

“十二五”国家重点出版物出版规划项目



现代统计学系列丛书

抽样： 理论与应用（第二版）

金勇进

**Sampling:
Theory and Application**

高等教育出版社

“十二五”国家重点出版物出版规划项目



现代统计学系列丛书

抽样： 理论与应用（第二版）

**Sampling:
Theory and Application**

金勇进

Chouyang: Lilun yu Yingyong

高等教育出版社·北京

内容提要

本书主要讲述如何进行抽样,才能得到一套科学的样本;怎样进行抽样设计,才能使抽样的效率最高;在不同抽样方法下,估计量的抽样误差如何计算以及怎样对抽样的效率进行评估;最后介绍抽样调查应用中经常遇到的问题。本书有理论阐述,也有实践中应用问题方法的讨论,主要特色表现在:抽样的方法体系完整;增添一些常用方法的内容;对复杂样本的方差估计方法进行了应用性讨论;增加统计分析的内容;对调查中的缺失数据进行了专门的讨论。

本书可以作为统计专业学生抽样调查课程的教科书,也可以作为非统计专业学生和各类人员学习抽样调查方法的教科书或学习参考书。

图书在版编目(CIP)数据

抽样:理论与应用/金勇进编著. --2版. --北京:高等教育出版社,2016.1
(现代统计学系列丛书)
ISBN 978-7-04-044193-2

I. ①抽… II. ①金… III. ①抽样调查统计-高等学校-教材 IV. ①C811

中国版本图书馆 CIP 数据核字(2015)第 275741 号

策划编辑 张晓丽
插图绘制 黄建英

责任编辑 张晓丽
责任校对 李大鹏

封面设计 赵阳
责任印制 韩刚

版式设计 马云

出版发行 高等教育出版社
社 址 北京市西城区德外大街 4 号
邮政编码 100120
印 刷 北京东君印刷有限公司
开 本 787mm × 960mm 1/16
印 张 26
字 数 470 千字
购书热线 010-58581118
咨询电话 400-810-0598

网 址 <http://www.hep.edu.cn>
<http://www.hep.com.cn>
网上订购 <http://www.landraco.com>
<http://www.landraco.com.cn>
版 次 2010 年 8 月第 1 版
2016 年 1 月第 2 版
印 次 2016 年 1 月第 1 次印刷
定 价 40.30 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换
版权所有 侵权必究
物 料 号 44193-00

现代统计学系列丛书编委会

(按姓氏笔画排序)

主 编：方开泰

副主编：史宁中 何书元 陈 敏 耿 直

编 委：马 洪 方开泰 史宁中 杨 虎 何书元 何晓群
张爱军 张崇岐 陈 敏 郑 明 赵彦云 耿 直
曾五一 缪柏其

总 序

统计学是一门收集、整理和分析数据的科学和艺术。这里的“数据”泛指“信息的载体”，涵盖了大千世界中的文本、图像、视频、时空数据、基因数据等。统计学是一门独立的学科，在历史上曾隶属于数学，但统计学与数学有着本质的区别，因此统计学教育有其自身的特点和要求，这些特点表现为：(1) 统计学研究的是随机现象，而数学研究的是确定性的规律；(2) 统计学是一门应用性很强的学科，许多概念和原理来自于实际的需要，不是数理逻辑的产物；(3) 数据在统计学中扮演了重要的角色。目前，统计学已被列为一级学科。

在过去的 30 年中，随着生命科学、信息科学、物质科学、资源环境、认知科学、工程技术、经济金融和人文科学等众多学科的发展，产生了许多新的统计学分支，如风险管理、数据挖掘、基因芯片分析等。此外，计算机及其有关软件在统计教育 and 应用中扮演了越来越重要的角色，它们提供了越来越多的图形表达和分析的方法，使得许多原来教科书中重要的内容，现在已变得无足轻重。统计教育必须要改革才能适应高速发展的形势。

