含有敏感因素二重抽样调查方法的改进

■何华芹　刘建平

一、问题的提出

现代市场调查中，有许许多多含有敏感因素的调查，如某减肥药品生产厂家，为了解该药品的效果，需要调查人的体重区间；又如某电器生产厂家，为了安排不同档次电器的生产，需要了解消费者的收入水平，这些调查，对一部分人来说，涉及到了他们的忌讳，是不愿意接受的，但是取得这些数据对企业来说，又是至关重要的。如何才能得到这些数据呢？显然，传统的处理敏感性问题的方法——随机化回答技术，在这里并不适用。因为无论是沃纳模型，西蒙斯模型，还是更为复杂的随机化回答模型，都是通过随机化装置，让被调查者回答两个或多个问题(其中一个为需要调查的敏感性问题)中的一个，而这些问题的答案只能为“是”或“否”，因而最终得到的也只是用“是”或“否”来描述的品质化敏感性问题的答案，而无法取得各种不同的数据。因而，必须采取调查问卷的方式，然而在进行问卷调查时，由于调查项目对某些人群来说是敏感性项目，就会产生不回答，而恰恰可能是这部分数据，对调查者来说是最重要的，因此，处理由于敏感因素带来的无回答，就显得格外重要。

在处理单元无回答的做法中，比较常见的是二重抽样法，其基本思想是先从总体中抽取一个较大的样本，采用比较便宜但无回答率可能较高的调查方法，然后在无回答的单元中再抽取一个较小的样本进行详细调查，取得无回答单元信息，最后把这两部分调查结果结合起来对总体作出推断估计。而对于项目无回答，一般采用插补法，主要思想是应用已有的数据代替缺失的数据，具体又分为随机插补和确定性插补。由于敏感性因素既可能带来单元无回答，也可能带来项目不回答，因而，这两种处理方法都非常重要。本文只对处理敏感性问题带来的单元无回答进行探讨。

在对含有敏感因素调查中的无回答的处理当中，普通的二重抽样并不有效，这是因为，在对无回答层进行抽样，对抽中单元进行面访时，由于无回答者中含有敏感人群，若直接提出该敏感性问题，被调查者可能仍然拒绝回答，而多次访问会造成被调查者的反感，厌恶，使调查工作更加难以开展。因此，处理这种情况下的无回答，应将处理敏感性问题的方法与处理无回答的方法结合起来运用，对二重抽样法进行改进。

二、改进二重抽样的实施步骤

设总体单元数为N，先抽取样本量为n的简单随机样本，n_1是回答的单元数，其样本均值为$\bar{y}_{n1}$，n_2是无回答的单元数，$n=n_1+n_2$，然后从n_2中抽取一个子样本n_2'，其抽样比为$f=n_2'/n_2$，为取得这样n_2'样本的均值$\bar{y}_{n2}'$，从而对总体进行估计，应采用特殊的调查方法。由于调查项目含有敏感性因素，不回答可能很大程度上是由于敏感性因素引起的，因而在对n_2'这个样本进行调查时，可针对被调查者的心理特点，相应对敏感因素进行分层弱化，降低敏感因素的敏感度，争取被调查者的合作。具体步骤为：

*首先，划分调查层。*具体方法为：在第二重调查时，在调查方案的设计中，调查人员以所要提出的敏感性问题为核心，设想该问题的直接提出会对不同的调查人群造成的心理影响及原因，在此基础上围绕核心问题根据被调查者的心理设计一个或一系列与敏感性问题息息相关的非敏感性问题作为分层标志，从而将n_2'样本中敏感人群，次敏感人群，非敏感人群划分出来。例如：某企业对减肥产品的调查中，急需被调查者的体重区间。如直接提出这一问题，有些人(尤其是女性）可能出于个人隐私的考虑而不愿意回答。分析被调查者的心理，产品的使用效果会对被调查者产生不同的影响：使用该产品效果好的被调查人群非常高兴，比较愿意配合调查；而使用产品效果差的被调查者因对该产品失望不愿回答。因此，在第二次调查时，可采用被调查者对产品的评价作为非敏感分层标志，提出“您对该产品的效果满意吗？”等诸如此类的问题作为分层标志，将被调查者分为两类：满意者划入非敏感人群，不满意者划入敏感人群。

