·基础理论与方法·

不同复杂抽样调查的设计效率比较分析

王建生 冯国双 于石成 马林茂 周脉耕 刘诗瑶

【导读】 为比较常用的几种复杂抽样调查的设计效率，以“2002年中国居民营养与健康状况调查”的数据为依据构造抽样总体，利用统计模拟的方法，估计出不同复杂抽样调查的设计效率值。结果显示，不同复杂抽样的设计效率值差别较大，样本量越大、抽样阶段数越多、分层数越少，复杂抽样的设计效率值越大。实际中采用复杂抽样时，可尽量减少抽样阶段数，细化分层类别，以降低设计效率值，提高设计效率。

【关键词】 复杂抽样调查；设计效率

Comparison of the Designing Effects (DE) among different designs related to complex sampling methods

WANG Jian-sheng[1], *FENG Guo-shuang*[2], *YU Shi-cheng*[2], *MA Lin-mao*[2], *ZHOU Mai-geng*[2], *LIU Shi-yao*[2]. 1 *Policy Research Center for Environment and Economy, Ministry of Environment, P.R. China, Beijing* 100029, China; 2 *National Center for Public Health Surveillance and Information Services, Chinese Center for Disease Control and Prevention*

Corresponding authors: ZHOU Mai-geng, Email: maigengzhou@126.com; *LIU Shi-yao*, Email: liushiyao641@163.com

【Introduction】 To compare the designing effects (DE) among different complex sampling designing programs. Data from the '2002 Chinese Nutrition and Health Survey' was used as an example to generate the sampling population, and statistical simulation method was used to estimate the values of DEs from six complex sampling designing programs. It was found that the values of DEs varied among the six complex sampling designing programs. The values of the DEs were associated with the sample sizes in a positive way, with more sample stages and less stratified categories. Reduction of the numbers of sample stages and detailing stratified categories could decrease the DE values so as to improve the DE.

【Key words】 Complex sampling design; Design effect

大型流行病学研究中，经常采用复杂抽样调查，即综合应用分层、整群、不等概率等多种抽样方式进行调查[1]，而非简单地采用某一种抽样方式。复杂抽样调查通常采用设计效率(design effect)这一指标来评价该抽样设计的好坏。但从目前国内应用情况看，不少复杂抽样的流行病学调查中，均未考虑设计效率这一问题。对于不同复杂抽样调查，设计效率是否有差别，差别有多大，这一问题尚无明确的答案。本研究以“2002年中国居民营养与健康状况调查”(Chinese Nutrition and Health Survey, CNHS2002)数据为依据，利用统计模拟的方法，对几种常见复杂抽样设计及指标的设计效率值进行模拟估算并比较，为合理选择复杂抽样提供参考。

基本原理

设计效率的概念首先由Kish和Frankel[2,3]提出，是指对于同一目标量，在调查单位相同时，所考虑的抽样设计估计量方差与完全随机抽样设计(不放回)估计量方差的比值。计算公式

$$deff = V(\hat{\theta})/V_{SRS}(\hat{\theta})$$

式中，$V(\hat{\theta})$为所考虑的抽样设计的估计量方差，$V_{SRS}(\hat{\theta})$为相同样本量的完全随机抽样设计的估计量方差。

从计算公式可见，设计效率值越大，所考虑的抽样设计的误差越大，即该设计的效率较低。如果设计效率值>1，表明所考虑的抽样设计效率不如完全随机抽样；反之，所考虑的抽样设计效率高于完全随机抽样。

统计分析主要采用统计模拟技术，利用SAS 9.2统计软件自行编程，基于构造的总体对6种复杂抽样重复抽取100次，并采用完全随机抽样方法抽取

DOI: 10.3760/cma.j.issn.0254-6450.2012.10.020

作者单位：100029 北京，环境保护部环境与经济政策研究中心(王建生)；中国疾病预防控制中心公共卫生监测与信息服务中心(冯国双、于石成、马林茂、周脉耕、刘诗瑶)

