DOI:10.19503/j.cnki. 1000－2529. 2022. 01. 007

概率调查和非概率调查:权数的构建与调整

邹宇春,李建栋

摘　要:对于调查数据来说,合理有效地实现权数的构建与调整,是提高调查数据推论准确度的重要方法之一。由于调查数据通常会存在几类影响调查数据代表性的问题,本研究从权数构建与调整的角度提出有针对性的解决方案:其一,对于概率抽样调查,就调查设计、实施、完成过程中出现的不等概率、无应答、覆盖率不足等问题,提出包括广义回归在内的几种权数构建和调整方法;其二,对于非概率抽样调查中的选择性偏差问题,着重探讨如何用倾向性得分的方法来实现样本权数的构建和应用。

关键词:权数;概率抽样;非概率抽样

作者简介:邹宇春,中国社会科学院社会学研究所副研究员,社会学博士(北京 100732)
李建栋,中央财经大学文化与传媒学院讲师,金融学博士(北京 102206)

社会调查研究是采用特定数据收集方式获取相关数据,并通过统计分析认识社会现象及其规律的研究方法。调查数据的质量成为做好社会调查研究的关键之一。然而,由于调查数据的收集方法存在多样性、复杂性,数据使用者需深入了解调查数据在实际收集过程中可能存在的不足,比如拒访率高、空缺值多、低覆盖率以及样本有偏等问题,并针对这些不足进行合理科学的数据完善,才能提高调查数据的使用质量从而有效实现研究目标。

从已有的文献来看,调查数据的权数的构建与调整,是一项重要的数据完善方法。在社会调查研究中,为使调查的样本数据能更准确地反映目标总体的特征,研究者常常需要基于抽样方法、调查数据的不足对样本数据进行权数调整。可以说,权数是在样本推断总体时可以用来反映每个样本单元数据能够反映目标总体的程度。

遗憾的是,在使用调查数据进行目标总体的研究分析时,部分研究者由于不了解调查数据的问题以及相对应的权数构建方法,常出现研究结果发生偏差而不自知的现象。因此,基于调查数据的几类主要问题,有针对性地进行权数的建构和调整,将有助于提高基于调查数据的研究的质量。然而,当前国内有关社会调查研究方法的文献中,尽管已有学者专门介绍了调查数据加权方法,处理了抽样时的不等概率抽样、无应答和抽取的样本与总体不符的问题(亓昕,2003[1];金勇进和张喆,2014[2];王小宁,2019[3]),但对调查数据各类问题进行全面的权数分析的探讨还有待丰富,尤其对非概率抽样的样本加权讨论尤显不足。随着网络使用率的提升,基于非概率抽样的网络调查兴起,非概率抽样数据如何加权以做出更加有效的统计分析也有待更丰富的研究讨论。此外,刘展和金勇进(2017)[4]等针对某些非概率抽样提出了有很强参考价值的加权方法,比如当存在固定样本或另一个相关随机抽样样本时可为非概率样本构建权数的方法,但由于统计方法较为复杂而未能引起足够关注。

本研究分别对概率抽样数据和非概率抽样数据的权数的建构和调整予以分析,回答“如何对调查数据进行权数构建和调整”问题。首先,以多阶段混合概率抽样为例,针对概率抽样中常见的不等概率、拒访、覆盖有偏等问题提出相对应的权数构建与调整方法,并通过示例予以解释说明。同时,还就非概率抽样的加权问题进行讨论及示例说明。通过对这些权数构建方法的讨论,呈现较为全面的样本

权数分析框架。

一、概率抽样:针对抽样设计的权数

当无法收集所有研究对象的数据信息时,调查研究者会倾向采用概率抽样或非概率抽样来采集数据。概率抽样是指在总体调查对象中每个样本都有可能被抽到且以已知的、不为零的概率进入样本的抽样方法。这是一种基于概率理论和随机原则为依据的样本数据收集方法,总体中每个单位被抽中的概率可以通过样本设计来规定,通过某种随机化操作来实现,其目的是力求得到一个能代表目标总体的调查样本。

最简单的能够近似代表预先设定总体的样本被称为自加权样本。这种样本是通过等概率抽样获得的,每个样本的入样概率一样,因而其权数也一样。例如年级主任想检查年级内的学生是否完成了周末作业,事先对每个学生进行编号,用简单随机抽样(类似抓阄)的方式在年级学生名单中抽取学生来检查。此时,每位学生被查到的概率相等,在样本中其权数也就相同。当经济学家研究股票市场的时候,每一天的股票收益率可以当成一个随机变量的实现值,因此股票市场的时间序列数据也是自加权样本。自加权样本的"自"字,意味着样本的结构分布本身已自带加权信息,无须额外考虑样本的权数调整问题。