大学的统计教育可分为两大类，一类是非统计学专业的课程，另一类是统计学专业的教学设计。非统计学专业的学生学习统计的目的是为了应用，在大学阶段，课程不多，主要是学习基础的统计概念和方法，学会使用统计软件，培养其解决实际问题的能力。统计学专业的课程设置十分重要，应向国际靠拢，对教师队伍的要求也较高。虽然这两类学生的教育有很多共同点，但在课程设置中必须加以区分。

我国的统计教育在过去受苏联的影响很深，把统计学作为数学的一个分支，在内容上偏理论，少应用，过于强调概率论在统计中的作用。统计学是一门应用性很强的学科，应从实际问题、从数据出发，通过统计的工具来揭示数据内部的规律。用“建模”的思路来教统计，使学生能更加容易理解统计的概念和方法，知道如何将实际问题抽象为统计模型，反过来又指导实践。对非统计学专业的学生，要强调统计的应用。学生要能熟练地使用至少一个统计软件包。对于统计学专业的学生，要培养学生对实际问题的建模能力。有些实际问题可直接应用现有的统计方法来解决，如问卷调查的统计分析。有些问题在初次接触时并不像一个统计问题，必须有坚实的统计基础和对实际问题的洞察力，才能从中发

掘出统计模型。要培养学生的这种能力及统计思想(统计思想是统计文化的一部分,是用统计学的逻辑思考问题)。教师在授课中要结合较多的应用例子,要求学生做案例研究,鼓励学生参加建模比赛,参加企业的实际项目。

为满足我国统计教育发展的需要,我们计划编写一套面向高校本科生、特别是一般院校,适用于统计学专业和非统计学专业的系列教材。系列教材的编写宗旨是:突出教学内容的现代化,重视统计思想的介绍,适应现代统计教育的特点及时代发展的新要求;以统计软件为支撑,注重统计知识的应用;内容简明扼要,生动活泼,通俗易懂。编写原则为:(1)从数据出发,不是从假设、定理出发;(2)从归纳出发,不是从演绎出发;(3)强调案例分析;(4)重统计思想的阐述,弱化数学证明的推导。系列教材分为两个方向,一个面对统计学专业,另一个面对非统计学专业和应用统计工作者。

系列教材是适应形势的要求,由高等教育出版社邀请专家组成“现代统计学系列丛书编委会”负责选题、审稿,由高等教育出版社出版。

以上是我们编写这套教材的背景和理念,希望得到读者的支持,特别是高校领导和教学一线教师的支持。我们希望使用这套教材的师生和读者多提宝贵意见,使教材不断完善。

现代统计学系列丛书编委会



扫描二维码,获取更多丛书信息

第二版前言

《抽样:理论与应用》第二版是在第一版的基础上修订而成的。自本书第一版出版发行以来,受到广大读者的关注,许多学校把该书作为抽样调查课程的教材,使我们深受鼓舞,特此表示感谢!第一版出版发行至今已经4年有余,在使用过程中我们也发现了教材中的一些问题,需要进行修订。

在知识体系方面第二版没有做变动,保持了与第一版相同的框架结构和内容。我们认为,本书包括了抽样理论中最经典和最常见的一些抽样方法;此外,在第十章其他抽样方法中,本书还纳入了捕获再捕获抽样,随机化回答抽样技术,稀有事件抽样,以及自加权抽样设计和样本轮换设计,这些内容在其他同类教科书中还不多见。本书中的复杂样本方差估计和复杂抽样设计下的统计分析内容略有些难,但体现了现代统计方法在抽样领域的应用。如果由于课程时间限制,在讲授中这部分可以略去。本书中对调查中的缺失数据问题用一章篇幅进行了讨论,提出了一些处理方法。

本书的修订主要是针对一些内容的表述,修订后的表述更为准确;修订中还包括对一些公式符号,特别是对第一版公式中上下标和大小写错误的改正;此外,第一版中的一些例题不够合理,个别例题有计算错误,对此我们也进行了改正。

在这里我要感谢中国人民大学统计学院的研究生宋丹、朱新景、张婉璐、王俊、唐箬穗。他们对书中的大部分内容进行了认真的校阅,发现了公式和例题计算中的一些错误,并进行了修改;对个别有欠缺的例题进行了重新设计。正是在他们的协助下,第二版的修订工作才进行得非常顺利!在此,我谨向这些同学表示深深的感谢!

