

社会调查中的样本量管理

——事前样本放大还是事后样本追加？

王 俊 吕 萍*

内容摘要：本文首先介绍样本放大法并梳理部分国内外现有调查项目中的实践经验，然后对事后样本追加法的理论基础及权数计算方法进行讨论，并推导单位有效回答的成本，最后在模拟研究中设置不同的缺失数据模式对两种样本管理方法进行评估。模拟研究结果表明：样本放大法中最终有效回答样本数、总成本受实际回答率和预估回答率之间差异的影响较大。若实际回答率低于预估回答率，将会扩大无应答调整后权数的设计效应、增加总发放样本的样本量，无法达到目标样本量；若实际回答率高于预估回答率，样本放大法将会降低无应答调整后权数的设计效应，但增加了最终样本量，从而增加总访问成本。无论总体实际回答率如何，样本追加法获得的最终回答样本的总样本量维持在目标样本量附近，当无回答的单位成本较低时，总成本受实际应答率的影响相对较小，当总体各层的回答率差异较大时，样本追加法优于样本放大法。

关键词：目标样本量；样本放大；样本追加；回答率；权数

中图分类号：C811 **文献标识码：**A **文章编号：**1004-7794(2025)05-0085-12

DOI: 10.13778/j.cnki.11-3705/c.2025.05.008

一、引 言

大型社会调查中如何对目标样本量进行管理十分重要，直接决定了利用最终有效回答样本估计总体能否达到预定的估计精度（Valliant et al., 2018）。在抽样设计阶段，通常严格根据抽样理论确定目标样本量（金勇进等，2021；王争和杜子芳，2022）。但实际调查执行过程中，会由于拒访、空户、无法联系等原因产生无应答的情况（金勇进和张喆，2014；王晓焘和风笑天，2018），导致访员入户的样本量和最终有效回答样本量之间存在差异，若有效回答样本的样本量远少于目标样本量，将会降低利用其估计总体的精度。因此，调查实际执行过程中需要做好目标样本量管理，尽量保证最终有效回答样本数接近目标样本量。根据对相关调查项目技术文档的梳理、与调查负责方的访谈（国内调查项目），实地目标样本量管理主要分为两种方法。一是事前多发样本，即样本放大法。使用样本放大法时，在抽样设计阶段，

* 王俊，2019年毕业于中国人民大学，获统计学博士学位，现为中国社会科学院人口与劳动经济研究所、中国社会科学院人力资源研究中心，助理研究员，研究方向为抽样调查、应用统计、人口与劳动力统计。吕萍（通讯作者），2009年毕业于中国人民大学，获统计学博士学位，现为北京大学中国社会科学调查中心副研究员，研究方向为抽样调查、复杂数据分析，邮箱：issslvp@pku.edu.cn。本研究得到国家社科基金项目“复杂抽样数据的统计推断方法及其应用研究”（19BTJ012）、中国社会科学院重大经济社会调查项目“中国城市劳动力调查”（GQDC2023022）、中国社会科学院研究所实验室综合资助项目“人口仿真与就业监测实验室”（2024SYZH008）、中国社会科学院青启计划项目“构建社会调查中的实时质量监测体系”（2025QQJH57）、全国统计科学研究项目“噪声惩罚下的多视角聚类研究”（2023LY093）的资助。

使用预估的无应答率的倒数作为放大系数，放大发放给访员入户访问名单的样本量。如预估住户回答率为 50%，目标样本量为 10000 户，则需要发放 20000 户入户名单。然而实际应用样本放大法中存在问题：首先，回答率在城乡、性别等群体内存在差异，使用单一或者简单的无应答放大系数是无效率的；其次，该方法可能会导致对回应率低的子总体的统计推断精度达不到指定要求；此外，高估或者低估实际回答率将会对调查成本产生重要的影响。二是事后样本追加法，简称样本追加法。样本追加法分多批发放入户样本，追加样本的样本量根据前几批样本的回答率进行计算。中国社会科学院社会所组织实施的中国社会综合状况调查（China General Survey，简称 CGS）、中国社会科学院人口与劳动经济研究所组织实施的中国城镇劳动力调查（China Urban Labor Survey，简称 CULS）均采用事后样本追加法（中国城市劳动力调查课题组，2024），其他部分大型社会调查虽采用样本放大法对目标样本量进行管理，但实地访问过程也或多或少使用该种方法管理末端样本量。实践中虽会经常使用事后样本追加的方法管理末端样本量，但如何计算样本单元的入样概率还缺乏理论上的讨论。在抽样调查理论中，李莉莉等（2007）、艾小青和金勇进（2009）在域推断背景下，万舒晨（2020）、谭馨和陈光慧（2024）在多层次推断背景下讨论样本追加方法，本研究将基于样本追加理论对事后样本追加法进行讨论。

二、符号说明

在多阶段抽样中，给定上一阶段抽样结果的条件下各阶段抽样过程相互独立。因此，当抽中末端总体后，末端总体内的住户抽样、一阶包含概率的计算不受前阶段的影响。如在三阶段抽样设计中，一阶段抽取区县，二阶段抽取末端总体村（居）委会，三阶段在村（居）委会内抽家庭户。不失一般性，本文后续均以分层简单随机抽样为例进行讨论，层对应复杂抽样中的末端总体。令总体记为 $U = \{1, 2, \dots, N\}$ ，共分为 H 层，层规模表示为 $N_1, N_2, \dots, N_H$ 。研究变量为 Y ， $Y = \{y_1, y_2, \dots, y_N\}$ ，研究变量的总体均值 $\bar{Y} = \frac{1}{N} \sum_{i=1}^N y_i$ 。样本记为 s ， $s = \{I_1, I_2, \dots, I_N\}$ ， $I_i = 1$ 表示总体 U 中第 i 个单元被选入样本，样本量记为 n ， $n = \sum_{i=1}^N I_i$ 。层内根据简单随机抽样方法抽取样本单元，层内目标样本量为 $D_h (h=1, 2, \dots, H)$ ，则最终样本单元的设计权数表示为：

$$w_{hi} = w_i^1 \times w_{hi}^2 \quad (1)$$

其中，在分层简单随机抽样中 $w_i^1 = 1$ ，如果为多阶段抽样， w_i^1 替换为末端总体抽样前各阶段抽样单元的设计权数的乘积即可； w_{hi}^2 为层 h 中单元 i 的设计权数。

