

应用统计学系列教材 Texts in Applied Statistics

抽样技术及其应用

Sampling Techniques and Practices

杜子芳 编著

Du Zifang



清华大学出版社



Springer

应用统计学系列教材 Texts in Applied Statistics

抽样技术及其应用

Sampling Techniques and Practices

杜子芳 编著

Du Zifang



清华大学出版社

北京



Springer

内 容 简 介

全书共分9章,第1章介绍了学习抽样理论需要的预备知识;第2章介绍了抽样理论的若干基本概念;第3章到第7章分别对常规的几种抽样方法——简单随机抽样、分层随机抽样、二阶及多阶抽样、整群抽样和系统抽样,围绕其基本概念、特点、适用场合、具体抽样步骤和相应的简单估计、比率估计和回归估计三种估计量的性质等几个方面,进行了深入详尽的介绍,并且特别对每一种抽样方法的三种估计量的性质融会在一起进行讨论、比较;第8章介绍了实际中极为常见的非概率抽样的各种方式,重点阐释了这些方式与概率抽样的几种抽样方法间的对应关系;第9章是关于与抽样有关但又不属于抽样的一些问题的讨论,帮助实际工作者开阔眼界,避免失误.出于同样的目的,在本书的重要章节还至少给出了一个相应的案例.

本书可作为统计专业的本科生、非统计专业的研究生教材,也可供相关专业的教师及实际从事抽样工作的人员参考.

版权所有,翻印必究.举报电话:010-62782989 13501256678 13801310933

图书在版编目(CIP)数据

抽样技术及其应用/杜子芳编著. —北京:清华大学出版社,2005.8

(应用统计学系列教材)

ISBN 7-302-11039-5

I. 抽… II. 杜… III. 抽样调查—高等学校—教材 IV. C811

中国版本图书馆 CIP 数据核字(2005)第 050112 号

出 版 者: 清华大学出版社

<http://www.tup.com.cn>

杜 总 机: 010-62770175

地 址: 北京清华大学学研大厦

邮 编: 100084

客户服务: 010-62776969

责任编辑: 王海燕

封面设计: 常雪影

印 装 者: 清华大学印刷厂

发 行 者: 新华书店总店北京发行所

开 本: 170×230 印张: 32.75 字数: 586 千字

版 次: 2005 年 8 月第 1 版 2005 年 8 月第 1 次印刷

书 号: ISBN 7-302-11039-5/O·463

印 数: 1~3000

定 价: 46.00 元

应用统计学系列教材

Texts in Applied Statistics

编审委员会

主 任：吴喜之

委 员：（按姓氏拼音字母排序）

杜子芳 冯士雍 耿 直 何书元 贾俊平

金勇进 易丹辉 袁 卫 张 波 赵彦云

序

随着社会经济的飞速发展,统计学课程设置的不断调整,统计学教材已经有了很大的变化。为了适应这些变化,我们从2000年开始编写面向21世纪统计学系列教材,经过近4年的实践,该系列教材取得了较好的效果,基本实现了预定的目标。然而目前学科的发展和社会的进步速度相当快,其中的一些教材已经需要进一步修订,也有部分内容成熟、适合教学需要的教材没有列入编写计划。

为满足应用统计科学和我国高等教育迅速发展的需求,清华大学出版社和施普林格出版社(Springer-Verlag)合作,倡议出版这一套“应用统计学系列教材”,作为对现有统计学教材的全面补充和修订。这套教材具有以下特点:

1. 此套丛书属于开放式的,一旦有好的选题,即可列入出版计划。

2. 在教材选择上,拓宽了范围。有些教材主要面向经济类统计学专业,包括金融统计、风险管理与精算方面的教材。部分教材面向人文社科专业,而另外一些教材则面向自然科学领域,包括生物统计、医学统计、公共卫生统计等。