由于作者水平有限,尽管我们做了很大努力,但书中的错误和疏漏在所难免,敬请读者提出宝贵意见,以便进一步修改和完善。

金勇进

于中国人民大学

2015年6月

第一版前言

抽样的理论与方法是统计学的一个重要分支。作为一门课程,它主要讲授如何进行抽样,才能得到一套科学的样本;怎样进行抽样设计,才能使抽样的效率最高;在不同抽样方法下,估计量的抽样误差如何计算;怎样对抽样的效率进行评估等。目前,抽样调查在我国的应用越来越广泛,已经成为人们获取信息的最主要方式之一。但我们也必须承认,在许多的抽样调查项目中,抽样方案的设计不尽科学,用样本数据推估总体方法不妥,更缺乏对抽样误差的计算,难以对推断结果进行评价。这说明,我们在抽样技术的应用方面,就整体而言,与国外相比还有不小的差距,这也从一个角度反映出,我们在抽样调查的理论与方法的知识推广方面还有许多工作可做。

教科书的主要功能是进行现代科学知识和技术的推广。在我国,有关抽样调查的教科书也有一些,这些教科书为传授抽样方法发挥了重要作用。但有相当一部分的教科书,过于追求数学公式的罗列和推导,缺乏结合应用背景的论述和讨论;还有一些教科书,沿袭多年传统的体系,缺乏现代新技术、新方法的引进,给人以知识老化之感。本书试图在原有教科书内容的基础上,努力做出一些探索。本书的特色主要表现在:

1. 抽样的方法体系完整。本书除第一章绪论外,从第二章到第九章分别介绍简单随机抽样、分层抽样、比率估计和回归估计、不等概抽样、整群抽样、系统抽样、多阶段抽样和二重抽样,这些内容涵盖了传统抽样理论中的经典抽样方法,也是一般教科书中所包含的内容。在对这些经典方法的介绍中,本书力求全面和精炼,既讲清楚每种方法的基本原理,又尽量避免复杂的数学公式推导,努力结合实际应用的背景进行论述。考虑到读者的不同数理知识基础,一些公式的证明放到附录中,使核心内容的介绍更紧凑。

2. 增添一些常用方法的内容。本书第十章“其他抽样方法”中,主要介绍了对总体单元数的估计,包括捕获再捕获法、逆抽样法;对敏感问题进行调查的技术,包括属性特征敏感问题的抽样设计和数量特征敏感问题的抽样设计;对稀有事件进行抽样的方法;如何进行自加权抽样设计及样本轮换设计。这些内容在实际调查中会经常遇到,但是在我们已有的教科书中却很少提及。本书作者在参考大量文献的基础上,对这些内容进行了提炼和归纳,希望这些内容对抽样调

查的实践有所帮助。

3. 对复杂样本的方差估计方法进行了应用性讨论。本书第十一章是复杂样本的方差估计。目前许多抽样调查的教科书中,这部分内容是没有的。抽样调查的实际情况是,许多调查项目特别是大型的抽样调查项目,往往是复杂的抽样设计,从经典的抽样方法中,无法找到直接对应的估计量方差计算公式,这就需要针对特定的抽样设计,进行方差估计的公式推导,或者采用其他方法进行方差估计。在国外现代教科书中,这部分内容已经变得越来越不可或缺。国内一些学者意识到这个部分的重要,一些教科书中出现了这部分内容。本书在前人已经做出成果的基础上,试图有所前进。本章中,作者对复杂样本的方差估计方法进行了分类,对直接推导法进行了提炼式归纳,对不同方差估计方法的应用特点进行了比较性分析,希望能够丰富这个部分的内容。