令各层内实际抽中发放的样本量表示为 n_h 。实际调查组织实施过程中，由于空户、拒访等原因的存在，最终发放给访问员的样本量往往不低于层内目标样本量，即 $n_h \geq D_h$ 。式（1）中的 w_{hi}^2 的计算方法根据发放样本的方式的不同而不同。

三、事前样本放大法

（一）事前样本放大法

事前样本放大法（Sample Inflation）根据预估的总体或者子总体的回答率在目标样本量的基础上放大发放的样本量，预估的回答率通过前期的（预）调查经验确定。为方便讨论，假定抽样总体可分为 C 类，类 c 的预估回答率为 $\hat{r}_c$ ，层 h 内该子总体样本量为 D_{hc} ，则该子总体实际需要

访问的样本量为 $\left[D_{hc} / \widehat{rp}_c \right]$ ，实际访问的总样本量 $n_h = \sum_c \left[D_{hc} / \widehat{rp}_c \right]$ 。此时，样本单元的一阶包含概率 π_{hi}^{pr} 表示为：

$$\pi_{hi}^{pr} = I(i \in U_c) \left[\frac{D_{hc}}{\widehat{rp}_c} \right] / N_{hc} \tag{2}$$

其中， $I(i \in U_c) / \widehat{rp}_c$ 也称为类 c 的样本放大系数。
在分层简单随机抽样的抽样设计时，总体均值 $\bar{Y}$ 的分层估计量表示为：

$$\hat{\bar{Y}}^{pr} = \frac{1}{N} \sum_h \sum_{i \in U_h} w_{hi}^{pr} y_{ji} \tag{3}$$

其中， $w_{hi}^{pr} = 1 / \pi_{hi}^{pr}$ ，为层 h 样本单元 i 的权数。

根据公开调查项目的技术报告，中国综合社会调查（Chinese General Social Survey，简称 CGSS）根据城乡对村委会、居委会的实际访问样本量进行放大，城镇按预估回答率为 65% 调整村居样本量、农村按 85% 的预估回答率调整村居样本量。2011 年中国家庭追踪调查（China Family Panel Survey，简称 CFPS）的抽样设计中，项目组根据前期经验将全国区（县）分为低应答率地区、一般应答率地区和高应答率地区三类：低应答率地区居委会和其他村委会的回答率估计为 60% 和 70%；一般应答率地区居委会和其他村委会的回答率估计为 70% 和 80%；高应答率地区居委会和村委会的回答率估计为 80% 和 90%（谢宇等，2012）。事前样本放大法中如何准确估计总体（子总体）的回答率十分重要，若高估回答率将会导致最终样本量无法满足调查项目的推断精度，若低估回答率将会增加整个调查的访问成本。受社会经济发展水平等因素的影响，不同子总体、区域、社区层面回答率均存在一定的差异，若预估回答率分层过粗，一方面会导致无应答调整是无效率的，另一方面可能导致对回答率较低的子总体的统计推断无法满足精度要求。中国健康与养老调查（China Health and Retirement Longitudinal Survey，简称 CHARLS）在基线抽样时，首先为每个抽中的村（居）发放 40 户样本户，访员根据这 40 户地址联系受访者，但不访问。然后使用这 40 户预估该村（居）的无应答率，并根据预估的无应答率放大正式入户访问样本的样本量（赵耀辉等，2013）。CHARL 基线调查中的这种方法虽然能够考虑各个村（居）应答率的异质性，但增加了调查访问的成本。澳大利亚住户、收入和劳动力动态（Household Income and Labour Dynamics in Austrilia, HILDA）也是通过事前样本放大法对目标样本量进行管理（Watson & Wooden，2002）。

表 1 部分大型调查中末端样本量管理方法

调查项目	抽样组织单位	末端样本量管理方式
中国家庭追踪调查（CFPS）	北京大学中国社会科学调查中心	样本放大法：按城乡将村（居）分类预估回答率
中国养老追踪调查（CHARLS）	北京大学中国社会科学调查中心	样本放大法：预估计各村（居）无应答率
中国综合社会调查（CGSS）	中国人民大学中国调查与数据中心	样本放大法：按城乡分类预估行无应答调整
中国社会状况调查综合（CGS）	中国社会科学院社会学研究所	样本追加法
中国城市劳动力调查（CULS）	中国社会科学院人口与劳动经济研究所	样本追加法
澳大利亚住户、收入和劳动力动态（HILDA）	墨尔本大学（University of Merbourne）	样本放大法