3. 本套教材的编写者都是活跃在教学、科研第一线的教师,他们能够积极地、广泛地吸收国内外最新的优秀成果。能够在教学中反复对教材进行补充修订和完善。

4. 强调与计算机应用的结合,在教材编写中,注重计算机软件的应用,特别是可编程软件的应用。对于那些仅限于应用方法的教材,充分考虑读者的需求,尽量介绍简单易学的“傻瓜”

软件。

5. 本套教材包括部分优秀国外教材译著,对于目前急需,而国内尚属空白的教材,选择部分国外具有广泛影响的教材,进行翻译出版。

我们希望这套系列教材的出版能够对我国应用统计科学的教育和我国统计事业的健康发展起到积极作用。感谢参与教材编写的中国人民大学统计学院和兄弟院校的教师以及进行审阅的同行专家。让我们大家共同努力,创造我国应用统计学科新的辉煌。

易丹辉

2004年1月

前言

由于决策科学化程度的提高,调查这种系统地搜集信息的活动已经成为各种研究的基础,“没有调查就没有发言权”的法
则现今在任何领域、任何层面均被普遍认同和遵守。而在任何较大规模的调查中,抽样技术的选择与应用皆是不可避免的。不幸的是,采用抽样技术的做法的普及反而凸显了相关知识普及的不足,除了某些官方机构和学术机构组织的抽样调查比较规范以外,即使那些在大众媒体上亮相频繁、抢尽风头的专业调查公司所公布的抽样方案,在科学意义上也很少称得上规范合格。

笔者在中国人民大学讲授抽样理论十几年,期间也断断续续主持或参与了不少政府或企业的调查设计工作。长期的教学与实践使我感受到,虽然前辈同行迄今做出了巨大努力:国际上一些抽样方面的名著悉数得到翻译,国内也出版了一些专著和教科书,但所有这些相对于人们对抽样技术理论和方法的了解的迫切需要程度来讲,仍相去甚远。坦率地说,国内现存的有关抽样技术的书籍,要么过于专业,内容艰涩难懂;要么过于业余,内容零碎不成系统,其中少数出版物更满是将“总体”混同“人口”之类的低级错误,使人愕然。

鉴于此,本书主要面向那些在大学里讲授抽样技术课程的教师、统计专业的本科生、非统计专业的研究生和相关部门的中高级人士。内容力求详尽而不失精练,结构力求系统而不失紧凑。理论证明步骤明晰,浅显易懂,对数理统计知识要求不多,以便教师备课和读者自学,例题、习题选题广泛,兼顾明理与实用的双重需要。

全书共分9章,第1章介绍了学习抽样理论需要的预备知识,包括排列组合、中心极限定理、参数的区间估计和调查概论;第2章介绍了抽样理论的若干基本概念;第3章到第7章分别对常规的几种抽样方法——简单随机抽样、分层随机抽样、二阶及多阶抽样、整群抽样和系统抽样,围绕其基本概念、特点、适用场合、具体抽样步骤和相应的简单估计、比率估计和回归估计三种估计量的性质等几个方面,进行了深入详尽的介绍。特别地,对每一抽样方法的三种估计量的性质融会在一起进行讨论、比较,构成了本书的特色之一。本书的另一个特色是对不等概率抽样的处理,对其不做独立的讨论,但强调在次级或基本单元的层次上仍属于等概率抽样的特点,这样有助于读者更好地理解所谓不等概率抽样的实质,避免无谓的混淆。第8章介绍了实际中极为常见的非概率抽样的各种方式,重点阐释了这些方式与概率抽样的几种抽样方法间的对应关系。这个问题是极其重要的,可惜国内外绝大多数抽样文献都未给予必要的重视。最后一章是关于与抽样有关但不属于抽样的一些问题的讨论,帮助实际工作者开阔眼界,避免失误。出于同样的目的,在本书的重要章节还至少给出了一个相应的案例。

在本书写作过程中,我的一些研究生朱华鹏、姜莉莉、江智杰、徐瑶、徐晓菊、牛奇和王欣等同学给予了许多帮助,他们不仅参与了部分章节的撰写、习题案例的寻找,还帮助调整了全部的格式。特别是朱华鹏和姜莉莉,对本书的结构提出了很多宝贵建议,又反复阅读和校对了全部书稿,保证了稿件的按时完成。在此谨表示诚挚的谢意。最后,作者还要特别感谢我的同事张波教授,他对本书的顺利完成助益良多,付出了很多心血与努力。

