中存在大量的这么一种现象: 问题的解析表达式能够清晰地写出,但因为所使用的函数是隐函数而无法得出精确的数值计算结果。这时有两种近似的解决办法: 其一是数学上的数值收敛迭代法,即先赋予自变量一个主观初始值,利用隐性方程找出其初始解,再以此初始解为二级初始值继续代入,得到一个二级初始解,又将此二级初始解代入……直至最后解收敛为止; 其二是统计上的试验模拟法,即先构造出一种统计模型,该模型中含有需要估计的参数,也含有至少一个随机变量,通过统计试验的方式让该随机变量取得一系列的取值,当取值数量不断增大时,极限理论可以证明此待估参数逐渐收敛于其真值。本书讨论的就是第二种方法,统计上称之为蒙特卡洛模拟运算方法(简记作蒙特卡洛马尔可夫链方法) 。

蒙特卡洛马尔可夫链方法的基本思想最初源于蒲丰在 1777 年提出的著名“蒲丰投针问题”的一项实验(Dörrie,1965) 。在这个著名的实验中,实验者向平行线网格间距为 D 的平面上投一长度为 l 的针($D>1$) ,理想条件下很容易计算出针与任意一条平行线相交的概率(假设为 π) 。因而,如果假设 p_N 为 N 次投针实验中针与平行线“相交” 的比率,则 $\hat{\pi}=\frac{(2l)}{(p_N D)}$ 可作为 π 的一个估计,并且当 N 趋于无穷时 $\hat{\pi}$ 收敛到 π。确实还真有一些研究者用

此方法来估算 π 的值。另一个典型例子是可以利用大数定律来求多维空间的定积分，如：$\lim_{n\to\infty}\int\cdots\int_{G_n} dx_1\cdots dx_n$。其中，$G_n=\{(x_1,\cdots,x_n):x_1^2+\cdots+x_n^2\leqslant\frac{n}{2},0\leqslant x_1,\cdots,x_n\leqslant 1\}$。这只要设想从一个均匀分布的总体抽样就可以，同样的方法可以应用于维尔斯特拉斯著名的多项式逼近定理的证明，且证明过程十分精巧。像这样借助模拟随机过程来估计某一有兴趣的量的思想现已成为科学计算的重要组成部分。

蒙特卡洛马尔可夫链的方法在现实科学问题中的系统应用始于电子计算的早期时代(1945—1955)，并伴随着世界上第一台可编程的“超大”计算机——MANIAC(数学分析机，数值积分器和计算机)——于第二次世界大战期间在洛斯阿拉莫斯(Los Alamos)的发展而不断发展。为了更好地使用这些具有快速计算能力的机器，科学家们(Stanislaw Ulam，John von Neumann，Nicholas Metropolis，Enrico Fermi 等)提出了一种基于统计抽样技术的方法，用以解决原子弹设计中的有关易裂变物质随机中子扩散的数值计算问题和估计 Schrodinger 方程中的特征根问题。这一方法的基本思想首先由 Ulam 提出，然后在他与 von Neumann 驾车从洛斯阿拉莫斯到拉米(Lamy)的途中，经两人仔细考虑后正式提出。据说，是 Nick Metropolis 将此方法冠名为“蒙特卡洛马尔可夫链(简称 MCMC)”的，该名称为推广使用这一方法起到了十分重要的作用。

早在 20 世纪 50 年代，统计物理学家们(N. Metropolis，Rosenbluth，M. Rosenbluth，A. Teller 和 E. Teller)就为简单流体的模拟引入了基于马尔可夫链的动态蒙特卡洛方法。这一方法被推广覆盖到越来越复杂的物理系统中，包括自旋玻璃(spin glass)模型、谐波型晶体和多聚体模型等。在 20 世纪 80 年代，统计学家与计算机科学家发展了用以解决诸如组合优化、非参数统计推断(如刀切法和自助法)、带有缺失观测值的似然计算、统计遗传分析、贝叶斯建模与计算问题的基于蒙特卡洛马尔可夫链的方法。在 20 世纪 90 年代，蒙特卡洛马尔可夫链方法在计算生物学中开始发挥重要作用，而且它

