·基础理论与方法·

多阶段抽样调查资料的加权估计法

侯晓艳 魏永越 陈峰

【导读】 多阶段抽样技术广泛应用于流行病学现况调查中，但针对其所得资料的统计分析方法往往选择不当。文中介绍一种用于多阶段抽样调查资料的统计分析方法——加权估计法，以推广针对此类资料的恰当的分析方法。在介绍加权估计法基本原理的基础上通过两个二阶段分层整群抽样的实际调查资料实现这种算法。加权估计法可以校正由多阶段抽样导致的三种效应：群效应、分层效应、不等概率性，给出无偏点估计和比较客观的误差估计，并作出正确的统计推断。

【关键词】 多阶段抽样；复杂调查资料；泰勒线性估计法；加权点估计

Weighted estimation methods for multistage sampling survey data *HOU Xiao-yan*, WEI Yong-yue, CHEN Feng. *Nantong Center for Disease Control and Prevention, Nantong* 226009, China
Corresponding author: CHEN Feng, Email: fengchen@njmu.edu.cn; *Department of Epidemiology and Biostatistics, School of Public Health, Nanjing Medical University, Nanjing* 210029, China

【Introduction】 Multistage sampling techniques are widely applied in the cross-sectional study of epidemiology, while methods based on independent assumption are still used to analyze such complex survey data. This paper aims to introduce the application of weighted estimation methods for the complex survey data. A brief overview of basic theory is described, and then a practical analysis is illustrated to apply to the weighted estimation algorithm in a stratified two-stage clustered sampling data. For multistage sampling survey data, weighted estimation method can be used to obtain unbiased point estimation and more reasonable variance estimation, and so make proper statistical inference by correcting the clustering, stratification and unequal probability effects.

【Key words】 Multistage sampling; Complex survey data; Linearized/Robust variance estimation; Weighted point estimation

多阶段抽样也称多级抽样、多阶抽样、套抽样，是在抽取样本时，分为两个或两个以上的阶段从总体中抽取样本的一种抽样调查方法，属于复杂抽样的一种。随着复杂抽样技术尤其是多阶段抽样在大型流行病学现况调查中不断推广应用，针对其所得样本，即复杂样本的统计分析方法也日益受到关注。目前适用的统计分析方法大致分为两类：一是随机化推断模式[1]，本文中称之为加权估计法；二是基于模型的分析模式[2]。尽管国内大型的多阶段抽样调查比比皆是，然而，对调查研究所得资料的分析仍然沿用传统的统计分析方法[3-5]。事实上，对于多阶段抽样调查资料如果采用传统的分析方法，所得参数估计的误差偏小，估计的可信区间偏窄，导致区间覆盖率降低。大量的统计学模拟研究结果表明，当样本量含量足够大时，复杂样本的数据结构并不会影响到总体均数、比值以及回归系数等一阶总体参数的点估计(point estimation)，但可以影响总体参数估计的准确度和精度，即误差，从而影响统计推断的结论；而当群数或者群含量不够大时，点估计和区间估计皆受到影响。为此本文介绍总体均数和总体率的加权估计方法，即Horvitz-Thompson点估计、Taylor线性化及再抽样两种常见的误差估计法，并通过实例展示其估计结果，揭示其与传统方法之间的差异，进一步说明传统方法在处理复杂样本时的不合理性。

基本原理

假设某复杂样本有三个水平，分别为区县、街道和居民。现假定某定量变量X的总体均数、样本均数、总体方差，样本方差分别为μ_x、$\bar{X}$、σ_x^2、s_x^2；二分类变量Y的总体率、样本率、总体方差、样本方差分别为ρ、p、σ_p^2、s_p^2。x_{hji}表示第h个区县(层)第j个街道(群)中第i个居民(个体)的测得值，y_{hji}表示上述个体发生所关心的结局事件(例如，患某病)，$y_{hji}=1$表示患病，$y_{hji}=0$表示未患病。

1. 总体均数和总体率的Horvitz-Thompson点估计：复杂样本的点估计与样本含量和抽样方案有关。当样本含量足够大时，无论何种抽样方案，使用

DOI:10.3760/cma.j.issn.0254-6450.2009.06.024
作者单位：226009 南通市疾病预防控制中心(侯晓艳)；南京医科大学公共卫生学院流行病与卫生统计学系(魏永越、陈峰)
通信作者：陈峰，Email：fengchen@njmu.edu.cn