（二）回答率的一个案例

本部分以中国社会科学院人口与劳动经济研究所组织实施的第五轮中国城市劳动力调查为

例，展示回答率在不同城市、区县、社区层面的差异。CULS 的涵盖范围包括 8 个城市，分别为上海、成都、广州、武汉、西安、沈阳、福州、贵阳。每个城市的抽样设计均为三阶段 PPS 抽样，独立组织实施，各城市的样本对该城市具有代表性。从上海市、广州市、成都市、沈阳市、西安市和贵阳市 6 个城市的实地访问情况来看^①，成都市的整体回答率较低，仅 68% 的样本单元完成了问卷访问，最高为沈阳市完访占比为 91%，其他城市的回答率在 75%~90% 之间，城市间的回答率存在一定差异。分城区统计回答率来看，西安市、沈阳市各区的回答率差异较小，回答率上四分位数和下四分位数差分别为 4% 和 6%，而成都市、广州市、上海市、贵阳市差异较大，该差值分别为 21%、19%、13%、13%。按社区统计回答率来看，除成都市外，其他城市各社区回答率的上四分位数均在 80% 以上，但也存在一些回答率较低的社区，如上海市社区层面回答率最低为 26%、贵阳市最低为 28%。从访员身份来看，社区工作人员上门入户的城市回答率相对较高，而学生上门入户的城市回答率相对较低。

表 2 第五轮中国城市与劳动力调查按城市分回答率

城市	访员身份	城市分回答率	社区层面回答率				区层面回答率			
			Min	Q25	Q50	Q75	Min	Q25	Q50	Q75
上海市	社区人员	0.85	0.26	0.83	0.92	1.00	0.66	0.80	0.83	0.93
广州市	学生	0.78	0.30	0.72	0.86	1.00	0.68	0.73	0.85	0.92
成都市	学生	0.68	0.30	0.55	0.75	0.96	0.51	0.64	0.74	0.85
沈阳市	社区人员	0.91	0.49	0.89	1.00	1.00	0.72	0.89	0.93	0.95
西安市	学生	0.88	0.53	0.83	0.98	1.00	0.83	0.85	0.88	0.89
贵阳市	学生	0.78	0.28	0.74	0.81	0.93	0.69	0.72	0.81	0.85

然而，中国城市劳动力调查的目标对象为城市城区的常住人口，一方面这些地方的回答率可能与其他城市、非城市、非城区社区的回答率存在差异，另一方面，不同机构的调查组织方式、培训方式、访员的素质等因素也会影响最终各层级汇总的回答率，需要汇总多方调查机构的实地执行并行数据进行经验上的分析。

四、事后样本追加

（一）事后样本追加法

事后追加方法先发放一个小的样本让访员入户调查，然后基于第一批访问样本的回答率估算所在末端总体的回答率，并根据估算的回答率确定追加的样本量。如果第一批和第二批样本的总有效回答数依然达不到目标样本量，可以基于前两批样本继续追加第三批样本，追加的样本需要全部发放给访员按规则入户，不然将会破坏样本的随机性。然而，若追加样本次数较多，一方面不利于实地访问进程管理，另一方面也会让样本单元权数的计算过程变得复杂，实践中建议只追加一次样本。本文基于一次样本追加方法讨论权数的计算过程。

令层 h 中第一批样本记为 s_h^1 ，目标样本量为 D_h ，层内总规模为 N_h 。由于各层第一批样本的回答率未知，可以将第一批样本的样本量设置为目标样本量 D_h ，避免错误估计无应答情况导致层内样本多发的情况。此时，层 h 中单元 i 被选入第一批样本的包含概率为 $\pi_{hi}^1 = P(i \in s_h^1) =$

^① 由于武汉市和福州市的调查组织管理模式和其他城市存在一定差异，本文仅以上海市、广州市、成都市、沈阳市、西安市和贵阳市的数据为例。

D_h / N_h 。假定第一批样本访问结束后，拒访的户数为 R_h ($R_h > 0$)，则该层总体需要发放第二批样本 s_h^2 ，第二批样本的样本量为：

$$\left\lceil \frac{R_h}{resp_h^1} \right\rceil \quad (4)$$

其中， $resp_h^1 = (D_h - R_h) / D_h$ 为末端总体 h 第一批样本的应答率； $(D_h - R_h)$ 为第一批样本的有效回答数。最终样本记为 $s_h = s_h^1 \cup s_h^2$ 。

(二) 一阶包含概率与权数

追加样本后层 h 中单元 i 被选入样本 s_h 的概率 π_{hi} 表示为：

$$\pi_{hi} = P(i \in s_h) = P(i \in s_h^1 \cup s_h^2) = P(i \in s_h^1) + P(i \in s_h^2) \quad (5)$$

其中， $P(i \in s_h^1)$ 、 $P(i \in s_h^2)$ 分别表示层 h 中任一单元 i 被选入第一批和第二批样本的概率。

由全概率公式可得：

$$P(i \in s_h^2) = P(i \in s_h^2 | i \in s_h^1) P(i \in s_h^1) + P(i \in s_h^2 | i \notin s_h^1) P(i \notin s_h^1) \quad (6)$$

在无放回抽样的条件下，如果单元 i 被抽中至第一批样本，则其被抽中至第二批样本的概率为零。因此， $P(i \in s_h^2 | i \in s_h^1) = 0$ ，则 π_{hi} 可以写为：

$$\pi_{hi} = P(i \in s_h^1) + P(i \in s_h^2 | i \notin s_h^1) P(i \notin s_h^1) \quad (7)$$

根据全概率公式，式 (7) 中 $P(i \in s_h^2 | i \notin s_h^1)$ 可表示为：

$$P(i \in s_h^2 | i \notin s_h^1) = \sum_{R_h} P(i \in s_h^2 | i \notin s_h^1, \sum_{i \in s_h^1} I_{hi} = D_h - R_h) \times P(\sum_{i \in s_h^1} I_{hi} = D_h - R_h) \quad (8)$$