然而,在以人为调查对象的调查研究中,对于较大规模的调查样本,不易获得完整的样本框,难以通过一次性的简单随机抽样得到调查所需的样本。为了获得更有代表性的样本数据,研究者需要有意识地去设计抽样方法,并通过调整权数的方式达到"每个样本能够代表被调查总体的程度基本相同"的目标,以便于调查数据能更好地推论总体。一般来说,抽样方法在正式实地调查之前已由研究者设计好,每个样本的入样概率能事先计算得到,而该样本的权数便是入样概率的倒数,因而被称为设计权数。抽样设计权数与样本的入样概率成反比,也就是说样本的入样概率越大,其权数就越小。例如,以概率 1/50 选择的 1 个样本表示目标总体中的 50 个单位。样本权数之和提供了对目标人群中个人总数的公正估计。

在全国性的抽样调查中,通常采用多阶段概率抽样的方法。其设计权数等于各个阶段的样本单元入样概率的倒数之乘积。比如,中国社会状况综合调查(Chinese Social Survey,简称 CSS)抽样设计分为四个阶段(李炜、张丽萍,2014)[5]。第一阶段在全国的区/县/市名单内按照隐含分层的方式排序并采用概率比例规模抽样(Probability Proportional to Size sampling,简称 PPS)抽取 151 个区市县(初级抽样单元,Primary Sampling Unit,简称 PSU);第二阶段在抽中的区市县内抽取一定量的村委会/居委会(二级抽样单元,Secondary Sampling Unit,简称 SSU);第三阶段在抽中的村/居委会里抽取一定量的家庭户;第四阶段在每个抽中的家庭户内抽取 1 位符合调查要求的居民。此时 CSS 数据的设计权数为①:

$$W_{ijk}^{1}=\frac{N}{N_i}\times\frac{N_i}{N_{ij}}\times\frac{M}{S}\times\frac{N_{ijk}}{1}$$

其中,W_{ijk}^{1} 表示第 i 个被抽中的 PSU(区/县/县级市)中的第 j 个被抽中的 SSU(村/居委会)中第 k 个家庭中的被调查者的设计权数。N 是全国的总人口,N_i 是第 i 个被抽中的 PSU 的总人口数,N_{ij} 是第 i 个被抽中的 PSU 中的第 j 个被抽中的 SSU 总人口数,N_{ijk} 是第 i 个被抽中的 PSU 中的第 j 个被抽中的 SSU 中第 k 个家庭的总人口数。M 是被抽中的 SSU 的总家庭户数,S 是 SSU 计划调查的目标样本量。

在上面公式中,最后一步是根据各家庭内不同的人口数所进行的权数调整。在各类大型入户问卷调查中,研究者能提前掌握村居层面的抽样信息并计算出抽至此层次的设计权数,但由于家庭人口数往往在访问员进入抽中的家庭户进行户内抽样及问卷访问时才能获知,基于家庭人口对设计权数进行调整是设计权数的最后一步。由于家庭人口数存在差异,每位被抽中的受访者入选概率不同,所代表的家庭户信息的程度存在差异。假定家庭人口数为 N_{ijk},简单随机抽取一位家庭成员作为受访者,那么,在家庭中,某一个成人入选样本的概率为$(1/N_{ijk})$,在户内抽样阶段的每位受访者的权数为入选概率的倒数。

下面以此步为例,来具体说明样本权数的调整对提高调查数据研究质量的重要性。假定研究者已完成抽样设计,在抽中的 4 个家庭户进行户内抽样得到家庭人口数据,并随机抽取一位家庭成员作为受访者,询问是否有大学文凭以及家庭收入数据。同时,假定前三阶段的设计权数为 W_0,均为 50,根据家庭人口调整后的设计权数 W_a。表 1 是虚拟完成的调查数据。

表 1　基于抽样设计与抽样实施的权数调整

受采访者	家庭成人数 N_{ijk}	有大学文凭人数 Y_1	家庭收入(万元) Y_2	前几阶段的设计权数 W_0	W_0*Y_1	W_0*Y_2	$W_0*Y_1*Y_2$	根据家庭人口调整后的设计权数 W_a	W_a*Y_1	W_a*Y_2	$W_a*Y_1*Y_2$
#1	2	1	30	50	50	1 500	1 500	100	100	3 000	3 000
#2	1	0	20	50	0	1 000	0	50	0	1 000	0
#3	2	1	90	50	50	4 500	4 500	100	100	9 000	9 000
#4	4	1	50	50	50	2 500	2 500	200	200	10 000	10 000
总数	9	3	190	200	150	9 500	8 500	450	400	23 000	22 000

依据表 1 可发现，在调查样本中：(1)未根据家庭人口对设计权数进行调整时，有大学文凭的受访者比例为 75%(150/200)，受访者家庭收入的估计为 47.5 万元(9 500/200)，有大学文凭的家庭收入估计为 55.67 万元(8 500/150)，高出总体 17.2%(55.67/47.5－1)；(2)但是，依据家庭人口数对设计权数做应有的调整后，上面四个调查数据分析结果变为：89%(400/450)，51.11 万元(23 000/450)，55 万元(22 000/400)，7.6%(55/51.11－1)。对比未进行权数调整的调查数据和权数调整后的调查数据，两者分析结果呈现较大的差异。因此，在使用调查数据进行研究分析时，需及时根据抽样过程的信息对设计权数进行构建和调整。同时，该设计权数也是后续进一步权数调整的出发点，应当引起足够重视和应用。

