

经济和社会事务部
统计司

方法研究

F辑第98号

住户调查 样本设计： 实用准则



联合国
纽约，2010年

经济和社会事务部

联合国秘书处经济和社会事务部是经济、社会和环境领域的全球政策与国家行动之间一个至关重要的结合部。该部从事三个主要相互关联领域的工作：（一）汇编、提供和分析范围广泛的经济、社会和环境数据与资料，联合国会员国依靠它们来审查共同的问题和评估各种政策选择；（二）便利会员国在许多政府间机构中谈判各种联合行动方针以应对正面临或新出现的全球挑战；和（三）对有关政府采用何种方式方法将在联合国各次会议和首脑会议上形成的政策纲要转变为国家一级的方案提出建议并通过技术援助帮助建设国家能力。

说 明

本出版物所采用的名称及材料编写方式，并不意味着联合国秘书处对任何国家、领土、城市或地区或其当局的法律地位，或对其边界或界限的划分表示任何意见。

本报告文内使用的“国家”一词适当时也指领土或地区。

采用“发达”和“发展中”国家或地区和“较发达”、“较不发达”及“最不发达”区域是为了方便统计工作，不一定表示对某一国家或地区达到的发展阶段做出判断。

联合国文件都用英文大写字母附加数字编号。凡提到此种编号即指联合国某一文件。

ST/ESA/STAT/SER.F/98

ISBN 978-92-1-730224-4

联合国出版物
出售品编号：C.06.XVII.13

版权©联合国，2008年
版权所有

序 言

《住户调查样本设计：实用准则》的主要目的是当作一本包罗抽样调查主要设计问题的手册，以便于实际参加各国抽样调查工作和活动的国家统计人员、研究工作者和分析家们参考。本手册以统计学理论为基础，使用了在方法论上扎实可靠的技巧，在样本选择过程的每一阶段运用概率抽样法。设计完善、稳妥实施的住户调查可以用较低的成本快速产生高质量、高精度的必要信息。

本出版物的内容，亦可在提供应用统计学尤其是调查方法论课程的统计工作培训机构有选择地用作导论课程的培训指南。

此外，本出版物也是为了补充有关抽样调查方法论的联合国其他出版物而编写的，比如最近出版的《发展中国家和转型期国家住户调查》*以及在《国家住户调查能力方案》(NHSCP)项下推出的系列出版物。

具体来讲，本手册的目的在于：

- (a) 利用一份出版物提供主要在国家一级住户调查中所需样本设计的基本概念和方法论上的可靠程序，重点放在住户样本设计的应用方面；
- (b) 在设计和实施有效住户样本调查方面充当调查工作者的实用指南；
- (c) 说明样本设计、数据收集、估算、处理和分析之间的相互关系；
- (d) 强调住户抽样调查中控制和减少非抽样误差的重要性。

对于本手册的使用者来说，虽然了解抽样背景很重要，但凡是懂得统计和数学概念一般知识的其他人想必也可以在很少需要或根本无需协助的情况下使用本手册并应用其中的内容。这是因为本手册的主要目的之一就是以注重实际、鼓励亲自实践的方式提供材料，而不偏重于抽样理论方面。不过，必要时也提供一些理论基础。只要基本掌握代数知识即可很容易地跟上讲解并应用相关的技巧。为此提供了许多范例来说明概念、方法和技巧。

有不少专家参与了本手册的编写工作。取样顾问Anthony Turner先生起草了第三、第四和第五章，并审查了最终合并文件；国际公务员制度委员会生活费用司副司长Ibrahim Yansaneh先生起草了第六和第七章；南部非洲发展共同体秘书处的统计专家Maphion Jambwa先生起草了第九章。

* 《方法研究》第96号(联合国出版物，出售品编号：E.05.XVII.6)。

担任本项目总编辑和技术协调人的联合国统计司Jeremiah Banda先生是第一、第二和第八章及附件一的作者。Clare Menozzi女士协助编辑了各章初稿；而Bizugenet Kassa女士提供了宝贵的秘书协助，同时Pansy Benjamin女士帮助协调了版式。

由统计司召集于2003年12月3日至5日举行的专家组会议审议了各章初稿。与会者名单列于附件二。此外，本手册经过墨西哥国家统计学、地理学和信息学研究所(INEGI)的Alfredo Bustos博士、Ana María Landeros女士和Eduardo Ríos先生等同行专家的审议，他们提出了非常宝贵的意见。

联合国经济和社会事务部

统计司

司长

Paul Cheung

目 录

	页次
序言	iii
第一章	
社会及人口统计资料来源	
1.1. 导言	1
1.2. 资料来源	1
1.2.1. 住户调查	1
1.2.2. 人口和住房普查	3
1.2.3. 行政案卷	4
1.2.4. 三种数据源的互补性	5
1.2.5. 结论意见	6
参考资料和补充读物	6
第二章	
调查计划与执行	
2.1. 调查计划	9
2.1.1. 调查目标	9
2.1.2. 调查全域	11
2.1.3. 拟收集的信息	11
2.1.4. 调查预算	11
2.2. 调查的执行	12
2.2.1. 收集数据的方法	12
2.2.2. 问卷的设计	17
2.2.3. 制表和分析计划	19
2.2.4. 现场工作实施	19
参考资料和补充读物	22
第三章	
抽样策略	
3.1. 导言	23
3.1.1. 概览	23
3.1.2. 抽样词汇及相关术语	24
3.1.3. 符号	24

	页次
3.2. 住户调查的概率抽样法与其他抽样法的比较	26
3.2.1. 概率抽样	27
3.2.2. 非概率抽样法	29
3.3. 确定住户调查的样本规模	31
3.3.1. 调查估计数的量级	31
3.3.2. 目标人口	32
3.3.3. 精度和统计置信度	33
3.3.4. 分析群体: 域	33
3.3.5. 聚类效应	35
3.3.6. 为预计的不答复率调整样本规模	36
3.3.7. 主样本的样本规模	36
3.3.8. 估计变化或水平	37
3.3.9. 调查预算	37
3.3.10. 样本规模计算	38
3.4. 分层	40
3.4.1. 分层与样本分配	40
3.4.2. 分层的规则	41
3.4.3. 隐含分层法	42
3.5. 整群抽样	42
3.5.1. 整群抽样的特性	44
3.5.2. 类集的设计效果	44
3.5.3. 类集规模	45
3.5.4. 设计效果 (<i>deff</i>) 的计算	46
3.5.5. 类集的数量	46
3.6. 分阶段抽样	46
3.6.1. 分阶段抽样的好处	47
3.6.2. 虚拟阶段的使用	47
3.6.3. 两段设计	50
3.7. 与规模成正比的概率抽样和与估计规模成正比的概率抽样	50
3.7.1. 与规模成正比的概率抽样	50
3.7.2. 与估计规模成正比的概率抽样	52
3.8. 抽样选项	54
3.8.1. 等概率抽样, 与规模成正比的概率抽样, 固定规模抽样和固定比率抽样	55
3.8.2. 人口和健康调查 (DHS)	58
3.8.3. 修改的类集设计: 多指标类集调查 (MICS)	59
3.9. 特别话题: 两阶段样本和趋势抽样	60
3.9.1. 两阶段抽样	60
3.9.2. 旨在估计变化或趋势的抽样	61
3.10. 当实施工作出毛病的时候	64
3.10.1. 目标人口的定义和覆盖面	64
3.10.2. 就调查预算而言样本规模过大	65
3.10.3. 类集规模大于或小于预期	65
3.10.4. 处理不答复的情况	65

	页次
3.11. 准则概要	66
参考资料和补充读物	67
第四章	
抽样标架和主样本	
4.1. 住户调查中的抽样标架	69
4.1.1. 抽样标架的定义	69
4.1.2. 抽样标架的特征	70
4.1.3. 地域标架	72
4.1.4. 清单标架	72
4.1.5. 多元标架	73
4.1.6. 两段设计中的典型标架	75
4.1.7. 主抽样标架	75
4.1.8. 标架的普遍问题和推荐的纠正办法	75
4.2. 主抽样标架	78
4.2.1. 主样本的定义和使用	78
4.2.2. 主抽样标架初级抽样单位的理想特征	79
4.2.3. 用主样本支持调查	79
4.2.4. 域间(政区间, 等等)的分配	81
4.2.5. 主样本的维持与更新	81
4.2.6. 主样本中初级抽样单位的轮换	82
4.3. 准则概要	88
参考资料和补充读物	88
第五章	
样本设计的文件编制与评价工作	
5.1. 导言	91
5.2. 样本文件编制与评价工作的必要性和类别	91
5.3. 设计变量的标号	92
5.4. 选择概率	93
5.5. 各个选样阶段的反应率和覆盖率	94
5.6. 加权: 基权数、不答复和其他调整	95
5.7. 关于抽样和调查实施成本的信息	95
5.8. 评估: 调查数据的缺陷	96
5.9. 准则概要	97
参考资料和补充读物	98
第六章	
样本权数的构建与使用	
6.1. 导言	99
6.2. 抽样权数的必要性	99

	页次
6.2.1. 概览.....	99
6.3. 抽样权数的开发.....	100
6.3.1. 未知被选资格的样本权数调整.....	100
6.3.2. 重复样本的权数调整.....	101
6.4. 不均等选择概率的加权.....	102
6.4.1. 构建权数方面的案例研究: 越南全国健康调查, 2001年.....	105
6.4.2. 自我加权的样本.....	106
6.5. 针对不答复的样本权数调整.....	107
6.5.1. 在住户调查中减少不答复偏差.....	107
6.5.2. 不答复的补偿.....	107
6.5.3. 针对不答复问题的样本权数调整.....	108
6.6. 针对未覆盖问题的样本权数调整.....	109
6.6.1. 住户调查中出现未覆盖问题的根源.....	110
6.6.2. 住户调查中对未覆盖的补偿.....	111
6.7. 加权导致抽样方差的增加.....	112
6.8. 权数的修整.....	113
6.9. 结论意见.....	114
参考资料和补充读物.....	115

第七章

调查数据抽样误差的估算

7.1. 导言.....	117
7.1.1. 复杂调查数据的抽样误差估算.....	117
7.1.2. 概览.....	118
7.2. 简单随机抽样项下的抽样方差.....	118
7.3. 抽样误差的其他量度.....	123
7.3.1. 标准误差.....	123
7.3.2. 变异系数.....	123
7.3.3. 设计效果.....	123
7.4. 计算其他标准设计的抽样方差.....	124
7.4.1. 分层抽样.....	124
7.5. 住户调查样本设计和数据的共同特征.....	127
7.5.1. 住户调查设计偏离简单随机抽样的差幅.....	127
7.5.2. 供分析用数据文档的准备.....	127
7.5.3. 调查估计数的类别.....	128
7.6. 抽样误差信息的表述准则.....	129
7.6.1. 确定报告什么内容.....	129
7.6.2. 如何报告抽样误差信息.....	129
7.6.3. 报告标准误差方面的经验法则.....	130
7.7. 住户调查中估计方差的方法.....	130
7.7.1. 精确法.....	130
7.7.2. 最终类集法.....	131

	页次
7.7.3. 线性化近似法.....	134
7.7.4. 重复法.....	136
7.7.5. 一些重复技巧.....	137
7.8. 在使用标准统计软件包分析住户调查数据方面易犯的错误.....	142
7.9. 抽样误差估算用的计算机软件.....	142
7.10. 各种软件包的一般比较.....	146
7.11. 结论意见.....	146
参考资料和补充读物.....	147

第八章

住户调查的非抽样误差

8.1. 导言.....	149
8.2. 偏差和变动误差.....	150
8.2.1. 变动分量.....	152
8.2.2. 系统误差(偏差).....	152
8.2.3. 抽样偏差.....	152
8.2.4. 偏差与变动误差的进一步比较.....	152
8.3. 非抽样误差来源.....	153
8.4. 非抽样误差的各种分量.....	154
8.4.1. 规格误差.....	154
8.4.2. 覆盖面或标架误差.....	154
8.4.3. 不答复.....	156
8.4.4. 计量误差.....	157
8.4.5. 数据处理误差.....	158
8.4.6. 估算误差.....	158
8.5. 非抽样误差的评估.....	158
8.5.1. 检查一致性.....	158
8.5.2. 样本检查/验证.....	158
8.5.3. 调查后的检查或再访谈检查.....	159
8.5.4. 质量控制方法.....	159
8.5.5. 关于记忆误差的研究.....	160
8.5.6. 交叉再抽样.....	160
8.6. 结论意见.....	160
参考文献和补充读物.....	161

第九章

住户调查的数据处理

9.1. 导言.....	163
9.2. 住户调查流程.....	163
9.3. 调查策划和数据处理系统.....	164
9.3.1. 调查目的和内容.....	164

	页次
9.3.2. 调查程序和工具	165
9.3.3. 住户调查的数据处理系统设计	167
9.4. 调查工作和数据处理	170
9.4.1. 标架创立和样本设计	170
9.4.2. 数据收集和数据管理	172
9.4.3. 数据准备	172
附录	186
微软办公软件	186
直观基本系统(Visual Basic)	187
抽样方差计算软件(CENVAR)	187
PC CARP	187
普查与调查数据处理系统(CSPro)	187
抽样误差实用数据计算和列举(CLUSTERS)	187
集成调查分析系统(ISSA)	188
统计分析系统(SAS)	188
社会科学统计软件包(SPSS)	188
调查数据分析	188
参考文献和补充读物	188

附件一

调查样本设计的基本原理

A.1. 导言	193
A.2. 调查单位和概念	193
A.3. 样本设计	194
A.3.1. 概率样本设计的基本要求	194
A.3.2. 概率抽样对大规模住户调查的重要性	195
A.3.3. 选样、实施和估算的程序	195
A.4. 概率抽样策略的基本原则	195
A.4.1. 简单随机抽样	195
A.4.2. 系统抽样	198
A.4.3. 分层抽样	202
A.4.4. 整群抽样	207

附件二

专家名单

联合国审查住户抽样调查设计手册草稿专家组会议的与会专家名单, 纽约, 2005年12月3日至5日	211
---	-----

表

表3.1. 抽样词汇及相关术语	25
表3.2. 选定用于总体值和样本特征的符号	27

	页次
表3.3. 各种同类相关 δ 和类集规模 n 设计效果的聚类因素之间的比较.....	45
表3.4. 备选样本计划：最后两个选择阶段.....	55
表6.1. 某次调查的反应分类.....	101
表6.2. 在不均等选择概率情况下的权数.....	103
表6.3. 不答复加权调整.....	109
表6.4. 对覆盖面的后分层加权调整.....	111
表6.5. 层方差参数.....	113
表6.6. 权数修整.....	114
表7.1. 每月每户食品支出额(美元).....	119
表7.2. 计算平均参数 \hat{Y} 的实际抽样方差.....	119
表7.3. 选定的人口特征估计数及其方差.....	121
表7.4. 被抽样的住户每周食品支出情况和电视机拥有情况.....	122
表7.5. 分层样本设计数据实例.....	125
表7.6. 作为关切变量在10个计数区内接种疫苗的学龄儿童比例.....	126
表7.7. 按层每周每户食品支出额.....	133
表7.8. 实施最终类集方差估算方法中的各项步骤.....	134
表7.9. 重复方法的数据文档结构.....	137
表7.10. 各种重复方法的方差公式中的常数因子值.....	139
表7.11. 对一个小样本及其子样本应用方差估算刀切重复法.....	139
表7.12. 整个样本：按层支出.....	140
表7.13. 刀切重复法(从第1层删除第2个初级抽样单位).....	140
表7.14. 基于重复品的估计数.....	141
表7.15. 均衡多次重复法(从第1、第3层删除第2个初级抽样单位；从第2层删除第1个初级 抽样单位).....	141
表7.16. 使用各种软件包估计调查估计数方差：在最近生育过的布隆迪妇女中血清反应阳 性妇女的百分比(1988-1989年).....	143
表8.1. 调查误差分类.....	150
表9.1. 调查分析对象/单位示例(取自1987年津巴布韦两次普查之间的人口调查).....	169
表9.2. 1987年津巴布韦两次普查之间人口调查使用的住户和个体文件.....	181
表9.3. 住户预算调查典型文件.....	181
表9.4. 1987年津巴布韦两次普查之间人口调查住户文件使用的平面文件格式.....	181
表9.5. 包含住户调查变量最终数据的观察文件.....	184
表A.1. 按雇员人数分层的学校数.....	206
图2.1. X国住户调查活动时间表.....	10
图2.2. 住户调查方案成本工作表范例.....	13
图3.1. 隐含分层法的行政区安排.....	43
图3.2. 使用与规模成正比的概率系统选择类集的范例.....	53
图8.1. 作为调查误差总量组成部分的抽样误差和非抽样误差之间的关系.....	150
图8.2. 调查误差总量及其各分量.....	151
图8.3. 减少的调查误差总量.....	151

	页次
图9.1. 住户调查流程.....	164
图A.1. 线性系统抽样(样本选择).....	199
图A.2. 圆形等距系统样本选择.....	200
图A.3. 单调线性趋势.....	201
图A.4. 阶段性波动.....	202

第一章

社会及人口统计资料来源

1.1. 引言

1. 在许多国家，住户调查是社会和人口统计的三大资料来源之一。一般认为人口和住房普查也是社会统计的一个主要资料来源，但这种调查通常间隔较长，大约10年左右进行一次。行政案卷系统是第三种资料来源，然而对大多数国家来说，这种资料来源用于健康和生命统计要比用在社会统计方面更多一些。就获取数据的及时性而言，住户调查相对于人口普查来说是一种比较便宜的选项；而相对于行政案卷系统而言则是一种相关性较强而且方便的选项。这种调查用于收集有关人们的生活条件、福利状况、从事的活动，以及人口特征和影响行为举止及社会经济变革的文化因素等方面详细而多样化的社会人口数据。但这并不排除结合使用比如人口普查和行政案卷的其他资料来源补充利用住户调查所生成的数据。

1.2. 资料来源

2. 上述三种主要社会和人口资料来源如能精心策划、圆满执行，或者对于行政案卷来说已有良好基础的话，即可在一体化的数据收集汇编方案中互为补充、相辅相成。社会和人口统计工作对于规划和监测社会经济发展方案至关重要。按年龄和性别分列并包含地理分布数据的人口构成统计资料是描述一个人口或人口子群体所必不可少的基本数据组成部分。这些基本特性为研究有关社会现象——诸如教育、残疾、劳动力参与、健康状况、营养状况、刑事受害发案率、生育率、死亡率和移民情况——的其他重要信息提供了背景环境。

1.2.1. 住户调查

3. 以往60–70年来，住户抽样调查已成为社会现象的一个主要数据源。这种调查所采用的数据收集方法当属最灵活的方法。从理论上讲，几乎任何基于人口的主题都可以通过住户调查来进行调研。在大多数基于地区的抽样策略中，通常把住户当作第二阶段抽样单位来使用（见本手册第三章和第四章）。在抽样调查中选择一部分人口，对其进行观察或收集数据；然后据此做出推断并扩展到整个人口。由于在抽样调查中访调员的工作负荷比较小，用于收集数据的时段比较长，所以大多数题目都可以比在普查中调查得更为详细些。此外，由于所需实地工作人员要比普查工作少得多，所以就有可能招募较为合格的人并加大对他们的培训力度。现实情况是，一个国家所有资料需求并非都可以通过普查得到满

足, 因此, 住户调查提供了一个能够持续不断地满足一些附加和紧急需求的机制。住户调查的灵活性使其成为能够满足用户对统计信息需求的绝佳选择, 因为否则的话此类信息要么不能获得, 要么不够充分。

1.2.1.1. 住户调查的种类

4. 许多国家都制定了住户调查方案, 既有阶段调查又有临时调查。最好将住户调查方案纳入国家的一体化统计数据收集体系。在社会和人口统计领域, 两次普查间的住户调查可以构成这一体系的组成部分。

5. 《国家住户调查能力方案》(NHSCP)有助于发展中国家建立统计和调查能力, 以便从住户方面获取必不可少的社会经济和人口信息。从1979年到1992年, 该方案实施了近14年。到它结束的时候, 共有50个国家参加了方案。它的主要成就是各国促成并采纳了持续进行的多主题一体化住户调查。此外, 该项目还促进了抽样调查能力建设, 尤其在非洲国家。

6. 可以进行各种不同类型的住户调查来收集社会和人口统计数据, 其中包括: 专项调查; 多阶段调查; 多主题调查; 和纵向调查。适当调查类别的选择取决于包括主题要求、资源状况和后勤方面考虑在内的诸多因素。

7. 专项调查是针对单一主题或问题的调查, 比如用时调查或营养状况调查。这种调查可以是阶段性的, 也可以是临时性的。

8. 多阶段调查需要在连续几个阶段收集统计信息, 而每个阶段都是下一阶段的前身。初期阶段一般比以后各阶段的样本规模要大。其功能主要是针对某些特征对各样本单位进行扫描, 以确定这些单位是否适合在以后各阶段使用。这些调查对于确定目标人口, 借以在后续阶段收集关于某个关切主题的详尽信息来说, 是一个成本-效益较高的办法。比如像残疾人和孤儿之类的主题就适合采取此种做法来研究。

9. 在多主题调查中, 单独一项调查涵盖多个不同的主题。这个办法一般要比进行一系列单一主题调查的成本-效益高。

10. 在纵向调查中, 在特定时段内从相同的抽样单位收集数据。调查间隔可以定为一个月、一个季度或一年。这种调查的目的在于衡量相同人口的某些特征在特定时段内的变化。而这种调查的主要问题是受访者人数的缩减率比较高。另外还有调节效应的问题。

1.2.1.2. 住户调查相对于人口普查的优点和局限性

11. 虽然住户调查费用不像普查那样昂贵, 但如果不得不在相对较低层次的行政领域(比如省或地区)单独出结果的话, 这种调查也会相当费钱。抽样调查不同于人口普查: 后者是从数百万住户收集数据, 而前者由于受经费制约一般局限于数千住户的样本, 从而严重限制了出产小区域可靠数据的能力。关于样本容量与小区域数据可信度之间的关系问题, 本手册用了连续几章的篇幅对此进行探讨。

12. 关于住户调查相较于人口普查的优势, 归纳如下:

- (a) 如前所述，总的调查费用一般要比普查低，因为后者需要大量人力、财力、后勤和物力资源。适当选择和调查的概率样本会产生准确可靠的结果，在此基础上可对整个人口做出推断。因此对于某些估计数来说（比如生育率）就没有什么迫切需要进行人口普查；
- (b) 一般来说，抽样调查产生的统计信息质量较高，因为前面讲过，这种调查有较大的可行性来雇用素质较高、较为训练有素的访调员。同时也比较容易更好地实施监督，因为一般监督员都经过良好训练，监督员/访调员的比率可能会高达1比4。此外，还有可能在现场使用较好的技术设备进行必要的实际测量。在人口普查中由于调查规模太大，有时候不得不在数据质量方面做出一些牺牲。这就在质量保证上导致出现各个阶段的瑕疵和疏漏，从而造成很高的非抽样误差；
- (c) 就调查深度和问卷中的项目数量而言，抽样调查要比普查范围大，也较为灵活。在人口普查中不大可能收集比较专业化的信息，因为专家人数有限，或者因为受到调研所需设备数量的制约。举例来说，在营养调研中就不宜进行食品加权和其他测量。也不宜对相关人口中的每个人都进行医学检查以测定比如艾滋病毒/艾滋病感染之类的发病率。另外，还有可能在住户抽样调查中增加一些对普查来说比较复杂的项目。

13. 抽样调查比较适合在全国和相对较大地理范围就一些需要深入考察的多方面问题——比如残疾、家庭支出、劳动力活动和刑事受害等方面——收集数据。这与人口普查形成鲜明对比，普查是相对较一般性的覆盖较小领域的信息源。

14. 总的来说，住户调查统计工作的优点包括：一是数据收集手段灵活，适合针对各种各样、数量较多的问题进行调查；二是有可能对一些可与人口和住房普查参数作比较的参数做出估计。

1.2.2. 人口和住房普查

15. 人口普查(以下简称“普查”)包含特定时期覆盖全国或明确界定的国内一个或多个政区内所有个人的人口、社会和其他数据的收集、汇编、评价和传播的全过程。它是社会统计的一个主要数据来源，其主要优势在于能提供较小地理单位的可靠数据——亦即不受取样误差影响的数据。普查是提供有关人口规模、构成和空间分布的信息以及社会经济和人口特征的一个理想手段。一般来说，普查旨在收集一个住户和每套住房中每个成员的信息，通常调查范围涵盖全国或明确界定的国内若干政区。

1.2.2.1. 传统人口和住房普查的基本特征

- 16. 传统的人口和住房普查有下述特征：
 - (a) 人口中的个体和每套住房中的个人要单独调查，因此相关特征也是分开记录的；

- (b) 目标是覆盖明确界定的领土范围内的整个人口。它意在包含现有的每一个人和/或通常居民, 这要取决于人口计数的类型是事实上的还是法律上的。在没有综合人口登记册或行政登记册的情况下, 普查就是小区域统计数据的主要来源;
- (c) 全国范围的调查一般要尽可能做到同步。所有个人和住所都以同样的所涉期计数;
- (d) 一般都是按规定的间隔进行调查。大多数国家都是每10年搞一次普查; 其他国家每5年搞一次。这样做便于产生固定间隔的可比信息。

1.2.2.2. 普查成果的用途

17. 普查成果一般有如下用途:
- (a) 普查成果提供有关人口规模、构成和空间分布的信息以及人口和社会特征信息;
 - (b) 普查是小区域统计的一个数据源;
 - (c) 普查计数区域是住户调查抽样标架的主要数据源。在普查中收集的数据往往用作分层样本和改进住户调查估算的辅助信息。

1.2.2.3. 人口普查的主要局限性

18. 由于其无以伦比的地域覆盖面, 普查通常是人口特征基线数据的主要来源。因此不宜涉及许多相当详细的主题。举例来说, 普查可能不是有关经济活动的详细信息的最理想来源。此类信息需要详尽的提问和探索。

19. 由于普查的访问调查工作在很大程度上有赖于代理答卷人, 所以它不一定能准确捕捉到只有当事人才知道的特征信息, 比如职业、工作时数、收入, 等等。

20. 过去几十年来许多国家都进行了人口普查。例如在2000年那一轮(1995-2004年), 共有184个国家和地区进行了普查。

1.2.3. 行政案卷

21. 有许多种类的社会统计是根据各种行政案卷汇编而成的行政程序副产品。这方面的实例包括根据医院的案卷汇编的保健统计数据, 根据职业介绍所的数据汇编的就业统计, 根据民事登记系统的生命统计数据, 以及根据教育部的入学报告汇编的教育统计数据。根据行政案卷汇编的统计数据的可靠性取决于行政案卷的完整性以及定义和概念的连贯性。

22. 虽然行政案卷可能是成本-效益很高的数据源, 但是这些行政系统尚未在大多数发展中国家建立起来。这就意味着, 在大多数情况下此类数据都不准确。即使为了行政的目的继续执行相关行政记录的程序, 在大多数情况下对于多数组织来说统计数据的汇编也属于

次要关切事项，从而影响到数据质量。通常不考虑、也不遵守需要达到的统计要求，比如概念和定义的标准化、覆盖面的及时性和完整性。

23. 对大多数国家来说，出自行政案卷的信息往往在内容上有限，因为它们对法律或行政目的比较有用。民事登记系统在许多国家发展起来的行政(案卷)系统的实例。但并非所有国家在这方面的工作都是成功的。拥有完备人口动态登记系统的国家能够提供阶段性人口动态事件报告，比如按性别分列的活产儿人数、出生日期和地点，按年龄和性别分列的死亡人数、死亡地点和死亡原因，以及结婚和离婚等。

24. 人口登记簿保存了一个国家的每个人和每个住户的生命数据库。该登记簿随着个人和/或住户的变化而不断更新。如果该登记簿与其他社会登记簿相结合，即可构成一个非常丰富的信息源。已经发展了此类系统的国家有丹麦、挪威、荷兰、德国和瑞典。这些国家大多将普查建基于这种登记系统。

25. 在许多发展中国家，虽然可以利用各种社会方案的行政案卷作为成本-效益较高的数据源，而且这也是一个很有吸引力的主张，但是这些数据源还有待发展。行政案卷往往在内容上有局限性；就概念或主题细节而言，它们通常也难以同住户调查相匹配。在这种情况下，能否利用行政案卷来辅助其他数据来源就构成一大挑战，因为缺乏标准化的概念和分类系统，加之覆盖范围有选择性且覆盖不足。

1.2.4. 三种数据源的互补性

26. 本章着重提出了配合使用普查、调查和行政案卷系统的各种方法。本节探讨依互补的方式较详细地综合利用出自不同数据来源的信息这一课题。之所以关注这方面的问题，是因为有必要限制普查和调查成本并减轻答卷的负担，以便提供较低层域(比如调查数据可能涉及不到的域)的数据，同时最大限度地利用国家的可用资料。

27. 由于普查不能频繁重复进行，所以住户调查就为某些普查信息的更新、尤其在国家一级和其他大域层级的信息更新提供了依据。在大多数情况下，在普查中只调查相对比较简单的问题，而且一般提问的数量有限。因此可以利用住户调查的规模小和具有潜在灵活性的特点，用其收集到的关于复杂问题的详细信息来补充普查信息。

28. 普查和住户调查在很多情况下是相互补充的。在普查期间收集根据某个住户样本追加问题的相关信息，是拓展普查范围以满足扩大社会统计需求的一种成本-效益较高的做法。通过使用抽样方法和技巧可以产生急需的、精度达到可接受标准的数据；否则的话，在时间和经费有限的情况下不可能通过完整的人口清查获得这些数据。

29. 普查也可以提供在进行住户调查时所需要的抽样标架、统计基础结构、统计能力和基准统计数字。通用的做法是在普查的背景范围内提取一个住户样本，以便收集较为复杂问题的信息，比如残疾率、产妇死亡率、经济活动和生育率。

30. 普查通过提供抽样标架支持住户调查：普查提供一个所有区域单位的明晰清单，比如人口计数区，一般在住户调查选择过程中当作第一阶段单位使用。另外，从普查中获取

的一些辅助信息可用于调查的有效设计。再者，普查产生的辅助信息还可以用于通过回归和比率估计改进样本估计数，从而提高调查估计数的精度。

31. 为了实现数据来源一体化，有必要确认计数单位并且在通过各种数据源收集和报告统计数据的时候采用连贯一致的地域单位。此外，在包括行政案卷在内的各种不同数据源之间采取通用的定义、概念和分类法也是至关重要的。

32. 住户调查所产生的数据亦可用于核查人口普查的覆盖面和内容，目的在于确定误差的大小与方向。例如，在赞比亚和柬埔寨2000年那一轮普查期间，计数后的调查就用于评价覆盖面误差这个目的。同样，普查数据也可以用于评价某些调查结果。

33. 由于对可靠的小区域估计量的需求日益增长而备受重视的小区域估计是一个同时使用调查数据和行政案卷生产估计数的领域。传统的地区性直接估计数不能提供足够的精度，因为小区域的样本规模很少有足够大的。小区域估计一般基于一系列用以生产区域估计数的统计技巧，而这些区域的传统调查估计是不可靠的或者是不能计算的。这些技巧涉及到通过补充或辅助数据(比如较近期的人口普查数据)与相关小区域相联系的模型。因此，小区程序的基本思路就是借用并结合各种不同数据源的相对长处，以便产生较为准确可靠的估计数。

34. 在拥有完善民事登记系统的国家，普查和调查数据可以成功地连同行政案卷一起使用。例如，在1990年新加坡的人口普查中，访调员掌握了根据行政案卷预先填好的关于每个住户成员基本信息。这个办法减少了访问调查时间和调查成本。基于登记簿的普查只是提供人口总清点及其基本特征，而详细的社会经济特征要在抽样基础上来收集。

35. 出自行政案卷的数据可以用来核查及评价调查和普查结果。举例来说，在有完备人口动态登记系统的国家，出自普查的生育率和死亡率数据可以跟源自登记系统的同类数据进行交叉核对。

1.2.5. 结论意见

36. 总之，应把住户调查、普查和行政案卷这三种数据源视为相互补充的关系。这就是说，在制订普查和调查计划的时候要尽可能使用通用的概念和定义。还要定期检查行政程序，以确保使用通用的概念和定义。

37. 一个国家的住户调查方案应该是包括普查和行政案卷在内的一体化统计数据收集系统的组成部分，这样才能充分满足对社会人口统计的总体需求。

参考资料和补充读物

Ambler, R. 等(2001年)。《结合失业津贴数据和劳动力调查(LFS)估计国际劳工局小区域失业情况：修改的Fay-Herriot方法的应用》。征文。国际统计学会届会，首尔。

- Banda, J. (2003年)。《社会统计现状：问题与关切事项综述》。提交联合国统计司与锡耶纳社会统计小组联合举行的关于确定社会统计范围专家组会议，纽约，2003年5月6-9日。
- Bee-Geok, L. 和K. Eng-Chuan (2001年)。文件号：ESA/STAT/AC. 88/05, 4月7日。《新加坡2000年人口普查中调查与行政数据相结合》。征文。国际统计学会届会，首尔。
- Kiregyera, B. (1999年)。《抽样调查：特别关于非洲》。坎帕拉：PHIDAM企业。
- Rao, J. N. K. (1999年)。基于模型的小区域估计的最新进展。《调查方法论》(加拿大统计局，渥太华)，第25卷，第2期，第175-186页。
- Singh, R. 和N. Mangat (1996年)。《调查抽样要素》。马萨诸塞州波士顿：克伦威尔学术出版社。
- 加拿大统计局(2003年)。《调查方法与实践》。渥太华。
- 联合国(1982年)。《国家住户调查能力方案：住户调查中的非抽样误差：来源、评估和控制》(初版)，DP/UN/INT-81-041/2，纽约：联合国发展技术合作部和统计办公室。
- _____ (1984年)。《住户调查手册》(修订版)，《方法研究》，第31号，出售品编号：E. 83. XVII. 13。
- _____ (1998年)。《人口与住房普查的原则和建议》，首次修订。统计文件，文件号：67/Rev. 1，出售品编号：E. 98. XVII. 8。
- _____ (2001年)。《生命统计系统的原则和建议》，第2次修订，出售品编号：E. 01. XVII. 10。文件号：ST/ESA/STAT/SER. M/19/Rev. 2。
- _____ (2002年)。《关于在人口普查中收集经济特征数据的技术报告》。纽约和日内瓦：社会经济事务部统计司和国际劳工局统计司。文件号：ST/ESA/STAT/119(仅英语文本)。
- Whitfold, D. 和J. Banda (2001年)。《计数后的调查(PES's)：是否值得？》。在关于2000年轮次人口与住房普查：十年中期评估和未来展望全球审查研讨会上发表，纽约，8月7-10日，由联合国秘书处经济和社会事务部统计司召开。文件号：Symposium 2001/10(仅英语文本)。

第二章

调查计划与执行

1. 虽然本手册的重点放在住户调查的抽样方面，但是有必要对住户调查的规划、运作和实施作一概述，以便将有关抽样的章节置于适当的背景环境之中。已有许多教科书、手册和指南相当详尽地处理住户调查计划与执行的问题，请读者参阅这些资料以获得更多信息。不过，本章突出概括了包括计划与执行关键特征在内的诸多要点，但样本设计和选样除外，这些问题在第三和第四章及附件一中有专题讨论。

2.1. 调查计划

2. 为了使调查获得预期的成果，需要特别注意启动现场调查之前的筹备工作。若在这方面取得成功，所有调查都需要做好仔细而审慎的准备工作。不过，计划工作量多少因所需调查、材料和信息类别而异。鉴于制订完善的调查计划需要有足够的时间和资源(见图2.1)，两年计划周期并非罕见(关于调查计划的详细讨论，见联合国, 1984年)。

图2.1是一份时间表图解。

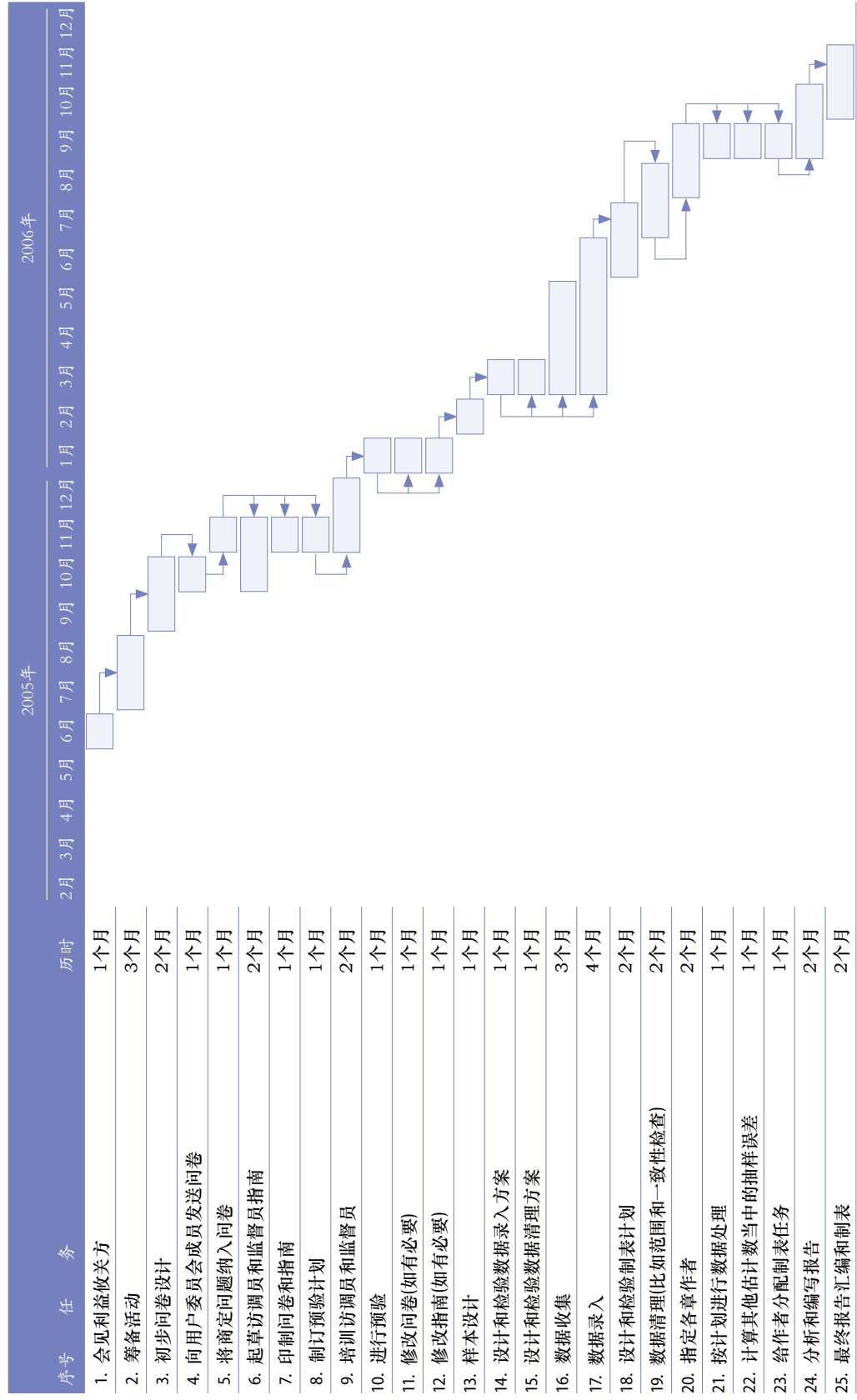
2.1.1. 调查目标

3. 必须从项目一开始就清楚地说明调查的目标。应当有一项明晰阐述的关于预期信息的统计声明，清楚地描述人口和地域覆盖面。还有必要在这一阶段说明准备如何使用调查成果。既定的调查预算应指导调查统计人员量身定制所要达到的目标。恰当认知预算的制约因素，会有助于成功地制订计划和执行调查。

4. 有些情况下没有说明调查目的。譬如讲，可能要求一个调查组织执行一项关于非正规部门活动的调研。如果没有说明目的，就必须由统计学家或调查管理者用适用于调查工作的措辞给非正规部门下定义，详细阐述最接近于体现主管机构要求的特定经济活动。应该指出的是，雄心勃勃但目标模糊的调查工作非常有可能造成很高的非抽样误差率。

5. 让利益攸关方——亦即统计数据的各种用户和生产者——参与界定调查目标及其范围和覆盖面至关重要。协商的过程有助于在诸如需要什么样的数据、需要数据的方式、分解的层次、传播策略和数据收集的频率等问题上取得一致或达成妥协。

图2.1
X国住户调查活动时间表



6. 由调查组织进行的一些调查有明确的目标。例如1983年赞比亚进行的试验性人力资源调查就有下述目标：

- (a) 收集有关正规部门现有工作人口规模与构成的信息；
- (b) 评估人力需求与供给；
- (c) 为特定职业进行人力预测奠定基础；
- (d) 在经济发展关键领域协助规划扩大教育工作。

7. 应当指出的是，有明确申明的目标是确定需要给出统计答案的调查问题的第一步。

2.1.2. 调查全域

8. 在计划一项调查的时候，有必要界定行将涵盖的地理区域和目标人口。譬如讲，在住户收入和支出调查中，调查可能覆盖城市地区而将农村地区排除在外。

9. 在界定全域方面，应确定拟抽样的确切人口。在住户收入和支出调查中，第一阶段单位的全域将是计数区（这些是——譬如讲——遍及全国的地域单位），而第二阶段单位将是选定计数区内的住户（第三章和第四章较为详细地讨论了聚类问题；关于类集的定义，另见附件一）。

10. 不过应该指出的是，在实际中目标人口要比构成全域的人口小一些。通常有若干理由限制目标人口。有的调查中可能把兵营中的某些军事住户排除于调查之外。在劳动力调查中，规定年龄以下的儿童可能显示为被调查住户成员，但不算劳动力的组成部分。

11. 应着重指出的是，当实际人口与目标人口有出入的时候，调查结果将适用于从中提取样本的特定人口。第四章将讨论在每个选择阶段构建综合性的相互排除标架。

2.1.3. 拟收集的信息

12. 根据一份需要给出统计答案的提问清单可以编制一份项目清单，它可以围绕调查的问题提供真实信息。任何时候都要牢记，某些需要的数据可以从现有的数据源获得。在编制项目清单的时候要规定包括收集与主项目相关联的辅助项目。比如在就业和收入调查中，可以稍带收集关于年龄、性别和教育程度的辅助信息。这些信息有助于更深入地理解相关问题，从而丰富分析的内涵。

13. 我们可以补充一点，即在制订调查计划的时候要做出制表计划。应通过传阅空白表格征求意见，加以改进。

2.1.4. 调查预算

14. 调查预算表明对行将进行的调查提出的财务要求。预算是为支持和指导实施调查及编制产生调查成果的时间表所需要的。成本估计一定要尽可能详细。因此有必要了解调查

工作各有关步骤的全部细节。预算表明人员、设备和所有其他支出项目的成本费用。如果预设了可用经费上限(通常都是如此),那么总的调查预算就不得超出这个预设的框架。还可以按照资助机构在编制预算方面的一般准则行事。这样做会有利于概算的审批。如果需要偏离预先确定的预算,必须获得相关组织核准。调查经费申请应在早期阶段提出。一般来说,预算在很大程度上取决于调查设计、要求达到的精度和地域覆盖面积。下面图2.2展示了可能的成本工作表。

15. 在正在进行调查的组织内建立有效的成本控制系统至关重要。在大多数大规模调查工作中都存在着一旦启动现场工作就会在监督资金支出方面出现失控的高风险。在这种情况下,可能有大量资金划拨给与主要调查工作无关的方面。审慎的成本控制有助于通过在估计费用和实际完成工作之间作比较来监测实际支出情况。负责调查工作的管理层必须确保经费开支的问责制。这可以大大提高调查组织的信誉。

2.2. 调查的执行

2.2.1. 收集数据的方法

16. 现有许多收集数据的方法,其中包括:直接观察和计量法;邮寄问卷法;以及电话和个人采访法。

17. **直接观察和计量法:**直接观察和测量是最理想的方法,因为通常此种方法比较客观。无需担心答卷人或调研员会记错,也无需担心他们会带有主观性。采取直接观察方法的领域有如下实例:

- (a) 食品消费调查的某些方面;
- (b) 调查员可以通过购买产品和记录价格来收集物价。

18. 这种方法虽然很有用,但缺点是既耗费资源又费时间。大多数情况下调研员都必须使用一些设备。经验表明,直接观察和计量法一般在样本和人口规模相对较小的情况下比较有用。

19. **邮寄问卷法:**采用邮寄问卷的方法相当便宜和快捷。在收集数据阶段主要花费是邮资。在完成问卷设计和印制后,将其邮寄给答卷人(预期这些人会填完答卷)。假定答卷人识字,因为指望他们会亲自填写问卷。而这种假设可能是错误的——尤其在发展中国家,那里的识字水平还很低。这种方法的主要缺点是不回答问卷的比率非常高(即未填写问卷的人和/或只回答了其中某些问题的人占了很大比重),这也许是因为所使用的问卷太复杂了。可是也不能完全否认不答复者对调查持冷漠态度。有些情况下,问卷回应的情况不错,但是个别项目不回答的比率较高。

20. 为了力争提高反应率,可能需要向不答复者(即没有回答调查问题的人)发出提醒函。不过,最好是从不答复者当中选择一个二段样本,对其采取个人采访的方法。这种做法可能是必要的,因为不答复单位的特征也许完全不同于答复者的特征(见第三章和附件一关于分层及其优点的讨论)。在这种情况下,答复单位和不答复单位是作为两个后层次/域来处

图2.2
住户调查方案成本工作表范例

活 动	估计的 工作单位 (人-月, 其他表 示除外)	单位成本 (每人-月的 相关货币 单位, 其他 表示除外)	估计总成本 (相关货币 单位)
一、 计划和准备活动			
A. 初步计划和后续监测 (高级人员)			
B. 主题的选择和说明			
1. 主题的计划			
2. 制表计划的准备			
3. 秘书和其他服务			
C. 调查设计的发展			
1. 初步设计的计划: 调查结构, 人口覆盖面, 抽样程序, 数据 收集方法, 等等(专业人员)			
2. 抽样材料的发展:			
a) 制图材料(假定有可用的普查材料): 人员费用 地图和供给品			
b) 现场住户列表(2 000个计数区): 人员费用(主要是访调员) 差旅费			
c) 根据现场名单选择和准备样本			
D. 问卷和其他材料的设计与印制			
1. 专业人员			
2. 秘书和其他服务			
3. 印制费(预验后)			
E. 预验			
1. 专业人员计划:			
a) 初步准备			
b) 成果分析和材料修订			
2. 现场监督员:			
a) 人员费用			
b) 差旅费			
3. 访调员:			
a) 人员费用			
b) 差旅费			
F. 现场用的指导和培训材料			
1. 专业人员			
2. 秘书和其他服务			
3. 复制费			

图2.2
住户调查方案成本工作表范例(续)

活 动	估计的 工作单位 (人-月, 其他表 示除外)	单位成本 (每人-月的 相关货币 单位, 其他 表示除外)	估计总成本 (相关货币 单位)
G. 杂项计划活动 (比如公关和宣传)			
H. 小计要素			
1. 高级人员			
2. 专业人员			
3. 技术人员			
4. 服务人员			
5. 差旅			
6. 印制			
7. 制图和杂项			
小 计			
二、现场工作			
A. 现场监督员的培训			
1. 人员费用			
2. 食宿费用			
3. 差旅费用			
B. 访调员的培训			
1. 监督员费用			
2. 访调员费用:			
(a) 人员费用			
(b) 差旅费用			
C. 数据收集(包括质量控制)			
1. 监督员费用			
2. 访调员费用:			
(a) 人员费用			
(b) 差旅费用			
D. 现场管理			
1. 现场指导			
2. 差旅			
3. 其他费用 (比如材料控制和运输)			
E. 小计要素			
1. 专业人员			
2. 技术人员			
3. 服务人员			
4. 差旅			
5. 差旅补贴			
6. 访问调查			
7. 杂项			
小 计			

图2.2
住户调查方案成本工作表范例(续)

活 动	估计的 工作单位 (人-月, 其他表 示除外)	单位成本 (每人-月的 相关货币 单位, 其他 表示除外)	估计总成本 (相关货币 单位)
三、 数据处理			
A. 系统计划			
B. 计算机编程			
C. 文书编码			
1. 初步编码			
2. 质量控制			
3. 监督			
D. 键盘-磁盘操作			
1. 初步键入			
2. 质量控制			
3. 监督			
E. 计算机时间 (包括操作员和维修费用)			
F. 杂项处理费用 (供给品, 等等)			
G. 小计要素			
1. 专业人员			
2. 技术人员			
3. 质量管理员			
4. 服务人员			
5. 计算			
6. 杂项			
小 计			
四、 数据审查和出版			
A. 专业时间			
B. 出版成本			
五、 调查指导与协调 (持续监督一切活动)			
六、 小 计			
七、 评价研究和方法论研究 (可以估计占累计数的10%)			
八、 一般间接费用 (可以估计占行政管理费、场所租用费和一般供给品等 方面累计费用的15%)			
九、 总 计			

资料来源：联合国(1984年)。

理的，即在准备估计数的时候必须予以差别加权(后面的章节尤其是第六章进一步讨论了调查加权的问题)。为了提高反应率，邮寄的问卷应尽可能简短明了、富有吸引力。加盖“邮资已付”邮戳且印好回函地址，可能有助于提高反应率。

21. 为了能够令人满意地使用这个方法, 还必须有一个尽可能时新的抽样标架。因此, 答卷人的地址必须是最近的。调查组织还必须确信答卷人能够自行填写问卷。

22. 下面是邮寄问卷调查方法的一些优点和局限性:

优点:

- (a) 费用较低;
- (b) 样本可以推广;
- (c) 排除了访调员的偏差;
- (d) 快速运转。

局限性:

- (a) 通常不答复率较高;
- (b) 对问题的回答只能取其字面数值, 因为没有机会深究;
- (c) 在态度调查中很难查明答卷人是否在没人帮助的情况下回答问题;
- (d) 该方法只有在问卷相当简单明了的情况下才有用, 因此不适合调查复杂问题。

23. 个人采访法: 这是发展中国家通过大规模抽样调查收集数据所采用的最普通方法。个人采访法之所以适合采用, 除了它所产生的反应率较高之外, 还因为在一些调查国家文盲率普遍都很高。这种方法就是访调员去到选定的受访者家里通过提问收集信息。其主要好处就是访调员可以(通过动员)说服受访者回答问题并可以解释调查的目的。另外, 当采用个人采访法的时候有较大可能在概念上存在困难的项目上收集到统计信息, 而在邮寄问卷中此类项目得到的回答很可能是模棱两可的。

24. 不过, 使用个人采访的方法也存在着固有的局限性。比如:

- (a) 不同的访调员可能对相关问题做出不同的解释, 从而给调查结果带来偏差, 因为很少有访调员坚持参照指导手册;
- (b) 在问答过程中有些访调员可能会向受访者推荐答案;
- (c) 访调员的个人特征——比如年龄、性别、有时甚至种族——也会对受访者的态度产生影响;
- (d) 访调员可能会读错问题, 因为他们在访谈和做记录之间有可能分散注意力。

25. 总之, 上述局限性是造成所谓访调员偏差的主要原因, 而正如各种研究表明, 此种偏差可能会在调查中造成严重的非抽样误差(见第八章关于非抽样误差的讨论)。

26. 在向受访者提问的时候要考虑到下列要点:

- (a) 访调员要清楚地了解访调员手册中解释的每个问题的目的所在。访调员有必要经常参照指导手册;

- (b) 经验表明，访调员最好按顺序提问调查问卷中的问题。大多数问卷都仔细考虑了各项问题的排序，考虑到受访者的动机，话题之间的联系，启发受访者对往事的回忆，以及最敏感的问题；
- (c) 访调员应当想方设法避免向受访者推荐答案；
- (d) 所有问题都要问到，以避免留下未回答的项目。另外，问卷中任何项目都不得留有空白，除非这符合略过方式。如果某个问题与特定受访者无关，那么应加备注。这种做法可以使调查管理者确信，问卷所包含的所有问题均已得到执行。

2.2.2. 问卷的设计

27. 调查目标和制表计划一经敲定，即可编拟相关问卷了。问卷在调查过程中起核心作用，在此过程中信息从掌握信息者(受访者)手中传递到需要信息者(用户)手中。问卷是用可操作词语表达用户对信息需求的手段，也是向特定调查数据处理系统录入数据的主要依据。

28. 问卷的大小和格式需要非常仔细的考虑。最好在制订调查计划的时候设计问卷。如果需要向答卷人邮寄问卷的话，问卷应简明且富有吸引力。这样可以提高反应率。另一方面，准备用于访调员在现场记录受访者回答内容的问卷应当坚实，不受调查操作的影响。

29. 理想的情况是，设计的问卷应该有利于收集相关的准确数据。为了提高调查精度，应特别考虑问卷中的项目顺序安排及其措词。要调动受访者的积极性。要妥善安排问卷，以方便受访者或访调员阅读问题。每一份问卷都应载有明确的指示，对这一点怎么强调都不过分。

30. 因此，调查队要特别注意准确界定所要收集的数据并确切说明如何将数据要求及相关概念转化成可操作的提问。就此而言，问卷的预验就成为一项常规而且一般来讲必须要做的一项工作，除非问卷业经历次调查的充分验证。

31. 总之，好的问卷应当：

- (a) 能够收集精确的信息以便及时满足潜在数据用户的需求；
- (b) 便于数据收集、数据处理和制表；
- (c) 确保数据收集经济实惠，亦即要避免收集任何不必要的信息；
- (d) 要能够据以进行综合而富有意义的分析并且有目的地利用所收集的数据。

32. 这就意味着，拟订的调查问卷要能产生尽可能高质量的信息，尤其要注重信息的针对性、及时性和准确性。为了能够高效率地完成调查任务，必须把收集必要信息所涉及的成本和负担降至最低限度。

2.2.2.1. 提问的设计

33. 在抽样调查问卷中既有无确定答案的问题又有限定答案的问题。在无确定答案的问题中，受访者对问题给出自己的答案。在态度调查中，可能请受访者界定他们认为什么

是好的生活质量。很明显，不同的受访者会用自己的方式界定什么是好的生活质量。另一方面，限定答案的问题则把受访者限制于从调查队给出的一份清单中选择答案。下面是限定答案的问题实例：

您是否有影响您日常生活的永久性心智残障问题？

是 不是

您如何评价自己的视力(甚至戴眼镜或隐形眼镜片——如果用的话)？

1. 失明
2. 严重永久性视力障碍
3. 某种程度的永久性视力障碍
4. 无障碍

34. 使用限定答案的提问方式好处是：(a) 答案比较统一；(b) 易于数据处理。这种提问的主要局限性是须由调查设计者设定可能的答案。在这种情况下，可能的重要回答或许忽略不计。在大多数调查中，与态度和观念有关的或许不得而知的复杂问题和提问采用无限定答案的方法最合适。

2.2.2.2. 提问的措辞

35. 提问应当明白、准确、毫不含糊。不能让受访人去猜测访调员想从他(她)这里听到什么。所使用的定义和概念也许对调查管理者来说似乎显而易见，但是对受访者来说可不是这样。于是，受访者就可能运用自己的判断力来回答问题。最终结果可能就是滋生非抽样误差。举一个简单例子。比如“您家的住址在哪里？”这个问题，在许多非洲国家如果不指明界定“家”的含义的话就会产生混淆，尤其对城镇人口来说更是如此。有些受访者会把“家”理解为他们“老家”的村庄。

2.2.2.3. “别有用意的”问题

36. 一个所谓别有用意的问题试图诱导受访者以某种方式回答问题。这就是说，该问题倾向于支持某种回答。下面举一例在健康调查方面别有用意的提问：“您每周有几天喝啤酒超过两瓶？”这个问题强迫受访者承认他(她)喝啤酒——实际上每天至少两瓶。如此提问使得问题的答案带有偏差。避免制造数据很重要：调查的目的只不过是收集数据。

2.2.2.4. 问题的相关性

37. 问卷的目的在于引导出信息，用以研究境况。因此，调查组织必须提问相关的问题才能获得所研究的特定境况的真实图景。载入问卷中的问题应当与大多数受访者有关。譬如讲，在当今大多数非洲国家的典型农村环境执行的问卷，如果夹杂了有关个人在高等(大学)教育方面的业绩如何的提问，那就毫无意义。同样，在生育率调查中也不适合包括10岁

以下的女孩，并且问她们生过几个孩子，或是否结过婚、离过婚，或者守寡之类的问题。这些问题与一定年龄以上的女性有关，但是跟尚未达到生育年龄的女孩无关。

2.2.2.5. 提问的顺序

38. 问卷中的项目排序应当有利于促动和方便受访者回忆并有助于引导出准确的信息。建议前面的问题要容易回答，令人感兴趣而且不敏感。这样有利于受访者树立信心，从而使他(她)能够跟着采访进程走下去；而多数情况下他们都是自愿参与的。在住户调查中一般的次序是先从确认样本单位入手，比如家庭住址；然后引导描述住户及其个体成员，比如其中包括各项人口特征；最后再逐一提问构成调查主题的各个细节问题——这样的次序现已成为一种相当标准的做法。¹一般而言，敏感的问题肯定属于最后提出的问题之列。这里要强调一点，就是在各个问题——尤其是那些不确定的或有问题——之间必须要有逻辑联系。

2.2.3. 制表和分析计划

39. 有一种有用的技巧可以帮助调查设计者提高其所设计工具的精度，以满足用户的信息需求，这些需求体现在一整套调查问题或调查目标之中。这种技巧涉及到制表计划和虚拟表格。虚拟表格是草拟的表格，包括除了实际数据外的任何内容。制表大纲至少要具体说明表格标题和栏目设置，确定拟纳入表中的实质性变量，拟用于分类的背景变量，以及各种表格所适用的人口群体(调查对象、要素或单位)。另外，最好也尽可能详细地展示分类的类别，尽管当日后更好地了解答复分类的样本分布时，这些类别可能要做出调整。

40. 可以从几个方面来看制表计划的重要性。譬如讲，制定的虚拟表格能够表明待收集的数据是否会产生可用的表格。这些表格不仅能够指出缺少什么内容，也能反映出多余的内容。另外，在制作虚拟表格方面额外花费的时间，通常会在数据制表阶段通过缩短设计和制作实际表格所用的时间而得到加倍的补偿。

41. 还需要考虑调查所要利用的抽样设计和制表计划之间的密切关系。譬如讲，只有在设计的样本允许特定地域划分的情况下，才可以在表格中如此划分地域。再者，样本的规模可能会制约交叉制表中的方格数量，以免产生过分稀疏的表格。有时候在制表工作过程中可能不得不对计划加以修改。可能需要合并一些类别以减少空格；或者在数据草案中有了令人感兴趣的新发现之后有必要制作新表格。在住户调查中收集到的数据被用来回答问题(实现目标)的方式方法可以一般化地叫做“数据分析计划”。这样的计划详细说明了实现调查目标需要什么数据。调查设计者在编制调查问卷的过程中必须经常参照这种数据分析计划。不言而喻，这种分析计划也应该是指导分析调查成果的主要参考点。

2.2.4. 现场工作实施

42. 在许多发展中国家，现场工作的实施受到资源匮乏的严重制约。然而，要想进行

¹ 见联合国，1984年。

调查就应妥善组织和实施现场工作，以便使得调查队所能支配的有限资源能够得到有效利用。为了成功开展调查工作，参与调查作业设计的人应当清楚地了解调查主题的概念。另外，访调员必须透彻掌握实际业务程序，才有可能成功收集准确数据。为了圆满完成调查工作，永远需要妥善有效地做好现场组织工作。

2.2.4.1. 设备和材料

43. 在许多发展中国家，需要预先准备好车、船、自行车等交通设备并使其处于随时待用状态。还需要有一些备用零部件。机动车和自行车便于团队领导人、监督员及访调员快速流动。

44. 要充分供应适当的材料，比如文件夹、笔记板、铅笔、铅笔刀、笔记本和(车用)燃油等，供调查工作期间使用。

2.2.4.2. 调查工作管理

45. 大规模抽样调查是一项要求苛刻的复杂工作。因此，各个层级都有必要实施审慎、有效和高效率的活动管理，不可对此掉以轻心。

46. 从调查管理人到访调员之间，必须建立明确规定的指挥路线。值得指出的是，现已发现监测调查进展的控制形式很有用。

2.2.4.3. 宣传工作

47. 有些调查取得的成功很有限，究其原因，部分是因为人们拒绝回应而造成了较高的不答复率。因此，调查组织者有义务为开展调查搞一些宣传运动。经验表明，宣传工作可以在争取受访者合作方面发挥重要作用，尽管某些资助组织或机构认为宣传方面的开支是浪费资源。

48. 因主流社会环境各异，可以采取不同的宣传做法。比如在一些国家的城市地区，广播、电视和报纸信息可以辅助海报，而在农村地区，可以发挥广播信息和海报的作用。

49. 另外，也许有必要和当地选区的舆论带头人一起安排一些会议。可在会上向人们宣讲调查目的。另外可以要求这些领导人说服本地区的老百姓向访调员提供必要的信息。

50. 在进入现场之前，要公布有关开展调查的法律规定。公告中除了提供其他信息之外，要包括调查的目标和期限以及行将涉及的话题。

2.2.4.4. 访调员的遴选

51. 访调员处在与受访者的交界面。他(她)代表调查组织，总是与受访者保持接触，这就清楚地说明了为什么访调员的工作对于调查方案成功与否如此至关重要。因此应高度重视并认真进行访调员的遴选工作。访调员要能与受访者进行有效沟通。他(她)应具备在合理的时限内引导出合乎精度要求的所有信息所需要的素质。

52. 因不同的调查类别而异，访调员应具有适当的学历。此外，访调员要能如实记录信息，而不得“添油加醋”。被录用的访调员要遵照指示，使用访调员现场指南所规定的定义和概念。

53. 下述程序可能有助于选择合适的访调员：

- (a) 预期的访调员应填报一份申请，注明本人年龄、婚姻状况、现住址、教育程度和就业简历；
- (b) 初选合格者可能要经过智力测验和简单数字计算的附加测验；
- (c) 通常，除了笔试以外还需要对候选人进行面试。应由一个独立专家小组进行面试，给候选人评级。在给候选人评级方面需要考虑的一些特征包括：待人友好、对工作感兴趣、表达能力强、思维敏捷。

54. 现场工作可能是单调乏味的，可能要艰难穿行崎岖的山路，因此入选的访调员应承诺，准备在艰苦条件下工作。

2.2.4.5. 访调员的培训

55. 入选的访调员在被派往现场之前须经过充分培训。培训项目的主要目的是使调查的采访程序一致。这对于防止访调员对调查定义、概念和目标做出不同的解释，从而最大限度减少访调员工作出现偏差自然是必要的。

56. 应由合格的培训教员负责执行培训工作。很明显，这些教员要熟知调查的宗旨和目标。他们最好也是执行调查任务的调查队成员。

57. 应向访调员仔细教授调查宗旨和打算如何使用调查结果。为了使访调员能够恰当地理解调查的各项目标，他们必须在问卷所使用的各种概念和定义方面受到很好的训练。

58. 作为培训过程的组成部分，访调员们要当着培训教员的面轮流向其他学员解释问卷中的各个项目。练习课应安排在教室和现场两种环境下进行。比如，访调员可以在教室环境里彼此轮流提问题，然后带他们到附近的左邻右舍搞一次现场练习，学员们可以在那儿采访几个住户。培训教员应一直在场向学员提供指导或纠错。现场采访完之后，学员们要在教员指导下进行讨论。培训方案的执行结果应导致调查管理人决定哪些学员还需要继续培训，以及他们中间是否有人根本不适合做这项工作。

2.2.4.6. 现场监督

59. 培训工作被普遍认为是有效圆满完成现场工作的先决条件。可是，缺乏适当监督的培训也不会产生预期的成果。现场工作的成功有赖于高级工作人员持续不断、专注而有效地发挥监督作用，他们比较有经验，比访调员资格老。监督员要经过调查工作的全面培训。不要小看监督员的工作，他们是数据收集组织和访调员之间的重要联系纽带。一般认为监督员的职责是组织访调员的工作，确定现场分派任务和地点；由他(她)负责检查已完成的工作，并维持访调员对调查方案的高水准承诺。我们建议，如有可能应相对提高监管人员与访

调员的比率。据认为, 1名监督员对4-5名访调员这个比例对大多数住户调查都比较理想。不过, 这只不过是一项准则而已。

2.2.4.7. 应对不答复情况的后续行动

60. 在大多数调查中肯定会有不答复的情况发生(请参见关于非抽样误差问题的第八章)。一些受访者也许会拒绝跟访调员配合; 有些情况下, 问卷中的某些问题可能被跳过去了。当监督员接获一份关于不答复单位的报告时, 他(她)必须与该样本单位取得联系, 尽力从中获得所需要的信息, 因为他(她)作为监督员资格较老, 比较有经验。任何住户调查都有一项争取实现尽可能高反应率的目标, 因此建议对起初不答复者进行再抽样来收集信息。在这种情况下, 调查工作要重新定位到二段样本, 最好由监督员来做访调员。

2.2.4.8. 降低不答复率

61. 在设计和执行住户调查的过程中, 有必要发展旨在最大限度提高反应率的良好调查程序。我们强调要有减少拒绝答复的办法, 比如在受访者方便时安排返回再访。另外也应该向不愿答复者仔细讲解调查的目的和用途, 以争取他们的配合。承诺保密也有助于减少受访者的顾虑, 他们也许会担心, 自己的回答有可能被用于调查规定以外的目的。

62. 如果家中无人, 应在一天之内反复回访。建议尝试多达四次的回访。

63. 避免找不到选定的抽样单位, 这个问题也很重要, 因为这可能是导致不答复的一个重要因素。可以通过使用新近确定的抽样标架来解决这个问题(见第四章中的详细讨论)。

参考资料和补充读物

Kiregyera, B. (1999年)。《抽样调查: 特别关于非洲》。坎帕拉: PHIDAM企业。

加拿大统计局(2003年)。《调查方法与实践》。渥太华。

联合国(1982年)。《国家住户调查能力方案: 住户调查中的非抽样误差: 来源、评估和控制》(初版), DP/UN/INT-81-041/2。纽约: 联合国发展技术合作部和统计办公室。