其中， $P(i \in s_h^2 | i \notin s_h^1, \sum_{i \in s_h^1} I_{hi} = D_h - R_h)$ 表示给定回答数为 $D_h - R_h$ 、且单元 i 未被选入第一批样本的条件下，层 h 余下抽样框中单元 i 被选入第二批样本中的概率。

式 (8) 中 $P(i \in s_h^2 | i \notin s_h^1, \sum_{i \in s_h^1} I_{hi} = D_h - R_h)$ 写为：

$$P(i \in s_h^2 | i \notin s_h^1, \sum_{i \in s_h^1} I_{hi} = D_h - R_h) = \frac{R_h / resp_h^1}{N_h - D_h} = \frac{R_h}{N_h - D_h} \times \frac{D_h}{D_h - R_h} \quad (9)$$

则对层 h 中任一单元 i ，未被选入第一批样本、且第一批样本回答数为 $D_h - R_h$ 的条件下，被选入第二批样本的概率 $P(i \in s_h^2 | i \notin s_h^1)$ 为：

$$\begin{aligned} P(i \in s_h^2 | i \notin s_h^1) &= \sum_{R_h} P(i \in s_h^2 | i \notin s_h^1, \sum_{i \in s_h^1} I_{hi} = D_h - R_h) \times P(\sum_{i \in s_h^1} I_{hi} = D_h - R_h) \\ &= \sum_{R_h} \frac{R_h}{N_h - D_h} \times \frac{D_h}{D_h - R_h} \times P(\sum_{i \in s_h^1} I_{hi} = D_h - R_h) = \frac{D_h}{N_h - D_h} \times E\left(\frac{R_h}{D_h - R_h}\right) \end{aligned} \quad (10)$$

其中，当层 h 期望应答率为 1 时，则无无应答情况， $P(i \in s_h^2 | i \notin s_h^1) = 0$ 。

由于 $P(i \in s_h^1) = \pi_{hi}$ ，可知 $P(i \notin s_h^1) = 1 - \pi_{hi}$ 。则给定层 j ，任一单元被选至样本 s_h 的概率为：

$$\pi_{hi} = \pi_{hi}^1 + \frac{D_h}{N_h - D_h} \times E\left(\frac{R_h}{D_h - R_h}\right) \times (1 - \pi_{hi}^1) \quad (11)$$

由于单元 i 被选入第一批样本的概率 $\pi_{hi}^1 = \frac{D_h}{N_h}$ ，将其带入式 (11)：

$$\begin{aligned}\pi_{hi} &= \pi_{hi}^1 + \frac{D_h}{N_h - D_h} \times E\left(\frac{R_h - D_h + D_h}{D_h - R_h}\right) \times (1 - \pi_{hi}^1) = \frac{D_h}{N_h} + \frac{D_h}{N_h - D_h} \left(-1 + E\left(\frac{D_h}{D_h - R_h}\right)\right) \times \left(1 - \frac{D_h}{N_h}\right) \\ &= \frac{D_h}{N_h} + \frac{D_h}{N_h} \left(-1 + E\left(\frac{D_h}{D_h - R_h}\right)\right) = \frac{D_h}{N_h} E\left(\frac{D_h}{D_h - R_h}\right) = \frac{D_h / E(1 / \text{resp}_h)}{N_h}\end{aligned}\quad (12)$$

根据抽样调查理论, 若层 h 总体规模 $N_h \rightarrow +\infty$, $D_h \rightarrow +\infty$ 时, $E(y^2) < +\infty$, 且 $0 < \pi_{hi}^1 \leq 1$ 时, 各层实际拒访数 $R_h \xrightarrow{D} E(R_h)$ 。易证 $\frac{R_h}{D_h - R_h} \xrightarrow{D} E\left(\frac{R_h}{D_h - R_h}\right)$ 。若各层实际拒访数替换式 (12) 中 R_h , 则

$$\begin{aligned}\pi_{hi}^{es} &= \frac{N_h D_h \frac{R_h}{D_h - R_h} + D_h (N_h - D_h) - D_h D_h \frac{R_h}{D_h - R_h}}{N_h (N_h - D_h)} \\ &= \frac{N_h D_h R_h + D_h (N_h D_h - N_h R_h - D_h D_h + D_h R_h) - D_h D_h R_h}{N_h (N_h - D_h) (D_h - R_h)} = \frac{D_h}{N_h} \times \frac{D_h}{D_h - R_h}\end{aligned}$$

易证 $\pi_{hi} = \pi_{hi}^{es} + o(1/n_h^{1/2})$ 。在分层一阶段简单随机抽样的抽样设计时, 总体均值 $\bar{Y}$ 的分层估计量表示为:

$$\hat{\bar{Y}} = \frac{1}{N} \sum_h \frac{1}{n_h} \sum_{i \in U_h} w_{hi} y_{ji} \quad (13)$$

其中, $w_{hi} = 1/\pi_{hi}^{es}$, 为层 h 样本单元 i 的权数。易证 $\hat{\bar{Y}} \xrightarrow{D} \bar{Y}$ 。

(三) 事后追加中的常见问题

1. 第一批样本的样本量多少合适。

由于无法事前确定总体或子总体的回答率, 如果第一批发放的样本量过多, 可能会增加调查执行的成本。若第一批发放的样本量过少, 将无法有效估计层内回答率。第五轮中国城市劳动力调查中, 抽中居委会的目标样本量为 24 (或 25) 户, 第一批各居委会发放的样本量为其目标样本量。居委会完成第一批样本访问后, 抽样组根据该居委会第一批样本的回答率确定追加样本的数量和名单, 并发放第二批样本。第二批样本名单需要全部按照入户规则上门访问。中国健康与养老调查的基线调查中, 虽然采用的是样本放大法进行目标样本量管理, 但该调查首先为各抽中村 (居) 委会发放 40 户住户名单, 调查员对这 40 户进行排摸但不参加问卷调查, 然后基于这 40 户计算各村 (居) 委会的回答率, 在村 (居) 委会层面放大实际发放的样本量。