二、概率抽样：针对拒访的事后加权

概率抽样调查进入实施阶段后，会遇到很多在前期设计环节无法控制的问题。邹宇春等(2019)[6]分析了社会调查执行过程中的种种障碍与挑战，对数据质量伤害最大的是拒访和对某些问题拒绝回答。这些均被称为无应答，在这些情况下如何对调查数据实行进一步的权数调整变得极为重要。

具体而言，调查实施过程中，被抽中的样本存在两种无应答情况：一是全部问题不回答(拒访、家中总是没有人或者没有符合调查资格的受访者)；另一种是部分问题不回答。前者被称为单位无应答(unit non-response)；后者被称为项目无应答(item non-response)，比较典型的项目无应答例子是高收入家庭不愿意透露收入数据。这两种无应答均会降低样本分析结果的有效性，很可能产生无应答偏差。即，如果无应答者与应答者之间存在系统性差异，基于应答者数据得到的分析结果在总体代表性上将有所削弱。

因此，调查实施阶段要尽量降低无应答的比例，并采用相应的实施方式缩小应答者和无应答者之间的差距。对于单位无应答，研究者会采用两种应对方式来减少无应答率：一是抽取并接触多于计划所需的样本量以减少无应答的样本②；二是采用在关键变量上具有与无应答样本相似特点的家庭户来替换无应答样本。对于项目无应答，研究者常采用追访、回访、补访甚至重访等措施补足样本信息。但这些应对方式并不能完全消除无应答情况，同时，如果没有做好严格的质控，替换样本很可能会增大调查误差。

为此，对于单位无应答，有必要针对样本缺失情况进行相应的权数调整，以降低无应答数据的负面影响。对于项目无应答，通常该样本信息仍保留在数据中，相应无应答项目以缺失值形式呈现，数据使用者常常直接舍弃样本或用插补(imputing)方法补足缺失数据。鉴于本研究的关注点是样本权数，本部分重点讨论单位无应答的权数建构与调整。为此需要进行的工作是，通过调查记录的样本信息(如平行数据)计算无应答率，并基于无应答率进行权数调整。

下面以一个虚拟的调查数据来分析如何根据拒访情况调整样本的权数。假设多阶段抽样，经过前几步抽样后，抽出了 4 个村居(见表 2 列 1)，在每个村居计划抽取不定量的家庭户(见表 2 列 2)，但由于各种原因最终实际接受访问的数量(见表 2 列 3)不同于计划访问数。假设调查目的是向每个受访者询问受访户家庭的汽车数量(见表 2 列 4)，假设前几个抽样阶段的基础权数为 W_a(见表 2 列 5)。

表 2　基于调查实施中拒访情况的权数调整

小区	计划抽取家庭户数 n	接受访问数 r	有汽车家庭数 y	基础权数 W_a	W_a*r	W_a*y	受访比例 p	调整后的权数 W_b	W_b*r	W_b*y
#1	10	8	7	100	800	700	80%	125.00	1 000	875
#2	20	17	12	100	1 700	1 200	85%	117.65	2 000	1 412
#3	30	15	10	200	3 000	2 000	50%	400.00	6 000	4 000
#4	40	36	18	200	7 200	3 600	90%	222.22	8 000	4 000
总数	100	76	47	600	12 700	7 500			17 000	10 287

接受访问数越小,则拒访率越高,即单位无应答率越高。这意味着能最后进入样本的数据代表着更多的小区家庭户,受访家庭者的权数相应地变得更高。针对拒访,权数调整公式应为:

$$W_b = W_a / P$$

其中 P 为受访比例。以第 3 个小区为例,受访比例为 $P=50\%$,其因拒访而得到的权数 W_b 为 200/(0.5)=400,即从 200 变成 400。分析调查数据可见:(1)当忽略任何权数,把样本当成自加权样本时,有汽车的家庭比例为 47/76=61.84%;(2)当使用基础权数,但未对拒访进行权数调整时,有汽车的家庭比例为 7 500/12 700=59.06%;(3)当使用基础权数,并根据拒访率进行权数调整时,有汽车的家庭比例为 10 287/17 000=60.51%。对比使用不同权数的结果,可发现存在明显差异,针对拒访的权数调整极为重要。尤其当研究者发现拒访人群的特征与研究问题的关键变量相关时,更应注重拒访权数的建构和使用,比如,拒访人群都是偏高收入者,而数据使用者的研究目标是分析受访者的收入差异。

三、概率抽样:针对覆盖有偏的事后加权

在设计权数和针对拒访率的权数调整的基础上,数据使用者在分析数据时仍可能发现调查数据的某些指标并不理想。重点表现为,数据覆盖的样本群体在某些特征上有偏,比如某些少数民族、高收入家庭或某特定人群偏低,出现样本特征与总体特征偏离的情况。