_____ (1984年)。《住户调查手册》(修订版), 《方法研究》, 第31号, 出售品编号: E. 83. XVII. 13。

_____ (1998年)。《人口与住房普查的原则和建议》, 首次修订。统计文件, 文件号: 67/Rev. 1, 出售品编号: E. 98. XVII. 8。

_____ (2002年)。《关于在人口普查中收集经济特征数据的技术报告》。纽约和日内瓦: 文件号: ST/ESA/STAT/119(仅英语文本)。社会经济事务部统计司和国际劳工局统计司。

Zanutto, E. 和A. Zaslavsky (2002年)。使用行政案卷改进小区域估计: 美国十年普查的一个实例。《官方统计学刊》(瑞典统计局), 第18卷, 第4期, 第559-576页。

第三章 抽样策略

3.1. 引言

1. 前面关于调查计划的第二章一般概述了住户调查工作的各个不同阶段，而本章则是专注于抽样方面的几章——也是本手册主要聚焦点——的第一章。本章通过对比简要讨论概率抽样和非概率抽样，并且论证为什么在住户调查中一定要采用前一种方法。把相当大的注意力放在了样本规模问题上——即决定样本大小的诸多参数和如何计算样本规模。提供了在住户调查中实现抽样效率的一些技巧，其中包括分层抽样、整群抽样和分段抽样，尤其侧重于二级抽样设计方法(见表3.1和附件一中关于这些概念的定义和说明)。提供了各种抽样选项，并且详细描述了在许多国家广为采用的两种主要样本设计方法。另外还讨论了两个专题：(a) 两阶段抽样以覆盖“稀疏”人口；和(b) 通过抽样估计人口变化与趋势。本章结尾概括了有关建议。

3.1.1. 概览

2. 不论在发展中国家还是在发达国家，几乎所有住户调查的样本设计都因其多阶段、分层和聚类的特征而显得十分复杂。此外，国家级的住户抽样调查通常在范围上都是为了一般性的目的，涵盖政府所关心的多个主题，这就更增加了样本设计的复杂性。因此本手册集中处理多级抽样的策略问题。

3. 为了产生预期的成果，对于住户调查而言好的样本设计——亦即交响乐似的和声安排——必须和谐地组合无数的要素。要分阶段选择样本以准确敲定采访地点和高效率地选择住户。分层设计要确保实际选择的样本在地域分区和人口分组上布局适当。抽样计划须采用类集，它们通常是选择住户的地域限定单位，以便将费用保持在可管理的水平上。同时要避免被过分聚类，因为被过分聚类的计划对可靠性有破坏作用(见第3.3.5节关于聚类效应的讨论)。样本容量要考虑到竞争的需要，以便使调查成本和数据精度达到最佳平衡。样本容量还必须满足用户对各种“域”——即分组人口或分区——数据的迫切需求。样本设计必须通过两种方式最大限度地追求精度：第一，所使用(或构建)的抽样标架必须尽可能完整、正确和时新；第二，所使用的样本选择方法要尽可能减少实施人员有时无意识造成的偏差。样本设计还应该可以自我评估，换言之，根据这种设计可以估计抽样误差，以便指导使用者测定关键成果的可靠性。之所以会产生抽样误差，是因为人口特征估计所依据的数据只是出自部分人口而不是总人口。

4. 调查的主要目的是为了能够根据某个随机样本对目标群体进行推论。通常调查者/研究者进行这种推论的目的在于估计某些未知的人口特征。估计的普通人口特征/参数有总数、平均数、比例和方差。譬如讲, 如果 $Y_1, Y_2, Y_3, \dots, Y_N$ 是人口变量 y 的数值, 那么

$$\text{人口平均数即为 } \bar{Y} = \frac{1}{N} \sum Y_i \quad (3.1)$$

$$\text{人口方差即为 } \sigma^2 = \frac{1}{N} (\sum Y_i^2 - N\bar{Y}^2) \quad (3.2)$$

在大多数情况下, 样本估计数被用于估计人口参数。例如, 对于通过置换选择的简单随机样本来说, 一个规模为 n 的样本的平均数和方差由下列方程式给出:

$$\bar{y} = \frac{1}{n} \sum y_i \quad (3.3)$$

$$s^2 = \frac{1}{n-1} (\sum y_i^2 - n\bar{y}^2) \quad (3.4)$$

式中, $y_1, y_2, y_3, \dots, y_n$ 表示样本中 n 单位的变量 y 的值。在抽样调查中, 研究者计算选定随机变量的方差, 用以测定估计量中抽样误差的量(见表3.1中的抽样误差定义; 关于抽样误差的更多讨论, 见第七章和附件一)。影响抽样方差量级的因素包括所研究变量的不同成分、样本规模和样本设计(本章各节和第七章分别讨论了这些方面; 附件一提供了调查抽样基本原则)。

5. 第三章和第四章详细讨论了设计住户调查的适当样本所涉及的每一种特征。总的来说重点放在国家级的调查方面, 不过, 其中描述的所有方法也适用于国内各分区的大型调查, 比如限于一个或多个地区、省、地方管区和市的住户调查。鉴于抽样标架对于实现良好抽样做法至关重要, 第四章全部用来讨论这个问题。

3.1.2. 抽样词汇及相关术语

6. 我们以本章和其余各章所用术语的词汇表来开头(见表3.1)。该词汇表无意提供各种抽样术语(其中有些是数学术语)的正式定义。它只是结合本手册来描述相关术语的用途, 当然, 重点放在住户调查的应用方面。

3.1.3. 符号

7. 本手册的本章以及后面各章均使用标准符号(见表3.2)。一般而言, 大写字母表示总体值, 小写字母表示样本观测值。例如, \bar{Y} 表示总体值。而 \wedge 则一般用于表示样本值。从上面可以明显看出, N 代表人口规模, 而 n 代表样本规模。这里有必要指出, 人口参数既可以用英文大写字母表示, 也可以用希腊字母表示。例如, \bar{Y} 和 σ 分别表示人口平均值和标准

表3.1
抽样词汇及相关术语

术 语	用 途
精度(有效性)	见“非抽样误差”
分区抽样	选择构成抽样标架的地域单位(可以包括按地图标绘的行政分区的细分区段)
查核	一种“覆盖”地理区域以确定住所和/或住户位置的方法,通常用于更新抽样标架的工作
聚类;被聚类	指样本单位(个人或住户)具有类似特征的倾向性
整群抽样	在倒数第二阶段包含某种地理界定单位——比如普查计数区(EA)——的抽样
类集规模	在类集中抽样单位(个人或住户)的(平均)数量
紧凑类集	由地域邻近住户构成的样本类集
复杂样本设计	是指在住户调查样本中使用多阶段、类集法和分层法,与简单概率随机抽样的含义恰好相反
置信度	形容统计置信程度,借此可获得调查估计数的精度或容许误差,一般以95%为标准
设计效果(<i>deff</i>)	在样本规模相同的情况下,复杂样本设计方差对简单随机样本方差之比;有时叫做“类集效应”,尽管设计效果既包括分层效应也包括类集效应
域	要求单独提供估计数的地域单位
虚拟选择阶段	为简化人工鉴别最终定位样本类集分区的工作任务而虚设的样本选择阶段
等概率抽样	用同等的概率进行抽样
估计量	对特定样本设计而言,估计量是根据样本数据估计人口参数的方法,例如,样本的算术平均就是一个估计量
隐含分层	通过抽样标架的地域分类并结合概率与规模成正比的系统抽样进行分层的手段
同类相关	用同类相关系数估量类集中各种要素的同质相似度
名单抽样	从构成抽样标架的单位名单中选择
主样本	一种“超级”样本,用以进行多阶段调查和/或同一项调查的多轮次调查(一般每轮次为期10年)
规模量度(MOS)	在多级抽样中,特定阶段每个单位的规模(比如人数)的计数或估计数
非紧凑型类集	由地域分散的住户构成的样本类集
非概率抽样法	见第3.2.2节关于此类方法的实例解说,其中包括:比例抽样;判断性或有目的的抽样;方便抽样;随机走动抽样
非抽样误差	由于设计和实施中的误差造成的调查估计偏差;系指估计数的精确性或有效性,与之相对的是它的可靠性或精度
初级抽样单位(PSU)	在第一阶段抽样过程中选定的限定地域的行政单位
概率抽样	一种选样方法,采用此种方法每个人口单位(个人、住户等)都有已知的非零机会被列入样本
快速清点	是指在粗略清点住所的时候进行更新以提供当前的样本规模;另见查核
相对标准误差(不均匀系数)	用调查估计数的百分比表示的标准误差,换言之,亦即估计数除以标准误差

表3.1
抽样词汇及相关术语(续)

术 语	用 途
可靠性(精度, 误差限度)	是指与特定调查估计数有关的抽样误差程度
抽样标架	据以实际选样的成套材料, 比如一份名单或一套分区, 亦即一套人口单位汇编
抽样率	样本规模对人口单位总数之比
样本规模	所选住户数量或人数
抽样误差(标准误差)	由于调查的是样本而不是整个人口这一事实所造成的调查估计数中的随机误差; 抽样方差的平方根
分段抽样; 亦称“复式抽样法”或“后分层抽样法”	(一般)在两个时段选择样本, 通常第二阶段样本是第一阶段样本的二段样本; 不要与趋势抽样法(见下文)相混淆
分阶段抽样	是指在连续若干阶段选择行政分区和住户/个人样本, 以确定进行调查的地理位置
抽样方差	标准误差或抽样误差的平方
与规模成正比的概率抽样(PPS)	用于选择第一(第二, ……)阶段的单位; 在各阶段每个单位都用与规模成正比的概率进行抽样
区段	在地图上标示的较大类集的细分部分
自我加权	一种样本设计, 其中所有情况都有相同的权数
SRS	简单随机抽样(在住户调查中很少使用)
分层抽样	在子群里安排抽样标架的一种方法, 这些子群内部是同类的, 外部是异类的, 以确保选样工作在重要人口亚群适当“铺开”
分段(分组或组块)	一种常用的现场作业方法: 把异常大的类集拆分, 以减轻登记工作负荷
系统抽样	采用随机起点、预设选样间隔的办法, 从清单中连续进行选择
目标人口	确定为调查覆盖人口的定义; 亦称“覆盖全域”
趋势抽样法	旨在估计从一个时段到另一时段发生的变化了的样本设计
权数	选择概率的倒数; 适用于原始数据的膨胀因子; 亦称设计权数

偏差。人口参数估计量在大写符号上 \bar{y} 方有符号, 例如 $Y_1, Y_2, \dots, Y_p, \dots, Y_N$ 和用小写字母标示的 $F_1, F_2, \dots, F_p, \dots, F_n$ 。

3.2 住户调查的概率抽样法与其他抽样法的比较

8. 虽然关于概率理论的讨论超出了本手册的范围, 但是有必要解释一下为什么概率方法在住户调查的抽样方面发挥了不可或缺的作用。本节提出了关于概率抽样的简要定义和说明, 并解释它何以如此重要。同时也概略提到其他一些不具备概率抽样条件的抽样方法, 诸如判断性或有意抽样、随机走动抽样、比例抽样和方便抽样等, 并且讨论了之所以不推荐使用这些方法进行住户调查的原因所在。

表3.2
选定用于总体值和样本特征的符号

特 征	代表符号		
	总 体	样 本	
		概念上方带有符号 “^”的估计数	使用小写 字母概念
单位	N	\hat{n}	n
观测值	$Y_1, Y_2, \dots, Y_p, \dots, Y_N$	$F_1, F_2, \dots, F_p, \dots, F_n$	$y_1, y_2, \dots, y_p, \dots, y_n$
平均值	\bar{Y}	$\hat{\bar{Y}}$	\bar{y}
比例	P	\hat{P}	p
参数和估计量	θ	$\hat{\theta}$	$\hat{\theta}$
y的方差	$\sigma^2(y)$	$\hat{\sigma}^2(y)$	$s^2(y)$
y的标准偏差	$\sigma(y)$	$\hat{\sigma}(y)$	$s(y)$
同类相关	δ	$\hat{\delta}$	$\hat{\delta}$
比率	R	\hat{R}	r
总和	$\sum_{i=1}^N$	$\sum_{i=1}^{\hat{n}}$	$\sum_{i=1}^n$

3.2.1. 概率抽样

9. 在住户调查背景下的概率抽样涉及到选择纳入调查的目标人口诸要素——地域单位、住户和个人——所需的各种手段。概率抽样需要：(a) 每个要素都有一个已知的被选中的数学机会；(b) 该机会大于零；以及(c) 该机会在数字上是可计算的。这里需要着重指出的是，每个要素被选中的机会不需要均等，而可以根据特定调查目标有所差异。

10. 正是因为概率抽样法所具有的数学性质，才有可能通过调查获得有科学依据的估计数。更重要的是，这是得出如下推论的基础，即：样本估计数代表了从中提取样本的总人口。可以根据从样本个案收集的数据来估计抽样误差——这是在调查中进行概率抽样的一个至关重要的副产品。而这些特征，非概率抽样方法一个也不具备。有鉴于此，我们极力主张在住户调查中一定要采用概率抽样法——即便调查成本高于那些不科学的非概率方法也在所不惜。

3.2.1.1. 分阶段概率抽样

11. 如上所述，必须在选样过程的每一阶段都采取概率抽样法，才能达到要求。譬如讲，选样的第一阶段一般涉及到选择限定地域的单位，比如村庄。最后阶段涉及到选择行将访谈的特定住户或个人。这两个阶段和任何中间阶段都必须使用概率方法进行适当的抽样。下面提供一个简化的说明范例。

范 例

假设在一个农村省份从100个村庄当中选定了一个由10个村子组成的简单随机样本(SRS)。进一步假设每个样本村都有一份完整的住户清单。不论列入各村清单的住户有多少,一律按五分之一的比例从清单中系统选择拟进行调查访谈的住户。这是一个概率样本设计,其中分两个阶段选样:第一阶段概率为10/100,第二阶段概率为1/5。为调查选择一个特定住户的总概率为1/50,亦即10/100乘以1/5。

12. 虽然上述样本不是特别有效,但其样本设计说明了两个阶段的样本如何使用概率抽样法。因此,可以通过在调查工作的数据分析阶段适当应用选择概率以无偏方式估计调查结果(见第六章关于调查加权的讨论)。

3.2.1.2. 概率计算

13. 上面范例也说明了如何达到概率抽样的另外两项要求。首先,该省的每个村都获得一个非零入选机会。对比之下,如果一个或多个村子出于任何考虑(比如安全关切)而被否决,这些村子的入选机会就是零,因而就违背了样本的或然性。上面范例中的住户也是用非零概率选择的。可是如果其中有些住户由于某种原因——比如不易接近——而被有目的地排除在外,他们的入选概率即为零,而样本的实施就变成了非概率设计。第3.2.1.3节讨论如何处理有关分区被排除在调查之外的情况。

14. 其次,选择村庄和住户的概率可能实际上是根据现有信息计算的。在选择村庄的情况下,样本规模(10)和总体规模(100)均为已知,而它们是界定概率的参数,10/100。对于住户来说,概率的计算略有不同,因为我们在调查之前不知道每个样本村选多少住户。相关的指示只不过要求我们每5户选1户。这样的话,如果A村总共有100户,B村有75户,那么我们会分别在这两个村子里选择20户和15户。而不论按总体规模还是按样本规模选择一个住户的概率依然各为1/5($20/100=1/5$,但是 $15/75$ 也等于1/5)。

15. 仍以上面的说明为例,第二阶段的选样概率作为完成调查后的交叉核对来计算。设 m_i 和 M_i 为已知数值,其中 m_i 和 M_i 分别是样本住户数和第 i 个村的住户总数,那么,概率将等于 m_i/M_i 。会有10个这种概率——每个样本村一个概率。可是如前所述,对特定设计而言这个比率总是1/5。因此,若仅为计算第二阶段概率之目的获取样本数和总住户数是多余的。但是为了质量管理的目的,获得这两种计数对于确保准确贯彻1/5的抽样率则是有益的。

3.2.1.3. 如果目标人口界定有误

16. 有时候,由于调查所要覆盖的目标人口的界定标准太宽泛而违反了概率抽样的必备条件。譬如讲,预期的目标人口也许是全国所有住户。而在设计和实施调查的时候往往有一些人口亚群——比如游牧住户、以船为家的人以及因为地形艰困而不易接近的整个地区的人口等——被故意排除在外了。在还有些情况下,目标人口涵盖特别限定的人口,比如结过婚的妇女或25岁以下的年轻人,由于各种原因可能会排除某些重要亚群。举例来说,一个意

在覆盖25岁以下年轻人的目标人口也许排除在军队服役的人或在押犯人，或者被收容在专门机构的人。

17. 凡是在调查所实际覆盖的目标人口与本来打算覆盖的人口有出入的时候，调查队就应注意更精确地重新界定目标人口。这样做的必要性不仅在于为用户澄清调查结果，而且在于满足概率抽样的必要条件。在前述25岁以下年轻人的那个实例中，目标人口应更确切地描述和定义为25岁以下未被专门机构收容的平民年轻人。否则，调查覆盖面就应扩大到包括被排除在外的亚群。

18. 可见，有必要非常小心地界定目标人口，以便使调查只涵盖那些将实际获得入选机会的成员。在有意识地将某些亚群排除在外的情况下，对构成调查范围的实际人口应用概率方法当然是至关重要的。另外，在发布调查成果的时候调查指导者有义务向用户清楚地说明调查包含了哪些区段的人口，而同时排除了哪些区段。

3.2.2. 非概率抽样法

19. 跟概率抽样不同的是，没有统计学理论来指导非概率样本的使用。它们只能通过主观评价来评估。因此，不适用概率技巧就意味着调查估计数可能会出现偏差。另外也不知道这些偏差的大小，其估计偏低或偏高的倾向性往往也不得而知。如前所述，使用概率抽样法可以估计抽样估计数的精确度，亦即其标准误差。这是为用户估量调查估计数的可靠性和建立调查估计数的置信区间所必要的。在有些情况下，比如当需要使调查人口分布符合其他控制因素的时候，采用概率抽样法也会出现估计数偏差(关于这个问题的进一步讨论，见第六章)。

20. 尽管非概率样本存在理论上的缺陷，但在各种环境状况下还要经常使用这些方法。实际工作者为此提出的理由一般基于成本和方便方面的考虑，甚至调查队认为“随机”样本或许不能恰当代表目标人口。下面我们结合住户调查，主要通过实例说明来简要讨论各种类型的非概率样本，并提出一些理由说明为什么不该使用此类方法。

3.2.2.1. 判断性抽样

21. 判断性抽样是依靠“专家”选择样本成分的一种方法。支持者声称，在运用随机技巧的情况下，判断性抽样可以避免选定一个“坏”样本或奇怪的样本，譬如讲，在一个样本里所有样本要素都不幸地属于西北地区。

范 例

适用于住户调查的判断性抽样可以举这样一个例子：在一个样本计划中一组专家有意识地选择一些地理区域当作第一阶段的选择要素，他们作决定的依据是看哪些地区在某种意义上或在某种背景下是典型的或者有代表性。

22. 此类抽样的主要难点在于确定构成一套有代表性地区的要素所带有的主观性。具有讽刺意味的是，这种选择也在很大程度上取决于专家本身的选择。对比之下，采用概率

抽样法, 首先需要使用设计团队所想要施加的任何规范对这些地区进行分层。请注意, 就连分层规范也可能带有主观性, 不过, 在应用较为客观的分层规范方面有专门准则(参见关于分层的第3.4节)。然后可以从每个层次选取一个概率样本(采用任何一种方法进行选择)。请注意, 分层可以大大减少像上面提到的那样选择奇数样本的可能性。这正是发明分层法的原因所在。有了这种分层样本, 每个区都会有已知的非零入选的机会, 而且这种机会无偏差, 不受主观意念的影响(甚至在层次本身是主观界定的情况下也不受影响)。另一方面, 判断性的抽样法既不包含确保每个区都有非零入选机会的机制, 也不保证计算最终入选的概率。

3.2.2.2. 随机走动抽样和比例抽样

23. 另一类广泛使用的非概率抽样, 是在住户调查的最后阶段进行的所谓随机走动程序。即便在各阶段之前就通过合法的概率方法选定了样本要素, 这种随机走动的抽样技巧也经常采用。下面的说明展示了随机走动抽样与比例抽样相结合。后者是另一种非概率技巧, 访调员借此获得作为访谈对象的某些类别的个人配额。

范 例

该方法的例证如下: 指示访调员在某个随机地点(比如一个村子)开始采访过程; 沿着既定的步行路线, 选择拟定采访的住户。这可能需要选择每个第 n 个住户, 或者沿着步行路线筛检每个住户以查明家中是否有特定目标人口(比如5岁以下儿童)。在后一种情况下, 每个合格的住户都将被访问调查, 直至达到预定的访谈比例。

24. 这种方法的合理性往往在于, 它可以避免在选定访谈住户之前先对样板区(村庄、类集和区段)的所有住户进行登记造册的一个既费资金又费时间的前期阶段。它的正当性还在于可以避免出现不答复的情况, 因为访调员会在不答复的住户以外坚持不懈地进行访谈, 直到有足够的答复住户达到了配额指标。另外, 它的支持者还声称, 只要沿着步行路线的起点是随机确定的, 这种方法就不会出现偏差。他们还断言, 通过用入选的住户数除以村子里的总住户数, 可以恰当计算出选样概率, 而后者要么是已知的, 要么可以非常近似地约计。

25. 从上面直接提出的条件来看, 理论上是可以获得概率样本的。可是在实际当中是否真的实现过还是令人怀疑的。这个办法通常是失败的, 这是因为: (a) 访调员的行为表现; (b) 对不答复住户(包括潜在的不答复住户)的处理方式。已有无数研究表明, 如果授予访调员以现场选样控制权的话, 就会导致出现样本偏差。譬如讲, 样本住户的平均规模(人数)通常小于住户总体规模。¹ 对于访调员来说, 回避一个可能因任何缘故被认为难以接触的住户是符合基本人性的。为此, 如果一个住户家有恶狗或门户森严、难以接近的话, 简单做法就是干脆让过这一家而进入没有此类问题的下一住户。

¹ 许多调查组织现在的标准做法是确保作为一项行政业务来指定入选样本的住户, 这样比较容易对选样工作实施监控。另外, 选择有关样本的人不应该是在选样前参加过编制住户清单的人, 或者说, 应该是不了解现场实际情况的人。

26. 用答复的住户替代不答复的住户，就会使样本偏向愿意合作、容易接触的住户。很明显，住户特征与其是否愿意和是否有机会参加调查有密切关系。采用比例抽样方法，难以接触或不愿意参加调查的人的代表性不足的可能性要高于概率样本中的此种情况。在后一种情况下，一般要求访调员对暂时见不到家中成员的住户多次回头再访。另外，在基于概率方法的调查中，通常访调员都经过培训，要求他们倍加努力，说服不愿参加调查的住户同意采访。

3.2.2.3. 方便抽样法

27. 方便抽样法因为实施简单，也广为使用。虽然方便取样在住户调查中不常使用，但可以列举许多使用例证，例如在一个有意选择的学校样本中进行学校青少年调查，调查对象很容易接近，知道他们愿意合作，就是说，抽样方便。另一个时下很流行的应用实例，就是由互联网站即时民意测验，人们只要登录该网站就被问及对各种话题的看法。这也许就是为什么此类样本天生带有明显偏差的原因所在，因此不适合用以进行关于总人口的推断。

3.3. 确定住户调查的样本规模

28. 本节用了很大篇幅来详尽讨论样本规模的问题，因为它对整个调查工作和成本十分重要。它的重要性不仅在于采访了多少住户，而且还在于提取了多少地区的初级抽样单位(PSUs)，雇佣了多少访调员，每个访调员的工作负荷有多大，等等。在确定样本规模方面必须要考虑的因素和参数有很多，但是它们主要围绕调查的度量目标反复出现。我们将结合所预期的关键估计数、目标人口、为涵盖必要的目标人口所须抽样的住户数量、预订达到的精确度和置信度水平、各种估算域(不论估量水平还是变化)、类集效应、不答复的容许度、以及可用预算等事项，来讨论有关确定样本规模的问题。显而易见，样本规模是整个样本设计的关键特征。

3.3.1. 调查估计数的量级

29. 在住户调查中，不论其是为一般目的还是为某个专题，比如健康或经济活动，调查所要产生的每个估计数(经常叫指标)都需要一个不同的样本规模，才能进行可靠的度量。样本大小取决于估计数的大小，亦即它在总人口中所占的比例。譬如讲，可靠地估计有安全饮用水的住户比例所需的样本规模就不同于估计当前不工作的成年人比例所需的样本规模。

30. 样本规模计算的表达式系基于概率声明，即实际人口参数包含在一个既定概率区间(置信度)。该区间的宽度(或精确度)取决于表3.2中提到的人口方差；取决于置信度；也取决于样本规模。一般来说，人口异质程度或期望的置信度越高，概率区间越宽。另一方面，该区间宽度随着样本规模加大而缩小。第七章第22段给出了置信区间的实例。下式代表了一个人口平均数 \hat{Y} 的一个置信区间，其中考虑到基于规模为 n 的无置换简单随机样本人口平均数的估计量 \bar{Y} 。

$$P \left[\hat{Y} - z_{1-\alpha} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2(y)}{n}} \leq \bar{Y} \leq \hat{Y} + z_{1-\alpha} \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2(y)}{n}} \right] = (1 - \alpha) 100\% \quad (3.5)$$

式中, $1-\alpha$ 是该区间的置信系数。另请注意, 在考虑到估算比率 p 的情况下, $s^2(y) = p(1-p)$ 。

31. 实际当中, 调查本身只能有一个样本规模。为了计算样本规模, 必须从行将在调查中测量的诸多估计数当中作一选择。譬如讲, 如果失业率是关键估计数, 那么样本规模的计算就应以失业率为基础。² 在有許多指标的情况下, 有时适用的一种惯例就是计算每个指标所需的样本规模, 然后使用能够产生最大样本的指标。一般来讲, 这个指标所代表的基本人口就其在总人口中所占比例而言是最小的“次级目标人口”。当然, 要考虑到预期的精度(见下文)。如果样本规模以这样一个估计数为基础的话, 那么其他关键估计数每个都要用同样的或更大的可靠度来计量。

32. 作为一个选项, 样本规模可以建基于相对较小的目标人口比例, 而不以某个特定指标为基础。在有多个互不相干主题上的一般目的住户调查中, 这也有可能成为一种最佳办法, 因为在那种情况下把样本规模建立在属于单一主题的某个指标的基础上也许不切实际, 或者不够谨慎。因此, 调查管理者可以决定将样本规模建基于可靠计量5%(或10%)的人口具有的某个特征的能力上, 确切的抉择取决于预算方面的考虑。

3.3.2. 目标人口

33. 样本规模也取决于预计被调查所涵盖的目标人口。跟相关指标的情况一样, 在住户调查中也往往有好几个目标人口。比如健康调查就可能将目标放在: (a) 住户, 以评估安全饮用水源和公共卫生设施状况; (b) 所有个人, 以估计慢性病和急性病状况; (c) 年龄在14-49岁的妇女, 以摸清生育健康指标; 以及(d) 5岁以下的儿童, 以便对其进行身高和体重方面的人体测量。

34. 因此, 样本规模计算必须考虑到每一个目标人口。如前所述, 住户调查往往有多个目标人口, 就调查的各项度量目标而言, 每个目标人口都同样重要。在这方面也可以在确定样本规模的时候把重点聚焦于最小的目标人口。举例来说, 如果5岁以下的儿童是一个重要目标群体的话, 那么样本规模就应以这个群体为基础。利用第32段所述概念, 调查管理团队可以决定计算样本规模, 以估计10%的5岁以下儿童所具有的某个特征。如此计算得出的样本规模可能会远远大于一个由所有人或所有住户构成的目标群体所需的样本规模。

² 为了计算样本规模而要求通过它的公式知道所要计量的估计数的近似值, 这多少有些荒谬。不过, 可以通过各种方法来“猜测”这个近似值, 比如使用出自普查或类似调查的数据、来自邻国的数据、试点调查获得的数据, 等等。

3.3.3. 精度和统计置信度

35. 上面建议，调查估计数、尤其是关键指标的估计数必须可靠。特别是样本规模的确定有赖于指标所要求的精度。调查估计数所需具备的精度或置信水平越高，就越需要较大的样本规模——而且是成数量级地增大。譬如讲，置信度翻一番，就可能需要有四倍的样本规模。调查管理者必须清楚地意识到过分严苛的精度要求对样本规模、从而对调查成本造成的影响。反过来，他们也必须当心不要选择过小的样本规模，以至于主要指标过分缺乏可信度，不适合用以进行信息灵通的分析或富有意义的计划工作。

36. 同样，随着期望的统计置信度的提高，样本规模增大，以保持既定的精度。几乎普遍把95%的置信度视为标准，并相应计算为达到这一标准所需要的样本规模(见前面第30段)。

37. 作为许多精心策划的调查所奉行的一个惯例，留意相关指标就是在95%的置信水平上使用10%的相对误差限度作为对所要估计的关键指标的精度要求，就是说，一个关键指标的标准误差不得大于估计数本身的5%，计算为 $(2 \cdot 0.05x)$ ，式中， x 是调查估计数)。譬如讲，如果劳动力中的个人比例估计为65%，那么它的标准误差不得大于3.25个百分点，亦即0.65乘以0.05。0.0325的两倍，即0.065，是在95%的置信水平上的相对误差限度。例如，在上面第30段我们有：

$$\sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2(y)}{n}} = 0.05 \hat{Y} \quad (3.6)$$

38. 因此，达到10%相对误差限度规范所需的样本规模就是在相对误差限度定为5%的情况下占到总数的四分之一。一般把20%的相对误差限度视为重要指标的最大容许误差(但是我们并不推荐它)。这是因为对于大多数分析或政策需求所要达到的富有意义的成果来说，容许较大误差的估计数周围的置信区间太过宽泛了。一般而言，对于主要指标——预算允许范围——来说，我们建议5%-10%的相对误差。

3.3.4. 分析群体：域

39. 对样本规模有很大影响的另一个重要因素就是域的数量。一般把“域”定义为需要有同等可靠数据的分析子群。样本规模的增大系数大体相当于所想要的“域”数。³不过，如果每个域都显示了差不多的可变性的话，情况就是这样(见脚注3的进一步澄清)。这是因为特定精度水平的样本规模不取决于人口本身的多少，除非它占总人口较大的百分比——譬如讲5%或以上(而在住户调查中此种情况很罕见)。因此，单独一个省份所需的样本规模(如果调查仅限于一省范围内)，就会相当于整个国家所需的样本规模。这是一个时常被调查工作者误解的极端重要的要点，他们错误地以为，人口越多，肯定样本规模就越大。

³ 每当需要各个域都达到同等可靠程度的时候，情况大致如此。

40. 可见, 如果只需要获得国家一级的数据, 就只有一个域, 而如此计算的样本规模适用于全国的样本。可是如果决定城乡地区分开进行调查但预期获得同样可靠的结果的话, 那就必须单独计算每个域的样本规模, 才能产生可靠的结果。一般而言, 每个相关域的样本规模的计算方法是: 如果有 D_1, D_2, \dots, D_k 个域, 就需要有 n_1, n_2, \dots, n_k 这么大的样本规模, 这将取决于每个域内相关特征的可变性以及既定的置信度和精度。因此, 调查的总样本规模将是 $n = n_1 + n_2 + \dots + n_k$ 。

3.3.4.1. 为获取域估计数的过度抽样

41. 各种域的同等可靠性要求的一个重要含义, 就是必须使用不成比例的抽样法。于是, 当抽样分布比例不是50-50的时候, 譬如像城市域和农村域可能发生的情况那样, 故意在比如说城市域过度抽样, 对大多数国家实现同等可靠性来说很可能是必要的。不过要强调一点, 在一个全国调查中是否需要在在一个研究域过度抽样主要取决于在特定置信水平上获取估计数的需求。

42. 有必要提请注意故意过度提取子样本可能带来的两个后果——不论对域还是对层次而言。首先是因此而需要使用补偿调查权数才能产生国家级的估计数。第二个、也是更重要的后果是, 国家级的估计数可能不如在子群体之间按比例分配抽样的估计数那样可靠。

3.3.4.2. 域的选择

43. 地理分区当然很重要, 而且总是有压力要求为了估计的目的把它们当作域来对待。比如在进行全国调查的时候, 成员用户往往不但想要获取每个大区的数据, 而且还要获取每个省的数据。很明显, 必须仔细考虑域的数量并慎重选择构成这些域的估算群体。一个似乎有道理的对策是确定哪些估算群体尽管很重要但是不需要在调查估量方面具有同等的可靠性。倒是应该在分析中把各种估算群体当作跟域相对的主要制表类别。这样, 每个估算群体的样本规模就会比将其当作域对待的样本规模小许多; 结果, 它们的可靠程度也会有所降低。然而值得注意的是, 在特定域过度抽样的动机也许是需要在该域获取具有独立于国家级标准之外的特定置信度和精度的估计数。

范 例

下面的范例显示, 如果把城乡当作制表群体, 而不是当作域对待的话, 应如何抽样及其对可靠性的影响。设人口分布比例为农村占60%, 城市占40%。如果为了达到特定精度要求所计算的样本规模确定为8 000住户, 那么, 如果城市域和农村域分开抽样的话, 就得提取16 000个样本——两个域各8 000户。而若将其作为制表群体处理的话, 8 000个住户的国家级样本按城乡比例抽样, 就会分别选取4 800个住户和3 200个住户。另外根据8 000住户的样本再设某个10%特征的标准误差为0.7个百分点。这是适用于国家级估计数(或者, 如果城乡两域各抽样8 000住户的话, 则分别是城市和农村的估计数)的标准误差。对于一个按比例分别从城乡提取的8 000住户国家级样本来说, 相应的农村地区标准误差约为0.9个百分点, 这是计算得出的样本规模与国家估计数标准误差之比的平方根积数, 或者

$$\left(\sqrt{\frac{8\,000}{4\,800}}\right)^*0.7)$$

对于城市地区来说，标准误差约为1.1个百分点，或者

$$\left(\sqrt{\frac{4\,800}{3\,200}}\right)^*0.7)$$

评价该效应的另一种方法就是从下述事实来看：所有农村估计数的标准误差都比国家估计数的标准误差高约29%

$$\left(\sqrt{\frac{8\,000}{4\,800}}\right)$$

对于城市地区来说，其标准误差大约比国家级的标准误差高58%

$$\left(\sqrt{\frac{8\,000}{3\,200}}\right)。$$

44. 请注意，不管国家一级的标准误差大小，上面范例的最后一句话都适用。换言之，它适用于在调查中被纳入表格的所有估计数。因此对于可被视为域的各种子群来说，有可能在抽样之前分析对可靠性的影响。调查队可以用这种方法获得信息，以便于决定是否应该把潜在的域当作制表群体来处理。前面暗示过，这就意味着将使用按比例分配样本的方法而不是均等分配样本的方法。举例来说，如果一个城市人口仅占20%的国家打算进行全国调查的话，那么它的城市地区的样本规模就只相当于总样本规模的20%。这样一来，城市估计数的抽样误差就会相当于农村估计数抽样误差的两倍(0.8n/0.2n的平方根)，而比国家估计数的误差高2.25倍(n/0.2n的平方根)。在这种情况下，调查管理者或可决定有必要对城市地区进行过度抽样，⁴ 以便有效创建单独的城市域和农村域。

45. 同样，可以对标准误差和域与制表群体之间的关系进行分析，以便围绕是否适用地区或其他国内分区的地理单位作为“域”，以及如果适用的话需要多少个域的问题指导决策过程。在域所需的样本规模相等的情况下，使用10个地区，所需样本规模就会10倍于国家样本规模，但是如果只有5个地区可以被确认为符合政策需要的话，这个样本规模可以削减一半。同样，如果把地区当作制表群体对待的话，国家级的样本就会在有关地区之间按比例分配。在这种情况下，如果有10个地区，平均地区标准误差大约比国家估计数的标准误差高3.2倍；但是如果只有5个地区，则前者只相当于后者的两倍。

3.3.5. 聚类效应

46. 本节讨论的是关于样本规模的决定如何受到影响(第3.5节提供了关于整群抽样的详细讨论)。住户调查样本的被聚类程度对估计数的可靠性或精度有影响，因而也影响到样本规模。住户调查中的类集效应产生于：(a) 倒数第二个抽样单位，一般叫做“类集”，

⁴ 譬如讲，如果预料的(任何)主要城市指标的相对标准误差大于7.5%的话，就会做出这种决定(本手册建议95%置信度的最大容许误差为15%)。

可能是一些村庄,也可能是一批城市街区;(b) 样本住户;(c) 类集的规模和/或可变性;以及(d) 在选定的类集范围内的住户抽样方法。可以通过设计效果(或简称 d_{eff})用数字估量聚类以及分层效应,它可以表示分层类集样本跟一个同样大小的简单随机样本相比,前者的抽样方差(即标准误差的平方)比后者大多少。分层一般会减小抽样方差,但减幅很小。而对比之下,聚类则可以大幅度增加方差。因此,设计效果主要表明调查样本中的聚类作用有多大。

47. 有效的样本设计要求使用类集来控制成本,但是也要把设计效果保持在尽可能低的水平上,以便使调查成果既可用又可靠。不幸的是,在进行调查以前设计效果是不知道的,而只有到后来根据数据本身才能估计出来。在以往进行过调查或别的国家进行过类似调查的情况下,或可在计算公式中利用这些调查所获取的 d_{eff} 值来估计样本规模。

48. 为了把设计效果保持在尽可能低的水平上,样本设计应遵循下述一般原则(另见本章末准则概要):

- (a) 尽可能多地使用类集;
- (b) 就户数而言使用尽可能小的类集规模;
- (c) 使用不变的而不用可变的类集规模;
- (d) 在最后阶段选择一个住户系统样本;要选择地域上分散的而不要地域上接近的住户。

49. 因此,对于一个12 000户的样本来说,最好选择600个类集,每集20个住户;而不是选择400个类集,每集30个住户。在前者情况下,抽样设计效果要低得多。另外,如果从一个类集的所有住户当中有系统地选择住户,而不是在地理上接近的分区段中进行选择,即可降低设计效果。若遵循这些经验法则,就可能适当降低设计效果。

3.3.6. 为预计的不答复率调整样本规模

50. 调查中的一般做法是因不答复率相对增加样本规模。这种做法可以确保调查中的实际采访数量大体上接近目标样本规模。

51. 调查中的不答复程度因国家和调查类别不同而有很大差别。在下述计算实例中,我们设不答复率为10%。当然,各国应使用能更准确反映其最近全国调查经验的数字。

3.3.7. 主样本的样本规模

52. 第四章详细讨论了主样本的问题。本节集中讨论主样本计划的样本规模问题。在拥有较大规模综合性连续调查方案的国家,主样本是一种由初级抽样单位构成的大型样本。这种大型样本意在提供足够的“库存”样本案例,用以在几年的间隔期内支持多项调查,而无需重复采访相同的受访者。

53. 由于主样本可供多项调查和多种课题所用,所以当然要涉及无数目标人口和关键估计数。就此而言,大多数国家确定的样本规模基于两方面考虑。首先受预算制约,这是不

言而喻的。其次是个别调查的预期样本规模，这些样本规模可能要在主样本所涵盖的整个间隔期内使用，而人口普查的间隔期往往长达10年。似乎可靠的主样本的样本规模非常大，可达50 000住户以上。利用整个住户数据库的计划都是精心制订的。

范 例

设A国主样本由50 000住户构成。该主样本意在用于业已计划的三项调查以及尚未计划的另外两项可能的调查。其中一项是住户收入和支出调查，拟在10年内反复进行三次——分别在第1、第5和第8年实施。按照此项调查设计，每三年调查8 000住户。可是在第5年，将用一个4 000户的取代样本来置换第1年采访的8 000住户的一半。同样，第8年将用4 000新住户置换第1年采访的其余4 000个住户。这样，收入和支出调查就将总共使用16 000个住户。计划的第二项调查是健康调查，预计大约使用10 000住户；而关于劳动力参与率的第三项调查将使用大约12 000住户。这三项调查总共保留了38 000个住户。于是，仍有12 000个住户留待其他调查在必要时使用。

3.3.8. 估计变化或水平

54. 在定期重复的调查中，一个关键计量目标就是估计调查之间的变化。就统计而言，第一时间获取的调查估计数提供特定指标的水平，而该估计数水平与第二时间获取的估计数水平之间的差异是所估计的变化。为了得出可靠的结论，在估计变化的时候一般需要一个较大的样本规模，它要比只估计指标水平所需的样本规模大得多。在估量较小变化的时候尤为如此。不过，在估计变化方面也有一些抽样技巧旨在削减样本规模(从而降低成本)(见第3.9.2节)。

3.3.9. 调查预算

55. 在住户调查中确定样本规模的时候不可忽视调查预算的问题，这大概是不言自明的。虽然预算在样本规模的数学计算中并非惹人注目的参数，但在现实当中它确实占居突出的地位。

56. 统计学家考虑到本章所讨论的每一种参数，是他们最初计算了样本规模。可是往往有这样的情况：事实证明样本规模超过了预算所能支持的限度。当出现这种情况的时候，调查队必须要么寻求增加资金，要么修订其度量目标——降低精度要求，或者减少域的数量。

57. 抽样技术人员有责任帮助指导有关成本对精度问题的讨论。每当因为预算方面的考虑而不得不缩小样本规模的时候，抽样技术人员应当解释限制域的数量所产生的抵消作用(对用户的用途小了)，或降低精度要求带来的负面影响(减少了关键指标的可靠性)。应参考上面给出的有关精度和域的实例展开讨论。类集数量也是决定调查成本和调查精度的关键因素之一，抽样人员在指导调查队的时候必须仔细权衡这一事实(第3.5.5节进一步提供了这方面的讨论)。

3.3.10. 样本规模计算

58. 我们在本节提供样本规模的计算公式，其中考虑到前面讨论的各种参数。因为我们集中于住户调查方面，所以按照所须选取的住户数来计算样本规模。同时提供了相关图解。

59. 一般情况下，当包含比率 p 的时候，样本规模 n_h 的估计公式⁵如下：

$$n_h = (z^2) (1-r) (f) (k) / (r) (p) (\bar{n}) (e^2) \quad (3.7)$$

式中， n_h 是需要计算的参数，并且是用所应选取的住户数量表示的样本规模； z 是界定预期置信水平的统计量； r 是行将通过调查计量的某个关键指标的估计量； f 是样本设计效果 $deff$ ，设定为2.0(缺省值)； k 是计算预料不答复率所需要的一个倍数； p 是参数 r 所依据的目标人口占总人口的比率； \bar{n} 是平均住户规模(即每户的人数)； e 是必须达到的误差限度。某些参数的推荐值如下。

60. 在95%的置信水平上拟使用的统计量 z 应为1.96(与之相对的是，譬如讲，90%置信水平的1.645)。一般把前者视为评估住户调查误差限度所期望的指定置信度的标准统计量。样本设计效果的缺省值一般定为2.0，除非根据以往调查或相关调查产生的辅助经验数据提出了不同的值。所选择的不答复率的倍数， k ，应当体现本国在不答复率方面的经验——在发展中国家一般低于10%。因此，1.1的 k 值是个保守选择。参数 p 通常可以根据最近的普查结果计算得出。在大多数发展中国家，参数 \bar{n} 通常约为6.0，但是应该用最近的普查所提供的确切数值。至于误差限度， e ，建议精度水平定为 r 的10%，于是 $e=0.10r$ 。采用不太严格的误差限度， $e=0.15r$ ，即可减小样本规模，但是调查结果的可靠性当然也就小多了。置换一些选定的值，就有：

$$n_h = (3.84) (1-r) (1.2) (1.1) / (r) (p) (6) (0.01) \quad (3.8)$$

方程式(3.2)进一步简化为：

$$n_h = (84.5) (1-r) / (r) (p) \quad (3.9)$$

61. 每当所有建议的参数缺省值都被用以替代出自本国经验的较精确数值的时候，即可使用这种简化方程式。

范 例

在B国，决定把失业率定为需要计量的主要调查指标，据认为该指标为民用劳动力的10%。“民用劳动力”被定义为“14岁或以上的人口”，约占全国总人口的65%。在这种情况下， $r=0.1$ ，而 $p=0.65$ 。假定我们想要估计在95%置信水平上相对误差限度为10%

⁵ 样本规模公式也可能包含一个因子，即所谓有限倍数，当计算出来的样本规模证明为占了总体规模的很大比例的时候，必须考虑到这个倍数。不过，在本手册所考虑的这类大规模住户调查中很少包含此种状况。因此设想有限倍数的值为1.0，所以在公式3.5中被忽略了。

的失业率；那么，按照上面的建议， $e=0.10r$ （亦即0.01的标准误差）。另外，预料不答复率、设计效果及平均住户规模等数值有如我们的建议。接下来，我们可以使用公式(3.9)计算得出1170个住户 $[(84.5*0.9)/(0.1*0.65)]$ 。这是个相当小的样本规模，主要是因为基本人口占了总人口那么大的比例，譬如讲65%。忆及所计算的样本规模是单个域的——在本案例中是全国一级的。如果度量目标包括获取城乡同等可靠的数据，那么样本规模就得翻一番，并假定公式(3.8)和(3.9)的参数同样适用于城市和农村。鉴于城乡之间有差异（比如，城市住户的住户规模可能有别于农村住户；预料城市和农村的不答复率也不相同），所以如果城乡分开计算样本规模的话，就应使用更精确的数值。其结果当然是不同的。

62. 下面的范例设计较小的基本人口——5岁以下的儿童。

范 例

在C国，将5岁以下儿童死亡率确定为调查指标，据认为该指标约为5个百分点。在本案例中， $r=0.05$ ，而 p 估计约为0.15，或 $0.03*5$ 。这里，我们也想在相对误差限度为10%的情况下来估计这方面的死亡率，于是， $e=0.10r$ （或者说标准误差为0.005）。预料不答复率、设计效果及平均住户规模也是有如我们所建议的那样。用公式(3.9)计算出将近10704住户， $(84.5*0.95)/(0.05*0.15)$ ，这个样本规模要比前例的样本规模大多了。究其原因，也主要与基本人口（即5岁以下儿童）仅占总人口15%的规模有关。同样，由于估计的参数值 r 很小，加之 p 值也很小，所以迫使样本规模变得很大。

63. 最后这个范例是一个将总人口确定为目标人口的案例。在这种情况下， $p=1$ ，可以忽略不计；但如果采用推荐参数值的话，公式(3.8)和(3.9)仍可使用。

范 例

在D国，把主要调查指标确定为在上一周内患过急性病的人占总人口的比例。据认为这个比例为5%–10%，在本案例中使用了较小的值，因为它会导致较大的样本规模（一种保守的选择）。在本案例中， $r=0.05$ ，而 p 当然是1.0。这里，我们也想在相对误差限度为10%的情况下来估计这种急症率： $e=0.10r^6$ （或者说标准误差为0.005）。预料不答复率、设计效果及平均住户规模也是有如我们所建议的那样。用公式(3.9)计算出1600多一点儿的住户， $(84.5*0.95)/(0.05)$ 。

64. 如前所述，可以通过计算若干关键指标的样本规模来最终确定调查样本规模并且将此决定建立在能带来最大样本规模的指标基础之上。此外，在作最后决定之前还必须考虑到调查域的数量和调查预算。

65. 对于那些不适合上面讨论的一种或多种假定的国家来说，可以很容易地在公式(3.7)中进行简单替换来获取较准确的样本规模数字。譬如讲，平均住户规模可以大于或小于6.0；不答复率可以预料为5%而不是10%；而特定国家的 p 值一般可以通过使用普查数据更精确地计算。

⁶ 由于 r 适用于本案例的整个人口，所以它等于 p ，故 e 等于 $0.10p$ 。

66. 但是建议不要改变统计量 z 值1.96, 因为这是通用的标准值。从实际考虑, 设计效果, f , 也应保持在2.0, 除非如前所述从其他来源的最近调查数据提出了其他数值。另外还建议把 e 界定为 $0.10r$, 除非预算不能维持作为结果确定的样本规模。在那种情况下, 它或许可以增至 $0.12r$ 或 $0.15r$ 。可是, 如此扩大误差限度将会大幅提高抽样误差率。

3.4. 分层

67. 在设计住户调查的时候, 选样之前对行将调查的人口进行分层是一个通用方法。它的作用是根据有关总人口的已知辅助信息将整体人口划分成各种分组人口——层次。然后按照符合调查度量目标的方式从每个层次选取样本要素。

3.4.1. 分层与样本分配

68. 采用分层抽样法, 每个层次内的样本规模经由抽样技术员掌控, 而不是通过抽样过程随机确定。分层的人口可以从每层精确提取 n_s 个单位的样本, 这里的 n_s 是预期在第 s 层提取的样本单位数。对比之下, 不分层的样本会产生多少有别于 n_s 的第 s 个分组人口的样本规模。

范 例

假设某调查的样本设计分为两层——城市和农村。人口普查提供了把地理行政单位归入城市类或农村类的信息, 因而有可能按照这种标准对人口进行分层。鉴于该人口分布比例为农村占60%, 城市占40%, 现决定在每个层次选择一个成比例(相对于不成比例而言)的样本。如果样本规模是5 000住户的话, 按层单独选样就能保证其中3 000住户是农村的, 2 000住户是城市的。如果未经实现分层而随机选择样本的话, 样本中的住户分布就会不同于3 000-2 000分割, 尽管那可能是它的预期分布。如果运气不佳的话, 不分层的样本可能会产生一个——譬如讲——3 200农村户和2 800城市户的样本。

69. 由此可见, 分层的一个理由就是减少运气不佳、从一个据认为对分析很重要的分组人口选取不成比例地过多(或过少)样本单位的机会。进行分层的目的在于确保分组人口群体的适当代表性, 以避免选样工作出现偏差。不过有必要提请注意: 适当代表性并不意味着按比例抽样。在实施分层的许多情况下也可以把一层或多层当作估算域(前面讨论过)。在那种情况下, 或许有必要在受影响的层次选择同样大小的样本, 从而按层产生一个不成比例的样本。因此, 层次间的样本单位分配不论按比例还是不按比例, 都是分层样本的正当设计特征, 而选择结果取决于调查的度量目标。

70. 正如上面直接暗指的那样, 前面那句话表明分层亦可提供样本的隐含配置手段——一种比最优配置更简便、更实用的方法。⁷换言之, 采用按比例分层抽样法就无需事先计算分配给每一层次的样本案例数了。

⁷ 最优配置系指基于成本函数和层内方差(异质量度)的配置。本手册未予讨论, 因为这种方法在发展中国家很少使用。这也许是因为缺乏确切的调查工作成本数据的缘故。读者可在本章提供的许多参考资料中发现有关最优配置的信息。

范 例

假设样本设计的目标之一是确保将总样本规模按比例精确分配给全国10个省份。如果A省人口占全国总人口的12%，那么，若预期的类集规模恒久不变的话，就有12%的样本类集从该省选出。进而假设拟在全国选定的类集总数为400个。许多国家经常使用的一个方法是将48 (0.12×400) 个类集指定分配给A省。而采用适当分层法，这个程序就没必要了。分层法把每个省视为选择过程中的一个单独层次来对待。然后运用概率与规模成正比的系统抽样法(见表3.1)，按照单一的抽样间隔，就会自动得出A省的48个类集数。下文第3.4.3节将进一步讨论这种分层法及其在简化分配办法方面发挥的作用。

3.4.2. 分层的规则

71. 在进行人口分层的时候需遵守两条基本规则。无论如何起码必须遵守其中一条规则。另一条规则一般应遵守，但若不遵守也不会给样本设计带来多大损害。要求必须遵守的规则是，至少要有一个样本单位选自创建的每一层次。这些层次基本上是相互独立和互补包容的人口子集：一个人口的每个要素都必须在一层且仅在一层。因为这一特性，每一层都必须抽样以便整个人口都被抽样并且计算出人口平均数的无偏差估计数。由于理论上每一层都可以在样本设计中单独处理，所以层次的创建不一定需要使用客观标准；如果想要的话，亦可使用主观标准。准则应该是：构成一个层次的单位就研究变量而言应该是尽可能相似的，以便减少每一层的可变性。

72. 分层的第二条规则是，创建的每一层次都要尽可能有别于其他层次。因此，各层之间的异质性和各层内部的内同质性应构成指导分层的基本特征。因此不难理解，为何在住户调查中往往把城市地区和农村地区分开确定为两个层次。如上所述，城乡人口在许多方面各不相同(就业类别、收入来源和多少、平均住户规模、生育率，等等)，而在每个子群中的人口成员则是相似的。

73. 在确定创建多少层次的时候，异质特征是个有益的指南。就用以界定层次的标准而言，层次数不应多于可鉴别的分组人口数。譬如讲，如果一个国家划分为8个行政区，而其中两个地区就某项拟议调查的主题而言非常相似，那就可以通过建立7个层次来进行样本设计(合并两个相似地区)。举例说，如果建10个层次即可提供同样多的异质子群的话，即便使用20个层次也是一点儿额外收益得不到。

74. 有必要提请注意，就按比例选样而言，所选的样本至少和同样大小的简单随机样本一样精确。可见，分层法可在调查估计数的精度上或可靠性上获益，而且在各层最大限度呈现异质特征的情况下收益最大。正是分层抽样的这一特点决定了，即使是糟糕的分层⁸也不会减损调查估计数的可靠性。

⁸ 当创建了不必要的层次或者在误将某些人口要素归入不当层次的时候，就可能表现为糟糕的分层。

75. 另一要点涉及到抽样误差的估计方面。虽然每层只选一个单位就足以在理论上满足分层抽样的要求, 但是考虑到抽样结果将被用来计算调查估计数的抽样误差, 必须每层至少选两个单位。

76. 有时候分层过程中可能需要使用许多变量。在这种情况下我们的行动需要考虑下述准则: 各分层变量之间最好互不相干但与调查变量相关联; 没有必要追求完整和形成方格(较小的和不太重要的方格可以合并); 一般来说, 对许多变量进行粗略分割显然要比对一个变量的精细分割收益大。

3.4.3. 隐含分层法

77. 如前所述, 可用于创建层次的信息选择取决于调查的度量目标。对于大规模的综合住户调查来说, 有一种特别有用的方法, 叫做隐含分层法。其基本标准是地理范围, 这一事实就足以把样本推广到重要的人口子群, 诸如城市和农村、各行政地区、族群分组人口、社会经济群体, 等等。由于具有这种地理特征, 即使在调查主题集中在单一话题的情况下——不论是劳动力、住户经济活动、贫困度量、保健还是收入与支出问题——隐含分层法也会非常有用。这种技巧之所以受到高度推崇, 除了上述理由之外, 还因为它应用简便。

78. 为了正确应用起见, 隐含分层法要求在抽样第一阶段先进行系统选择。实施做法很简单, 首先按地理次序安排初级抽样单位建档。在许多国家, 这种地理次序一般排列为: 城市地区-省-区, 接下来是: 农村地区-省-区。下一步是从分类档案中有系统地选取初级抽样单位。系统选择既可采用同等概率抽样法, 又可采用概率与规模成正比的抽样法, 后者较为普遍。

79. 前面讲过, 隐含分层法的一个重要好处是无需建立明晰地理层次; 反过来也无需给这些层次分配样本, 尤其在采用按比例抽样法的情况下。另一个好处是上段讲到的简便法, 因为该方法只需档案分类和应用抽样间隔。在地域分类的第一层级也可能很容易应用不成比例的抽样法。譬如讲, 如果城乡构成第一层级, 那么, 一个直截了当的做法就是分别对城市和农村应用不同的抽样率。图3.1提供了一个隐含分层系统抽样图解。下面第3.6节进一步讨论了概率与规模成正比的抽样方法。

3.5. 整群抽样

80. “整群抽样”这个词语原本是指对某个群体的全体成员进行抽样的样本设计。群体本身被界定为类集。比如, 可在第一阶段选定学校样本, 在第二阶段选定教室样本。如果要调查每个教室的所有成员, 那么我们就需要有一个教室类集样本。在住户调查中选择一批城市街区, 对所选街区的所有住户进行调查采访——可以用这样的范例来说明整群抽样的本来含义。可是近年来, “整群抽样”已扩大到一般指调查中有一个倒数第二阶段抽样, 即选择(并界定)类集, 比如村庄、普查计数区或城市街区等。最后阶段抽样, 在每个选定的类集范围内对次级住户样本进行调查, 而不是调查所有住户。在本手册中一般使用的是该词语的后一种含义。

图3.1
隐含分层法的行政区安排

城 市		
省 01		
区 01		
		计数区 001
		计数区 002
		计数区 003
		计数区 004
区 02		
		计数区 005
		计数区 006
		计数区 007
省 02		
区 01		
		计数区 008
		计数区 009
区 02		
		计数区 010
		计数区 011
		计数区 012
省 03, 等等		
农 村		
省 01		
区 01		
		计数区 101
		计数区 102
		计数区 103
		计数区 104
区 02		
		计数区 105
		计数区 106
		计数区 107
省 02		
区 01		
		计数区 108
		计数区 109
		计数区 110
		计数区 111
区 02		
		计数区 112
		计数区 113
		计数区 114
省 03, 等等		

81. 在住户调查中, 样本设计总是需要利用一些整群抽样形式, 以确保节制调查成本。如前所述, 举例来说, 在50个地方调查1 000个住户(每个类集20户)要比调查从整个人口中随机选取的1 000个住户便宜多了。遗憾的是, 样本的聚类会降低它的可靠性, 因为生活在同一类集的人们有可能趋于同质性, 或多少具有类似特性。必须在样本设计过程中通过补偿性地增加样本规模对这种所谓“聚类效应”给予补偿。

3.5.1. 整群抽样的特性

82. 整群抽样和分层抽样在两个方面大不相同。⁹ 对于后者, 所有层次都在样本中有代表, 因为每一层都选出了单位样本。在整群抽样中, 对类集本身进行了选择, 因此, 样本中的类集代表了不在样本内的类集。分层抽样和整群抽样的上述显著区别导致了它们之间的第二方面差别。如前所述, 就所要计量的调查变量而言, 理想的是所创建的各层次应该是层内同质、层外异质为好。对类集来说恰好相反: 就样本精度而言, 类集内部尽可能地异质是比较有利的。

83. 在住户调查中, 基本上聚类总是被界定为地域单位, 比如村庄或部分村庄, 这就意味着——很遗憾——在类集内部一般达不到高度异质。的确, 就诸如就业类别(比如务农)、收入水平之类变量而言, 按地域界定的类集更有可能是内部同质而不是异质。因此, 对于特定变量而言, 类集的同质程度决定了据说一个样本“聚类”到何种程度。样本的聚类程度愈高, 其可靠性愈低。

3.5.2. 类集的设计效果

84. 一个样本的类集效应可以部分地用设计效果(*deff*)来计量。可是*deff*也体现了分层效果。样本设计团队有责任确保样本计划谋求在成本最小化和精度最大化之间实现最佳平衡。通过最大限度减少或尽可能控制设计效果可以实现上述目标。为了确定如何最大限度减少或控制设计效果*deff*的聚类成分, 看一看它的数学定义是有益的:

$$deff = 1 + \delta(\bar{n} - 1) \quad (3.10)$$

式中, δ 是同类(或群内)相关, 亦即一个类集中的两个单位与在人口中随机选出的两个单位相比较可能在数值上相似的程度; 而 \bar{n} 是该类集内目标人口单位的数量。

85. 严格来讲, 方程式(3.10)不是*deff*的公式, 因为忽略了分层效应; 在类集规模不统一的情况下引进的另一因子也被忽略了。但由于聚类成分是*deff*中的主要因子, 它仍然可以被用作一种近似形式, 以表明聚类如何影响样本设计以及怎样才能对它加以控制。

⁹ 有必要提请注意, 分层抽样和整群抽样在样本设计中并非相互竞争的选项, 因为在住户调查抽样中二者都是非用不可的。

86. 从以上表述可知, d_{eff} 是同类相关 δ 和类集规模 \bar{n} 这两个变量的倍增函数。 d_{eff} 随着 δ 和 \bar{n} 的增长而增长。虽然抽样人员不能控制考虑之中的任何变量的同类相关,但是他(她)可以在设计样本的时候上下调整类集规模,从而在很大程度上控制设计效果。

范 例

假设某人口在慢性病方面有一同类相关0.03,相当小。并且设想,在样本规模为5000住户的情况下,几位样本计划人正在辩论是采用10个住户的类集设计,还是采用20个住户的类集设计。为了简化说明,进一步假设所有住户的人口一样多:都是5口人。然后确定 \bar{n} 值:10户的类集为50,20户的类集为100。对方程式(3.4)进行简单替代,得 d_{eff} 的近似值为 $[1+0.03(49)]$,或者,10住户类集设计为2.5,但20住户类设计为4.0。由此可知,类集规模较大者,其设计效果大约比规模较小者高60%。接下来,就需要调查队来决定哪一种选择较好:要么采用10住户、类集数加倍(500)的选择,可在较为可接受的水平上保持可靠性;要么选择住户数为250的较便宜的设计,而以大幅增加抽样方差为代价。当然,也可以选择介于10住户和20住户之间的其他选项。

87. 有好几种解释设计效果的方法:作为(行将)在调查中使用的实际样本设计的抽样方差大于简单随机样本(SRS)的抽样方差的因子;只是作为衡量实际样本计划比简单随机样本在精度上差多少的量度;或者表明,为了实现同等的抽样方差水平,相较于简单随机样本而言,还需要在计划的样本设计中额外选择多少样本案例。譬如讲, d_{eff} 为2.0就意味着,若想实现简单随机样本所能达到的可靠性,需要选择的案例数就得翻一番。因此,对于主要指标来说, d_{eff} 值远大于2.5-3.0的样本设计是不可取的。

3.5.3. 类集规模

88. 前面提请注意抽样人不能控制同类相关。另外,对于大多数调查变量来说,很少有(即便有的话)试图估计此类相关值的经验研究。理论上,同类相关差异范围在-1和+1之间,不过很难设想许多住户变量的同类相关为负值。因此,对于抽样人来说,保持 d_{eff} 最小化的唯一可能就是在预算允许范围内要求类集规模越小越好。表3.3展示了各种同类相关值的 d_{eff} 和一个恒定的类集规模。

表3.3 各种同类相关 δ 和类集规模 \bar{n} 设计效果的聚类因素之间的比较

\bar{n}	δ						
	0.02	0.05	0.10	0.15	0.20	0.35	0.50
5	1.08	1.20	1.40	1.60	1.80	2.40	3.00
10	1.18	1.45	1.90	2.35	2.80	4.15	5.50
20	1.38	1.95	2.90	3.85	4.80	6.65	10.50
30	1.58	2.45	3.90	5.35	6.80	11.15	15.50
50	1.98	3.45	5.90	8.35	10.80	18.15	25.50
75	2.48	4.70	8.40	12.10	15.80	26.90	38.00

89. 从表3.3可以清楚地看出,若类集规模超过20,则 $deff$ 值(大于3.0)会高得不可接受,除非同类相关值相当小。在评价表内的数值时,有必要记住, \bar{n} 系指目标人口中的单位数,而不是住户数。就此而言,行将使用的 \bar{n} 值等于类集中的住户数乘以目标群体中的平均人数。譬如讲,如果目标群体是14-49岁的妇女,该群体一般每户有一位妇女,在这种情况下 b 住户的类集规模就有大约同等数量的14-49岁的妇女。换言之,目标群体的 \bar{n} 和 b 大体相等;按这种样子,表3.3适用。在下面的范例中,住户数和类集中的目标人口不相等。

范 例

假定目标人口为所有人,在以估计急性病和慢性病为目标的健康调查中就是如此。进一步假设该调查打算使用10住户的类集。在本案例中 \bar{n} 值是平均住户规模的10倍;如果后者是5.0,那么 \bar{n} 值为50。这样,必须在表3.3中用以估量其潜在 $deff$ 的 \bar{n} 值就是50。表3.3揭示了设计效果 $deff$ 非常大,除非在 δ 值约为0.02的情况下。这说明,对于一项特征调查比如传染病状况调查来说,一个设计为使用少到每类集10个住户的类集样本所获取的调查结果是非常不可靠的,因为后者的 δ 值可能会非常大。

90. 这个范例说明了为什么在设计住户调查的时候考虑到类集规模是如此之重要,尤其对行将计量的主要指标来说更是这样。另外要记住,样本设计中申明的类集规模一般是指住户数,而为了估量设计效果,类集规模必须考虑目标人口。

3.5.4. 设计效果($deff$)的计算

91. 分析家所指的调查变量的实际设计效果可以在完成调查以后来计算。这就需要估算选定变量的抽样方差(第七章讨论相关方法),然后计算每个变量方差对同一总体样本规模的单纯随机样本方差之比。这项计算是对包括分层效应以及类集规模变异在内的“完整”设计效果的估计,而不仅仅是对聚类因素的估计。

92. 根据方差比率的平方根可以得出标准误差比率(亦称 $deft$),在实践当中这方面的数据往往计算得出并且在调查技术文献中予以发表,譬如像“人口与健康调查”(DHS)。

3.5.5. 类集的数量

93. 有必要牢记,类集的重要性超出了它对抽样精度和也与总体样本规模有关问题的影响,因为类集规模决定了在调查中必须访问的不同地点的数量。这当然对调查成本有重要影响,因而这也是为什么首先要使用类集样本的缘故。这样,一个采用10户类集的10 000-住户样本,将需要有1 000个类集;而采用20户类集则只需要500个类集。前面强调指出,在对样本设计的这方面特征作决定的时候,至关重要是既要考虑到成本因素又要考虑到精度因素。

3.6. 分阶段抽样

94. 在理论层面上,完善的住户调查样本计划必须包括从构成住户总体 N 的适当分层中随机选择住户样本, n 。如此获取的分层随机样本将提供最大限度的精度。然而,此类样

本的使用费用过分高昂，实际行不通。¹⁰ 我们已在有关整群抽样成本效益的讨论中强调说明了这个问题。

3.6.1. 分阶段抽样的好处

95. 分阶段选择样本对选样过程本身有实际好处。它可以让抽样人能在连续步骤中隔离调查工作——尤其是编列住户清单和执行访谈活动——的地理位置。如果因为抽样标架过时需要重新编列住户清单的话，可以引进一个选样阶段以限制纳入清单的区域规模。

96. 若采取整群抽样法，一般来讲至少有两个选样阶段——首先是选择类集，然后是选择住户。在住户调查中类集总是被界定为某种地域单位。如果这些单位无论在地域上还是在总体规模上都足够小，并且它们有一个时新、完整又准确的清单可供抽样使用的话，那么有两个阶段就能满足样本计划的需要了。如果现有最小的地域单位对于有效用作一个类集还嫌太大的话，可能有必要分三个阶段选择。

