概率抽样条件下样本代表性事后评估方法探讨*

宋子轩　冷　燮　陈瑶瑶

内容提要：样本代表性直接牵涉到统计数据质量和统计公布引起的民众反响，目前社会上不乏对政府统计数据的质疑之声，最终影响到政府统计机关的公信力，因此有必要重新审视现行样本代表性的研究。目前相关文献普遍强调不同抽样方式下的样本代表性的相对性内涵，确保样本的代表性仅从抽样方法和样本量两个维度入手，缺乏对既定抽样方法下样本代表性的事后评估体系的探索，以及多样本之间样本代表性优劣的比较方法研究。鉴于此，本文在结合人口普查数据基础上尝试从样本—总体整体分布和内部属性结构两个方面构建样本代表性事后评估的一整套指标和假设检验，进而找到一种多样本代表性比较的依据。最后对浦东新区 2010 年城镇居民收入调查样本进行了代表性检验的尝试。

关键词：抽样；样本代表性；事后评估；分布检验；列联系数
中图分类号：C811　　文献标识码：A　　文章编号：1002－4565(2012)07－0096－05

Discussions on the Post-Evaluation of Sample's Representativeness in the Condition of Probability Sampling

Song Zixuan　Leng Xie　Chen Yaoyao

Abstract: The sample's representativeness directly involves the quality of statistics and its public repressiveness, but nowadays sound of doubts arises frequently, which further influences the credibility of governmental statistical offices, so it's necessary to reconsider the research on the sample's representativeness. The current related literature has always been focused on its relativity between different sampling methods, to guarantee the sample's representativeness generally set an emphasis on the sampling design and sample size, which lacks the research on post-evaluation of sample's representativeness under the given sampling method and the comparable methods of multiple samples' representativeness. This paper combining with the population census data to attempt to set up a series of indicators and hypothesis tests from the aspects of sample-population whole distribution and internal attribute structure in order to find a comparable basis between multiple samples. In the end, we make an attempt to evaluate the sample's representativeness of the 2010 urban resident earnings survey in Pudong District.

Key words: Sampling; Sample Representativeness; Post-evaluation; Distribution Test; Contingency Coefficient

一、问题的提出

1994 年以后抽样调查方法在我国统计工作中得到普遍应用，至今抽样调查已经成为我国政府统计工作中非全面调查的首选。但在实施抽样调查的过程中依然存在有悖于抽样调查原则的问题，具体到抽样技术及应用环节，存在的问题主要有以下四方面：一是抽样框的变化使最终确定调查个体（家庭户）的名录也相应地出现了变动，必要的样本单位数不足，在实际调查中应对这类问题的做法是变随机抽样为随意抽样，进而造成了抽样误差的大幅扩大。二是在抽样实施过程中缺乏对于无回答或无效回答的预防和有效处理措施。三是样本量的确定缺乏抽样的理论计算，或者选用不恰当的方差估计

* 本文获上海市浦东新区统计局项目“概率抽样样本代表性研究”（2011shpd0701）资助。

方法。四是按照行政区划作为分层的依据默认的前提是同一区域内调查单位的差异减小,而区域之间的差异大,实地调研工作中墨守成规地遵循这一既定前提,对抽样误差也会有一定程度的影响。

上述存在的种种问题必然有损样本对总体的代表性,民众对统计数据的现实差异感较大,质疑呼声高,最终影响到政府统计机关的公信力,因此有必要重新审视现行样本代表性的研究。

二、对样本代表性的再认识

目前学界专门研究样本代表性问题的文献屈指可数,并且多半文献讨论的焦点集中在对样本代表性的理解上,另外一部分文献的焦点则是国内统计工作制度的改革、统计数据的误差来源。如果将审计、农业、医药卫生等领域的调查研究文献包含在内,确保样本代表性的切入角度有抽样设计和样本量两方面。