4. 增加统计分析的内容。本书第十二章为“复杂抽样设计下的统计分析”。关于数据的统计分析方法常见于统计学的教科书中,本不属于抽样调查的教科书中的内容,所以,以往的教科书中是没有这部分内容的。但是,统计学教科书中所讨论的分析方法,都是针对简单随机抽样下产生的数据,而抽样调查的实践中,仅仅单纯使用简单随机抽样的情况是非常稀少的,实际常采用的是多种抽样方法的结合。这里把多种抽样方法结合产生的数据称为“复杂数据”。复杂调查数据通常很难满足观测值相互独立、服从于同一个总体分布等这些假定,因而这样的调查数据被录入计算机,不加任何处理,直接用标准模式的统计软件进行运算,无法保证结论的科学和正确。本章就是专门讨论如何对复杂数据进行统计分析的方法。这应该是抽样调查内容在数据处理、分析过程中的延续。过去,我们恰恰是忽略了数据采集和数据分析这个衔接点上存在的问题。

5. 对调查中的缺失数据进行了专门的讨论。本书第十三章是“调查中的缺失数据”,用一章篇幅专门讨论缺失数据是因为,缺失数据已经成为调查实践中面临的一个重要挑战,如果对缺失数据不进行任何有效处理,就有可能出现估计偏差和增大估计方差,如果问题严重,抽样调查的成果将付之东流。以往,我们对缺失数据问题重视不够,也研究不够,教科书中鲜有这方面的专门性讨论,现在看来,这个问题已无法回避,也不应回避。本书在介绍调查中缺失数据产生的原因、缺失数据的统计影响、缺失数据与样本量之间的关系之后,在缺失机制的基础上,重点讨论了处理缺失数据的两类方法,分别是加权调整法和插补法,以期对调查中的缺失数据起到补救作用。

本书可以作为统计专业抽样调查课程的教科书,也可以作为非统计专业学生和各类人员学习抽样调查方法的教科书或学习参考书。由于本书涉及内容较多,有理论阐述,也有实践中应用问题方法的讨论,不同背景的读者可以根据需

求自行取舍。

本书是集体劳动的成果,由我和我的博士生团队共同完成。其中,我负责全书编写大纲的设计,并负责第一章和第十三章的写作;其余各章由博士生们分别提供初稿,具体为:艾小青编写第二、五章,谢佳斌编写第三、十一、十二章,汤琰编写第四、十章,陶然编写第六、七章,孙欣、汤琰编写第八章,孙欣、陶然编写第九章,各章完成后又经过集体讨论和修改,最后由我做全书的修订和总纂。在本书写作中参阅许多文献,主要观点和方法的直接引用在书中有注释,并列于参考文献中,特向有关作者和出版社表示谢意。

在本书出版过程中,高等教育出版社数学分社社长李艳馥女士、编辑李蕊女士给予了大力支持,在此对她们表示由衷的感谢。

尽管我们尽了努力,但书中一定会有不少错误和缺憾。对于书中的不足,恳请各位专家和读者提出宝贵意见。

金勇进

2009年7月

目 录

第一章 绪论	1
1.1 调查与抽样调查	1
1.1.1 调查	1
1.1.2 抽样调查	2
1.1.3 抽样调查的作用	6
1.1.4 抽样调查与普查	6
1.1.5 抽样调查应用领域	7
1.2 基本概念	8
1.2.1 目标总体与抽样总体	8
1.2.2 抽样框与抽样单元	9
1.2.3 总体参数与统计量	10
1.2.4 估计量方差、偏倚、均方误差	11
1.2.5 抽样误差与非抽样误差	13
1.2.6 精度与费用	14
1.3 几种基本的抽样方法	14
1.3.1 简单随机抽样	14
1.3.2 分层抽样	15
1.3.3 整群抽样	15
1.3.4 多阶段抽样	15
1.3.5 系统抽样	16
1.4 抽样调查步骤	16
1.4.1 确定调研问题	16
1.4.2 抽样方案设计	17
1.4.3 问卷设计	17
1.4.4 实施调查过程	17
1.4.5 数据处理分析	18
1.4.6 撰写调查报告	18
习题一	19