范 例

假定某国想要将其类集界定为普选计数区(EAs)，因为这是行政上现有最小的地域单位。计数区标架(见第四章关于标架的较详细讨论)是完整的，因为整个国家划分成诸多计数区。这种标架是精确的，因为按照定义每个住户住在一个也是唯一的计数区内。另外，因为它以最近的普查为基础，所以只要普查以后没有在计数区规定上有任何变化，它也是够时新的。进一步假定最近的普查已经过了两年。因此决定需要编制一份更新的样本计数区住户清单，而不再使用已经用了两年的普查住户清单。平均每个计数区有200住户，但为采访目的设定的类集规模要求每个类集由15户构成。调查队的计算结果表明，以每15户构成一个最终抽样的类集(超过13比1)的比率，编制200个住户清单的成本太高了。于是抽样人决定实施一个较便宜的现场工作方案：将每个样本计数区分成规模大致相等的四等份，每份50个住户。然后将样本计划修订为从每个准备进行编列清单工作的样本计数区选择一个四分之一等份或区段，从而使编列清单的工作负荷减少了四分之三。在这样的设计中我们三个阶段：第一阶段，计数区的选择；第二阶段，计数区中的区段选择；第三阶段，住户的选择。

3.6.2. 虚拟阶段的使用

97. 在选样过程中往往使用所谓“虚拟阶段”，以免不得不在倒数第二阶段从庞大的单位档案中提取样本。相关档案也许包含了大量的单位并且可能太难处理，不是通过单调乏味的手工选择所能应付得了的。即便档案经过计算机处理，仍有可能量太大，很难从中进行有效的样本选择。¹¹ 虚拟阶段可以通过利用一个国家行政分区的等级性质，使我们能够把次级域的数量缩减到应付得了的水平。

¹⁰ 有一、两个例外情况，那就是地域上非常小的国家，比如科威特，那里的住户随机样本的选取在差旅费用方面会很低。

¹¹ 不过，有可能通过把非常庞大的计算机文档分解成若干对应于每个层次或行政区(比如地区或省级)的

98. 以孟加拉国的农村调查为例, 一般把村庄纳入倒数第二阶段的选样范围。孟加拉全国有100 000多个村庄, 多到远远超过能设法有效进行选样的地步。举例来说, 如果样本计划的设计是在倒数第二阶段选择600个村子, 那么在孟加拉全国大约每167个村子当中才能入选一个村。为了裁减选样档案规模, 可以决定利用孟加拉国行政区划的地方单位等级——专区、乡和村, 来分阶段选择样本。选择样本的步骤是: 首先使用与规模成正比的概率选择600个专区(该方法在第3.7节有详细讨论)。接着, 还是使用与规模成正比的概率从每个样本专区仅选一个乡: 于是样本中就有了600个乡。第三步, 使用与规模成正比的概率从每个样本乡各选一个村子, 结果选出600个村。最后一步, 从每个样本村选择住户样本。一般情况下, 这将导致每个样本村产生一个所有住户的系统化样本。

99. 上面讲的样本选择方法实际上是村和住户的两段抽样法, 但是从使用两个虚拟阶段入手, 先选择专区和乡, 然后再据此选择村。在这个例证中有必要用数学方法检查每个选择阶段的概率和总体选择概率, 来说明前两阶段的虚拟特征。

3.6.2.1. 第一选择阶段: 专区

100. 专区是用与规模成正比的概率方法选定的。这一阶段的概率由下式给出:

$$P_1 = \frac{am_t}{\sum m_t} \quad (3.11)$$

式中, P_1 是选择特定专区的概率; a 是预计选出的专区数(本例中是600); 而 m_t 是根据所用抽样标架(比如最近的人口普查)的数据在第 t 个专区辖下的农村住户数。¹²

101. 因子 $\sum m_t$ 是该国所有专区的农村住户总数。请注意, 选中的实际专区数也许不到600个。每当有一、两个专区两次入选的时候, 就可能发生上述情况, 对于规模超过抽样区间的任何专区来说都有这种可能。选择专区用的抽样区间由 $\sum m_t \div a$ 给出。于是, 如果设抽样区间为12 500个住户而相关专区共有13 800个住户的话, 该专区就自动被选中一次, 并且有1 300/12 500的二次入选机会(分子等于13 800-12 500)。

3.6.2.2. 第二选择阶段: 乡

102. 在第二阶段, 还是使用与规模成正比的概率方法, 从每个样本专区选择一个乡。实际做法是: 编列入选专区辖下所有乡的清単, 累积它们的规模量度 m_w , 并随机选择介于1和样本专区规模量度 m_t 之间的一个随机数。累积量的值是等于或大于这个随机数的最小的数, 用它来确定选中的乡(或者用等值惯例来确定入选的乡)。如果一个专区在第一阶段不止一次入选, 那就再从中选择同样数量的乡。第二阶段的选择概率由下式给出:

单独子文档, 使之比较容易地供抽样使用。

¹² 这是规模量度, 它也可以是相关专区的人口规模, 只要所用的数量和每个阶段的所有规模量度相一致即可。

$$P_2 = (1) \binom{m_u}{m_t} / m_t \quad (3.12)$$

式中, P_2 是在样本专区选中一个特定乡的概率; (1)表示只选一个乡; 而 m_u 是根据标架第 u 个乡的住户数。

3.6.2.3. 第三选择阶段: 村

103. 在第三阶段, 用与规模成正比的概率方法从每个样本乡选择一个村。第三阶段的选择概率由下式给出:

$$P_3 = (1) \binom{m_v}{m_u} / m_u \quad (3.13)$$

式中, P_3 是在样本乡中选择一个特定村的概率; (1)表示只选一个村; 而 m_v 是根据标架第 v 个村的住户数。

3.6.2.4. 第四选择阶段: 住户

104. 在第四阶段, 我们将假定每个入选村都有标架住户清单可用, 故可以从这些清单中有系统地选取住户样本。从每个样本村选择固定数量的住户; 这个固定数就是预先确定的类集规模。第四阶段的选择概率由下式给出:

$$P_4 = \binom{b}{m_v} / m_v \quad (3.14)$$

式中, P_4 在样本村中选择特定住户的概率; b 是每个村入选住户的固定数。

3.6.2.5. 总体选择概率

105. 作为每个阶段各种概率积数的总体概率用下式给出:

$$P = P_1 P_2 P_3 P_4 \quad (3.15)$$

通过置换, 我们有:

$$P = \left[\binom{am_t}{\sum m_t} \right] \left[(1) \binom{m_u}{m_t} \right] \left[(1) \binom{m_v}{m_u} \right] \left[b / m_v \right] = \left[(a)(b) \right] / \sum m_t \quad (3.16)$$

106. 请注意, P_2 和 P_3 完全相互抵消, 说明了这“四个”阶段的选择过程的虚拟性质。因此, 专区和乡虽然实际“选择”了, 但是仅起到确定样本村所在位置的作用。

3.6.3. 两段设计

107. 近来, 发展中国家把很大注意力放在使用两段样本设计方面。它是1990年代以来联合国儿童基金会(儿童基金会)在100多个国家进行的多指标类集调查(MICS)的样本设计。这些设计还主要运用在人口与健康调查(DHS)方面。

108. 通常, 两段设计简单来说是由经过第一阶段适当分层、采用与规模成正比的概率方法选择的数百个地域单位组成的样本。可在第一阶段样本单位中编列一份最新住户清单, 这要取决于是否有包括住户住址和/或所在位置在内的相关信息可供使用, 以及这种信息是不是最新的。接下来的第二阶段需要提取一个住户固定数的系统化样本。通常叫做“类集”的地域单位一般界定为农村地区的村或普查计数区和城市地区的街区。

109. 上面讲的两段设计在许多方面很有吸引力, 但主要是因为它简便易行。为了减少样本实施过程中的潜在非抽样误差起见, 在样本设计方面追求简化总比追求复杂更为有利。两段设计所具有的有益特征使它相对来说比较简便可取。比如:

- 如前所述, 样本设计是自我加权(样本中的所有住户都用同一概率选择), 或近似于自我加权(关于用与规模成正比的概率选择的样本和用与估计规模成正比的概率选择的样本之间的区别, 见第3.7.1节和第3.7.2节);
- 在大多数国家, 用计数区或城市街区界定的“类集”就规模而言是比较方便的(不算太大), 特别是如果在最后选择阶段之前必须重新编列住户清单的话, 更是如此;
- 通常计数区、城市街区和大多数村庄都经过测绘(不论为普查目的还是为其他目的), 因此界线清晰。

3.7. 与规模成正比的概率抽样和与估计规模成正比的概率抽样

110. 第3.5节提供了一个关于与规模成正比的概率抽样并突出类集选择特征的范例。本节进一步详细讨论与规模成正比的概率抽样。

3.7.1. 与规模成正比的概率抽样

111. 采用与规模成正比的概率抽样法可以使抽样人对类集调查的最终样本规模实施更有力的控制。在所有类集规模相同或大体相同的情况下使用与规模成正比的概率抽样法显示不出有什么优越性。比如说, 某一城市的每个街区正好有100个住户, 而我们想要一个分布在50个街区的1 000住户的样本。显然, 样本计划就得选择一个包含50个街区的简单随机样本(SRS), 亦即一个等概率样本, 然后系统地每个街区恰好每5个住户选择1户(也是一个等概率样本)。其结果将是一个每个街区恰好20住户的样本或总共1 000户的样本。该案例的选择方程式是: $P=(50/M)(1/5)$ 式中, p 是选择一个住户的概率; $(50/M)$ 是选择一个街区的概率; M 是该市的街区总数; 而 $(1/5)$ 是在特定样本街区选择一个住户的概率。

112. P 减少到 $10/M$ 。由于 M 是个常数, 所以选择每个样本住户的概率等于10除以街区数, M 。

113. 然而在实际情况中，可能用作住户调查类集的街区或其他地域单位很少在规模上如此一致。比如在上面的范例中，街区的户数可能在25户到200户不等。一个等概率街区样本可能结果“运气不佳”：多半是小街区或大街区。在这种情况下，结果就会是一个与范例中讨论的1 000个住户期望值大不相同的总体样本规模。有一个办法可以降低出现这种变异很大的样本规模的可能性，那就是根据类集规模进行分层并从每一层次选择一个样本。一般情况下不建议使用此种方法，因为它可能会减少使用样本设计中的其他分层要素或者使其复杂化。与规模成正比的概率抽样是可取的解决办法，因为它可以加大对最终样本规模的控制力度而无需按规模分层。

114. 为了说明与规模成正比的概率抽样，我们先从上面讲到的选择方程式入手，但更正式地表示两段设计¹³如下：

$$P(\alpha\beta) = P(\alpha)P(\beta|\alpha) \quad (3.17)$$

式中， $P(\alpha\beta)$ 是在类集 α 中选择住户 β 的概率； $P(\alpha)$ 是选择类集 α 的概率； $P(\beta|\alpha)$ 是在第二阶段选择住户 β 的条件概率，假定类集 α 是在第一阶段选取的。

115. 为了固定就住户数而言的样本规模，我们要求从 N 住户总人口当中提取一个 n 住户的等概率样本。这样，总体抽样率就是 n/N ，等于下面界定的 $P(\alpha\beta)$ 。另外，如果把行将抽样的类集数规定为 a ，那么理想上我们需要从每个类集选择 b 个住户，而不管被选中类集的规模大小。如果我们规定 m_i 为第 i 类集规模，那么我们需要 $P(\beta|\alpha)$ 等于 b/m_i 。于是，

$$P(\alpha\beta) = [P(\alpha)][b/m_i]$$

既然 $n=ab$ ，所以我们有

$$ab/N = [P(\alpha)][b/m_i]$$

为 $P(\alpha)$ 求解后面这个方程式，我们得到

$$P(\alpha) = (a)(m_i)/N \quad (3.18)$$

116. 请注意 $N = \sum m_i$ ，于是选择一个类集的概率与其规模成正比。因此，在第一阶段单位中选择与规模成正比的概率样本(但其中最终单位用等概率选择)的方程式是：

$$P(\alpha\beta) = [(a)(m_i)/\sum m_i][b/m_i] \quad (3.19)$$

$$= [(ab)/\sum m_i] \quad (3.20)$$

¹³ 关于这种标志法和有关与规模成正比的概率抽样法的补充讨论，见Kalton, 1983年，第38-47页。

117. 从方程式(3.19)可以看出,如此完成的样本设计是自我加权的,因为该方程式各项均为常数;忆及:虽然 m_i 是一个变量,但总和 Σm_i 则是一个等于 N 的常数。下面的图3.2提供了一个关于如何使用与规模成正比的概率选择类集样本的实例。

118. 说到实际选择样本,请注意,在图3.2中抽样区间, I ,先后七次(或 $a-1$ 次,式中, a 是拟选定的类集数)依次附加到随机起点 RS 上,结果选中的类集数是311.2(这是随机起点)、878.8、1 446.4、2 014、2 581.63、149.2、3 716.8和4 284.4。在每一种情况下,为这八个选择数抽样的类集是:其累积规模量度是等于或大于选择数的最小值。于是,类集03就被选中了,因为377是等于或大于311.2的最小累积量,而类集26被选中是因为3 744是等于或大于3 716.8的最小累积量。

119. 虽然该图解没有确定性地表明这一点(因为只选定了8个类集),但是与规模成正比的概率抽样一般倾向于选择较大而不是较小的类集。这也许是显而易见的,因为从公式(3.17)可以看出,选择一个类集的概率与类集规模成正比;因此,一个拥有200个住户的类集被选中的可能性是拥有100个住户的类集的两倍。所以请注意,如果一个类集的规模量度超过其抽样区间 I 的话,同一个类集就有可能不止一次被选中。然而,该图中没有一个类集符合这一条件;可是一旦发生了此种情况,行将在该类集中入选的住户数,若“二次选择”就会增至二倍,若三次选择就会增至三倍,如此类推。

3.7.2. 与估计规模成正比的概率抽样

120. 上一节讲述的与规模成正比的概率抽样方法多少有些理想化,在实际当中多数情况下可能实现不了。这是因为,在第一阶段用以确定类集选择概率的规模量度在第二阶段选择住户样本的时候往往不是实际的规模量度。

121. 在住户调查中,第一阶段用以选择初级抽样单位或类集的规模量度一般出自最近普查的住户(或人口)计数。即便普查刚刚过去不久,调查时的实际住户数也可能有小小的出入。不过,如果第二阶段的住户选择是直接出自用以确定规模量度的同一标架的话,会有例外(关于抽样标架的进一步讨论,见第四章)。

范 例

假定人口普查结束三个月后进行一次住户调查。调查队没有对入选类集的住户重新编列清单,而是决定在两段样本的第二阶段使用普查住户清单,因为据认为普查清单对于一切使用目的来说都是时新的和准确的。这种想法似乎有理。在第一阶段,使用普查时的住户计数作为每个村的规模量度选择了村一级的样本。对每个样本村来说,规模量度, m_i ,与选择样本所依据的实际住户数是一致的。因此,如果A村入选,据普查该村有235个住户,那么,据以选择调查住户样本的清单上也包含235个住户。

122. 可是,许多基于普查标架的住户调查中,调查是在普查结束好几个月、有时甚至好几年以后才进行的(见第四章关于抽样标架更新的进一步讨论)。在这种情况下,往往决定进行现场作业,以便在第一阶段选入样本的类集中准备一份更新的住户清单。然后根据更新的清单为调查选择一个住户样本。

图3.2
使用与规模成正比的概率系统选择类集的范围

类集/初级抽样 单位数	规模量度 (住户数)	累积量	样本选择
001	215	215	
002	73	288	
003	89	377	311.2
004	231	608	
005	120	728	
006	58	786	
007	99	885	878.8
008	165	1 050	
009	195	1 245	
010	202	1 447	1 446.4
011	77	1 524	
012	59	1 583	
013	245	1 828	
014	171	1 999	
015	99	2 098	2 014.0
016	88	2 186	
017	124	2 310	
018	78	2 388	
019	89	2 477	
020	60	2 537	
021	222	2 759	2 581.6
022	137	2 896	
023	199	3 095	
024	210	3 305	3 149.2
025	165	3 470	
026	274	3 744	3 716.8
027	209	3 953	
028	230	4 183	
029	67	4 250	
030	72	4 322	4 284.4
031	108	4 430	
032	111	4 541	

样本说明：使用与规模成正比的概率从总体32个当中选择8个初级抽样单位(类集)；因此选择区间(I)等于 $4\,541/8$ ，或567.6，式中，4 541是所有类集的总累积规模量度，而8是拟选定的类集数；随机起点(RS)是从随机数值表中选取的介于0.1和567.6之间的一个随机数；在本图解中，RS=311.2。

123. 用以选择类集的规模量度, m_i , 是上面范例中讨论的普查住户计数。可是据以选择住户样本的实际清单可能会与此有出入。当然, 它会在一定程度上有一个不同的规模量度, 这要取决于从普查结束到编列调查住户清单的间隔长短。在迁入迁出相关类集、建造新住所和拆除老住所、婚后分户(有时仍和父母住在同一住所)以及死亡等情况下, 都会导致住户数的变异。如果使用与估计规模成正比的概率选择样本, 其基于选择方程式的概率就是:

$$P(\alpha\beta) = \left[(a)(m_i) / \sum m_i \right] \left[b/m_i' \right] \quad (3.21)$$

式中, m_i' 是基于编列清单工作的住户计数; 其余项依照前面的规定。

124. 鉴于对大多数(若非所有)样本类集来说 m_i' 和 m_i 可能会有所不同, 所以在计算选择概率(因此选择其权数、亦即概率倒数)的时候应考虑到这种差别。如方程式3.20所示, 每个类集都会有一个不同的权数, 因而排除了自我加权的样本设计。

125. 通过使用精确权数来补偿普查和调查规模量度之间的差异, 即可使由此得出的调查估计数不带有偏差。若不相应调整权数, 就会产生有偏差的估计数, 其差异幅度无疑将随着普查和调查间隔拉长而加大。不过应该指出的是, 在 m_i' 和 m_i 差异不大的情况下, 样本基本上自我加权的; 而在某些情况下,¹⁴ 不用加权地产生调查估计数或许是谨慎的做法, 因为这样可以忽略不计偏差。但是在就此行动方向做出抉择之前, 务请逐一检查每个类集的 m_i' 和 m_i 值, 以便凭经验评估是否真的差异不大。

126. 每当采用与估计规模成正比的概率抽样法的时候, 在最后阶段可以使用一种备选方法来选择住户, 即样本实际上自我加权的方法。它涉及到在每个类集范围内根据其实际规模使用可变的比率来选择住户(下节讨论)。

3.8. 抽样选项

127. 本节讨论可供设计通用住户调查的适当样本时考虑的诸多选项当中的某些选项, 主要聚焦在有关倒数第二阶段和最后阶段的选样策略, 因为这两个阶段都有好几个备选方案。其中考察了倒数第二阶段选择等概率类集抽样法或与规模成正比的概率类集抽样法以及最后阶段选择固定比率住户抽样法对固定规模住户抽样的问题; 并且在一定程度上总结了前几节的样本规模控制问题, 自我加权对非自我加权的设计问题, 以及比如像访调员工作负荷之类的其他问题。此外还审查了目前正在广为使用的专项设计, 比如“人口和健康调查”以及联合国儿童基金会“十年中期多指标类集调查”的设计。这些设计提供了值得考虑的更多选项, 其中包括使用紧凑类集和非紧凑类集。

¹⁴ 在估计数局限于比例、比率或比的调查中, 这是一种适当的策略; 但在期望获得估计总数或绝对值的调查中, 就必须使用加权, 而不管相关样本是自我加权、近似自我加权、抑或非自我加权。

3.8.1. 等概率抽样，与规模成正比的概率抽样， 固定规模抽样和固定比率抽样

128. 在表3.4中，潜在的设计提供了一个讨论各种样本计划的做法、条件、优点和局限性的框架。

表3.4
备选样本计划：最后两个选择阶段

倒数第二阶段 单位的选择	固定类集规模 (住户数)	每个类集中的 固定选择率
与规模成正比的概率	计划1	计划2[未推荐]
与估计规模成正比的概率	计划3	计划4[未推荐]
等概率	计划5	计划6

129. 我们讨论了为什么初级抽样单位或类集的与规模成正比的概率抽样是一种能够比等概率抽样更准确地控制最终样本规模的手段——这是其主要优势——尤其在每个类集所包含的住户数差别很大的情况下。样本规模的控制不但对控制成本十分重要，而且对于调查管理者在实施调查工作之前准确计划访调员的工作量也十分重要。另一方面，等概率样本选择在执行上要比与规模成正比的概率抽样来得简单，在类集的规模量度(MOS)大体相当或相差不大的情况下这种方法是有意义的。在实际当中，只要实际的规模量度与抽样标架中给出的规模量度有出入，就要用与估计规模成正比的概率抽样替代与规模成正比的概率抽样。

130. 在每个样本类集中选择固定数量的住户有两个非常重要的好处：一是可以精确控制样本规模；二是这种方法可以使调查管理者能够精确分配访调员的工作负荷并且根据访调员的要求平衡他们的工作量。可是，固定规模抽样有点儿复杂，因为它需要计算每个类集的不同抽样区间。采用不同的抽样区间可能造成混淆，容易出现误差。不过，有一种内置质量控制校正机制，因为事先知道所要选择的住户数量。但因为不得不纠正选择中的误差，各种复杂因素仍有可能导致浪费时间并因此造成效率降低。

131. 从定义可知，固定规模抽样需要编列住户清单，以便据此指定和鉴别选中的住户。大多数情况下，这项编列清单的工作是实施调查前现场工作的一部分，所以是反映最新状况的。有必要由一个中心办公室来执行样本住户的选择，而不是由登记造册者本人做这项工作，以便最大限度地减少选择过程中出现偏差的可能性。

132. 或者，可以在每个类集中以固定比率对住户进行抽样。在这种情况下选择工作就简单多了，也不大容易出现偏差。在现场的一个好处是可以在访调员查核类集以获取最新住户清单的同时进行抽样。具体做法是在设计登记表的时候就表明预定的鉴别样本住户的方式方法。在一次访问中同时完成登记造册和抽样两项工作，这显然对节省成本来说是有利的，但是有重要局限性。

133. 固定比率抽样的一个局限性就是它基本上不对样本规模或访调员的工作量施加任何控制，除非每个类集的规模量度都差不多。另一局限性、也是更为严重的缺点就是：如果委托访调员来实际选择样本住户——亦即确定哪些住户符合列入样本的要求——的话，那

往往会导致选择上的偏差。迄今进行的无数研究都表明,凡是由访调员实施控制的住户选择一般规模都比较小,这说明访调员们可能有意无意地选择受访者较少的住户以减少工作量。

134. 设计是否自我加权,这取决于每个选择阶段抽样方法的特定组合方式。于是,一个采用与规模成正比的概率的整群抽样和固定规模住户抽样相结合的两段设计是自我加权的,而采用与规模成正比的概率抽样和固定比率抽样相结合的设计就不是自我加权的。下面的讨论说明表3.4中的哪些方案是自我加权的。

3.8.1.1 计划1: 与规模成正比的概率, 固定类集规模

条 件

- 可变的类集全域规模量度;
- 从用作规模量度的同一清单(比如普查住户清单)选择住户。

优 点

- 可以控制总体样本规模,从而控制成本;
- 可以控制访调员的工作负荷;
- 自我加权。

局 限 性

- 与规模成正比的概率抽样法或多或少要比等概率抽样法难以应用;
- 从每个类集选取住户的比率不同,有可能产生误差。

3.8.1.2 计划2: 与规模成正比的概率, 固定选择

135. 没有似乎可能的条件使用这种设计。如果类集规模是可变的,那么,结合固定类集规模的与规模成正比的概率抽样就是可用的适当方案。如果各个类集规模近似相等,那么固定比率抽样法就是适合的,但类集本身的选择应使用等概率抽样法。

3.8.1.3 计划3: 与估计规模成正比的概率, 固定类集规模

条 件

- 可变的类集全域规模量度;
- 用最新的住户清单更新标架清单,据以确定原始规模量度。

优 点

- 可以控制总体样本规模,从而控制成本;
- 可以控制访调员的工作负荷;

- 因为住户清单是时新的，所以对于特定标架来说要比与规模成正比的概率抽样法精确度高。

局 限 性

- 与估计规模成正比的概率要比等概率选择方法的应用难度大一些；
- 从每个类集选取住户的比率不同，有可能产生误差；
- 没有自我加权。

3.8.1.4 计划4：与估计规模成正比的概率、固定选择率

136. 由于上面方案2中陈述的理由，没有似乎可能的条件使用方案4中的这种设计。

3.8.1.5 计划5：等概率，固定类集规模

条 件

- 类集全域的规模量度大致相等或变化很小。

优 点

- 可以控制总体样本规模(但作用不如计划1那么大)，从而控制成本；
- 可以控制访调员的工作量，但也是作用不如计划1那么大；
- 等概率抽样法要比与规模成正比的概率或与估计规模成正比的概率抽样法容易应用。

局 限 性

- 从每个类集的选取住户的比率不同，有可能产生误差；
- 没有自我加权。

3.8.1.6 计划6：等概率，固定选择率

条 件

- 类集全域的规模量度基本相等。

优 点

- 自我加权；
- 在两个阶段选择样本都极为简便。

局 限 性

- 对总体样本规模的控制能力很差，尤其是如果时新规模量度与标架规模量度有很大出入的话，在成本和可靠性上可能产生不良后果；如果样本比目标人口小太多的话，也会带来可靠性的问题；

- 基本上不对访调员的工作量施加控制。

3.8.2. 人口和健康调查(DHS)

137. 虽然人口和健康调查(DHS)的聚焦点是育龄妇女,但是其样本设计适合于一般调查。

138. 自1984年以来已有许多发展中国家在实施人口和健康调查,因为它在抽样手册中提倡使用标准区段设计,¹⁵其方法简便,实用性强。标准区段按规模界定,一般为500人。国家给属于抽样标架的每个地域单位都指定了一个规模量度,这种规模量度是用各地域单位的人口除以500计算得出的(或者为有关国家确定任何一种标准区段规模)。经过四舍五入凑成整数,其计算结果就是有关地域单位的标准区段数。

139. 地域单位的与规模成正比的概率样本是使用标准区段数作为规模量度来选择的。因为用于这一阶段样本的地域单位一般是计数区(EAs)、城市街区或者村一级,所以其中占很大比例的地域单位规模量度等于1或2。对规模量度大于1的任何入选地域单位都组织进行测绘以建立地理区段,这种区段数就等于相关地域单位的规模量度。于是,对规模量度为3的样本地域单位进行测绘,以便在自然边界允许的情况下按照各区段居民人数(而不是按地理范围)将其划分为三个大体相等的区段。

140. 规模量度为1的每个地域单位被自动纳入样本,而在其余各地域单位则采用等概率抽样法随机选择一个区段。然后,详细检查包括自动入选区段在内的所有样本区段,以编列最新的住户清单。按照固定的份额(比率)有系统地每个样本“类集”选择住户,进行人口和健康调查采访。由于所有区段的规模都差不多大小,所以抽样程序便产生了两段等概率抽样区段样本和住户样本。

141. 人口和健康调查标准区段设计类似于上面讲的计划6:在样本类集(也是等概率)范围内的等概率类集选择和固定比率住户选择。不过,它通过标准分割程序避免了上述计划6的局限性:总体样本规模以及访调员的工作量得到近乎精确的控制。

142. 标准区段设计的一个主要好处是大大减少了倒数第二选样阶段的登记造册工作量。对于每个由 s 个区段构成的地域单位而言,造册工作量减少到 $1/s$ (如果只有一个区段,工作量就没有可减的了)。举例来说,如果一个特定地域单位包含四个区段,那么登记造册的工作量就只相当于在整个地区进行登记造册所需工作量的四分之一。由于这一特征,准备样本的成本也会相应降低。

143. 虽然住户造册成本降低了,但是却在另一方面付出代价。标准区段设计的一个局限性就是对于规模量度大于1的区段必须进行测绘工作。测绘工作是单调乏味、成本高昂的,它需要进行细致的培训,并且难免出现误差。自然边界往往未经恰当界定,这种情况阻碍了在相关地域单位范围内合理地勾勒区段。这方面的缺陷就使得日后访问该区段的访调员

¹⁵ 在1980到1990年代的泛阿拉伯儿童发展项目(PAPCHILD)中也使用了标准区段设计——见阿拉伯国家联盟(1990年)。

很难确切地找到被选中住户的住址。不过，后面这个问题或多或少可以通过在登记阶段将户主名字写进去而得到缓解，在这种情况下，边界划得不准确的问题就不那么严重了。

3.8.3. 修改的类集设计：多指标类集调查 (MICS)

144. 调查工作者普遍抱怨，在倒数第二阶段选定的样本类集中对住户进行登记造册花费了太多的金钱和时间。大多数调查——包括前面讲到的人口和健康调查标准区段设计方法——都需要进行登记造册的工作，以编制最新住户清单，从中选取调查采访对象。尤其在抽样标架已经存在了一年以上的情况下，这项工作更是至关重要。住户登记造册的工作牵涉大量的调查经费和加工处理过程，而这些往往在编列预算和策划调查期间被忽视了。除了调查采访所需要的访问之外，还需要单独进行一次实地访问，以完成造册。另外，需要列入清单的住户比率往往高达5-10倍于拟列选的户数。举例来说，假定样本计划欲选择类集规模为25户的300个初级抽样单位，拟总共采访7 500个住户。如果倒数第二阶段初级抽样单位平均每单位有150个住户，那就需要把45 000个住户列入清单。

145. “扩大免疫方案(EPI)类集调查”(世界卫生组织, 1991年)采用的抽样策略是由疾病控制中心和世界卫生组织制定的，其部分目的是为了在编列清单的过程中花费金钱和时间。扩大免疫方案类集调查意在估计儿童免疫覆盖面，过去20多年来已有许多发展中国家进行了此项调查。一个重要统计问题(Turner, Magnani和Shuaib, 1996年)关系到抽样方法。尽管第一阶段单位(村或邻里)通常是按照概率抽样原则选定的，但类集调查方法则在第二选样阶段使用比例样本。经常使用的比例样本方法虽然有变异，但是它必须在调查开始的时候先在选定的村子里采访某个中心点，然后按照随机确定的方向继续采访住户，直到达到特定比例为止。按照扩大免疫方案类集调查的变异继续访问住户，直到在目标年龄组发现7名儿童为止。虽然在使用此类方法方面没有出现故意的偏差，但长期以来许多统计学家就此提出了批评意见，其中包括Kalton(1987年)、Scott(1993年)和Bennett(1993年)。主要批评是这种方法不能产生概率样本(关于为什么在住户调查中推荐使用概率抽样法的讨论，见第3.2节关于概率抽样法与其他抽样法的对比论述)。

146. 扩大免疫方案类集调查方法的一个变异，即所谓“修改的类集调查”(MCS)设计，是为满足避免登记造册作业的抽样策略需要而开发的，但是却以概率抽样法为依据。在联合国儿童基金会的支持下，全球开展的多指标类集调查(MICS)，旨在监测世界儿童问题首脑会议确定的有关儿童和妇女状况的某些目标和指标，在这项调查中从各种不同的方面应用了修改的类集调查设计以及其他设计(联合国国际儿童教育基金会，2000年)。

147. 修改的类集调查设计是一种最低限度要求的抽样策略。它采用了简单的两段设计，运用仔细的分层并辅以快速查核及区域分割。不进行登记造册。修改的类集调查样本设计有如下基本特征：

- 使用与规模成正比的概率或等概率抽样方法(依初级抽样单位规模量度的可变程度而定)，选择第一阶段的一个地域单位(如村或城市街区)样本。即便普查标架已经用了好几年，仍可沿用老的规模量度，但该标架必须完全覆盖有关的人口(不论其为国家一级的还是地方一级的)；

- 使用现有的地图或草图，按照预定的区段数或者相当于普查规模量度除以想要(或预期)的类集规模所得商数的区段数，逐一快速查核每个样本地域单位并进行区域分割。创建的各区段在总体规模上大体相当；
- 从每个样本初级抽样单位选择一个概率相同的区段；
- 采访每个选定区段内的所有住户。

148. 在不必要登记造册的情况下划分区段是规模量度设计的关键优点。这有别于人口和健康调查的标准区段设计，后者要求每个区段都要编列住户清单。通过划分区段的工作也部分地补偿了使用可能过时标架带来的缺陷。虽然这样做的好处是可以产生无偏差的估计数，但是它也有局限性，就是对最终样本规模的控制能力较差，因为由于增长缘故，一个选定区段的实际规模可能会远远超过标架给出的数据。

149. 不过，对于规模量度设计来说测绘还是必要的，这跟人口和健康调查的标准区段设计要求没有什么两样。这种设计有着前面讲到的人口和健康调查方法的所有局限性。此外，在没有自然划界的小区域可能难以准确勾画规模同类集一样大的小区段。最后一个局限性是：就采访的区段是个紧凑型的类集——即所有住户在地域上相互毗邻——而言，与标准区段设计的非紧凑型类集相比，这种区段的设计效果较大，因为相对来讲它的同类相关程度较高。

3.9. 特别话题：两阶段样本和趋势抽样

150. 本节涉及有关住户调查样本设计的两个特别话题：(a) 两阶段抽样：其中第一阶段用来进行简短采访，目的是在住户居民中筛选属于目标人口的个人，第二阶段需要选择一个由符合标准的人构成的样本；(b) 一种抽样方法：为估计变化或趋势反复进行一项调查。

3.9.1. 两阶段抽样

151. 在现有信息量不足以有效选择相关目标人口样本的情况下，在住户调查中需要一种特殊的样本设计。这是因为，一般而言调查的目标人口是个分组人口——往往是人口稀少的分组——其成员仅占住户的一小部分。比如某个特殊族群的成员、孤儿群体或者收入高于或低于特定水平的人群。通常可以采用仔细分层的方法来鉴别某个族群或高收入人群比较集中的地域单位；但是如果这些人群相当杂乱地分散在人口当中，或者目标群体(比如孤儿)非常稀少的话，就不适合采取分层策略，而必须使用其他方法来对其进行抽样了。

152. 经常使用的一种方法是两阶段抽样，亦称“后分层抽样法”或“复式抽样法”。这种方法分为四步：

- (a) 选择一个“大型”住户样本；
- (b) 进行简短的筛选采访，以确定有目标人口成员居住的住户；
- (c) 根据筛选采访对大样本进行后分层，将其分成两类；

- (d) 从这两个层次各选出一个次级住户样本，以便对目标群体进行第二次、也是较长时间的采访。

153. 两阶段抽样法的目的在于通过在初始的大样本中进行简短的筛选采访节省调查成本。接下来在日后进行时间较长的采访，但是只对符合标准的住户进行采访。为此，初始样本往往是为别的目的选择的样本，而筛选采访是作为主要调查的“附加项目”捎带进行的。于是，这种做法就把大部分资源分配给了第二阶段抽样和采访，而筛选阶段只需要不太多的预算。

范 例

假设正在计划调查800名与幸存的单亲家长或其他亲属一起生活的孤儿(与之相对的是生活在机构环境的孤儿)。进而假设，据估计若想给800名孤儿定位需要对16 000个住户进行抽样——大约每20户有一个孤儿。考虑到，仅为800户访谈设计和执行一个16 000个住户的样本不切实际，所以决定利用一项也在计划之中的关于健康问题的一般调查。该健康调查设计为一个20 000个住户的抽样。这两项调查的管理者商定，在健康调查中单独附加一个问题：这家住户是否有17岁或以下、其母亲或父亲已故或父母双亡的人？预期通过这一附加问题将会找出有大约1 000名孤儿的住户。然后，孤儿调查管理者将会按80%的抽样率对这些住户进行次级抽样，以供详细采访。

154. 上面的范例还说明了在什么情况下适合采用两阶段抽样法。请注意，该解说中所瞄准的样本规模只有800名孤儿，但是就住户数而言为找到这么多孤儿所必要的样本规模是16 000户。因此在计算后者的时候(见公式3.7)，抽样技术人员和调查管理者会得出这样的结论：两阶段抽样法是最实用的成本-效益高的抽样设计。

155. 有两个原因说明第一阶段样本的后分层很重要。供筛选用的问题几乎总是简明扼要的，因为这种问题附加于另一项调查，而毫无疑问该调查本来已经是非常耗费时间的采访了。主调查的管理者不大可能会同意附加一套详尽的筛选问题。因此，上面范例中为查找孤儿选定的某些住户实际上很可能家中没有孤儿，或者相反——有孤儿的住户没有入选。这种分类不当的错误说明，应设立两个层次：一个层次筛选结果为正面的住户，另一层面是筛选结果为负面的住户。分别从每个层次提取样本以供充分采访，理由是可能在一定程度上发生分类不当的失误。“正面”层次的抽样率将非常高——达100%，而“负面”层次的抽样率就小多了。

3.9.2. 旨在估计变化或趋势的抽样

156. 在许多国家，住户调查是为双重目的设计的，即为了估计：(a) 关于调查首次实施时机的基线指标(它们的水平)；(b) 在第二次即后来实施调查的过程中这些指标的变化。如果调查重复一次以上，还要计量各项指标的的变化趋势。随着重复调查，会对样本设计产生各种影响，而在进行一次性截面调查的时候就没有这样的影响。尤其令人关切的是变化估计的可靠性问题以及本次调查和下次调查的相同或不同住户的使用适当结合的问题。与

后面这个问题有关联的是担心在同一住户被反复采访的情况下可能会产生偏差和增加其答复的负担。

157. 考察可靠性问题也需要数学示范。我们先来看估计变化的方差, $d=p1-p2$, 表示为:

$$\sigma_d^2 = \sigma_{p1}^2 + \sigma_{p2}^2 - 2\sigma_{p1,p2} = \sigma_{p1}^2 + \sigma_{p2}^2 - 2\rho\sigma_{p1}\sigma_{p2} \quad (3.22)$$

式中, p 值是正在被估计的比率; σ_d^2 是差异的方差; σ_p^2 是第一次或第二次调查(用1或2表示) p 的方差; $\sigma_{p1,p2}$ 是介于 $p1$ 和 $p2$ 之间的协方差; 而 ρ 是两次调查观察值 $p1$ 和 $p2$ 之间的相关系数。

每当估计变化较小的时候(往往如此), 我们就有

$$\sigma_{p1}^2 \approx \sigma_{p2}^2$$

于是, $\sigma_d^2 = 2\sigma_p^2 - 2\rho\sigma_p^2$ (我们可以去掉下标, 1和2)。因此:

$$\sigma_d^2 = 2\sigma_p^2 (1-\rho) \quad (3.23)$$

158. 评价方程式(3.22), 我们注意到在某次类集调查中, σ_p^2 的估计值是一个简单随机样本SRS的估计数乘以样本设计效果, $deff$, 的积数。在使用同一住户样本的时候, 相关系数, ρ , 最高, 可能是0.8或更高。在这种情况下, σ_d^2 的估计量 s_d^2 由下式给出:

$$s_d^2 = 2[(pq)f/n](0.2) \quad \text{或} \quad 0.4(pq)f/n \quad (3.24)$$

159. 如果使用相同的类集而用不同的住户, ρ 仍为正数但大大变小——或许约为0.25到0.35, 那我们就有(对于0.3的 ρ):

$$s_d^2 = 2[(pq)f/n](0.7) \quad \text{或} \quad 1.4(pq)f/n \quad (3.25)$$

160. 最后, 用第二次调查的一个完全独立的样本, 使用不同的类集和不同的住户, ρ 为零, 我们有:

$$s_d^2 = 2[(pq)f/n] \quad (3.26)$$

使用典型的 $deff$ 值2.0, 公式3.19得出:

$$s_d^2 = 4[(pq)/n] \quad (3.27)$$

161. 对于部分重叠的重复调查, 比如用50%的相同类集/住户, 而用50%的新类集/住户, ρ 就必须乘以等于重叠样本比率的因数 F 。在这种情况下方程式3.16变成:

$$\sigma_d^2 = 2\sigma_p^2 (1-F\rho) \quad (3.28)$$

162. 从以上讨论可以概括几个令人感兴趣的要点：首先，使用相同住户样本的两次调查之间较小估计变化的估计方差仅约为第一次或第二次调查平均方差的40%。使用相同类集但是不同住户所产生的一个变化方差估计数比平均方差估计数高40%。独立样本产生的估计方差比平均值高一倍。

163. 可见，就可靠性而言，在重复调查中使用相同的住户有很大优越性。如果做不到这一点，那么通过下述两个办法仍可做出重要改进：(a) 使用一部分相同住户，或者(b) 使用相同的类集但不同的住户。相较于最缺乏吸引力的选项——即使用完全独立的样本而言，上述两种做法所产生的估计数方差都比较小。

164. 关于非抽样误差问题，有两种负面反应效果：一是不答复，二是有条件的反应。重复使用相同住户样本越频繁，上述负面效果越明显。不但受访者越来越不愿意配合，从而增加了后来调查轮次的不答复率，而且他们还受到条件作用的影响，致使其回答质量或准确度随着重复访谈而愈益下降。

165. 与条件作用有关联的是一种所谓“受访次数”偏差现象，就是基于受访者就同一时段报告得出的调查估计数由于接受不同程度的调查而有不同的预期值。围绕这种现象进行了广泛研究的结果表明，许多专题调查都存在这个问题，诸如劳动力、支出、收入和刑事定罪等方面的调查。例如在美国，那里的劳动力调查受访者一般受到8次采访；首次受访者的失业估计数向来都比整个8次采访中受访者提供的平均值高7%左右。多年来在美国一直存在这种情况。在解释此种偏差方面，专家们主要提出如下一些看法：

- 访调员不要在后来的采访中向受访者提供跟初次采访一样的激励手段；
- 受访者可能知道有些回答会引发更多的提问，所以他们回避给出肯定的答案；
- 首次采访可能涉及所涉期以外的事件，而在后来的采访中事件是“限定的”；
- 受访者可能会因为接受调查而改变其行为举止；
- 一旦受访者对调查过程感到厌倦了，他们在后来的采访中也许就不那么细心提供准确答案了(Kasprzyk, 1989年)。

166. 值得注意的是，上面引述的大多数道理对同一调查的重复采访是适用的；但是如果不同的调查中使用相同的住户，就会发生某些相同的反应。

167. 从以上可以看出，在以下几方面的使用上存在相互竞争的效应：

- (a) 每次调查使用同一住户样本；
- (b) 部分样本使用置换住户；
- (c) 每次实施调查使用一个新住户样本。

168. 若从(a)向(c)推进，有关变化估计数的抽样误差就呈增长趋势，而非抽样误差则呈递减趋势。在每次调查都使用同一样本住户的情况下，抽样误差最少，因为观察数值间

的相关性最高。对比之下, 同样住户的使用会增加抽样偏差。如果每次调查都使用新的住户样本, 情况正好相反。

169. 一般认为, 选项(b)在权衡抽样误差和非抽样偏差方面提出一个折中办法。如果逐年保留部分样本, 那么选项(c)的抽样误差和选项(a)的非抽样误差都会得到改善。如果某项调查只进行两次, 那么选项(a)有可能是最佳选择。在一个样本只用两次的情况下, 答复方面的作用不会对整个调查误差产生太大的破坏性影响。不过, 如果重复进行三次或三次以上调查, 那就最好采用选项(b)。有一个方便的办法, 那就是每次调查轮流置换50%的样本(关于主样本中的轮流抽样, 见第四章)。

3.10. 当实施工作出毛病的时候

170. 本节概括归纳在实施样本计划遇到障碍的时候需要采取的各种行动, 其中大多已在上文讨论过或间接提到了。可是本章和下一章所强调的重要原则之一, 就是通过在构想样本设计的时候只要小心仔细地进行策划, 实施当中的许多障碍是可以预防的。但计划得再好, 仍会出现预料不到的问题。

3.10.1. 目标人口的定义和覆盖面

171. 由于各种缘故, 当调查所覆盖的实际人口不是意想中的目标人口的时候, 往往就会出问题。

范 例

让我们来考虑一项拟覆盖全国所有人的典型目标人口的调查。由于下述一系列原因中的任何一种, 实际覆盖人口(即从中选择样本的人口)都往往会少于总人口:

- 生活在机构住处(比如医院、监狱和军营等)的人未被抽样;
- 居住在某些地区的人可能被有意排除在覆盖范围之外。其中可能包括地形崎岖难以通达的地区, 受自然灾害影响的地区, 由于内乱或内战被宣布为禁止进入的地区, 难民或其他外国劳工居住的围居区或帐篷区, 等等;
- 没有永久住所的个人被定为“超出调查范围”的人。其中可能包括游牧民族、船上人家、过往的劳工, 等等。

172. 就样本计划而言, 上述各类分组人口的问题在于, 通常他们在调查之前没有被预先确定为应排除在外的人。于是问题就来了: 在实施选择样本的时候偶尔会选到, 譬如讲, (a) 一个劳改所、监狱或集体宿舍而不是“传统”意义上的居民区; 或者(b) 是一个地处山区交通不便的初级抽样单位。在这些情况下经常会采取的“解决办法”是用另外一个初级抽样单位来替换。但这是一个会带来偏差的办法。

173. 可接受的解决办法是在样本设计阶段避免出现问题。首先要仔细界定目标人口, 不仅确定调查所要包括的分组人口, 而且确定需要排除在外的分组人口。其次, 要修订

抽样标架，从中删除任何不准备调查的地理区域。这个做法也适用于应当排除在外的任何特别计数区——比如劳改营。第三，应从修订的标架选择样本。第四章将更详细地讨论抽样标架的问题。

174. 还应记住一点，就是上面建议的解决办法有助于更确切地界定目标人口。准确描述目标人口对于用户获取适当信息十分重要。

3.10.2. 就调查预算而言样本规模过大

175. 另一个问题是在计算的样本规模过大、以致调查预算难以支撑的情况下发生的。当出现这种情况的时候，调查队必须要么寻求增加资金，要么修改其度量目标：降低精确度要求，或者减少域的数量。

176. 降低精确度(增加抽样误差)从而大幅度降低成本的一个办法，就是在保持总体样本规模的前提下减少选择的初级抽样单位数。譬如讲，不用600个初级抽样单位、每单位15个住户($n=9\ 000$)，可将样本计划改为选择400个初级抽样单位、每单位22或23个住户($n\approx 9\ 000$)。至于域，一个解决办法或许是——譬如讲——选定全国4个大区而不是10个省份。

3.10.3. 类集规模大于或小于预期

177. 一个经常发生的问题是，由于譬如进行了新的住房建设，尤其是如果抽样标架已经过时的话，一个样本类集可能远远大于其规模量度。调查队也许预料某个特定类集有125个住户，但是可能在登记造册阶段实际发现了400个住户。在这种情况下一个可能的解决办法就是把该类集细分为人口规模大致相等的若干地域小区段。区段数应相当于当前住户数除以原来的规模量度，经四舍五入凑整数。在我们的范例中这个商数是 $400/125$ 或3.2，四舍五入为3个区段。通过测绘和速查住所(与住户相对而言)产生这些区段。然后随机选择一个区段以供造册。

178. 也可能发生相反的问题。由于拆除或破坏、自然灾害或其他原因，一个类集或许远远小于预期。在这种情况下往往试图用另一个类集来取而代之，但是这种做法会带来偏差。不如就按其现在的状况进行抽样。虽然这样做的结果是最终样本规模会小于目标规模，但抽样误差不会增加多少，除非涉及许多这种类集。原封不动地起用(或者不替换)较小的类集却可以产生无偏差的估计数，因为该类集“代表了”自从原标架确立以来当前人口曾经发生过的变化。

3.10.4. 处理不答复的情况

179. 与实施抽样相比较，虽然不答复的情况更多地属于实施调查的问题，但它是个可以毁了住户调查估计数的严重问题(见第六章和第八章关于不答复问题的详细讨论)。如果允许不答复率超过样本个案10%–15%的话，所得估计数的偏差就非常令人怀疑了。对此，许多国家也倾向于通过用答复的住户替代不答复的住户来“解决”问题。这个办法本身就有偏差，因为被替代的住户仍然只代表答复住户而非不答复住户。据了解，答复群体和不答复群

体特征在一些重要调查变量方面是不同的, 尤其是与社会经济地位有关联的变量。而比较可取的解决办法是努力争取从起初不答复的住户获得答复(不幸的是, 这个办法绝不可能100%取得成功)。这需从一开始就计划好通过一系列连续返回再访住户不答复者, 以便争取他们的合作(对拒绝受访者而言), 或者在家里找到他们(对不在家者或因为别的缘故未能取得联系者而言)。可能需要多达五次回访, 但最少不得低于三次。

3.11. 准则概要

180. 本节概括从本章归纳的主要准则。虽然有些准则几乎在任何情况下都适用(比如“使用概率抽样法”), 但是也有一些准则适合于例外情况, 这要取决于一个国家的特殊情况、资源状况和需求。因此, 本着“经验法则”的精神, 以核对表的格式而不是作为固定的、不可动摇的建议提出这些准则。参与调查者应承诺:

- 在选样的各个阶段使用概率抽样法;
- 样本设计要尽可能追求简单方便而摒弃复杂繁琐;
- 尽量使用能够产生带有自我加权或近似自我加权的域内样本或总体样本(假如设计不包括域的话)的选样方法;
- 若可能的话使用两段样本设计法;
- 使用像公式(3.5)那样的方程式来计算样本规模, 必要时调整固定参数值(比如预计的不答复率和平均住户人口)以反映全国情况;
- 使用设计效果值2.0作为样本规模公式中的缺省值, 除非相关国家有更好的信息可用;
- 样本规模要以据认为是调查行将覆盖的所有关键估计数当中人口百分比最小的关键估计数为基础;
- 在预算允许的范围内选定关键估计数(见上条)的容许误差或精度水平, 即估计数的10%, 亦即10%的相对误差, 置信区间为95%; 否则选定12%–15%的相对误差;
- 如果方便并且适当的话, 将普查计数区(EAs)确定为初级抽样单位(PSUs);
- 使用隐含分层法并尽可能结合使用系统化的与规模成正比的概率抽样法(*pps*), 尤其对多种目的的综合设计更是如此;
- 将估算域限制在绝对必要的数量上(以便使样本规模保持在可管理的水平上);
- 竭力争取选用大量(数百个)类集(或者初级抽样单位——若是两阶段抽样的话): 数量越多越好;
- 使用小规模类集: 10–15个住户: 规模越小越好;
- 使用固定不变而不是可变的类集规模, 就是说, 要使用固定的住户数, 而不是固定的抽样率;

- 至于域，力争每个域至少有50个初级抽样单位；
- 计划至少要回访三次——但最好是五次，以转变不答复住户的态度；
- 对于稀少人口，考虑采用两阶段抽样法，即在一项已经计划好的现有大型调查中附加一个“捎带”提问，以便找出目标个人；接着对二段样本进行强化采访；
- 对于旨在计量变化的调查，如果计划只进行两次采访的话，那就两次都采访同一住户；如果准备进行三次或更多次采访的话，那就采用部分重叠的方案，每次在样本中轮番加入新住户。

参考资料和补充读物

Bennett, S. (1993年)。《疫苗接种扩大方案整群抽样方法：关键评估》。征文。国际统计学会届会，意大利佛罗伦萨。

Cochran, W. (1977年)。《抽样方法》(第三版)。纽约：威利国际出版公司。

Hansen, M., W. Hurwitz和W. Madow(1953年)。《抽样调查方法和理论》。纽约：威利国际出版公司。

Hussmans, R., F. Mehran和V. Verma(1990年)。《从事经济活动人口就业、失业和未充分就业情况的调查：劳工局概念和方法手册》。日内瓦：国际劳工局。第11章，“样本设计”。

国际统计研究所(1975年)。《样本设计手册》。世界生育率基本文献。荷兰福尔堡。

Kalton, G. (1983年)。《调查抽样导论》。加利福尼亚贝弗利山：赛奇出版社。

_____ (1987年)。《世卫组织估计疫苗接种覆盖面简化整群抽样方法评估》。提交儿童基金会的报告，纽约。

_____ (1993年)。《稀少人口和难以通达人口的抽样调查》。经济社会信息和政策分析部统计司和国家住户调查能力方案，INT-92-P80-16E。纽约：联合国。

Kasprzyk, D. 等编辑(1989年)。《群组调查》。纽约：约翰·威利国际出版公司，第1章。

Kish, L. (1965年)。《调查抽样》。纽约：威利国际出版公司。

Krewski, D., R. Platek, 和J. N. K. Rao编辑(1981年)。《调查抽样的最新话题》。纽约：学术出版公司。

Le, T. 和V. Verma(1997年)。《人口和健康调查样本设计和抽样误差分析》，《人口和健康调查分析报告》，第3号。马里兰州卡尔夫顿：宏大国际咨询公司。

阿拉伯国家联盟(1990年)。《阿拉伯母婴健康调查抽样手册》，基本文献第5号。开罗：泛阿拉伯儿童发展项目。

宏大国际咨询公司(1996年)。《抽样手册》，第三次人口和健康调查基本文献第6号。马里兰州卡尔夫顿：宏大国际咨询公司。

- Namboodiri, N. 编辑(1978年)。《调查抽样和计量》。纽约: 学术出版公司。
- Raj, D. (1972年)。《抽样调查设计》。纽约: 麦格劳-西尔出版公司。
- Scott, C. (1993年)。“在发展中国家花钱不多的调查方法”会议与会者的评论意见。征文。国际统计学会届会, 意大利佛罗伦萨。
- Som, R. (1966年)。《实用抽样方法》(第2版)。纽约: 马塞尔·德克尔出版公司。
- Turner, A., R. Magnani, 和 M. Shuaib (1996年)。“疫苗接种扩大方案类集调查设计的一个不太快但顺畅得多的选项”, 《国际流行病学刊》(英国利物浦)第25卷第1期。
- 联合国(1984年)。《住户调查手册》(修订版), 《方法研究》, 第31号, 出售品编号: E. 83. XVII. 13。
- _____ (1986年)。“一体化住户调查方案抽样标架和样本设计”, 《国家住户调查能力方案》。纽约: 联合国发展技术合作部和统计处。
- _____ (2005年)。《发展中国家和转型期国家住户调查》, 《方法研究》, 第96号, 出售品编号: E. 05. XVII. 6。
- 联合国儿童基金会(2000年)。《十年末多指标调查手册》。第4章, 题为“样本设计和选择”及附件7, 题为“抽样细节”。纽约: 联合国儿童基金会。
- 美国人口普查局(1978年)。《现时人口调查设计和方法论》, 技术文件第40号。哥伦比亚特区华盛顿: 人口普查局。
- Verma, V. (1991年)。《抽样方法》, 培训手册。东京: 亚洲和太平洋统计学会。
- Waksberg, J. (1978年)。“随机数字拨号抽样方法”, 《美国统计协会会刊》, 第73卷, 第40-46页。
- 世界银行(1999年)。《核心福利指标问卷(CWIQ)手册》。哥伦比亚特区华盛顿: 世界银行, 第4章。
- 世界卫生组织(1991年)。《疫苗接种扩大方案, 中层管理人员培训: 覆盖面调查》。WHO/EPI/MLM91. 10, 日内瓦。

第四章

抽样标架和主样本

4.1. 住户调查中的抽样标架

1. 上一章涵盖了样本设计的多方面特征和住户调查样本设计的一些可用的选项，只是没有讲抽样标架的问题。而在住户调查中样本设计的一个最重要的方面就是抽样标架，所以才单独设置一章来讨论这方面问题。

2. 抽样标架对无论住户调查还是其他任何调查的费用和质量都有重要影响。在住户调查中，有缺陷的抽样标架是非抽样误差、尤其是重要人口亚群漏查问题的共同来源。本章试图在考虑到各种不同抽样阶段的基础上阐述标架的构建和使用方面的最佳做法。它分两节：第一节讨论有关标架及其建立的一般问题，重点放在住户调查的多极样本设计方面；第二节讨论在使用主抽样标架方面引起的特殊问题。

4.1.1. 抽样标架的定义¹

3. 抽样标架的一个简单实务定义就是：选择样本所依据的一套原始资料。该定义还包含了抽样标架的目的，即为选择行将在调查中接受采访的特定目标人口成员提供一种手段。可能需要不止一套资料。一般在住户调查中也是如此，这是因为它的多阶段性质。在住户调查的初期阶段，一般都是从地域标架提取样本。在最后阶段，既可以从地域标架抽样，也可以从清单标架抽样(见下文关于地域标架和清单标架的讨论)。

4.1.1.1. 抽样标架与目标人口

4. 在确定住户调查所使用的适当标架方面的一个重要考虑，就是调查的目标人口与抽样单位之间的关系问题。抽样单位决定了标架。它还决定了最后阶段的抽样概率。

范 例

举例来说，在一项以婴幼儿为目标人口的调查中，调查队可能要考虑两种潜在的标架：一是过去12个月来记录出生人口的医疗机构；二是家里有12个月以下婴幼儿的住户清单。在第一种情况下，标架可能由两部分组成，每一部分代表一个抽样阶段：首先是给婴儿接生的医院和诊所清单；其次是过去12个月所有在这些医院和诊所降生的

¹ 关于第三章和第四章使用的有关抽样的术语，请读者参见第三章表3.1。

婴幼儿清单。抽样单位将是第一阶段的医疗机构和第二阶段的婴幼儿。于是在抽样的最后阶段抽样单位和目标人口就是同义语了。然而在第二种情况下,有可能(在抽样的稍后阶段)把标架确定为小地域单位(比如村或城市街区)范围内的一份住户清单。在实施样本计划的时候要对住户进行选择 and 筛查,以查明家里是否有0-12个月大的婴幼儿。在这种情况下,住户就是选择概率所依据的抽样单位。不过请注意,在为查明目标人口的存在而对住户进行筛选以前,并没有对目标人口成员进行实际鉴别和调查。可见,就住户标架而言,抽样单位和目标人口是不同的。

5. 在住户调查(本手册的主题)中,抽样单位以及样本设计所依据的单位是住户。而目标人口即使在一般目的的调查中也会因度量目标不同而变化。住户收入和支出调查除外,通常目标人口不是指住户本身的人口。例如在就业调查中,目标人口一般是10岁(或14岁)以上的人,这就完全排除了幼小儿童;在关于妇女生育健康问题的调查中目标人口由14-49岁的妇女组成(而且往往只包括这个年龄组结婚的妇女),等等。

4.1.2. 抽样标架的特征

6. 前面讨论过,就统计而言,抽样标架当然要捕捉目标人口。除此之外,完美的抽样标架应当是完整的、准确的、最新的。这些理想特征在住户调查中是难以实现的。但是一定要竭尽全力去争取实现,不论是从零开始构建标架,还是利用现有的标架。可以根据一个标架的理想特征与目标人口的相关程度来评估它的质量。回顾前面第三章,我们提出的概率样本定义——即样本中目标人口的每个成员都有一个已知的非零入选机会的概率样本——是评判标架质量的有用晴雨表。

7. 因在实现上述各个理想特征方面的失误程度而异,调查结果会以各种不同的方式发生偏差,但往往倾向于对目标人口估计不足。

4.1.2.1. 完整性

8. 据认为,标架的完整性就看是否目标人口的所有成员(即全域)均为标架所覆盖。因此,目标人口的覆盖面是判断标架是否适用于调查的一个至关重要的特征。如果它不适合,那么调查队就必须评估是否可以对其加以修整或进一步开发以使之较为完整。在前面的范例中,如果把医疗机构当作唯一抽样标架来使用的话,那么在家里或在医疗机构以外其他地方出生的婴幼儿就可能被排除在调查范围之外。因此,在此例证中就会有为数众多的目标人口入选样本的机会为零,因而违反了概率样本的条件。结果,按照医疗机构这个标架婴幼儿人数就被低估了。另外,在医疗机构出生的婴幼儿特征也与在家里出生的婴幼儿有很大不同。因此,机构标架就会产生带有偏差的有关婴幼儿或其保育状况的重要指标分布数据。

9. 覆盖不充分也是住户调查中的一个潜在问题。举例来说,某项全国调查计划可能打算通过一次住户调查覆盖整个人口。可是有各种各样的人口子群,比如生活在机构里的人,以及没有固定住所的游牧民和船上人家等。在这种情况下,通过住户调查显然达不到总

人口的覆盖率。还需要追加开发覆盖非住户群体的抽样标架，以便使他们的成员获得非零入选概率。若做不到这一点，就得修改实际目标人口，以求更仔细地界定其内涵。这样，用户就能清楚地知道人口的哪些部分已被排除于调查范围之外。

4.1.2.2. 准确性

10. 准确性是抽样标架的另一个重要特征，但在用于住户调查以外的标架中发生不准确情况的可能性比较大。如果目标人口的每个成员被纳入一个标架一次且仅仅一次，那就可以说这个标架是准确的。让我们以一份雇用50名以上工人的企业清单为例来说明。在下列情况下可能出现误差：(a) 清单上有任何企业雇用49名以下的工人；(b) 该清单漏掉了雇用50名以上工人的任何企业；或者(c) 某个企业不止一次地被列入清单(或许用不同的名称)。

11. 在住户调查中就较少可能会遇到这种包含不准确性的标架。可是有些情况下也可能发生不准确的情况，比如：(a) 一个由计数区(EAs)计算机文档构成的标架遗漏了某些要素；(b) 一个村庄的住户清单标架漏掉了该村周边的某些住户；(c) 在一个地域单位住户清单标架中，有些住户被列入不止一个抽样单位；或者(d) 一个过时的住户清单标架没有包括新建的住所。最后提到的这种情况，即过时标架的问题，将在下文进一步讨论。

12. 在一个地域单位中漏掉计数区或列入清单的住户显然会使受影响的住户没有机会入选样本。同样，这也不符合产生准确概率样本所需具备的条件。重复列入清单也违反了概率标准，除非考虑到这种重复列入的情况才能计算出准确的选择概率。不幸的是，往往没有查出来提到的这些类型的遗漏和错误重复。可见，抽样技术人员也许不知道需要在从中抽样之前订正标架。另一方面，如果一个标架中发生遗漏或重复的情况所占比重很小，通常不会在调查估计数中造成明显的——或值得注意的——非抽样偏差。

4.1.2.3. 最新标架

13. 当然，一个标架首先要最新才能具备另外两项特征——完整和准确。过时的标架显然包含不准确的情况，从而导致样本不完整，尤其在住户调查中更是如此。可以把已经过去好几年的人口普查当作一个过时标架的典型范例。陈旧的普查资料不会准确反映新建或拆除住所、迁入迁出居住单元、出生与死亡等变化情况。这些缺陷不符合概率样本的要求，即目标人口的每个成员都有一个已知的入选机会。

范 例

假定标架是由根据最近普查确定的计数区构成的，而这次普查迄今已时隔4年，没有进行过标架更新。同时假定，在都市郊区以前普查时没有或基本没有人的计数区如雨后春笋般地新建了不计其数占用公地的棚户区。样本设计不会向生活在以前空旷的计数区的住户提供入选机会的，因此不符合概率样本条件。在从前基本无人居住的计数区，即便这些计数区不在技术上违反概率抽样的要求，也会引起另外一个严重问题。毫无疑问会采用与规模成正比的概率来选择样本，其规模量度按照普查人口或住户计数。由于在普查时人口很少，所以如果采用与规模成正比的概率抽样法的话，任何这

种高速增长计数区都只有很小的入选机会。因此样本就会产生高得不能接受的抽样方差。

4.1.3. 地域标架

14. 我们将在本节和下一节讨论在抽样中使用的两类标架，而不论其用于住户调查还是别的用途。有必要提请注意，在多阶段设计中，必须把每个阶段的标架视为一个单独的成分。每个阶段的具体标架各不相同。住户调查的样本设计有可能既在初期阶段使用地域标架(本节讨论)，又在最后阶段使用清单标架(下节讨论)。

15. 在住户调查中，一个地域抽样标架是由一个国家按等级排序的各种地域单位构成的。这些地域单位在不同国家的行政上有不同的叫法，但是一般自上而下包括如下名称：省或县；行政区；人口普查区；地方选区和村(农村地区)或街区(城市地区)。为了人口普查的目的，行政分区还进一步划分为诸如船员领导人责任区和计数区之类的实体。普查计数区往往是一个国家中界定和划分的最小地域单位。

16. 为了调查的目的，对样本设计十分重要的地域单位有四个显著特征：

- (a) 它们通常覆盖一个国家的整个陆地区域；
- (b) 它们的边界清楚地划定；
- (c) 它们有人口数字记载；
- (d) 它们经过测绘。

17. 如前所述，覆盖整个国家的地理区域是十分重要的，因为它是实现真正概率样本的标准之一。在地图上明确划分的边界对于实施样本非常宝贵，因为这些界线表明了进行现场工作的地点。良好的边界信息也有助于访调员寻找最终被确定为访谈对象的样本住户。样本设计需要有人口数字来确定规模量度和计算选择概率。

18. 通常构建住户调查区域标架的起点是根据上述四个因素进行全国人口普查。此外，计数区是稍后抽样阶段(即两段设计中的倒数第二阶段)规模适当的地域单位。在大多数国家，有意识地建立包含住户数量大致相等(通常约100户左右)的计数区，目的在于为普查人员规定可比工作量。

19. 似乎自相矛盾的是，一个区域标架也是一个清单标架，因为我们必须从编列一个人口的行政区域单位清单入手来实施住户调查样本选择的早期阶段。这就需要讨论清单标架的问题。

4.1.4. 清单标架

20. 简单来说，清单抽样标架是由目标人口单位清单构成的标架。理论上，每个国家在完成了人口普查之后马上就有了一个住户调查清单标架。新近的普查原则上提供了按地域分布的全国每个住户(或居住单元)的登记清单。

21. 最近完成的普查清单是理想的住户抽样标架，因为就任何住户清单而言，它可能提供最时新、最完整和最准确的信息。由于普查清单的地域安排，所以按照适当的地域分布样本对其进行分层在做法上相当简单。因此，当需要进行一次普查后续样本调查以获取补充信息或者比普查所能有效提供的信息更详尽信息的时候，新近的普查清单在理想上就适合当做清单标架使用。可是有必要承认，新的普查清单只能在短暂的时期内充当时新标架。很明显，普查与后续调查之间相隔时间越久，普查清单作为标架来源的用途就越小。

22. 还有其他一些清单或可根据其质量被视为住户调查的适当抽样标架，比如公民登记和公用设施入户登记等。在一些保有公民及其住址详细记录的国家，公民登记可以作为候选标架使用。有些情况下它们可能比基于地区的普查标架更有用，因为这种登记可能不断地更新。每当一个国家的普查资料严重过时的时候，公用设施(通常指供电设施)的入户登记也许可以当做抽样标架使用，但是当然要经过评价以估量潜在的问题及其影响。一个导致覆盖不充分的明显问题就是不通电的住户；另一个需要弄清的问题就是存在多个住户共用供电线路的现象。