覆盖率有偏的原因有很多。比如,我国城市化进程加快而出现的地址变化大,拆迁改造、新建、并居等情况会导致一些抽中的地址空缺,或者本意是全国调查,却由于特殊原因如地震、暴雨、疾疫等排除了某些地区[③]。再比如,访问员对如何从建筑物中抽取受访者出现困难,尤其遇到集体户或多户并居、三代同堂、未婚同居、居民住宅进行商业经营等情况时,极易误判抽样单元而导致抽中错误的受访者。这些不足,仅通过设计加权或拒访加权的方法,覆盖率有偏的情况仍不能消除。

因此,针对覆盖率低或覆盖率有偏的不足,需在前两类权数调整的基础上进一步做权数调整。此阶段的权数调整所依据的基础不再是该调查本身的平行数据信息,比如家庭户人口数或调查现场的拒访数据,而是已发布的人口数据,比如少数民族人口数、男女比例、年龄分段人数、已婚人口等。这些数据可称为控制变量,来自权威的已发布数据库,比如全国人口普查、国家统计局统计年鉴、民政局资料等。本节介绍已知控制变量分布情况下的事后分层法与只知控制变量数值情况下的广义回归法。

(一)事后分层法

通常情况下,我们可以知道总体的某些变量的分布,比如年龄的分布、性别的分布等。当样本的分布并不与总体的分布相同时,可以通过对样本进行权数调整,让样本分布尽可能趋近总体分布。这种方法称为事后分层法(post-stratification),通过已知分布的变量构成不同的格子单元(子样本),使每一格子单元占总样本的比例都接近总体的相应比例。下面用虚拟例子说明。表 3 是某项调查数据发布之前的受访者的年龄结构(青年、中年、老年)与性别(男、女)组成,表 4 是从更权威数据源得到的总体年龄结构与性别组成。

对比表 3、表 4 发现,样本中低年龄段明显不足,高年龄段明显偏多。这与入户访问的特点有关,高年龄段有更高的可能性被采访到。同时,从性别分布看,样本中女性不足。权数调整的方法是行列迭代法,步骤如下:

步骤 1，行调整。把总体行的比例除以样本对应行的比例得到行调整系数，每一行都乘以该行的调整系数。例如，第 1 行调整系数为 30%/20%＝1.5，原第 1 行的 15%，5%乘以 1.5 后得到 22.5%，7.5%。第 2 行调整系数为 45%/40%＝1.13，调整后变为 28.13%，16.88%。第 3 行调整系数为 25%/40%＝0.63，调整后变为 12.50%，12.50%。行调整后，每一行的样本比例就等于总体的比例，男女的比例变为 63.13%，36.88%。

表 3　样本的年龄－性别结构

年龄组	男	女	总和
18～34 岁	15%	5%	20%
35～59 岁	25%	15%	40%
60 岁及以上	20%	20%	40%
总和	60%	40%	

表 4　总体的年龄－性别结构

年龄组	男	女	总和
18～34 岁	15%	15%	30%
35～59 岁	20%	25%	45%
60 岁及以上	15%	10%	25%
总和	50%	50%	

步骤 2，列调整。把总体列的比例除以上一步骤结果对应列的比例得到列调整系数，每一列都乘以该列的调整系数。例如，第 1 列调整系数为 50%/63.13%＝0.79，第 1 列从 22.50%，28.13%，12.50%乘以 0.79，变为 17.82%，22.28%，9.90%。第 2 列调整系数为 50%/36.88%＝1.36，第 2 列调整后变为 10.17%，22.88%，16.95%。列调整后，每一列的样本比例就等于总体的比例，三个年龄段比例为 27.99%，45.16%，26.85%。

步骤 3，再次行调整。把总体行的比例除以上一步骤结果对应行的比例得到行调整系数，每一行都乘以该行的调整系数。例如，第 1 行调整系数为 30%/27.99%＝1.07。调整后男女的比例变为 50.52%，49.48%。

步骤 4，再次列调整。在上一步骤结果上再次进行列调整。第 1 列调整系数为 50%/50.52%＝0.99。

如此，行、列调整不断进行直到满意的结果。通过上述步骤，我们发现样本的分布已收敛于总体的分布，我们就可以把每一步骤的系数相乘，得到综合调整系数，赋予相对应的格子单元。权数调整公式为：

$$W_{c,i,j}=W_{b,i,j}*A_{i,j}$$

其中，$A_{i,j}$ 是上述分层中第 i 行 j 列样本的综合调整系数。例如，年龄 18～34 岁的男性样本（第 1 行第 1 列），调整系数 1.260 3（＝1.5＊0.79＊1.07＊0.99）。所有系数如表 5 所示。

表 5　每一个年龄－性别结构格子单元的综合权数调整系数

年龄组	男	女
18～34 岁	1.260	2.202
35～59 岁	0.878	1.536
60 岁及以上	0.456	0.797

在该例中，样本中的低年龄段明显不足，同时女性不足。因此低年龄段、女性的综合调整系数为 2.202，大于 1，表示要让该子样本代表更多的总体。同样道理，高年龄段、男性格子的调整系数为 0.456，小于 1，表示要降低该子样本的代表性。其他格子单元的权数调整系数则介于两者之间。