在抽样设计与样本代表性的研究方面,冯士雍(2001)认为对于样本对总体的代表性不能理解为一个具体的样本对总体的代表性,应该从判定目标估计量的优良准则来考察,这其实就是从样本数据的获取方式上(抽样方法)考察总体目标估计量是否是无偏的、有效的、一致的以及均方误是不是最小的,而概率抽样得到的样本一般都具有上面所述的优良性,因此他认为基于概率抽样下的样本对总体的代表性都是有保证的。俞纯权等(2003)认为即便是同一抽样方法下抽取的样本也是随机的,因此比较同一抽样方式下的各个样本的代表性没有可比的价值,因此从相对性的角度考察样本的代表性其比较的前提是不同抽样设计下抽取的样本。在样本量与样本代表性的研究方面,李文华(2006)认为虽然从纯理论上讲,样本的代表性与样本容量没有必然的关系,但是在实际中,总体中个体之间的差异是固然存在的,所以一般情况下样本容量越大,抽样误差越小,其对总体的代表性也可能越大。显然这种代表性考察的逻辑准则也是基于总体估计量的优良性指标。

无论是从抽样设计还是样本量上确保调查样本代表性可以称之为事前保证方法,对于已经从总体抽取出来的样本验证其是否具有较高的代表性,这属于事后评估的范畴。对事后评估的研究文献目前比较少并且存在争议,对于样本代表性的理解常被提及的一种观点是认为样本的结构与总体的结构尽可能相近或者说样本分布尽可能与总体分布相一致即意味着样本具有良好的代表性,这就与上述从抽样设计和样本量来确保样本代表性的事前保证产生了区别,因此这种事后检验的指标受到有些学者的质疑和批判。

通过对这些文献的归纳整理,发现争论的焦点集中在三个方面: 一是认为如果强调样本的结构尽可能和总体的结构保持一致才足以使得样本具有较好的代表性,那么有可能会陷入"代表性抽样"的历史错误之中,破坏样本获得方式上遵循的概率抽样的原则,将导致抽样误差的扩大。二是认为由于事后检验的指标中含有关于总体的未知参数,因此在实际操作层面上这类检验的方法不具有可操作性。三是认为评估样本代表性的关键是判断样本是否是一个概率样本,而样本特征与总体特征之间具体差异的大小,并非与概率样本以及代表性高低有必然联系,且事后检验指标的构造思想存在逻辑上的错误,即使样本的各种结构特征和总体保持一致甚至都不能排除该样本是来自于其他的非目标调查总体。

纵观现有关于样本代表性的文献,目前普遍集中在事前保证的研究方面,其中多数文献又是分别从抽样设计、样本量、辅助信息、辅助变量四个方面展开,并形成了普遍的共识,研究的空间狭小。因此,本文在评述已有相关文献的基础上着眼于抽样实施之后的样本评估方法研究,从而能够保证样本代表性从调查前后两个节点展开评估,即事前保证和事后评估。

三、样本代表性的事后评估方法

(一) 单样本评估方法的构建

前面已经提到当采用一种抽样方法获得的多个样本之间比较优劣,其评估可依据样本的分布和总体的分布是否一致来衡量。基于上述思路我们从样本和总体的内部中具体标志的结构和样本和总体的综合分布两个角度展开对样本事后评估方法的探讨,见表 2。

(二) 多样本代表性比较方法

多样本之间比较代表性优劣的前提是这些样本均出自于同一个总体,依据样本的抽样方法是否相同可以划分为两种情况。当多个样本按照不同的抽样方法获得时,样本代表性的比较可以通过抽样方式下总体目标推断统计量的优良性准则加以判定;

当多个样本采用同一种抽样方法获得时，样本之间代表性的优劣可以按照各个样本与总体的整体差异率大小来判定，差异率越小则说明该样本的代表性越好。

针对后一种情况，计算整体差异率基本的思路是将样本与总体的各个属性之间的差异率（GCR 或 DI）加权汇总成一个确切的数值，权重的设计从属性变量相对于抽样调查的目标推断统计量之间的相关关系入手，相关程度越高则表明该属性相对于调查目的的重要性就越高，所以以这种相关程度的大小作为权重的替代选择是符合逻辑的。实际操作中由于调查目的各异，总体的目标推断统计量又有连续型和离散型之分，选择的相关系数也必须根据实际情况恰当地选用。若两个属性变量均为定序数据时，可以选择 Spearman 秩相关系数和 Kendall 秩相关系数；若两个属性变量仅为定类数据时，可以选择列联系数；若两个属性变量中一个为定类或定序数据另外一个为定距或定比数据时，先将定距或定比数据通过分组使其离散化，再根据前两种情况具体选择。