第二章 简单随机抽样	21
2.1 引言	21
2.1.1 定义	21
2.1.2 样本性质	21
2.1.3 符号	23
2.1.4 抽样方法	24
2.2 估计量及性质	25
2.2.1 总体目标量的估计	25
2.2.2 估计量的性质	25
2.2.3 放回简单随机抽样下的估计	31
2.3 样本量的确定	33
2.3.1 费用限制	33
2.3.2 精度要求	33
2.3.3 影响样本量的其他因素	36
2.3.4 总体参数的预估计	36
2.4 子总体的估计	37
2.4.1 相关定义及符号	37
2.4.2 子总体均值的估计	38
2.4.3 子总体总量的估计	39
习题二	41
第三章 分层抽样	43
3.1 引言	43
3.1.1 定义	43
3.1.2 作用	43
3.1.3 应用场合	44
3.1.4 符号说明	45
3.1.5 实施方法	45
3.2 估计量及性质	46
3.2.1 总体均值估计	46
3.2.2 总体总量估计	48
3.2.3 总体比例估计	49
3.2.4 总体特征单元数估计	50
3.3 样本量在各层的分配	52
3.3.1 样本分配对精度与费用的影响	52

3.3.2 比例分配	54
3.3.3 最优分配	54
3.3.4 奈曼分配	55
3.3.5 某些层大于 100% 抽样时的情况	56
3.4 总样本量的确定	57
3.4.1 一般公式	58
3.4.2 不同应用场合	58
3.5 分层抽样的效率分析	61
3.5.1 分层抽样与简单随机抽样的比较	61
3.5.2 分层抽样的设计效应	63
3.5.3 比例分配与奈曼分配的效率比较	65
3.5.4 偏离最优分配时对方差的影响	66
3.6 划分层的问题	67
3.6.1 层的界限	67
3.6.2 层数的确定	69
3.7 其他分层技术	70
3.7.1 目录抽样	70
3.7.2 事后分层	71
3.7.3 关于多指标分层抽样	73
习题三	74
第四章 比率估计和回归估计	79
4.1 引言	79
4.1.1 定义	79
4.1.2 应用条件	80
4.1.3 符号	80
4.2 简单随机抽样下的比率估计	81
4.2.1 比率估计量及其性质	81
4.2.2 样本量的确定	86
4.2.3 比率估计量的偏倚及纠偏方法	88
4.3 分层抽样下的比率估计	91
4.3.1 分别比率估计	91
4.3.2 联合比率估计	91
4.3.3 分别比率估计与联合比率估计的比较	92
4.4 简单随机抽样下的回归估计	94

4.4.1 回归估计量及其性质	95
4.4.2 差值估计——回归估计量的一种特殊情形	98
4.5 分层抽样下的回归估计	99
4.5.1 分别回归估计	99
4.5.2 联合回归估计	100
4.5.3 分别回归估计与联合回归估计的比较	101
4.6 比率估计、回归估计的效率问题	103
4.6.1 比率估计与简单估计的比较	104
4.6.2 回归估计与简单估计的比较	104
4.6.3 比率估计与回归估计的比较	105
本章附录	105
习题四	108
第五章 不等概抽样	114
5.1 引言	114
5.1.1 定义	114
5.1.2 种类	115
5.1.3 区域抽样	116
5.2 放回不等概抽样	116
5.2.1 PPS 抽样的性质	117
5.2.2 PPS 抽样的实施	117
5.2.3 汉森-赫维茨估计量	119
5.3 不放回不等概抽样	121
5.3.1 π PS 抽样的性质	121
5.3.2 霍维茨-汤普森估计量	123
5.3.3 几种严格的 π PS 抽样	125
5.3.4 几种非严格的 π PS 抽样	129
习题五	130
第六章 整群抽样	133
6.1 引言	133
6.1.1 定义	133
6.1.2 应用特点	134
6.1.3 群的划分	135
6.1.4 符号	136
6.2 群规模相等的情形	137