（二）广义回归法

当我们有控制变量的总体水平数据却不知道其分布时，覆盖率不足的问题已无法采用行列迭代法予以解决。但我们仍然要面对低覆盖率问题，找出合理的办法进行权数调整。下面用关于某项家庭消费的虚拟调查数据例子来说明。

假设数据已完成家庭户人口、拒访率、已知分布的变量的覆盖率等问题的权数调整，设为前期权数 W_c（表 6 第 2 列）。此外，假设从更准确权威的数据已知某居住小区目标总体有 60 岁及以上男性 300 人和 60 岁及以上女性 300 人，这两个数据将作为权数调整的控制变量。而此样本数据显示，60 岁及以上男性数为 280 名（Σw_c*n_1），60 岁及以上女性数为 220 名（Σw_c*n_2）。这与已知两个数据存在明显差异，说明 60 岁及以上男性和女性都存在覆盖率不足的问题，但女性更严重。需对数据进行权数调整，使样本数据与总体数据一致。

可以考虑如下步骤。第 1 步,首先考虑第 1 个控制变量,即 60 岁及以上男性人数。此时,权数调整的比例为:300/280=1.071 42,将该数乘以前期权数 W_c,我们得到一组新的权数 W_{d1}。依此新权数,我们会得到虚拟的家庭单元为 343 家。60 岁及以上男性人数恰好为 300。但 60 岁及以上女性人数只有 214。第 2 步,60 岁及以上女性人数远远不符合总体 300 的要求,覆盖率不足。既然女性人数不足,那么就要增加女性多男性少的样本的权数。但是需要满足的约束条件只有两个,可以调整的样本种类大于两个,因此,有无穷多个调整办法。例如,对样本中无 60 岁及以上男性的 2 个样本(样本第 4,样本第 6)权数进行调整,乘以 1.889,得到新的一组权数 W_{d2}。只调整这两个样本不会影响第 1 步的结果。此时我们发现,样本代表的 60 岁及以上男性为 300 人,代表的 60 岁及以上女性亦为 300 人(表中最右侧两列。)覆盖率不足的问题得到了解决(calibration)。

但是这样的调整大大改变了原来的权数结构,基于新权数的估计值的性质无从谈起。因此,需要对新权数的构建提出一些要求。基本要求就是尽量不改变原来权数的结构。下面用广义回归法来进行权数调整。广义回归法是一种特殊的权数校准办法。

表 6　多变量校正法示例

受访者编号	前期权数 W_c	家里 60 岁及以上男性人数 n_1	家里 60 岁及以上女性人数 n_2	某项家庭消费(元)	权数调整 W_{d1}	60 岁及以上男性人数	60 岁及以上女性人数	权数调整 W_{d2}	60 岁及以上男性人数	60 岁及以上女性人数
1	50	1	1	200	53.571	53.571	53.571	53.571	53.571	53.571
2	50	1	0	500	53.571	53.571	0	53.571	53.571	0
3	30	1	2	600	32.143	32.143	64.286	32.143	32.143	64.286
4	40	0	1	200	42.857	0	42.857	80.952	0	80.952
5	50	2	0	400	53.571	107.14	0	53.571	107.14	0
6	50	0	1	500	53.571	0	53.571	101.19	0	101.19
7	50	1	0	300	53.571	53.571	0	53.571	53.571	0
代表家庭户数	320				343			429		
代表人数		280	200			300	214		300	300

为了构建回归型权数,先定义几个数学符号如下(参考 Zieschang(1990)[7]):

Ω 是 $n\times 1$ 的初始权数向量,代表着样本被选择的概率的倒数,n 是样本数。Ω 既可以是最初设计权数,也可以是在调查执行时由于拒访(单元无应答)等原因而调整后的权数。总之,我们的前提是根据 Ω 可以得到无偏的估计。

X 是 $n\times k$ 的控制变量的矩阵,k 是这些控制变量的个数。根据假定,这些 k 个控制变量总体水平的数据是已知的。这些控制变量通常是年龄、性别、人种等分类变量。

Nx 是 $k\times 1$ 的向量,代表着刚才说的这些控制变量的总体水平。

W 是 $n\times 1$ 向量,代表着调整后的权数,也被称作 GLS 权数(Generalized Least Square,GLS)。根据这个权数,可实现 $X'W=N_X$,即根据样本计算的控制变量总体值与实际相符。

Λ 是 $n\times n$ 矩阵,代表着 Ω 与 W 两个权数向量的协方差矩阵。为简单计,它被设定为对角矩阵,其对角线上的元素为 Ω 的元素,即 $\Lambda=\mathrm{diag}(\Omega)$。