由于作者水平有限,写作亦较仓促,虽然诚惶诚恐,殚精竭虑,但不足之处仍然在所难免,甚至可能不少,敬请有关专家、同行和广大读者赐教、斧正。

杜子芳

2005年2月

目 录

第 1 章 预备知识	1
1.1 排列组合	1
1.1.1 基本原理	1
1.1.2 排列	2
1.1.3 组合	2
1.2 概率统计中的一些基本原理	3
1.2.1 大数定律	3
1.2.2 中心极限定理	4
1.2.3 参数估计	6
1.3 调查概论	11
1.3.1 调查与测量	11
1.3.2 真值、测量值与误差	16
1.3.3 信度、效度与精度	22
思考与练习	26
参考文献	27
第 2 章 基本概念	28
2.1 抽样调查与非抽样调查	28
2.2 总体与样本	34
2.2.1 总体	34
2.2.2 抽样框与抽样单元	35
2.2.3 抽样与样本	36
2.3 总体特征与估计量	37
2.3.1 总体特征	37
2.3.2 估计量和估计方法	39

2.3.3	抽样分布	40
2.4	误差与精度	42
2.4.1	均方误差与偏倚	43
2.4.2	置信区间与误差限	44
2.4.3	费用与效率	45
2.5	几种基本的抽样方法	46
2.5.1	简单随机抽样	46
2.5.2	分层抽样	46
2.5.3	整群抽样	48
2.5.4	系统抽样	48
2.5.5	多阶段抽样	49
* 2.6	抽样调查的实施步骤	50
	思考与练习	53
	参考文献	54
第3章	简单随机抽样	55
3.1	定义与符号	56
3.2	简单估计量及其性质	64
3.2.1	简单估计的性质	65
3.2.2	简单估计量 $\bar{y}$ 的方差与协方差	69
3.2.3	方差与协方差的估计	76
3.3	比率估计量及其性质	81
3.3.1	比率估计量的性质	82
3.3.2	比率估计量的方差估计	86
3.3.3	比率估计的其他问题	89
3.4	回归估计量及其性质	93
3.4.1	回归估计的性质	94
3.4.2	多变量回归估计	98
3.4.3	各种估计量的精度比较	99
3.5	简单随机抽样的实施	101
3.5.1	样本容量的确定原理	101
3.5.2	样本量的确定步骤	104
3.5.3	简单随机抽样的实施	105

附件 A 中华人民共和国国家标准	
利用随机数骰子进行随机抽样的方法	110
附件 B 中华人民共和国国家标准	
利用电子随机数抽样器进行随机抽样的方法	114
思考与练习	120
参考文献	123
第 4 章 分层随机抽样	124
4.1 定义与符号	124
4.1.1 定义	124
4.1.2 符号	126
4.2 简单估计量及其性质	126
4.2.1 总体均值的简单估计量及其性质	126
4.2.2 总体总量的简单估计量及其性质	129
4.2.3 总体比例的简单估计量及其性质	132
4.3 比率估计量及其性质	133
4.3.1 分别比估计	134
4.3.2 联合比估计	135
4.3.3 分别比估计与联合比估计的比较	137
4.3.4 分层抽样中采取比率估计时的最优分配	138
4.4 回归估计量及其性质	142
4.4.1 分别回归估计	142
4.4.2 联合回归估计	145
4.4.3 分别回归估计与联合回归估计的比较	148
4.4.4 比率估计与回归估计小结	154
4.5 各层样本量的分配	155
4.5.1 比例分配	156
4.5.2 最优分配	157
4.5.3 奈曼最优分配	160
4.5.4 某些层需要进行大于 100% 抽样的修正	162
4.5.5 偏离最优分配时对精度的影响	164
4.5.6 多变量情况下样本量在各层的分配	166
4.6 总样本量的确定	170