*其次，弱化各调查层的敏感度。*无论采用何种方法，目的在于通过对敏感性问题的层层剥离，分解出一系列与敏感性问题密切相关的非敏感性问题或敏感度很低的问题逐步提出，以降低其敏感度。前一步已将n_2'个样本分成了敏感层，次敏感层，非敏感层。对非敏感层，问卷可直接提出敏感性问题。对敏感层和次敏感层，就必须采用弱化敏感度的方法。在上例中，在设计各层的调查问题是时，对回答“效果不满意”的人群，可结合使用联想技法，投影技法，层层剥离体重这一核心，设立有内在逻辑关系的问题，如“您认为理想的减肥产品应在一月内体重减少____公斤，达到您满意的体重标准疗效不超过____月，您理想的体重标准是____公斤”等问题，通过对敏感性问题的分解，可减少直接提出该敏感性问题给敏感被调查者造成的突兀感和排斥感。

值得注意的是，该方法要想在运用中取得良好的效果，还必须解决一个关键的问题：做好调查人员的选取和培训工作，使其具有一定的谈话技巧和询问技术。只有这样，才能打消被调查者的疑虑，从而取得他们的合作。

*最后，解决“难啃的硬核”。*通过弱化敏感度的方法，仍然可能有部分被调查者不回答，这时，若进行多次访问，可能效果不佳。因为，在敏感因素已被弱化的情况下，不回答者可能是永久拒绝者，也就是Cochran (1977),Erison (1967),Kish (1965)所称的“难啃的硬核”。要想取得所需资料，可考虑替换样本。设在n_2'样本中，有n_2''个样本无回答，从第一重调查无回答层中，剔除第二重调查的样本，在剩下的n_2-n_2'个单元中，选取n_2''个单元，重新进行调查，其抽样过程如下图。

三、估计量与样本最优分配

《统计学原理》教材内容体系的优化

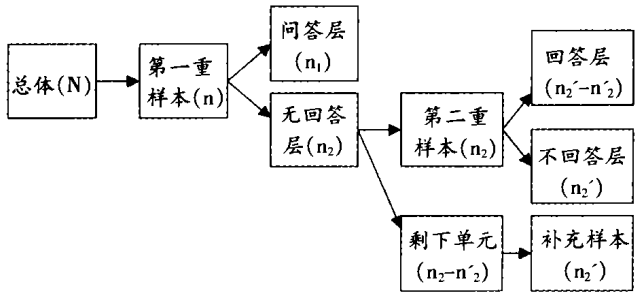
■刘翠杰

《统计学原理》教材一直以来都是多种版本并存的，特别是近几年中，各种不同名称的版本大量出现，除《统计学原理》以外，还有《新概念统计学》、《统计学》、《统计学基础》、《新编统计学》等等，虽然书名不同，但基本内容是一致的，有些教材中加入一些国民经济核算中的基本内容。在内容体系设置上，最少包括十章左右、最多包括十四章左右，各章的顺序安排有很大区别。

目前，随着教学课程体系改革的不断深入，课程体系的设置以培养学生的综合素质和动手能力为主要目标，课程体系的内容包括的模块越来越多，这样就必然要削减每门课的理论教学时数。统计学原理课程也不例外，特别是非统计专业的课时安排在逐渐减少，如我院的信息管理与信息系统、财务会计、国际贸易、审计、市场营销等专业统计学原理教学时数大概在40学时，随着专业选修课的增加，还可能更少。这就要求我们在教学内容体系的安排和教学实施过程中，力求理论性与实践性密切结合，尽量压缩统计描述部分的内容，尽量避免内容的重复，增加现代统计分析及应用的介绍，合理配置各章节的内容，依据内容的关联性安排内容的时序，优化教材的内容体系，使统计教材真正成为全面推进素质教育，培养统计创新人才的基本保证。

《统计学原理》教材内容体系可顺序地设置为以下八章内容：统计总论、统计调查、统计资料整理、集中趋势与离中趋势指标的测定与分析、时间数列与统计预测、统计指数因素分析、相关与回归分析、抽样推断与假设检验等。与一般教材的内容体系相比较，变化较大的有以下几个方面：