调整权数的思路为:目标是使 W 与 Ω 之间的距离非常小,如果根据 Ω 可以得到无偏的估计,根据 W 进行的估计也会保持无偏的性质。如何定义距离在数学中有很多选择。Deville 和 Sarndal (1992)[8]进行了总结。应用最广泛的距离定义是 $(\Omega-W)'\Lambda^{-1}(\Omega-W)$。这样 W 的计算就变成这样一个问题:在满足 $X'W=N_X$ 约束下,如何最小化 $(\Omega-W)'\Lambda^{-1}(\Omega-W)$。此经典问题的答案是:

$W=\Omega+\Lambda X\ (X'\Lambda X)^{-1}(N_X-X'\Omega)$。

设定有一个感兴趣的变量 Y,在构建出 W 之后,如何估计 Y 的总体值呢?自然地,估计值是 $Y'W$。下面是 GLS 权数调整的方法。

表 7　GLS 调整权数

受访者编号	最初权数 W_c	家里 60 岁及以上男性人数 n_1	家里 60 岁及以上女性人数 n_2	某项家庭消费 y(元)	GLS 权数 W_{GLS}	60 岁及以上男性人数	60 岁及以上女性人数	家庭总体消费额 $W_0 * y$	权数调整后家庭总体消费额 $W_{GLS} * y$
1	50	1	1	200	67.301	67.301	67.301	10 000	13 460
2	50	1	0	500	46.655	46.655	0	25 000	23 327
3	30	1	2	600	52.768	52.768	105.536	18 000	31 660
4	40	0	1	200	56.517	0	56.517	8 000	11 303
5	50	2	0	400	43.31	86.62	0	20 000	17 324
6	50	0	1	500	70.646	0	70.646	25 000	35 323
7	50	1	0	300	46.655	46.655	0	15 000	13 996
家庭总数	320				384			121 000	146 395
人总数		280	200			300	300		

从上表中得到两个辅助变量，就是男性人数与女性人数，这两个辅助变量是进行权数调整的依据。由此：

$$X=\begin{pmatrix}1&1\\1&0\\1&2\\0&1\\2&0\\0&1\\1&0\end{pmatrix},\Omega=\begin{pmatrix}50\\50\\30\\40\\50\\50\\50\end{pmatrix},\Lambda=diag\begin{pmatrix}50\\50\\30\\40\\50\\50\\50\end{pmatrix},N_X=\begin{pmatrix}300\\300\end{pmatrix},\mathrm{Y}=\begin{pmatrix}200\\500\\600\\200\\400\\500\\300\end{pmatrix}$$

利用公式　$W=\Omega+\Lambda X\ (X'\Lambda X)^{-1}(N_X-X'\Omega)$，得到 GLS 权数

$$\mathrm{W}=\begin{pmatrix}67.301\\46.655\\52.768\\56.517\\43.310\\70.646\\46.655\end{pmatrix}$$

此 GLS 权数保证了男性总体总数 300 与女性总体总数 300 的约束条件，毕竟 GLS 权数就是依据此约束条件计算得到的。GSL 权数的计算直接，一步到位，同时满足了多个约束条件。从公式来看，W 是直接计算得出的，不需要循环代入。但如果数据样本较多，矩阵的维度上升，会带来计算上的耗时成本。

如果研究的变量是“某项家庭消费额”，可以根据 GLS 权数直接估计。家庭总体消费额为 $Y'W=146\ 395$ 元。而此时家庭总体数为 384，因此家庭平均消费额为 146 395/384＝381.38 元。对比最初设计权数未调整情况下，总体消费额为 12 100 元，总体家庭数 320 个，家庭平均消费额为 12 100/320＝378.13 元。

关于这样权数的构建，可以从另外一个角度来理解。把 W 的公式带入 $Y'W$，得到：

$$\mathrm{Y}'W=Y'[\Omega+\Lambda X(X'\Lambda X)^{-1}(N_X-X'\Omega)]=Y'\Omega+(X'\Lambda X)^{-1}(X'\Lambda Y)(N_X-X'\Omega)$$

其中 $(X'\Lambda X)^{-1}(X'\Lambda Y)$ 恰好就是 Y 对 X 做加权线性回归的系数。因此这样的权数调整方法也称为基于回归的权数构建，记为 GLS 权数调整(Generalized Least Square)。

四、非概率抽样：倾向性得分匹配的权数

非概率抽样的调查数据不同于概率抽样的调查数据。由于前者不是严格地按随机抽样原则来抽取的

样本数据,失去了大数定律的存在基础,因而无法确定非概率抽样误差。加之其数据统计特征不明确,导致很大程度上无法有效地计算出样本的统计值在多大程度上适合于总体。非概率抽样实施方法较为多样,包括方便抽样、定额抽样、滚雪球抽样,等等。因此,理论上,很难找出通用的解决办法来提高非概率抽样调查数据的统计代表性。

不过,随着互联网使用的日益广泛,网络调查成为较广泛使用的一种非随机抽样调查。这类调查常常交给有固定样本群的专业数据调查公司来实施。这种情况,为使用倾向性得分匹配(Propensity Score Matching,简称 PSM)的方法来处理非概率抽样样本成为可能。处理方法有两类:一是在实施调查之前的设计阶段就从固定样本群中选配样本,使答题者的参与有"概率"可循;二是在实施调查之后,进行权数构建,也一样获得了"概率",为后面的统计提供支撑。本节着重后者,即讨论数据调查完成之后用 PSM 建立权数。