4.6.1	估计总体均值时样本量的确定	170
4.6.2	估计总体总量时样本量的确定	175
4.6.3	估计总体比例时样本量的确定	178
4.6.4	总费用给定时总样本量的确定	180
4.7	层的构成和分层界限的确定	181
4.7.1	分层标志的选择	181
4.7.2	分层界限的确定	181
4.7.3	确定分层界限的其他方法	187
4.8	层数的确定	187
4.8.1	根据目标变量 Y 值确定层数	188
4.8.2	根据其他变量 X 值确定层数	189
4.8.3	考虑费用因素时层数的确定	190
4.9	分层随机抽样的精度研究	191
4.9.1	分层随机抽样与简单随机抽样精度比较	191
4.9.2	最优分配在精度上的改进	193
4.9.3	分层随机抽样精度反比简单随机抽样差的情况	194
4.9.4	从分层样本来估计分层随机抽样的效果	196
4.9.5	层权误差对估计量的影响	202
4.9.6	每层只抽取一个单元时的方差估计	207
4.10	分层抽样的其他方面	210
4.10.1	多重分层	210
4.10.2	事后分层	215
4.10.3	子总体的估计	220
4.10.4	定额抽样	230
4.11	案例	231
附件 C		234
附件 D	1993 年人口变动情况抽样调查样本设计参考资料	235
附件 E	省级人口变动情况抽样误差计算公式和总体人口出生率的 估计方法	238
思考与练习		241
参考文献		246

第 5 章 多阶(段)抽样	247
5.1 概述	248
5.1.1 定义及特点	248
5.1.2 二阶(段)抽样的预备知识	249
5.2 初级单元大小相等时的二阶(段)抽样	250
5.2.1 符号	250
5.2.2 总体均值 $\bar{Y}$ 的估计量及其性质	251
5.2.3 总体比例的估计	256
5.2.4 最优样本量 m 与 n 的确定	259
5.2.5 分层二阶(段)抽样	262
5.3 初级单元大小不等情形($n=1$)	265
5.3.1 一般说明和符号	265
5.3.2 等概率抽取初级单元	267
5.3.3 不等概率抽取初级单元	269
5.3.4 $n=1$ 时五种方法的总结	272
5.4 初级单元大小不等时的二阶(段)抽样($n>1$)	275
5.4.1 采用放回的抽样方式——按不等概率抽取 初级单元	276
5.4.2 采用不放回的抽样方式	281
5.4.3 估计比例的二阶(段)抽样	289
5.5 二阶(段)抽样的效率	292
5.5.1 二阶(段)抽样与简单随机抽样比较	292
5.5.2 二阶(段)抽样与分层抽样比较	293
5.5.3 二阶(段)抽样与整群抽样比较	293
5.5.4 小结	294
5.6 三阶(段)及多阶抽样	294
5.6.1 各级单元大小相等的三阶(段)抽样	295
5.6.2 各级单元大小不等的三阶(段)抽样	297
5.7 案例	300
案例 1 “网民知多少?”——中国互联网络信息中心全国调查 抽样方案设计	300
案例 2 第二次国家卫生服务调查设计方案	306

附件 F 国家卫生服务总调查样本地区和样本个体的抽取方法	312
思考与练习	323
参考文献	328
第 6 章 整群抽样	329
6.1 概述	329
6.1.1 整群抽样的基本概念	329
6.1.2 整群抽样的特点及适用场合	331
6.2 群规模大小相等的情形	333
6.2.1 符号说明	333
6.2.2 估计量及其性质	334
6.2.3 总体方差 S^2 的估计	336
6.2.4 群内相关系数与设计效应	341
6.2.5 整群抽样的效率分析	345
6.3 群规模大小不等的情形	349
6.3.1 符号说明	349
6.3.2 对群进行简单随机抽样的简单估计	350
6.3.3 对群进行简单随机抽样的比估计	352
6.3.4 对群进行不等概率抽样	354
6.4 对比例估计的整群抽样	356
6.4.1 问题的提出	356
6.4.2 群规模相等的情形	356
6.4.3 群规模不相等的情形	358
6.5 案例	366
思考与练习	373
参考文献	377
第 7 章 系统抽样	378
7.1 定义与实施方法	379
7.2 等概率情形的估计量及其性质	382
7.2.1 符号说明	382
7.2.2 估计量的性质	384
7.3 方差估计及其改进	394