23. 另一种在发达国家广泛使用的清单标架是电话订户登记。通过随机数字拨号(RDD)方法进行抽样，以确保未公布电话号码的订户有适当的机会入选。可是在电话拥有普及率低的国家，我们不推荐使用通过随机数字拨号进行抽样的方法。

24. 在常规住户调查中，最后抽样阶段都不可避免地建基于清单标架概念。前面我们讨论了如何通过倒数第二阶段设计产生将所有住户汇编其中的类集样本。从这种清单中选择样本住户。这样，我们就有一个区域标架用以界定样本类集，又有一个清单标架来确定类集中的样本住户。

4.1.5. 多元标架

25. 第三章讨论了住户调查中的两阶段抽样，就是在第一阶段使用筛选法确定一个特定目标群体，接着在第二阶段采访由入选者构成的二段样本。另一种抽样方法可以产生大体差不多的最终结果，它需要使用不止一个抽样标架。通常这种方法只用两种标架，为此我们有双重标架设计；但偶尔也可能使用三个或三个以上的标架(多元标架设计)。譬如讲，可以从一个由好几个清单组合而成的人口标架选择一个独立的样本，在这种情况下每个次级标架构成一个层次(见第三章第3.4.2节和附件一中关于分层法的讨论)。不过，这种标架存在的一个共同问题是重复。

4.1.5.1. 住户调查中的典型双重标架

26. 为了陈述简便，我们来讨论双重标架设计，尽管相关的原则与多元标架设计原则很相似。通常，这种方法需要把一般人口区域标架和一个由作为研究对象的特定目标人口成员的已知个人构成的清单标架结合起来使用。譬如让我们来考虑一项打算研究失业者特征的调查。该调查可以建基于住户的区域标架，但或可辅以一个由已在社会服务部门登记的当前失业者构成的清单标架样本。这种双重标架样本的目的在于用具有很高概率列入目标人口的

人构成样本规模。这是一种可用以替代两阶段抽样法的成本较低而效果较好的办法。有必要使用通用住户标架来计量未被列入清单的目标人口。在这个例子中, 这些人是在没有在社会服务机构登记的失业者。

27. 可是, 双重标架设计有几个方面的局限性。一是清单标架必须基本上是现时的。如果从清单选择的人有很大一部分在状况上发生了变化, 以致使他们不再属于目标人口的话, 那么相关清单标架的用途就变得无效了。在我们举出的这个例子中, 任何失业者一旦在调查以前找到了工作, 他(她)就没有资格被列入清单了, 这个事实说明了为什么清单标架必须是最时新的。

28. 另一方面的局限性是如下事实造成的: 清单标架上的个人居住地点可能分散在整个社区, 因此由于交通费用问题使得对这些人的采访成本高昂。这当然会与基于区域的住户标架形成鲜明对比: 在后一种情况下可以从类集中选择样本, 从而降低采访成本。

29. 与双重标架设计有联系的一个严重问题就是重复现象。一般来说, 纳入清单标架的人也被列入了区域标架。在我们的例证中也是如此: 从登记中选择的失业者也是住户成员。因此, 在这两种标架同时使用的情况下, 他们就会有重复的入选机会。可以通过适当调整来处理重复问题。可是这就关系到调查问卷的内容了。在我们的实例中, 住户样本中每个被采访的失业者都要被问到是否在社会服务机构进行过失业登记。对做出肯定回答的人需要进一步做工作, 把他们的姓名和清单标架进行核对, 而这个过程容易产生误差, 且充满复杂因素。如果核对无误, 须将有关个人的调查权数改为 $(1/P_h + 1/P_l)$ 以反映如下事实: 他(她)有一个从住户标架被选中的概率, P_h , 和一个从清单标架被选中的概率, P_l 。请注意, 必须对照整个清单标架进行核对, 而不是仅仅对照标架上那些碰巧从样本选出的人组成的标架进行核对。这是因为概率(和权数)是不为实际选择所左右的选择机会的函数。

4.1.5.2. 不同类居住单元的多元标架

30. 如果目标人口居住在不同类、不重叠的居住单元的话, 所进行的抽样就是另一类双重标架抽样。譬如讲, 一项关于孤儿的调查很有可能被设计成包括生活在如下两种住房安排之一的孤儿: 一是像孤儿院之类的机构, 构成一种标架; 二是家有孤儿跟幸存的父亲或母亲、其他亲属或非亲属一起生活的住户, 构成另一种抽样标架。于是, 双重标架设计就包括了住户标架和机构标架, 而这当然不发生重叠。

31. 此类设计的目的在于尽可能充分覆盖目标人口(尽量接近100%)。在有大量人口分别生活在这两种居住单元之一的情况下, 如果样本仅限于两种标架之一, 就会发生严重偏差。比如, 一个仅以生活在住户中的孤儿为基础的孤儿样本, 不但会低估孤儿人口, 而且在估计其特征的时候也会发生偏差。仅以生活在机构中的孤儿为基础的调查也会发生类似的偏差。

32. 上面讨论的有关重复现象的局限性与基于不重复双重标架的设计无关。所以, 这后一种设计管理起来就容易多了。

4.1.6. 两段设计中的典型标架

33. 第三章强调了两段样本设计的使用价值。本节讨论两段设计中常用的标架。

34. 在抽样第一阶段选择的地域单位(或称类集)在农村地区一般被界定为村(或部分村)或普查计数区,在城市地区被界定为街区。然后就由研究全域中的所有地域单位构成标架,而不论这种全域是如何界定的——是整个国家,还是一个省或者一批省,抑或都会城市。抽样的做法是:汇编单位清单,检查其是否完整,以适当方式(一般按地域)对清单进行分层,然后(通常使用与规模成正比的概率)选择一个系统化的单位样本。

35. 前一章讨论过,如果全域中的类集档案非常庞大,可能需要在全域设立若干中间虚拟选择阶段。在这种情况下,每个虚拟阶段的标架单位定义就各不相同了。在前面关于孟加拉国的范例中,两个虚拟阶段的标架单位被界定为专区和乡。

36. 在两段设计中的第二阶段标架就是第一阶段样本类集中的住户。如果它们是从住户清单中取样的,那么根据定义标架就是一个清单标架。它们也可能是作为紧凑型区段被抽样的,这些区段是通过把类集进一步划分成细小且互不包容的地域部分而产生的。在这种情况下,第二阶段标架就是一个区域标架。

4.1.7. 主抽样标架

37. 这里,我们只是简单讲一讲主抽样标架的概念。下面第4.2节再详细讨论。

38. 主抽样标架是在调查中用以选择样本的标架,既可用于内容各不相同的综合调查,又可用于连续调查或阶段性调查不同轮次。阶段性的必要更新除外,抽样标架本身在各种调查之间或者在同一调查的不同轮次之间并无差异。所不同的是,要把主抽样标架设计并构建成为在一个相当长的时期内特定调查或同一调查的不同轮次用以选择二段样本所需要的稳定框架;这也是主抽样标架的显著特征。

4.1.8. 标架的普遍问题和推荐的纠正办法

39. 在住户调查中因为标架有缺陷所引起的问题包括非抽样偏差和抽样方差。前面暗示过,如果抽样标架过时了、不准确或不完整,就会带来一些共同的问题。在绝大多数国家通用调查中都把最近的人口普查当作基本标架,这也是本节所指的标架。基于普查的标架中往往同时存在过时、不准确和不完整的问题。普查和调查之间相隔时间越久,这些方面的问题越严重。

40. 如前所述,标架必须时新才能体现人口现状。举例来说,一个以五年前的普查为基础的标架不可能恰当反映人口增长和迁移状况。即便是现时的普查标架,如果它没有覆盖不是生活在传统住户条件下的军营、船上人家、游牧民和其他重要分组人口的话,也可能不完整并经由住户调查发生问题。不论时新的还是过时的标架,不准确的数据都会带来各种各样的问题,其中涉及到住户被重复列入清单的问题,遗漏住户的问题,或者住户被编入错误的计数区编号或编码的问题。

41. 能否恰当处理普查标架存在的过时、不准确或不完整问题的策略, 部分地取决于: (a) 调查目标和(b) 标架的新旧程度。在度量目标方面, 举例来说, 如果把调查有意识地设计为仅覆盖定居住户, 那么把游牧住户排除在外的普查标架就足够用了。另一方面, 如果调查打算涵盖游牧住户的话(在有游牧住户的国家), 那就应该制订相关程序以创建游牧住户标架。就此而言, 一个普查标架是否完整, 取决于行将被调查所涵盖的目标人口或分组人口的定义。

42. 纠正过时和不准确问题的具体办法因普查距今相隔时间长短而异。由于各国的国情千差万别, 虽然提出确切的规则或许尚需慎重考虑, 但是有一条经验法则或可用于指导修订或更新原有标架, 那就是确定有关的普查标架是否超过了两年。至于第40段提及的那类准确性问题, 可以采用第4.1.2和4.1.8小节提出的解决办法。

4.1.8.1. 两年以上的普查标架

43. 第一种情况适用于普查数据陈旧——在两年或两年以上——的国家。正是这些陈旧的标架在住户调查样本设计方面提出了最大的挑战, 尤其对快速发展的城市地区更是如此。理想的解决办法是把全国的陈旧普查标架全部更新, 因为要是能够成功实施的话, 它将确保由此获取的调查数据就调查覆盖面而言尽可能地准确而可靠。但遗憾的是, 这也是极为费钱又费时的工作, 因此是不切实际的。然而对于那些普查资料严重过时的国家来说, 恐怕也没有更好的替代办法。

44. 有一个不用全部更新的折中办法, 那就是只更新某些目标地区的标架, 这些地区是由熟悉国内增长模式和人口迁移情况的专家来确定的。普查标架的更新是相当简单的事: 所需要的只不过是当前的规模量度。为了标架更新的目的, 要把规模量度定义为居住单元数, 而不是住户数或人数。

45. 有必要承认, 抽样方法的有效性无需规模量度很精确。举例来说, 如果根据上一次普查认为一个特定计数区有122个住户, 那就完全没有必要关注它目前是否有115户或132户。因此, 试图在几十年来没有什么变化的既有的老邻居中更新标架是徒劳无益的, 尽管其中个别居民有进有出。值得关注的倒是与上一次普查相比而言现状的大幅度变化——比如说, 本来预期有100户而实际有250户。此类情况有可能发生在大规模开发或拆迁区附近, 比如城市郊区的棚户社区、高层建筑群或拆迁区等。这些区域将构成更新标架的目标区。要依靠国内合作者和专家来帮助鉴别目标区; 当然了, 只包括那些普查后发生变化的区域。

46. 更新标架的工作一般有如下几个步骤: (a) 确定构成目标区的计数区; (b) 快速清点有关的计数区以掌握当前的规模量度; (c) 修订普查档案以体现更新的规模量度。前面讲了, 只需近似的规模量度就够了, 这就是为什么要进行速查以摸清居住单元数而不是住户数的原因所在。在快速清点查核过程中无需敲门清点住所, 除非在多单元的住所里, 若不进入建筑物就看不清单元数。

47. 按照上述方式更新陈旧的普查标架，稳定倒数第二阶段单元的选择概率，从而提高调查估计数的可靠性是必要的。实际上，这种更新不但有助于控制样本规模，而且有助于控制现场工作人员进行登记造册和采访的工作量。另外，它还减少了在现场遭遇特大类集的可能性，实际发现该类集规模之大远远出乎意料，因而不得不对其进行二段抽样或采取某种别的适当行动。与后面这一点有关的是，二段抽样需要加权调整，从而给数据处理工作增加了复杂因素。这种可能性会减少到这样的地步：经过倒数第二阶段抽样之后就不会再遇到大得出乎意料的类集了。

48. 住户调查的标准样本设计可能需要汇编样本类集中的现时住户列表。在这种情况下，倒数第二抽样阶段也可以进行标架更新了(见下节关于与估计规模成正比的概率抽样的论述，此法在这里也适用)。因此，样本类集中未经更新的现时列表可能非常接近于普查清单(但并非肯定如此)。不过可以预料，根据普查标架的更新部分形成的样本类集将产生与普查清单大不相同的现时清单——不论就住户总数而言还是就其特定识别特征而言。

49. 关于使用陈旧普查的问题最后需要强调一点，就是这个问题关系到样本有效性而不是样本方差。如前所述，类集通常是采用与规模成正比的概率抽样法选择的。如果没有在选择样本之前更新高增长类集的规模量度，就会导致那些在普查时曾经住户很少但后来大幅度增长地区的严重代表性不足。调查结果将出现偏差并且当然会产生误导作用，因为生活在这些高增长地区的人在特征上可能与居住在较稳定地区的人会有很大的不同。

4.1.8.2 两年或两年以下的普查标架

50. 本节适用于在以往一两年内刚刚进行过人口普查的国家，它们不需要从总体上更新标架。在这种情况下可以使用原有的普查规模量度来选择类集，预料这会相当准确。只是在倒数第二抽样阶段，当现场工作人员对样本类集中的当前住户进行登记列表的时候，标架本身会进行更新。将从现时列表选定样本住户，并依照第3.7.2节讨论的与估计规模成正比的概率抽样方法对抽样权数加以必要的调整。

51. 虽然在标架全域中有些类集可能在完成普查后有了很大发展，但是预料这种案例的数量未必会多到严重影响现场运作或调查精度的地步。任何碰巧被选入样本的此类类集都可以在必要时进行再分段。再分段，或称“组块”，是旨在减少登记列表工作负荷的一种现场操作方法。这种方法涉及到：(a) 把原有的类集分割成若干部分，通常分为四部分；(b) 随机选择一部分以供列表；(c) 从该区段选择拟采访的住户。再分段不会改进抽样可靠性，因为每个样本块都携带一个相等于类集内块数的附加调查加权因子——如果类集分为四部分，那么该加权因子即为4。可是，再分段的确有助于节制现场运作成本。即便普查是最近完成的，也可能需要在普查以来发生了根本变化的高增长计数区再进行组块作业。当然，如果是刚刚搞完普查的话，这种区域会非常少。

52. 应着重指出的是，前面讲到的各种不精确的问题(比如住户的重复或缺漏，错误的计数区归属，等等)可以在更新标架的时候得到部分纠正，从而导致在倒数第二阶段产生时新的住户列表。这当然是在调查中获取时新住户列表的另一重要原因。

4.1.8.3. 标架被移作他用的情况

53. 调查管理者有时对专为一种调查构建的住户标架能否用于其他调查没有把握。举例来说, 一个劳动力调查方面的抽样标架能否用于样本设计来计量健康、残疾、贫困或农业财产等方面的状况? 不过, 一般情况下不是标架本身有问题, 而是它的分层方法成问题。在大多数情况下标架可以用于不同的调查, 除非它们对于所要进行的调查不够完整、不够精确, 而且过时了。关于不能满足多用途需要的标架, 请见下面的范例。举例来说, 如果一项集中在生活费用方面的调查仅以城市社区为基础(实际情况往往如此), 那么抽样标架就会把农村地区排除在外。显然, 这样一个标架就不适合在贫困基本上以农村为主的国家用来估计国内贫困状况。

54. 可是, 大多数住户调查不论就其内容还是就其样本设计而言都具普通用途。举例来说, 一项劳动力调查通常包含有关人口统计、教育程度及其他话题的辅助信息。在这种情况下, 相应的样本设计也是通用性质的, 就是说, 使用惯常抽样标架——一个覆盖全国所有住户的标架。可以根据专用于计量劳动力的一个变量对标架进行分层。比如, 或可根据最近一次普查中的失业百分比这个变量来对计数区进行分类。然后分别创建高、中、低三个层次就业率的计数区。如上所述, 这是个分层决策, 而标架本身不受影响。如果想把这个标架用到另一种调查——比如健康调查, 那解决方案或许是对该标架“不分层”。

55. 抽样统计工作者的一项至关重要的任务, 就是对准备用于另一种调查的现行抽样标架进行评估。这种评估要确保既有的标架能够满足拟议调查的度量目标需要, 特别要根据(贯穿于本章始末的)完整、准确和时新三原则来进行评估。

4.2. 主抽样标架

56. 主样本可以做到不但有成本效益、花费值得, 而且投资效益好, 但条件是: 一个国家要有足够数量的专项调查或同一调查的阶段不同轮次来维持这种样本的使用。或许不言而喻的是它们必须设计完善, 但长期适当维持使用也十分重要。对此, 联合国的有关文件(见联合国, 1986年)就主抽样标架及其使用问题提供了非常全面的处理办法。

4.2.1. 主样本的定义和使用

57. 在住户调查中, 第一抽样阶段的抽样标架(或称标架)必须涵盖整个目标人口。当这种标架被用于多目标调查或同一调查的多轮次的时候, 它就叫做主抽样标架或简称主样本。

58. 对任何拥有普查间大规模连续住户调查方案的国家来说, 使用主抽样标架都是首选策略。反之, 如果没有连续调查方案, 则一般不推荐使用主样本。有一些规模经济长期使用同样的标架单位, 因为大部分抽样成本发生在构建主标架的作业过程而不是每次现场实施调查的过程。另一方面, 对于那些只是偶尔在两次普查之间进行全国调查的国家来说, 不会从利用主样本设计中获得多么大的实惠。

59. 主样本的特征包括第一选择阶段的单位数量、规模和种类。一般来说，主样本包括初级抽样单位(PSUs)的初步选择，而初级抽样单位对每个二段样本都保持固定不变。请注意后期阶段通常是可变的。譬如讲，在最后选择阶段被选中接受采访的特定住户通常不同于作为独立调查对象的住户，而在各次重复调查中，住户是相同的或部分重叠的。

4.2.2. 主抽样标架初级抽样单位的理想特征

60. 有关构建主抽样标架的准则与一般抽样标架的准则几乎没有什么区别。主标架应尽可能完整、准确和时新。住户调查的主抽样标架正如通常抽样标架那样，一般都是在最近普查的基础上开发的。然而，由于主标架可能在整个普查间隔期内使用，所以通常它会需要阶段性的定期更新，比如每两三年更新一次。这一点恰好与通常的标架形成鲜明对比：后者多半在临时的基础上并且只是在筹划一次专项调查的时候才进行更新。

61. 可以预言，适合开发主标架的特征总的来说同抽样标架差不多。譬如讲，拟用作初级抽样单位的单位定义需要满足的限制条件是，它们应该是业经测绘的地域单位。不过这并不是很严厉的限制，因为标架单位将不可避免地被界定为行政区域单位，而这种单位已经为普查目的构建好了。不过有一项重要的要求有别于普通抽样标架，那就是，初级抽样单位的规模必须足够大才能满足多目标调查的需要，而不必反复采访同一受访者；但就连这一特征在某些实际应用中也是可以放松要求的。

范 例

一个已在某些环境中使用的特定种类主标架系基于两段设计。第一阶段涉及一个计数区大样本(或类似的在地图上标明的小地域单位)。从使用该标架的每个独立调查中选择一个主计数区样本的二段样本(或称子样本)。对每个二段样本进行列表或再分段，以便在积极计划调查工作的时候供其使用。进一步说明：主样本可以由10 000个计数区组成，其中，为进行就业调查对1 000个计数区进行了二段抽样。在1 000个计数区进行了住户登记造册，在每个计数区为调查选取15个住户组成第二阶段样本。第二年，又从主样本中选取了由800个计数区组成的另一个二段样本，以便用于健康调查，等等。采取这种方式，没有任何计数区被使用超过一次，因此与初级抽样单位的规模无关。

62. 可是，如果从主标架产生的所有二段样本都必须出自同一套初级抽样单位的话，那么初级抽样单位的规模就十分重要了。在上面的范例中，计数区是初级抽样单位，而每个二段样本使用一个不同的计数区子集。在主样本中选择初级抽样单位不算特殊问题，因为无论对主样本还是对任何其他样本来说，所采用的方法都是一样的。一般都使用与规模成正比的概率(*pps*)方法，但初级抽样单位规模大体相等的罕见情况除外；在后一种情况下可以使用一种等概率初级抽样单位样本。

4.2.3. 用主样本支持调查

63. 第3.3.7节讨论了为什么主样本需要大样本才能提供足够的住户，以便在好几年内支持多目的调查而无需对相同的受访者反复进行采访。对有可能利用主抽样标架的所有拟

议的和潜在的调查来说, 预期的样本规模是其构架设计中的关键参数。举例来说, 如果预计在使用主样本的各种调查中对50 000个住户进行采访, 那么抽样工作队就必须掌握所需要的基本信息资料才能决定初级抽样单位的数量和规模。另外, 可以结合使用主样本, 制订一个如下例所示的调查实施计划(另见第3.3.7节所作的说明, 以便比较)。

范 例

跟第3.3.7节的范例一样, A国的主样本有50 000个住户。计划的三次调查将有如下样本和类集规模: 收支调查, 16 000个住户的样本, 每个类集有6户; 劳动力调查, 12 000个住户的样本, 每个类集有12户; 以及健康调查, 10 000个住户的样本, 每个类集有20户。选用不同的类集规模, 是为了应付(不同类调查的)不同设计效果。此外还持有12 000个住户的样本, 以备必要时供其他调查使用。计划内的三次调查总共需要4 167个初级抽样单位($16\,000/6 + 12\,000/12 + 10\,000/20$)。由于或许使用这种保留二段样本的调查内容不得而知, 所以决定基于12户类集规模制订计划, 这就增加了1 000个初级抽样单位, 总数增至5 167个。因此主样本设计团队决定构建一个由5 200个初级抽样单位组成的主样本。初级抽样单位的定义必须考虑到拟采访的住户数。在本例证中, 每个初级抽样单位都要达到足以产生50个受访住户。有了这种信息, 抽样团队即可确定哪些地域单位最适合界定初级抽样单位。如果A国平均每个计数区有100个住户, 且围绕这一平均数差异很小, 那就可以使用计数区作为初级抽样单位。

4.2.3.1. 主抽样标架的多用途优点

64. 使用主样本有明显优越性。首要的一点, 就是可以把主样本计划当作国家统计方案项下的一系列政府部门和其他利益攸关者之间的协调工具。这个优点体现在调查事件的诸多方面, 而不仅限于抽样本身的考虑, 主要表现在控制成本方面以及制订各种跨部门的标准程序, 诸如各项统计定义、调查问题措词、数据编码程序, 等等。

65. 主样本方案的一个关键优点就是利用同样的初级抽样单位。可以在主样本的生命周期之内组建并且维持一个现场工作班子。比如, 可以雇用和培训访调员, 他们在主样本方案一开始就投入工作, 知道各个初级抽样单位的位置在哪里; 主标架可供所有调查使用多年, 譬如讲10年。在必要的情况下, 可以在主样本的初级抽样单位范围内或附近雇用本地访调员。像初级抽样单位地图和住户名册之类的调查材料可以在主样本方案开始的时候制做出来, 以便节省时间和分期注销所有预期调查的一大部分活动经费。此外, 样本的多用途提供了更有效控制非抽样误差甚至不答复的机会。这是因为, 反复访问相同的受访者有助于记录这些受访者的态度, 同时注意到各方面的问题, 以便制订日后调查的纠正措施。不过, 这里也有必要强调指出: 只有在主样本频繁使用的情况下, 这些优点才能得以发挥。

66. 总的来说, 主样本的其他优点——无论是在三段设计中使用相同的初级抽样单位还是在两段调查中使用不同的初级抽样单位, 都包括如下潜在可能性: (a) 在分析的基础上, 对出自两个或两个以上不同内容主样本用途的数据进行整合; (b) 对意料之外的收集数据需求做出快速反应。

4.2.3.2. 多用途主抽样标架的局限性

67. 主样本可能存在一些局限性，比如初级抽样单位枯竭，就是说，假如过度使用主标架，住户就会被耗尽。不过，这个问题可以通过完善的事先计划予以防范。虽然不可能预见主样本在其整个生命周期的所有用途，但只要主样本够大，就有可能给保留二段子样本派上用场。

68. 另一局限性是：如果不随着主样本老化对其进行适当更新，它所产生的偏差就会变得越来越大。最后一点，主样本不大适合提供“有特殊要求的”调查数据，比如可能引起关注的特定省份或稀少分组人口的调查。

4.2.4. 域间(政区间, 等等)的分配

69. 在国内重要行政分区(比如大区、省和大城市)住户调查数据的列表和分析方面，各国的统计部门正在面临越来越大的压力。有些国家(比如越南)甚至预期在地区一级例行提供数据。这些要求和期望受到合法政策需求的驱动，一般理由是：社会经济方案一般聚焦于地方一级而不是全国一级，并且是为地方发展而制定的。

70. 如前所述，按照统计抽样的说法这些是域的估计数。由于取得可靠调查结果所必要的样本规模十分庞大，需要付出的代价如此之高，以至于往往超过政府通常所能承受的调查预算。域数据的需求也影响到主抽样标架的发展。

71. 第3.3.4节关于样本规模的讨论考虑了行将建立的域的数量和种类，这里就不重复讨论了。一俟做出抉择，即可构建主抽样标架了。比如说，一个国家可能决定在其主样本方案项下进行的调查仅提供两种域——城市和农村——的数据。另一国家有12个省份，它希望提供每个省的估计数。它断定，如果把这些省当作域对待，那么它的调查资源可以支撑所需的额外样本规模。第三个国家有50个省份，它可能认为如果为每个省提供估计数花费太大。所以它可能决定把50个省划分为8大行政区，而将这些大区分别界定为域。第四个国家可能决定不建立域，而是按地区、省、城市、农村以及选定的大城市，按比例分配其国家级的样本并列成表格，以期发表这些次区域的公共数据。据认为这些次区域的样本规模大到足以提供相当可靠的调查结果。

72. 就以上各例而言，域的分配无关紧要，因为样本是按比例分配给有关的国内各分区的。对前三个例子来说，须采取特别步骤来适当分配主样本的初级抽样单位。既然域的估算意味着每个被界定为域的分组人口群体或地区都有同等的可靠性，那就应该从每个域选择同等数量的样本初级抽样单位——无论样本设计是否基于主样本，此项要求都适用。

4.2.5. 主样本的维持与更新

73. 就其对人口覆盖面的影响而言，适当维持一个主抽样标架是这种标架的发展及其使用计划方面的一个关键要素。一个特定国家的主样本通常在两次普查间隔期间使用10年，在此期间可能会发生深刻的人口变迁。有必要定期更新标架，以反映人口变化，从而使之继续具有“代表性”。

74. 有两种更新十分重要。第一种比较简单, 它需要编列在倒数第二阶段选出的样本类集住户的新清单, 本手册从头到尾都在推荐此种方法, 不论是用于主样本还是用于单一用途样本设计。这样, 样本类集经过更新就能自动反映迁移、出生和死亡等状况了。这种仅限于样本类集的更新有助于最大限度地减少覆盖面(非抽样)误差, 但是除非整个标架都更新了, 否则抽样方差就会随着时间推移增加。

75. 也有必要对整个标架进行阶段性更新, 以便适当说明上次普查以来的大规模人口增长情况。如前所述, 这种人口增长是在高层住宅建设和市郊棚户区扩展的情况下发生的。在构建主抽样标架的时候, 发生这种高增长情况的计数区就不可避免地小多了。结果, 它们的规模量度就随着增长的发展而被严重低估了, 因而大大减少了它们在与规模成正比的概率设计中被选中的机会。一旦碰巧选中了这种计数区, 由此对抽样方差产生很大的影响, 因为当前的规模量度可能要比原来大许多数量级。

76. 通过定期修订标架(譬如讲每两三年更新一次), 可以显著减少上述高增长区的问题及其对主样本的影响。

4.2.6. 主样本中初级抽样单位的轮换

77. 关于样本重叠问题的详细讨论, 请参见第三章第3.9.2节, 题为“旨在估计变化或趋势的抽样”。这个问题与旨在估量变化或趋势的反复调查或连续调查有联系。所谓重叠的样本, 意思就是在重复进行调查的时候利用置换住户的样本方案。这里有必要再次强调, 对于旨在估量一两年变化的调查来说, 使用重叠样本当属首选方法。一种取代方法就是样本轮换, 亦即不同的调查或不同的调查轮次之间有部分住户重叠。

78. 上一节指出, 在每个调查轮次使用相同住户的情况下, 无论抽样可靠性(合乎需要)还是非抽样误差(不合需要)的几率都是最大的。因此, 一般都是寻求一种折中办法, 即在不同调查轮次之间使用部分重叠样本, 尤其在一项调查反复进行三次或三次以上的情况下(关于部分重叠的基本原理, 见第3.9.2节)。

79. 引进部分重叠样本的一个方法, 就是置换或轮换样本的初级抽样单位(而不是置换样本初级抽样单位里面的住户)。如果一个初级抽样单位主样本不但用于同一调查的不同轮次, 而且用于多种用途的调查, 也同样必须仔细考虑轮换这些初级抽样单位。

80. 为了使轮换计划便于实施并取得有价值的成果, 时段之间的重叠程度要保持一致并且经久不变。举例来说, 如果第1、2年之间的重叠率是K%, 那么第2、3年、第3、4年……之间的重叠率也应该是K%。当全部初级抽样单位轮换一遍之后, 这一特征需要相应地纳入轮换设计。

4.2.6.1. 关于主样本的国家范例

81. 本节介绍四个发展中国家的主样本, 它们是: 柬埔寨、阿拉伯联合酋长国、越南和莫桑比克。每个国家的范例用来解释主样本的一些特点和原则, 比如主样本单位的一段或两段抽样, 以及本章讨论的特别调查的灵活应用。此外还解释了本手册所强调的其他样本设

计特征，其中除其他外，包括隐含分层，旨在减少设计效果类集规模的最佳选择，以及域的样本分配等。

4.2.6.2. 柬埔寨，1998-1999年

82. 柬埔寨的主样本说明了两段设计的用途。初级抽样单位大样本提供了经过再抽样的一个第二阶段区段主清单，以供专项调查使用。

83. 柬埔寨国家统计局于1999年开发了一个主样本，供政府的普查间隔期住户调查方案使用。该方案包括一项阶段性社会经济调查，以及潜在的健康调查、劳动力调查、收支调查、人口调查和其他临时调查。1997年的人口普查充当了主样本设计标架，调查分两阶段进行。第一阶段是用与规模成正比的概率选择被界定为村的初级抽样单位；规模量度是普查住户计数。初级抽样单位的选择是通过计算机操作进行的。第二阶段是在选定的初级抽样单位内创建若干区段，系手工操作。

84. 之所以决定使用村作为初级抽样单位，是因为这些单位都比较大，有足够的住户（在城市地区平均有245户，农村地区平均有155户），可供普查间隔期间进行好几次调查。这样就免除了反复采访相同受访者的负担。曾经考虑过使用普查计数区，但是放弃了，因为平均起来每个计数区只相当于村子规模的一半。样本中不包括暂住人口或生活在机构中的人口，也不包括军营人口。

85. 总共在主样本中选择了600个初级抽样单位，因为据认为这个数字足以通过全国的分布充分代表所有省份。在选择样本中使用了隐含分层法，通过将村一级的档案按照地域次序——即按省、地区和社区所辖城乡地区——进行分类来实施。于是，主样本便按比例自动分配到各个省份的城市和农村地区。

86. 柬埔寨主样本的一个引人瞩目的特征就是第二阶段抽样作业。前面讲到，这一阶段需要在每个选定的初级抽样单位范围内创建一些区段。应该指出的是，第二阶段主样本建设的概念不要和属于特定调查住户选择的第二阶段样本选择相混淆。作为一项办公室的文书工作，在每个选定的主样本初级抽样单位中按照（平均）10个住户的规模划分区段。在这方面除个别情况外不需要进行任何现场工作，因为使用了1997年的普查登记造册和现有的区划草图。在每个主样本初级抽样单位创建的区段数按普查住户数除以10计算，四舍五入凑整数。举例来说，一个村子截至1997年普查前共有187个住户，故划分成19个区段。

87. 如此创建的区段构成行将结合特定调查进行抽样或再抽样的建筑地段。使用主样本的每项调查或每个调查轮次从全部初级抽样单位或其中一个子集中选择一个或多个区段。按照前述说明创建主样本的一个重要好处，是依样本设计的具体细节而定，为使用主样本的每项特定调查提供了自我加重的机会。

88. 主样本设计的一个关键优点，就是在如何进行二段抽样以供特定调查使用方面有很大灵活性。每次调查的类集选择（亦即区段的选择）会如愿产生一个不同的集。典型的初级抽样单位包含18-30个区段，在每个初级调查单位中提供一定数量的区段，足以维持所有调

查。另外, 每年重复一次的社会经济调查可以每年使用一套不同的区段。通过逐年轮换保留一部分区段(譬如讲25%)可以实现重叠使用样本的目的, 这样每年可以置换75%的区段。

89. 主样本设计的一个局限性是因为使用紧凑型类集(即样本区段中的所有住户都相互毗邻)造成的。相较于非紧凑型区段——亦即较大类集中的一个系统住户样本——而言, 这种局限性或多或少增加了设计效果, 因为据认为通过限制类集规模, 可以在一定程度上减少设计效果。所以把区段规模限制在只有10个住户。

90. 预计每两三年更新一次样本。虽然承认最好更新整个主样本, 但还是决定为了支持当时正在计划的特定样本调查只对初级抽样单位进行更新。更新工作包括进行现场访问, 以便在受影响的区段编列新的住户清单。对包含原有住户集的这些区段的同一地区进行了重新登记列表, 这就是为什么区段边界如此重要的一个原因。

91. 值得注意的是, 在柬埔寨的主样本设计方面, 更新工作谋求与村长的合作。据了解, 他们精心保管本村所有住户的登记表, 并且日常保持登记随时更新。大多数情况下他们的表册都相当准确。此外, 村长们还是辨认和寻找适合于任何特定区段地域的宝贵信息来源。

4.2.6.3. 阿拉伯联合酋长国, 1999年

92. 阿拉伯联合酋长国的主样本体现了两个重要设计特征。首先, 主样本设计运用了特殊分层法来处理阿拉伯联合酋长国的两种不同人口——公民和非公民。其次, 该设计说明了如何利用标准区段设计(见第3.8.2节)来处理变化多端的计数区规模和陈旧的普查资料的问题。

93. 据阿拉伯联合酋长国计划部描述, 该国主样本系以1995年人口普查为基础并以此作为抽样标架由500个初级抽样单位构成的超级样本。该国想要利用这个主样本进行特定住户调查, 直到下一次人口普查。初级抽样单位被界定为普查计数区或其中部分计数区, 这样, 平均每个初级抽样单位有60个住户——既有公民又有非公民。

94. 在选择初级抽样单位之前构建了两个层次。第一层由计数区构成, 其中三分之一或三分之一以上的住户在普查时是公民住户。第二层由所有其他计数区组成。总共对第一层的1 686个计数区和第二层的2 986个计数区进行了分类。采用系统化的与规模成正比的概率抽样法, 两层各提取一个由250个初级抽样单位组成的样本, 全国共计选择500个初级抽样单位。预期这个主样本将产生数量上大体相等的公民住户和非公民住户。首先对大的初级抽样单位(90个住户以上)进行分段, 并且在每个大的初级抽样单位随机选取一个区段。在这500个初级抽样单位进行一次新的当前住户登记造册, 以更新标架。这次列表工作将大约30 000个住户纳入样本初级抽样单位, 以便作为各种组合供特定调查使用。为了便于灵活应用, 把每个样本初级抽样单位中的住户划分为12个子集, 预期每个子集5户。

95. 主样本的一个值得注意的特征是, 它不是自我加权的, 因为这两个层次规模不等。首次使用主样本进行的调查是1999年卫生部主办的全国糖尿病调查。预期实施的其他调查包括一次劳动力调查和一次收支调查(或者叫做家庭预算调查)。

96. 下面较为详细地介绍一下决定了阿拉伯联合酋长国如何建设其主样本设计的某些特殊情况。有两个方面的考虑经过仔细思考摆在了头等重要的位置，这就是目标人口和抽样标架的问题。

97. 如前所述，该国有两个重要目标人口——即公民和非公民。根据1995年的人口普查，虽然前者约占总人口的43%，但是仅占全国住户的四分之一左右。这是因为就每个住户的人口而言非公民住户数量少多了。对样本设计来说这就意味着，如果按比例选择全国住户样本，就会有将近四分之三的答复住户属于非公民住户。上述情况的进一步含义是，由此产生的非公民住户的调查估计数可靠性大约要相当于公民住户的三倍。鉴于调查结果要被用来制订政策和筹划方案，因此认为在估计数的可靠性上存在如此之大的差异是不可取的或者说无益的。就样本设计而言，解决办法就是应用上面所讲的适当分层法，把这两个不均衡不平等的目标人口分别当作单独的实体来对待。

98. 同时，还使用了另一等级的分层，即地域分层法，以确保样本分别按照酋长国和城乡的适当分布。在选择之前，计数区档案按照如下次序进行分类：首先是公民层，按城市分类；而在城市中，按酋长国分类；而在酋长国内，按照自下而上的公民百分比顺序安排的计数区编码进行分类；接下来是农村、酋长国和计数区编码；然后是非公民层次，按照同样的次序分类。

99. 据认为，主抽样标架的一个关键特征是必须有一套清晰的地图，上面勾画了被指定为样本区即初级抽样单位的地域单位。地域单位必须小到便于列入清单，但同时又必须大到能够作为自然边界清楚地加以界定(即容易寻找)。据认为，普查计数区是能够满足这双重标准的唯一可行的地域单位。但遗憾的是在人口普查中没有使用地图，因此现有计数区的边界没有清楚地加以界定。所以有必要明确提供主样本初级抽样单位(计数区)的边界信息。

100. 初级抽样单位主样本的准备采用了第三章描述的“标准区段设计”。这个方法已在许多国家用于人口和健康调查方案，还有泛阿拉伯母婴健康调查(PAPCHILD)。

101. 之所以决定使用标准区段设计，是因为阿拉伯联合酋长国的普查计数区在规模上参差不齐，差别很大。建立了大约60个住户的标准区段。每个样本初级抽样单位都计算了标准区段数，即住户总数(公民加非公民)除以60，所得商数四舍五入凑整数。

102. 区段数(即规模量度)为两个或两个以上的情况下，初级抽样单位被分割成地域区段。这需要现场工作，实地访问计数区(初级抽样单位)，并通过快速清点和图上标出住所(不是住户)的位置来绘制一份草图。分段以后，从每个初级抽样单位随机选取一个区段。该区段就成为主样本中实际进行抽样的地域。另一次实地访问是为了获取每个样本区段的现时住户清单。后面的程序是为了更新已使用三年之久的主抽样标架，后来被认为是抽样工作的重要组成部分。

103. 主样本的最后一步工作是把每个样本区段的新列表住户划分为12个系统子集或群组。用一个或多个群组进行特定调查。既然区段的平均规模为60个住户，所以每个群组平均有5个住户。

104. 之所以规定划分12个群组, 主要是因为这个数目在调查中组合使用较为灵活。在特定调查中的实际选择取决于各种因素, 其中包括调查目标、所期望的类集规模以及调查所需的总体样本规模。例如, 将有五分之二的首级抽样单位被用于全国糖尿病调查。于是这些首级抽样单位内的12个住户群组中有4个包含进来。这种组合产生了一项涉及200个首级抽样单位、类集规模为20户(即 4×5 个住户)、总样本规模约为4 000个住户的总体样本计划。

4.2.6.4. 越南, 2001年

105. 下面介绍越南主样本的两个突出特点。一是说明采取的主样本两阶段抽样法和利用第三阶段实施特定调查。二是说明如何向地理区域分配主样本。

106. 基于1999年普查并以此作抽样标架的主样本采取了两段设计。把首级抽样单位界定为农村地区的公社和城市地区的市内行政区。如此界定的原因是已经决定, 主样本的每个首级抽样单位至少要有300个住户。也曾考虑把计数区当作首级抽样单位, 但是它们太小了, 必须与相邻计数区相结合才能符合作为首级抽样单位的要求。据认为后面这个选项太费时费力。另一方面, 因为规模小而需要组合的公社或市政区的数量, 在10 000多个里面只有529个。

107. 采取与规模成正比的概率抽样法总共为主样本选择了3 000个首级抽样单位。平均每个样本首级抽样单位在城市地区有25个计数区, 在农村地区有14个计数区。在第二阶段抽样使用与规模成正比的概率, 每个样本首级抽样单位选择3个计数区。根据1999年普查, 第二阶段的单位平均每个计数区约有100个住户: 城市地区平均有105户, 农村地区平均有99户。

108. 主样本的目标之一, 就是能够为越南的8大行政区各提供相当可靠的数据。每个省独立进行样本选择。这样, 各省便充当了主样本的层次。期望在人口很少的个别小省份额外多选一些样本。因此采取了与有关省份的规模平方根成正比的概率方法, 在各省之间分配样本。同时采取了城乡之间按比例分配的办法。

109. 除了上面讲的省一级的分层之外, 还采用了省内隐含地域分层法。在将主样本应用于具体调查方面, 要使用计数区子集——例如其中三分之一用于综合住户调查。在调查应用方面, 实施了一个第三抽样阶段, 其中, 要从每个样本计数区选择固定数量的住户。这个住户数因具体调查和城乡不同地区而异。举例来讲, 对于农村计数区可以每个计数区选择20个住户; 每个城市计数区可选择10个住户。

4.2.6.5. 莫桑比克, 1998-1999年

110. 莫桑比克的主样本案例说明了一种单一阶段首级抽样单位选择方法, 它适用于政府主办的国家普查间隔期内住户调查方案项下的所有全国调查。它还说明了一个灵活的主样本如何能经过调整满足特定调查度量目标的需要。

111. 莫桑比克的主样本首级抽样单位直截了当地按照本手册各部分所描述的方法进行界定和抽样。以1997年的人口普查为抽样标架并据此构建了首级抽样单位。这些首级抽样

单位一般由3-7个普查计数区构成地域分组，平均每个计数区约有100个住户。采用与规模成正比的概率法选择主样本初级抽样单位。

112. 总共选择了1 511个初级抽样单位，为莫桑比克综合住户调查体系提供了实施抽样的架构。将主样本初级抽样单位划分成群组，每个群组构成一个系统子集，从而凭其本身的条件成为一个概率样本。共有10个这种群组，每个群组有151个初级抽样单位。在2000-2004年的五年计划期间，世界银行拟订的《核心福利指标问卷》调查是首次使用该主样本的调查。

113. 《核心福利指标问卷》样本计划的设计牢记两项计量目标。首先是获取勾勒莫桑比克个人和家庭幸福状况所需要的相关指标。其次是分别提供这些指数在国家一级、城市和农村地区以及全国11个省份的可靠估计数。《核心福利指标问卷》的抽样方法是利用莫桑比克的主样本在分层、类集的设计中选取14 500个住户。因此，第一阶段当然就是选择主样本的初级抽样单位了。

114. 《核心福利指标问卷》的第二阶段是选择主样本初级抽样单位的一个二段样本。对主样本中的1 511个初级抽样单位进行了再抽样，总共选定了675个初级抽样单位。这些单位是采用等概率方法系统选出的，并且在莫桑比克的11个省份均等分配。第三阶段在每个初级抽样单位为《核心福利指标问卷》选出一个计数区样本。于是在《问卷》样本中就有了675个类集——农村有475个，城市有200个。计数区是用等概率选出的，因为它们各自的规模大体相等——前面讲了，平均100户左右，尽管有多有少。最后选样阶段是在国家统计局访问类集、汇编新的住户清单以更新1997年抽样标架之后进行的。从新汇编的清单中选出了一个系统样本——含农村住户20个，城市住户25个——供《问卷》调查采访之用。因此，《问卷》的样本选择是一个四段选择过程，不过它所依据的主样本是个单一阶段。

115. 《核心福利指标问卷》设计的两个特征说明主样本的灵活性：它可以经过调整适应调查应用的特定要求。

116. 首先，在将主样本用于《核心福利指标问卷》方面，国家统计局会有兴趣利用上述已指定的群组。鉴于预期《问卷》大约有600个初级抽样单位，所以希望能够使用其中4个群组。可是当意识到对每个省份来说所需群组数不尽相同的时候，便放弃了这个想法，因为各省的计量目标需要差不多一样的样本规模。结果从整个主样本档案中选择了《问卷》所需要的675个初级抽样单位，而没有考虑使用群组。这就背离了主样本的本来意图，即按比例提供各省的样本。之所以如此，是因为在省一级都需要同样的《问卷》可靠性，这与主样本的原始设计大相径庭。在主样本计划中，本来设想先产生国家级的估计数。

117. 关于《问卷》样本设计的第二个重要问题是类集的规模。一致同意城乡住户的类集规模应有所区别，理由是农村地区的样本设计效果，或称 $deff$ ，比城市地区高，因为农村的经济生活大多出自生计农业。换言之，农村地区的住户很可能具有十分相似的特征。主样本提供了在最后阶段城乡选择不同固定数量的住户（即城市地区和农村地区分别选取25户和20户）的可能性。

4.3. 准则概要

118. 本节总结了本章所讨论的各项主要准则。跟第三章一样，现以对照表的形式将其作为经验法则而不是刻板建议概括如下：

- 尽可能使用完整、准确而且最新的抽样标架；
- 确保抽样标架覆盖所期望的目标人口；
- 在住户调查中尽可能使用最近的普查做标架；
- 按照地域单位(比如普查计数区)用地图上标出的清晰边界来界定标架中的初级抽样单位，并且注明其人口数；
- 只有在最近完成普查(一般不超过一年)的情况下，才在最后阶段使用普查住户清单作标架；
- 在确保防止重复的程序到位的情况下，谨慎使用双重或多元标架；
- 在全国范围或在已知高增长的特定目标区域更新两年以上的普查标架，以便：
 - 通过快速清点核查更新陈旧标架；
 - 在样本类集通过重新编列住户清单来更新标架；
- 如果普查标架没超过2年，就只更新样本类集，以便：
 - 通过重新编列住户清单来进行更新；
- 只有在计划了或正在执行大规模连续调查方案的情况下才使用主样本或主抽样标架；
- 界定足以在普查间隔期内维持许多调查或重复调查轮次的规模够大或数量够多的主样本初级抽样单位；
- 使用上面推荐的相同准则，为单一调查标架更新主抽样标架；
- 在使用主样本重复进行调查的时候实行样本轮换制度——要么轮换住户，要么轮换初级抽样单位。

参考资料和补充读物

Cochran, W. G. (1977年)。《抽样方法》(第3版)。纽约：威利国际出版公司。

Hansen, M. H., W. N. Hurwitz和W. G. Madow(1953年)。《抽样调查方法和理论》。纽约：威利国际出版公司。

国际统计学会(1975年)。《样本设计手册》。世界生育率调查基本文献。荷兰福尔堡。

Kalton, G. (1983年)。《调查抽样导论》。加利福尼亚州贝弗利山：赛奇出版社。

Kish, L. (1965年)。《调查抽样》。纽约：威利国际出版公司。

- 阿拉伯国家联盟(1990年)。《阿拉伯母婴健康调查抽样手册》，基本文献，第5号，开罗：泛阿拉伯儿童发展项目。
- 宏大国际咨询公司(1996年)。《抽样手册》，第三次人口和健康调查基本文献第6号。马里兰州卡尔夫顿：宏大国际咨询公司。
- Pettersson, Hans (2001年)。《特派团报告：关于越南住户调查主样本设计的建议》(未出版)。统计局，河内。11月25日。
- _____ (2005年)。“发展中国家住户调查主抽样标架和主样本的设计”，《发展中国家和转型期国家住户调查》。《方法研究》第96号，出售品编号：E. 05. XVII. 6。
- Turner, A. (1998年)。《特派团提交柬埔寨王国全国统计学会的报告》，11月11-24日。全国统计学会，Phnom Penh：未出版。
- _____ (1999年)。《特派团提交阿拉伯联合酋长国卫生部和中央统计局的报告》，1月23日-2月3日。阿布扎比：未出版。中央统计局。
- _____ (2000年)。《特派团提交莫桑比克全国统计学会的报告》，8月13-26日。全国统计学会，马普托：未出版。
- 联合国统计处(1984年)。《住户调查手册》(修订版)，《方法研究》第31号，出售品编号：E. 83. XVII. 13。
- _____ (1986年)。《国家住户调查能力方案：综合住户调查方案的抽样标架和样本设计》(初版)。DP/UN/INT-84-014/5E。纽约：联合国发展技术合作部和统计处。
- 联合国儿童基金会(2000年)。《十年末多指标调查手册》。纽约：联合国儿童基金会，第4章。
- 美国人口普查局(1978年)。《现时人口调查设计和方法论》，技术文件，第40号。哥伦比亚特区华盛顿：人口普查局。
- Verma, Vijay (1991年)。《抽样方法》，培训手册。东京：亚洲和太平洋统计学会。
- 世界银行(1999年)。《核心福利指数问卷手册》。哥伦比亚特区华盛顿：世界银行，第4章。

第五章

样本设计的文件编制与评价工作

5.1. 引言

1. 本章虽然篇幅相对较短，但在本手册中起到核心作用。在匆忙争取早日发表调查成果时，往往忽视了特定样本设计的文件编制与评价工作和一般的调查方法问题。尤其在缺乏住户调查经验的国家更是如此，那里以调查工作单和报告的文献形式所载的元数据残缺不全。有些国家，包括实施抽样在内的调查程序文件编制基本上未受重视。因此，在需要惯例的地方却没有惯例。虽然研究人员承诺对调查所产生的数据进行分析，但是他们也应该有兴趣了解与样本设计有关的各种方法。因此应编制若干文件，以详细说明调查中采取的相关程序。

2. 调查成果的评价时常被完全忽视，结果在调查分析中夹杂了误差。一般来讲，这是因为在预算限制下往往不能开展任何正式研究或拟定适当的方法来评估在住户调查中可能发生的非抽样误差的多寡。然而，还有其他一些衡量数据质量的现成指标(比如不答复率)，但这方面也往往在调查报告中很少提及。

3. 本章还强调了有必要向用户提供有关已知数据缺点的信息——即便尚未进行正式评估；但在这方面需要强调一点，就是关于正式评估调查方法的技巧方面的讨论超出了本章的主题范围，因为相关的研究成果太多了。¹所以本章着力于探讨应当向用户提供什么样的信息以帮助他们评估调查质量，尤其集中在抽样方面。

5.2. 样本文件编制与评价工作的必要性和类别

4. 在住户调查中需要有两种文件编制工作。一种是在实施调查的过程中仔细记录调查和取样程序的运作情况。若没有这种文献，调查分析中就会夹杂误差。比如说，要是没有细心保管记录，在进行分析的时候就可能不太了解选择概率。

5. 因此，抽样技术人员应采取必要措施仔细编制文件，不但要记录特定调查的抽样计划，而且要记录它的实施过程。由于在实施调查的过程中可能出现意外情况，所以往往需要在现场工作的各种不同阶段对样本设计作适应性调整。有必要一步一步地记录执行抽样计

¹ 旨在评估调查中的非抽样误差具体类别的专题研究包括：再采访调查(以评估反应率的变量)；计数后的调查(以评估覆盖率和调查内容)；互相贯穿的样本(以评估访调员的变量)；反向核对记录(以评估受访者的记忆误差)，等等。

划中使用的所有程序, 以确保严格按设计行事。假如没按设计执行, 那就更有必要记录所有偏离设计的操作, 就连小的偏离也不要漏掉。在日后分析阶段一旦需要做出调整的时候, 上述信息是必要的。另外, 此类信息对于规划将来的调查也是必不可少的。

6. 第二种文献工作就是编写报告。每次调查需要编写两类报告: 一种是方便用户的关于包括抽样计划及其实施在内的调查方法的简要说明; 另一种是关于调查方法的详细描述。前者一般涵盖各种实质性报告的“技术”部分(或称附件部分), 发表的目的在于讨论和解释实质性调查结果,²其中包括一小节关于已知数据缺陷的描述(见下文)。

7. 第二种技术报告主要是为专业研究人员、社会科学家和统计学家而不是决策者或一般公众提供的, 这种报告应包含对调查方法的较详细描述, 并且应当独立发表而不是作为实质性系列报告的一部分发表。美国人口普查局(1978年)提供了此类报告的一个很好版本。当然, 最好还是提供详细的技术报告和同时发表的阶段性调查报告, 不过前者一般都是很晚以后才发表(如果有的话)。在统计刊物上发表技术报告或某种缩写本以延长其使用寿命也是个有益的办法。

8. 这两种报告都很重要, 建议各国的统计部门指定一个专门办公机构或官员作为例行工作编写此类住户调查报告。

5.3. 设计变量的标号

9. 本节和第5.2-5.7节讨论第一类文献编制工作, 其中涉及到与抽样有关的记录保管问题。

10. 每个阶段确定的选样单位都必须清楚地单独标号。在多阶段设计中, 这就需要建立初级、二段、三段和最终抽样单位编码(取决于设计中有多少阶段)。通常, 对初级抽样阶段来说, 四位数编码就够用了, 其余阶段三位编码也足够了。地理区域也必须适当标号。此外, 识别抽样单位所属地域和行政区结构的行政编码应属于标号程序的组成部分。分析单位也要有明确的标识。

范 例

假定一个被界定为普查计数区的样本由1 200个初级抽样单位组成, 被选定用于一个两阶段设计, 在城、乡两个域中各有600个初级抽样单位。给初级抽样单位编码的一个简便方法就是从0001号编到1 200号。另外, 把这些同样顺序的编码指定用于选择初级抽样单位也很有用。可能需要把这一特征用于计算抽样方差。因此, 如果农村的初级抽样单位先抽样, 它们就会按照从0001到0600编码, 而城市的初级抽样单位就会从0601编到1 200。这种编码方法有两个好处: 一是每个初级抽样单位都是独一无二的编号和标识, 二是分析家从识别码一眼就能看出一个初级抽样单位是城市的还是农村的。在抽样第二阶段, 每个初级抽样单位都被列表, 并选出了20个住户作为采访对象。在这

² 关于一项抽样调查报告可以包含哪些内容, 可参见联合国统计抽样小组委员会的一份报告(联合国, 1964年)。

个阶段，所有列选的住户都给一个三位编码(或者四位编码，如果某些计数区包含999个以上住户的话)，仍然按其编入清单的顺序。鉴于样本住户将保留如此指定的编码，就不再给被选中的住户指定编码了；譬如讲，编码01-20。最后，必要时可指定行政编码。这样，样本住户080的编码09 003 008 0128就把该户鉴别为：在属于09省003区第008民事法庭辖下之初级抽样单位0128中被列入清单(并被选定为采访对象)的第80户。另外，从初级抽样单位编号一眼就能看出它属于农村域。如果调查获得了有关住户成员的信息，那么每个住户成员也都有一个从01到99独一无二的两位数编码。

11. 恰当标号首先对质量管理是必不可少的：随着给访调员分派任务和从现场返回问卷，可以对照主清单对其逐一核对，以确保所有样本住户都访问到了。其次，独一无二编码系统对数据处理人员也是很有价值的，因为可以据此按照地理位置列表。

12. 对于有综合调查方案的国家来说，所有调查都采用统一的标准式样对设计变量进行标号是非常可取的。借此消除数据生产者 and 数据使用者当中存在的混乱，这在数据处理和成果表达方面都有明显的好处。

13. 从后者观点来看，正如前面范例中所述，通过给初级抽样单位全域、而不仅仅是样本单位指定初级抽样单位编码，将使综合调查方案受益。这是因为，不同的初级抽样单位往往在不同的调查中被抽样，当使用主样本的时候就经常是这种情况。

14. 一般来说，主样本甚至比一次性调查样本更需要设计标号。前面讨论过，主样本的一个关键用途就是重复进行不同轮次的同一调查。设计变量的恰当标号可借以识别各个选样阶段，这对于在不同调查之间保持跟踪样本的重叠部分至关重要。为了便于识别拟在后续调查轮次中置换的单位(住户、类集或初级抽样单位)，往往指定轮换组群(整体样本的系统子集)。它们当然需要有各自的群组识别码。另外，在阶段性更新主样本的过程中追加的住户也必须予以适当编码。编码方案的设计要便于在新老住户之间加以区分。

5.4. 选择概率

15. 在样本文件编制方面有一项信息往往被忽视，那就是记录各个不同阶段的选择概率。在有信息存在的情况下，往往局限于每个样本案例的总体样本权数(据此可以很容易地计算出总体概率)。

16. 关于适当的文件编制，必须考虑到一个特别重要的细节，就是在实地收集数据期间，如果样本区段或类集过大，就可能需要进行二段抽样。举例来讲，前面讨论过，一个出乎预料的大类集可能需要细分成大体相等的四部分。然后随机选定一部分以供列表和采访。在这种情况下，抽样区段(以及在其中选定的住户/个人)的总体概率就是原类集的四分之一；而它的权数就是四分之一的倒数，或为4。在分析数据的时候，这个加权因子必须反映在计算值里。

17. 在一个住所里有不止一个住户的情况下，(如果该住所是列表单位的话)可能也需要进行二段抽样。对于调查管理者来说，一个不偏不倚的选项就是采访所发现的所有住户：如果只有两户的话通常就采取这个办法。可是，譬如讲，这个住所里有5户，而需要选取的只有一户。预算限制决定了他们当中只有一户(当然通过随机任择)将接受采访。这里也同样

必须仔细记录二段抽样率(在本例中是五分之一),以便抽样工作人员能够准确计算有关住户的选择概率,并且对权重加以适当调整(用因子5倍增)。

18. 正如本节开头所述,记录每个选样阶段的选择概率也是有益的。譬如讲,每当使用与样本规模成正比的概率抽样法的时候每个初级抽样单位的选择概率就不相同。即便总体样本设计能自我加权也是如此。如果忽视或错误记录了初级抽样单位的选择概率就不可能确定总体抽样权数。比如,了解原始选择概率对于准确敲定二段抽样程序很有用。

5.5. 各个选样阶段的反应率和覆盖率

19. 作为样本调查实施评价过程的组成部分,向用户提供有关反应率和覆盖率的信息是至关重要的。因此,不但有必要提供反应率(或其不足数——不答复率),而且有必要列表明不答复的原因。不答复的情况大概分为如下类别:

- 家中无人;
- 居住单元空置;
- 居住单元已拆迁或不能居住;
- 拒绝受访或答复;
- 暂时外出(度假等)。

20. 反应率的定义,就其所包含的类别而言,各国可能不尽相同。不过,一般来说完整的答复包含上述第一、第四和第五类情况。这几类情况下应尽可能获得答复。空置和拆迁的单位(在计算反应率时)通常忽略不计,因为按定义不可能在这种单位获得答复。于是,举例来讲,一个国家可能选择5 000个住户,结果如下:4 772户完成了采访;75户“家中无人”;31户住所空置;17户已拆迁;12户拒绝答复;93户“暂时外出”。计算反应率的时候通常将空置和拆迁单位排除在外,结果是: $4\,772 / (5\,000 - 31 - 17)$,或反应率为96.4%。

21. 在调查的目标人口包括住户(比如为了调查家庭收入或享受服务的机会)和个人(比如为了调查成年妇女的健康状况)的情况下,通常的惯例是在住户和个人两个层面计算反应率。举例来讲,可能有98%的住户答复;但是在答复的住户内,可能有一小部分个人是不答复者。

22. 由于种种原因往往整个类集没有进行采访,其中包括安全问题,比如民事冲突或暴乱;以及因为地势崎岖或天气恶劣而难以通达。当发生此类问题的时候,往往选择替换类集,这种办法会带来严重偏差,因为替代类集的居民几乎总是可能在很大程度上有别于被取代类集中的居民。但是当实施这种替代的时候,调查工作队必须记录这些类集的数量和地点。另外还有必要提供一些有关在这种情况下覆盖不足的信息。这大概需要尽可能估计据认为在目标人口中有多少人居住在取代和被代表类集的地区。

23. 值得注意的是,上面直接提及的那类问题的解决办法是,可在选样之前确定国内有哪些地区由于安全或难以通达的缘故而“超出了调查采访的范围”。确定的这些地区应计入文件并被排除在调查全域之外。被排除在外的地区应该写入报告,并且声明调查结果对这些除外地区不适用。

5.6. 加权：基权数、不答复和其他调整

24. 第六章讨论调查权数的计算问题。本章强调关于这些计算值的文件编制的重要性。

25. 住户调查的加权一般涉及到三个方面的操作——基权数或设计权数的计算；不答复的调整；以及分层后的调整。在许多应用中只使用设计权数，而在其他应用中可以用一个附加因子来调整设计权数，以体现不答复的情况。在相对很少的应用中，加权还可以反映另一个因素，而不论是否进行不答复的调整。这种加权意在调整基于样本的人口分布情况，以便使其与基于另一独立数据源（比如最近的普查）的人口分布相一致。这种加权通常叫做分层后的加权。有些应用根本不进行加权；这必须具备两个条件：一是样本完全自我加权，二是所产生的数据仅限于百分比分布、比例和比率，而不是估计的总数或绝对值。

26. 在使用加权的时候，当然有必要仔细记录计算值。如前所述，应计算和记录每个选样阶段的权数（或概率）。在每个数据操作阶段的单独权数也要记录，此类权数包括：(a) 设计权数，(b) 经不答复调整因子倍增之后的设计权数，以及(c) 在应用了后分层调整因子之后(b)项中的设计权数。

27. 有必要着重强调一点，即：每当样本设计包含域估算的时候，每个域的设计权数各不相同。换言之，甚至当样本在域内自我加权的时候，每个域也都会有各自不同的权数。再者，如果设计不是域内自我加权，那么每个域都会有一套不同的权数。此外，值得一提的是，不答复调整往往由重要地理次区域（比如大政区）单独使用，而不管设计中有没有域的估算。最后，设计权数本身可由一个特定类集或住户因子倍增，每当使用二段抽样法（见第5.3节）的时候就可以这么做。

5.7. 关于抽样和调查实施成本的信息

28. 虽然通常都要为住户调查仔细编造预算，但同样重要的是保管各项活动期间实际支出的记录。譬如讲，调查成本记录对规划主样本很有用，而保管各种调查活动记录对于制订未来调查计划也十分重要。

29. 在利用主样本的时候，需要支出很大一笔启动费来实现它的开发。一般包括以下几方面：(a) 用计算机处理普查文卷以建立抽样标架；(b) 通过绘制地图或制图工作创建初级抽样单位；以及(c) 用计算机选择样本的初级抽样单位。上一章讲过，启动业务经费往往由行将在主样本生命周期利用主样本的各行政部门来分摊。这些费用也应分配到需要使用主样本的所有调查中去，但条件是所有这些调查项目都事先为人所知。因此非常仔细地保管有关主样本开发和未来调查抽样计划方面的记录至关重要。

30. 一俟主样本到位，就需要汇编有关维持主样本的成本记录。如前所述，主样本的更新工作是阶段性的，当然也需要仔细监测这方面的成本。

31. 应定期获取抽样业务成本数据。下面列举的各种成本项目既适用于一次性样本调查，也适用于主样本：

- (a) 样本设计(包括任何外部顾问)工资；
- (b) 更新抽样标架的现场成本，包括人工费和准备各种辅助材料(如地图)的费用；
- (c) 为选择初级抽样单位样本用计算机准备抽样标架的各项费用；
- (d) 为选择初级抽样单位样本的人工费用(如果不是用计算机的话)；
- (e) 在倒数第二阶段抽样单位进行登记造册的现场费用，包括人工费和准备材料(如类集文件夹)的费用；
- (f) 从样本类集中的住户样本收集数据的人工费用。

32. 除了调查抽样费用之外，还须保管与实施成本有关的各项记录。此类费用可能包括：调查员和主管人员工资；调查组织正规工作人员的日常生活补贴和现场开销；差旅费；办公用品；培训开支；燃料费用；通信服务和数据处理费用。

5.8. 评估：调查数据的缺陷

33. 上面讨论的关于妥善保管记录的大部分文件编制工作不但对处理调查结果十分重要，而且对评估样本设计和调查实施情况也很有用。譬如讲，关于反应率的信息就有助于评估因为不答复所导致的偏差是否严重。抽样成本信息可用于评估样本设计的“经济”效益如何及其对未来调查的用途。

34. 如前所述，对样本调查的正式评价涉及到非抽样误差的方方面面，远远超出了本手册的范围(有关这个问题的综合处理，见联合国(1984年))。不过值得一提的是，非抽样误差的评估应包括譬如讲现场运作和数据处理方面的活动。另一方面，抽样误差是可以估算的，下面继续讨论。

35. 尽管不是经常对住户调查进行正式评价研究，但在调查文件编制工作中包含有关数据缺陷的信息是至关重要的。关于这个论题，应该在实质性调查成果报告中专设一节，通常题目就叫“调查数据的缺陷”。要在这一部分告知读者有关抽样和非抽样调查误差的信息。

36. 美国人口普查局的一份有价值的出版物(1974年)讨论了如何提供有关调查误差的信息。从该出版物摘取的如下示范段落(附件一第I-1页)说明了在发表调查成果的时候应该向读者提供哪种信息：

载于本报告的统计数据是从一次样本调查中推导出来的估计数。在基于样本调查的估计中可能会有两类误差——即抽样误差和非抽样误差。发生抽样误差的原因是观察对象仅限于样本而不是整个人口。非抽样误差(第八章讨论)可以归咎于许多原因，诸如：不能获取有关样本中所有案例的信息；定义方面存在困难；在提问的解释方面存在差异；受访者不能或不乐意提供正确信息；在所获数据的记录或编码方面发生错误；以及在收集、应对、处理、覆盖和估算短缺的数据等其他方面出现的误差。完整的人口普查也会发生非抽样误差。调查结果准确与否，取决于抽样和非抽样误差的综合效应。

本次调查中使用的特定样本是所有可能用同样的样本设计选定的同样规模的大量样本之一。从不同样本推导出来的估计数会彼此各不相同。一个样本估计数偏离所有可能样本的平均值，叫做抽样误差。一个调查估计数的标准误差是从各种可能的样本得出的估计数之间的一个差异量度，因而也是出自某一特定样本的估计数借以逼近所有可能样本平均结果的一个精确度的量度。相对标准误差的定义是：估计值除以标准误差。

据本报告计算，标准误差亦可部分地据以计量非抽样误差效应，但是不能估量数据中的任何系统偏差。偏差是所有可能样本的平均估计数与期望值之间的差异。显而易见，一个调查结果的精确性既取决于抽样误差和用标准误差计量的非抽样误差，又取决于偏差和不用标准误差计量的其他非抽样误差。

37. 上文暗示，样本误差的估算是对主要调查估计数进行样本评价的一个重要组成部分。前面讨论过，概率样本的突出特点之一就是样本自身可以用来估计标准误差。第六章详细讨论方差和标准误差的估算方法。此外，还要尽可能利用现有各种各样可靠的软件包来估计标准误差。