王建生、冯国双同为第一作者

通信作者：周脉耕，Email: maigengzhou@126.com；刘诗瑶，Email: liushiyao641@163.com

与之样本量相同的100份样本。分别计算出身高、体重、BMI、超重率、肥胖率、收缩压、舒张压、高血压患病率共8个指标的复杂抽样设计方差和完全随机抽样方差，计算出不同复杂抽样、不同指标的设计效率。

复杂抽样的方差估计采用泰勒级数线性化法[4,5]，令$Y=(Y_1, Y_2, \cdots, Yp)$表示总体参数的一个p维向量，相应的估计值向量用$\hat{Y}=(\hat{Y}_1, \hat{Y}_2, \cdots, \hat{Y}p)$表示。由于要估计的不是参数本身，考虑$Y$的函数形式$\theta=f(Y)$，相应的估计量为$\hat{\theta}=f(\hat{Y})$。对于函数$f(Y)$，假定在含有$Y$和$\hat{Y}$的开放空间存在连续的二阶导数，使用泰勒展开式的线性项，得

$$\hat{\theta}-\theta \approx \sum_{j=1}^{p} \frac{\partial f(Y)}{\partial y_j}(\hat{Y}_j - Y_j)$$

则$\hat{\theta}$的近似方差可写成

$$V(\hat{\theta}) \approx V\Big(\sum_{j=1}^{p} \frac{\partial f(Y)}{\partial y_j}(\hat{Y}_j - Y_j)\Big)$$

$$= \sum_{j=1}^{p}\sum_{l=1}^{p} \frac{\partial f(Y)}{\partial y_j}\frac{\partial f(Y)}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l)$$

式中，$V(\hat{Y}_j, \hat{Y}_l)$表示估计值$\hat{Y}_j$和$\hat{Y}_l$的方差与协方差，因此将非线性估计值$\hat{\theta}$的方差简化成p个线性估计值$\hat{Y}_j$的方差与协方差的函数。使用方差与协方差的估计值$\hat{v}(\hat{Y}_j, \hat{Y}_l)$代替$V(\hat{Y}_j, \hat{Y}_l)$，便可得到方差估计值$\hat{v}(\hat{\theta})$。

实例分析

1. 总体构造：本研究以CNHS2002的数据为依据构造总体。CNHS2002采用多阶段分层整群随机抽样方法，通过样本估计总体[6]。样本县(市、区)的抽取是按经济发展水平及类型将全国各县/区划分为大城市、中小城市、一类农村、二类农村、三类农村、四类农村共6类地区。抽样共分4个阶段：第一阶段在6类地区抽取，每类地区抽取22个调查县/区，共132个调查县/区；第二阶段从抽到的样本县/区中抽取3个乡/街道；第三阶段采取整群随机抽样方法从样本乡镇/街道中抽取2个村/居委会；第四阶段根据所需要的样本例数，采用整群抽样法，在抽中的村中随机抽取90户家庭为调查样本户，被抽中的样本户中的人口资料全部纳入。

构造总体的基本方法：

第一步，如果(省编码+00+县编码)是11的倍数或者被11除余数为10，则县编码加10，生成新的变量，加入到原始数据库中。

第二步，如果(省编码+00+县编码+地区类型编码+00+街道编码)是11的倍数或者被11除余数为10，则街道编码加10，生成新的变量，加入到数据库中。

第三步，如果(省编码+00+县编码+地区类型编码+00+街道编码+00+居委会编码)是11的倍数或者被11除余数为10，则居委会编码加10，生成新的变量，加入到数据库中。

经过上述步骤构造出的总体的样本含量为372 502。每层的县/区数、乡/街道数、村/居委会数、户数和人数不相等，不同分层、不同性别和不同年龄组的比例和CNHS2002数据库的基本情况类似。