7.3.1 方差的近似估计	394
7.3.2 线性排列情形抽样与估计的改进	396
7.4 不等概率系统抽样	403
7.4.1 概述及实施方法	403
7.4.2 估计量的方差估计	404
7.5 案例	407
思考与练习	412
参考文献	415
第 8 章 非概率抽样	416
8.1 非概率抽样概述	416
8.1.1 概念及适用场合	416
8.1.2 具体的抽样方法	418
8.2 非概率抽样的理论基础	423
8.2.1 非概率抽样与模型抽样	423
8.2.2 非概率抽样的效度与信度	424
8.2.3 非概率抽样与概率抽样的对应关系	430
8.3 非概率抽样中的估计	432
8.3.1 非概率样本的估计问题	432
8.3.2 捕获-再捕获中的估计	433
8.4 样本容量与抽样实施	436
8.4.1 样本容量确定的出发点	436
8.4.2 确定样本容量的几种思路	438
8.4.3 配额抽样设计样本量的确定	441
8.4.4 非概率抽样方法的实施	445
8.5 案例	448
案例 1 关于某市宾馆、饭店、娱乐场所扰民噪声的抽样 调查	448
案例 2 1997 年我国进行的城市环境状况和居民环境 意识的配额抽样调查	450
思考与练习	453
参考文献	454

第 9 章 其他专题	455
9.1 敏感性问题的处理	455
9.1.1 沃纳随机化回答模型	455
9.1.2 西蒙斯随机化回答模型	457
9.1.3 使用随机化回答方法应注意的问题	459
9.1.4 随机截尾模型	459
9.2 无回答误差处理	461
9.2.1 无回答的概念及类型	461
9.2.2 无回答的影响	462
9.2.3 降低无回答率的方法	463
9.2.4 对无回答的调整	464
9.3 捕获-再捕获抽样	469
9.3.1 直接抽样法	469
9.3.2 逆抽样方法	472
9.4 样本轮换	473
9.4.1 样本轮换的最优比例	474
9.4.2 样本轮换应遵循的原则	476
9.4.3 案例说明	476
思考与练习	477
参考文献	478
练习题参考答案	479



第 1 章

预 备 知 识

本章将简要地介绍排列组合、概率统计、调查等方面的有关知识. 这些知识属于抽样技术的基础知识或预备知识, 在以后各章的叙述中会用到, 但又不属于本书中任何其他章节的介绍范围. 补充这些知识供读者在学习时参考, 能够使读者更加顺利地学好以后各章节的内容. 建议读者特别是那些没有系统学习过调查和数理统计的读者, 对本章所介绍的内容能够认真阅读, 因为这些内容不仅是进一步学习抽样技术的准备性知识, 同时这些知识本身也是十分重要和有用的.

1.1 排列组合

1.1.1 基本原理

首先叙述两条基本原理, 这两条原理在排列组合分析公式的推导中起着重要作用.

(1) 加法原理: 假如完成一件事有 m 种方式, 第一种方式有 n_1 种方法, 第二种方式有 n_2 种方法, $\dots$, 第 m 种方式有 n_m 种方法, 而无论通过哪种方式方法都可以完成这件事, 则完成这件事总共有 $n_1 + n_2 + \dots + n_m$ 种方法.

(2) 乘法原理: 假如完成一件事有 m 个步骤, 第一个步骤有 n_1 种方法, 第二个步骤有 n_2 种方法, $\dots$, 第 m 个步骤有 n_m 种方法, 而每一步骤都不可缺少也不可重复, 则完成这件事共有 $n_1 n_2 \dots n_m$ 种方法.