38. 一般来讲，标准误差估计数是为调查中令人关注的主要特征准备的，因为计算所有项目的标准误差估计数既不现实也无必要。当然，标准误差为读者提供的平均值是为了评价调查估计数的可靠性，并且在点估计数上下构建置信区间。

39. 标准误差也可以用来评价样本设计本身。这种做法的一个特别有用的统计数据，就是样本的设计效果， $deff$ ，或者更确切地说，是 $deff$ 的平方根， $deft$ 。需要估计标准误差的每个数据项的 $deft$ 的计算方法非常直截了当：对于一个特定项目而言，只需用基于一个同样规模的简单随机样本的标准误差除以估计的标准误差，即： pq/n ，式中， p 是估计的比率； q 是 $1-p$ ；而 n 是样本规模。应用此种方法的目的在于确认或否定在设计样本时假定的设计效果，因为在进行了调查、处理了数据和估算了标准误差之前，很难知道实际的 $deffs$ 或 $defts$ 。

40. 抽样技术人员可以使用计算出来的设计效果来确定类集的规模对于关键数据项是否合适，并且在必要时采取纠正措施。举例来讲，对于特定关键项目来说，如果 $deft$ 比预想的大得多，那就可以用小一点的类集规模来设计未来的调查。

5.9. 准则概要

41. 本节以对照表的形式总结本章所讨论的各项主要准则。跟前面各章一样，我们将这些准则视为经验法则而不是刻板建议：

- 以两种方式编制调查抽样方面的各种文件：一是恰当保管书面记录，二是向用户提供技术信息；
- 保管包括成本在内各种抽样过程的详细记录；
- 为样本设计变量编码，变量包括：行政区、初级抽样单位、类集、住户、个人，等等；
- 力争实现设计变量的编码标准化，并在所有调查中保持一致；

- 记录在实施调查过程中发生的偏移或背离其他原始样本计划的情况;
- 计算并记录每一抽样阶段的选择概率;
- 尤其要记录关于发生在现场工作中的二段抽样的信息;
- 记录关于不答复数量及类型的信息;
- 记录设计权数、不答复调整和分层后的调整情况;
- 保管样本设计和实施中的每项业务活动成本的详细记录;
- 在主样本方面, 保管有关其开发(启动)及其维持的成本记录;
- 为用户编写关于抽样和调查方法的技术报告;
- 为报告调查成果的所有出版物编写有关数据缺陷的简要报告;
- 就抽样方法的方方面面编写较为详细的技术报告;
- 计算并在技术报告中提供关键变量的抽样误差;
- 计算关键变量的设计效果(*deff*或*defi*);
- 任命一位主管文件编制工作的官员。

参考资料和补充读物

Casley, D. J. 和D. A. Lury (1981年)。《发展中国家的数据收集工作》。英国牛津: 克拉伦登出版公司。

国际统计学会(1975年)。《样本设计手册》。世界生育率调查基本文献。荷兰福尔堡。

阿拉伯国家联盟(1990年)。《阿拉伯母婴健康调查抽样手册》, 基本文献第5号。开罗: 泛阿拉伯儿童发展项目。

宏大国际咨询公司(1996年)。《抽样手册》, 第三次人口和健康调查基本文献第6号。马里兰州卡尔夫顿: 宏大国际咨询公司。

联合国(1964年)。“关于编写抽样调查报告的建议”(临时刊出), 《统计论文集》C辑, 第1号, 第二修订版。

_____ (1984年)。《住户调查手册》(修订版), 《方法研究》第31号, 出售品编号: E. 83. XVII. 13。

美国人口普查局(1974年)。《数据误差的讨论和展示标准》, 技术文件第32号。哥伦比亚特区华盛顿: 人口普查局。

_____ (1978年)。《现时人口调查设计和方法论》, 技术文件第40号。哥伦比亚特区华盛顿: 人口普查局。

世界银行(1999年)。《核心福利指标问卷手册》。哥伦比亚特区华盛顿: 世界银行。

第六章

样本权数的构建与使用

6.1. 引言

1. 本章讨论样本权数开发及其用于住户调查数据相关特征估计数计算的各个阶段。尤其是描述了为补偿入选样本的种种不足之处而进行样本权数调整的方法。注意力局限于在大多数调查报告广泛产生的描述性估计数方面。使用当前发展中国家进行的调查实例或模拟实际调查境况的例证来说明一些重要理念。

6.2. 抽样权数的必要性

2. 住户调查一般建立在复杂样本设计的基础之上，主要是为了控制成本。设计出来的样本可能会有各种不足之处，从而导致样本与参照人口之间的偏差和其他差异。此类不完善之处包括用不均等的概率选择单位、人口未覆盖和不答复等。为了弥补这些缺陷就需要有样本权数并借此推导出相关特征的适当估计数。总之，加权的目的在于：

- (a) 补偿不均等的选择概率；
- (b) 补偿(样本单位的)不答复；
- (c) 调整所关注的关键变量的加权样本分布(比如年龄、种族和性别)，以便使其与已知的人口分布相一致。

3. 下面各节讨论在上述每一种情况下采用的程序。一旦需要对样本不足之处给予补偿，可以在估算相关人口特征方面以及估算所产生的调查估计数抽样误差方面使用权数。

4. 如果不使用权数来补偿层内区别选择率(在样本如此设计的情况下)和上述样本不足之处的话，一般得出的人口参数估计数就会出现偏差。关于在每种情况下使用加权程序的实例，每种情况下加权估计数和不加权估计数的比较，见第6.3节、6.4节和6.5节。

6.2.1. 概览

5. 第6.3节结合多级样本设计处理抽样权数开发的问题，其中包括调整样本权数以补偿样本重复及在选样时尚不知是否有资格参加调查的抽样单位的情况。第6.4节讨论不均等选择概率的加权；提供了几个数字实例，其中包括一个关于开发全国住户调查权数的案例研究；并且在末尾讨论了自我加权的实例。第6.5节和第6.6节分别讨论了住户调查中的不答复

和未覆盖的问题。讨论了不答复和未覆盖的原因与后果。同时提出了对不答复和未覆盖的补偿方法, 其中包括说明不答复和未覆盖样本权数调整的数字实例。第6.7节讨论了由于在住户调查数据分析中使用样本权数致使调查估计数方差膨胀的问题。还提供了一个数字实例来说明如何计算因为加权导致的方差增加量。第6.8节讨论了权数整理问题, 并提出了一个整理程序的实例, 借以对经过整理的权数进行重新整理, 以将其添加到原始权数的总和中去。第6.9节提出一些结论。

6.3. 抽样权数的开发

6. 一俟确定了抽样单位的选择概率, 即可开始构建抽样权数。一个抽样单位的选择概率取决于用来选择单位的样本设计。第三章描述了最常用的抽样设计与与这些设计相对应的选择概率。自始至终假设已经确定了选择概率。

7. 抽样权数的开发有时被认为是分析调查数据的第一步。它通常从构建每个样本单位的基权数或设计权数入手, 以反映其不均等的选择概率。一个样本单位的基权数是其纳入样本的选择概率的倒数。如果一个单位以概率 p_i 被纳入样本, 用符号 w_i 来表示其基权数, 那么它的数学表达式就是

$$w_i = 1/p_i \quad (6.1)$$

8. 举例来说, 一个以1/50的概率被选取的样本单位代表从中提取样本的人口中的50个单位。于是, 样本权数就充当了设计为代表调查人口单位数的膨胀因子, 而这个人口是由被指定了权数的样本单位所代表的。样本权数总和提供了目标人口中单位总数的一个无偏差的估计数。

9. 对于多阶段设计来说, 基权数必须反映每个阶段的选择概率。举例来讲, 在两阶段设计情况下, 第 i 个初级抽样单位在第一阶段以 p_i 的概率被选中, 而在这个被选中的初级抽样单位中的第 j 个住户在第二阶段以 $p_{j(i)}$ 的概率被选中, 那么, 该样本中每个住户的总体选择概率(p_{ij})即可由这两个概率的乘积给出, 或者表示为

$$p_{ij} = p_i \times p_{j(i)} \quad (6.2)$$

而住户的总体基权数, 如前所述, 则取其总体选择概率的倒数。相应地, 如果第 j 个住户的基权数是 $w_{ij,b}$, 可归于不答复补偿的权数是 $w_{ij,nr}$, 而可归于未覆盖补偿的权数是 $w_{ij,nc}$, 那么该住户的总体权数可由下式给出:

$$w_{ij} = w_{ij,b} \times w_{ij,nr} \times w_{ij,nc} \quad (6.3)$$

6.3.1 未知被选资格的样本权数调整

10. 在住户调查收集数据期间, 有时候住户的被选资格可能会成问题。譬如讲, 访调员可能会在收集数据时或反复多次回访时发现一个被抽样的居住单元一个人也没有。在这种情况下

下不知道该居住单元是否有人住。如果确实有人住，那就应该将其归类为“不答复”居住单元(属于“家中无人”类)。否则，它就超出了调查范围，因此没有资格被计入样本单位之列。有时候访调员假定，如果反复回访都发现一个居住单元家中无人，那么该居住单元就无人居住，因此也就无资格被选。一般来说这种假定是不对的，它往往导致错误地使反应率膨胀。

11. 如果不知道某些被抽样的居住单元是否有资格被选，那就必须调整它们的权数以反映这一事实。这个想法就是做出一些假定，以便估算未知是否合格而实际合格的居住单元的比例。最简单的办法就是确定已知要么合格要么不合格的已抽样居住单元的比例，然后将此比例应用于未知是否合格的居住单元。譬如讲，假定一个由300个居住单元组成的样本对调查的反应分类有如表6.1所示：

表6.1
某次调查的反应分类

反应分类	居住单元数
完成采访	215
有资格被选而不答复	25
无资格被选	10
未知是否有资格被选	50

12. 请注意，已知其是否合格的居住单元中实际合格的比例是 $(215+25)/(215+25+10) = 0.96$ 。因此我们可以假定有同样比例(0.96)的未知其是否合格的居住单元可以被认为合格。换言之，在50个未知是否合格的居住单元中，实际合格的占96%(或有48个居住单元)。然后我们使用按如下定义的调整因子，来调整合格居住单元(完成采访者和有资格被选而不答复者)的权数：

$$F_{ue} = \frac{\sum_c w_{ij,b} + \sum_{nr} w_{ij,b} + \varepsilon \times \sum_{ue} w_{ij,b}}{\sum_c w_{ij,b} + \sum_{nr} w_{ij,b}} \tag{6.4}$$

式中， ε 表示未知是否合格但估计为合格的情况(在本例中 $\varepsilon=0.96$)。上面公式中 c 、 nr 和 ue 的各自总计分别为完成采访的住所、合格而不答复的住所以及未知是否合格住所的基权数之总和。然后，通过用初始基权数 $w_{ij,b}$ 乘以加权因子 F_{ue} ，可获得完成采访的住所及合格而不答复的住所的调整后的基权数。

6.3.2. 重复样本的权数调整

13. 假定知道某些样本单位在标架上有重复，那么，如果这种单位最终入选样本的话，可以通过给它们指定加权因子——即重复列入标架的数量的倒数——来补偿其膨胀的选择概率。然而，往往在完成抽样之后才发现重复现象，并且有必要对这种被抽样单位的选择概率加以调整来说明重复现象。此种调整的实施程序如下：假定第 i 个抽样单位的选择概率

为 p_{i1} , 并且假定在抽样标架上又发现有 $k-1$ 个记录与该抽样单位重复, 其各自的选择概率分别为 p_{i2}, \dots, p_{ik} 。那么, 该抽样单位的调整后的选择概率由下式给出:

$$p_i = 1 - (1 - p_{i1})(1 - p_{i2}) \cdots (1 - p_{ik}) \quad (6.5)$$

然后, 该抽样单位便经由 $1/p_i$ 得到相应加权。

14. 下面通过具体实例来说明在上述种种情况下构建样本权数的程序。

6.4. 不均等选择概率的加权

15. 为了便于说明, 让我们来考虑一个以普查计数区为其初级抽样单位、而以住户为其第二阶段单位的两阶段设计。假定第一阶段从 N 个单位总数中选取了一个由 n 个初级抽样单位组成的等概率样本, 然后又从每个被选中的初级抽样单位选取 m 个住户。很明显, 一个住户的选择概率将取决于它所在初级抽样单位的总户数。设 M_i 表示初级抽样单位 PSU_i 中的住户数。那么, 一个初级抽样单位的选择概率即为 n/N , 而在第 i 个初级抽样单位中, 选择一个住户的有条件概率则为 m/M_i 。因此, 一个住户的总体选择概率可由下式给出:

$$p_{ij} = p_i \times p_{j(i)} = \frac{n}{N} \times \frac{m}{M_i} = \frac{nm}{N} \times \frac{1}{M_i} \quad (6.6)$$

根据这种设计, 一个被抽样住户的权数亦可由下式给出:

$$w_i = \frac{1}{p_{ij}} = \frac{N}{nm} \times M_i \quad (6.7)$$

范 例 1

假设从250户中选定了—个由5个住户组成的等概率样本。从每个住户随机选择一个成年人。记录了第 i 个住户的这位被选中的第 j 个成年人的月收入(y_{ij})和受教育程度(如果受过中等或高等教育, $z_{ij}=1$; 否则为0)。设 M_i 表示住户 i 中的成年人数量。那么, 一个被抽样成年人的总体选择概率可由下式给出:

$$p_{ij} = p_i \times p_{j(i)} = \frac{5}{250} \times \frac{1}{M_i} = \frac{1}{50} \times \frac{1}{M_i}$$

因此, 一个被抽样的成年人的权数可由下式给出:

$$w_i = \frac{1}{p_{ij}} = 50 \times M_i$$

16. 现在我们来解释在上述设计中如何计算基本估计数。假定在第一阶段5个住户组成的样本中从每个住户单独选定的成年人那里收集的数据有如下面表6.2所示。请注意, 每个住户的成年人数量和从每户选定的成年人的相应总体权数分别在第二栏和第三栏内给出。

表6.2
在不均等选择概率情况下的权数

被抽样的住户	M_i	w_i	y_{ij}	z_{ij}	$w_i y_{ij}$	$w_i z_{ij}$	$w_i z_{ij} y_{ij}$
1	3	150	70	1	10 500	150	10 500
2	1	50	30	0	1 500	0	0
3	3	150	90	1	13 500	150	13 500
4	5	250	50	1	12 500	250	12 500
5	4	200	60	0	12 000	0	0
共 计	16	800	300	3	50 000	550	36 500

17. 然后，可从表6.2获得下列各项特征估计数：

平均月收入估计数是

$$\bar{y}_w = \frac{\sum w_i y_{ij}}{\sum w_i} = \frac{50\,000}{800} = 62.5$$

如果没有使用权数，这个估计数应为60(或300/5)。

受过中等或高等教育的人的比例估计数为

$$\bar{z}_w = \frac{\sum w_i z_{ij}}{\sum w_i} = \frac{550}{800} = 0.6875 \text{ 或 } 68.75\%$$

如果没有使用权数，这个估计数应为3/5或0.60或60%。

受过中等或高等教育的总人数估计数为

$$\hat{t} = \sum w_i z_{ij} = 550$$

受过中等或高等教育的成年人平均月收入估计数为

$$\bar{y}_w = \frac{\sum w_i z_{ij} y_{ij}}{\sum w_i z_{ij}} = \frac{36\,500}{550} = 66.36$$

18. 有时候，抽样权数是“赋范的”，即权数要乘以如下比率

$$\frac{\text{受访者人数}}{\text{所有受访者权数之总和}} \quad (6.8)$$

19. 可见，赋范权数的总和就是供分析的实际样本规模(即受访者人数)。注意，赋范权数不能用于估计总数，比如受过中等或高等教育的成年人总数。在这种情况下，被抽样的单位需要用它们的选择概率的倒数来加权，也就是说，必须使用正常的抽样权数。可是对于估算平均数和比例来说，只要权数与选择概率的倒数成比例即可。换言之，无论使用正常权数还是使用(与正常权数成比例的)赋范权数来获取人口参数平均估算数(比如享受初级保健服务的育龄妇女平均人数或比例)的估计数都无所谓。这两类权数会产生同样的结果。

20. 举例来讲, 在前面的实例中, 权数 w_i 与 M_i ($w_i = 50 \times M_i$)成比例。如果我们用 M_i 作权数, 那么受过中等或高等教育的人口比例估计数就是

$$\hat{p} = \frac{\sum M_i z_{ij}}{\sum M_i} = \frac{3 \times 1 + 1 \times 0 + 3 \times 1 + 5 \times 1 + 4 \times 0}{3 + 1 + 3 + 5 + 4} = \frac{11}{16} = 0.6875$$

或68.75%, 跟以前完全一样。可是对于受过中等或高等教育的成年人总人数的估计数来说, 就必须使用正常的抽样权数($w_i = 50 \times M_i$)来获得正确结果, 也就是说,

$$\hat{t}_s = \sum (50 \times M_i) z_{ij} = 50 \sum M_i z_{ij} = 50 \times 11 = 550$$

在某国农村地区选择了一个两阶段住户样本。在第一阶段, 对50个村子进行了抽样, 选择概率与上一次普查时的住户数成比例。上次普查时农村地区住户总数为300 000户。继第一阶段样本选择之后进行了登记造册, 以编列每个选定村子的居住单元清单。有时候发现, 一个居住单元内住着不止一个住户。

21. 现在我们来考虑各种再抽样选项(为了从选定的居住单元选择住户), 并且具体指定计算一个住户被纳入样本的总体选择概率的方程式。设 D_i 表示 i 村的居住单元数, 并且设 H_{ij} 表示 i 村居住单元 j 内的住户数。于是, 一个村子的住户总数 H_i 即可表示为

$$H_i = \sum_j H_{ij} \quad \text{请注意} \quad \sum_i H_i = \sum_i \sum_j H_{ij} = 300\,000$$

这里计算的选择概率系基于第三章介绍的公式。

设计选项1

22. 采用无置换简单随机抽样法(SRSWOR)从每个入选村庄的清单中选择了15个居住单元。住在入选居住单元中的所有住户均已纳入样本, 因此只有两阶段选择: 一是选择村庄, 二是选择居住单元。根据这一设计, 一个住户被纳入样本的总体选择概率由如下方程式给出:

$$P_{ij} = pr(\text{选定了 } i \text{ 村}) * pr(\text{在选定了 } i \text{ 村的前提下选定了居住单元 } j)$$

$$\text{那么, } p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{15}{D_i} = \frac{750}{\sum_i M_i} \times \frac{H_i}{D_i} \quad \text{基权数可由下式给出: } w_{ij} = \frac{\sum_i M_i}{750} \times \frac{D_i}{H_i}$$

23. 请注意, 各村的总体选择概率各不相同, 这要取决于住户数对居住单元数的比率 H_i/D_i 。因此我们的结论是, 该设计不是自我加权的(关于自我加权设计的进一步讨论, 见下文第6.4.2节)。如果每个居住单元只有一个住户的话, 也就是说, 如果所有抽样村的比率 H_i/D_i 相同, 则可以进行自我加权。

设计选项2

24. 在选定的村子里对住所进行系统抽样，一个村的抽样率与其上次普查时的住户数成反比。选定住所内的所有住户都包含在样本中。跟前面一样，只有两个选择阶段：一是选村庄，二是选住所。在选定的*i*村中选择一个住所的有条件概率可以表示为 k/H_i ，式中， k 是比例的常数。因此，根据这种设计一个住户被纳入样本的总体选择概率的选择方程式可由下式给出：

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} = \frac{50 \times k}{\sum_i M_i}$$

而基权数可由下式给出：

$$w_{ij} = \frac{\sum_i M_i}{50 \times k}$$

这是一个常数。因此我们得出结论：设计选项2是一个自我加权设计。

设计选项3

25. 在选定的村子里对住所进行系统抽样，一个村的抽样率与其上次普查时的住户数成反比。在每个入选的住所随机选择一个住户。在这种情况下，有三个选择阶段：分别选择村、住所和住户。因此，根据这种设计一个住户被纳入样本的总体选择概率的选择方程式可由下式给出：

$$p_{ij} = 50 \times \frac{H_i}{\sum_i H_i} \times \frac{k}{H_i} \times \frac{1}{H_{ij}}$$

而基权数由下式给出： $w_{ij} = \frac{\sum_i M_i}{50} \times \frac{H_{ij}}{k}$

这个基权数在各个居住单元之间各不相同，这要取决于特定居住单元中的住户数。因此我们的结论是：设计选项3不能自我加权。

6.4.1. 构建权数方面的案例研究：越南全国健康调查，2001年

26. 现在我们来解释一次实际调查——即越南2001年进行的全国健康调查——是如何构建抽样权数的。这次调查系基于分层三段抽样设计。总共有122个层次，按照61个省内的城市域和农村域来界定。然后在每个层次内独立进行样本选择。第一阶段，使用与规模（1999年人口与住房普查时的住户数）成正比的概率抽样法对公社或市内政区进行选择。第二阶段，按照与被抽样的公社或市内政区的计数区数成反比的抽样率，采取系统抽样法，从每个被抽样的公社或市内政区选定两个计数区。在第三阶段也是最后阶段，从每个计数区选取15个住户，同样采取系统抽样法。

27. 按照全国健康调查设计, 被抽样住户的基本样本权数可以按下述程序构建。设 H_i 和 E_i 分别表示 i 公社的住户数和计数区数(按1999年普查时计), 并且设 H_{ij} 表示公社 i 计数区 j 内的住户数。那么, 在公社 i 计数区 j 内住户 k 的总体选择概率可由下式给出:

$$p_{ijk} = n_c \times \frac{H_i}{\sum_i H_i} \times \frac{2}{E_i} \times \frac{15}{H_{ij}} \quad \text{式中, } n_c \text{ 是在某一特定层次选定的公社数, 而}$$

$\sum_i H_i$ 是该层次的住户总数。

住户抽样权数(w_{ijk})是选择概率的倒数, 亦即:

$$w_{ijk} = \frac{E_i \times H_{ij} \times \sum_i M_i}{30 \times n_c \times H_i} \quad (6.9)$$

6.4.2. 自我加权的样本

28. 当所有被抽样单位权数一样的时候, 该样本就叫做自我加权样本。尽管由于抽样效率的缘故较高阶段的单位往往可以用各种变化不定的概率被选中, 但这些变化不定的概率可能被后来阶段的选择概率所抵消。上面第2个范例中的设计选项2就提供了此种情况的一个实例。

29. 然而在实际当中, 由于下述几方面的原因, 住户调查样本很少可以在国家一级自我加权。首先, 按照设计, 抽样单位往往是采用不均等的选择概率选定的。的确, 虽然时常可以用与规模成正比的概率来选择初级抽样单位, 并且在初级抽样单位中可按适当比率选择住户以实现自我加权的设计, 但可以通过在每个被抽样的住户选定一个人作采访对象而使得这种设计失效。其次, 被选定的样本往往存在包括“不答复”(第6.5节)和“未覆盖”(第6.6节)在内的一些缺陷。第三, 为了获取各种域和特别分组人口的精确估计数, 往往需要对这些域超额抽样, 才能使样本规模达到满足预定精度要求的需要。第四, 如果样本设计需要编列选定类集(初级抽样单位或二段抽样单位)的当前住户清单, 并且预先确定拟在每个类集选出的固定住户数的话, 那么该住户的实际选择概率就会与设计概率有所不同, 因为设计概率系基于标架计数而不是当前的住户计数; 结果, 即便追求自我加权的设计, 也会产生不均等的选择概率。

30. 尽管存在种种障碍, 仍然要把获得自我加权样本当作每项样本设计实践的目标, 因为这些样本不论对实施设计还是对分析设计所产生的数据都是有利的。有了自我加权样本, 就可以从未加权的数据推导出调查估计数; 然后, 如果必要的话, 可以用一个常数因子补充结果, 以获得适当的人口参数估计数。另外, 基于自我加权样本的分析更直截了当, 分析结果更容易被非专业统计工作者和普通百姓所理解。

6.5. 针对不答复的样本权数调整

31. 从调查的所有抽样单位获得全部想要的信息，这种情况十分罕见。譬如讲，有些住户可能根本不提供任何数据，有些住户可能只提供部分数据，就是说，有关所调查的某些问题但不是所有问题的数据。前一类不答复叫做单位不答复或总体不答复，而后一类则叫做项目不答复。如果在答复者与不答复者之间有任何系统差异的话，那么，天真地仅以答复者的答案为依据得出的估计数就会带有偏差。本手册自始至终所强调的良好调查做法的一个关键点，就是必须尽可能把调查中的不答复现象减少到最低程度。为了降低因为没有把人口的特定部分包括在内(或者只包括这部分人口的少得不成比例的一小部分)而致使调查估计数出现偏差的可能性，有必要最大限度地减少不答复。譬如讲，居住在城市地区收入较高的人或许不大可能参加包括收入模式在内的综合调查。如果不能从这部分人口的一大块获得答复，就会影响到有关全国平均住户收入、教育程度、识字率等方面的估计数。

6.5.1. 在住户调查中减少不答复偏差

32. 譬如讲，对于样本平均值来说，不答复的偏差率高低就是两个因素的函数：

- 不答复的人口比例；
- 在有关特征的人口平均值方面，答复群体和不答复群体之间的差异大小。

33. 因此，要想减少不答复造成的偏差就必须：要么降低不答复率，要么缩小答复住户和不答复住户及个人之间的差异。通过妥善保管每个在调查中被选定为抽样单位的记录，有可能直接通过调查数据估计整个样本及相关次级域的不答复率。另外，可以通过精心设计的调研来评估答复者和不答复者之间的差异(Groves和Couper, 1998年)。

34. 与一次性调查或截面调查相比较而言，在群组调查(指在特定时期内从同一抽样单位群组反复收集数据)中，调查设计者可以获得较多的数据，用以研究和调整潜在的不答复误差。这里，发生不答复的原因可能是因为在调查过程中样本单位(资料)遗失，也可能是因为受访者感到厌倦或其他原因而拒绝参加后来几个轮次的调查，等等。这样的话，从以往轮次群组调查中获取的数据就可以用来进一步了解答复者与不答复者之间的差异，并可充当下述调整的依据。关于用来补偿调研中不答复的各种方法，详见Brick和Kalton(1996年)、Lepkowski(2003年)以及其中援引的参考资料。

6.5.2. 不答复的补偿

35. 在住户调查中可以采取多种方法来提高反应率，从而减少与不答复有关联的偏差。其中一种方法是通过“反复回访”来转变拒绝态度，这期间访调员不是一次而是多次尝试完成对一个抽样住户的采访。通过改进访调员的培训也可以提高反应率。然而，无论在提高反应率方面花费多大气力，不答复将总是任何住户调查的一个不可避免的特征。因此，调查设计者往往做一些调整来补偿不答复。针对单位不答复，有如下两种基本调整方法：

(a) 通过提取多于所需户数的初始样本来调整样本规模, 以补偿预期的不答复;

(b) 通过调整样本权数来补偿不答复。

36. 对于住户调查中的单位不答复来说, 最好总是通过调整样本权数来补偿住户不答复。第6.5.3节提供了关于针对不答复进行样本权数调整的步骤概要, 同时提供了一个数字范例。

37. 有几个与替代有关的问题, 它等同于不答复单位整体记录的推算(Kalton, 1983年)。首先是它增加了潜在替代单位的选择概率, 因为接近不答复抽样住户的非抽样住户要比接近于答复抽样住户的非抽样住户有较高的选择概率。其次, 试图替代不答复住户的努力很费时间, 容易出现错误和偏差, 而且非常难以检查和监督。譬如讲, 可能会使用一个方便的住户而不用特别指定为不答复替补单位的住户, 这就带来了产生偏差的另一根源。鉴于所有这些问题, 不应当用替代办法来补偿住户调查中的不答复, 除非有充分理由在特殊情况下采用此法。

38. 对于部分不答复或项目不答复, 标准补偿方法是插补, 但这不属于本手册的涵盖范围。

6.5.3. 针对不答复问题的样本权数调整

39. 在大型住户调查中往往使用样本权数调整程序来补偿不答复的偏差。这种调整基本上是将所有具有被选资格的不答复抽样单位的基权数转移到答复单位, 其实施步骤如下:

- 步骤1. 应用初始设计权数(进行不均等选择概率调整和上几节讨论的其他适用调整);
- 步骤2. 将样本分解成子群并计算每个子群的加权反应率;
- 步骤3. 使用子群反应率的倒数进行不答复调整;
- 步骤4. 计算第*i*个样本的不答复调整权数:

$$w_i = w_{1i} \times w_{2i}$$

(6.10)

式中, w_{1i} 是初始权数, 而 w_{2i} 是不答复调整权数。注意: 加权不答复率可以定义为: 在样本合格的情况下完成的加权采访数与加权合格样本数之比。

范 例

从一个国家的(南、北)两个地区选定了一个由1 000住户组成的分层多级样本。住户抽样率, 在北部地区按1/100, 在南部地区按1/200。城市地区的反应率低于农村地区。设 n_b 表示在*b*层被抽样的住户数; r_b 表示有被选资格且答复调查的住户数; t_b 表示答复调查且享受初级保健服务的住户数。那么, *b*层的住户不答复调整权数可由下式给出:

$$w_b = w_{1b} \times w_{2b} \tag{6.11}$$

式中， $w_{2b} = n_b/r_b$ 。假定给出的分层数据如表6.3所示。

表6.3
不答复加权调整

层 次	n_h	r_h	t_h	W_{1h}	W_{2h}	W_h	$W_h r_h$	$W_h t_h$
北部-城市	100	80	70	100	1.25	125	10 000	8 750
北部-农村	300	120	100	100	2.50	250	30 000	25 000
南部-城市	200	170	150	200	1.18	236	40 120	35 400
南部-农村	400	360	180	200	1.11	222	79 920	39 960
共 计	1 000	730	500				160 040	109 110

因此，享受初级保健服务的估计住户比例是：

$$\hat{p} = \frac{\sum w_h t_h}{\sum w_h r_h} = \frac{109\ 110}{160\ 040} = 0.682 \text{ 或 } 68.2\%$$

而享受初级保健服务的估计住户数是：

$$\hat{t} = \sum w_h t_h = 109\ 110 = 160\ 040 \text{ 户的 } 68.2\%$$

请注意，仅根据答复的数据，享受初级保健服务的未加权估计住户比例是：

$$\hat{p}_{unw} = \frac{\sum t_h}{\sum r_h} = \frac{500}{730} = 0.685 \text{ 或 } 68.5\%$$

而使用初始权数但未经不答复调整的估计比例是：

$$\hat{p}_1 = \frac{\sum w_{1h} t_h}{\sum w_{1h} r_h} = \frac{83\ 000}{126\ 000} = 0.659 \text{ 或 } 65.9\%$$

40. 提供上面范例的另一目的也是为了说明如何通过调整初始权数来补偿不答复。结果显示，在仅使用初始权数的估计比例和使用不答复调整权数的估计比例之间存在相当大的差异，但是未加权的比例和不答复调整比例之间的差异则似乎可以忽略不计。

41. 经过不答复权数调整之后，可以对权数作进一步的适当调整。下一节来考虑通过权数调整来补偿未覆盖。

6.6. 针对未覆盖问题的样本权数调整

42. 未覆盖系指抽样标架未能覆盖所有目标人口，致使一些人口单位没有被纳入住户调查选定样本的选择概率。这只是用于选择调查样本的抽样标架的诸多缺陷之一(关于抽样标架的详细讨论，请参阅第四章)。

43. 未覆盖是住户调查中的一个主要关注问题, 尤其对发展中国家更是如此。根据某些发展中国家的调查得出的人口样本估计数远远低于其他来源的人口估计数, 这一事实证明了未覆盖所可能产生的影响。因此, 鉴别、评价和控制住户调查中的未覆盖问题的方法学应该是国家统计局的一个主要工作和培训领域。

44. 本节讨论住户调查中出现未覆盖问题的根源和用以补偿未覆盖的办法之一, 即通过分层后的统计权数调整。

6.6.1. 住户调查中出现未覆盖问题的根源

45. 在发展中国家, 大多数住户调查都建基于分层多阶段区域概率设计。第一阶段单位, 或称初级抽样单位, 通常是地域单位。第二阶段建立住户清单或居住单元清单, 以便从中选取住户样本。最后阶段产生住所成员或居民清单, 以便从中选取个人样本。由此可见, 未覆盖问题可在三层当中的任何一个层次发生: 在初级抽样单位层次、住户层次或个人层次。

46. 鉴于初级抽样单位一般以前一次人口和住房普查时所确定和使用的计数区为基础, 所以它们有望覆盖目标人口的整个地域范围。因此初级抽样单位的未覆盖面一般很小。对于发展中国家的住户调查来说, 初级抽样单位的未覆盖问题一般不像后面设计阶段的未覆盖那样严重。可是的确大多数调查都存在初级抽样单位未覆盖的问题。譬如讲, 虽然可能把某项调查设计为提供一个国家或该国一个地区整个人口的估计数, 但是因为发生内战或骚乱、自然灾害或其他方面的缘故致使国内有些地区不能通达, 所以在设计阶段就故意把某些初级抽样单位排除在外了。另外, 人烟稀少的边远地区有时也被排除在住户调查的抽样标架之外, 因为覆盖此类地区的成本太高; 由于他们只代表很小一部分人口, 所以他们对总体估计数的影响是很小的(关于住户调查中初级抽样单位的未覆盖问题的详细讨论连同许多范例, 请参见第四章)。在这种调查的成果报告中, 必须明确告知读者本报告已排除了此类地区。不要给人造成这样一种印象: 即调查成果适用于整个国家或地区, 而实际上有部分人口没有包括在内。必须在调查报告中充分指明未覆盖的调查特性。

47. 在住户层次上, 未覆盖的问题就比较严重了。大多数调查认为, 住户通常是由以某种方式相互关联并且通常住在同一住所或居住单元的个人组合而成的集体。这里有些尚待解决的定义问题, 比如什么人可被视为常住居民以及居住单元的构成要件是什么。如何处理多单元结构(比如公寓楼房)和一个居住单元内有多个住户的问题? 也许辨认一个居住单元比较容易, 但是由于社会结构复杂, 要想分辨同一居住单元内的各家住户可能就很困难。因此对于不同的访调员来说, 或者在不同国家或不同的文化环境, 就很可能对这些概念产生误解或有不同的解释。不论在任何情况下, 都需要在诸如什么人可以被视为住户成员以及构成居住单元的要件是什么的问题上对访调员提供业务指导。

48. 导致发生未覆盖问题的其他因素包括: 在现场编列清单时因为疏忽造成的居住单元的遗漏或相关人口分组(比如幼儿或老年人)的遗漏, 由于计量误差造成的遗漏, 没有把不在家的住户成员包括在内, 以及由于未能恰当理解调查概念而造成的遗漏。这个问题还有其时间范围的一面, 就是说, 有的居住单元也许在登记列表的时候没有人住或正在建造之中, 但在收集数据时已经有人住了。对于发展中国家的住户调查来说, 由于它们的普查作为构建

抽样标架的唯一基础大多不能提供住户和个人层次的详细抽样单位地址，因而使得未覆盖的问题变得更加严峻。往往使用的是过了时的或不准确的户口登记，而住户内的个人被有意地或偶然性地从常住户口登记中漏掉了。关于产生未覆盖问题的根源，Lepkowski (2003年) 及其中援引的参考资料提供了更详细的资料。

6.6.2. 住户调查中对未覆盖的补偿

49. 有好几种方法可用于处理住户调查中的未覆盖问题(Lepkowski, 2003年)。其中包括：

- (a) 改进现场工作程序，比如使用多标架的程序，同时改进登记造册的程序；
- (b) 通过统计权数调整补偿未覆盖。

50. 关于第二种方法，如果有关于整个人口和特定人口子群的控制总量数据可用，可以尝试通过使权数总和与特定子群中的控制总量相匹配来调整样本单位权数。这些子群叫做后层次，而这种统计调整程序叫做后分层法。这种程序通过调整某些变量的加权抽样分布以使其同已知的人口分布相一致，来补偿未覆盖(关于如何使用后分层法分析调查数据的一些实际例证，见Lehtonen和Pahkinen(1995年))。下面提供一个简单例证。

范 例

假定在前面的范例中，根据某个独立资料来源——比如现时民事登记册——已知住户数为：北部地区45 025户，南部地区115 800户。进一步设这两个地区的加权样本总数分别为40 000个和120 040个。执行下述两个步骤：

- 步骤1. 计算后分层因子：
 北部地区，我们有： $w_{3b} = \frac{45\ 025}{40\ 000} = 1.126$
 南部地区，我们有： $w_{3b} = \frac{115\ 800}{120\ 040} = 0.965$
- 步骤2. 计算经过调整的最终权数： $w_f = w_b \times w_{3b}$

数字结果归纳在表6.4内：

表6.4
对覆盖面的后分层加权调整

层 次	r_h	t_h	w_h	w_m	$w_m r_h$	$w_m t_h$
北部-城市	80	70	125	140.75	11 260	9 852
北部-农村	120	100	250	281.40	33 768	28 140
南部-城市	170	150	236	227.77	38 721	34 166
南部-农村	360	180	222	214.20	77 112	38 556
共 计	730	500			160 861	110 714

估计能享受初级保健服务的住户比例为:

$$\hat{p}_f = \frac{\sum w_{fb} \hat{p}_b}{\sum w_{fb} n_b} = \frac{110\,714}{160\,861} = 0.69 \text{ 或 } 69\%$$

51. 请注意: 经过后分层加权调整之后, 北部地区和南部地区的加权样本数分别是: 45 024 (11 256+33 768) 和 115 821 (38 709+77 112), 这些数字与上面给出的独立控制总量非常近似。

6.7. 加权导致抽样方差的增加

52. 虽然在调查数据分析中使用权数一般可以减少估计数的偏差, 但是这也会增加这些估计数的方差。为了简化讨论, 让我们来考虑一个使用层内等概率样本的分层单一阶段设计。如果每一层的方差 (亦即层内各单位之间的方差) 各不相同, 那么在各层之间使用不均等的层权数 (比如与层方差成反比的权数) 或许会产生较准确的调查估计数。可是, 如果每一层的层方差都相同的话, 那么, 使用不均等的权数就会导致调查估计数中的方差高于使用均等权数所产生的方差。

53. 使用权数的作用是以如下因子增加估计人口平均数的方差:

$$L = n \times \frac{\sum_b n_b w_b^2}{(\sum_b n_b w_b)^2} \quad (6.12)$$

式中, $n = \sum_b n_b$ 是实现的总体样本规模, w_b 是最终权数, 而 n_b 是在 b 层次上实现的样本规模。上面的公式亦可按照权数变异系数写成:

$$L = n \times \frac{\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j) \quad (6.13)$$

$$\text{式中, } CV^2(w_j) = \frac{\sum_j w_j^2}{(\sum_j w_j)^2} \left[\sum_j w_j^2 - \frac{1}{n} (\sum_j w_j)^2 \right] = \frac{\text{权数方差}}{(\text{平均权数})^2}$$

范 例

现在我们使用第6.6.2节的范例中使用的数据, 来计算方差膨胀因子; 最终的权数 w_b 和实现的各层样本规模 n_b (参见表6.5)。

$$\text{因此, } L = 730 \times \frac{36\,424\,009}{(160\,861)^2} = 1.03$$

换言之, 由于使用权数, 调查估计数中的方差增加了3%。

表6.5
层方差参数

层 次	r_h	w_{fh}	$w_{fh} r_h$	$w_{fh}^2 r_h$
北部-城市	80	140.75	11 260	1 584 845
北部-农村	120	281.40	33 768	9 502 315
南部-城市	170	227.77	38 721	8 819 459
南部-农村	360	214.20	77 112	16 517 390
共 计	730		160 861	36 424 009

6.8. 权数的修整

54. 计算并调整了权数以补偿上述缺陷之后，最好要检查一遍经过调整的权数的分布情况。一些特别大的权数虽然只对一小部分抽样有影响，但是却能导致大幅度增加调查估计数的方差。因此，一般的做法是把特别大的权数削减到某个最大值，以约束相关的权数变异（从而减少调查估计数的方差），同时预防少数抽样单位在总体估计数中占支配地位。在针对不答复调整权数之后，最经常使用的是权数修整法。

55. 虽然权数修整一般能够减少估计数的方差，但是它也会带来估计量偏差。在有的情况下，由于削减特大权数所达到的减少方差的效果可能不止于抵消发生的偏差而已，它还减少了调查估计量的均方误差。在实际当中，只有在适当的时候才进行权数修整，也就是说，在可以证实由于使用修整的权数（而不是原始权数）造成的偏差对总的均方误差产生的影响小于因修整权数导致方差相应减少的情况下。

56. 对于任何分层设计而言，理想上权数修整过程应在每个层次范围内进行。该过程从确定原始权数的上限入手，然后调整整套权数，使修整权数的总和与原始权数的总和相等。设 w_{hi} 表示 h 层第 i 个单位的最终权数，而设 w_{hB} 表示被确定为 h 层权数的上限。那么， h 层第 i 个样本单位的经修整的权数可以界定为：

$$w_{hi}^{(T)} = \begin{cases} w_{hi} & \text{如果 } w_{hi} < w_{hB} \\ w_{hB} & \text{如果 } w_{hi} \geq w_{hB} \end{cases} \tag{6.14}$$

57. 现在，可对整个样本经修整的权数作进一步调整，以便使其总和恰好等于原始权数的总和。为了便于说明起见，我们要假设层内权数固定不变，并且在余下的讨论中省略下标 i 。设 F_T 表示原始权数总和与修整权数总和之比，换言之，

$$F_T = \frac{\sum_b n_b w_b}{\sum_b n_b w_b^{(T)}} \tag{6.15}$$

式中，比率中的两方总和跨越了所有层次和样本中所有单位。如果我们把 h 层内经过调整和修整的权数定义为

$$w_{b(T)}^* = F_{Tb} \times w_{(T)} \quad (6.16)$$

那么很明显, $\sum_b n_b w_{b(T)}^* = \sum_b n_b w_b$ 正如预期。

下面提供的实例意在解释并帮助理解修整的程序。

58. 下面表6.6的前两栏分别给出了7个层次中每一层的单位总数和最终权数。选定最大权数为250, 故原始权数被截至250, 如表内第3栏所示。

表6.6
权数修整

	n_h	w_h	$w_{h(T)}$	$n_h w_h$	$n_h w_{h(T)}$	$n_h w_{h(T)}^*$
	80	140.75	140.75	11 260	11 260	11 823.00
	100	150.25	150.25	15 025	15 025	15 776.25
	125	175.00	175.00	21 875	21 875	22 968.75
	150	200.00	200.00	30 000	30 000	31 500.00
	120	250.00	250.00	30 000	30 000	31 500.00
	120	275.13	250.00	33 015	30 000	31 500.00
	170	285.40	250.00	48 518	42 500	44 625.00
共 计	865			189 693	180 660	189 693.00

注意, 在这种情况下,

$$F_T = \frac{\sum_i n_b w_{bi}}{\sum_i n_b w_{bi(T)}} = \frac{189\,693}{180\,660} = 1.05$$

修整的权数业经重新调整, 每个权数按 $F_T=1.05$ 倍增, 总计为原始权数之和 $\sum_b n_b w_b = 189\,693$ 。

6.9. 结论意见

59. 发展中国家跟世界上其他国家一样, 现在已将样本权数视为住户调查数据分析的组成部分。现在大多数调查方案都主张甚至在涉及自我加权样本的罕见情况下也使用权数(在这种情况下权数为1)。在过去, 调查设计人员曾经花费很大力气想方设法达到基本上达不到的目标——即获取自我加权的样本, 从而实现无需在分析调查数据的时候使用权数。常识告诉人们, 使用权数会使分析过于复杂, 而且非常缺乏(如果有的话)用于加权分析的计算基础设施。然而, 以往10年来计算机技术的进步推翻了这一论断。现在许多发展中国家都买得起计算机硬件和软件。此外, 现在还有许多专门用于分析调查数据的专用计算机软件包。下面第七章对这些软件包进行了综合评述和比较。

60. 如前所述,使用权数可以减少因为样本中与未覆盖和不答复有关的缺陷造成的偏差。不答复和未覆盖是由于所设计的调查未能从目标群体中的某些单位获取信息而产生的不同类型的误差。对于发展中国家的住户调查来说,未覆盖是个比不答复更严重的问题。本章提供了开发并在统计上调整基本权数以补偿某些与住户调查有关的问题,并且在相关参数估算中使用调整权数的程序实例。快速计算机的出现,加之人们已能买得起或可以免费下载的统计软件供应,定会使得甚至发展中国家都能在住户调查数据分析的日常工作中利用权数。不过,正如本章所说明的那样,抽样权数的发展从各种不同方面增加了调查工作的复杂性。比如,需要计算每个选样阶段的权数;然后需要对这些权数加以调整以补偿样本中的各种不足之处;最后还需要妥善储存并用于所有日后的分析。因此,必须精心开展加权业务并实际计算调查分析中要使用的权数。

参考资料和补充读物

- Brick, J. M. 和G. Kalton(1996年)。“调研中遗漏数据的处理”,《医学研究中的统计方法》,第5卷,第215-238页。
- Cochran, W. G. (1977年)。《抽样方法》(第3版)。纽约:约翰·威利国际出版公司。
- Groves, R. M. 和M. P. Couper(1998年)。《住户采访调查中的不答复问题》。纽约:约翰·威利国际出版公司。
- Groves, R. M. 等(2002年)。《调查中的不答复》。纽约:约翰·威利国际出版公司。
- D. Kasprzyk(1986年)。“遗漏调查数据的处理”,《调查方法》,第12卷,第1-16页。
- Kalton, G. (1983年)。《遗漏调查数据的补偿》。密歇根州安阿伯:密歇根大学调查研究中心。
- Kish, L. (1965年)。《调查抽样》。纽约:威利国际出版公司。
- I. Hess(1950年)。“关于住所抽样未覆盖的问题”,《美国统计协会会刊》,第53卷,第509-524页。
- Lehtonen, R. 和E. J. Pahkinen(1995年)。《复杂样本设计与分析实用方法》。纽约:威利国际出版公司。
- Lepkowski, James(2005年)。“发展中国家住户调查中的非观察误差”,《发展中国家和转型期国家住户样本调查》,《方法研究》,第96号,出售品编号:E. 05. XVII. 6。
- Lessler, J. 和W. Kalsbeek(1992年)。《调查中的非抽样误差》。纽约:约翰·威利国际出版公司。
- Levy, P. S. 和S. Lemeshow(1999年)。《人口抽样:方法和应用》(第3版)。纽约:约翰·威利国际出版公司。
- Lohr, S. (1999年)。《抽样:设计与分析》。加利福尼亚州, Pacific Grove: 达克斯伯里出版公司。

Yansaneh, I. S. (2004年)。《发展中国家住户调查样本设计问题概述》，“发展中国家和转型期国家住户调查：技术报告”。联合国，纽约。

第七章

调查数据抽样误差的估算

7.1. 引言

1. 本章简要综述用于估算各种样本设计产生的住户调查数据抽样误差的各种方法，这些样本设计包括从可在有关抽样理论的任何入门教科书(比如Cochran(1977年))中找到的标准设计到用于大规模住户调查的较为复杂的设计，范围很广。在标准样本设计方面，提供了公式及数字例证，以解释如何进行抽样误差的估算、置信区间的构建，以及设计效果和有效样本规模的计算。然后提出一些较复杂设计的抽样误差估算方法。讨论了每一种方法的优缺点并提供了数字例证，用以说明如何实施相关程序。根据实际调查数据提供了一个实例来说明这样一个事实：有些标准统计软件包低估了调查估计数中存在的抽样误差，从而导致对调查所关切的某些参数得出错误结论。为了避免此类问题，本章强烈建议使用专用统计软件包，因为这些软件考虑到住户调查通用设计的复杂性质。本文对其中几种软件包作了介绍和比较。

7.1.1. 复杂调查数据的抽样误差估算

2. 近年来，对于精心设计的住户调查来说，其分析目标已经超出了基本汇总计数表或相关参数总计的范畴。当今的分析家们还对假设的形成与检验或模型构造感兴趣。举例来讲，现在的分析家们不是简单地估计贫困人口或受过中等或高等教育的人口比例，而是想要评估相关的政策影响，或者探索某个关键答复变量——比如在校儿童的学业成绩，或一个住户的贫困程度——如何受到各种因素的影响，诸如地区因素、社会经济因素、性别和年龄等方面的因素。

3. 回答此类问题，需要详细分析住户或个人层次的数据。发表这些分析结果的时候必须包括从调查数据中推导出来的各种估计数的精度或准确度的计量标准。为了正确使用和解释调查结果，同时为了评价和改进样本设计和程序，就需要有关于调查估计数精度的信息。尤其在大规模全国调查方案的情况下这种对样本设计的监测与评估更加重要，因为这种调查方案往往是有关各种各样议题的详细信息的唯一来源。

4. 在样本调查精度方面，一个关键量度就是抽样方差(第三章介绍了这个概念)，这是一个在假定调查收集的信息正确的前提条件下，通过选择一个样本而无需查点整个人口所引进的可变性指标。抽样方差是一个估计量抽样分布的可变性量度。标准误差，或方差平

方根, 用来计量抽样误差。对于任何特定调查来说, 这种抽样误差的一个估计量是可以评估的, 并可用来显示估计数的精度。

5. 方差估计量的形式和如何对它进行评估取决于基本样本设计。对于标准设计来说, 这些估计量往往是用简单公式来评估的。可是对于住户调查所用的复杂设计而言, 由于经常涉及到分层、聚类和非均等概率抽样, 所以这些估计量的形式就比较复杂, 难以进行评估。在这种情况下, 计算抽样误差就需要所用的程序要考虑到产生数据的样本设计的复杂程度, 而这往往又会反过来要求使用适当的计算机软件。

6. 在许多发展中国家, 住户调查数据的分析经常局限于基本表格分析, 连带给出平均数、比例和总计的估计数, 但是不提供这些估计数精度或准确度的任何指标。甚至在拥有广泛统计数据收集和处理基础设施的国家统计机构, 人们也往往发现缺乏详细分析低层次数据的专门技能。有些调查设计者或分析家, 当他们得知下述事实的时候会感到意外, 比如: 由要素聚类带来的要素间的相关会减损与其惯常分析的简单随机样本有关的估计数的精度; 或者, 使用权数进行分析一般会使抽样误差膨胀; 或者, 他们日常在工作中使用的标准软件包不能适当地弥补上述精度方面的损失。

7. 本章通过简要综述在通常用于发展中国家住户调查的复杂设计中计算抽样误差估计数的各种方法, 以及这种调查分析中所使用的各种统计软件包, 试图来纠正上面讲的那些情况。提供了几个数字实例来解释所讨论的方差程序。

7.1.2. 概览

8. 第7.2节提供了简单随机抽样情况下抽样方差的一个第一原则定义, 其中包括用数字实例来说明如何计算抽样方差和构建置信区间。第7.3节提供了关于抽样误差其他量度的定义。第7.4节提出了在各种标准设计——比如分层抽样和整群抽样——中计算抽样方差的公式; 为便于理解相关概念, 提供了几个数字实例。第7.5节讨论住户调查设计的共同特征, 以及为适当估算抽样误差所需的调查数据的内容与结构。另外还提出了住户调查所关注的估计数的一般形式。第7.6节提供了关于陈述抽样误差信息的简要指南, 而第7.7节描述了计算较为复杂设计中抽样误差的实用方法。这些方法往往需要专门的程序并使用专用计算机软件包。第7.8节以1989年布隆迪进行的关于免疫接种覆盖面调查所获数据为例, 讨论了在使用标准统计软件分析调查数据方面容易犯的错误。第7.9节和7.10节对现有的一些公开用于估算住户调查数据抽样误差的软件包进行了评述和比较。第7.11节提出本章的结论性意见。

7.2. 简单随机抽样项下的抽样方差

9. 一个估计数的抽样方差可以定义为该估计数平均值的均方偏差, 其中“平均”的意思系指所有可能的样本的平均。正如第三章指出的那样, 简单随机抽样是最基本的抽样方法, 但在大规模的调查中很少使用, 因为它的实施效率很低, 而且费用高昂, 令人望而却步。

10. 为了便于理解抽样方差的概念, 让我们来考虑一个5住户的小人口($N=5$), 采用无替代的简单随机抽样法(SRSWOR)从其中选取一个2住户的样本(规模 $n=2$)。设关切变量为每月家庭食品支出水平, 并且假定5个住户中每户的支出额如下面表7.1所示:

表7.1
每月每户食品支出额(美元)

住户(i)	按美元计算食品支出额(Y_i)
1	10
2	20
3	30
4	40
5	50

11. 首先请注意, 既然已知人口中所有住户的关切变量值, 我们就能计算对应于月平均住户食品支出的参数值, 亦即:

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^N Y_i = \frac{10+20+30+40+50}{5} = \frac{150}{5} = 30$$

无替代的简单随机抽样法的月平均食品支出估计量为:

$$\hat{Y} = \frac{1}{2} \sum_{i \in S} Y_i$$

式中, 总和是样本所选取的各个单位相加之和。显然, 所得估计数取决于所选样本。下面的表7.2显示了所有可能的样本, 基于每种样本的估计数, 每个样本估计数偏离人口平均数的差幅, 以及平方偏差。注意, \hat{Y}_{ave} 表示所有基于样本的估计数的平均数。并且注意, \bar{Y} 代表人口平均值, 而 \hat{Y} 代表人口平均值的估计数, 叫做样本平均值(见表7.2)。

表7.2
计算平均参数 \bar{Y} 的实际抽样方差

样本1	样本单位	样本估计数(\hat{Y})	$\hat{Y} - \hat{Y}_{ave}$	$(\hat{Y} - \hat{Y}_{ave})^2$
1	(1,2)	15	-15	225
2	(1,3)	20	-10	100
3	(1,4)	25	-5	25
4	(1,5)	30	0	0
5	(2,3)	25	-5	25
6	(2,4)	30	0	0
7	(2,5)	35	5	25
8	(3,4)	35	5	25
9	(3,5)	40	10	100
10	(4,5)	45	15	225
平均值		30	0	750

注意, 基于所有可能的样本的估计数平均值为:

$$\hat{Y}_{ave} = \frac{1}{10} \sum_{i=1}^{10} \hat{Y}_i = \frac{15+20+25+30+25+30+35+35+40+45}{10} = \frac{300}{10} = 30 = \bar{Y}$$

12. 换言之, 基于所有可能的样本的估计数平均值等于人口平均值。带有此种特性的估计对于它所估计的参数来说叫做无偏估计。

13. 采用无替代的简单随机抽样法对该人口中的规模 $n=2$ 的两户样本月平均食品支出估计量的实际抽样方差为

$$Var(\hat{Y}) = \frac{1}{10} \sum_{i=1}^{10} (\hat{Y}_i - \hat{Y}_{ave})^2 = \frac{750}{10} = 75$$

14. 采用上述方法的问题出在从人口中选择所有可能的样本不切实际。实际上只选了一个样本, 而且该人口的各种数值不得而知。一个比较实用的方法是使用计算方差的公式。所有标准抽样设计中都有这种公式。

15. 按照无替代的简单随机抽样法, 一个基于规模为 n 的样本的估计平均值 (\hat{Y}) 的抽样方差由下式给出:

$$Var(\hat{Y}) = \left(1 - \frac{n}{N}\right) \frac{\delta^2}{n} \quad (7.1)$$

$$\text{式中, } \delta^2 = \frac{\sum_{i=1}^N (Y_i - \bar{Y})^2}{N-1}$$

是相关特征可变性的量度(人口方差 Y)。通常, δ^2 是未知的, 只能根据样本估计(见下面方程式 7.2)。从以上公式可以清楚地看出, 抽样方差取决于下列因素:

- (a) 相关特征的人口方差;
- (b) 人口规模;
- (c) 样本规模;
- (d) 样本设计和估算方法。

16. 样本中的人口比例, n/N , 叫做抽样份额(用 f 表示); 而没有包括在样本中的那部分人口比例, 因子 $[1 - (n/N)]$, 或 $1-f$, 叫做有限人口纠正因子 (fpc)。 fpc 代表对估计数的标准误差所作的调整, 以说明该样本是在没有来自有限人口替代的情况下选取的这一事实。不过请注意, 在抽样份额很小的情况下 fpc 可以忽略不计。实际当中, 如果 fpc 在 5% 以下便可忽略不计 (Cochran, 1977年)。

17. 上面的公式表明, 抽样方差与样本规模成反比。随着样本容量增大, 抽样方差便减小; 而在普查或完全调查(即 $n=N$) 的情况下, 不会出现抽样方差。注意, 不答复确实会减小样本规模, 从而增加抽样可变性。

18. 估计平均值抽样方差的无偏估计数可由下式给出:

$$v(\hat{Y}) = (1 - \frac{n}{N}) \frac{s^2}{n} \quad (7.2)$$

$$\text{式中, } s^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{n-1}$$

是基于样本人口方差估计数, δ^2 。这叫做样本方差。人口平均值的95%置信区间(见第三章第30段)可由下式给出:

$$\hat{Y} \pm 1.96\sqrt{v(\hat{Y})} \quad (7.3)$$

19. 对于一个人口比例而言, 基于样本的估计数和估计方差可分别由下面两式给出, 即:

$$\hat{P} = \frac{\text{有特征的单位数}}{n} \quad (7.4)$$

$$\text{和 } v(\hat{P}) = (1 - \frac{n}{N}) \frac{\hat{P}(1 - \hat{P})}{n-1} \quad (7.5)$$

20. 下面表7.3归纳了在无替代的简单随机抽样下各种人口数量的估计数和估计数方差。

表7.3
选定的人口特征估计数及其方差

参 数	估 计 数	估计数方差
人口平均值(\bar{Y})	$\hat{Y} = \frac{1}{n} \sum_{i \in \text{Sample}} Y_i$	$v(\hat{Y}) = (1 - \frac{n}{N}) \frac{s^2}{n}$
人口总数	$\hat{T} = N\hat{Y}$	$v(\hat{T}) = N^2 v(\hat{Y})$
特定类别的人口比例	$\hat{P} = \frac{\text{特定类别抽样单位数}}{n}$	$v(\hat{P}) = (1 - \frac{n}{N}) \frac{\hat{P}(1 - \hat{P})}{n-1}$

21. 一般情况下, 人口平均值的 $(1-\alpha)\%$ 的置信区间由下式给出:

$$\text{估计数} \pm z_{1-\alpha/2} \sqrt{\text{估计的估计数方差}} \quad (7.6)$$

式中, $z_{1-\alpha/2}$ 是标准正常分布的第 $(1-\alpha/2)$ -th个百分位数。

22. 下面这个范例旨在解释基于选定样本的抽样方差的估算方法。

范 例 1

让我们考虑从一个 $N=20\ 000$ 住户的大人口中选取的一个 $n=20$ 的简单随机样本。收集的数据列于下面表7.4, 其中 Y 表示每周住户食品支出额, 而变量 Z 则表明住户有无电视机($z=1$ 表示“有”, 0 则表示“无”)。

表7.4
被抽样的住户每周食品支出情况和电视机拥有情况

住户 (i)	Y_i	Z_i	i	Y_i	Z_i
1	5	0	11	7	1
2	10	1	12	8	1
3	5	0	13	9	1
4	9	1	14	10	1
5	5	1	15	8	1
6	6	1	16	8	0
7	7	0	17	5	0
8	15	1	18	7	0
9	12	1	19	12	1
10	8	0	20	4	0
总 计		160	12		

人口平均每月住户食品支出估计数为:

$$\hat{Y} = \frac{1}{20} \sum_{i=1}^{20} Y_i = \frac{5+10+\dots+12+4}{20} = \frac{160}{20} = 8$$

估计平均数的估计方差为:

$$v(\hat{Y}) = \left(1 - \frac{20}{20\ 000}\right) \left\{ \frac{(5-8)^2 + (10-8)^2 + \dots + (12-8)^2 + (4-8)^2}{19} \right\} = 7.87$$

人口平均值的95%置信区间是:

$$8 \pm 1.96 \times \sqrt{7.87} = (2.50, 13.50)$$

整个人口每月住户食品支出总额估计为:

$$\hat{Y} = N\hat{Y} = 20\ 000 \times 8 = 160\ 000$$

估计总额的估计方差为:

$$v(\hat{Y}) = 20\ 000^2 \times 7.87 = 3\ 148\ 000\ 000$$

人口平均值的95%置信区间是:

$$160\ 000 \pm 1.96 \times \sqrt{3\ 148\ 000\ 000} = (50\ 030, 269\ 970)$$

家有电视机的住户在人口中所占比例的估计数为:

$$\hat{P} = \frac{1}{20} \sum_{i=1}^{20} Z_i = \frac{12}{20} = 0.6$$

家有电视机的住户在人口中所占比例估计数的估计方差为：

$$v(\hat{P}) = \left(1 - \frac{20}{20\,000}\right) \frac{0.6(1-0.6)}{19} = 0.0126$$

人口平均值的95%置信区间为：

$$0.6 \pm 1.96 \times \sqrt{0.0126} = (0.38, 0.82)$$

7.3. 抽样误差的其他量度

23. 除了抽样方差之外，还有其他一些抽样误差量度。其中包括标准误差、变异系数和设计效果。这些量度都是在代数上相互关联的，因为可以使用简单代数运算从其他量度推导出其中任何一种量度的表达式。

7.3.1. 标准误差

24. 一个估计量的标准误差是其抽样方差的平方根。这个量度比较好解释，因为它表明了使用与估计数同等尺度的抽样误差，而方差则建基于平方差异。

25. 在调查设计方面经常提出的一个问题是，据认为多大的标准误差是可以接受的。对这个问题的回答取决于估计数的大小。比如，100的标准误差在估计年收入的时候可能被认为是小的，但在估计人们的平均体重的时候就被认为是大的了。同样，对于上面范例1中所得160 000的估计总数来说，也可以认为 $\sqrt{3\,148\,000\,000} = 56\,107$ 的标准误差是过大的。

7.3.2. 变异系数

26. 一个估计数的变异系数(CV)是其标准误差对估计数本身平均数之比。于是，变异系数提供了一个与所计量特性有关的抽样误差的量度。通常以百分比表示。

27. 变异系数在比较规模或尺度各不相同的调查估计数精度方面很有用。但是对于实值可能为零或负数的特征估计数——包括变异估计量，比如两年内平均收入的变化——来说，没有什么用途。

7.3.3. 设计效果

28. 设计效果(表示为 d_{eff})的定义是：在特定设计项下，一个估计量的抽样方差对基于同等规模简单随机样本估计量的抽样方差之比。可以把它当作一个因数，必须用它乘以基

于同等规模简单随机样本的估计数方差, 以便考虑到由于分层、聚类和加权等因素造成的实际样本设计的复杂性。

29. 换言之, 一个基于出自规模为 n 的复杂样本的数据估计量之方差, 等同于根据出自规模为 $n/deff$ 的简单随机样本的数据计算的估计量之方差。因此, 有时候把 $n/deff$ 这个比率叫做为估算基于出自复杂设计的数据所需的有效样本规模。关于“有效样本规模”的一般讨论, 见Kish(1995年)和Potthoff、Woodbury和Manton(1992年)以及其中援引的参考资料。另见第三章有关各节中关于设计效果及其在样本设计中使用的较详细讨论。

7.4. 计算其他标准设计的抽样方差

30. 对于简单设计和诸如平均值、比例和总计之类的简单线性估计数而言, 推导出公式来计算估计数方差, 通常是可能的。然而对与住户调查有关的复杂设计和估计数来说, 这就往往很困难或者不可能。我们将在本节用实例来说明如何计算分层的和单一阶段类集样本设计的抽样方差。有关的教科书(比如Cochran, 1977年; 和Kish, 1965年)中提供了其他标准样本设计关于计算方差的公式和范例。

7.4.1. 分层抽样

31. 第三章和附件一提供了对分层抽样的详细描述。我们在本节将只是关注此种设计项下的方差估算问题。让我们来考虑一项有 H 个层次的分层设计, 这些层次的各项人口平均值的样本估计数分别由 $\bar{y}_1, \bar{y}_2, \dots, \bar{y}_H$ 给出, 而这些层次的人口方差的样本估计数分别由 $S_1^2, S_2^2, \dots, S_H^2$ 给出。本设计项下的一个人口平均数估计量由下式给出:

$$\hat{Y}_{st} = \sum_{b=1}^H \hat{Y}_b \quad (7.7)$$

式中, \hat{Y}_b 是基于样本的估计数 $\bar{y}_b, b=1, \dots, H$ 。估计量的方差由下式给出:

$$v(\hat{Y}_{st}) = \sum_{b=1}^H v(\hat{Y}_b) \quad (7.8)$$

采用分层随机抽样法, 估计量及其估计方差由下式给出:

$$\hat{Y}_{st} = \sum_{b=1}^H \frac{N_b}{N} \bar{y}_b = \sum_{b=1}^H W_b \bar{y}_b \quad (7.9)$$

式中, \bar{y}_b 是 b 层的样本平均值; N_b 是 b 层的总体规模; 而 $W_b = \frac{N_b}{N}, b=1, \dots, H$ 。

这个在分层随机抽样设计项下估计数的估计方差由下式给出：

$$v(\hat{Y}_{st}) = \sum_{b=1}^H W_b^2 v(\bar{y}_b) = \sum_{b=1}^H \left(\frac{N_b}{N} \right)^2 \left(1 - \frac{n_b}{N_b} \right) \frac{s_b^2}{n_b} \quad (7.10)$$

式中， n_b 是 h 层的样本规模，而 s_b^2 是一个基于样本的估计数 S_b^2 ， $b=1, \dots, H$ 的样本方差。

范 例 2

现在我们将上述结果应用到涉及三个层次的分层设计的一个实例，各项参数如下面表7.5所示。假定我们想要根据一个规模为1 500的总体样本来估算人口平均数。

表7.5
分层样本设计数据实例

参 数	人 口	第1层 (首都)	第2层 (省辖城市)	第3层 (省辖农村)
规模	$N=1\,000\,000$	$N_1=300\,000$	$N_2=500\,000$	$N_3=200\,000$
方差	$S^2=75\,000$	$S_1^2=?$	$S_2^2=?$	$S_3^2=?$
平均值	$\bar{Y}=?$	$\bar{Y}_1=?$	$\bar{Y}_2=?$	$\bar{Y}_3=?$
每个单位的成本	N/A	$C_1=1$	$C_2=4$	$C_3=16$
最佳分配下的样本规模*	$n=1\,500$	$n_1=857$	$n_2=595$	$n_3=48$
样本平均值	N/A	$\bar{y}_1=4\,000$	$\bar{y}_2=2\,500$	$\bar{y}_3=1\,000$
样本方差	N/A	$s_1^2=90\,000$	$s_2^2=62\,500$	$s_3^2=10\,000$