PSM 的思想是利用 Logit 模型,把是否参与调查当成 Logit 模型的因变量,把一些控制变量当成 Logit 模型的自变量,通过模型得到参与调查的倾向性得分,倾向性得分的倒数可以当作权数。Logit 模型如下:

$$p_i=\frac{\exp(\beta_0+\beta_i x_i)}{1+\exp(\beta_0+\beta_i x_i)},\ i=1,2,\cdots,c$$

其中,p_i 是参与网络调查的概率,β_0 是常数项,x_i 是影响参与网络调查的众多因素,β_i 是相应的系数。

下面通过例子来说明如何使用 PSM 实现非概率抽样的权数构建(具体见表 8)。假设在一次网络自愿调查中,有 13 个参与者。由于是自愿调查,因此存在着选择偏差问题,即某些特征的人更愿意来参与调查。这是一种非概率抽样,因此基于这 13 个简单数据,关于某项政策满意度(y)的统计量并不能反映那些"沉默的民众"的观点。

表 8　用倾向性得分来构建权数

受访者	性别	年龄	满意度 y	参与概率 p	权数 $w=1/p$	$w*y$	其他受访者	性别	年龄
#1	0	23	4	0.431	2.318	9.274	#14	0	18
#2	0	23	5	0.431	2.318	11.592	#15	0	21
#3	0	28	6	0.558	1.791	10.744	#16	0	31
#4	0	36	8	0.741	1.349	10.791	#17	0	48
#5	0	42	5	0.841	1.189	5.944	#18	1	18
#6	0	46	5	0.889	1.125	5.627	#19	1	20
#7	0	55	9	0.952	1.050	9.450	#20	1	25
#8	0	56	10	0.957	1.045	10.451	#21	1	27
#9	0	58	9	0.965	1.037	9.331	#22	1	28
#10	1	30	5	0.144	6.956	34.780	#23	1	32
#11	1	45	9	0.438	2.285	20.562	#24	1	33
#12	1	52	8	0.614	1.628	13.023	#25	1	33
#13	1	56	9	0.706	1.417	12.754	#26	1	38
平均值			7.077				#27	1	41
总值					25.508	164.323	#28	1	48

如果数据调查公司正好有另一个样本,且是概率抽样样本(假定其样本数为 28),其中恰好包含了这 13 个受访者,那么这个概率抽样样本就可以帮助实现 PSM 权数构建。这个概率抽样样本的其他 15 个人的信息放在表格的右侧。

以参与这次调查为事件(因变量),性别、年龄为自变量,对整个 28 个观测值运行 Logit 回归模型,可以发现这两个自变量都显著,结果如下:

$$\hat{p}_i=\frac{\exp(-2.628-2.223Gender+0.102Age)}{1+\exp(-2.628-2.223Gender+0.102Age)}$$

由此计算出这 13 个主动参与调查者的参与概率 p(即倾向性得分),取倒数得到其"伪权数"$w=1/p$。

根据此权数，政策的满意度平均值为 164.323/25.508＝6.442，而非没有权数时的 7.077。

通过 PSM 办法构建权数，可以让非概率抽样产生样本“伪概率”，增加了估计的准确性。当然，用 PSM 办法构建权数的前提是，必须寻找到一个概率抽样样本包含着该非概率抽样样本，这其实是一个很苛刻的条件。另外，在计算倾向性得分时，不同的自变量选择也会影响最后的权数结果，Logit 模型是基于数据的模型，缺少理论支撑。对非概率抽样样本用 PSM 办法构建权数只能是一种并不最理想但理论上基本可行的做法。

现有文献中用 PSM 办法构建权数时，还有另外一种处理办法，参考 Lee et al.（2009）[9]，Valliant 与 Dever（2011）[10]，刘展、金勇进（2017）[4]，均是把倾向性得分值相同的所有受访者放在一组内，先对组赋予权数，然后在组内调整，使得落在这一组的网络参与者占所有网络参与者的比例等同于该组未参与者占所有未参与者的比例（即在组内，网络调查的分布与概率抽样的分布相同）。本文不再赘述此方法。

五、总结

本研究梳理了几类影响调查数据推断总体的问题，并提供了相应的权数构建和调整方法。对于随机抽样数据，权数分析有很大必要性，它保证了样本数据推断总体的质量。尤其是 GLS 方法，在数据调查的最后阶段，可以通过 GLS 方法综合解决调查数据中产生的大部分问题，诸如无应答、覆盖率低等。对于非概率抽样数据，在有概率抽样样本做参考的情况下，可以用 PSM 方法来构建权数。本研究的贡献在于，在已有文献的基础上作了较系统的调查数据的权数构建和调整方法的梳理，结合调查数据的执行特点和虚拟案例，详细解释了权数的构建思路及应用，尤其是基于广义回归的权数构建与基于 PSM 方法的权数构建，从方法学的角度为研究者提供了借鉴和参考。