以城镇居民收入调查样本为例，人口或家庭属性为定类数据，而收入作为调查的总体目标推断统计量，首先需要将收入数据分组处理，然后通过计算列联系数作为权重的替代，列联系数的计算公式如式（1）：

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}} \qquad (1)$$

其中 n 为样本容量，χ^2 统计量是列联表的检验统计量，用来检验两个属性是否相互独立。

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$

$$\sim \chi^2((r-1)(c-1)) \qquad (2)$$

其中 r、c 分别为两个属性的类别/水平数目，假设这两个属性的分别为 A、B，则 n_{ij} 表示 n 个个体中既属于 A 的第 i 类又属于 B 的第 j 类的数目。可见，列联系数 C 的取值在 0～1 之间，当 C = 0 时，表示两个属性 A 和 B 之间没有关联；当 C 接近于 1 时，表示两个属性 A 和 B 之间关联性很强。

（三）数据

本文拟通过上海市浦东新区社情民意调查中心的城乡居民收入状况调查的城镇居民样本，结合 2010 年浦东新区第六次人口普查数据进行实际测算，比较样本代表性事后评估指标和分布检验的实际应用效果。需要说明的是由于人口普查中没有涉及到居民收入的数据，但是考虑到居民收入和人口属性有着一定的内在相关关系，例如性别、从事的职业和行业都会引起收入上的差异，因此此次评估是从区域分布、家庭属性以及个人属性三个方面展开的，总类下又有细分，其中城镇居民的样本容量为 600 户，共计 1692 人，评估角度分为所在街道、家庭人口规模、户口状况、性别、文化程度、行业、职业七个方面。人口普查数据中城镇居民中剔除了集体户，总数为 2971809 人。

（四）属性权重的计算结果

上海市浦东新区城镇居民的月收入介于区间 0 至 70000 元，结合本区实际情况，利用数据库软件 SQL Server 2008 编程提取收入数据并将收入等距划分为七组，组距 10000 元，然后在 SPSS 中选择交叉表分析计算所在街道/乡镇、家庭人口规模、户口状况、性别、文化程度、行业、职业与家庭成员月收入两两之间的列联系数，如表 1，所有的列联系数在 5% 的水平下均通过了显著性检验，其中与收入相关程度最高的是文化程度属性为 32.6%，其次是职业属性为 30%，继而是区域分布为 24.4%，区域分布之所以和收入有较高的相关程度原因在于浦东新区开发的历史相对较短，区域内部的经济发展水平和经济结构差异较大，因此计算的结果符合实际情况的预期。最后将列联系数归一化得到实际所需要的权重。

表 1 列联系数和权重

	年龄	区域分布	家庭人口规模	户口状况	性别	文化程度	行业	职业	合计
与收入分组的列联系数	0.186	0.244	0.1	0.157	0.057	0.326	0.215	0.3	1.585
城镇居民各属性权重	0.117	0.154	0.063	0.099	0.036	0.206	0.136	0.189	