从实地访问样本管理的便捷性来看, 将第一批样本的样本量设置为目标样本量 D_h 具有一定的优势: 当层 h 不存在无应答问题时, 则无需追加样本; 如存在无应答情况, 则根据式 (4) 确定层内需要追加的样本, 并从余下的末端抽样框中随机抽取追加样本。

2. 追加样本量多于末端抽样总体的规模。

实践中可能会存在末端总体 h 第一批样本的回答率过低的情况, 此时根据式 (4) 确定的追加的样本量较大或多于末端总体剩余单元规模的情况。若让访员对这些追加的样本名单全部入户访问, 不仅会增加访员的工作量, 还会增加调查的成本。实践中, 可以将追加的样本量设置上限, 以避免部分末端总体追加样本量过大的情况。第五轮中国城市劳动力调查中, 根据前四轮的经验, 将居委会追加样本量的上限定为 20 户, 此时需要各层单元的入样概率进行修改。假定追加的样本量的上限为 D_h^{max} , 则层 h 内单元的入样概率表示为:

$$\pi_{hi} = \frac{1}{N_h} \times \max \left(D_h^{max}, D_h \frac{D_h}{D_h - R_h} \right) \quad (14)$$

五、成本分析

样本放大法和样本追加法两种方法除了在末端抽样的组织方式、对调查组织效率的要求不一致外，更重要的是体现在费用成本上。调查实践中，通常同一个末端抽样总体内的单位完成访问成本、单位未完成访问的成本是固定的。令假定末端总体确定后，与交通等相关的总费用已确定，为 C^{trans} 。末端抽样总体记为 $h(h=1,2,\dots,H)$ ，则总调查费用表示为：

$$C_T = C^{trans} + \sum_h \left(c_h^r n_h^r + c_h^{nr} (n_h - n_h^r) \right) \quad (15)$$

其中， c_h^r 为层 h 内接受访问单元的单位成本， c_h^{nr} 为层 h 内无应答单元的单位成本； n_h^r 为层 h 内回答的样本数。

C^{trans} 不受末端总体入户人数的限制，因此在分析中，将此项设置为 0。则有效回答样本单位成本 ($\bar{C}_T$) 定义为总费用除回答样本数 ($n_r = \sum_h n_h^r$)：

$$\begin{aligned} \bar{C}_T &= \frac{1}{\sum_h n_h^r} \sum_h \left(c_h^r n_h^r + c_h^{nr} (n_h - n_h^r) \right) = \sum_h \frac{n_h^r}{\sum_h n_h^r} \left(c_h^r + c_h^{nr} \left(\frac{n_h}{n_h^r} - 1 \right) \right) \\ &= \bar{c}_r + \frac{\sum_h (n_h - n_h^r)}{\sum_h n_h^r} \sum_h \frac{(n_h - n_h^r)}{\sum_h (n_h - n_h^r)} c_h^{nr} = \bar{c}_r + \frac{\sum_h (n_h - n_h^r)}{\sum_h n_h^r} \bar{c}_{nr} = \bar{c}_r + \left(\frac{n}{n^r} - 1 \right) \bar{c}_{nr} \end{aligned} \quad (16)$$

其中， $\bar{c}_r$ 为层回答样本数加权的平均单位回答样本的费用， $\bar{c}_r = \sum_h \frac{n_h^r}{\sum_h n_h^r} c_h^r$ ； $\bar{c}_{nr}$ 为层无应答数加权的平均无应答样本的费用， $\bar{c}_{nr} = \sum_h \frac{(n_h - n_h^r)}{\sum_h (n_h - n_h^r)} c_h^{nr}$ ； n 为实际发放样本的样本量， $n = \sum_h n_h$ 。

从式 (16) 可见，最终有效回答样本的单位成本仅取决于层回答数加权的平均回答费用、无应答样本的层无应答数加权平均费用，以及总体实际有效回答率。当单位回答成本费用和单位无应答成本费用为固定额度时，如每成功访问一位受访者给予访问员报酬 100 元、未完成访问的受访者则给予访问员 10 元，则最终单位有效回答成本仅取决于总体实际回答率。实际回答率越低，则单位有效回答成本越大，反之单位有效成本越低。为考察单位成本的变化，设置各层单位回答成本一样，即 $c_h^r = 10$ ， c_h^{nr} 分别取值为 6、4、2、1 四种情形。由表 3 可见，当总体

表 3 完全随机情形下：期望单位成本

	$c_h^r = 10, c_h^{nr} = 6$	$c_h^r = 10, c_h^{nr} = 4$	$c_h^r = 10, c_h^{nr} = 2$	$c_h^r = 10, c_h^{nr} = 1$
0.5	16.00	14.00	12.00	11.00
0.6	14.00	12.67	11.33	10.67
0.7	12.57	11.71	10.86	10.43
0.8	11.50	11.00	10.50	10.25
0.9	10.67	10.44	10.22	10.11

真实无应答率较低时，单位有效回答的成本受单位无应答费用的影响较大，当真实无应答率较高时，单位有效回答的成本受单位无应答费用的影响较小。因此，在实际调查执行过程中，当给定单位回答费用、单位无应答费用、总体真实回答率后，最终有效回答率直接决定了调查实地访问的总成本。