2. 复杂抽样设计：选择较为常用的6种复杂抽样调查设计，分别以M1～M6表示(表1)。M1～M3均为四阶段分层整群抽样，但最后阶段抽样比例不同；M4为多阶段分层随机抽样；M5为二阶段分层整群抽样；M6也为四阶段分层整群抽样，但分层仅为两层。

M1采用四阶段分层整群抽样方法。样本县(市、区)的抽取是按经济发展水平及类型将全国各县/区划分为大城市、中小城市、一类农村、二类农村、三类农村、四类农村共6类地区。抽样为四阶段抽样，第一阶段在6类地区抽取，每类地区抽取15个调查县/区；第二阶段从抽到的样本县/区中抽取2个乡/街道；第三阶段采取整群随机抽样方法从样本乡

表1 复杂抽样设计中M1～M6基本情况

抽样设计	M1～M3	M4	M5	M6
抽样阶段	4个阶段	4个阶段	2个阶段	4个阶段
分层方法	6类经济地区	6类经济地区	6类经济地区	城乡
第一阶段	有放回抽取15个调查县/区	有放回抽取15个调查县/区	有放回抽取30个村/居委会	有放回抽取45个调查县/区
第二阶段	有放回抽取2个乡/街道	有放回抽取2个乡/街道	有放回抽取40户	有放回抽取2个乡/街道
第三阶段	有放回抽取1个村/居委会	有放回抽取1个村/居委会	-	有放回抽取1个村/居委会
第四阶段	M1有放回抽取40户；M2有放回抽取20户；M3有放回抽取10户	有放回抽取95人	-	有放回抽取40户
最后一阶段抽样方法	整群有放回抽样	完全随机有放回抽样	整群有放回抽样	整群有放回抽样
抽样比例	M1为6.5%；M2为3.3%；M3为1.6%	6.5%	6.5%	6.5%

镇/街道中抽取1个村/居委会;第四阶段采用整群抽样法,在抽中的村中随机抽取40户家庭为调查样本户。抽样比例约为6.5%。

M2和M3也为四阶段分层整群抽样,分层和抽样方法与M1相同,只是在最后一阶段抽取的户数不同。M2的最后一阶段随机抽取20户家庭,抽样的比例约为3.3%。M3在最后一阶段随机抽取10户家庭,抽样的比例约为1.6%。

M4为四阶段分层随机抽样,分层方法及前三阶段的抽样方法、抽样比例与M1相同。但在最后一阶段不是采用整群抽样抽取户,而是随机抽取95人。

M5采用二阶段分层整群抽样方法,第一阶段在6类地区中,采取整群随机抽样方法抽取30个村/居委会;第二阶段采用整群抽样法,在抽中的村中随机抽取40户家庭。抽样比例约为6.5%。

M6为四阶段分层整群抽样,但未按6类地区分层,而是按城市和农村分两层。第一阶段在城市和农村中,分别抽取45个调查县/区;第二阶段从抽到的样本县/区中抽取2个乡/街道;第三阶段采取整群随机抽样方法从样本乡镇/街道中抽取1个村/居委会;第四阶段采用整群抽样法,在抽中的村中随机抽取40户家庭。抽样比例约为6.5%。

3. 统计模拟结果:通过重复抽样及计算,分别求得身高、体重、BMI、超重率、肥胖率、收缩压、舒张压、高血压患病率8个指标的设计效率值(表2)。

根据调查设计的类型,分别比较M1~M3(主要反映样本量的差异)、M1和M4(主要反映最后阶段抽样方式的差异)、M1和M5(主要反映抽样阶段数的差异)、M1和M6(主要反映分层方式的差异)。由于指标较多,以8个指标的设计效率值的均值来综合反映不同复杂抽样的设计效率。

M1、M2、M3的8个指标的设计效率值均值分别为11.67、5.17、3.74。除BMI外,复杂抽样M1的其余7个指标设计效率值均远远高于M2和M3。

M1和M4的设计效率值均值分别为11.67和11.56,二者差别很小。从8个指标的值来看,M1和M4两种复杂抽样中8个指标的设计效率值各有高低,而且各指标的值差别均较小,可以认为M1和M4的设计效率值差别不大。