四、评估结果

（一）样本—总体年龄分布的一致性检验

检验之前须对年龄变量加权，权数为每个年龄的人数并做归一化处理，检验的年龄区间界定在 1～100 之间。从样本和总体的描述统计量来看，两

表 2　　事后评估方法及说明

评估层次	指标或检验	指标或假设检验说明	计算公式
属性水平	平均数代表性检验系数	适用于连续型特点的人口属性,比较样本—总体属性中心位置的偏离程度,经验要求控制在 3% 以内。	$\frac{\lvert \bar{x} - \bar{X} \rvert}{\bar{X}} \times 100\%$
	结构代表性检验的差异率	适用于离散型特点的人口属性,比较样本—总体属性水平之间的差异程度,经验要求控制在 5% 以内。	$\frac{\lvert p - P \rvert}{P} \times 100\%$
	样本—总体偏离指数	度量样本—总体属性水平上的差异程度,该指标采用作商处理,理想值为 1。	$SPDI_i = \frac{p_i}{P_i}$
属性整体	偏离指数	该指标对属性的每个水平的结构差异率做加权平均,用来反映整体属性的差异度,DI 的取值范围为 0 ~ 100%,经验认为当 DI 值 < 10% 时,认为样本的分布与总体的分布统计上无显著差异。	$DI = \sum_{i=1}^{m} \lvert p_i - P_i \rvert = \sum_{i=1}^{m} \left\lvert \frac{p_i - P_i}{P_i} P_i \right\rvert$
	基尼集中比	常用于社会学和人类地理学中测量人群的居住和地理分布,介于 0 ~ 1 之间,值越小表明样本构成与总体构成的相似度越高。	$GCR = \left\lvert 1 - \sum_{i=1}^{m} (P'_i + P'_{i-1})(p'_i - p'_{i-1}) \right\rvert$
分布一致性检验	*Kolmogorov–Smirnov* 检验	基本思想是检验两个样本是否来自于同一个总体,检验的思路是计算来自两个独立总体样本的累积频数之间的差值,并计算最大差值,当两个独立总体分布相同时,最大差值应当较小。	$D_N = \max\{\max_i(\lvert F_m(x_i) - G_n(x_i) \rvert), \max_j(\lvert F_m(y_j) - G_n(y_j) \rvert)\}$
	Wald–Wolfowitz 游程检验	检验的思路是把两个独立样本的观测值混合,然后给每个观测值进行评分并按照升序进行排列,然后确定排列中的游程数量。	$Z = \frac{r - \left(\frac{2mn}{m+n} + 1\right)}{\sqrt{\frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)}}} \sim N(0,1)$

者的均值、标准差、峰度以及中位数差别不大,但是众数和偏度存在较为明显的差异,样本呈现左偏倾向,总体呈现轻微右偏,见表 3。计算的平均数代表性检验系数为 8.15%,大于 3% 的临界值,说明样本和总体之间的年龄结构存在比较明显的差异。

表 3　　城镇居民样本和总体的描述统计量

	均值	标准差	众数	中位数	偏度	峰度	平均数检验系数(%)
样本	43.7742	18.35854	58.00	46.0000	-0.284	-0.642	
总体	40.4770	19.81015	28.00	39.0000	0.096	-0.538	8.15

Kolmogorov-Smirnov 检验统计量为 1.677,伴随概率为 0.007,在 95% 的置信水平下检验的结果表明样本和总体的分布之间不一致。Wald-Wolfowitz 游程检验的统计量为 -1.575,单尾概率为 0.058,在 95% 的置信水平下检验的结果表明样本和总体之间的分布是一致的,这与 Kolmogorov-Smirnov 检验的结果正好相反,但是当我们放宽显著性水平时,如在 10% 时,两个检验的结构均可拒绝原假设,认为样本和总体的分布存在差异。

(二) 样本—总体各属性的结构检验

从每个大类属性下的细分水平看,结构差异率和 SPDI 指标值显示有些水平有抽样不足和过度抽样的情况,以地域分布属性为例,没有抽选到的街道或乡镇其结构差异率均为 100%,SPDI 指标值为 0,如申港街道、高桥镇等,而有些街道或乡镇结构差异率和 SPDI 指标值明显偏大,如新场镇,其结构差异率达到 490.657%,SPDI 指标值 5.907 大于 2,存在明显的过度抽样的现象。

(三) 整体差异率

本文分别采用 GCR 和 DI 两个指标经加权计算得到整体差异率,结果如表 4,城镇居民 DI 计算的整体差异率为 4.302%,GCR 计算的整体差异率为 9.350%。从人口属性的整体来看,户口状况的 DI 值最大,为 11.691%,超过了经验值 10%,对应的 GCR 值为 0.208,在各个属性的 GCR 值中也是最大的,因此认为户口状况属性的代表性效果相对较差。代表性最好的为性别属性,DI 值仅为 0.068%,GCR 接近于 0,为 0.001。