六、模拟研究

本部分通过设置不同情形无应答模式进行模拟研究，从实际发放样本量、最终有效完成样本量、权数设计效应、总成本四个方面，对事前样本放大法和事后样本追加法的实际效果进行比较与评估。模拟总体规模为 1000000，分为 200 层，每层规模为 $N_h (h=1,2,\dots,H)$ 。辅助变量 $X = \{X_1, X_2, \dots, X_{10}\}$ ，由多元正态分布 $N(0,1)$ 生成，研究变量 $Y = \beta_0 + \beta X + a_h I_h$ 生成， β 由 $N(1,1)$ 生成， a_h 由 $U(-1,1)$ 生成，为减少常数项对估计结果的影响，本研究模拟研究中令 $\beta_0 = 0$ 。

（一）模拟研究步骤

1. 抽样设计。

由于样本放大法和样本追加法的样本发放过程存在差异，本文模拟研究中将分别对样本放大法和样本追加法进行模拟研究，两种方法的基本模拟研究设计如下。

样本放大法模拟研究设计：目标样本量为每层 40 个基本抽样单元，合计目标样本量为 8000。为了便于分析，本文模拟研究中统一设置样本放大法的预估回答率为 0.7，则每层实际发放的样本量为 57，总发放样本量为 11400。

样本追加法模拟研究设计：为和样本放大法比较，样本追加法模拟研究中目标样本量同设置为每层 40 个基本抽样单元，合计目标样本量为 8000。模拟研究过程中，首先从各层抽取 40 个基本抽样单元作为首批样本，然后根据预设的回答机制生成有效回答样本，再根据各层有效回答样本估计各层回答率并根据式（4）计算追加的样本量。

2. 模拟研究步骤。

为评估事前样本放大法和事后样本追加法两种方法，本文设置了完全随机无应答、分层完全随机无应答两种情形分别进行模拟研究。为了对两种样本管理方法进行评价，本研究使用实际发放样本量、有效回答样本量、设计效应、总成本四个指标进行衡量。设计效应由式（17）计算，总样本量为各种方法下实际发放的样本量，总成本由式（15）确定。

设计效应定义为：

$$\text{设计效应} = \frac{n \sum_h \sum_i w_{hi}^2}{\left(\sum_h \sum_i w_{hi} \right)^2} \quad (17)$$

具体模拟研究步骤如下。

步骤一：根据模拟研究设计（样本放大法或样本追加法模拟研究设计）从总体中获取有效回答样本。

步骤二：计算各层样本的设计权数，并根据各层实际有效回答率对设计权数进行无应答调整得到无应答调整后权数。

步骤三：根据无应答调整后权数得到总体均值的加权估计，并计算无应答调整后权数的设计效应、计算实际发放的样本量、有效回答样本量、各成本配置条件下的总成本。

步骤四：重复上述过程 500 次，并计算无应答调整后样本权数的设计效应、平均计算实际发放的样本量、平均最终有效回答样本量、平均总成本。

(二) 结果分析

1. 完全随机无应答。

在完全随机无应答模拟情形中，假定总体中所有单元回答和不回答的概率都是一样的。为考察高估、低估总体回答率对最终样本管理效果的影响，本部分模拟研究分别设置有效回答率为 0.5、0.6、0.7、0.8、0.9 五种情形。

从最终有效回答样本数上来看，样本追加法表现整体优于样本放大法，获得的有效样本量始终接近目标样本量 8000，受真实有效回答率的影响较小，有效样本量与总发放的样本量的比值接近真实回答率，即真实回答率越低需要发放的总样本量越大。给定目标样本量、预估回答率条件下，样本放大法最终获得的有效样本量取决于总体真实回答率。当预估回答率高于总体回答率时，最终有效样本量低于目标样本量，预估回答率比真实回答率越高，最终有效样本量比目标样本量越低，导致最终样本的样本量无法满足指定的估计精度。当预估回答率低于实际回答率时，最终有效样本量将高于目标样本量，预估回答率比实际回答率越低，实际有效回答样本量比目标样本量越高，从而导致调查总成本的增加。从无应答调整后权数的设计效应上看，当层回答率不存在差异时，样本放大法的设计效应一致低于样本追加法，主要原因在于样本追加法中各层第一批样本回答率可能会存在一定差异，从而增加了权数的波动。随着实际应答率的增加，样本放大法和样本追加法的权数设计效应均有下降的趋势。

表 4 完全随机情形下：总样本量和权数设计效应

实际回答率	样本放大法（预估回答率 0.7）			样本追加法（两次）		
真实回答率	发放样本量	有效完成样本量	权数设计效应	发放样本量	有效完成样本量	权数设计效应
0.5	11400	5698	1.02	16345	8188	1.09
0.6	11400	6269	1.01	13490	8096	1.06
0.7	11400	6839	1.01	11503	8050	1.04
0.8	11400	7408	1.01	10005	8000	1.03
0.9	11400	7981	1.00	8848	7966	1.01

表 5 完全随机情形下：总成本

方法	真实回答率 成本	0.5	0.6	0.7	0.8	0.9
样本放大法（预估回答率 0.7）	$c_h^r = 10, c_h^{nr} = 6$	91192.95	95756.77	100328.93	104869.89	109421.57
	$c_h^r = 10, c_h^{nr} = 4$	79788.63	86634.51	93489.28	100306.59	107134.38
	$c_h^r = 10, c_h^{nr} = 2$	68384.32	77512.26	86649.64	95743.30	104847.19
	$c_h^r = 10, c_h^{nr} = 1$	62682.16	72951.13	83229.82	93461.65	103703.59
样本追加法	$c_h^r = 10, c_h^{nr} = 6$	130974.45	113476.65	101291.27	92309.70	85344.92
	$c_h^r = 10, c_h^{nr} = 4$	114590.60	102640.10	94343.78	88203.20	83448.15
	$c_h^r = 10, c_h^{nr} = 2$	98206.75	91803.55	87396.29	84096.70	81551.37
	$c_h^r = 10, c_h^{nr} = 1$	90014.83	86385.27	83922.54	82043.45	80602.99