M1和M5的设计效率值均值分别为11.67和6.60,除BMI之外,复杂抽样M1中的其余指标设计效率值均高于M5。

M1和M6的设计效率值均值分别为11.67和20.50,复杂抽样M6有5个指标的设计效率值远远高于M1,其余3个指标则显示M1高于M6。总的来看,M6中指标的设计效率值要高于M1。

讨 论

本研究模拟比较的6种复杂抽样均为多阶段分层整群抽样和多阶段分层随机抽样,这在大型流行病学调查中应用十分广泛。但这两种复杂抽样的效率是否一致,不同抽样阶段、分层方式是否会影响到复杂抽样的效率,以往的研究尚无确切结论。

本次模拟结果显示,在抽样阶段、分层方式和样

表2 复杂抽样设计M1~M6模拟抽样的设计效率值

抽样设计	设计效率	身高	体重	BMI	超重率	肥胖率	收缩压	舒张压	高血压患病率
M1	$V(\hat{\theta})$	5.03	6.28	0.06	0.68	0.47	0.11	2.53	0.82
	$V_{SRS}(\hat{\theta})$	0.22	0.28	0.02	0.11	0.06	0.01	0.22	0.11
	deff	23.08	22.81	2.80	5.93	8.21	11.77	11.60	7.19
M2	$V(\hat{\theta})$	6.11	7.41	0.06	0.72	0.51	0.14	2.77	0.93
	$V_{SRS}(\hat{\theta})$	0.62	0.71	0.05	0.35	0.15	0.02	0.62	0.25
	deff	9.86	10.45	1.29	2.06	3.30	6.32	4.47	3.65
M3	$V(\hat{\theta})$	5.74	7.05	0.17	1.02	0.69	0.16	3.60	1.10
	$V_{SRS}(\hat{\theta})$	1.03	1.23	0.05	0.60	0.21	0.04	1.03	0.41
	deff	5.59	5.75	3.17	1.69	3.26	4.30	3.51	2.64
M4	$V(\hat{\theta})$	5.38	6.57	0.08	0.71	0.43	0.13	2.61	0.86
	$V_{SRS}(\hat{\theta})$	0.26	0.29	0.02	0.10	0.06	0.01	0.26	0.11
	deff	20.61	22.36	3.19	7.00	6.76	14.76	10.01	7.77
M5	$V(\hat{\theta})$	3.12	4.11	0.06	0.75	0.33	0.08	1.92	0.51
	$V_{SRS}(\hat{\theta})$	0.32	0.37	0.02	0.11	0.06	0.01	0.32	0.14
	deff	9.80	10.97	3.61	6.76	5.62	6.31	6.05	3.70
M6	$V(\hat{\theta})$	11.54	13.87	0.05	1.34	0.47	0.14	1.99	0.69
	$V_{SRS}(\hat{\theta})$	0.19	0.29	0.02	0.13	0.05	0.01	0.19	0.11
	deff	60.02	48.63	2.35	10.43	9.16	16.72	10.37	6.33

本量相同的情况下，多阶段分层整群抽样与多阶段分层随机抽样的设计效率并无差异。实际应用时，可根据研究目的等选择较为适合的抽样设计。

采用复杂抽样设计时，设计效率作为方差校正因子，在调查设计中用途很广。实际中最常见的用途是计算复杂抽样设计的样本含量，复杂抽样的样本量等于完全随机抽样样本量与设计效率值的乘积[7]。本文模拟结果验证了样本量与设计效率值的正比关系。M1、M2和M3三种抽样均为四阶段整群随机抽样，分层方式也完全一致，只是样本量不同，M1样本量最大，M2次之，M3最少。这三种复杂抽样设计的设计效率比较结果显示，样本量越大，设计效率值越大。