注：N/A标记不适用。

* 见第三章。

人口平均值估计数是：

$$\hat{Y}_{st} = \frac{300\,000}{1\,000\,000} \times 4\,000 + \frac{500\,000}{1\,000\,000} \times 2\,500 + \frac{200\,000}{1\,000\,000} \times 1\,000 = 2\,650$$

上述估计数的估计方差是：

$$v(\hat{Y}_{st}) = \left(\frac{300\,000}{1\,000\,000} \right)^2 \left(1 - \frac{857}{300\,000} \right) \left(\frac{90\,000}{857} \right) + \left(\frac{500\,000}{1\,000\,000} \right)^2 \left(1 - \frac{595}{500\,000} \right) \left(\frac{62\,500}{595} \right) + \left(\frac{200\,000}{1\,000\,000} \right)^2 \left(1 - \frac{48}{200\,000} \right) \left(\frac{10\,000}{48} \right) = 43.98516$$

人口平均值的95%置信区间是：

$$2\,650 \pm 1.96 \times \sqrt{43.98516} = (2\,637, 2\,663)$$

注意, 简单随机抽样设计项下估计平均值的估计方差是:

$$v(\hat{Y}_{SRS}) = \left(1 - \frac{1\,500}{1\,000\,000}\right) \times \frac{75\,000}{1\,500} = 49.925$$

因此, 此项分层设计的设计效果是 $\frac{43.98516}{49.925} = 0.88$, 而有效样本规模是 $\frac{1\,500}{0.88} = 1\,705$ 。

这就意味着, 基于规模为1 500的分层随机样本的估计数方差与基于规模为1 705的简单随机样本的估计数方差是一样的。

32. 第三章详细描述了整群抽样方法。我们在本节提供一个简单实例来说明在单一阶段整群抽样的特殊情况下如何计算抽样误差。

范 例 3

假定我们想要估计某省接种脑脊髓灰质炎的学龄儿童在人口中所占的比例。为简化起见, 设该省有500个大小相等的计数区, 每个区有25名学龄儿童。在本例中计数区充当类集。假设我们使用无替代的简单随机抽样法从该省500个计数区选取了10个计数区, 并且在每个抽样计数区计量接种疫苗的学龄儿童比例, 结果如下面表7.6所示。

表7.6
作为关切变量在10个计数区内接种疫苗的学龄儿童比例

抽样的计数区(i)	1	2	3	4	5	6	7	8	9	10
样本比例(\hat{P}_i)	$\frac{8}{25}$	$\frac{10}{25}$	$\frac{12}{25}$	$\frac{14}{25}$	$\frac{15}{25}$	$\frac{17}{25}$	$\frac{20}{25}$	$\frac{20}{25}$	$\frac{21}{25}$	$\frac{23}{25}$

对于本例而言, 该省接种疫苗的学龄儿童比例估计数为:

$$\hat{P} = \frac{160}{250} = 0.64 \text{ 或 } 64\%$$

另外, 样本方差是:

$$s_p^2 = \frac{1}{10-1} \sum_{i=1}^{10} (\hat{P}_i - \hat{P})^2 = 0.040533$$

因此, 估计比例方差是:

$$v(\hat{P}) = \left(1 - \frac{10}{500}\right) \times \frac{0.040533}{10} = 0.003972$$

注意, 采用简单任意抽样法, 估计比例的估计方差为:

$$v(\hat{P}_{SRS}) = \left(1 - \frac{250}{12\,500}\right) \times \frac{0.64(1-0.64)}{250-1} = 0.0009078$$

因此, 该类集样本的设计效果是 $\frac{0.003972}{0.0009078} = 4.38$, 而有效样本规模是 $\frac{250}{4.38} = 57$ 。

这就意味着，基于规模为250的类集样本的估计数方差与基于规模为57的简单随机样本的估计数方差是一样的。

7.5. 住户调查样本设计和数据的共同特征

7.5.1. 住户调查设计偏离简单随机抽样的差幅

33. 如前所述，在大规模住户调查中很少实际运用简单随机抽样法，因为其执行费用过于昂贵。可是有必要透彻理解此种设计，因为它构成了较复杂样本设计的理论基础。由于下述三个特征中的一种或多种原因，大多数住户调查样本设计都偏离了简单随机抽样法：

- (a) 在一个或多个抽样阶段分层；
- (b) 在一个或多个抽样阶段实施单位聚类，这样可以降低成本，但是由于在同一类集存在各种相关，会增大估计数方差；
- (c) 为补偿非均等选择概率、不答复和未覆盖之类的样本缺陷而进行加权(详见第六章)。

34. 如果一个样本设计有上面讲的一个或更多特征，就算是复杂设计。大多数住户调查设计都是复杂设计，因此不符合简单随机抽样的假设条件。所以，假如把住户调查数据当作简单随机样本设计所产生的数据来分析，就会导致在分析中和基于此种数据的推论中出现误差。另外，如前所述，大多数住户调查中的有关估计数都不能表示为观测值的线性函数，因此不大可能有任何闭合型估计方差的公式。下面几节在考虑到上述复杂因素的情况下，处理住户调查设计中有关估计方差的方法问题。

7.5.2. 供分析用数据文档的准备

35. 发展中国家收集的调查数据有时候不适合用来进行超出基本频率和列表的分析。有几方面的原因。第一，有关调查样本设计的技术文献可能很有限或根本没有。第二，数据文档可能没有适合进行任何复杂分析的格式、结构和必要信息。第三，可能缺少适当的计算机软件和技术专长。

36. 为了能对样本调查数据进行适当分析，有关的数据库必须载有反映样本选择过程的所有信息(详见第五章讨论)。尤其是数据库应包含诸如“样本设计分层”、“初级抽样单位”(PSUs)、“二次抽样单位”(SSUs)……之类的标签。有时候，为了进行方差估计需要对实际分层和用于选择调查样本的初级抽样单位加以修改。为了使实际样本设计适合采用至少一个统计分析软件包所能提供的某种样本设计选项，上述修改是必要的(见第7.9节)。为估计方差而创造的层次和初级抽样单位有时叫做“假层次”(或方差层)和“假初级抽样单位”(或方差初级抽样单位)。相关的样本设计变量以及为了估计方差而创造的变量应连同关于如何定义和使用这些变量的相应文件一起包含在数据文档之中。方差估计至少需要一套三个变量，即：样本权数、层(或假层)和初级抽样单位(或假初级抽样单位)。这三个

变量概括了样本设计, 把它们纳入调查数据集即可进行适当的数据分析, 并说明样本设计的复杂因素。

37. 另外, 要为数据文档中的每个抽样单位拟定样本权数。这些权数要能反映每个抽样单位的选择概率并且补偿调查不答复及样本中的其他缺陷。调查估计数可变性的适当估算需要有样本权数和设计变量的标签。第六章和本章的前几节提到, 样本权数不仅对产生适当的调查估计数是重要的, 而且对估算这些估计数的抽样误差也是必不可少的。因此, 必须把有关权数的全部信息纳入数据文档。尤其是, 每当进行不答复、后分层或其他类调整的时候, 调查文件中都必须载有关于这些调整程序的描述。

7.5.3. 调查估计数的类别

38. 对大多数住户调查来说, 最常用的重要调查估计数均以总计和比率的形式表示。假设一个分层三阶段设计: 第一阶段是初级抽样单位; 第二阶段是二级抽样单位; 第三阶段是住户。总体调查估计数可以表示为:

$$\hat{Y} = \sum_{b=1}^H \sum_{i=1}^{n_b} \sum_{j=1}^{m_i} \sum_{k=1}^{l_j} W_{bijk} Y_{bijk} \quad (7.11)$$

式中, W_{bijk} 是在***b***层($b=1, \dots, H$)第***i***初级抽样单位($i=1, \dots, n_b$)第***j***二级抽样单位($j=1, \dots, m_i$)的第***k***个住户($k=1, \dots, l_j$)的最终权数; Y_{bijk} 是在***b***层第***i***初级抽样单位第***j***二级抽样单位选择的第***k***个住户的变量***Y***的数值。

39. 在最基本层面, 与样本单位有关的权数与纳入该样本的各单位的选择概率成反比。不过, 在计算准备用于分析的权数方面往往使用较复杂的方法。其中有些方法已在第六章及其援引的参考资料中讨论过。

40. 一个比率的调查估计数由下式界定:

$$\hat{R} = \frac{\hat{Y}}{\hat{X}} \quad (7.12)$$

式中, \hat{Y} 和 \hat{X} 分别是按照上面方程式(7.12)计算的变量***Y***和***X***的总计估计数。

41. 在多级抽样的情况下, 平均数和比例只不过是比率估计量的特殊情况。在平均数的情况下, 在比率的分子部分, 变量***X***是一个计数变量, 对每个要素来说都界定为等于1, 所以分母就是权数的总和。在比例的情况下, 分母中的变量***X***也被界定为对所有要素来说都等于1。而在分子中的变量***Y***是一个二项变量, 被界定为要么等于零, 要么等于1, 这要取决于所观察的单位是否具有其比例正在被估计的特性。在大多数住户调查中, 比率估计量中的分母定义各种各样, 诸如“总人口”、“女性总人数”、“男性总人数”、“农村总人口”、“某省或某区总人口”, 等等。

7.6. 抽样误差信息的表述准则

7.6.1. 确定报告什么内容

42. 对于有着无计其数的关切变量和域、且有好几个往往相互竞争的目标的大规模全国调查来说，若要求每个估计数都连同其相关的抽样误差一起提供，那是不切实际的。这不仅会大大增加出版量，而且有可能使重要成果的陈述变得杂乱无章。从抽样误差估计数本身的预期可变性来看，若提供太多个别变量结果的话，可能会导致混乱并且使人感觉所收集的调查数据总体质量缺乏一致性。一种有用得多的做法是只对一些属于头等关切事项的最重要特征提供抽样误差信息，而将其余抽样误差汇编为一个附录。

43. 在提供有关抽样误差信息的时候，有必要留心它对调查成果的解释以及根据这种解释制定政策的决定有何潜在影响。无论何时都要把抽样误差信息视为只是总调查误差的一部分，而且不一定是最重要的部分。在某些调查情况下，非抽样误差(见第八章)对调查数据总体质量的影响比抽样误差还要大。因此建议有关抽样误差的信息要包括有关非抽样误差的主要来源的讨论及其对调查数据总体质量影响的定性评估。鉴于抽样误差在较低分解层次变得至关重要，所以还建议包含一些关于调查数据可以分解到何种程度的告诫性提示。

44. 一般来说，关于抽样误差的信息应包含足够的细节，以便于纠正调查结果的解释，并且满足广大数据用户群体的要求，其中包括：从一般数据用户或决策者(他们的利益在于使用调查结果制订政策)和独立分析家(他们从事调查结果的进一步分析和报告)，到抽样统计工作者(他们关注的是相对于其他备选方案而言特定设计的统计有效性，以及或许可用于未来调查的设计特征)。

7.6.2. 如何报告抽样误差信息

45. 可以采用下述三种不同的形式来表述抽样误差：

- (a) 标准误差的绝对值；
- (b) 相对标准误差(对方差的平方根)；
- (c) 置信区间。

46. 这三种表述形式当中选择哪一种，取决于估计数的性质。在估计数大小和计量单位不尽相同的情况下，同样的标准误差值可能适用于用相对值表达的估计数。因此，提供相对标准误差可能较为有效。可是一般来说，绝对标准误差在理解上和与估计数的联系上要容易多了，尤其在涉及百分比、比例和比率的时候更是如此。使用置信区间需要选择一个置信度(比如规定为90%、95%或99%)。由于置信度因调查目的和估计数的精度要求不同而异，所以在表述抽样误差信息的时候有必要说明所使用的置信度，并且在确定调查结果重要性的全过程要始终坚持这一置信度。如前所述，实际当中最常用的置信区间是95%(见第三章第30段和第七章第22段)，亦即：

$$\text{估计数} \pm 1.96 \times \text{标准误差}$$

(7.13)

47. 关于抽样误差的表述, 包括对各类用户的具体指导及诸多例证, 详见联合国(1993年)及其中援引的参考资料。

7.6.3. 报告标准误差方面的经验法则

48. 在报告标准误差中经常运用的经验法则是标准误差报告到两位有效数字, 然后将相应的点估计报告到与标准误差相同的小数位。比如:

1. 如果点估计为73 456, 标准误差为2 345, 那么我们将点估计报告为73 500, 将标准误差报告为2 300。
2. 如果点估计为1. 54328, 标准误差为0. 01356, 那么我们将点估计报告为1. 543, 将标准误差报告为0. 014。

49. 经验法则背后的论证似乎与t-统计有关。标准误差中的两位有效数字和点估计中的相应数位确保了t-统计结果中没有太大的四舍五入误差效应, 同时又避免了报告的点估计通过太多无关紧要的数位提供的过高精度。不过请注意, 在t-统计不是主要关切事项的情况下这种经验法则未必奏效。

7.7. 住户调查中估计方差的方法

50. 我们在本节简要描述一些用来估算基于调查数据估计数的方差或抽样误差的常规方法。住户调查抽样误差的估算方法可以分为四大类:

- (a) 精确法;
- (b) 最终类集法;
- (c) 线性化近似法;
- (d) 重复法。

下面, 我们将逐一简略讨论这些方法。有兴趣的读者可以从如下参考资料获取更详细的资料: Kish和Frankel(1974年); Wolter(1985年); 以及Lehtonen和Pahkinen(1995年)。

7.7.1. 精确法

51. 第7.2节和7.4节提供了估算标准样本设计方差的精确法的几个例子。这些方法构成在适用情况下处理方差估算问题的最佳途径。然而, 用这些方法来计算基于住户调查数据的估计数抽样方差因为涉及到几方面的因素而变得复杂化了。首先, 大多数住户调查所使用的样本设计都比简单随机抽样法复杂(见上面第7.5.1节)。其次, 相关的估计数或许不是以观察值的简单线性函数形式出现, 致使抽样方差往往不能像简单随机抽样设计或分层抽样设计中的样本平均数公式那样, 用一个闭合性的公式来表达。另外, 精确法有赖于所考虑的特定制度的样本设计、相关的估计数以及所使用的加权程序。

52. 下面各节, 我们来讨论住户调查中常用的几种样本设计方差估算方法。这些方法旨在克服精确法的一些缺陷。

7.7.2. 最终类集法

53. 最终类集方差估算方法(见Hansen、Hurwitz和Madow, 1953年, 第257-259页)用来估算基于复杂样本设计所产生的一个样本调查估计数的方差。按照这种方法, 最终类集是由出自一个初级抽样单位的整个样本组成, 而不管多级设计的后来阶段抽样如何。仅使用初级抽样单位之间的总数来计算方差估计数, 而无须计算每个选择阶段的方差成分。

54. 假定从 h 层选定了由 n_h 个初级抽样单位组成的一个样本(初级抽样单位内的选样阶段数不限)。然后, h 层的总估计数可由下式给出:

$$\hat{Y}_b = \sum_{i=1}^{n_b} \hat{Y}_{bi}, \quad \text{式中, } \hat{Y}_{bi} = \sum_{j=1}^{m_i} W_{hijk} Y_{hijk} \quad (7.14)$$

注意, 初级抽样单位层次的估计数 \hat{Y}_{bi} 是 $\frac{\hat{Y}_b}{n_b}$ 的一个估计数。因此, 单个初级抽样单位层次的估计数方差可由下式给出:

$$v(\hat{Y}_{bi}) = \frac{1}{n_b - 1} \sum_{i=1}^{n_b} \left(\hat{Y}_{bi} - \frac{\hat{Y}_b}{n_b} \right)^2 \quad (7.15)$$

而根据一个规模为 n_b (h 层总人口估计量)的随机样本估计的它们的总数 \hat{Y}_b (即层次一级的总数)的方差则由下式给出:

$$v(\hat{Y}_b) = \frac{n_b}{n_b - 1} \sum_{i=1}^{n_b} \left(\hat{Y}_{bi} - \frac{\hat{Y}_b}{n_b} \right)^2 \quad (7.16)$$

55. 注意, 通过简单代数运算产生了 h 层总人口的方差估计量的如下同等表达式:

$$v(\hat{Y}_b) = \frac{n_b}{n_b - 1} \left\{ \sum_{i=1}^{n_b} \hat{Y}_{bi}^2 - \frac{\left(\sum_{i=1}^{n_b} \hat{Y}_{bi} \right)^2}{n_b} \right\} \quad (7.17)$$

56. 最后, 通过跨层次独立抽样, 取层次一级的各总数方差之和, 获得总人口的方差估计量, 亦即:

$$v(\hat{Y}) = \sum_{b=1}^H v(\hat{Y}_b) \quad (7.18)$$

注意, 有时候在上面公式中使用一个有限人口校正系数 $(1 - n_b/N_b)$ 。

57. 估计总数方差只是经过适当加权的初级抽样单位总数 \hat{Y}_{bi} 的一个函数, 它与初级抽样单位内的抽样结构和方式不相干; 在这个意义上, 方程式 (7.18) 值得注意。这就大大简化了方差估算公式, 因为它不需要计算归属于初级抽样单位范围内其他抽样阶段的方差成分。这一特征使得最终类集方法在处理不同样本设计方面有了很大灵活性, 而且这的确是该方法的主要优点之一, 也是它广泛用于调查工作的一个主要原因。

58. 现在, 比率 $\hat{R} = \frac{\hat{Y}}{\hat{X}}$ 的方差估计量由下式给出:

$$v(\hat{R}) = \frac{1}{\hat{X}^2} \left\{ v(\hat{Y}) + R^2 v(\hat{X}) - 2\hat{R} \text{cov}(\hat{Y}, \hat{X}) \right\} \quad (7.19)$$

式中, $v(\hat{Y})$ 和 $v(\hat{X})$ 是根据估计总数方差公式计算的, 而

$$\text{cov}(\hat{Y}, \hat{X}) = \sum_{b=1}^H \left\{ \frac{n_b}{n_b-1} \sum_{i=1}^{n_b} \left(\hat{X}_{bi} - \frac{\hat{X}_b}{n_b} \right) \left(\hat{Y}_{bi} - \frac{\hat{Y}_b}{n_b} \right) \right\} \quad (7.20)$$

或者等同于:

$$\text{cov}(\hat{Y}, \hat{X}) = \frac{n_b}{n_b-1} \left\{ \sum_{i=1}^{n_b} \hat{X}_{bi} \hat{Y}_{bi} - \frac{\left(\sum_{i=1}^{n_b} \hat{X}_{bi} \right) \left(\sum_{i=1}^{n_b} \hat{Y}_{bi} \right)}{n_b} \right\} \quad (7.21)$$

59. 注意, 可以通过利用相对比率方差约等于分子和分母相对方差之间的差异这一事实来简化上面的比率方差公式。忆及: 一个估计量的相对方差是其方差对其平方之比。于是, 对于一个估计比率 \hat{R} 而言, 用 $relvar(\hat{R})$ 表示的相对方差可由下式给出:

$$relvar(\hat{R}) = \frac{v(\hat{R})}{\hat{R}^2} \quad (7.22)$$

因此, 比率方差的估计数可由下式给出:

$$v(\hat{R}) = \hat{R}^2 relvar(\hat{R}) = \hat{R}^2 \left\{ relvar(\hat{Y}) - relvar(\hat{X}) \right\} \quad (7.23)$$

60. 用于计算估计总数和比率的抽样误差的最终类集方法, 可以系统化概括为下列步骤:

- 步骤1. 每一层单独计算每个初级抽样单位关切特征 Y 的加权估计数 \hat{Y}_{bi} (按照第六章指明的加权程序);
- 步骤2. 计算步骤1得出的每个初级抽样单位估计值的平方值;

- 步骤3. 计算步骤2得出的该层所有初级抽样单位估计值的总和;
 - 步骤4. 计算从步骤1得出的所有初级抽样单位估计总数的总和;
 - 步骤5. 求步骤4的结果的平方值并除以该层初级抽样单位数 n_b ;
 - 步骤6. 从步骤3的结果中减去步骤5的结果, 并且用因子 $n_b/(n_b-1)$ 乘以这一差额。这是在层次一级的特征估计方差;
 - 步骤7. 合计所有各层执行步骤6的结果, 以获得关切特征的总体估计方差;
 - 步骤8. 计算步骤7所得结果的平方根, 以获得关切特征的估计抽样误差。
61. 为了计算估计比率(比如估计比例)的抽样误差, 我们接着执行以下步骤:
- 步骤9. 通过用相关分子估计数的平方值除以步骤7的结果, 计算该分子 \hat{y} 的相对方差;
 - 步骤10. 重复步骤9, 以获得分母 \hat{x} 的相对方差;
 - 步骤11. 从步骤9的结果中减去步骤10的结果;
 - 步骤12. 用估计比率 \hat{r} 的平方值乘以步骤11的结果。这是 \hat{r} 的估计方差;
 - 步骤13. 计算步骤12所得结果的平方根, 以获得 \hat{r} 的估计抽样方差。

范 例 4

现在我们来考虑一个假定范例, 以说明最终类集方差估计方法。假定我们想要估计A市住户每周食品支出总额。我们利用一个分层三阶段类集设计法来设计此项调查, 其中涉及到三个层次, 每层选取两个初级抽样单位, 从每个被抽样的初级抽样单位选定两个住户。然后在调查中记录每个被抽样的住户每周的食品支出费用。表7.7列出了这项调查所获数据, 其中包括权数(W_{hij})和每个被抽样的住户每周食品支出额(Y_{hij})(以美元计)。

表7.7
按层每周每户食品支出额

层	初级抽样 单位	住 户	权数 (W_{hij})	以美元计 支出额 (Y_{hij})	$W_{hij} * Y_{hij}$
1	1	1	1	30	30
		2	1	28	28
	2	1	3	12	36
		2	3	15	45
2	1	1	5	6	30
		2	5	7	35
	2	1	2	16	32
		2	2	18	36
3	1	1	6	7	42
		2	6	8	48
	2	1	4	13	52
		2	4	15	60
共 计			42		474

62. 根据第六章, 该市住户的每周食品支出总额由下式给出:

$$\hat{Y} = \sum_{b=1}^3 \sum_{i=1}^2 \sum_{j=1}^2 W_{bij} Y_{bij} = 474$$

另外还可以由下式给出住户平均每周食品支出额:

$$\hat{\bar{Y}} = \frac{\sum_{b=1}^3 \sum_{i=1}^2 \sum_{j=1}^2 W_{bij} Y_{bij}}{\sum_{b=1}^3 \sum_{i=1}^2 \sum_{j=1}^2 W_{bij}} = \frac{474}{42} = 11 \text{ (最接近的美元)}$$

63. 现在我们来实施列入下面表7.8的最终类集方差估算方法的各个步骤。其中栏目数对应于上面概括的各步骤。

表7.8
实施最终类集方差估算方法中的各项步骤

层 次	初级抽样							
	单 位	步骤1	步骤2	步骤3	步骤4	步骤5	步骤6	步骤7
1	1	58	3 364	9 925	139	9 660.5	529	
	2	81	6 561	-	-	-	-	
2	1	65	4 225	8 849	133	8 844.5	9	
	2	68	4 624	-	-	-	-	
3	1	90	8 100	20 644	202	20 402	484	
	2	112	12 544	-	-	-	-	
总 计								1 022

64. 层次的方差估计数, 第1层是529; 第2层是9; 第3层是484。通过连加各个层级的估计数, 其总和为1 022, 获得每周住户估计总收入(在我们的方案中是步骤7)的方差估计数。

7.7.3. 线性化近似法

65. 住户调查中的大多数关切估计数都是非线性的。这方面的实例有: 一个国家学龄儿童的体格指数, 特定城市居民住房费用占其收入的比例, 一个人口子集具有某种特征的可能性与另一分组人口具有同样特征的可能性之比, 等等。在线性化方法中, 使用一种泰勒级数展开式将这种非线性估计数“线性化”。这一过程涉及到按照泰勒级数展开式来表达估计数, 然后使用前面几节讨论的精确方法, 用该展开式的一阶或线性部分的方差来逼近估计数方差。

66. 假定我们想要估算一个参数 Z 的估计值 z 的方差, 并且假设 z 是参数 Y_1, Y_2, \dots, Y_m 的简单估计数 y_1, y_2, \dots, y_m 的一个非线性函数, 亦即:

$$z = f(y_1, y_2, \dots, y_m) \quad (7.24)$$

设 z 接近于 Z ，那么按照一级 $z-Z$ ，泰勒级数展开式 z 就是：

$$z = Z + \sum_{i=1}^m d_i (y_i - Y_i) \quad (7.25)$$

式中， d_i' 是就 y_i' 而言 z 的偏导数，亦即，

$$d_i = \frac{\partial z}{\partial y_i}$$

它是基本估计数 y_i 的一个函数。这就是说，可以用上面方程式(7.24)中的线性函数方差逼近 z 的方差，我们知道如何用前面几节提出的精确方法来计算它的值，亦即：

$$v(z) = v\left(\sum_{i=1}^m d_i y_i\right) = \sum_{i=1}^m d_i^2 v(y_i) + \sum_{i \neq j} d_i d_j \text{cov}(y_i, y_j) \quad (7.26)$$

67. 方程式(7.26)涉及到包括 m 个基本估计数 y_1, y_2, \dots, y_m 并具有 m 方差项和 $m(m-1)/2$ 相同协方差项的一个 $(m \times m)$ 协方差矩阵，可以根据前几节讨论的线性统计精确方法对其加以评估。

范例5(比率方差)

为了说明线性化方法，让我们来考虑比率方差的估算：

$$z = r = \frac{y}{x} \quad (7.27)$$

注意，在这种情况下， $\frac{\partial r}{\partial y} = \frac{1}{x}$ 而 $\frac{\partial r}{\partial x} = -\frac{y}{x^2} = -\frac{r}{x}$ 因此，

$$v(r) = \frac{1}{x^2} \{v(y) + r^2 v(x) - 2r \text{cov}(y, x)\} \quad (7.28)$$

这是在大多数抽样教科书中都能找到的人们所熟悉的比率方差表达式。

68. 线性化在实践中广为应用，因为它几乎可以用于所有样本设计和可以线性化——亦即作为人们熟悉的统计数字(比如平均数或总数)的一个线性函数——且带有来自泰勒级数展开式所需偏导数的系数的任何一项统计。一俟非线性估计数方差被线性化，即可使用上述精确方法对其予以逼近(关于线性化程序技术细节连同相关例证，详见Cochran(1977年)和Lohr(1999年))。

7.7.3.1 优点

69. 由于线性化方差估算方法业经长期使用，它已在理论上有了扎实的发展，并且在样本设计方面比在重复方法方面适用范围更广(见下文)。如果已知偏导数，而泰勒级数展开

式中的二次及更高阶项小到可以忽略不计, 那么, 线性化即可为差不多所有相关线性估计量 (比如比率和回归系数) 产生近似方差估计数。

7.7.3.2 局限性

70. 只有在上述有关偏导数和高阶项的假设正确的情况下, 线性化才能奏效。否则就会导致估计数出现严重偏差。另外, 很难把这个方法应用于涉及权数的复杂函数。必须为每一种估计量单独研拟公式, 而这可能需要专门编程。这个方法不能应用于不构成人口总数或平均数的修匀函数的统计数据, 比如中位数和其他百分位数。

71. 再者, 也很难将线性化方法应用于不答复和未覆盖的调整补偿方面, 这要取决于样本设计、关切估计数和加权程序。它还要求在数据文档中包含相关的样本设计信息 (诸如分层、初级抽样单位、权数等)。

7.7.4. 重复法

72. 重复法是这样一类方法, 即从数据中重复选取子样本 (或称重复品), 重新计算每个重复品及整个样本的加权调查估计数, 然后计算作为这些重复品估计数偏离整个样本的差幅函数的方差。此法可以概括为下列步骤:

- 步骤1. 从整个样本中删除不相同的子样本, 以形成重复品样本;
- 步骤2. 通过重复每个重复品样本的估算过程产生重复权数;
- 步骤3. 从整个样本中以及从每套重复权数中产生一个估计数;
- 步骤4. 计算估计数方差, 作为出自整个样本估计数的重复估计数的平方偏差。

73. 举例来讲, 从一个样本产生了 k 个重复品, 各有一个参数 θ 的估计数 $\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k$, 并且假定基于整个样本的估计数是 $\hat{\theta}_0$ 。那么, 基于重复法的方差估计数可由下式给出:

$$\text{var}(\hat{\theta}) = \frac{1}{c} \sum_{r=1}^k (\hat{\theta}_r - \hat{\theta}_0)^2 \quad (7.29)$$

式中, c 是一个取决于估算方法的常数。各种重复方法在该常数值以及形成重复品的方式上不尽相同 (关于最常用的重复技巧, 见第 7.7.5 节的简要评述)。

7.7.4.1. 数据文档结构

74. 不管采用何种重复技巧, 数据文档结构都要保持跟下面表 7.9 所示结构相一致。

7.7.4.2. 优点

75. 重复法相对于线性方法的主要优点, 就是重复法使用相同基础估计数方法, 不管所估计的统计数字如何 (因为方差估计数是一个样本函数, 而不是估计数的函数); 而线性

表7.9
重复方法的数据文档结构

案卷	数据	整个样本 权数	重复品权数				
			1	2	3	k
1	数据1	w_1	w_{11}	0	w_{13}	w_{1k}
2	数据2	w_2	0	w_{22}	w_{23}	w_{2k}
3	数据3	w_3	w_{31}	w_{32}	0	w_{3k}
.....
.....
.....
N	数据n	w_n	w_{n1}	w_{n2}	w_{n3}	0

注：表内省略号表示数列。

化近似法必须在分析上用于每个统计数据，由于在大规模住户调查中有大量关切特征需要调查，所以线性方法会非常耗费人工。另外，重复法使用方便，且几乎所有统计数据不论线性还是非线性皆适用。使用重复法可以很容易地计算分组人口的估计数，并可在重复品权数中体现不答复调整和其他调整。

7.7.4.3 局限性

76. 重复法是属于计算机密集型的，主要因为此类方法需要计算一套重复权数，这些是为每个选中的重复品重新计算的分析权数，以便于每个重复品都能适当代表跟整个样本一样的人口。另外，重复品的形成也可能因为在样本设计方面的局限性而被复杂化(见下面第7.7.5节)，因而有时可能会导致过高估计抽样误差。

77. 关于抽样误差估算的线性化方法与重复法的全面比较，我们的结论是：这两种方法不会产生相同的抽样误差估计数。而经验调查结果(见Kish和Frankel, 1974年)表明，对于大样本和许多统计数据来说，这两种方法所产生的结果之间存在的差异小到可以忽略不计。

7.7.5. 一些重复技巧

78. 最常用的重复技巧有：

- (a) 随机组合法；
- (b) 均衡多次重复法(BRR)；
- (c) 刀切重复法(JK1、JK2和JKn)；
- (d) 自助法。

现在，我们来逐一简要讨论其中每种技巧。

7.7.5.1. 随机组合法

79. 随机组合法需要在保留调查设计的情况下将整个样本划分为 k 个小组, 也就是说, 每组代表调查的一个微型化版本, 同时能反映样本设计。举例来讲, 如果整个样本是一个规模为 n 的简单随机样本, 那么就可以通过将 n 个观察数据按比例随机分配到 k 个小组来组成随机组, 每组的规模为 n/k 。如果它是一个类集样本, 那么就将初级抽样单位随机划分到 k 个小组, 让每个初级抽样单位都能保留其所有观察数据, 因此每个随机组仍然是个类集样本。如果整个样本是个分层多级样本, 那就可以通过从每层选取一个初级抽样单位来组成随机组。注意, 在这种情况下组成的随机组数不得超过在最小层次选取的初级抽样单位数。

80. 随机组合法可以很容易地用于估计线性统计数据(如平均数和总数及其修匀函数)和非线性统计数据(如比率和百分位数)的抽样误差。不需要专用软件来估计抽样误差, 因为这只是基于整个样本的随机组合估计数的标准偏离。但在复杂样本设计中可能很难产生随机组, 因为每个随机组必须保留完整调查的设计结构。另外, 调查设计本身也可能会制约随机组的数量。譬如讲, 对于每层有两个初级抽样单位的设计来讲, 只能形成两个随机组, 而一般来说随机组数少就会导致抽样误差估计数不准确。一般经验法则是, 至少要有10个随机组才能获得较稳定的抽样误差估计数。

7.7.5.2. 均衡多次重复法

81. 均衡多次重复法(BRR)设想了一种每层有两个初级抽样单位的设计。形成重复品需要将每个层次分成两个初级抽样单位, 并且依照规定的模式在每层选定两个初级抽样单位之一来代表整个层次。通过将初级抽样单位组合成假层, 每层两个初级抽样单位, 可以使这种均衡多次重复法适应其他设计。

7.7.5.3. 刀切重复法

82. 跟均衡多次重复法一样, 刀切重复法是随机组合法的一种一般化方式, 它可以允许重复组重叠。有三种类型的刀切法, 即: JK1法、JK2法和JK n 法。

83. JK1法是简单随机抽样设计的典型“删一单位”刀切法。不过, 它亦可用于其他设计, 条件是抽样单位被组合成随机子集, 每个子集都与整个样本相类似。

84. JK2法类似于均衡多次重复法, 即设定为每层两个初级抽样单位的设计。在自我代表初级抽样单位的情况下, 可以产生成对的二次抽样单位(SSUs)。跟均衡多次重复法一样, JK2法也可以通过把初级抽样单位组合成假层、每层有两个初级抽样单位来适用于其他设计。然后依次从每层随机删除一个初级抽样单位, 以便形成重复品。

85. JK n 法是分层设计的典型“删一单位”刀切法。为了产生重复品, 依次从每层删除一个初级抽样单位。该层剩下的初级抽样单位再加权以估算该层总体。重复品数与初级抽样单位数(或假初级抽样单位数)相等。

7.7.5.4. 自助法

86. 自助法从选择复制了整个人口的所有重要特征的完整样本入手。然后把整个样本当作整个人口，从中抽取子样本。跟前面讲的一样，根据基于再抽样的估计数偏离基于整个样本的估计数的标准差幅来计算抽样误差的估计数。

87. 自助法非常适合一般样本设计和未修匀函数如百分位数。不过，与其他重复方法相比而言，它需要较多的计算。

88. 下面表7.10列举了与各种重复方法相对应的方差公式(方程式(7.28))中的常数 c 的值。

表7.10
各种重复方法的方差公式中的常数因子值

重复方法	方程式(7.28)中的常数 c 值
随机组合	$k(k-1)$
均衡多次重复(BRR)	k
JK1法	1
JK2法	2
JK n 法	$k/(k-1)$
自助法	$k-1$

范例6(刀切重复法)

现在，我们来提供一个应用方差估算刀切重复法的简单数字范例。假定我们有一个规模为3的样本。我们可以创建3个子样本；通过从整个样本中依次删除一个单位，每个子样本规模为2。下面表7.11给出了变量(Y)的值。对于这3个子样本而言，一个“X”表示样本的那些单位包含在子样本中。

表7.11
对一个小样本及其子样本应用方差估算刀切重复法

样本单位	变量(Y)	子样本(g)		
	Y	1	2	3
1	5	X	X	
2	7	X		X
3	9		X	X
样本总数	21			
样本平均数	7	6	7	8

$$\text{样本方差: } s^2 = \frac{(5-7)^2 + (7-7)^2 + (9-7)^2}{3-1} = 4$$

$$\text{样本平均数方差估计数(忽略} fpc \text{): } \frac{s^2}{n} = \frac{4}{3}$$

各子样本平均值的平均数: $\frac{\bar{y}_1 + \bar{y}_2 + \bar{y}_3}{3} = \frac{6+7+8}{3} = 7$

样本平均数的刀切方差估计数由下式给出:

$$V_j(\bar{y}) = \frac{n-1}{n} \sum_{g=1}^3 (\bar{y}_g - \bar{y})^2 = \frac{3-1}{3} [(6-7)^2 + (7-7)^2 + (8-7)^2] = \frac{4}{3}$$

结果与上面计算的样本平均数估计方差完全相同。

范例7(重复品的形成)

表7.12使用范例4(第7.7.2节)中的数据来说明各种重复方法形成重复品的情况以及使用刀切重复法计算方差的情况。

表7.12
整个样本: 按层支出

层 次	初级抽样单位	住 户	权数(W_{hj})	支出(Y_{hj})	$W_{hj} * Y_{hj}$
1	1	1	1	30	30
1	1	2	1	28	28
1	2	1	3	12	36
1	2	2	3	15	45
2	1	1	5	6	30
2	1	2	5	7	35
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
3	2	1	4	13	52
3	2	2	4	15	60
共 计			42		474

$$\text{基于整个样本的估计平均数} = \hat{Y}_0 = \frac{\sum_{b=1}^3 \sum_{i=1}^2 \sum_{j=1}^2 W_{bij} Y_{bij}}{\sum_{b=1}^3 \sum_{i=1}^2 \sum_{j=1}^2 W_{bij}} = \frac{474}{42} = 11$$

表7.13
刀切重复法(从第1层删除第2个初级抽样单位)

层 次	初级抽样单位	住 户	权数(W_{hj})	支出(Y_{hj})	$W_{hj} * Y_{hj}$
1	1	1	1	30	30
1	1	2	1	28	28
2	1	1	5	6	30
2	1	2	5	7	35
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
3	2	1	4	13	52
3	2	2	4	15	60
共 计			36		393

基于上面重复品样本的估计平均数 $= \hat{Y}_1 = \frac{393}{36} = 11$

我们可以继续这个程序，从每一层依次删除一个初级抽样单位。这样可以形成总共6个重复品样本。下面表7.14列举了基于这6个重复品样本的平均每周住户收入估计数。

表7.14
基于重复品的估计数

重复品 j	删除的初级抽样单位 (PSU)	估计数 \hat{Y}_j	$\hat{Y}_j - \hat{Y}_0$	$(\hat{Y}_j - \hat{Y}_0)^2$
1	PSU2, 第1层	11	0	0
2	PSU1, 第1层	10	-1	1
3	PSU2, 第2层	10	-1	1
4	PSU1, 第2层	12	1	1
5	PSU2, 第3层	12	1	1
6	PSU1, 第3层	13	2	4
共 计				8

估计平均数方差的刀切重复估计数由下式给出：

$$\text{var}_{JK}(\hat{Y}) = \sum_{b=1}^H \left\{ \frac{n_b - 1}{n_b} \sum_{j=1}^{n_b} (\hat{Y}_j - \hat{Y}_0)^2 \right\} = \frac{1}{2} \times 8 = 4$$

(注：在本例中， $H=3$ ；对所有 b 而言 $n_b=2$ 。)

89. 我们在本节的结尾再举一个使用均衡多次重复法形成重复品样本的范例。下面表7.15列出了采用标题中规定的初级抽样单位删除方式得出的结果。

表7.15
均衡多次重复法
(从第1、第3层删除第2个初级抽样单位；从第2层删除第1个初级抽样单位)

层 次	初级抽样单位	住 户	权数 (W_{hj})	支出 (Y_{hj})	$W_{hj} * Y_{hj}$
1	1	1	1	30	30
1	1	2	1	28	28
2	2	1	2	16	32
2	2	2	2	18	36
3	1	1	6	7	42
3	1	2	6	8	48
共 计			18		216

基于以上均衡多次重复样本的估计平均数 $= \hat{Y}_{1,BRR} = \frac{216}{18} = 12$ 。

7.8. 在使用标准统计软件包分析 住户调查数据方面易犯的错误

90. 对住户调查数据的适当分析需要在考虑到产生这些数据的设计复杂性的情况下计算估计数的抽样误差。这种复杂因素包括分层、聚类、不均等概率抽样、不答复和其他方面的样本权数调整(关于权数的开发与调整, 详见第六章)。标准统计软件包不能应付这些复杂因素, 因为它们一般假定样本要素是采用简单随机抽样法从人口中选取的。正如第六章所说明的那样, 人口参数的点估计受到与每项观察有关的样本权数的影响。这些权数取决于选择概率以及诸如分层和聚类等其他调查设计特征。当标准软件包忽视了样本权数的时候, 它们就会产生带有偏差的点估计。使用这些软件包进行加权分析可以在一定程度上减少点估计偏差, 但即便如此, 点估计的抽样误差也往往在总体上被低估, 因为方差估算程序一般不考虑诸如分层和聚类等其他设计特征。这就意味着从这些分析中得出的推论会产生误导作用, 譬如讲, 可能会把人群之间的差异错误地宣布为非常之大, 或者错误地拒绝一些假设。比如住户调查数据分析所产生的推论, 就会对国家和地区一级的资源分配和政策制订产生重大影响。

91. 现在我们借用Brogan(2004年)中的一个实例来说明使用标准统计软件包可能会导致点估计偏差、不适当的标准误差和置信区间, 以及带有误导性的重要性测试等。该实例系基于1989年在布隆迪进行的破伤风类毒素(TT)疫苗接种覆盖面抽样调查所产生的数据集。这次调查的目的之一就是比较血清反应阳性(其定义是有破伤风类毒素疫苗接种史的滴定率不少于0.01国际单位/毫升(IU/ml)的破伤风抗毒素)。关于此项调查的方法论及其发表成果, 见Brogan(2004年)和其中援引的参考资料。下面表7.16列出了血清反应阳性妇女的百分比估计数以及相关的标准误差和置信区间。

92. 注意, 对于使用权数的所有方案来说点估计数都是一样的, 但是在加权和未加权的估计数之间有明显差别。另外, 由适当软件产生的标准误差几乎是假定简单随机抽样的标准软件包产生标准误差的两倍。换言之, 标准软件包严重低估了调查估计数方差。这或许会产生重要的政策影响。譬如讲, 如果计划根据65%以下的血清反应阳性百分比采取某种干预措施, 那么, 根据使用专用软件包的分析结果有可能实施这种干预, 但是根据使用标准软件包的分析结果也可能不实施干预。表7.16表明, 能适当估计调查估计数方差的软件包产生的结果大体相同。下一节我们将简略评述现有的一些公开用于住户调查数据分析的统计软件包。

7.9. 抽样误差估算用的计算机软件

93. 上述抽样误差估算方面的各种方法已在发达国家长期使用, 主要通过由官方统计机构、学术机构和私人调查组织开发的专用化计算机算法来实施。近年来计算机技术的进步导致开发了多种软件包来实施这些技术。其中有不少软件包现已用在个人计算机上。软件包只把一种普通方法用于第7.7节讨论的方差估计。这些软件包大多产生最常用的估计数, 诸如平均数、比例、比率和线性回归系数等。有些软件包还包括一系列广泛估计量的约计法, 比如逻辑回归系数。

表7.16
使用各种软件包估计调查估计数方差：在最近生育过的
布隆迪妇女中血清反应阳性妇女的百分比(1988-1989年)

软件包	血清反应阳性百分比	标准误差	95%置信区间
SAS 8.2 无加权	74.9	2.1	(70.8, 79.0)
SAS 8.2 加权	67.2	2.3	(62.7, 71.7)
SAS 8.2 调查平均数	67.2	4.3	(58.8, 75.6)
SUDAAN 8.0	67.2	4.3	(58.8, 75.6)
STATA 7.0	67.2	4.3	(58.8, 75.6)
EPI INFO 6.04d	67.2	4.3	(58.8, 75.6)
WESVAR 4.1	67.2	4.3	(58.8, 75.6)

94. 在本节我们来简要评述现有的一些公开用于住户调查数据抽样误差估计的软件。这绝非一份现有可用程序和软件包的详尽无遗的清单。我们只想集中讨论现有可供一般调查数据分析人员在个人计算机上使用的一些统计软件包。对每一种软件包都进行简要评述，指明适用的样本设计和方差估计方法。同时突出使用相关软件的优缺点。无意详细介绍构成相关软件包基础的技术和计算程序。这些信息均可从相关软件包的网站以及本章末尾提供的参考资料中获得。

95. 这里审查的六个软件包是：CENVAR、EPI INFO、PC CARP、STATA、SUDAAN和WESVAR。其中，SUDAAN(Shah, Barnwell和Bieler, 1998年)、STATA(StataCorp, 1996年)、PC CARP(Fuller等, 1989年)和CENVAR这几种软件包都使用线性化方法估计非线性统计数据的抽样误差。WESVAR仅使用重复法。SUDAAN的最近版本也应用均衡多次重复法和刀切重复法。另外，SAS和SPSS(本文未予评述)这两个软件为调查数据分析开发了新的程序块。基于重复法的程序提供了许多基本方法，但不包括自助法。我们只对这些软件包的通用特征加以简略比较。深入的比较需要在各种不同规模的样本调查之间进行更广泛的比较并且掌握更多的统计数据，这就超越了本节的评述范围。

96. 关于本章评述的几种统计软件包的互联网链接以及其他许多软件包的链接，均可登陆如下网站查阅：www.fas.harvard.edu/~stats/survey-soft/survey-soft.html。

97. Brogan(2004年)根据上述布隆迪住户调查提供了若干种统计软件包的详细比较，其中包括本文评述的这些软件包。

98. 现在我们来简要评述各软件包。有兴趣的读者可以从时新的各种手册或从下面指明的网站获取更详尽的资料。

CENVAR

美国人口普查局；联系单位：国际方案中心

United States Bureau of the Census; contact International Programs Center

United States Bureau of the Census

Washington, D.C. 20233-8860

E-mail: IMPS@census.gov

网址：www.census.gov/ipc/www/imps

99. CENVAR是美国人口普查局设计用于复杂调查数据处理、管理和分析的一个统计软件系统的组件, 整个系统叫做“综合微型计算机处理系统”(IMPS)。它适用于大多数样本设计, 诸如简单随机抽样、分层随机抽样, 以及多级整群抽样等。CENVAR使用线性化近似法进行方差估计。

100. CENVAR所产生的估计数包括整体样本的平均数、比例和总数以及列表配置的特定子集。除了抽样误差之外, 该软件还提供95%置信区间限度、变异系数、设计效果和无加权样本规模等。

EPI INFO

美国疾病预防控制中心

United States Centers for Disease Control and Prevention

Epidemiology Program Office, Mailstop C08

Centers for Disease Control and Prevention

Atlanta, GA 30333

E-mail: EpiInfo@cdc1.cdc.gov

网址: <http://www.cdc.gov/epiinfo/>

101. EPI INFO是美国疾病预防控制中心设计的一个统计软件系统, 用于处理、管理和分析包括复杂调查数据在内的流行病学数据(CSAMPLE的组成部分)。相关的文件可登陆该程序网站查阅, 并可逐章打印。它专门用于通过最终整群抽样模型进行分层多级整群抽样。

102. EPI INFO所产生的抽样误差估计数包括整体样本的平均数和比例以及双向设计的子集。打印输出品包括无加权频率、加权比例或平均数、标准误差、95%置信区间限度和设计效果。

PC CARP

艾奥瓦州立大学

Iowa State University

Statistical Laboratory

219 Snedecor Hall

Ames, IA 50011

网址: <http://cssm.iastate.edu/software/pccarp.html>

103. PC CARP是在艾奥瓦州立大学开发的一个统计软件包, 用于估计平均数、比例、分位数、比率、比率差异以及双向列联表分析的标准误差。该程序是为处理分层多级类集样本而设计的。PC CARP使用方差估计的线性化方法。

STATA

斯塔塔公司

Stata Corporation

702 University Drive East
College Station, TX 77840
E-mail: stata@stata.com
网址: <http://www.stata.com>

104. STATA是为平均数、总数、比率、比例、线性回归、逻辑回归以及概率单位分析程序等方面的抽样误差估计而设计的一个统计分析软件包。其附加能力包括参数的线性组合估计和假设测试, 还有分位数估计、列联表分析、遗漏数据弥补以及其他分析。STATA使用方差估计的线性化方法。

SUDAAN

研究三角研究所
Research Triangle Institute
Statistical Software Center
Research Triangle Institute
3040 Cornwallis Road
Research Triangle Park, NC 27709-2194
E-mail: SUDAAN@rti.org
网址: <http://www.rti.org/patents/sudaan.html>

105. SUDAAN是一个用于分析包括复杂调查数据在内相关数据的统计软件包。它适用于多种多样的设计, 其中包括简单随机抽样和多级分层设计。它为估计一系列统计数据及其相关抽样误差提供了方便, 其中包括平均数、比例、比率、分位数、横列表和机会比率; 线性回归模型、逻辑回归模型和比例危险回归模型; 以及列联表分析。该程序使用方差估计的线性化方法。

WESVAR

韦斯塔特有限公司
Westat, Inc.
1650 Research Blvd.
Rockville, MD 20850-3129
E-mail: WESVAR@westat.com
网址: <http://www.westat.com/wesvar/>

106. WESVAR是韦斯塔特公司设计的一个统计软件系统, 用于分析复杂调查数据, 其中包括列联表分析、回归和逻辑回归。它适用于大多数样本设计, 但专门是为基于最终整群抽样模型的分层多级类集样本设计的。

107. WESVAR使用方差估计的重复方法, 其中包括刀切法、均衡半样本和对均衡半样本法的Fay修正。它需要在一特定WESVAR格式中创建新版本的数据集并确定重复品权数。

7.10. 各种软件包的一般比较

108. 这里评述的软件包有许多共同特征。所有程序都需要具体规定每一样本要素的权数、层次和抽样单位。它们不能以无偏方式处理所有可想象的样本设计。譬如讲, 大多数分层多级样本设计中的初级抽样单位都是采用无替代的、与规模成正比的概率选择的。清单中只有一种程序, 即SUDAAN, 有用于明确处理此类设计的特征。但所有列入清单的程序都可以使用一种最终类集选样模型(见上面第7.7节)来处理这种设计。另外, SUDAAN还具有在初级抽样单位无替代选择的设计中估计方差的特征。STATA是具有处理设计中使用的分层和高级选样估计特征的唯一软件包。

109. 所有这些软件包都能估计有关总体样本的平均数、总数和比例、总体样本的子集, 以及各子集间的差异的抽样方差和相关统计数据(设计效果、同类相关等)。它们大多能估计回归和逻辑回归统计数据的抽样方差。它们全都能根据所产生的抽样方差估计测试统计数据。

110. CENVAR、EPI INFO、PC CARP和WESVAR均可免费或象征性付费获取。有兴趣的用户可通过本文所提供的e-mail和其他联系地址获取关于如何获得软件及相关文件资料的信息。

7.11. 结论意见

111. 本章对住户调查所使用的基于标准样本设计和较复杂设计的估计数抽样误差计算程序进行了简略评述。抽样误差的计算是分析和报告住户调查结果的一个至关重要的方面。按理讲, 一揽子调查制表中的所有特征都要计算抽样误差, 然而实际当中, 只为每个域指定一套关键性的关切特征来计算抽样误差。选定的特征应该是据认为对调查具有根本重要性的特征, 但是也应当包括有代表性地选择某些统计特征, 即据认为高度聚类的特征(比如表明族群或享受服务的机会的变量)和据认为有低聚类效应的特征(比如婚姻状况)。此外, 还应该用其他特征来指导选择, 比如所选特征占人口比例高低, 或重要的关切域。

112. 本章还主张使用专用计算机软件来估计调查数据的抽样误差。我们就使用标准统计软件包估计抽样误差是产生严重误差的原因提供了例证。一般来说, 使用标准统计软件包分析住户调查数据, 会低估调查估计数的实际可变性。这种对标准误差的偏低估计可能会导致对调查评估结果得出误导性的结论, 比如错误地宣称两个群体平均数之间存在重大差异, 或者错误地拒绝某种假设。

113. 本章还提供了现有的一些公开使用的统计软件包清单, 并附有基本通讯联系信息及其应用的简略评述。缺乏抽样误差估计方面的知识或专门技能, 是不利于发展中国家进行复杂数据分析的一大障碍。许多分析工作者不知道需要使用专用软件, 或者即便知道, 也宁愿不参加有关使用新软件包的培训。

114. 必须强调指出, 对于复杂调查数据方差估计这一广泛的新兴领域而言, 本章只能算是入门的引言。希望读者参阅本章末援引的一些参考资料, 以探求对此课题的更详细而系统的论述。关于这些及其他软件包的评述, 其中包括现有一些软件的计算机编码与输出, 请参阅Brogan(2004年)和其中援引的参考资料。

115. 最后, 必须承认, 随着技术的飞速进步, 许多软件包将会在相对较短的时间内要么变得过时, 要么经过改进超越了本评述所提供的说明。的确, 到本手册出版之时, 其中一些说明可能已被淘汰。因此务请记住, 在使用这些软件包的时候, 要从各自的用户指南或网站获取最准确的相关信息。

参考资料和补充读物

- An, A. 和D. Watts(2001年)。“抽样调查数据分析用的新SAS程序”, 《SUGI论文》, 第23期。北卡罗来纳州加里: SAS协会有限公司。可上网查阅: <http://support.sas.com/rnd/app/papers/survey.pdf>。
- Binder D. A. (1983年)。“关于出自复杂调查的渐进正常估计量方差”, 《国际统计评论》, 第51卷, 第279-292页。
- Brick J. M. 等(1996年)。《WesVarPC: 用户指南》。马里兰州罗克韦尔: 维斯塔特有限公司。
- Brogan, Donna(2005年)。“调查数据的抽样误差估计”, 《发展中国家和转型期国家住户样本调查》, 《方法研究》, 第96号, 出售品编号: E. 05. XVII. 6。
- Burt, V. L. 和S. B. Cohen(1984年)。“复杂调查数据方差估计备选策略比较”, 《美国统计协会调查研究方法论文集》。
- Carlson, B. L., A. E. Johnson和S. B. Cohen(1993年)。“关于使用个人计算机进行复杂调查数据方差估计的评估”, 《官方统计学刊》第9卷, 第4期, 第795-814页。
- Dippo, C. S., R. E. Fay和D. H. Morganstein(1984年)。“使用重复品权数法计算复杂样本方差”, 《美国统计协会调查研究方法论文集》。
- Fuller, Wayne等(1989年)。《PC CARP: 用户指南》。艾奥瓦, 埃姆斯: 艾奥瓦州立大学统计实验室。可上网查阅: <http://cssm.lastate.edu/software>。
- Hansen, M. H., W. N. Hurwitz和W. G. Madow(1953年)。《抽样调查方法和理论》, 第一卷, 《方法与应用》。纽约: 威利国际出版公司, 第10. 16节。
- Hansen M. H., W. G. Madow和B. J. Tepping(1983年)。“抽样调查中依靠模型的概率抽样推论评估”, 《美国统计学会会刊》, 第78卷, 第384期, 第776-793页。
- Kist, Leslie(1965年)。《调查抽样》。纽约: 约翰·威利国际出版公司。
- _____ (1995年), 《Leslie Kish论文选集》。Steven Heeringa 和 Graham Kalton编辑。新泽西州霍博肯: 约翰·威利国际出版公司。
- Kish, L., 和M. R. Frankel(1974年)。“根据复杂样本的推论”《皇家统计学会会刊: B辑》, 第36卷, 第1-37页。

- Landis, J. R. 等(1982年)。“一种分析复杂调查数据的统计方法: 首次全国健康和营养检测调查”, 《生命与健康统计》, 第2卷, 第92期。哥伦比亚特区华盛顿: 卫生、教育和福利部。
- Lehtonen, R. 和E. J. Pahkinen(1995年)。《复杂调查设计和分析实用方法》。纽约: 威利国际出版公司。
- Lepkowski, J. M., J. A. Bromberg和J. R. Landis(1981年)。“复杂样本调查多元分类数据分析程序”。《美国统计协会统计计算论文集》。
- Levy, Paul S. 和Stanley Lemeshow(1999年)。《人口抽样: 方法与应用》(第3版)。纽约: 约翰·威利国际出版公司。
- Lohr, Sharon(1999年)。《抽样: 设计与分析》。加利福尼亚州, Pacific Grove: 达克斯伯里出版公司。
- Potthoff, R. F., M. A. Woodbury和K. G. Manton(1992年)。“为使用超级人口模型项下的调查权重进行推论而改进‘等值样本规模’和‘等值自由度’”, 《美国统计协会会刊》, 第87卷, 第383-396页。
- Rust, K. (1985年)。“抽样调查中复杂估计量的方差估计”, 《官方统计学刊》1(4), 第381-397页。
- Rust, K. F. 和J. N. K. Rao(1996年)。“使用重复方法进行复杂调查方差估计”, 《医学研究统计方法》, 第5卷, 第283-310页。
- Shah, Babhai V. (1998年)。“方差估计的线性化方法”。《生物统计百科全书》, 第3卷, Peter Armitage和Theodore Colton编辑。纽约: 约翰·威利国际出版公司, 第2276-2279页。
- Shah, B. V., B. G. Barnwell和G. S. Bieler(1996年)。《SUDAAN用户指南: 发布7.0》。北卡罗莱纳州研究三角园: 研究三角研究所。
- Tepping, B. J. (1968年)。“复杂调查中的方差估计”, 《美国统计协会社会统计论文集》, 第11-18页。
- 联合国(1993年)。《住户调查中的抽样误差》。UNFPA/UN/INT-92-P80-15E。纽约: 经济和社会信息与政策分析部统计司和国家住户调查能力方案。
- Wolter, K. M. (1985年)。《方差估计导论》。纽约: 施普林格出版公司。
- Woodruff, R. S. (1971年)。“一个逼近复杂估计数方差的简单方法”, 《美国统计协会会刊》, 第66卷, 第334期, 第411-414页。

第八章

住户调查的非抽样误差

8.1. 引言

1. 不论抽样误差还是非抽样误差，都需要加以控制并减少到不至于因其存在而使最终调查结果丧失用途的地步。在前面有关样本设计和估算方法的章节中，把主要注意力放在了抽样误差方面，而对不答复和未覆盖之类其他来源的调查偏差则注意较少。这后一类误差统称为非抽样误差。如果非抽样误差不是随机发生的，它就特别有害，因为由此造成了调查估计数的偏差。

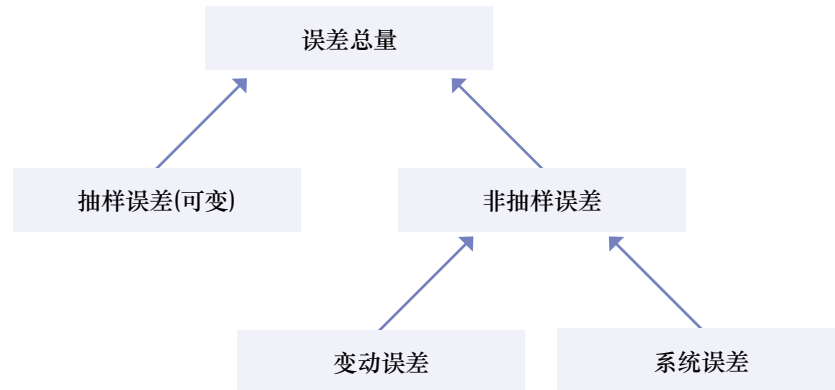
2. 所有调查数据都有可能产生各种不同来源的误差。在计量过程中产生的误差和抽样误差——亦即人口样本计量所产生的人口估算值误差——之间存在着根本性的巨大差别。

3. 在前面各章，假定一个人口中的每个单位 Y_i 都与一个被认为是该单位特征 y 的真值的数值 y_i 有关。同时假设，每当样本中有 Y_i 的时候，就此报告或观测到的 y 值就是 y_i 。有些情况下这一假设是成立的，但并非所有情况下都成立。举例来讲，在通过出生证明保有综合有效登记的国家，当 y_i 代表年龄的时候，很容易获取这种“真”值。可是在另一些情况下，比如当涉及对一个人的定性评估的时候，若想获取甚或定义真值，那就困难多了。譬如讲，一个病人就可能属于视情况而定者之列。

4. 调查实践中，关于单位 Y_i 所报告或观测到的数值永远是 y_i ，而不管是谁报告的，也不管是在什么情况下获得的，这种推测是不可靠的。实际调查经验提供的无数实例表明，每当进行调查的时候就可能会出现计量或观察误差以及因为错误回答、不答复和其他原因造成的误差。

5. 除了回答误差之外，调查中还可能覆盖面误差、数据处理误差，等等。一个人口参数样本估计量的质量是包括抽样误差和非抽样误差在内调查误差总量的函数。如前所述，产生抽样误差的唯一来源是概率抽样而不是完整的调查计数。而另一方面，非抽样误差则主要与数据收集和处理程序有关联。图8.1勾勒了作为调查误差总量组成部分的抽样误差与非抽样误差之间的关系。

图8.1
作为调查误差总量组成部分的抽样误差和非抽样误差之间的关系



6. 由此可见，非抽样误差主要是由于下列原因造成的：无效定义和概念；不准确的抽样标架；不令人满意的问卷；有缺陷的数据收集、制表和编码方法；以及不完整的抽样单位覆盖面。这些误差不可预测且不易控制。此类误差跟抽样误差不一样，它们可能随着样本容量增大而增加。若不加以适当控制，非抽样误差对大规模住户调查的破坏性可能比抽样误差还要大。

8.2. 偏差和变动误差

7. 如表8.1所示，调查误差可以分为变动误差(VE)和偏差两大类。变动误差基本上出自抽样误差；非抽样误差主要通过编码和键入等数据处理作业产生，但也促使产生变动误差。对比之下，偏差主要是由于下列因素导致非抽样误差造成的结果，诸如：无效的定义；错误的计量程序；回答不正确；目标人口覆盖不足，等等。有些种类的偏差亦可归因于抽样误差：这些误差是因为使用一个不能有效反映样本设计的方差估计量来计算抽样方差而引起的，由此导致抽样误差估计偏高或偏低。

8. 偏差一般是指系统误差，这种误差对于按照某种带有同样固定误差的具体抽样调查设计进行的调查有影响。前段暗示，通常一项调查的大多数变动误差都属于抽样误差，而偏差则主要是由非抽样来源产生的。因此可以说，偏差是因为在基本调查设计和程序中出了问题才产生的；而造成变动误差的原因是没有始终如一地执行调查设计和程序。

表8.1
调查误差分类

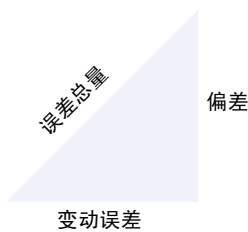
变动误差	抽样误差
	非抽样误差
偏差	抽样误差
	非抽样误差

9. 调查误差总量的统计项是均方误差(MSE)，相当于方差加上平方偏差(见图8.2)。因此，为便于论证起见，若偏差为零，则均方误差只不过是估计数的方差。但是在住户调查中偏差永远不会是零。不过如前所述，计量调查偏差总量几乎是不可能的，其部分原因是因为它的计算需要知道真实的人口值，而在大多数情况下这也是未知的。由于偏差来源如此之难以计数，其性质又是如此之复杂，因此很少有人试图估计偏差总量。

10. 下面图8.2中的三角形描绘了误差总量及其构成分量。三角形的高代表偏差，而基底代表变动误差。三角形的斜边是误差总量，它体现了这样一个概念：均方根误差(即误差总量)等于抽样方差积数加上平方偏差的平方根。因此，

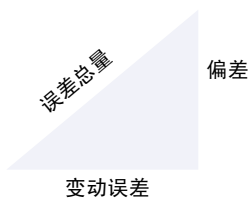
$$\text{均方根误差} = \sqrt{VE^2 + Bias^2} \quad (8.1)$$

图8.2
调查误差总量及其各分量



11. 不论减少变动误差还是减少偏差，误差总量都会相应减少。图8.3展示了一种变动误差和偏差都大幅度减少的情况。结果，通过拿此图中的三角斜边长度和图8.2中的三角斜边长度进行比较可以看出，误差总量也大大减少了。

图8.3
减少的调查误差总量



12. 争取实现周密的抽样调查设计并辅以妥善的实施策略，目的就在于减少变动误差和偏差，以取得相对准确的抽样结果。

13. 总的来说，高精度要靠周密设计的大型样本来实现，而只有最大限度地减少变动误差和偏差，才能达到准确性的要求。这就意味着，精确的设计倘若带有很大偏差，它也许是非常不准确的。就此而言有必要承认，经常包含在住户调查报告中的标准误差估计数低估了调查误差总量，因为这些估计数没有把偏差的影响考虑在内。

14. 实际当中, 可将非抽样误差进一步分解成变动误差和系统误差(见Biemer和Lyberg(2003年))。系统误差一般是非补偿误差, 因此倾向于一致(大体在同一方向上); 而变动误差是补偿误差, 故倾向于不一致(互相抵消)。

8.2.1. 变动分量

15. 一种误差的变动分量是由对调查的各种样本和重复有影响的机会因素(随机因素)引起的。在计量过程方面我们可以设想, 从访调员选择到数据收集、再到数据处理的一整套程序都可以使用同样的特定程序、按照同样的既定条件单独地重复, 而任何一次重复都不会影响另一次重复。重复的结果不但受系统因素的影响而且受随机因素的影响; 这些因素是由进行重复的条件引起的, 并且以同样的方式影响重复的结果。

16. 在变动误差(VE)只是由抽样误差引起的情况下, 变动误差平方等于抽样误差。调查平均值偏离实际人口值的差幅是偏差。无论变动误差还是偏差均可产生于抽样或非抽样业务。变动误差将计量估计量背离其预期值的偏差, 而它既包含抽样方差又包含非抽样方差。估计量的期望值与其真值的差异是总偏差, 它既包含抽样偏差又包含非抽样偏差。

17. 可以根据适当设计的在相同条件下进行的重复(复制)调查工作之间的比较来评估变动误差。变动误差的减少有赖于某些方面的增加, 比如扩大样本容量或启用更多的访调员。另一方面, 只有通过改进调查程序——比如在调查作业的各个阶段实行质量控制措施——才能减少偏差。

8.2.2. 系统误差(偏差)

18. 举例来讲, 如果调查中有一种普遍低报或者高报的倾向性, 就会出现系统误差。譬如在一些没有出生登记制度的社会, 一般男子倾向于高报本人年龄。这种习俗显然会导致发生系统偏差, 即过高估计男性人口的平均年龄。

8.2.3. 抽样偏差

19. 之所以发生抽样偏差, 要么是因为特定概率抽样在执行上不妥当或者有瑕疵, 要么是因为估算方法有缺陷。前者包括标架方面存在的缺陷、选样程序有误以及抽样单位的部分计数或不完整的计数。本手册第三章和第四章详尽讨论了可能因为不适当地实施抽样设计(即便是近乎完美的设计)而发生抽样偏差的诸多情况。