从总成本上看，在给定抽样设计的前提下，样本放大法的总成本随着总体回答率的增加而增加，随着单位回答成本和单位无回答成本比的增加而降低，波动较大，主要取决于预估回答率和总体真实回答率的差异。当预估回答率高于实际回答率时，虽然总成本较低，但其实际发放样本量和有效回答样本量均低于目标样本量。当预估回答率低于实际回答率时，导致实际发放样本量、

有效回答样本量均高于目标样本量,从而增加总成本。在样本追加法中,由于总样本量相对稳定,总成本随着总体回答率的上升而降低,随着单位回答成本和单位无回答成本比的增加而减少。当预估回答率和真实回答率相同时,样本放大法和样本追加法的总成本相差不大。若预估回答率低估真实总体回答率,样本追加法的总成本相对更低。

综合来看,当总体真实回答率在各层不存在差异时,样本放大法的设计效应一致低于样本追加法。然而,样本放大法的最终实际有效样本量、总成本受预估回答率的影响较大。当无法准确预估总体回答率时,样本追加法的表现相对稳定,但会在一定程度上增加估计量的标准差。

2. 社区层面完全随机无应答。

为考察不同层回答率存在差异的情形,本部分模拟实验中为各层设置差异化回答率,层内所有单元回答不回答的概率都是一样的。为此设置各层回答率分别由均匀分布 $U(0.3, 0.9)$ 、 $U(0.4, 0.9)$ 、 $U(0.5, 0.9)$ 、 $U(0.6, 0.9)$ 、 $U(0.7, 0.9)$ 5 种情形生成,对应 5 种情形的总体期望回答率分别为 0.6、0.65、0.7、0.75、0.8。由于前文设置预估的总体回答率为 0.7,则 5 种情况对应低估、准确估计、高估总体回答率的情况。

表 6 社区层面完全随机无应答:总样本量和权数设计效应

实际回答率	样本放大法 (预估回答率 0.7)			样本追加法 (两次)		
层回答率范围	发放样本量	有效完成样本量	权数设计效应	发放样本量	有效完成样本量	权数设计效应
$U(0.3, 0.9)$	11400	6757	1.39	14501	8934	1.42
$U(0.4, 0.9)$	11400	6760	1.39	13075	8629	1.31
$U(0.5, 0.9)$	11400	6727	1.08	11996	8306	1.19
$U(0.6, 0.9)$	11400	7291	1.06	10855	8132	1.10
$U(0.7, 0.9)$	11400	7878	1.03	10059	8037	1.05

表 7 层回答率存在差异情形下:总成本

方法	真实回答率	$U(0.3, 0.9)$	$U(0.4, 0.9)$	$U(0.5, 0.9)$	$U(0.6, 0.9)$	$U(0.7, 0.9)$
	成本					
样本放大法 (预估回答率 0.7)	$c_h^r = 10, c_h^{nr} = 6$	74911.51	77168.14	79513.54	82423.14	84538.47
	$c_h^r = 10, c_h^{nr} = 4$	72365.07	75750.32	79268.84	83633.73	86806.86
	$c_h^r = 10, c_h^{nr} = 2$	69818.64	74332.51	79024.15	84844.31	89075.26
	$c_h^r = 10, c_h^{nr} = 1$	68545.42	73623.61	78901.80	85449.61	90209.45
样本追加法	$c_h^r = 10, c_h^{nr} = 6$	96358.47	92676.39	88062.40	85070.90	83158.12
	$c_h^r = 10, c_h^{nr} = 4$	94018.75	90547.68	86395.68	83818.93	82229.08
	$c_h^r = 10, c_h^{nr} = 2$	91679.02	88418.96	84728.97	82566.97	81300.04
	$c_h^r = 10, c_h^{nr} = 1$	90509.16	87354.61	83895.61	81940.98	80835.52

从结果上看,使用样本放大法时,社区层面完全随机无应答情形和总体完全随机无应答情形的结果类似,当预估的总体回答率高于实际回答率时,实际回答率越低则获得的有效样本量越少,无应答调整后权数的设计效应越大,获取有效回答的总成本越低;反之,当预估总体回答率低于实际回答率时,实际回答率越高则获得的最终样本量越多,无应答调整后权数的设计效应越小,获取有效样本的总成本越高。使用样本追加法时,最终样本的样本量始终在 8000 左右,接近目标样本量。当预估回答率高于实际回答率时,样本追加法的权数效应整体而言低于样本放大法,

当预估回答率低于实际回答率时,样本追加法无应答调整后权数的设计效应虽高于样本放大法,但整体而言比较接近。总成本上看,和完全随机情况下类似,当预估回答率等于或低于真实总体期望回答率时,样本追加法的总成本的优势更为明显,这种优势随着有效回答单位成本和无回答单位成本比增加而增加。因此,在样本追加法中,适当降低无应答单位成本将有利于进一步节省总成本。