被用来解决序列基序识别和复杂的谱系(pedigree) 分析问题。现在,蒙特卡洛马尔可夫链方法的应用领域包括统计学(Efron, 1979; Gelfand & Smith, 1990; Rubin, 1987; Tanner & Wong, 1987) 、生物学(Petsko, 1990; Lawrence、Altschul、Boguski、Liu、Neuwald & Wootton, 1993) 、化学(Alder & Wainwright, 1959) 、计算机科学(Kirpatrik、Gelatt & Vecchi, 1983) 、经济学(Gourierourx & Monfort, 1977) 、工程学(Geman, 1984) 、材料科学(Frenkel & Smit, 1996) 、物理学(Metropolis、Posenbluth、Rosenbluth & Teller, 1953) 以及其他许多学科。在所有的蒙特卡洛马尔可夫链方法中,马尔可夫链蒙特卡洛理论为处理复杂的随机系统提供了巨大的机会,同时也是大分子学和其他物理系统研究中的中流砥柱。最近,由于蒙特卡洛马尔可夫链理论和技术能使统计学家考虑更复杂、更现实的统计模型,所以它引起了统计学家们的广泛关注。

许多不同学科领域的研究者由于受蒙特卡洛马尔可夫链方法的高度灵活性和超强功效性的吸引,都为它的发展做出了相应的贡献。然而,要了解任何一个领域的问题都需要大量丰富的特定专业领域的知识,这就大大地限制了不同领域中研究者的相互交流。近年来,大量的研究工作致力于重新发现在其他领域中已经发展出的各项技术。因此,迫切需要发展一个相对通用的框架,在此框架下,每个领域的科学家如理论化学家、统计物理学家、结构生物学家、统计学家、经济计量学家和计算机科学家,既能相互比较各自的蒙特卡洛马尔可夫链技术,又可以相互学习。许多把蒙特卡洛马尔可夫链模拟和有关全局优化技术(如模拟退火) 作为其研究工作中必不可少的工具的科学家和工程师们也需跟上蒙特卡洛马尔可夫链方法最新的发展步伐,同时也要了解各种蒙特卡洛马尔可夫链方法的性质和联系。本书的主要目的就是为读者提供一个有关蒙特卡洛马尔可夫链方法的自成体系的、统一的和最新的处理模式。

本书可以面向三类读者: 一是专门从事蒙特卡洛马尔可夫链算法研究的数理统计研究者; 二是对应用先进的蒙特卡洛马尔可夫链技术感兴趣的

人员;三是想学习蒙特卡洛马尔可夫链计算的统计学、计算生物学和计算机科学专业的研究生。要了解本书所述的方法至少必须具备下列知识:一个学期的概率理论课程和一个学期的统计理论课程,这两门课程的掌握都只需达到大学水平即可。当然,如果读者具有诸如人工智能、计算生物学、计算机视觉、工程学或者涉及繁重计算的贝叶斯统计等某一特定科学领域的一些背景知识则更为理想。本书特别适合作为大学高年级或研究生学习有关蒙特卡洛马尔可夫链方法课程的教材。该书重点阐述了蒙特卡洛马尔可夫链方法与科学和统计研究的关系。

本书的撰写工作得到许多人的帮助,包括本人所在学校的本科生与研究生,也包括本人的家人,是他们对本人的书稿进行不厌其烦的编辑和排版工作。可以这么说,若没有他们的帮助,本书的付梓出版至少要推迟半年。对此,本人表示深深的谢意。同时,由于涉足用统计抽样的方法解决数值计算问题这一领域时间不长,本人经验、水平均有限,书中难免有挂一漏万之处,欢迎各位同仁给予批评指正。

2013 年 10 月

目　录

第一章

模拟运算的统计学基础

1.1 统计抽样

抽样调查是一种非全面调查，它是从全部调查研究对象中抽选一部分单位进行调查，并据以对全部调查研究对象做出估计和推断的一种调查方法。显然，抽样调查虽然是非全面调查，但它的目的却在于取得反映总体情况的信息资料，因而，也可起到全面调查的作用。根据抽选样本的方法，抽样调查可以分为概率抽样和非概率抽样两类。概率抽样是按照概率论和数理统计的原理从调查研究的总体中，根据随机原则来抽选样本，并从数量上对总体的某些特征做出估计推断，对推断可能出现的误差可以从概率意义上加以控制。习惯上将概率抽样称为抽样调查。

1.1.1 抽样调查

抽样调查从研究对象的总体中抽取一部分个体作为样本进行调查，据此推断有关总体的数字特征。

抽样调查具有经济性好，实效性强，适应面广，准确性高等诸多特点。

抽样调查是根据部分实际调查结果来推断总体标志总量的一种统计调查方法，属于非全面调查的范畴。它是按照科学的原理和计算，从若干单位组成的事物总体中，抽取部分样本单位来进行调查、观察，用所得到的调查

标志的数据以代表总体，推断总体。

与其他调查一样，抽样调查也会遇到误差和偏误问题。通常抽样调查的误差有两种：一种是工作误差（也称登记误差或调查误差），一种是代表性误差（也称抽样误差）。但是，抽样调查可以通过抽样设计，通过计算并采用一系列科学的方法，把代表性误差控制在允许的范围之内；另外，由于调查单位少，代表性强，所需调查人员少，工作误差比全面调查要小。特别是在总体包括的调查单位较多的情况下，抽样调查结果的准确性一般高于全面调查。因此，抽样调查的结果是非常可靠的。