一、总量指标的内容处理



设第一重调查回答层的样本均值为$\bar{y}_{n1}$，第二重调查回答层的样本总量为$y_{n2-n'2}$，补充调查样本总量为$y_{n''2}$，设$\bar{y}_{n2}=\frac{y_{n2-n'2}+y_{n''2}}{n'_2}$，第二重抽样比为$f=\frac{n'_2}{n_2}$，$n=n_1+n_2$，第一重调查无回答率为$R_1$，则总体均值的一个估计量为：

$$\bar{y}=\frac{n_1\bar{y}_{n1}+n_2\bar{y}_{n''2}}{n}$$

因为补充样本和第二重调查中无回答单元都是从第一重调查中无回答的单元中抽取的，从而具有相似的特征，因此，用补充样本的数据代替第二重调查中无法取得的数据，不会造成误差，故$y_{n''2}$可直接看作是第二重样本的均值。由二重抽样的结论可知，$\bar{y}$是总体均值的无偏估计。

抽样方差为：

$$V(\bar{y})=(\frac{1}{n}-\frac{1}{N})S^2+\frac{R_1}{n}(\frac{1}{f}-1)S_2^2$$

其中S^2为总体方差，S_2^2为第一次调查无回答层方差。

$V(\bar{y})$的近似无偏估计为：

$$V(\bar{y})=(\frac{1}{n'_2}-\frac{1}{n_2})w_2^2s_2^2+(\frac{1}{n}-\frac{1}{N})w_2(\overline{y_{n'2}}-\bar{y})$$

其中w_2为第一重样本无回答层比重，s_2^2为第一重样本无回答层方差。

如果考虑二重抽样中的费用差异，总费用函数可记为：

$$C=c_0n+c_1n_1+c_2(n'_2-n''_2)+c_3n''_2$$
$$=c_0n+c_1n(1-R_1)+c_2(nfR_1-fnR_1R_2)+c_3fnR_1R_2$$

其中c_0为第一重样本每单元的调查费用，c_1为整理第一次调查中每个回答单元的费用，c_2是整理第二次调查中每个回答单元的费用，c_3为补充抽样每单元的费用。因而样本量的最优分配应极小化：

$$C\times(V+\frac{S^2}{N})$$
$$=[c_0n+c_1n(1-R_1)+c_2(nfR_1-fnR_1R_2)+c_3fnR_1R_2][\frac{S^2}{n}+\frac{R_1}{n}(\frac{1}{f}-1)S_2^2]$$
$$=[c_0+c_1(1-R_1)+c_2(fR_1-fR_1R_2)+c_3fR_1R_2]\{S_2+R_1S_2^2(\frac{1}{f}-1)\}$$
$$=\{[c_0+c_1(1-R_1)]+[c_2(fR_1-fR_1R_2)+c_3fR_1R_2]\}\{[S_2-R_1S_2^2]+R_1S_2^2\frac{1}{f}\}$$

运用柯西-许瓦兹不等式，当$\frac{\sqrt{c_0+c_1(1-R_1)}}{\sqrt{S_2-R_1S_2^2}}=\frac{\sqrt{c_2(fR_1-fR_1R_2+c_3fR_1R_2)}}{\sqrt{R_1S_2^2\frac{1}{f}}}$时，$C\times(V+\frac{S^2}{N})$取得最小值，此时

$$f=\frac{\sqrt{c_0+c_1(1-R_1)}}{\sqrt{S_2-R_1S_2^2}}\cdot\frac{S_2\sqrt{R_1}}{\sqrt{c_2(R_1-R_1R_2)+c_3R_1R_2}}$$

当总费用固定时，样本量为

$$n=\frac{C}{c_0+c_1(1-R_1)+c_2(fR_1-fR_1R_2)+c_3fR_1R_2}$$

将f代入上式，便可求出样本量。

通过对二重抽样进行改进，不仅提高了敏感性问题的回答率，提高了估计的精度，节省了费用，而且使得调查避免了访问被多次拒绝的情况，使得调查工作更加容易开展。

（作者单位/暨南大学统计学系）

（责任编辑/亦　民）