具体而言，在设计阶段，由于入户调查的性质，每个样本的设计权数需要根据受访家庭进行调整；在实施阶段，由于出现拒访、拒答问题，样本的权数要随之进行改变，原则上，那些出现拒访的家庭所在的小区应赋予更高的权数；在调查完成阶段，需要依据通常更准确的另外数据来源的性别、年龄、民族等数据来进行样本权数调整，以解决覆盖率不足、有偏的问题，使样本在这些性别、年龄、民族等数据上与总体保持一致。如果知道这些变量的联合分布，可以采用事后分层法（例如行列迭代法）；如果仅知道这些控制变量的总体水平，可以采用广义回归法。广义回归法的核心思想是在解决覆盖率不足的问题前提下，对旧的权数上做最小的改动。针对非概率抽样，在另有一较大的概率抽样样本，且较大样本包含着非概率抽样样本时，我们可以通过 PSM 办法得到某个体进入非概率抽样样本的近似概率，从而构建出每个个体的权数。

本研究仅在基础层面讨论了数据调查过程中权数的构建与构建方法。囿于篇幅，有些重要环节并未涉及，但这些环节对于调查数据的分析有重要影响，留待以后进一步研究。这些环节包括：（1）在使用了 GLS、PSM 等权数构建方法的前提下，有必要对估计值的质量做出评估。尽管这些权数构建的做法已约定俗成，但估计值是否会产生较大的变异有待统计分析。（2）对于非概率抽样，其种类多、差异大，但本研究仅提供了一种权数调整的途径，其他的途径也都是需要特定的条件才能实现，本研究并未能详尽展开论述。（3）在调查数据的研究前沿，除了本研究介绍的基本方法外，还有贝叶斯方法、伪权数、工具变量等探索。需要的读者可以参考 Elliott（2009）[11]、Kott（2006）[12]等文章。数据调查研究者有必要加强对这些方法的理论及应用情况的了解。

注释：

① 在实际执行中，CSS 扩大了接触样本量以便能完成目标样本量，故还应有实际执行调整权数。为节省篇幅并考虑到下文有对无应答权数调整的阐述，此处略去此权数论证。具体可参见李炜和张丽萍（2014）文章。

② 严格说来，重新抽样时，入样概率由于移除了拒访家庭会与最初设计有些许差异。

③ 因此，数据发布必须明确说明采样的范围、目标总体和实施的情况。

参考文献：

[1]亓昕.人口抽样调查数据分析中的加权方法[J]. 人口与经济，2003(1)：40-43.

[2]金勇进,张喆.抽样调查中的权数问题研究[J]. 统计研究,2014(9):79-84.
[3]王小宁.权数在人口抽样调查估计中的应用研究[J]. 统计与信息论坛,2019(12):9-15.
[4]刘展,金勇进.网络访问固定样本调查的统计推断研究[J]. 统计与信息论坛, 2017(2):3-10.
[5]李炜,张丽萍.全国居民纵贯调查抽样方案设计研究[J]. 科研信息化技术与应用, 2014(6):17-26.
[6]邹宇春,等.仗卷走天涯:全国大型社会调查之督导笔记[M].北京:社会科学文献出版社, 2019:125-126.
[7]Zieschang, Kimberly D. Sample Weighting Methods and Estimation of Totals in the Consumer Expenditure Survey[J]. Journal of the American Statistical Association, 1990,85(412): 986-1001.
[8]Deville J, C Sarndal. Calibration Estimators in Survey Sampling[J]. Journal of the American Statistical Association, 1992, 87(418): 376-382.
[9]Lee S, Valliant R. Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and CalibrationAdjustment[J]. Sociological Methods & Research, 2009:319-343.
[10]Valliant R, J A Dever. Estimating Propensity Adjustments for Volunteer Web Surveys[J]. Sociological Methods & Research, 2011, 40(1):105-137.
[11]Elliott M R. Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights[J]. Survey Practice, 2009(August):1-7.
[12]Kott P S. Using Calibration Wighting to Adjust for Nonresponse and Coverage Errors[J]. Survey Methodology, 2006, 32(2):133-142.

Construction and Adjustment of Weights in Probability Sampling and Non-probability Sampling Survey Data

ZOU Yuchun, LI Jiandong

Abstract: Whether sample weights can be reasonably and effectively constructed and adjusted is one of the important aspects to improve the inference accuracy in using survey data. Subject to problems affecting the representativeness of survey data, this study puts forward targeted solutions from the perspective of weight construction and adjustment. On the one hand, for probability sampling survey, this study puts forward several weight construction and adjustment methods including generalized regression method to solve the problems of unequal probability, no response and insufficient coverage in the process of survey design, implementation and completion. On the other hand, for non-probability sampling survey that has severe selection bias problem, this paper focuses on how to use propensity score method to realize the construction and application of sample weight.

Key words: weight; probability sampling; non-probability sampling

(责任编校:文晶)