8.2.4. 偏差与变动误差的进一步比较

20. 一般来说偏差难以计量, 正因为如此我们才强调在这方面要严加控制。只能通过与外部的可靠数据来源进行比较才能对偏差进行评估。另一方面, 可以通过样本细分或者通过相同条件下的重复调查来评估变动误差。通过改进调查程序可以减少偏差。

21. 根据Verma(1991年),有些来源的误差——其中包括覆盖面、不答复和样本选择问题——主要以偏差形式出现。另一方面,在编码和录入方面的误差则可能主要以变动误差的形式出现。

22. 虽然系统误差和变动误差都可以从总体上降低精确性和可靠性,但是偏差对于诸如总体均值、比例和总数之类的估计数危害更大。这些线性估计数是样本观测值的总和。前面强调过,跟抽样误差一样,也可以通过增加样本容量来减少非抽样变动误差。对于像相关系数、标准误差和回归估计数之类的非线性估计数来说,无论变动误差还是系统误差都可以导致严重偏差(Biemer和Lyberg,2003年)。在许多住户调查中,主要目的是提供诸如平均数、人口总数和比例等描述性的人口量度;因此在这些情况下重点就放在减少系统误差方面。

23. 总之,偏差产生于基本调查设计和程序方面存在的缺陷。偏差要比变动误差难以计量,而且只有根据与来自正常调查以外的较为可靠的数据进行比较或者与通过改进程序获取的信息进行比较才能对偏差进行评估。

8.3. 非抽样误差来源

24. 从一开始进行调查规划和设计到最后阶段进行数据处理和分析,始终存在各种各样无计其数的非抽样误差起因。

25. 可以把住户调查方案看作明确规定了各项业务工作的一套严格的规则。例如,这些规则描述了行将涵盖的目标人口,规定了拟在问卷中使用的主旨概念和定义,并且提出了收集数据的方法和所要进行的计量项目。假如调查作业依照提出的规则进行的话,理论上获得所研究特征的实值是可能的。但是非抽样误差把它变成了一个可望而不可及的空想。

26. 一般而言,下述一种或多种因素可能会引发非抽样误差:

- (a) 与调查目标有关的数据要求不适当并且/或者不一致;
- (b) 由于抽样单位的区域界限定义不确切而造成单位重叠或遗漏;
- (c) 抽样单位的识别特征不完整或不正确,¹或者采访的方法有缺陷;
- (d) 由于采用了含糊的问卷、定义或指示,致使采访、观察或计量的方法有欠妥当;
- (e) 缺少受过训练且有经验的现场访调员,包括缺少高素质的现场监督;
- (f) 对基本数据缺乏仔细核对以纠正明显错误;
- (g) 在数据处理操作中——比如编码、键入、检查、制表等方面——出现误差;
- (h) 在提供和出版列表成果的过程中出现误差。

但以上清单绝非详尽无遗。

¹ 请注意,这种误差即便是在选样作业中发生的,它也属于一种非抽样偏差。

8.4. 非抽样误差的各种分量

27. Biemer和Lyberg(2003年)鉴定了非抽样误差的五种分量,即:规格误差;标架误差;不答复误差;计量误差;数据处理误差。我们还可以补充另外一种误差——即估算误差,供考虑。下面简略讨论这些误差类别。

8.4.1. 规格误差

28. 规格误差是在提问所隐含的概念与应该计量的构成物不相符的情况下发生的。譬如讲,对于某人有多少子女这样一个简单提问,在一些文化中可能会有各种不同的解释。在有延伸家庭的住户中,可能不把同一户口内的受访者自己所生子女与其兄弟姊妹的子女区分开来。在残疾调查中,如果笼统地问受访者是否有残疾,那么依损伤程度或受访者对残疾概念的理解不同也会有各种不同的解释。有轻度残疾的人也许认为自己没有残疾。除非在问卷中包含正确的筛查问题,否则答案本身可能不能揭示残疾人的总人数。

8.4.2. 覆盖面或标架误差

29. 在大多数区域调查中,初级抽样单位是由比如普查计数区(EAs)之类的地域单位类集构成的(见第四章关于标架的全面讨论)。在普查测绘中计数区的划界不准确是常有的事。这样,住户就可能在第二阶段标架中被遗漏或发生重叠。这种缺陷可能导致调查估计数产生两个方向的偏差。如果本应在标架之中的单位实际没有列入,那么被遗漏单位的选择概率即为零,结果是估计数偏低。另一方面,如果某些单位发生重叠,就会产生过度覆盖,从而导致估计数偏高。

30. 因此,与标架有关的误差要么会导致过度覆盖,要么会导致覆盖不足。后者在非洲国家大规模调查中是最常见的结果。

31. 在多级住户调查中,抽样涉及到多个阶段,其中包括在一个阶段或分几个阶段选择区域单位;住户的列表和选择;以及入选住户中的个人的列表与选择(见第三章)。在其中的任何阶段都有可能发生覆盖面的误差。

32. 有必要再次强调,覆盖方面的误差无论就其大小还是就其影响而言都难以估计,因为它需要有不但对样本来说而且在定义上对所用抽样标架来说均属于外部的信息。

33. 正如前面的讨论所暗示的那样,未覆盖是指在某个定义的调查人口中有些单位没有被纳入抽样标架(关于包括未覆盖在内的覆盖面的误差,详见第六章的讨论)。由于这种单位的选择概率为零,所以它们实际上被排除于调查结果之外。

34. 应当指出的是,对于故意而且明确地将某个规模较大人口中的某些部分排除在调查人口之外的设计来说,这里所讲的覆盖面误差不适用。这种故意的排除是由调查目标和实际困难决定的。譬如关于婚姻问题的态度调查可能会把最低法定年龄以下的人排除在外。由于调查中的实际困难,往往也把机构居民排除在外。为了维护现场工作人员的安全,一个国家可能将埋设了地雷的一些地区排除在住户调查以外。在计算未覆盖率的时候,不应把故意

且明确排除在外的群体成员计入调查目标人口或未覆盖者之列。就此而言，调查目标人口的界定应该是明确声明的必要调查条件的组成部分(见第三章关于调查目标人口的讨论)。

35. 总覆盖误差这个词系指未覆盖率和过度覆盖误差率的绝对值的总和。未覆盖净值是指未覆盖量多于过度覆盖量的超额数，因此是它们的代数和。如果不存在过度覆盖的话，净覆盖值就只计量总覆盖值。大多数发展中国家在住户调查中遇到的主要问题是覆盖不足方面的误差。调查研究工作者普遍认为，在大多数社会调查中，覆盖不足的问题要远比过度覆盖的问题更普遍。对未覆盖的误差纠正和加权调整难度，远远超过了处理不答复问题的难度，因为覆盖率不能从样本自身获得，而只能从外部来源获得。

36. 正如第四章用了很大篇幅所讨论的那样，未覆盖误差可能是由于使用有缺陷的抽样单位标架而引起的。如果标架未经更新，而为了节省时间和经费使用了陈旧标架的话，就可能招致严重偏差。譬如讲，在一次住户调查中，如果老的居住单位清单自其初次编制以来(距眼前调查最长可达10年之久)从未经过更新，那么在选定的计数区内新增加的居住单位就不会成为第二阶段居住单位标架的组成部分。同样，已经废弃的居住单位仍将在标架中留着空白。在这种情况下，可能既有属于相关人口的单位遗漏，又夹杂着不属于相关人口的单位。

37. 有时候，样本中还有些单位没有找到或访问过。这个问题也是因为使用了不完整的清单引起的。此外，天气或交通条件差有时也会使得访调人员不能在指定的调查期限内通达某些抽样单位。

38. 第三章讨论过，住户抽样调查的根本目的是获取客观结果，以便根据样本中的单位观察对预期的目标人口做出有效推论。由此可知，如果地区之间以及诸如男性/女性、年龄类别、族群和社会经济分类等子群间的未覆盖率各不相同，就会扭曲调查结果。

39. 未覆盖误差不同于不答复误差。如前所述，后者是因为拒绝受访、未找到住址或受访者不在家、问卷丢失等缘故而导致未能获得观察数据。不答复率是通过选定的样本与完成访调的样本之间的比较，根据抽样结果来计量的。对比之下，正如前面所强调的，只能通过调查作业以外的某种检查才能估计出未覆盖率。

8.4.2.1. 抽样过程的实施误差

40. 抽样过程的实施误差是指在抽样范围内的遗漏或扭曲，比如选择率或操作程序的错误应用。另一个例子是不适当地用比较容易通达的住户或比较愿意合作的住户来替换已被选定的住户。

8.4.2.2. 减少覆盖面的误差

41. 减少覆盖面误差的最有效方法就是通过剔除错误和重复单位来改进标架；而改进标架的最佳途径，就是确保适当更新标架(见第四章的详细讨论)。另外，确保区域抽样单位和其中的住户容易找到也很重要。若想实现此项目标，最好的办法就是在构建原始标架——通常是最近的人口与住房普查——的时候，要有良好的测绘工作。

8.4.3. 不答复

42. 正如本手册反复强调的那样，不答复是指未能从一些样本单位获得答复。建议考虑将抽样人口划分为两个层次：一层是提供了答复的所有样本单位；另一层是不能从中获得答复的所有样本单位。

43. 大多数情况下，不答复现象在样本单位中间呈不均匀分布，但是过度集中在子群当中。由于不答复的情况有差别，致使子群中的调查完成率偏离选定样本中的完成率。如果调查变量也与子群有关的话，这种偏离就有可能招致不答复偏差。

44. 如果保存了样本范围内所有合格要素计数的话，不答复率是可以准确估计的。调查答复率的定义是：完成问卷的抽样单位数量与有资格被选抽样单位²数量之比(另见第六章)。建议在所有公开发表调查数据的出版物中都要报告不答复的情况，这对官方调查而言则应是强制的。不答复可能归因于入选样本的个人不在家，拒绝参加调查，或因丧失能力而不能回答问题。不答复也可能是因为调查表/问卷丢失，或者因为某些地区天气恶劣或地形险峻或缺乏安全保障等缘故而不能前往调查。所有不答复类别都是指有被选资格的受访者，因此应将脚注2所包含的无被选资格者除外。举例来讲，在生育率调查中，选定计数区内的目标人口就只包括适当年龄组的妇女，而将特定年龄组以外的女性和所有男性排除在外。

45. 第六章着重指出有两种不答复：一种是单位不答复，另一种是项目不答复。单位不答复是指没有从特定样本单位得到任何信息；而项目不答复是指从特定单位收集到一些但不是所有的信息。项目不答复表现在相应的抽样单位在数据记录上有差距，这些差距可以归因于受访者拒绝回答、访调员的疏忽或缺乏能力。预期的受访者拒绝参加调查可能受到多方面影响，其中包括缺乏动力，没有空闲时间，对调研的某些问题比较敏感，等等。Groves和Couper(1995年)提出了受访者抱持拒绝态度的一系列原因，包括调研的社会背景、受访者的特性、调查设计(包括受访者的负担)、访调员的特性以及访调员与受访者之间的互动。具体到项目不答复，受访者可能会觉得调查中的提问令人难堪、比较敏感或者/并且与申明的目的不相干。访调员也许会跳过一个提问或者漏记了一个回答。另外，有可能在编辑过程中排斥了某项答复。

46. 单位(或整体)不答复的数量多寡——除其他因素外——可以表明调查的一般接纳程度、复杂程度、组织和管理水平，因而也反映出问卷中探索的某个特定项目的复杂性、清晰度和可接受的程度以及访调员处理这些项目的工作质量。

47. 不答复的情况会给调查结果带来偏差；如果不答复单位对于答复单位不具“代表性”(这种情况比较常见)，那么此类偏差也可能会很严重。不答复不但通过减小样本容量而增加抽样误差，而且也会增加非抽样误差。

² 可能会发现一些被抽样的单位因为超出了调查范围而变得无被选资格，比如空置房屋、被没收的或者已经废弃的住所。

48. 旨在提高响应率的努力经常会促使改进包括选择调查工作在内的程序。例如在1978年赞比亚生育率调查中，为了提高响应率招募了女教师做访调员，以方便提问有关避孕等方面的问题。当时想到，假如使用年轻的男性访调员来做这项工作，拒绝答复率就会比较高，因为人们很忌讳年轻小伙子向年长的妇女提问，特别是包括避孕在内与性有关的问题。

49. 实际当中不可能彻底消除不答复的现象，但是可以通过采取说服教育的方法，反复访问“家中无人”的住户，以及其他方法，来最大限度地降低不答复率。关于处理调查数据方面不答复问题的相关信息，详见第六章和第九章。

8.4.4. 计量误差

50. 当观察或计量值背离抽样单位实值的时候，就会发生这种误差。这些误差围绕调查的实质性内容发生，比如调查目标的界定，它们转化为可用的提问，以及受访者答复的获取、记录、编码和加工处理。因此这些误差涉及到个体单位层面的计量精确度。

51. 譬如讲，在调查初期创立有关标架构建和问卷设计的错误或误导性的定义和概念，就会导致覆盖不完全和不同访调员所作的各种各样的解释，从而造成收集信息的种种误差。

52. 对现场工作人员缺乏指导是另一个误差来源。在一些调查中，含糊不清的指令需要访调员在现场执行任务时运用自己的判断力。访调员本身就可能是一个误差来源。有时候，从所有单位收集的关于某个特定项目的信息全都是错误的；这主要是因为现场工作者缺乏培训。

53. 在非洲，由于往上虚报年龄和数字喜好，致使自报年龄成为一个普遍性的计量问题。这些及其他计量误差的实例可以归因于受访者或访调员或两方面都有责任。有时候，可能这两者之间有互动，这就更增加了这方面的误差。同样，测量装置或技术里的缺陷也可能造成观测误差。

54. 受访者可能通过下述方式带来误差：

- 他们没有弄明白调查的问题；
- 由于受访者对调查目的缺乏了解而给出漫不经心的错误答复；尤其是受访者也许不愿花费足够的时间去考虑提问；
- 即便不知道正确的答案，他们也抱着“合作”的愿望回答提问；
- 在涉及敏感问题(比如收入问题和名声不好的疾病)的调查中，他们倾向于故意提供错误答案；
- 在参考期较长的情况下，他们的记忆力减退，在支出调查中收集关于非耐用商品信息的时候就是如此。

55. 起因于不同来源的各种误差可能产生相当大的累积效应，因为不同来源的误差不一定会被剔除。这种误差的净效应可能是个很大的偏差。

8.4.5. 数据处理误差

56. 数据处理误差主要包括:

- 编辑误差;
- 编码误差;
- 数据录入误差;
- 编程误差。

57. 上述误差是在数据处理阶段发生的。譬如讲, 在对有关经济特征的无确定答案进行编码的时候, 编码员可能会偏离编码手册中规定的程序, 而给职业指定错误的编码。

8.4.6. 估算误差

58. 估算误差主要是因为是在计算调查权数的时候没有运用正确的公式造成的。但即便使用了正确的公式, 如果权数计算有误也会产生误差。如果所使用的方差估计量与实际抽样设计不符, 从而在与调查点估计数有关的置信区间方面产生误差的话, 就会造成抽样估计数方差(抽样误差)。每当出现这种情况的时候, 结果就会产生偏差。

8.5. 非抽样误差的评估

59. 本章用了很大篇幅表明, 非抽样误差的来源无计其数而且多种多样。因此基本上不可能总体评估一次调查中所发生的非抽样误差。不过, 像下面的讨论那样, 研究和评估某些非抽样误差分量还是可能的。

8.5.1. 检查一致性

60. 在设计调查工具(问卷)的时候要特别留意包含一些辅助信息项目, 用以检查收集数据的质量。如果这些补充信息项目容易获得, 那就可以对调查涵盖的所有单位进行详查; 否则就只对抽样单位的一个子样本详查这些项目。

61. 举例来讲, 在对法定常住人口进行普查后的计数调查(PES)的同时, 收集事实上的人口信息也是有益的, 这样就有可能计算暂住人口和暂时不在本地的人口。通过这两个数字的比较可以看出数据的质量。同样, 包含某些相对稳定的比率项目(比如性别比率), 也许对评估调查数据的质量有益。

62. 在调查的评估阶段也应使用一致性的检查方法。可以引进交叉检查, 以确保——譬如讲——登记为户主的人不低于规定的年龄, 或者有生育史的女性年龄不低于13岁。

8.5.2. 样本检查/验证

63. 评估和控制调查中一些种类的非抽样误差的一个方法, 就是单独重复不同作业阶

段的工作，以便检查和验证误差。由于实际原因，可由一个经过良好训练和富有经验的工作人员小组只对一个工作样本进行这种重复检查。如果该样本设计恰当并且检查工作有效进行的话，不但能够查出误差，而且可以大概估计出误差的大小。如有可能彻底检查调查工作，即可大幅提高最终调查结果的质量。

64. 在样本检查方面，只能对检查过的样本进行验证。通过在不同的调查阶段将产出品——亦即填好的表格、编了码的表格、计算单，等等——划分成若干批量，然后检查每个批量的样本，这样或多或少可以减少上述局限性带来的影响。如果某个特定批量的误差率高于规定的水平，那就可以对该批量进行检查并纠正误差，从而提高最终结果的质量。

8.5.3. 调查后的检查或再访谈检查

65. 可用以评估答复误差的一项重要样本检查，就是选择一个子样本，或者在普查的情况下选择一个样本，并聘用经过较好培训和较有经验的工作人员(他们在这两方面都优于一般从事主要调查的人员)对其进行重新计数。为了使这种办法产生效力，有必要确保：

- 在主要调查结束后立即进行重新计数，以减少记忆误差。
- 采取措施尽量减少主要调查可能对调查后的检查工作产生的调节效应。

66. 通常，这种检查性的调查旨在便于评估覆盖面误差和内容误差。为此建议，最好先对高层次阶段样本中的所有单位——亦即计数区和村庄——进行再计数，以探查覆盖面误差，然后只对最终单位其中的一个样本进行再调查，同时要确保，这个样本对于在非抽样误差方面有特殊意义的人口各个部分都有代表性。

67. 这种检查性的调查有个特殊的优点，就是便于进行单位检查，其中包括：首先把两次计数所获的数据与检查样本所涵盖的单位挂钩，然后分析观察到的个别差异。在发现存在差异之后要努力找出其根源所在，并且深入分析这种非抽样误差的性质和类别。

68. 如果由于受时间和财力制约不能进行单位检查的话，可以采用一种叫做“总量检查”的效果较差的备用方法。这种方法就是用检查性调查数据所提供的参数估计数同主要调查的相应估计数作比较。总量检查只能给出净误差，即正负误差的结式。对比之下，单位检查所提供的既有净误差的信息又有总误差的信息。

69. 在调查后的检查中，应使用原始调查所使用的相同概念和定义。

8.5.4. 质量控制方法

70. 调查工作有充裕的空间运用各种统计质量控制方法，因为这种工作规模大，并且其业务有可重复性。在大规模调查中可以使用控制图和验收取样法来评估数据质量和提高最终结果的精确度。举例来讲，在初期阶段可对每位数据录入员的工作进行100%的检查，但是如果误差率降低到规定水平以下的话，从此以后只对其工作精确度进行抽样检查就行了。

8.5.5. 关于记忆误差的研究

71. 本章在前面讲到, 答复误差是由各种因素引起的, 诸如:

- 受访者对调查的态度;
- 访谈方法;
- 访调员的技能;
- 记忆误差。

72. 其中, 记忆误差尤其值得注意, 因为它带来了超乎受访者控制能力的特殊问题。它与报告期的长短有关, 也与报告期距调查日期的间隔时间长短有关。后面这个问题可以通过在调查日期以前为报告期选择一个适当的间隔期或者一个尽可能接近该间隔期的阶段来解决。

73. 研究记忆误差的一种方法, 是在一项调查所涵盖的单位样本或子样本中收集并分析与不止一个报告期有关的数据。此种方法的主要问题是产生了一定数量的调节效应, 这可能是因为针对一个报告期报告的数据对另一个报告期的报告数据有影响。为了防止出现调节效应, 可以从不同的样本单位收集研究中的不同时期的数据。请注意, 这种比较需要有大型样本。

74. 另一个方法就是收集一些补充信息, 以便获取不同报告期的估计数。譬如讲, 在人口调查中不但可以收集关于受访者年龄的信息, 而且可以收集关于其出生年月日的信息。这样, 数据中的差异就会揭示自报年龄中存在的任何记忆误差。

8.5.6. 交叉再抽样

75. 交叉再抽样法就是采用相同的方法从整体样本中抽取两个或更多子样本, 每个子样本都能提供一个人口参数的有效估计数。这种方法有助于信息质量评估, 因为交叉子样本可用于获取有关非抽样误差的信息, 比如不同访调员偏差所产生的差异, 诱导信息的各种方法, 等等。

76. 在子样本经过不同访调员小组的调查、又经过制表阶段不同工作团队的加工处理之后, 通过对基于该子样本的各种估计数进行比较, 即可对调查结果的质量进行一次大范围的检查。举例来讲, 假如我们对基于四个子样本的估计数进行比较, 这些子样本是分别由不同调查工作组调查处理的, 其中三个估计数彼此较为接近, 但是第四个估计数却相去甚远; 而该差异如此之大, 以致不能合理地归因于抽样误差, 那么, 在这种情况下就需要对这个差别较大的子样本工作质量提出质疑。

8.6. 结论意见

77. 在住户抽样调查中须对非抽样误差给予适当的注意, 因为这种误差若不加以控制, 就会造成调查结果的巨大偏差。大多数调查都很少注意这方面的误差, 而不惜承担或许

产生不可靠调查结果的风险。控制非抽样误差的最佳途径就是在所有调查活动中——从制订计划到选择样本直至分析结果——都要遵循正确的程序。尤其要关注并加强对现场工作人员的培训；同时，各种调查提问，尤其是未经以往调查工作验证的提问，都要经过充分的预先检验。

参考文献和补充读物

- Biemer, P. 和 L. Lyberg (2003年)。《调查质量导论》。威利调查方法论丛书。新泽西州霍博肯：威利国际出版公司。
- Biemer, P. 等编 (1991年)。《调查计量误差》。威利概率与数学统计丛书。纽约：威利国际出版公司。
- Cochran, W. (1963年)。《抽样方法》。纽约：威利国际出版公司。
- Groves, R. 和 M. Couper (1995年)。“住户调查中进行调查后不答复调整的理论动员”。《官方统计学刊》第11卷第1期，第93-106页。
- Groves, R. 等编 (2000年)。《调查中的不答复》。威利跨学科出版物。纽约：约翰·威利国际出版公司。
- Hansen M., W. Hurwitz 和 M. Bershad (2003年)。“普查和调查中的计量误差”，《调查统计方面的里程碑论文》。IASS周年纪念专辑。
- Kalton, G. 和 S. Heeringa (2003年)。《Leslie Kish: 论文选集》。威利调查方法论丛书。新泽西州霍博肯：威利国际出版公司。
- Kish, L. (1965年)。《调查抽样》。纽约：威利国际出版公司。
- Murthy, M. (1967年)。《抽样理论与方法》。印度加尔各答：统计出版学会。
- Onsembe, Jason (2003年)。《在2000年轮次人口和住房普查中改进数据质量》。埃塞俄比亚，亚的斯亚贝巴：联合国人口活动基金国家技术服务队。
- Raj, D. (1972年)。《抽样调查设计》。纽约：麦格劳-希尔图书出版公司。
- P. Chandhok (1998年)。《抽样调查理论》。伦敦：纳罗萨出版社。
- Shyam, U. (2004年)。《2003年土库曼斯坦生活水平调查》。技术报告，国家统计和信息全国协会，土库曼斯坦，Ashgabad。
- Som, R. (1996年)。《实用抽样方法》。纽约：马塞尔·德克尔出版公司。
- Sukhatme, P. 等 (1984年)。《抽样调查理论与实践》。艾奥瓦州埃姆斯和新德里：艾奥瓦州立大学出版社和印度农业统计学会。
- 联合国 (1982年)。《国家住户调查能力方案：住户调查中的非抽样误差：来源、评估和控制》。DP/UN/INT-81-041/2。纽约：联合国统计处。

Verma, V. (1991年)。《抽样方法: 培训手册》。东京亚洲和太平洋统计协会。

Whitfold, D. 和 J. Banda (2001年)。《计数后的调查: 值得吗? 》。联合国2000年人口和住房普查全球审查圆桌研讨会: 十年中期评估和未来展望。纽约。8月7-10日。

第九章

住户调查的数据处理

9.1. 引言

1. 本章讨论全国性住户调查的数据处理。本章将首先描述典型住户调查的流程，随后将分析数据处理的准备工作，准备工作是调查策划过程中不可缺少的一部分。

2. 在过去的20多年中，信息技术发展迅速。其发展又深刻地影响了统计调查处理系统的设计和施工技术。

3. 硬件方面的主要发展是主机系统转换成了个人计算机平台。不管在处理速度还是储存容量方面，个人计算机的功能都越来越强。个人计算机可以执行统计工作(从小规模调查到大规模统计作业，如人口普查和包含大量样本的住户调查)的各种过程。

4. 在硬件发展的同时，用于统计数据处理、分析和发布的软件质量也有了长足的发展。软件的发展就使一些处理任务有可能从电脑专家手中移交给领域专家。

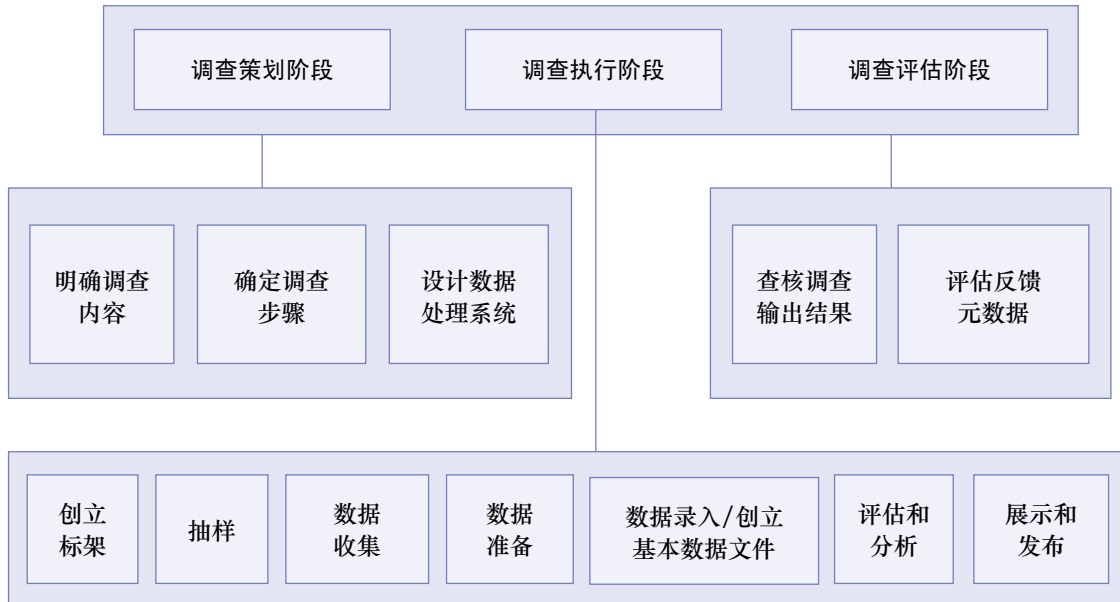
5. 过去几年出现了大量用于统计调查的软件包。每一种软件产品的优势都是相对于数据处理不同步骤的要求而言的。可参照本章附录部分来选择调查数据处理过程中不同步骤所用的软件；附录部分还对本章提及的所有软件产品分别进行了总结描述。

9.2. 住户调查流程

6. 图9.1描绘了住户调查的一般流程。大体上所有住户调查都经过相同流程，典型的流程如下：

- 调查策划阶段：调查的设计者决定调查结果的主要目的和用途、调查将主要得到何种输出结果以及需要什么输入、得到输入的过程(调查问卷的设计和准备，以及相关的调查工具)、把输入转换为输出结果的过程以及数据处理和记录系统的设计；
- 调查工作阶段：该阶段包括设计抽样标架、样本设计和选择、数据收集(计量)、数据准备(数据录入、编码、编辑以及插补)、创立观察(基本原始数据)文件、评估(包括计算权数)、创立派生变量、分析以及结果的展示和发布；
- 调查评估阶段：该阶段包括检查和评估是否得到了指定的终端产品、是否已经通过合适的方式发表并宣传调查成果、是否已经记录并保存元数据等。

图9.1
住户调查流程



资料来源: Sundgren (1991年)。

7. 在为某次调查设计和实施数据处理系统之前, 应该首先设想此次调查的整体系统, 这点非常重要。合适的调查执行顺序和过程顺序对于任何一次住户调查的成功执行都至关重要。在理想的排序中, 调查目的应决定调查输出结果的设计(如制表计划和数据库)。同时, 该设计将决定一系列后续活动, 包括调查设计、数据收集、数据准备和处理, 以及最终对于结果的分析 and 发布。

8. 数据处理可被视作将数据收集阶段所获得的调查回复转变为一种适合制表和分析的形式过程。数据处理需要自动化操作和人工操作。数据处理也可能是耗费时间、消耗资源的, 并且对于最终输出结果的质量和成本具有影响。

9.3. 调查策划和数据处理系统

9.3.1. 调查目的和内容

9. 正如第二章所述, 任何调查设计的第一步都应该是明确并记录其主要目标。住户调查提供人口中的住户信息, 旨在回答利益相关方关于目标人口的问题。任何一次调查的目标都反映在获取这些问题答案的尝试中, 所以, 调查问卷应该提供相关数据。

10. 通常情况下, 利益相关方需要通过使用住户调查数据所了解的问题可以分为以下几种。(见Glewwe, 2005年)。

11. 一类问题旨在确定调查人口的基本特征(人口中贫困人口的比例、失业率等)。

12. 另一类问题旨在评估介入活动对于住户特征的影响，或评估住户特征的整体发展情况(比如，参加某项目的住户比例、他们的特征与未参加该项目的住户有何不同、随着时间推移住户的生活条件是改善还是恶化了等)。

13. 最后一类问题讨论住户环境和特征的决定因素，或者两者之间存在的关系(即，讨论现有情况以及导致该情况发生原因的问题)。

9.3.2. 调查程序和工具

9.3.2.1. 制表计划和预期输出结果

14. 创立制表计划和虚拟表格可以有效地帮助调查设计者更准确地得到使用者所需的信息。虚拟表格是表格的草样，包括除了实际数据外的所有内容。制表大纲至少要确定表格标题和栏目设置，并明确制表需要哪些实质变量、用于分类的背景变量，以及适用于不同表格的人口群体(调查对象或要素或单位)(见第二章)。最好也尽可能详细地展示分类的类别，尽管当日后更好地了解答复分类的抽样分布时，这些分类可能要做出调整。

15. 可以从几个方面来看制表计划的重要性。制定虚拟表格能够表明待收集的数据是否会产生可用的表格。虚拟表格不仅能够指出缺少什么内容，也能反映出多余的内容。另外，在制作虚拟表格方面额外花费的时间，通常会在数据制表阶段通过缩短设计和制作实际表格的时间而得到加倍的补偿。

16. 调查的制表计划与抽样设计之间也存有紧密的联系。例如，只有设计的样本允许在特定地域划分的情况下，才可以在表格中如此地划分地域。

17. 联合国(1982年)更详尽地描述了制表计划所需要的内容以及它带来的各种好处。

18. 上述内容旨在强调：在特定调查的有效计划过程以及相应的数据处理系统中，制表计划可能发挥重要作用。但是，同时需要强调的是，制表计划只代表该调查所期望得到结果的基本部分。住户调查有可能产生大量信息。经过清洗的微观数据集可被视作主要和基本的输出结果。通常情况下，需要给这样的微观数据集进行打包，并通过合适的发布渠道，以用户可接受的形式提供给利益攸关方。

9.3.2.2. 表格设计和打印

19. 一旦确定了调查的目标和制表计划，就可以制作相关的调查问卷。调查问卷在调查过程中发挥关键作用，在此过程中，信息从持有者(被调查者)传递到需求者(使用者)。通过调查问卷这一工具，使用者的信息需求就以可操作的术语表述出来了。同时，问卷也是调查的数据处理系统输入的主要基础。

20. Lundell(2003年)认为，调查期间所使用问卷的布局对于数据收集具有影响，反之亦然。如果发现扫描技术便于数据收集，那么就需要使用特殊形式的问卷设计；设计还取决于使用键入磁盘输入技术还是人工数据录入技术，录入方式不同，设计也不同。

21. 不管使用哪种数据录入技术, 每一份调查问卷都必须是可唯一识别的。每一份表格上都应印有唯一的表格识别码。因为误读表格识别码可能会导致重复录入和其他问题, 所以应该采取措施使风险最小化。在使用扫描技术时, 采用条形码显然是最佳选择。如果考虑使用人工数据录入, 表格识别标志应该包括校验数位等信息, 以防止录入失误。识别码应对应可唯一识别的调查问卷, 并且应该是以数字形式表示的。通常情况下, 样本权数或膨胀因子分配信息(层级、初级取样单位、地域区段、制表所需的行政区域区分等)也附在表格识别码中。

22. 表格通常装订在工作簿中(例如, 每个调查地区或城区的工作簿等)。与表格一样, 每一份工作簿都必须是唯一可识别的。而且每一份工作簿与其包含的表格之间的关系必须明确, 以保证X表格一直属于Y工作簿, 而且只属于Y工作簿。在数据处理的整个过程中, 都要使用工作簿的识别信息。这个过程始于表明一个新的工作簿从现场来到储藏室, 持续到在需要时查找表格。例如, 在数据制表或分析数据过程中, 检查某项内容。因此, 就像必须减少误读某张表格识别信息的风险一样, 也必须减少误读工作簿识别信息的风险。

23. 必须为每一个填写区域设计充足空间, 使其可以容纳变量可能出现的最大字符数; 例如, 必须明确住户成员的最大规模, 以正确分配填写区域空间。

24. 保证调查问卷的观察单位的定义、跳答模式和其他方面没有缺陷是非常重要的。每一次住户调查都收集主要统计单位(基本对象)——住户的信息, 同时还有一系列住户范畴内的附属单位(关联单位)——人、预算项目、农场与庄稼等。调查问卷应该清晰明确地表述这些单位的定义, 并保证每一个观察单位都标有适当的唯一识别码。识别住户的典型方法是使用简单的序列号, 这也是人工录入系统的重要特点。将序列号写在或贴在调查问卷的封面上, 或者事先在打印店打印序列号。通常情况下, 序列号也表示表格的识别码。

25. 图像扫描和处理是一种快速获取数据的方式, 并越来越多地被采用。鉴于此, 对将用于图像扫描的问卷设计的特点进行一些讨论也是很重要的。当计划使用图像扫描仪如光学字符识别(OCR)、智能辨字系统(ICR)、光学阅读(OMR)等软件处理问卷, 就会给问卷设计带来一些问题。下面我们将讨论这些问题。

26. 调查问卷上使用两个条形码。第一个条形码表示调查问卷的页码。这个条形码非常重要, 尤其在问卷每一页的版式和布局都非常相似的情况下。这是成像软件区分问卷不同页码的主要方式。问卷每一页上都另有一个条形码以及相关的解释, 并且每一页上的这个条形码都是一样的, 但是该条形码与之后问卷的条形码顺序不同。第二个条形码把同一份问卷的不同页联系在一起, 因为扫描的准备过程经常把问卷的各页分开, 所以第二个条形码也非常重要。

27. 纸质表格填写区域的确切格式表示了将从问卷中得到信息的数据字典格式。举例来说, 如果目的是获取普查区的五位区号, 纸质问卷上则设有一处狭窄的填写区, 填写区印有五个隔开的方框, 用来填写该普查区的区号。如果信息收集界面是为人工数据录入而设计的, 那么问卷的打印就不必要非常精确, 只需要一个能够接收5位数的窗口管理器就足够。

但是，修改数据录入系统的设计是非常重要的，这样域就用数字表示，而且可以被自动填写，以免获得的数字产生偏差并防止误读数据。

28. 对于使用扫描方式进行数据录入和使用打孔方式进行数据录入，表格设计也会有很大区别。因为扫描仪完全依赖数据填写区域的位置来进行识别。与人工录入相反，通过扫描方式进行的数据收集只需要问卷每一页上的调整区域，而不需要纸质问卷上印有识别标志。为了使扫描系统能够高精度地解读图像，填写区域应该大小适中。在培训期间以及现场工作管理过程中，应鼓励调查者以数据填写区域为中心，清晰明确地进行填写。

29. 如果获取信息所用的方式不同，调查问卷的设计也会存在根本区别。以职业类别编码为例，如使用人工录入数据的人工编码过程来获取编码，就和使用包括屏幕职业码本的计算机辅助查找表的屏幕编码不同。如果将人工录入问卷数据，编码的格式就打印在问卷上，在负责问卷编码的人员看完问题（“您的职业是？”）的开放性回答之后，供其使用。如果问卷将被扫描，则不一定需要在纸质问卷中分配此区域，因为计算机内设计好的打印模板含有码表，而且验证者融合了编码者和数据录入文员的角色。在对职业类别编号进行输入时，验证者在看到扫描图像上开放性回答的同时，可以在索引职业类别编码的下拉列单中进行快速选择。

30. 当考虑使用扫描技术时，问卷的打印质量便是非常重要的问题。与肉眼相比，扫描仪对于打印缺陷更加敏感。譬如，如果使用了某几种颜色或某些颜色的合成色、色调和清晰度出现变化、打印歪斜或错位、或是页数自动编码出现错误并有装订错误，那么就可能出现这个问题。

31. 关于统计调查和普查处理中扫描技术的使用，Catherine(2003年)和Lundell(2003年)介绍了几个重要的问题。

9.3.3. 住户调查的数据处理系统设计

9.3.3.1. 住户调查数据处理的一般方法

32. 在住户调查策划阶段，系统设计是其中一项主要活动。基本上，此阶段将根据一些形式统一的方案，确定需要收集哪些调查数据并明确整个数据处理系统。

33. Jambwa、Parirenyatwa和Rosen(1989年)从以下几点讨论了在一个统计机构，为所有系统的设计、开发和记录都采用一套形式统一的方案的好处，尤其是在住户调查的过程中：

- (a) 该方案可能成为统计学家、领域专家以及系统分析/程序员进行合作的共同平台；
- (b) 后面的步骤将明确地描述、记录并提及所有的调查工作。所产生的记录(即元数据集)对于相应的统计生成系统的开发和维护都非常重要；
- (c) 鉴于需要很多不同的系统，统计机构用于系统开发和维护的成本通常很高。住户调查的结构通常具有相同的模式和准则。譬如，它们常常共有相同的文件和数据

结构以及编码系统等。因此, 后来的调查可以受益于之前调查所开发的数据处理系统, 而且有望降低开发和维护的成本;

- (d) 采用形式一致的方式对于调查的整合也很重要(譬如, 如果希望对于不同调查或者不同调查轮次的数据进行综合分析的话)。

9.3.3.2. 系统设计中形式一致方案的一般特征

- 34. 本部分对于上述形式一致的方案的一些总体和基本方面提供说明。

数据结构

35. 决定分析哪个社会问题、采用什么数据、应用哪种统计技术——这些都是好的分析的基本成分。然而, 确认并定义调查的分析对象或单位是更加基本的问题。此前在表格设计和打印部分已经强调过这个问题。在调查的数据处理系统设计过程中, 应该对分析单位(对象)和变量做出更加形式一致和具体的描述。

36. Sundgren(1986年)认为分析对象或单位可以被定义为用户希望得到其信息的任何具体或抽象的实体(物质实体、生物、组织、事件等)。对象的定义与社会事件(调查目的)紧密相连, 收集并分析数据正是为了该社会事件服务的。同样的, 对于住户调查来说, 利益相关方想得到信息的对象、项目、因素或单位是住户、人、土地等。在大多数情况下, 基本对象是住户, 而且随特定调查而定通常有很多与基本对象相关的关联对象。

37. 表9.1列出了1987年津巴布韦两次普查之间的人口调查中定义的对象/单位。住户是基本对象, 其关联对象是“个人”、“12岁或12岁以上的女性”以及“死者”(Lagerlof, 1988年)。

38. 每一个对象都有很多相关的变量。变量是对象的属性(特质或特点)。比如, “个人”这个对象就有年龄、收入、职业、婚姻状况等变量。变量可能是定性的, 也可能是量化的。

39. 每一个对象都应该有独一无二的身份。关联对象的身份表明了它所相关的基本对象。比如, “个人”与“住户”相联系, 而且住户身份和个人身份一起能够确认个人, 也就是说, 确认个人在住户名册内的序列号。

数据处理系统的输入

- 40. 输入包括调查者根据调查问卷获得和记录的数值。

数据处理系统的输出

41. 系统的输出结果主要包括统计表格(基于制表计划)、包括微观数据和宏观数据的数据库等。对象类型、变量类型以及统计方法的类型不同, 输出结果也会不同。被制表的变量通常是“原始”的, 但也有可能是从原始变量推导得来的。

表9.1
调查分析对象/单位示例(取自1987年津巴布韦两次普查之间的人口调查)

对象/ 单位	身份变量	对象/单位定义	重要变量	关联对象	
				对象	外键
住户	HID=住户身份(区域、地区, subdiv=分区, Eanr=普查计数区(EA)数量, Hhnr=户主数量)	一户人是指在通常情况下, 一起生活并一起吃饭的一群人, 而且不包括客人	SOH=住户规模 层 计数区	个人 死者	HID HID
个人	HID、PID PID=个人身份	个人是指住户的常规成员或是前一晚的客人	性别、年龄、MAR-STAT=婚姻状况, ETHNIC=民族, USMEM=住户的常规成员, RELTH=与户主关系	住户 女性≥ 12岁	HID HID、 PID
死者	HID DID=死者身份	死者在过去12个月中是住户的常规成员。	SEXD=死者性别 AGED=死者年龄	住户	HID
女性≥ 12岁	HID、DID	所有12岁以及12岁以上的女性, 并且是住户的常规成员或是前一晚的客人	出生婴儿数量	个人	HID、 DID

文件组织

42. 通常情况下, 在输入阶段以及制表前阶段, 有几种不同的文件结构。譬如, 住户调查的数据录入可能会选用变量长度(与平面文件相对), 因为住户的规模与构成各异, 所以在数据录入阶段需要变量长度记录。这种方法能够有效利用空间, 却不便于后来的处理工作。但是, 最后平面文件中经常会引用将要组织的数据, 以便于制表, 并使不同类型的通用软件得到最佳利用。

43. 系统流程图。住户调查应建立一张适度详细的流程图。该流程图非常重要, 原因诸多。首先, 流程图是创建时间表以及估计完成调查处理过程所需资源的工具。一般情况下, 任何调查的数据处理都包括以下主要活动:

- (a) 数据检查、编辑和编码;
- (b) 数据录入、验证和确认;
- (c) 将输入阶段采用的数据结构转换为制表适用的数据结构;
- (d) 制表。

44. 系统流程图还应包括基本的文件操作, 如选择、映射、文件整理和匹配、推导新变量、汇集、制表和图示。

记录系统

45. 全面清晰的记录(即元数据集)对于住户调查数据处理系统的开发和维护都非常重要。因此,记录文件和不同的操作非常重要,这样,没有参与最初系统执行的人也可以使用该系统。应使用标准化模板,并与其数据一起电子存档,以保证进行充分记录。

46. 在调查机构所做的所有不同调查中,如果代码的意义一样,数据处理系统的变量应尽可能使用相同的名称、代码以及数据格式。对于用于识别文件内记录(对象)的变量来说,使用相同格式尤其重要,因为当结合(集合)不同系统的数据时可能要用到这些变量。

47. 在设计数据录入界面或表格扫描时,软件中设有模板工具,辅助文件记录;需充分利用这些工具达到有效的记录。比如,普查与调查数据处理系统(CSPro)或者集成微机处理系统(IMPS)数据字典格式定义了每个变量在数据文件中的位置——起始点和结束点,不管变量是数字还是文字,也不管变量是否重复出现且重复出现多次。数据字典还标明了变量内含有的数值(St. Catherine, 2003年)(关于CSPro和IMPS的更多信息,见本章附录)。

9.4. 调查工作和数据处理

9.4.1. 标架创立和样本设计

48. 正如第三章和第四章所述,很多住户调查的第一阶段抽样单位是可用的最近一次全国性普查所定义的普查计数区(EAs)。首先创建包含国内所有普查计数区(EA)清单的电脑文件,这是创立第一阶段样本框架便捷而高效的方式,使用Excel这样的电子表格是最佳的创建方式,每一普查计数区(EA)对应表格一行,并设立多列包含所有需要的信息(参照本章附录,了解更多关于Excel的信息)。

49. 标架一定要易于进入,并便于进行各种操作,如排序、筛选以及概要统计,有助于样本设计和评估。Excel易于使用,而且很多人都知道如何使用;Excel有排序、筛选和汇总等功能,从标架中准备样本时需要用到这些功能。工作表能够轻易被导入到其他大多数软件包中。通常来说,为每一个样本层创建不同的工作表更加方便。

50. 标架单位记录应有如下内容:

- 应包括第一标识号(应为数字)。记录中应含有代码,独一无二地标识标架单位所在的行政区域和子区域。如果标架单位按地理顺序标号,就更有好处。通常,普查计数区(EA)代码已经有上述属性;
- 第二标识号是标架单位所在村庄(或其他子行政区域)的名字。第二标识号也很重要,用于在地图上和现场标识标架单位;
- 样本单位的一系列特征,如规模量度(人口、住户)、农村或城市、人口密度等都应该包括在文件中。这些特征可能被用作分层或者分配选择概率,并在评估中用作辅助变量;
- 应包括操作型数据,如关于单位变化的信息以及指明样本使用方法的信息。

51. 在每一个阶段，所有样本单位的选择步骤以及选择概率都必须完全记录。当使用标准样本时，应记录特定调查的样本使用了哪些标准样本单位。取样单位应该使用标准标识数字系统。

52. 基于1991年人口和住房调查的纳米比亚标准样本可作为以上所述内容的范例。(中央统计局，纳米比亚，1996年)

范 例

为了能够选择纳米比亚地理区域的随机样本，需为地理区域创建一个样本标架。为此目的，创建了地理区域标架——初级抽样单位。区域平均包括约100个住户，大多数包括80-100户。所选区域是基于1991年人口和住房普查的普查计数区。各个小规模普查计数区与临近的普查计数区结合在一起，形成足够规模的初级抽样单位。经验法则认为，一个初级抽样单位至少应包括80个住户。共计有1 685个这样的初级抽样单位，并根据地区、农村、小型城市和城市将其分类至初级抽样单位的各个层级。

53. 分层基于为1991年普查做准备时对于普查计数区的分类。共分为32层，层级内部按照地理顺序列出初级抽样单位。在城市和小型城市，也按照各地区的收入水平列出初级抽样单位。城市和小型城市层级的清单中，首先列出的是高收入地区，随后是中/低收入地区。

54. 中央统计局准备了初级抽样单位标准样本标架的Excel文件，标架包括：

- 区域；
- 独一无二的初级抽样单位编号；
- 收入水平(仅适用于城市)；
- 区；
- 普查计数区编号；
- 1991年普查统计的住户总数；
- 各层住户累积总数；
- 普查统计的两性人口；
- 标准样本状况(无论初级抽样单位是否在标准样本内)；
- 标准样本初级抽样单位编号(仅限标准样本中的初级抽样单位)；
- 权数(上升或膨胀因子)，仅限标准样本中的初级抽样单位。

55. 每一个区域都有一个Excel文件，每一个Excel文件都按照农村、小城市、城市——高收入；小城市、城市——中/高收入为初级抽样单位进行分组。

56. Pettersson(2003年)讨论了关于标准样本的具体问题。Munoz(2003年)进一步探讨了如上面描述的计算机化的标架在执行住户调查取样步骤时可能发挥作用，引导住户调查

完成其主要步骤: 组织第一阶段架构, 通常基于最新的人口和住房普查结果(EAs); 选择规模与概率成比例的初级取样单位(规模由住户总数、居住单位或人口来衡量); 并且基于所选住户以及计算选择概率和相应样本权重清单的基础上, 更新电子数据表。下文中9.4.3.5是关于点估计过程和权数计算的, 更详细地讨论了选择概率和相应权数的计算(另见第五章)。这些计算所需数据将从以上讨论的电子数据表中获得, 并且可以使用Munoz展示的电子数据表计算权重。

9.4.2. 数据收集和数据管理

57. 住户调查可能产生大量已完成的问卷。如果要避免混乱, 那么就需要在早期步骤中周详考虑并设定人工处理并解释这些大量文件的步骤。必须仔细计划人工处理(归档和检索)问卷的流程, 在现场传回数据之前, 该流程必须可以投入使用。该系统的一个重要方面就是需要估计拟得到的数据规模, 这样就可以准备需要使用的文件夹、盒子等, 而且可以分配架子或橱柜的空间。系统的第二部分是一份信息记录, 记载有关问卷的信息的到达情况, 并且记录数据在系统内的流动情况。这些是数据管理的重要方面, 也是任何一次调查成功管理并执行数据处理策略的重要前提。

58. 在此阶段, 所完成问卷的实物安全也很重要。正因如此, 在此阶段使用图像扫描就很有吸引力。如果问卷在到达办公室时就进行扫描, 则可以减少问卷数据意外丢失的危险。图像扫描技术能够在扫描完成后, 将问卷的扫描版保存在现场和现场外, 从而进一步增加了安全性(Edwin, 2003年)。但是, 需要注意的是, 采用这种技术是否成功, 很大程度上取决于有关机构如何组织并管理该技术的用法和相关过程。所以, 虽然有一些国家成功使用扫描, 同时另一些国家却未能成功使用。为了成功进行扫描操作, 在诸多因素之中, 必须考虑统计局的组织方式: 它的程序是集中化或分散化的; 受调查者的档案以及保证调查数据收集工具的质量。

9.4.3. 数据准备

59. 需要将收集的数据输入到数据文件中。将问卷上的数据转换为电脑可阅读的数据被称作数据录入。在此方面, 通常需要给不同的变量值分类, 这些值是开放性答案; 这个分类过程被称作编码。通过编辑获得的数据可以确认错误的数据。可以采取适当的措施检查有可能存在的错误, 譬如, 可以再次与信息来源进行联系。检查之后进行更新(更正)。处理步骤包括: 数据录入、编码、编辑、检查和更新/更正。汇总起来, 这些过程被称作调查处理的数据准备阶段。

9.4.3.1. 数据准备策略

60. Munoz(2003年)详细讨论了数据准备的各个方面和构型。住户调查最普遍的组织机制包括在现场收集数据之后集中进行数据准备。另一种安排则将数据录入融合到现场工作中。比较近的创新是计算机辅助访问技术。

集中化的数据准备

61. 在个人计算机出现之前，该方法是存在的唯一选择。发展中国家仍主要使用该方法，但是因为微型计算机的出现，该方法有一些改良。在此方式下，数据录入被视作调查之后在一个或多个地方进行的生产过程。数据录入可以在国家统计局总部进行或者在其地区办公室进行。

现场数据准备

62. 近来，将基于计算机的质量控制融入到现场操作中，这种方式已经被看作提高住户调查质量和及时性的关键之一。在这种战略下，数据录入和一致性控制被视作现场操作中不可缺少的部分。

63. 现场数据准备的形式之一是让数据录入员在固定地点用台式电脑工作(譬如，在国家统计局的地区办公室中)，并且组织现场工作，这样，团队的其他成员至少要访问调查地点(通常是初级抽样单位)两次，从而使数据录入员在几次访问期间，有时间录入数据并检查数据的一致性。在第二次和之后的访问中，调查者再次向相关住户询问数据录入程序发现存有错误、遗漏或不一致的问题。

64. 另一种方式是让数据录入员用笔记本电脑工作，并与其他成员一起访问调查地区。整个团队都留在该地，直至所有数据都录入完成，而且数据录入程序认证这些数据是完整正确的。

65. 融合数据收集过程和数据准备过程具有一些相对优势，包括：数据质量更高，因为调查者还在现场时就能够更正错误；在现场操作结束之后，可能立即生成数据库，并进行制表和分析；以及调查者有更大的空间来进行标准化数据收集过程。

66. 要采用上述两种方式，操作地点必须有持续不断的电源供应。在供电紧缺的国家，如大多数发展中国家，尤其是其农村地区，上述两种方式可能都不可行。应该注意的是，数据收集和准备所用的移动设备会遇到组织和物流挑战。要成功使用这个战略，前提条件有：拥有有效的管理系统；设备和数据的有效安全保护；能够得到足够的数据备用储存；而且有充足的消费品供应，如现场使用的备用电池。

计算机辅助访问

67. 计算机辅助个人访问(CAPI)是个人访问的一种方式。采用这种方式的调查者带着一台笔记本电脑或者便携式电脑，直接将数据录入数据库中，而不是在纸上填写问卷。这种方法节省了处理数据的时间，而且免去了调查者携带数百份问卷的负担。但是，虽然计算机辅助个人访问方式在很多年前就已适用，但发展中国家很少在复杂的调查中严格使用这种战略。设立此类数据收集方法可能价格不菲，而且需要调查者具有电脑和打字技能。基于计算机的访问还需要结构完好，有始有终。但是，发展中国家大多数调查都需要对每一户进行多次访问，并对于住户的每一个成员进行单独访问等。这样的过程没有严密的结构，而本质上是由调查者决定的。

9.4.3.2. 调查数据编码和编辑

68. 数据检查、编辑和编码可能是数据处理过程中最困难的方面。在数据管理和数据准备的组织阶段, 入行不久的调查专业人员经常会遇到很多困难。如果可能, 数据检查、编辑和编码过程能够通过自动化方式得到最佳执行。但是, 在不能自动分配代码的情况下, 自然就要考虑编码, 这样就需要人工分配代码。

编 码

69. 编码的目的是将数据处理为适合电脑录入的形式。编码操作主要是指给以文字方式表述的答案(例如, 地理位置、职业和行业等)分配数字代码。编码还包括转录, 将调查过程中已经分配和记录的数字代码转换到编码表中。

70. 应编写一本明确指导编码人员的手册。手册中应包括不相交分类, 涵盖调查问卷中问题所有可能的回答。在大规模住户调查中, 需要尽量扩大封闭式问题和预先编码问题的比例。

数据编辑和检查

71. 检查和/或编辑问卷的目的是使数据之间、各表格内部以及各表格之间达成一致性, 并且查找、核实、更正或排除异常值, 因为极端值是导致调查估计样本变异性的主要原因。

72. 编辑包括修订或更正问卷中的录入。编辑可以被视作一个确认过程, 发现并更正数据中不一致和不可能情况; 或者被视作一个统计过程, 基于数据的统计分析进行检查。不管是数据录入或是数据的特殊编辑运行, 现在都趋向于让计算机越来越多地进行编辑。这样的编辑运行可能是, 也可能不是交互式的, 也就是说, 操作人员可能会, 也可能不会马上更正错误。但是, 在校正更复杂的错误时, 需要更多的时间和更深入的分析才能找到正确的更正, 因此, 更适合使用非交互式的编辑运行方式。Olsson(1990年)的参考材料更加详细地讨论了调查数据检查和编辑的各个方面。

73. 检查和人工编辑。检查或人工编辑的主要任务是在下一步数据处理阶段开始之前, 发现问卷中遗漏、不一致以及其他明显错误。应该尽快, 而且要尽可能地在数据来源临近处, 譬如, 省级、区级或更低一级的办公室进行数据编辑。理想情况下, 应该在表格被送往处理中心之前, 在现场发现并更正数据中的大多数错误。因此, 培训和指导手册通常指导调查者和监督者在数据被送走之前, 现场检查问卷并更正错误。这项任务重要而困难, 完成好坏由现场工作质量、监督有效性以及调查管理等决定。

74. 计算机辅助的编辑。计算机编辑的执行方式有: (a) 在数据录入阶段交互执行, (b) 在数据录入之后使用成批处理方式, 或者(c) 结合(a)和(b)方法。交互式编辑一般在处理简单错误时较为有效(譬如, 键入错误): 如果某些错误需要咨询监督者意见, 那么交互式编辑便会延误数据收集过程。这种错误(包括不答复)应该留给独立的电脑编辑操作处理。

75. 计算机辅助编辑的程序一般由数据库程序设计, 如集成微机处理系统(IMPS)、集成调查分析系统(ISSA)、普查与调查数据处理系统(CSPro)、VB和微软Access(本章附录详细介绍了上述软件)。最简单的程序扫描数据、进行记录并基于写进程序的编辑规则发现不一致现象。更加复杂的编辑程序能够对于各文件间的变量(如身份变量)进行比较, 并可能发现不一致之处。系统的输出结果包括错误列表。一般会对照原始数据, 人工核对该列表。错误将在原始数据文件副本中得到更正。

检查方法的种类

76. 对调查问卷中的数据需要采用不同种类的检查方法进行检查, 典型的检查方法包括: 范围检查、对照参考数据检查、跳过检查、一致性检查及排版检查(Munoz, 2003年)。

77. 范围检查。范围检查旨在确保调查中的每项变量仅包含有限的有效值区间内的数据。绝对变量的值只能取自问卷中已预先规定的值(以性别为例, 只能采用男性为“1”、女性为“2”的编码方式)。时间变量应包含有效日期, 数字变量应该在规定的最小值与最大值之间(例如年龄在0-95岁之间)。来自两个或两个以上密切相关领域的的数据可以对照外部参照表进行检查, 这种情况是范围检查的一个特例。

78. 跳过检查。跳过检查旨在检验跳过类型和规则是否得到适当遵守。例如, 如果一名儿童在回答前面是否入学一题时答案为否, 那么对其问卷仅以学校+儿童为指向的问题就不做记录, 为确保这一点而进行的检查就是一次简单的检查。确保每位被调查者的答案被填写在问卷的正确部分, 这样的检查更加复杂。根据各自的年龄及性别, 住户各成员应回答(或跳过)问卷中的特定部分。例如, 15-49岁的女性可包括在生育能力部分, 男性则不可。

79. 一致性检查。一致性检查旨在确保从一个问题中得到的数值与另一个问题得到的数值保持一致。如果两个值来自于同一个统计单位, 例如某人的出生日期与年龄, 那么这种检查就是简单检查。更复杂的一致性检查则需要比较来自于两个或两个以上不同观察单位的信息。例如, 父母应至少比他们的子女年长15岁。

80. 排版检查。典型的排版错误包括在输入多个数字的时候将数字位置颠倒(比如将“41”输入为“14”)。如果在年龄上出现此类错误, 或许能够在对婚姻状况或家庭关系进行一致性检查时发现。例如, 一位41岁的已婚或丧偶的成年人, 若其年龄被错误输入为14, 那么在对其婚姻状况进行年龄检查时就会显示错误标记。然而, 同样的错误如果出现在每月肉类支出中可能就很容易逃过检查, 因为14美元或者41美元都是有效数量。应对此类错误的典型方法就是让两名操作人员将同一份问卷中的数据输入两次。

处理缺失数据

81. 当调查进入数据处理阶段时, 在大多数情况下仍然会有大量数据缺失。一些住户可能已经搬迁或者拒绝回答问题。问卷中的一些问题可能没有得到回答; 或者, 一些数据可能是捏造的或与问卷中的其他信息不一致。不管因为何种原因, 结果都是记录缺失、记录为空或记录部分为空。

82. 区分缺失数据与无效数据很重要,前者指的是本应存在但正确数值未知的数据。例如,因为住户拒绝参与,一份调查问卷可能空白;而住户未在地里种庄稼则可能造成另一份问卷部分空白。在第二种情况下,“种植面积”变量便无效。这些记录必须存档,以供分析和制表之用。

83. 如何处理确实缺失的数据取决于缺失数据的种类。由于某住户拒绝参与调查或回答问题的受调查住户不能回答调查问卷中的所有问题,这会导致选择样本要素全部缺失。在这种情况下,我们说发生了“单位不答复”。

84. 如果被调查者只能回答部分问题,那么这就属于“项目/部分不答复”情况,因为从被调查住户中只获取了部分数据,而不是全部数据。

85. 两种类型的缺失数据都会造成调查估计数出现偏差,这一点在本手册中已经多次强调。处理不答复的适当方法、包括调整方法详见第六章。

86. 在部分不答复的情况下,为确保总数的一致性,可能有必要用合理估计数值来代替缺失数值。这就是第六章中所谈到的插补。插补替代数值有若干种方法。下面列出其中几种:

- 平均值插补法:用平均值(在初级抽样单位或整个数据集)来插补缺失值;
- “热卡”插补法:从与不完整记录相似的donor记录中借用缺失值。donor记录应该已经通过所有编辑检查;
- 统计插补法:利用与从完整数据中得到的与其他变量有关的数值(回归、比率)来插补缺失值。

87. 上述只是几种可以应用的插补方法;还有几种其他的插补方法。插补的有效性当然取决于插补模型在多大程度上成功补充了不答复。在选择可用的辅助信息方面,很重要的一点是变量与待插补的变量相关(关于这方面的更多信息参照Olsson(1990年))。

9.4.3.3. 数据录入

88. 数据录入的目的是将纸质问卷上的信息转换为半成品(机器可阅读的文件)。必须通过编辑程序和文书过程进一步改善半成品,以获取最终产品,即所谓的清洁数据库。在数据录入的最初阶段,重点是速度以及保证文件上的信息与问卷收集的信息完全吻合。

89. 应该在早期阶段决定使用何种方式将问卷上的数据录入到电脑媒介中,因为所用方式将大幅影响基本工作流程、数据存储技术、表格设计以及员工构成。

键入磁盘式数据录入

90. 键入磁盘式数据录入是指将经过编码的数据键入磁盘、磁碟或光盘中。很多发展中国家的调查组织在这种数据录入模式的使用方面获取了大量经验。键入磁盘式数据录入法是现在使用的主要方式。而且由于个人计算机和相关软件的出现,键入磁盘式数据录入法得到进一步加强。

91. 数据录入应用程序。正常情况下，数据录入应用程序包括三个模块。第一个模块是信息输入模块。第二个模块用于验证录入数据。该模块证明录入信息质量良好，并且追踪数据录入员的表现。第三个模块用于更正输入信息，因为可能需要更改数据录入或核实过程中未发现的错误值。

92. 数据录入应用程序通常有一个主菜单。负责数据录入的人员可以在菜单上选择数据录入、验证和更改。在使用主菜单之前，使用者必须输入用户名和密码，证明他/她拥有进入该程序的许可。如果登录失败(即输入了错误的用户名或密码)，应用程序将立即关闭。所有用户名和密码都储存在后端的用户表中，密码在此得到加密处理。当用户使用有效密码登录系统，后端的表格就会得到更新。

93. 数据录入模块。数据录入模块是问卷和数据文件或数据库之间的纽带。输入系统必须易于数据录入员使用。以下是一些重要的条件，规定：

- 数据录入界面应该尽可能接近问卷相应页面的设计。录入员应能够非常迅速地从问卷上找到屏幕上相应的区域；
- 数据录入的速度非常重要。录入员不愿意等待系统评估每一个录入值。因此评估过程必须非常迅速；这意味着系统只能与服务器进行必要的联系，这也表示直到住户的所有数值都录入完成之后，数值才能保存到数据库。这样做的缺点是：如果由于某种原因，应用程序关闭，那么当前录入的住户信息将丢失。但是，相对来讲，高速度更加重要；
- 问卷中每一个数值都应该有一个数字代码，这样录入员就能够使用数字键区。这是达到高速的基础；
- 数据录入模块必须含有变量效度控制。当录入员键入无效值时，他/她能马上收到错误报告。效度控制也应该控制关联数值；譬如，如果“性别”的数值为“1”（男性），那么就必须使生育力信息失效；
- 当出现任何逻辑上不可能发生或不符合常理的情况(如一个女孩的年龄大于她的母亲)或者非常不可能的情况时(如一个女孩的年龄比她的母亲小15岁以内)，数据录入程序当然应该将这些情况标记为错误；
- 记录按键次数和数据录入时间很重要。后来的统计工作可利用该记录，如用来预计数据录入的全部时间。

94. 数据验证模块。验证系统的目的是提供录入数据的质量信息，以及每一个数据录入员的错误率。验证模块的界面设计与数据录入模块界面完全一致，不存在任何明显的差别。但是，两个模块主要的区别不仅在于总结的键入次数，还在于发现的错误次数。验证类型的可选项包括总体验证，验证所有的EA和一个EA内的问卷；或是样本验证，只验证一部分EA以及一部分问卷。

95. 数据更正模块。数据更正模块主要用来更正数据录入模块由于某种原因未能完成的信息。不管是一整户还是一个单独数值，更正模块都可以增加、删除或更新其信息。

96. 监督者管理应用程序。管理程序是监督者改变数据库的工具。管理程序主要用于更正批处理主文件(BMF)，还用于接收记录用户表现的报告。以下几点非常重要：

- 监督者能完全控制程序中的批处理主文件。他们应该可以添加、删除和更新批处理主文件信息；
- 可以添加或删除用户，并可以得到所有用户的完整列单。应该可以检查所有用户或某一个用户的当前状态；
- 可以看到并打印键入次数的统计数据。应该可以选择不同的时间段；
- 可以看到并打印某一个用户的错误率以及所有用户的平均错误率；
- 可以重设一个计数区的数据录入或数据验证；
- 可以在这个程序中得到监督者管理工作时所需的所有信息。

Svensson(1996年)详细描述了键入磁盘式数据录入系统的各个方面。

97. 键入磁盘式数据录入系统平台。市场上有很多数据录入和编辑程序的开发平台。例如，普查与调查数据处理系统以及它的前身——集成微机处理系统已经证明它们能够为很多发展中国家复杂的全国性住户调查开发有效的数据录入和编辑程序提供支持。事实证明，它们是易于获得和使用的平台(Munoz, 2003年)。

扫 描

98. 在各普查和调查的数据处理中，扫描技术的使用发展迅速。仅在几年前，主要的数据录入方式还是键盘操作系统。很多与之竞争的系统不能在市场上得到。如今，情况发生了变化，最畅销的数据输入系统都基于扫描技术。扫描技术有很多分类，每一种都有利有弊。最常用的技术包括：光学字符识别(OCR)，标志能够识别机器打印的文字；智能辨字系统(ICR)，标志能够识别手写文字；光学阅读(OMR)，标志能够识别预先确定位置(通常是标志框)的钢笔或铅笔标志；以及条形码识别(BCR)，标志能够识别印刷条形码表示的数据。