对于同一种复杂抽样设计，不同的抽样阶段数会影响到复杂抽样的设计效率。从M1（四阶段分层整群抽样）和M5（二阶段分层整群抽样）的比较结果可见，抽样阶段数越多，设计效率值越大，四阶段分层抽样的设计效率值约为二阶段设计效率值的2倍。这提示在进行复杂抽样设计时，如无必要，应尽量简化抽样阶段数，以提高设计效率。如果采用多阶段分层抽样，需要更大的样本量来保证结果的精度。

本文结果还显示，不同的分层方式也会产生不同的设计效率。M1和M6均为四阶段分层整群抽样，但M1分层方式较为详细，以6类经济地区划分层次，而M6仅以城市和农村分两层。M1的设计效率值总的来看要低于M6，提示分层的数目越多，设计效率值越低。因此在实际抽样时，应尽可能地获取分层因素的详细信息，细化分层因素。

本研究中估计的设计效率值均较大，这可能跟样本的聚集性有关。当抽样样本具有较明显的聚集性时，通常会导致方差变大，从而产生一个很大的设计效率值。可以看出，对于大规模的复杂抽样调查，其精度的损失要比完全随机抽样大得多，有的甚至可达几十倍。实际上也就是说，复杂抽样设计比完全随机抽样需要更大的样本量才能保证足够的精确度。即使同一种复杂抽样，不同的抽样阶段数和分层数所需的样本量也差别较大，抽样阶段数越多、分层数目越少，所需样本量越大。

本次模拟共选择了8个指标，其中身高、体重、收缩压、舒张压、BMI为连续变量，肥胖率、超重率、高血压患病率为分类变量。除BMI外，其余连续变量的设计效率值均要高于分类变量，尤其是身高、体重这两个指标，其设计效率值远远高于其他指标。提示如果复杂抽样的结局变量为连续变量，可能需要比分类变量更多的样本量。但由于BMI的结果并不符合这一规律，因此这一结论尚需进一步验证。

本研究模拟次数为100次，主要是考虑到每次模拟抽样的例数将近40万，如果模拟次数过多，普通计算机配置无法完成，因此最终共模拟了100次。尽管模拟次数不多，但由于本研究是在大规模实际调查数据的基础上进行的模拟，每次模拟的例数非常多，完全可以弥补模拟次数不足可能带来的不稳定性。因此可以认为这一基于大规模数据的模拟结果还是比较可靠。

参 考 文 献

[1] Lv J, He PP, Li LM. Data analysis from surveys using complex sampling methods. Chin J Epidemiol, 2008, 29(8): 832-835. (in Chinese)
吕筠，何平平，李立明. 复杂抽样调查数据实例分析. 中华流行病学杂志，2008，29(8)：832-835.

[2] Kish L. Survey sampling. New York: John Wiley & Sons, Inc. 1965.

[3] Kish L, Frankel MR. Inference from complex samples. J Royal Stat Soc Ser B, 1974, 36: 1-37.

[4] Jin YJ, Du ZF, Jiang Y. Sampling technique. Beijing: Chinese Renmin University Press, 2008. (in Chinese)
金勇进，杜子芳，蒋妍. 抽样技术. 北京：中国人民大学出版社，2008.

[5] Liu JH, Jin SG. Estimation of population quantities and their variances in complex sample survey. Chin J Health Stat, 2008, 25(4): 377-379. (in Chinese)
刘建华，金水高. 复杂抽样调查总体特征量及其方差的估计. 中国卫生统计，2008，25(4)：377-379.

[6] Wang LD. Chinese Nutrition and Health Survey Report—2002 comprehensive report. Beijing: People's Medical Publishing House, 2005. (in Chinese)
王陇德. 中国居民营养与健康状况调查报告之一：2002综合报告. 北京：人民卫生出版社，2005.

[7] Lohr SL. Sampling: Design and Analysis. Pacific Grove, CA: Duxbury Press, 1999: 239-241.

（收稿日期：2012-05-23）
（本文编辑：张林东）