6.2.1 估计量及性质	138
6.2.2 整群抽样效率分析	141
6.2.3 最佳群规模的确定	144
6.3 群规模不相等的情形	146
6.3.1 等概抽样:加权估计	146
6.3.2 等概抽样:比率估计	147
6.3.3 不等概抽样	153
6.3.4 群规模不等时三种估计方法的比较	155
本章附录	157
习题六	159
第七章 系统抽样	165
7.1 引言	165
7.1.1 定义	165
7.1.2 应用特点	166
7.1.3 不同系统抽样方法	167
7.2 等概率系统抽样	169
7.2.1 系统抽样与整群抽样、分层抽样的关系	169
7.2.2 估计量及性质	170
7.3 不同总体排列方式的方差估计	175
7.3.1 按无关标志排列的总体	176
7.3.2 按有关标志排列的总体	176
7.3.3 有周期波动的总体	181
7.4 其他类型的系统抽样	182
7.4.1 不等概率系统抽样	183
7.4.2 分层系统抽样	186
7.4.3 二维系统抽样	188
本章附录	189
习题七	191
第八章 多阶段抽样	195
8.1 引言	195
8.1.1 定义	195
8.1.2 应用特点	196
8.1.3 推断原理	197
8.1.4 符号	197

8.2 二阶抽样初级单元规模相等的情形	198
8.2.1 总体均值与总体总量的估计	198
8.2.2 总体比例的估计	201
8.2.3 样本的最优分配	202
8.3 二阶抽样初级单元规模不等的情形	205
8.3.1 按简单随机抽样抽取初级单元	205
8.3.2 按不等概率抽样抽取初级单元	208
8.3.3 总体比例的估计	211
8.4 二阶抽样的扩展	214
8.4.1 分层二阶抽样	214
8.4.2 三阶及多阶抽样	215
8.5 二阶抽样效率分析	218
8.5.1 与简单随机抽样比较	218
8.5.2 与分层抽样比较	219
8.5.3 与整群抽样比较	219
本章附录	220
习题八	223
第九章 二重抽样	228
9.1 引言	228
9.1.1 定义	228
9.1.2 二重抽样与两阶段抽样	229
9.1.3 二重抽样的作用	229
9.2 为分层的二重抽样	230
9.2.1 抽样方法	231
9.2.2 估计量及其性质	231
9.3 为比率估计的二重抽样	234
9.3.1 抽样方法	234
9.3.2 估计量及其性质	234
9.4 为回归估计的二重抽样	237
9.4.1 抽样方法	237
9.4.2 估计量及其性质	238
9.5 样本量的最优分配	239
9.5.1 为分层的二重抽样最优分配	240
9.5.2 为比率估计的二重抽样最优分配	241

9.5.3 为回归估计的二重抽样最优分配	241
本章附录	242
习题九	243
第十章 其他抽样方法	247
10.1 总体单元数的估计	247
10.1.1 捕获再捕获法	247
10.1.2 逆抽样法	249
10.2 敏感性问题的抽样方法	250
10.2.1 属性特征敏感性问题的抽样设计	250
10.2.2 数量特征敏感性问题的抽样设计	260
10.3 稀有事件抽样	262
10.3.1 按不等比例分配的分层抽样	262
10.3.2 二重抽样处理	263
10.4 自加权抽样设计	264
10.4.1 概述	264
10.4.2 主要抽样方法下的自加权设计	265
10.5 样本轮换设计	266
10.5.1 概述	266
10.5.2 样本轮换下的总体均值的估计	266
本章附录	269
习题十	271
第十一章 复杂样本的方差估计	274
11.1 引言	274
11.1.1 什么是复杂样本	274
11.1.2 基本方法概述	275
11.2 直接推导法	275
11.2.1 精确公式	276
11.2.2 近似公式	276
11.3 随机组方法	277
11.3.1 独立随机组情形	277
11.3.2 非独立随机组情形	280
11.3.3 随机组的划分和组数的确定	280
11.4 平衡半样本方法	282