七、结论与建议

当前我国大型社会调查有了一定的发展,但实践中的样本管理缺乏规范。一些调查虽然使用样本放大法管理目标样本量,但实际调查执行过程中,会根据末端样本回答情况追加样本,但没有科学的方法计算样本单元的包含概率和权数。本文梳理了部分国内外调查执行过程中样本管理的方法,基于分层抽样推导出了样本追加法中最终样本单元的入样概率和权数计算方法、分析了最终单位有效回答的成本,并通过模拟研究从总发放样本量、最终回答样本量、权数设计效应、总成本四个方面对事前样本放大法和事后样本追加法进行评估。成本分析表明,给定单位有效回答费用、单位无回答费用的前提下,最终有效回答样本的单位成本仅取决于总体实际回答率,实际总体回答率越低,则有效回答样本的单位成本越高,反之则越低。模拟研究结果表明:样本放大法中最终样本量受实际回答率和预估回答率之间的差异影响较大,而样本追加法表现相对稳定;当末端总体实际回答率不存在差异时,样本放大法的权数设计效应一致低于样本追加法。若末端总体真实回答率存在差异,则当预估回答率高估真实回答率时,样本追加法的权效应一致小于样本放大法,反之样本追加法的权效应高于样本放大法;从总成本上看,样本追加法由于根据第一批回答率调整后续发放的样本数,从而保证了最终样本的样本量始终接近目标样本量,在给定总体真实回答率的前提下,总成本更为稳定。样本追加法在调查执行过程中,根据前批样本的回答率动态调整所需发放的样本量,最大程度上降低了事前调整方法低估或者高估回答率带来的影响,从而保证最终有效回答样本的样本量始终在目标样本量附近,受实际回答率的影响相对较小,但当层回答率差异过大,且有较多层真实回答率较低的情况下,最终有效回答样本的样本量会略高于目标样本量。实践过程中,如果不能准确估计总体或子总体回答率的条件下,使用样本追加法是一种相对稳妥的样本量管理方法。

相较于发达国家,我国社会调查起步相对较晚,抽样理论研究和调查实践过程缺乏有机的统一,导致调查实地执行过程中遇到的很多问题的解决方式主观性较多,缺乏客观、科学的评估框架,本研究以社会调查中的目标样本量管理为例,对调查实践中可能遇到的问题进行科学地评估,为今后社会调查实践中的样本管理提供理论依据。

参考文献

- [1] 艾小青,金勇进.域内不放回样本追加的性质分析[J].统计与信息论坛,2009,24(4):7-11.
- [2] 金勇进,杜子芳,蒋妍.抽样技术[M].北京:中国人民大学出版社,2021:20-59.
- [3] 金勇进,张喆.抽样调查中的权数问题研究[J].统计研究,2014,31(9):79-84.
- [4] 李莉莉,冯士雍,秦怀振.不放回样本追加策略下域的估计[J].统计研究,2007(6):80-85.
- [5] 谭馨,陈光慧.不放回样本追加下基于永久随机数抽样技术的多层次推断方法及应用研究[J].统计研究,2024,41(8):139-149.
- [6] 万舒晨.小微企业抽样调查的样本追加研究[J].调研世界,2020(4):49-55.
- [7] 王晓焘,风笑天.社会调查中的应答率与无应答偏差[J].南京政治学院学报,2018,34(2):35-40.
- [8] 王争,杜子芳.多控制指标的劳动力调查初始样本量确定[J].调研世界,2022(8):33-42.

- [9] 谢宇, 邱泽琦, 吕萍. 中国家庭动态跟踪调查技术报告系列: CFPS-1. [OL]. <https://www.isss.pku.edu.cn/cfps/docs/20200520161539050175.pdf>.
- [10] 赵耀辉, John S, 杨功焕等. 中国健康与养老追踪调查-全国基线调查用户手册[OL]. <http://charls.pku.edu.cn/wenjian/jixiandiaochashujuyonghushiyongshouce2013-04-07.pdf>.
- [11] 中国城市劳动力调查课题组. 理解中国城市劳动力市场——中国城市劳动力调查(第五轮)报告[J]. 劳动经济研究, 2024, 12(1): 3-27.
- [12] VALLIANT R, DEVER J A, KREUTER F. Practical tools for designing and weighting survey samples[M]. Cham: Springer International Publishing, 2018: 163-184.
- [13] WATSON N, WOODEN M. The household, income and labour dynamics in Australia (HILDA) survey: Wave 1 survey methodology[OL]. <https://melbourneinstitute.unimelb.edu.au/assets/documents/hilda-bibliography/hilda-technical-papers/htec102.pdf>

Sample Size Management in Social Surveys

—Pre-survey Sample Inflation or Post-survey Sample Addition?

Wang Jun^{1,2} Lv Ping³

(1. Institute of Population and Labor Economics, Chinese Academy of Social Sciences;

2. Human Resources Research Center, Chinese Academy of Social Sciences;

3. Institute of Social Science Survey, Peking University)

Abstract: This article first reviews the sample inflation method and the practical experiences of some existing survey projects, both domestically and internationally, and then discusses the theoretical foundation and weight calculation methods of the post-survey sample addition method. It also derives the cost per valid response. Finally, the article evaluates the two sample management methods using different missing data patterns in a simulation study. The simulation results show that if the actual response rate is lower than the estimated response rate, the sample inflation method will increase the design effect in the non-response adjusted weights and increase the total number of distributed samples, falling short of the target sample size. If the actual response rate exceeds the estimated response rate, the sample inflation method will reduce the design effect in the non-response adjusted weights but increase the final sample size, thus increasing the total survey cost. Regardless of the overall actual response rate, the sample addition method consistently keeps the total number of responding samples close to the target sample size. When the cost per non-response is low, the total cost is less affected by actual response rate. If there are significant differences in response rates across strata, the sample addition method demonstrates better performance than the sample inflation method.

Key words: Target Sample Size; Sample Inflation; Sample Addition; Weights

(责任编辑: 黄 煌)