99. Lundell(2003年)认为，统计调查和普查采用扫描技术时，选择主要介于智能辨字系统和光学阅读之间。人口众多的国家通常倾向于使用光学阅读，而复杂的问卷则倾向于使用智能辨字系统。光学阅读限制了表格的设计，但是能够带来快速处理，而且对于员工技能的要求相对较低。智能辨字系统对于表格设计没有要求，但是处理过程对于电脑效率以及员工技能要求更高。条形码通常只用来打印和获取身份信息，如表格的编号，因为条形码包括校验数位信息，能够尽可能减少错误。

100. 在扫描过程中，问卷的扫描速度为每分钟40-90页(双面扫描)。选择使用扫描方式，而不是选择传统的数据录入形式，如键入数据，最重要的决定因素是速度。扫描软件使用智能辨字系统和光学阅读，确认问卷的页数并评估问卷的内容。存在疑点或者尚未编码的项目被传到验证者处。验证者检查书写不清楚的项目，并通过扫描模板嵌入的电子查找表格给开放式问卷编码。执行验证检查的方式有很大的灵活性，取决于扫描模板是如何设立的。可以对关键变量进行全部或部分检查，尽可能提高数据文件所获取回答的准确性。

101. 事实表明, 使用图像扫描过程能够将数据收集的效率提高70%(Edwin, 2003年)。由扫描导致的一些问题能够通过对过程进行适当的技术组织得以解决。例如, 在问卷上事先印刷条形码, 并将条形码用作连接问卷各页码之间的工具, 这种方式可以解决页码缺失和错配的问题。如果对于设备和软件有适当的维护和管理, 扫描操作的长期成本(包括设备和软件的购买)可能比数据键入操作的成本低得多。

102. 住户调查使用扫描技术的经验通常十分有限, 特别是在撒哈拉以南地区。但是, 该技术在2000年人口和住房普查中得到大幅使用, 而且可能标志了扫描技术得到普遍采用的转折点。例如, 肯尼亚、坦桑尼亚联合共和国、南非、纳米比亚和赞比亚在它们最近的普查中都使用了扫描技术。近来, 世界银行发起的所有“核心福利指标问卷”调查都使用了扫描技术。纳米比亚和南非等国也在本国的住户调查项目中使用了扫描技术。

9.4.3.4. 数据集的文件结构和组织

数据储存

103. 住户调查通常具有住户层面和个人层面的信息, 要有效使用存储空间, 就需要顺序文件或是变长文件, 因为不同的住户由不同人数的个人构成。平面文件不必要地占用了可用空间的很大部分, 只有当调查的所有问题都以住户为统计单位时, 平面文件才适用。但是, 如上所述, 所面临的情况并非如此。部分问题的统计单位是每一个住户内的次级单位, 数目不一, 如个人、庄稼、消费品等。如将住户每一个成员的年龄和性别存储为住户层面的不同变量, 则会浪费空间, 因为所需变量的数量将由最大的住户规模决定, 而非由住户的平均规模决定。

104. 住户调查的数据输入一般使用变长文件。因为住户的规模和构成不一, 所以数据输入过程中需要变长记录。虽然每种类型记录的长度和形式固定, 每一个文件内都有多种记录种类。每一个文件完成后实质上都是问卷的电子版本。问卷中的每一行或每一部分都形成一份记录。每一份记录开头都是一串识别符, 将记录与住户、观察单位等联系起来。这种方式有效利用空间, 但是不便于之后的处理。因为在之后处理过程中, 交叉引用不同文件中的数据变得非常关键。

105. 举例来说, 普查与调查数据处理系统所使用的文件结构能够很好的应对在处理很多不同统计单位时出现的复杂情况, 同时使储存要求最小化, 而且能够在分析阶段与统计软件有效配合。

106. 数据结构为每一种统计单位使用不同的记录类型, 维持每一个所观察的统计单位与电脑文件记录之间的一一对应。举例来说, 要对住户名册上的数据进行管理, 就会为名册上的变量定义一种记录类型, 而且与每个人所对应的数据将被储存在该类型一份独立的记录中。与此相似的, 在食物消费模块中, 一种记录类型将与各种食物项对应, 而对应于每一种食物的数据将被储存在该类型一份独立的记录中。

107. 允许每一种记录类型包含的记录数量不同。这种方法节省了所需的储存空间, 因为文件不必要使每一种类型都占用最大的空间。

108. 在包括标识符之后, 调查所记录的每一个单位的实际数据都被记录在固定长度的区域, 记录顺序与问卷上的问题顺序保持一致。所有数据以标准ASCII(美国信息互换标准代码)形式储存。

109. Munoz(2003年)和世界银行(1991年)更加详细地讨论了住户调查的文件管理。

为进一步操作重构数据集

110. 为了便于进行适当的分析, 相关的数据库一定要包括取样过程的所有信息; 样本设计层的标记、初级取样单位、二级取样单位等; 以及每一个取样单位的权数。在估计所需数据以及估算这些估计数的抽样误差时, 都要使用上述信息。

111. 在数据输入之后, 通常需要重构数据集, 生成新文件, 并且对于部分现有数据域重新进行编码, 以定义新的变量, 使其更加便于制表和分析。为了能够对数据进行一些操作, 包括估计过程, 就需要重构数据集。

112. 最初完整的调查数据文件实际上可能包括从不同调查总体取样得来的单位信息(Rosen, 1991年)。譬如, 在住户预算调查中, 同一份初始文件中可能既包括被抽样住户的数据, 又包括被抽样个人的数据。为了估计住户总体和个人总体的统计特征, 调查人员需要分别备有一份含有每一个被抽样住户记录的文件, 还有一份含有每一个被抽样个人记录的文件。基于住户单位(对象)的数据集或文件被用来生成私人住户的数据(表格等)。基于个人单位(对象)的数据集或文件被用来生成私人住户里个人的数据(表格)。

113. 上述内容已明确表示: 住户调查通常有两种主要文件类型: 住户文件和个体(具体个人)文件。大多数情况下的文件都是住户文件, 因为它们包含住户变量(与观察单位或对象“住户”相关的变量)数值。有一些文件是个体文件(个人文件), 因为它们包含个体变量(与观察单位或对象“个人”相关的变量)数值。完整和最终的数据文件(数据集)包括每一个所调查的初级取样单位包含的所有受访住户和个人信息。

114. 下表9.2以1987年津巴布韦两次普查之间的人口调查为例, 说明如何对该调查的大文件进行重新组织, 以便进一步处理。表9.3例二是一个典型的住户预算调查文件。下面两个例子都是以Lagerlof(1988年)和Rosen(1991年)的材料为基础。

115. 出于制表目的, 大多数统计软件包都需要平面文件。很多可用的普通软件需要平面格式的数据。在平面文件中, 所有记录具有相同的变量集或者域, 而且长度相同。当每一个被调查者具有完全相同的数据域时, 文件被称作平面文件。每一份记录内数据域的安排完全一致, 并且记录的数量固定、格式相同。表9.4展示了1987年津巴布韦两次普查之间人口调查住户文件使用的平面文件格式。

116. 住户文件包含每一个被观察住户的一份记录, 记录包含下列信息:

- 住户身份;
- 取样设计参数;
- (住户)变量观察值;

表9.2
1987年津巴布韦两次普查之间人口调查使用的住户和个体文件

文件	类型	内容
住户	住户文件	住户身份(地区、省份、区等) 关于住户的所有问题的答案 派生变量, 如住户规模(从成员文件中得出)等
个人	个体文件	住户身份(HID)和个人身份(PID) 人口学特征: 年龄、性别、婚姻状况、住户常规成员、与户主关系
死者	个体文件	住户身份(HID)和死者身份(DID) 曾是住户常规成员的死者的详细信息: 性别、终年(死亡年龄)
女性≥12岁	个体文件	住户身份、个人身份 住户中每一位至少12岁或12岁以上女性的详细信息

表9.3
住户预算调查典型文件

文件	类型	内容
住户	住户文件	住户身份(地区、省份、区等) 对与住户有关的所有问题的答复 派生变量, 如住户规模(从成员文件中得出)等
成员	个体文件	住户身份和成员身份 人口学特征: 成员的年龄、性别、婚姻状况、教育程度等 主要活动的信息: 就业情况、职业等。
收入	个体文件	住户身份加上个人身份加上收入情况 收入来源:
食物	住户文件	住户身份加上食物种类 食物支出:
其他非耐用品	住户文件	住户身份加上非耐用品种类 非耐用品支出:
耐用品	住户文件	住户身份加上耐用品种类 耐用品支出:
农业	住户文件	住户身份加上农业项 农业支出:
农业资本	住户文件	住户身份加上农业资本项 农业资本支出:

表9.4
1987年津巴布韦两次普查之间人口调查住户文件使用的平面文件格式

身 份		取样设计参数							变 量 值			权数变量		
层	分区	EA	Hh	S _h	a _h	R _h	b _{hr}	S _{hi}	M _{hi}	m _{hi}	x	y	z	w
h	r	i	h								X _{hrhj}	Y _{hrhj}	Z _{hrhj}	W _{hrhj}

- 权数变量。

住户身份: $brij$ 组合代表: 住户 j 属于 b 层 r 分区的普查计数区(EA)。

取样参数: 本例的参数如下:

- S_b = 1982年取样层住户数量;
- a_b = 取样层EA样本规模;
- R_b = 取样层中样本代表的分区数量;
- b_{br} = 分区中被取样EA的数量;
- S_{bi} = 1982年EA包括的住户数量;
- M_{hi} = 1987年EA包括的住户数量;
- m_{hi} = EA包括的住户样本规模。

变量观察值: x, y, z 表示住户变量。

权数变量值: w 表示住户的权数变量。

117. 个人文件的组织与上表列出的住户文件组织类似。细微的差别有: 身份是个人身份(PID), 以及指数(k)是针对个人, 而“变量”指的是个人变量。

118. 调查数据集只需被组织为单独的平面文件(一种记录类型对应一个文件)进行发布, 因为原始结构的固长域格式也适合用来将数据转换至标准数据库管理系统(DBMSs), 以进行进一步操作, 或适合于将数据转换至用于制表和分析的标准统计软件。将数据转换至数据库管理系统非常简单, 因为原始结构差不多可以直接转换为标准DBF格式, 所有系统都接受DBF格式的表格输入(在此情况下, 记录标识符的作用是联系各表格之间的关系)(Munoz, 2003年)。

9.4.3.5. 估计流程和权数计算

119. 第六章详细描述了住户调查数据权数计算的基本原理和方法(另见本章末Rosen编写的参考书目)。一种通过观察值估计统计特征的运算法则被称作点估计程序。点估计的第一步是计算每一个被调查对象的权数。然后, “总值”的估计值由加权后观察值的总和算出(观察值乘以相应的权数)。

120. Munoz(2003年)很好地描述了微软Excel电子数据表的计算机化系统如何被用来执行住户调查的取样过程。Excel引导取样过程完成主要的步骤: 组织第一阶段标架; 选择概率与规模成正比的初级抽样单位; 以及计算选择概率和相应的取样权数。

121. 加权估计量的实际建构是简单直接的。可以从原始样本的数据集开始, 将观察值与规定的权数值相乘, 创建新的数据集, 然后运用标准公式和加权数据集计算参数。

122. 但是, 需要注意的是, 准确的权数必须结合三个成分(Yansaneh, 2003年), 包括所需的各种调整(同样参照第六章)。基本权数依照调查最初设计的规定, 说明从不同住户组被选出的概率差异。第二个调整便是说明各域或子群体的不答复率的差异。最后, 在某些情况下, 可能需要做出分层后调整, 使调查数据与独立来源(如最新的人口普查)的发布信息一致。

123. 使估计过程复杂化的另一个因素是对于域层级数据的需求提高。正如第三章所述, 一个域是一个子集, 需要进行单独估算。通常情况下, 可能在样本设计阶段明确域, 但是也可能是由推导数据得到域的。域也可能是一层、层级的结合、行政地区(省、区、农村/城市层次等)。域可能以人口统计学或社会经济学特点(如年龄、性别、民族、贫困等)定义。接下来, 我们将试图描述如何建构便于估算域的数据集。

124. 首先, 我们要设想出一个如上文津巴布韦两次普查之间人口调查所展示的数据(观察)文件(譬如住户文件)。这类文件为每一个被抽样住户保存一份记录。在调查过程的最后, 文件应包含每一个住户的如下信息:

- (a) 住户身份;
- (b) 取样参数;
- (c) 研究变量 x 、 y 和 z 的数值;
- (d) 住户估算权数值;
- (e) 住户是否属于类别 c ;
- (f) 住户是否属于域 g 。

125. 这些信息(除取样参数之外)表示方式如下:

- HID =被抽样住户身份标记。为了简单表述, 我们采用序列标号: $1, 2, \dots, n$ 。所以, n 代表取样总规模;
- x 、 y 和 z 是住户 X 、 Y 和 Z 变量的观察值;
- 如果住户属于类别 c , 那么 $c = 1$, 否则为0;
- 如果住户属于域 g , 那么 $g = 1$, 否则为0;
- w = 住户的估算权数。

126. 指示变量 c 和 g 的数值通常由其他变量的数值推出, 而不能直接观察得出。例如, 我们可以让类别 c 代表“贫困线以下”。住户不会直接被问及他们是否属于类别 c 。类别是推算得来的, 例如, 通过住户的收入数据和规定的贫困线推算而来。与之相似的, 通常需要从其他变量推算得来的值决定某住户是否属于某一特定的研究域 g (例如, 域 g 可能由有3个以上的儿童住户组成)。在估算阶段, 观察文件中应该有这类指标的数值。

127. 当观察文件中能够得到所有数据时, 该文件大体与下表9.5类似, 但下表未包括取样参数。

表9.5
包含住户调查变量最终数据的观察文件

HID	X	Y	Z	C	G	W
1	x_1	y_1	z_1	c_1	g_1	w_1
2	x_2	y_2	z_2	c_2	g_2	w_2
3	x_3	y_3	z_3	c_3	g_3	w_3
.
w
n	x_n	y_n	.	c_n	g_n	w_n

128. 以上内容只限于讨论住户调查总体的统计特征估计。个人调查总体的统计特征估计遵循同样步骤。一般情况下，个人的估算权数与其所属住户的估算权数相同。因为一般用户的所有成员都列在问卷中，所以只有在个人所属住户包括在住户样本中时，才将该个人包括在个人样本中。因此，个人的包含概率与其所属住户的包含概率相同。但是，需要注意的是，如果进行住户内二次抽样，那么上述情况不适用。譬如，在一些样本设计中，过程可能只需要在每一户中选择一位成人或一位男性和一位女性；在这些情况下，所选个人的权数单独计算，而且与住户权数不相同。

129. 为了保持完整性，一次住户调查估计过程的部分必须包括提供调查的抽样(或标准)误差估算，特别是对于那些最重要的将公诸于众的数据。本手册的第七章专门讨论这个问题。

9.4.3.6. 制表和数据库所用表格和数据集

130. 以下是一次统计调查的三种主要基本输出(Sundgren, 1995年):

- 宏观数据: 代表某种统计特征估计的“数据”; 生成宏观数据是调查的首要目的;
- 微观数据: 即“个体对象观察值”, 是调查产生的宏观数据的基础; 微观数据对于今后使用以及解读调查结果至关重要;
- 元数据: “描述潜在宏观和微观数据的意义、准确性、可用性和其他重要特点的数据”; 元数据对于为一个特定问题正确确认和检索相关统计数据以及正确解读和(再次)使用统计数据至关重要。

为了使调查结果得到更广泛的使用, 同时也为了提高网站等引用调查结果的能力, 考虑设计多维数据表格(立方体)可能也很有用处。

131. 住户调查项目最终创造出的数据档案应该是基于微观以及宏观数据组合的。为了达到这个目标, 必须详细描述多个调查所收集的信息结构。

132. 应该从三个阶段考虑数据储存(Lundell, 2003年):

- 储存: 如前面所述, 在数据录入期间, 数据存储的方式应该首先与采用的数据录入方式和数据清洗方式配合良好;

- **存入数据仓库：**当数据录入完成并经过清洗后，应该将数据加入数据仓库。仓库的结构将根据分析和发布数据的工具和方式进行调整；
- **存档：**项目数据存档方式应该与长期标准一致，以保证今后能够简单地进行数据检索。

133. 包含清洁数据的数据仓库可以用多种方式创建，可采用如下方式中的一种(参照本章附录，可以得到更多关于这些软件包的信息)：

- 平面文件；
- 关系型数据库(比如微软结构化查询语言(SQL)服务器)；
- 统计软件(比如，统计分析系统(SAS)或者社会科学统计软件包(SPSS))。

134. 最终数据的长期存档有一个主要选择。数据必须以单一ASCII格式的平面文件形式储存，并附有记录描述。大多数数据库系统和统计软件可以轻松将数据导出到这些文件中，也可以从文件中轻松导入数据。

附 录

调查数据处理不同阶段的可用软件

操作类型	可用软件
数据库管理系统	微软结构化查询语言(SQL)服务器2000, 标准版 微软Access 统计分析系统(SAS)
数据录入和编辑	直观基本系统(Visual Basic) 微软Access 集成微机处理系统(IMPS) 普查与调查数据处理系统(CSPro)
数据检索	统计分析系统(SAS) 社会科学统计软件包(SPSS) 微软Access 微软Excel
制表、分析和展示	微软Word 微软Excel 统计分析系统(SAS) 社会科学统计软件包(SPSS)
方差估计	CENVAR: (IMPS)方差计算软件成分 抽样误差实用数据计算和列举(CLUSTERS) 集成调查分析系统(ISSA) 调查数据分析(SUDAAN) 数据分析系统(SAS) 社会科学统计软件包(SPSS) 类集分析与回归软件包(PC-CARP)

微软办公软件

微软办公软件由微软公司开发, 是包括各种软件程序的综合性软件包, 包括:

- 微软办公软件Access是Office数据库管理程序, 提供了改进的易用性以及导入、导出和同可扩展标记语言(XML)数据文件一起使用的扩展功能;
- 微软办公软件Excel是Office电子表格制作程序, 包含了对XML的支持以及一些使信息的分析和共享变得更加容易的新增特性;
- 微软办公软件Word是Office文字处理软件;
- 微软SQL Server 2000是一个服务器数据库, 以便实现完整的企业项目和资源管理功能;
- 微软办公软件Outlook是Office个人信息管理器和通信程序, 允许用户在同一个地方对于电子邮件、日历等进行管理。

网址：<http://www.microsoft.com/office/system/overview.msp#EDAA>。

直观基本系统(Visual Basic)

微软于1987年推出Visual Basic。Visual Basic不仅是一种编程语言，也是一种完整的图形开发环境。即使是编程经验有限的用户，也能在这种环境下快速开发有效的、能够使用OLE(对象连接与嵌入)对象的Windows应用程序(如Excel电子表格)。Visual Basic还能够开发被用作数据库前端应用程序的程序，可被用作用户界面，收集用户信息，以美观实用的形式展示格式化输出，而很多SQL版本不如VB的输出形式美观实用。

Visual Basic的主要特色在于：用户可以轻松创建外表美观的图形程序，而不需要程序员编写大量代码。Visual Basic的主要对象被称作“form”，便于数据录入界面的开发。

网址：<http://www.engin.umd.umich.edu/CIS/course.des/cis400/vbasic/vbasic.html>。

抽样方差计算软件(CENVAR)

CENVAR是集成微机处理系统(IMPS)的方差计算成分。IMPS是一系列软件包，用于普查和调查数据的录入、编辑、制表、估计、分析和发布。美国人口普查局开发了集成微机处理系统(IMPS)。

网址：<http://www.census.gov/ipc/www/imps/>。

PC CARP

CENVAR基于个人计算机类集分析与回归软件包(PCCARP)。该软件包最初由艾奥瓦州立大学开发。PC CARP使用线性化过程计算方差。

网址：<http://www.census.gov/ipc/www/imps/>。

普查与调查数据处理系统(CSPro)

CSPro(普查与调查数据处理系统)是用于普查和调查数据的输入、编辑、制表和映射的公共域软件。IMPS和ISSA的开发者：美国人口普查局、宏大国际咨询公司和联邦软件程序服务机构共同努力，设计并实施了CSPro。美国国际开发署人口办公室为本软件包的开发提供了资金。设计CSPro的目的是最终代替IMPS和ISSA。

网址：<http://www.census.gov/ipc/www/imps/>。

抽样误差实用数据计算和列举(CLUSTERS)

抽样误差实用数据计算和列举CLUSTERS软件包最初是用来计算世界生育率调查项目的抽样误差的。它采用泰勒线性化方法计算抽样误差，已被多个发展中国家的住户调查，特别是人口和健康调查项目下的调查用来计算抽样误差。

集成调查分析系统 (ISSA)

宏大国际咨询公司专门为人口和健康调查项目开发了集成调查分析系统 (ISSA)。ISSA 已被运用到数据处理的各个阶段中, 如数据录入、数据编辑和制表。ISSA 还含有一个抽样误差模块, 使用刀切法计算复杂的人口统计量度, 如生育率和死亡率的抽样误差 (见宏大国际咨询公司 (1996年))。

统计分析系统 (SAS)

统计分析系统 (SAS) 于1966年由SAS公司开发。SAS是一个用于数据分析、文件管理以及抽样误差计算的电脑软件包 (见An和Watts (2001年), 可以了解SAS的最新功能)。

社会科学统计软件包 (SPSS)

社会科学统计软件包 (SPSS) 由SPSS公司开发。SPSS是一个用于数据分析、文件处理等功能的电脑软件包 (参照SPSS公司 (1988年), 可以了解SPSS的最新功能)。

调查数据分析

调查数据分析 (SUDAAN) 由研究三角研究所开发 (北卡罗来纳州, 研究三角园)。SUDAAN是一个综合性样本调查 (相关数据) 软件包, 在描述性分析和模拟分析方面都具有分析优势 (Shah、Barnwell和Bieler (1996年) 有更加详细的介绍)。

参考文献和补充读物

An, A和D. Watts (2001年)。《抽样调查数据分析的新型SAS程序》。SUGI论文集, 第23期。

北卡罗来纳州, 卡瑞: 美国SAS软件研究所。

Arnic等 (2003年)。“欧洲元数据生成系统: 以斯洛文尼亚统计系统为例”, 元数据生成研讨会发表论文, 卢森堡。欧统局第3331号文件。

澳大利亚统计局 (2005年)。《劳动统计: 概念、来源和方法》。堪培拉: 统计学概念图书馆。

来自[www.abs.gov.au/AUSSTATS/abs@nsf/DirClassManually Catalogue/59D849DC7B01FCC-ECA257/10F00I_F6E5B](http://www.abs.gov.au/AUSSTATS/abs@nsf/DirClassManually+Catalogue/59D849DC7B01FCC-ECA257/10F00I_F6E5B)。开放文件目录编号: 6102.0.55.001。

Backlund, S. (1996年)。“信息技术问题的未来方向”。提交给国家统计局的考察报告, 老挝人民民主共和国, 万象。

Brogan, D. (2003年)。《适用于发展中国家调查的数据分析软件比较》, 联合国统计司, 纽约。

纳米比亚中央统计局 (1996年)。1993/1994年度全国家庭收支调查 (NHIES)。行政和技术报告, 温得和克: 国家计划委员会。

- Chromy, J和S. Abeysasekara(2003年)。《调查数据的分析用途》，联合国统计司，纽约。
- Chronholm, P.和Edsfeldt(1996年)。“系统设计课程和研讨会”，提交给中央统计局(CSS)的考察报告，比勒陀利亚。
- Giles, M.(1996年)。《将数据转换为信息：社会分析手册》。堪培拉：澳大利亚统计局。
- Glewwe, Paul(2005年)。“发展中国家住户调查问卷设计概览”。选自《转型期与发展中国家住户抽样调查》。《方法研究》，第96号。出售品编号：E.05.XVII.6。
- Graubard, B.和E.Korn(2002年)。《抽样权重在调查数据分析中的使用》，联合国统计司，纽约。
- 国际劳工局(1990年)。《关于经济活动人口、就业、失业和就业不足的调查：国际劳工局概念和方法手册》，日内瓦：国际劳工局。
- Jambwa, M.和L.Olsson(1987年)。“非洲背景下的数据库技术应用”。特约论文，国际统计学会第四十六届会议，东京。
- Jambwa, M.、C.Parirenyatwa和B.Rosen(1989年)。《中央统计局数据处理：近来的历史教训》。中央统计局，哈拉雷。
- Lagerlöf, Birgitta(1988年)。《全国性住户调查的系统设计发展》。瑞典统计局研发报告，第4号。斯德哥尔摩：瑞典统计局。
- Lehtonen, R.和E.Pahkinen(1995年)。《设计和分析复杂调查的实用方法》。纽约：约翰威利父子出版公司。
- Lundell, L.(1996年)。《中央统计局信息系统策略》。提交给中央统计局(CSS)的考察报告，比勒陀利亚。
- _____ (2003年)。《普查数据处理经验》。提交给中央统计局(CBS)的报告，温得和克。
- 宏大国际咨询公司(1996年)。《抽样手册》，第三次人口和健康调查基本文件第6号，马里兰州卡尔夫顿市：宏大国际咨询公司。
- Munoz, Juan(2003年)。“住户调查数据管理指南”。选自《转型期与发展中国家住户抽样调查》。《方法研究》，第96号。出售品编号：E.05.XVII.6。
- Olofsson, P.(1985年)。《莱索托王国调查设计方案》。提交给统计局的关于劳动力调查的短期考察报告，马塞卢。
- Olsson, Ulf.(1990年)a。《发展中国家农业统计方法：对于国际咨询办公室经验的评价》，第12号。斯德哥尔摩：瑞典统计局(瑞典统计局，国际咨询办公室)。7月20日。
- _____ (1990年)b。《应用统计学讲义：专题报告》。TAN 1990:1。斯德哥尔摩：瑞典统计局国际咨询办公室。

- Pettersson, Hans (2005年)。“发展中国家住户调查标准抽样标架和标准样本设计”。选自《转型期与发展中国家住户抽样调查》。《方法研究》，第96号，出售品编号：E. 05. XVII. 6。
- Puide, Annika (1995年)。《达累斯萨拉姆Takwimu任务报告，1994年11月21日-12月21日》。坦桑尼亚统计局1994:20 (1995年1月20日)。斯德哥尔摩：瑞典统计局，国际咨询办公室。
- Rauch, L. (2001年)。《数据发布网站设计的最佳实践》。欧洲统计学家方法论资料会议。日内瓦：联合国统计委员会和欧洲经济委员会。
- Rosen, B. 和B. Sundgren (1991年)。“瑞典统计局所做调查得出的微观数据再利用文件”。斯德哥尔摩：瑞典统计局研究与发展部门工作文件。
- _____ (2002年)a. “抽样任务：莱索托王国标准样本架构”。提交统计局的考察报告，莱索托马塞卢，2002年6月1-15日。莱索托统计局2002:7。斯德哥尔摩：瑞典统计局国际咨询办公室。
- _____ (2002年)b. “关于标准样本调查估计程序的短期考察报告”。马塞卢：莱索托王国统计局。
- Rosen, Beugt (1991年)。“收入、消费和支出调查中的估计”，津巴布韦统计局1991:8:1。
- Shah, B.、B. Barnwell和G. Bieler (1996年)。《SUDAAN用户手册：7.0版》，北卡罗来纳州，研究三角园。研究三角研究所。
- Silva, P. Pedro Luis do Nascimento. (2005年)。“巴西的调查非抽样误差报告和补偿：现行方法和今后挑战”。选自《转型期与发展中国家住户抽样调查》。《方法研究》，第96号。出售品编号：E. 05. XVII. 6。
- SPSS公司 (1988年)。《SPSS/PC+V2.0基础模块手册》。伊利诺伊州芝加哥：SPSS公司。
- St. Catherine, Edwin (2003年)。《住户调查样本设计：实用准则》“数据处理、分析和发布审查”。审查住户调查样本设计手册草本的联合国专家组会议，纽约，2003年12月3-5日。
- Sundgren, B. (1984年)。《数据库和信息系统概念设计》，P/ADB报告E19。斯德哥尔摩：瑞典统计局。
- _____ (1986年)。《瑞典统计局面向用户的系统开发》。U/ADB报告E24，斯德哥尔摩：瑞典统计局。
- _____ (1991年)。《国家和国际统计办公室信息系统结构：指导方针和建议》。日内瓦：联合国统计委员会和欧洲经济委员会。
- _____ (1995年)。《指导方针：数据和元数据建模》。日内瓦：联合国统计委员会和欧洲经济委员会。

- Svensson, R. (1996年)。“普查数据录入应用程序”。提交给中央统计局(CSS)的考察报告，比勒陀利亚。
- Thiel, Lisa Olson(2001年)。“网站的设计和开发”。提交给统计局的考察报告，莱索托马塞卢，2001年11月12-23日。莱索托统计局：2001:17。12月28日。斯德哥尔摩：瑞典统计局国际咨询办公室。
- 联合国(1982年)。国家住户调查能力项目：调查数据处理：问题和程序审查。DP/UN/INT-81-041/1。纽约：联合国技术合作促进发展部统计办公室。
- _____ (1985年)。国家住户调查能力项目：住户收入支出调查：一项技术研究。DP/UN/INT. 88-X01/6E。纽约：联合国技术合作促进发展部统计办公室。
- Verma, Vijay(1982年)。抽样误差的估计和展示，世界生育率调查，技术公告，第11号(12月)。海牙：国际统计学会。荷兰沃尔堡。
- Wallgren, Anders等(1996年)。《图示统计数值和数据：创建更好的图表》。加利福尼亚，千橡市：赛奇(Sage)出版公司。
- 世界银行(1991年)。SDA调查工具：调整所涉社会问题的工具。哥伦比亚特区华盛顿。贫困和社会政策司，技术部，非洲司。
- Yansaneh, I. (2005年)。《转型期与发展中国家住户调查样本设计问题概览》。联合国统计司，纽约。选自《转型期与发展中国家住户抽样调查》。《方法研究》，第96号，出售品编号：E. 05. XVII. 6。

附件 一

调查样本设计的基本原理

A. 1. 引言

1. 抽样是这样一种方法：选定一部分人口，然后根据从这部分人口获取的调查结果，对选出这部分人口或样本的整个人口做出推断。总的来讲有两类样本，即概率样本和非概率样本。本手册集中在概率样本方面。下面围绕调查单位、样本设计和各项基本抽样策略几个方面，结合例证进行概述。

A. 2. 调查单位和概念

2. 我们先从调查抽样中常用的调查单位和概念的定义入手。**要素**：一个人口的要素(单位)是指为其寻求信息的单位。它们可能是构成拟对其做出推断的人口的基本单位。譬如讲，在住户生育率调查中，育龄期的妇女通常就是最终要素。为了便于调查收集数据，确保要素定义明确并且在实际当中容易识别是绝对必要的。

3. **人口**：人口是上面定义的要素之总量。因此，要素是构成并界定人口的基本单位。有必要根据以下几方面界定人口：

- 内容，要求界定构成人口的要素种类和特征；
- 范围，系指与覆盖面有关的地域边界；
- 时间，系指人口存在的时间段。

4. **观察单位**：这些是从中获取观察数据的单位。在采访调查中，他们叫做受访者。报告单位是报告调查所获信息的要素。注意，有些情况下观察单位和报告单位也许各不相同。譬如讲，在5岁以下儿童调查中，须由父母作为代理人提供有关其子女的信息。在这种情况下，样本中选定的儿童是观察单位，而其父母是报告单位。

5. **抽样单位**：抽样单位用于选择准备纳入样本的要素。在要素抽样中，每个抽样单位包含一个要素；而在整群抽样中，一个抽样单位是由一个叫做类集的要素群体构成的。举例来讲，一个计数区(EA)作为第一阶段抽样单位包含一个住户类集。同一调查有可能使用不同的抽样单位。使用抽样单位等级层次的多级抽样法就是很好的例证(见第三章)。

6. **样本单位:** 选定的抽样单位可以叫做样本单位, 而所研究的样本单位特征值就叫做样本观察值。**分析单位:** 这是一个在制表和分析阶段使用的单位。它可以是一个要素单位, 也可以是要素单位群体。值得注意的是, 如前所述, 分析单位和报告单位无须相同。

7. **抽样标架:** 抽样标架用于鉴别并将抽样单位选入样本, 同时, 它也是根据抽样数据进行估计的依据。这就是说, 需要选样的人口必须以实物形态为代表。理想的是标架应该拥有属于研究人口并具有适当识别特征的所有抽样单位。这些标架应当是排他的, 最好也是相互排斥的(详见第四章)。调查中常用的标架有清单标架、区域标架和多元标架。

8. **清单标架:** 一个清单标架包含一个可以从中直接选取样本的抽样单位清单。建议这种标架应有每个抽样单位的相关而准确的信息, 比如规模和其他特征。这种补充信息有助于设计和/或选择有效的样本。

9. **区域标架:** 区域标架一般来说是住户调查中常用的多级标架。就此而言, 这种标架包含一个或多个阶段的区域单位。譬如讲, 在一个二级抽样设计中, 标架包含若干可以称之为“初级抽样单位”(PSUs)的类集; 在选定的初级抽样单位中, 住户清单就成了第二阶段标架。一般来讲, 每个选样阶段都需要有标架。沿着单位等级层次越往下, 标架的耐用期越短。

10. **区域单位:** 区域单位覆盖特定陆地地区, 有明确的边界; 这些边界可能是一些实物特征, 比如道路、街道、河流、铁路线等, 也可能是代表行政区划之间官方边界的虚线。通常在一个国家的较小行政单位内划定普查计数区。这有利于累积作为“域”的行政单位计数。

11. 凭借住户调查使用的标架, 应能访问调查人口中的所有抽样单位, 以便使每个单位都有一个已知的被选入样本的非零概率。通过从标架中抽样可以进行访问, 通常需要经过两个或两个以上的选样阶段。第一抽样阶段的标架必须包括所有指定的抽样单位。后来的阶段只需要前面阶段入选样本单位的选样标架。抽样标架可以储存在付印文本和/或电子媒体中。

A. 3. 样本设计

12. 一般而言, 样本设计是指样本选择和估计。因此这个问题就关系到如何选择需要纳入调查范围的一部分人口。实际当中, 样本设计需要确定样本容量和结构, 同时考虑到调查成本。最理想的样本设计就是在既定的调查成本范围内获取最高精确度的成果, 或者在达到特定精度水平的前提下最大限度地减少成本。

13. 不过, 从一开始就应强调指出, 样本设计不能脱离调查设计与实施的其他方面而独立存在。总的来说, 对于特定人口而言, 抽样理论所关注的就是调查估计数和与其有关的抽样误差与样本容量和结构的关系如何。

A. 3. 1. 概率样本设计的基本要求

- 必须明确界定目标人口;
- 必须有一个(或在多级样本中须有多个)抽样标架;

- 调查目标必须就调查内容、分析变量和分解层次(比如, 你是否需要国家级、城/乡、省级、区级的估计数?)等方面毫不含糊地做出明确规定;
- 应考虑到预算和现场工作方面的制约因素;
- 必须明确规定精度要求以确定样本容量。

A. 3. 2. 概率抽样对大规模住户调查的重要性

- 使用它可以在选样中覆盖整个目标人口;
- 使用它可以减少抽样偏差;
- 使用它可以把样本结果推广到从中选择样本的整个人口;
- 使用它可以计算样本误差, 作为衡量可靠性的尺度;
- 据论证, 使用这种抽样法可以使调查者免于为使用不科学的方法说抱歉。

A. 3. 3. 选样、实施和估算的程序

- 人口中的每个要素均应在从中选取样本的标架中有代表;
- 样本的选择应建基于随机方法, 这种方法赋予每个单位以特定的选择概率;
- 所有——且唯有——选定的单位必须计数;
- 在根据样本估计人口参数的时候, 出自每个单位/要素的数据必须根据其选择概率进行加权。

14. 单位的随机选择减少了获取非代表性样本的机会。因此, 随机化是克服意外偏差因素影响的一个安全方法。选样方法的使用取决于所采用的抽样方案。样本设计越复杂, 对选样程序的要求越高。

A. 4. 概率抽样策略的基本原则

15. 为了样本设计的目的, 现已开发了一系列概率抽样方法, 其中包括简单随机抽样、系统抽样、分层抽样, 还有整群抽样。下面我们结合例证来简略讨论这些方法。

A. 4. 1. 简单随机抽样

16. 简单随机抽样(SRS)是一种概率选样方法, 使用这种方法, 人口的每个要素都有平等的被选机会/概率。样本选择可以带置换, 亦可不带置换。大规模住户调查很少使用这种方法, 因为它在登记造册和差旅方面成本高昂。可以把它看作适用于以往没有人口结构信息可用的情况下进行概率抽样的基本形式。简单随机抽样之所以有一定吸引力, 是因为它在选样和估算(比如抽样误差的估算)程序方面比较简单。

17. 虽然简单随机抽样法使用不是很多,但它属于抽样理论基础部分,主要因为其简单的数学属性。因此,大多数统计理论和方法都假定要素的简单随机选择。的确,所有其他概率样本选择法都可以被视为对简单随机抽样的制约,这些方法抑制人口要素的某些组合。简单随机抽样有两个作用:

- 它为其他抽样方法的相对效力的比较提供了一条基线;
- 它可以在较为复杂的设计条件下——比如多级聚类抽样设计和分层抽样设计——被用作选择要素单位的最后方法。

下面的范例解释了简单随机抽样设计下的选择概率计算方法:

1. 首先,我们考虑一个拥有100个住户 $H_1, H_2, \dots, H_i, \dots, H_{100}$ 的有限人口,其收入值为 $X_1, X_2, \dots, X_i, \dots, X_{100}$ 。

在这个范例中,任何特定单位的入选概率均为 $\frac{1}{100}$ 。

2. 作为第二例,我们提请注意,为了提取一个住户样本,可以在标架/清单中对目标住户进行系列编号。譬如讲,使用随机编号可以选取一个规模为25个住户的样本。对于等概率选择方法(EPSEM)而言, f 是要素的总体抽样份额。

于是, $f = \frac{n}{N}$ 。

设样本容量为 $n=25$,住户总数为 $N=100$,那么,抽样份额即选择概率为

$$\frac{25}{100} = \frac{1}{4}。$$

A.4.1.1. 简单随机抽样法的样本选择类型

18. 简单随机抽样有两种常用的选样方法,即:

- (a) 带置换的简单随机抽样(SRSWR);
- (b) 不带置换的简单随机抽样(SRSWOR)。

带置换的简单随机抽样

19. 带置换的简单随机抽样的基础是每次取样后通过置换人口中的被选定要素进行人口随机选样。每次抽样后,每个要素的选择概率保持不变,任何被选定的独立样本都是彼此独立的。这一特性说明了为什么在许多理论统计研究中都把简单随机抽样用作缺省抽样方法。另外,由于简单随机抽样大大简化了估计量公式(比如方差估计量公式),所以把它作为参考来使用。我们在下面第20段根据带置换的简单随机抽样法分别给出估计平均值的公式(A. 1)和估计样本平均值方差的公式(A. 2)。这些公式都有数字实例说明。

20. 假定使用带置换的简单随机抽样法选定了一个由 n 个单位组成的样本,并且收集了关于其变量 x 的信息。那么,平均值和方差的公式分别为:

1. 平均值

$$\bar{x} = \frac{1}{n} \sum_i^n x_i = \frac{1}{n} [x_1 + x_2 + \dots + x_n] \quad (\text{A.1})$$

若 $x_1=24, x_2=30, x_3=27, x_4=36, x_5=31, x_6=38, x_7=23, x_8=40, x_9=25, x_{10}=32,$

$$\text{则 } \bar{x} = \frac{24+30+27+\dots+25+32}{10} = 30.6$$

2. 方差

$$V(\bar{x}) = \frac{s^2}{n} \quad (\text{A.2})$$

$$\text{式中, } s^2 = \frac{1}{n-1} \sum_i^n (x_i - \bar{x})^2 = \frac{1}{n-1} \left[\sum_i^n x_i^2 - \frac{x^2}{n} \right] = \frac{1}{n-1} (\sum_i^n x_i^2 - n\bar{x}^2) \quad (\text{A.3})$$

$$x^2 = (\sum_i^n x_i)^2 = 93\,636$$

用这些数值来计算,

$$s^2 = \frac{(9\,684 - 9\,364)}{9} = 35.56$$

$$V(\bar{x}) = \frac{35.56}{10} = 3.56$$

$$\text{Se}(\bar{x}) = \sqrt{3.56}$$

不带置换的简单随机抽样

21. 凭直觉, 最好不带置换地进行抽样, 这样可以获取较多的信息, 因为样本单位不可能重复。因此, 不带置换的简单随机抽样策略成为最经常使用的简单随机抽样法。使用这种方法, 选择过程一直持续到选出 n 个不同的单位, 而所有重复一律忽略。这就等于保留一个或多个已被选中的单位, 同时用同等的概率继续从人口的其余单位中选取下一个单位。

下面是不带置换的简单随机抽样法的一些特性:

- 它给出固定的样本容量;
- 其结果是每个要素/单位都有均等的选择概率 (EPSEM);
- 跟带置换的简单随机抽样法一样, 样本的平均值和方差是人口参数的无偏估计数。

22. 我们将在下面第24段提供用来在不带置换的简单随机抽样的设计中估计平均值和方差的公式 (A. 4和A. 5)。此外, 我们还提供数字实例来说明如何计算样本的平均值和方差。

23. 假定某地区的小学总数为275所。采用不带置换的简单随机抽样法定选了一个由55所小学组成的样本。下面的数字是每所入选学校的雇员人数 (y_i)。

5	10	32	6	8	2
15	16	35	7	50	6
2	6	47	20	20	6
7	6	35	6	16	2
21	2	48	4	15	2
7	5	46	6	7	
4	4	8	2	6	
7	2	7	8	2	
5	12	10	6	2	
2	40	7	7	19	

$\sum y_i = 688$, 雇员总人数

$\sum y_i^2 = 18\ 182$

1. 样本的平均值是

$$\bar{y} = \frac{\sum y_i}{n} \quad (\text{A.4})$$

式中, n 是样本容量。

使用这些数字计算,

$$\bar{y} = \frac{688}{55} = 12.5$$

2. 样本平均值的方差

$$V(\bar{y}) = 1 - f \frac{s_y^2}{n} \quad (\text{A.5})$$

式中, $1-f$ 是人口校正系数, 而

$$s_y^2 = \frac{1}{n-1} [\sum y_i^2 - n\bar{y}^2] = \frac{1}{54} [18\ 182 - 8\ 594] = 177.56 \quad (\text{A.6})$$

那么,

$$V(\bar{y}) = \left(1 - \frac{55}{275}\right) 177.56/55 = 2.58 \quad \text{同时} \quad \text{Se}(\bar{y}) = \sqrt{2.58}$$

A. 4. 2. 系统抽样

24. 系统抽样是一种概率样本选择方法, 其中通过选择每第 k 个人口要素来获取样本, 式中, k 是一个大于1的整数。样本的第一个数必须从第一个 k 要素中随机选出。样本是从一个按顺序排列的清单中选取的。这是一种常用的选择方法, 尤其在单位数量很多而且从1到 N 依次排列的情况下使用更为频繁。假定单位总数 N 是所需样本容量 n 的一个整数倍数, 而 k 是一个整数, 故 $N = nk$ 。然后在1和 k 之间选择一个随机数。设2为随机起点, 那么样本容量即为 n , 其单位系列编号为:

$$2, 2 + k, 2 + 2k, \dots, 2 + (n-1)k$$

由此可见，该样本包括第一个随机选择的单位和每第 k 个单位，直至获得所需的样本容量。抽样区间 k 将人口分割成一定数量的类集或群组。在本程序中我们用概率 $1/k$ 选择一个单位类集。既然第一个数是从1至 k 中随机提取的，所以在设定为均等的类集中的每个单位都有相同的被选概率， $1/k$ 。

A.4.2.1. 线性系统抽样

25. 如果单位总数 N 是预期样本容量的倍数，换言之，如果 $N = nk$ ，式中， n 是预期样本容量，而 k 是抽样区间，那么，每个可能的系统样本中的单位数为 n 。在这种情况下，该系统相当于把 N 个单位划分为 k 个样本，每个样本有 n 个单位并且按照 $1/k$ 的概率选定一个类集。当 $N = nk$ 的时候， \bar{y} 是人口平均值 \bar{Y} 的无偏估计量。另一方面，如果 N 不是 n 的倍数，则使用系统抽样法按照抽样区间 k 选取的单位数等于最接近于 N/k 的整数，但未必等于 n 。因此，如果 N 不等于 nk ，样本容量就不同了，而样本平均值将是人口平均值的一个有偏差的估计量。下面图A. 1解释了采用线性系统抽样法的样本选择。

图A. 1
线性系统抽样(样本选择)

1	2	3	4	5	6	7	8	9	10
11	12	13	14	15	16	17	18	19	20

上面的范例解释了从一个20个学生的班级选出一个由4个学生组成的样本。其随机起点为3， $N=20$ ， $n=4$ ，而 $k=5$ 。所得样本单位分别标记为3、8、13和18。

A.4.2.2. 圆形等距系统抽样

26. 前面我们提请注意，在线性系统抽样中实际样本容量可能跟预期的规模不一样；如果 N 不是 n 的倍数的话，样本平均值就是人口平均值的一个有偏差的估计量。可是，圆形等距抽样法克服了上述局限性。根据圆形等距系统选择方法，列表以圆形排列，最后单位接下来是第一单位。随机起点是从1至 N 选择而不是从1到 k 选择。然后附加上第 k 单位，直至恰好 n 个要素被选中。当选到清单结尾处的时候，继续从头选起。图A. 2提供了一个按照圆形等距系统抽样法进行样本选择的解释，其中， $N=20$ ， $n=4$ ， $k=5$ ，而随机起点是7。因此，选中的单位分别是7、12、17和2。

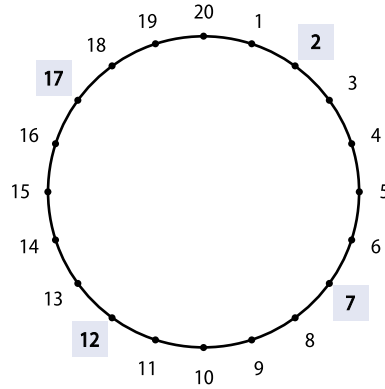
A.4.2.3. 系统抽样法的估算

27. 下面给出估算总数的公式(A. 7)、估算样本平均值的公式(A. 8)和估算方差的公式(A. 9)及相关的数字例证，以说明如何计算估计总数、样本平均值和方差。

1. 为了估计总数，需用样本总数乘以抽样区间，因此

$$\hat{Y} = k \sum y_i \quad (\text{A.7})$$

图A. 2
圆形等距系统样本选择



人口平均值的估计数为

$$\bar{y} = k \frac{\sum y_i}{N} \quad (\text{A.8})$$

2. 方差的估算很复杂, 不能单靠一个系统样本做出严谨的估计。解决难题的办法就是假定单位的编号是随机的; 在这种情况下可以把系统样本当作随机样本来对待。因此, 平均值的方差估计数可由下式给出:

$$V(\bar{y}) = \frac{1}{n} \left(1 - \frac{n}{N}\right) \sum s^2 \quad (\text{A.9})$$

式中, $s^2 = \frac{1}{n-1} \sum (y_i - \bar{y})^2$ 且 $\bar{y} = \frac{\sum y_i}{n}$

28. 一个系统样本无偏方差的严格估计数可以通过从特定人口选择不止一个系统样本来计算。

数字例证

29. 假定某省有180个商业农场, 各有30头或以上的牲畜。现以 $k=6$ 的抽样区间使用系统抽样法选取一个由30个农场组成的样本。

在这30个入选农场中, 给出的牲畜数(y_i)如下:

60	200	45	50	40	79	35	41	30	120
300	65	111	120	200	42	51	67	32	40
46	55	250	100	63	90	47	82	31	50

而 $\sum y_i = 2\,542$

1. 牲畜估计数是

$$\hat{Y} = k \sum y_i = 6 \times 2\,542 = 15\,252$$

2. 估计平均每个农场的牲畜数是

$$\bar{y} = k \frac{(\sum y_i)}{N} = 6 \times 2\,542 / 180 = 84.7 \approx 85$$

3. 在假定农场随机编号的基础上计算的样本平均值方差是

$$V(\bar{y}) = 1 - f \frac{s_y^2}{n} \tag{A.10}$$

$$\text{式中, } s_y^2 = \frac{1}{n-1} \left\{ \sum y_i^2 - \frac{(\sum y_i)^2}{n} \right\} = \frac{1}{29} (348\,700 - 215\,392.13) = 4\,596.80$$

$$\text{因此 } V(\bar{y}) = (0.833)(153.227) = 127.64$$

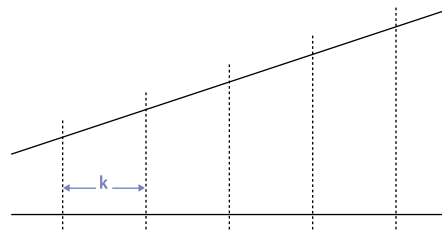
$$\text{而 } \text{Se}(\bar{y}) = \sqrt{127.64} = 11.30$$

30. 使用系统抽样法有利也有弊。

- (a) 好处:

- 第一单位的选择决定整个样本。这对现场作业是个好兆头，因为调查人员可以在编列单位清单的时候现场选定最终抽样单位；
- 由于标架中的各单位都有适当编号，所以样本在整个人口中分布均衡。不过，如果人口中有某种趋势的话，样本估计数会比较精确；
- 系统抽样可以提供隐含分层。下面图A. 3说明了通过单调线性趋势进行的隐含分层。

图A. 3
单调线性趋势

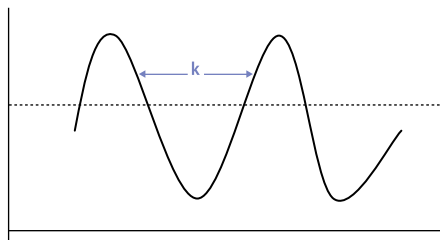


- (b) 弊端:

- 如果人口中有阶段性变化，系统抽样就会产生要么估计不足要么过高估计的结果。在这种情况下，抽样区间就会与数据取得一致。举例来讲，如果你在研究一个繁忙城市街道24小时的交通流量，而假如你的抽样区间恰好落在高峰时段，那么你就总会获得高数值。因此，研究就会产生过高估计结果。图A. 4说明了在使用系统抽样的情况下阶段性变异可能导致产生不可靠的估计数；
- 严格说来，你不能靠单一系统样本获取严谨的方差估计数；

- 这种选样方法容易被某些调查者/现场工作人员所滥用。

图A. 4
阶段性波动



A. 4. 3. 分层抽样

31. 在分层抽样方法中，人口中的抽样单位被分成叫做“层”的群体。通常，分层是为了把人口细分成内部同质的异质群体。总的来说，如果就辅助变量(叫做分层变量)而言抽样单位属于同质，那么层估计量的可变性一般就会减少。应着重指出的是，层次和层次之间在抽样和估算程序上不尽相同，在这个意义上，分层有很大的灵活性。

32. 因此，在分层抽样中我们把或多或少相似的单位/要素归类成组，以便使每个层次内的方差 δ_b^2 都比较小。与此同时，必须使不同层次的平均值 (\bar{x}_b) 尽可能有所差异。通过在层次范围内适当组合作为研究对象的特征估计量获取适当的整个人口估计数。

A.4.3.1. 分层抽样的好处

33. 分层抽样的主要好处是可以提高估计数的精度，并且有可能在不同层次使用不同的抽样程序。此外还发现在以下几方面分层是有益的：

- 如果人口分布偏斜，那么许多情况下可以采用大份额抽样，从一些较大单位选取样本。这种方法比较重视特大单位，而最终减少了层内抽样的可变性；
- 如果一个调查组织在全国不同行政区分设了几个现场办事处，那么把这些行政区当作层次来处理对于组织现场工作是有利的；
- 如果不论对于整个人口还是对于某些子群(比如省、城/乡、性别等)都要求估计数不得超出特定误差限度，那么，通过分层可以很方便地提供这种估计数；
- 如果抽样标架有地区级的或特定单位类别的次级标架的形式可用，那么把这种次级标架当作选择样本的层次来使用，不但操作方便，而且经济实惠。

A.4.3.2. 分层抽样步骤概要

- 抽样单位的整个人口被划分为内部同质而外部异质的分组人口；
- 在每一层次，从该层的所有抽样单位中单独选取一个样本；

- 根据每一层获取的样本，单独计算层均值(或任何其他统计数据)。然后，譬如讲，层均值经适当加权形成总体均值的组合估计数；
- 通常，如果调查目标是总体——譬如国家级——估计数并且调查目的是综合性的，那么在层内采用比例抽样法；
- 如果调查以子群域为主，譬如讲需要国内行政分区的、可变性均等的估计数，那么在这种情况下就使用非比例抽样法。

A.4.3.3. 标志法

34. 采用分层抽样法涉及到许多符号和下标。因此我们来界定此种抽样法常用的一些标志和符号。

人口值

对于 H 层次而言，每一层的要素总数由 $N_1, N_2, \dots, N_b, \dots, N_H$ 来表示。

这种信息通常是未知的。人口总值是

$$\sum_b^H N_b = N \quad (\text{A.11})$$

层均值是

$$\bar{X}_{bi} = \frac{1}{N} \sum_i^{N_b} X_{bi} = \frac{X_b}{N} \quad (\text{A.12})$$

式中， X_{bi} 是第 b 层的第 i 要素值，而 X_b 是第 b 层的总数。

A.4.3.4. 权数

35. 权数一般代表层次中的人口要素比例，而

$$W_b = \frac{N_b}{N} \quad (\text{A.13})$$

因此 $\sum W_b = 1$

$$S_b^2 = \frac{1}{N-1} \sum_{i=1}^{N_b} (X_{bi} - \bar{X})^2 \quad (\text{A.14})$$

A.4.3.5. 样本值

36. 样本值是根据层内选定的 n_b 个要素计算的一个估计数。本节描述分层抽样设计中常用的一些符号。

(a) 对于 H 层, 设每层的样本容量由 n_1, n_2, \dots, n_h 表示, 式中, $\sum n_b = n$ 是样本总体规模。

(b) 设 x_{bi} 为 b 层中的样本要素 i 。

(c) 那么

$$\bar{x}_b = \frac{1}{n_b} \sum_{i=1}^{n_b} x_{bi} \text{ 即为 } b \text{ 层的样本均值。} \quad (\text{A.15})$$

(d) 那么

$$\bar{x}_{st} = \sum W_b \bar{x}_b \text{ 即为总体样本均值。} \quad (\text{A.16})$$

(e) 那么

$$f_b = \frac{n_b}{N_b} \text{ 即为 } b \text{ 层的抽样份额。} \quad (\text{A.17})$$

第 b 层第 n_b 要素的方差由下式给出:

$$v(\bar{x}_b) = \sum \left[1 - \frac{n_b}{N_b} \right] \frac{s_b^2}{n_b} \quad (\text{A.18})$$

式中, s_b^2 是第 b 层的要素方差, 由下式给出:

$$s_b^2 = \frac{\sum (x_{bi} - \bar{x}_b)^2}{(n_b - 1)} \quad (\text{A.19})$$

样本均值方差由下式给出:

$$v(\bar{x}_{st}) = \sum W_b^2 (1 - f_b) \frac{s_b^2}{n_b} \quad (\text{A.20})$$

下面讨论两种分层抽样策略, 即比例分层和非比例分层。

A.4.3.6. 比例分层

37. 分层抽样中的比例分配需要所有层次使用统一的抽样份额。这就是说, 每层选取相同比例的单位。举例来讲, 如果我们决定选择一个占总体10%的样本, 这就意味着我们将从每层选取10%的单位。由于所有层次的抽样率都是一样的, 所以各个层次的样本中所选取的样本要素不尽相同。每层的样本容量都与该层的要素数成正比。

在这种情况下, 抽样份额由 $f_b = \frac{n_b}{N_b} = \frac{n}{N}$ 给出, 就是说, 这是个等概率选择(EPSEM)设计。

$$\text{样本平均值是 } \bar{x}_{st} = \sum W_b \bar{x}_b \quad (\text{A.21})$$

$$\text{总体均值方差是 } v(\bar{x}_{st}) = \frac{(1-f)}{n} \sum W_b s_b^2 \quad (\text{A.22})$$

A.4.3.7. 非比例分层

38. 非比例抽样法就是在各个层次使用不同的抽样率。目的在于，给各层次指定的抽样率要有利于获得每单位成本的总体均值最小方差。

39. 在使用这种方法的时候，特定层次的抽样率与该层的标准偏差成比例。这就是说，拟从任何层次选取的抽样单位数将不仅取决于要素总数，而且取决于辅助变量的标准偏差。

在非比例分配法中还引进了价值函数概念。比如，

$$C = C_o + \sum c_b n_b \quad (\text{A.23})$$

式中， C_o 是固定成本，而 c_b 是覆盖特定层样本的成本。

在许多情况下我们可以假定 c_b 是所有层次的一个常数。莱蒙分配公式是计算层次间样本非比例分配常用的公式之一。

在 c_b 为常数而 $n = \sum n_b$ 即总体样本容量是固定的情况下。

在一个层次内应选取的单位数由下式给出：

$$n_b = \frac{W_b s_b n}{\sum W_b s_b} \quad \text{或} \quad n_b = \frac{N_b s_b \cdot n}{\sum N_b s_b} \quad (\text{A.24})$$

方差由下式给出：

$$v(\bar{x}_{st}) = \frac{(\sum W_b s_b)^2}{n} - \frac{1}{N} \sum W_b s_b^2 \quad (\text{A.25})$$

减号右侧项是一个有限人口校正系数；如果是在一个非常大的人口中抽样，也就是说，如果抽样份额很小的话，这个校正系数可以删除。

A.4.3.8. 总的意见

- 一般情况下，总体值 S_b 和 C_b 都是未知的。因此可以根据以往的调查或试点抽样调查进行估计；
- 非比例分配对于选择比例不是很有效；
- 在综合调查中可能会在待优化的变量方面有冲突；

- 非比例分配的一般结果是最小方差。

40. 下面的例证说明了在比例和非比例分层抽样设计中的样本容量和方差的计算。在这个虚拟例证中, 各学校根据雇员人数进行分层。某省的小学校总数为275所。选取了一个由55所学校组成的样本并且根据雇员人数对其分层。

A.4.3.9. 层内样本容量的确定

41. 请参见下面表A. 1。

表A. 1
按雇员人数分层的学校数

层次	每所入选学校的雇员人数 (y_{hi})	每一层的学校总数 (N_h)	按层选定的学校数		W_h	s_h^2	s_h	$W_h s_h$	$W_h s_h^2$
			比例分配 (n_h)	非比例分配 (n_h)					
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1	2, 4, 2, 2, 4, 2, 2, 4, 2, 2, 2, 2, 2, 2, 5, 5	80	16	8	0.2909	1.663	1.289	0.3750	0.48
2	7, 7, 7, 6, 8, 7, 7, 6, 7, 6, 6, 8, 6, 7, 8, 6, 7, 6, 6, 6	100	20	6	0.3636	0.537	0.733	0.2665	0.19
3	10, 12, 10, 15, 21, 16, 20, 20, 16, 19, 15	55	11	18	0.2000	15.564	3.945	0.7890	3.11
4	32, 35, 35, 48, 46, 47, 50, 40	40	8	23	0.1455	48.836	6.989	1.0169	7.10
共 计		275	55	55	1.0000			2.4474	10.90

注: N = 是小学校总数

n = 是整个样本中的小学校总数

N_h = 是第 h 层的规模

n_h = 是第 h 层的样本容量

比例分配

比例分配法使用普通的抽样份额。因此, $\frac{n}{N} = f$ 是用于特定层次单位总数的总体抽样份额。

在上例中, $f = \frac{55}{275} = 0.2$ 或20%。

表A. 1第4栏列出了样本容量的分布情况; 比如, 第1层的样本容量是 $n_h = 0.2 \times 80 = 16$ 。

非比例分配

获得不同层次样本容量的公式由下式给出:

$$n_b = \frac{W_b s_b}{\sum W_b s_b} (n) \quad (\text{A.26})$$

例如第1层 $n_b = \frac{0.3750}{2.4474} \times 55 = 8$ 。

该表第5栏给出了其余结果。

A.4.3.10. 方差的计算

42. 比例分层和非比例分层的方差计算方法分别由公式A. 27和A. 28来说明。

比例分层：

$$V(\bar{y}_{prop}) = \frac{1-f}{n} \sum w_b s_b^2 = \frac{(1-0.2)}{55} (10.9) = 0.16 \quad (\text{A.27})$$

非比例分层：

$$V(\bar{y}_{opt}) = \frac{(\sum w_b s_b)^2}{n} - \frac{1}{N} \sum w_b s_b^2 = \frac{(2.4474)^2}{55} - \frac{10.9}{275} = 0.07 \quad (\text{A.28})$$

A.4.3.11. 总体上

$$v(\bar{x}_{st})_{OP} \leq v(\bar{x}_{st})_{PROP} \leq v(\bar{x}_{st}) \leq (\bar{x}_{st})_{SRS} \quad (\text{A.29})$$

A. 4. 4. 整群抽样

43. 前面几节讨论的抽样方法是把要素抽样单位视为按照标架清单排列的；在这种排列方式下，可以直接从标架选取个体单位。在整群抽样中，较高级别的选择单位——譬如计数区(见第三章)——包含不止一个要素单位。在这种情况下抽样单位是类集。举例来讲，在某城市选择一个由住户组成的随机样本的简单方法，可能需要有包括所有住户的一份清单。这或许办不到，因为实际上该市也许没有包括所有住户在内的完整标架。为了绕过这个问题，可以按街区的形式组成类集。然后选取一个由街区组成的样本，接着在选定的街区产生住户清单。如果需要的话，可按譬如10%的抽样率从每个街区提取一个住户样本。

A.4.4.1. 采用整群抽样法的理由

44. 下面是赞成使用整群抽样法——尤其是在多级样本设计的情况下——的一些理由。

- 通过聚类可以减少差旅费和与收集数据有关的其他费用；
- 它可以改进监督、控制、后续覆盖面以及对收集数据的质量有影响的其他方面；

- 因为分阶段实施, 它可以降低标架的构建成本。譬如就像第三章讨论的那样, 在多级抽样中, 只是在选择初级抽样单位(亦即第一阶段的类集)的时候才需要覆盖整个人口的标架。在任何较低的阶段, 都只需要上一阶段入选单位范围内的标架;
- 另外, 较大和较高阶段的标架一般都比较耐久, 因而可在较长时期内使用。小单位(比如住户、尤其是个人)清单则一般在很短的时期内就会变得陈旧过时;
- 在实施调查的过程中整群抽样法有利于行政管理。

45. 总的来说我们注意到, 通过拿一个类集样本和一个同样大小的要素样本作比较, 我们会发现, 在整群抽样中平均每个要素的成本较低, 因为要素的列表和/或定位成本比较低。另一方面, 由于类集内部的不规则同质性(同类相关)的缘故, 要素方差较高。让我们通过考虑单级设计来解释整群抽样(第三章详细介绍并讨论了多级设计)。

A.4.4.2. 单级整群抽样

46. 在特定行政区也许不能获得一份包括所有住户的清单并从中选取一个样本。可是有可能发现上一次调查编列的或者为行政目的保管的村庄清单。在这种情况下, 我们可以提取一个村庄样本, 然后在入选的村庄获得有关所有住户的信息。这就代表一个单级整群抽样设计, 因为在选定一个村庄样本之后, 对类集中的所有单位——此种情况下系指所有住户——进行详查。

47. 对聚类情况下的样本选择可作如下解释。假定使用等概率从村(类集)人口选取一个样本。对于一次单级整群抽样而言, 样本包括了选定村庄的所有住户。

假设:

A = 为村庄总数;

B = 为类集中的住户总数;

a = 为一村庄样本。

因此假定:

$aB = n$ 代表总样本中要素单位(即住户)数量,

而

$AB = N$ 是所有村庄的住户总数。

那么, 使用等概率选择一个要素的概率由下式给出:

$$\frac{a}{A} \times \frac{B}{B} = \frac{n}{N} = f \quad (\text{A.30})$$

式中, N 是要素单位总数, 而 f 是抽样份额。在这种情况下, 选择概率只不过是 $\frac{a}{A}$ 。

A.4.4.3. 样本平均值和方差公式

48. 下面给出样本平均值和方差公式:

样本平均值:

$$\bar{y} = \frac{1}{aB} = \sum_{\alpha=1}^{\alpha} \sum_{\beta=1}^{\beta} \bar{y}_{\alpha\beta} = \frac{1}{a} \sum_{\alpha=1}^{\alpha} \bar{y}_{\alpha} \quad (\text{A.31})$$

样本平均值是总体均值的一个无偏估计数:

$$E(\bar{y}) = \frac{1}{A} \sum_{\alpha=1}^{\alpha} \bar{y}_{\alpha} = \bar{Y} \quad (\text{A.32})$$

事实上, 因为样本容量是固定的 ($aB = n$) 而且是等概率选择, 所以平均值 (\bar{y}) 是总体均值 \bar{Y} 的一个无偏估计数。

方 差

如果各类集是用简单随机选择法选定的, 那么方差可以估计如下:

$$V(\bar{y}) = (1-f)s_{\alpha}^2 \quad (\text{A.33})$$

$$\text{式中, } s_{\alpha}^2 = \frac{1}{a-1} \sum_{\alpha=1}^{\alpha} (\bar{y}_{\alpha} - \bar{y})^2$$

49. 值得注意的是, 这些数值不带抽样误差, 因为它们的基础是 B 中的所有要素, 而不是一个样本。样本平均值方差只归因于类集平均值间的方差。

附件二 专家名单

联合国审查住户抽样调查设计手册草稿专家组会议的与会专家名单，
纽约，2005年12月3日至5日*

与会专家	
姓名	职务和所属关系
Oladejo Oyeleke Ajayi	统计顾问，尼日利亚
Beverly Carlson	拉丁美洲和加勒比经济委员会生产、生产力和管理处，智利圣地亚哥
Samir Farid	统计顾问，埃及
Maphion M. Jambwa	技术顾问，南部非洲发展共同体/欧洲联盟，博茨瓦纳哈博罗内
Udaya Shankar Mishra	副研究员，哈佛大学，美利坚合众国马萨诸塞州波士顿
Jan Kordos	教授，华沙经济学研究生院，波兰华沙
Edwin St. Catherine	局长，国家统计局，圣卢西亚
Anthony Turner	抽样顾问，美利坚合众国
Shyam Upadhyaya	局长，综合统计局(INSTAT)，尼泊尔
Ibrahim Yansaneh	副司长，联合国国际公务员制度委员会生活费用司，美利坚合众国纽约

* 专家组会议的报告见ESA/STAT/AC.93/L.4号文件。