传统方法也可以得到近似无偏的点估计值；当样本含量不够大，且为复杂抽样时，使用考虑抽样设计的加权估计即Horvitz-Thompson估计，可以得到近似无偏的点估计值。

该方法中权重定义为抽样概率的倒数，表示抽中的个体所能代表的同一层同一群中的个体数。如表1所示，不同抽样方案的权重和抽样概率以及样本含量之间的关系可表示为：

$$w_i = \frac{1}{f_i} = \frac{\hat{N}}{n},\ w_h = \frac{1}{f_h} = \frac{\hat{N}_h}{n_h},\ w_{hj} = \frac{1}{f_{hj}} = \frac{\hat{N}_h}{n_{hj}},$$

$$w_{hji} = w_h \times w_{hj} \times w_i = \frac{1}{f_h} \times \frac{1}{f_{hj}} \times \frac{1}{f_i}$$

表1　各种抽样方案的加权点估计值

抽样方案	总体参数估计值($\bar{X}$)
(1)简单随机抽样或者自加权抽样	$\bar{x} = \sum_{i=1}^{n} x_i / n$
(2)分层随机抽样	$\bar{x}_{str} = \sum_{h=1}^{H}\sum_{i=1}^{n_h} w_h x_{hi} \Big/ \sum_{h=1}^{H} w_h n_h$
(3)分层整群抽样(层间不等权重)	同上
(4)分层整群抽样(群间不等权重)	$\bar{x}_{clu} = \sum_{h=1}^{H}\sum_{j=1}^{m_h}\sum_{i=1}^{n_{hj}} \frac{w_{hj}}{m_h} x_{hj} \Big/ \sum_{h=1}^{H}\sum_{j=1}^{m_h} w_{hj} n_{hj}$
(5)多阶段分层整群抽样	$\bar{x}_{clu} = \sum_{h=1}^{H}\sum_{j=1}^{m_h}\sum_{i=1}^{n_{hj}} w_{hji} x_{hji} \Big/ \sum_{h=1}^{H}\sum_{j=1}^{m_h} w_{hji}$

注：当结果变量为二分类(例如患病与未患病)时，只需将式中变量x用y表示，$y=1$表示患病，$y=0$表示未患病，即可得到率的估计

n_h表示层样本含量，$\sum_{h=1}^{H} n_h = n$；n_{hj}表示第h层第j群中的个体数即群含量，m_h表示第h层的群数；f为抽样比；分母$\hat{N}$为总样本含量的估计值，$\sum_{h=1}^{H} w_h n_h = \hat{N}$或$\sum_{h=1}^{H}\sum_{j=1}^{m_h} w_{hj} n_{hj} = \hat{N}$或$\sum_{h=1}^{H}\sum_{j=1}^{m_h}\sum_{i=1}^{n_{hj}} w_{hji} = \hat{N}$；方案(3)同方案(2)的估计公式一致，但此时只要反应变量间存在群内相关性，则方案(3)和(4)一样，设计效应DEFF(design effects)往往都>1，即传统方法将低估误差。

从表1公式中可以看出，在复杂抽样中，不管是均数还是率，均已转换成两个随机变量的比值，分母不再是固定的样本含量n，所以必须使用Horvitz-Thompson估计法。此法还适用于后分层估计(poststratification)和有信息抽样(informative probability sampling)设计，只需在原有权重的基础上加以校正，校正后的权重可用通式表示成：$w_k^* = g_k w_k$，其中w_k表示原来样本权重，g_k即为校正系数(表2)。

假设后分层变量s(通常是性别、年龄、种族等人口学特征)将原样本分成L层；辅助信息变量为z，则校正系数g_k见表2。

表2　三种校正系数

	后分层估计	有信息设计(z为定性资料)	有信息设计(z为定量资料)
g_k	$N_s/\hat{N}_s$	$T_z/\hat{t}_z$	$\frac{N}{\hat{N}}\left[1 + \frac{\bar{Z} - \bar{z}}{\frac{n-1}{n} s_z^2}\right]$