表 4　　DI、GCR 及整体差异率计算结果表

	年龄	区域分布	家庭人口规模	户口状况	性别	文化程度	行业	职业	整体差异率(%)
DI	2.700	2.025	8.017	11.691	0.068	3.854	1.803	5.018	4.302
GCR	0.112	0.148	0.0828	0.208	0.001	0.089	0.097	0.001	9.350

五、结束语

本文认为单纯从抽样设计等事前保证样本代表性的做法是不完全的,样本数据采集上来之后,须同

步做好样本事后评估的配套工作，此环节不可或缺。

本文建议实施抽样调查之前根据当地经济社会的实际发展情况制定不同的抽样设计，然后实施预调查，根据总体目标估计量的优良性准则优选适合当地的抽样方法；在正式调查完成后，对采集上来的样本数据进一步检验样本的分布是否和总体的分布基本一致，如果差异度太大，出现抽样不足或者过度抽样则须重新设计抽样方法实施新一轮的抽样计划，直至差异率降到可以接受的程度，从而保证抽样数据从调查之初到调查之后的准确性和一致性。

然而样本代表性评估的方法论研究不是一蹴而就的，本文所提出的事后评估体系是一个尝试性、阶段性的探索研究成果，因而需要实时调整更新，不断充实完善，特别是指标经验值的规定需要根据实地情况和调查难度量力谨慎选择，同时分布的假设检验须配套使用，切忌囿于单个指标超界而武断地妄下结论。

参考文献

[1] 冯士雍. 关于样本对总体代表性问题的认识与讨论——兼论抽样调查中辅助变量的作用 [J]. 统计研究，2001(9)：30－33.

[2] 李文华. 社会调查研究中样本的代表性问题探讨 [J]. 统计与决策，2006(9)：157－159.

[3] 王萍萍. 农村住户调查县级样本代表性评估方法研究 [J]. 统计研究，2011(2)：71－75.

[4] 游正林. 应该如何评估样本的代表性? [J]. 华中师范大学学报(人文社会科学版)，2009(3)：45－49.

[5] 曹阳，陈洁，曹建文，蒋锋，钱军程. 世界健康调查项目中国预调查抽样效度分析 [J]. 中国公共卫生，2006(1)：47－49.

[6] 王亚雄，张鸿武，王芳. 我国抽样调查方法改革刍议 [J]. 统计与信息论坛，2005(3)：32－35.

[7] 俞纯权，王曰人. 论样本的代表性 [J]. 统计与信息论坛，2003(2)：12－14.

[8] 胡英. 从抽样误差看“五普”长表抽样方法 [J]. 统计研究，2004(9)：47－52.

[9] 詹绍康. 现场调查技术 [M]. 上海：复旦大学出版社，2003. 63－64.

[10] Christopher Berry. Land use regulation and residential segregation: dose zoning matter? [J]. American Law and Economic Review，2001，3(2)：251－274.

[11] Barrettl，Schneider A. The 2002 Alberta survey sampling report. University of Alberta technical report [EB]. 2003. http://www.ualberta.ca/PRL.

[12] 吴喜之. 非参数统计 [M]. 北京：中国统计出版社，1999. 120－128.

[13] 刘顺忠，荣丽敏，景丽芳. 非参数统计与 SPSS 软件应用 [M]. 武汉：武汉大学出版社，2008. 57－125.

作者简介

宋子轩，男，1986 年生，山东潍坊人，东华大学管理学院产业经济学硕士，国家统计局高级调查分析师，国家统计局浦东调查队实习生。研究方向为经济社会统计，产业经济，金融中介。

冷燮，男，1981 年生，上海人，华东师范大学理工学院统计学学士，中级统计师职称，上海市浦东新区社会统计调查中心住户调查科科长。研究方向为住户调查、社会调查。

陈瑶瑶，女，1986 年生，安徽池州人，中央财经大学财政学院财政学学士，国家统计局浦东调查队住户与社会调查处科员。研究方向为住户调查、社会调查。

（责任编辑：何锦义）