抽样调查数据之所以能用来代表和推算总体，主要是因为抽样调查本身具有其他非全面调查所不具备的特点，主要是：

（1）调查样本是按随机的原则抽取的，在总体中每一个单位被抽取的机会是均等的，因此，能够保证被抽中的单位在总体中的均匀分布，不致出现倾向性误差，代表性强。

（2）是以抽取的全部样本单位作为一个“代表团”，用整个“代表团”来代表总体，而不是用随意挑选的个别单位代表总体。

（3）所抽选的调查样本数量，是根据调查误差的要求，经过科学的计算确定的，在调查样本的数量上有可靠的保证。

（4）抽样调查的误差，是在调查前就可以根据调查样本数量和总体中各单位之间的差异程度进行计算，并控制在允许范围以内，调查结果的准确程度较高。

基于以上特点，抽样调查被公认为是非全面调查方法中用来推算和代表总体的最完善、最有科学根据的调查方法。

抽样调查的步骤：

（1）界定总体；

（2）制定抽样框；

（3）分割总体；

（4）决定样本规模；

(5) 确定调查的信度和效度;

(6) 决定抽样方式;

(7) 实施抽样调查并推测总体。

1.1.2 调查法的分类

(1) 随机抽样——简单随机抽样

简单随机抽样是一种最简单的一步抽样法,它是从总体中选择出抽样单位,从总体中抽取的每个可能样本均有同等被抽中的概率。抽样时,处于抽样总体中的抽样单位被编排成 1 ~ n 编码,然后利用随机数码表或专用的计算机程序确定处于 1 ~ n 间的随机数码,那些在总体中与随机数码吻合的单位便成为随机抽样的样本。

这种抽样方法简单,误差分析较容易,但是需要样本容量较多,适用于各个体之间差异较小的情况。

(2) 随机抽样——系统抽样法

这种方法又称顺序抽样法,是从随机点开始在总体中按照一定的间隔(即“每隔第几”的方式)抽取样本。此法的优点是抽样样本分布比较好,有好的理论,总体估计值容易计算。

(3) 随机抽样——分层抽样法

分层抽样是根据某些特定的特征,将总体分为同质、不相互重叠的若干层,再从各层中独立抽取样本,是一种不等概率抽样。分层抽样利用辅助信息分层,各层内应该同质,各层间差异尽可能大。这样的分层抽样能够提高样本的代表性、总体估计值的精度和抽样方案的效率,抽样的操作、管理比较方便。但是抽样框较复杂,费用较高,误差分析也较为复杂。此法适用于母体复杂、个体之间差异较大、数量较多的情况。

(4) 随机抽样——整群抽样法

整群抽样法是先将总体单元分群,可以按照自然分群或按照需要分群,在交通调查中可以按照地理特征进行分群,随机选择群体作为抽样样本,调查样本群中的所有单元。整群抽样样本比较集中,可以降低调查费用。例

如,在进行居民出行调查中,可以采用这种方法,以住宅区的不同将住户分群,然后随机选择群体为抽取的样本。此法优点是组织简单,缺点是样本代表性差。

(5) 随机抽样——多阶段抽样法

多阶段抽样是采取两个或多个连续阶段抽取样本的一种不等概率抽样。对阶段抽样的单元是分级的,每个阶段的抽样单元在结构上也不同,多阶段抽样的样本分布集中,能够节省时间和经费。调查的组织复杂,总体估计值的计算复杂。

(6) 非随机抽样——重点抽样

只对总体中为数不多但影响颇大(标志值在总体中所占比重颇大)的重点单位调查。

(7) 非随机抽样——典型抽样

挑选若干有代表性的单位进行研究。

(8) 非随机抽样——任意抽样

随意抽取调查单位进行调查(与随机抽样不同,不保证每个单位相等的入选机会),如:柜台访客调查,街头路边拦人调查。

(9) 非随机抽样——配额抽样

对总体作若干分类和样本容量既定情况下,按照配额从总体各部分进行抽取调查单位。

1.1.3 调查中的统计术语

(1) 总体

总体是指所要研究对象的全体。它是根据一定研究目的而规定的所要调查对象的全体所组成的集合,组成总体的各研究对象称之为总体单位。

(2) 样本

样本是总体的一部分,它是由从总体中按一定程序抽选出来的那部分总体单位所组成的集合。

(3) 抽样框

抽样框是指用以代表总体,并从中抽选样本的一个框架,其具体表现形