注：式中g_k表示后分层中第k层的总体含量(一般来自普查数据)，N_s为后分层的第g层的实际总人数，$\hat{N}_s$为加权估计的总人数；T_z表示研究总体中z为定性变量时，所有阳性结果($y_{hji}=1$)的总和，$\hat{t}_z$为相应样本估计得到的阳性结果的总和；$\hat{N} = \sum_{k=1}^{n} w_k$，$s_z^2$是$z$为定量变量时的样本方差

在选择辅助变量时，通常选择与分析变量相关性较强，且关系稳定的变量，即满足

$$\hat{r} = \sum_{i=1}^{H}\sum_{j=1}^{m_h}\sum_{i=1}^{n_{hj}} x_{hji} \Big/ \sum_{i=1}^{H}\sum_{j=1}^{m_h}\sum_{i=1}^{n_{hj}} z_{hji} \approx R = \frac{T_x}{T_z} \quad (1)$$

从上述总体均数和总体率的估计来看，复杂抽样中的点估计并不满足线性，同样，基于此类样本估计的相关系数、回归系数等也是非线性的，被称作为比估计值，其误差估计均无显式表达。因此，通常借助于特定的线性化法或再抽样法(resampling)估计其误差。

2. 泰勒线性化估计误差法：假设θ为点估计值(总体均数或者总体率)，如表1所示，在复杂样本中点估计值并非线性估计值，其分子分母均为变量，表达式可表示为：$\theta = f(Y) = Y_1/Y_2 = Y/N$，为了便于描述现设样本为分层抽样所得，其估计值为：

$$\hat{\theta} = \hat{Y}_1/\hat{Y}_2 = y/n = \Big(\sum_{h=1}^{H}\sum_{i=1}^{m_h} y_{hi}\Big) \Big/ \Big(\sum_{h=1}^{H}\sum_{i=1}^{m_h} n_{hi}\Big) \quad (2)$$

根据泰勒展开式，获得其线性部分为：

$$\hat{\theta} - \theta \approx \sum_{j=1}^{s} \frac{\partial f(Y)}{\partial y_j}(\hat{Y}_j - Y_j) \quad (3)$$

其中$\partial f(Y)/\partial y_i$为$f(Y)$对$y_j$的偏导数。由式(3)可得$\theta$的方差：

$$\begin{aligned} V(\hat{\theta}) &= V\Big(\sum_{j=1}^{s} \frac{\partial f(Y)}{\partial y_j}(\hat{Y}_j - Y_j)\Big) \\ &= \sum_{j=1}^{s}\sum_{l=1}^{s} \frac{\partial f(Y)}{\partial y_j} \frac{\partial f(Y)}{\partial y_l} V(\hat{Y}_j, \hat{Y}_l) \end{aligned} \quad (4)$$

此时，泰勒线性法将计算一个非线性估计值的方差转变成了与之有关系的参数Y_j的方差估计上。显然简单了很多，式中$V(\hat{Y}_j, \hat{Y}_l)$的是方差协方差估计值。

假设需要探索Y和X的关系，模拟回归模型时系数为β，此模型正规方程表示为：

$$G(\beta) = \sum_{j=1}^{M} S(\beta; Y_j, x_j) = 0 \quad (5)$$

若为简单随机抽样，则最小二乘法的常规等式为：

$$G(\beta) = X'Y - X'X\beta = 0 \quad (6)$$

而对复杂样本，其模型等式应为加权表达式：

$$G(\beta) = \sum_{j=1}^{M} w_j S(\beta; Y_j, x_j) = 0 \quad (7)$$

根据泰勒展开式，获得其线性部分为：

$$\hat{\beta} - \beta \approx -\Big\{\frac{\partial G(\beta)}{\partial \beta}\Big\}^{-1} G(\beta) \quad (8)$$

其中$\partial \hat{G}(\beta)/\partial\beta$为$G(\beta)$对$\beta$的偏导数。由上式可得$\hat{\beta}$的方差：

$$\hat{V}(\beta) = \left[\left\{\frac{\partial \hat{G}(\beta)}{\partial\beta}\right\}^{-1}\hat{V}\{\hat{G}(\beta)\}\left\{\frac{\partial \hat{G}(\beta)}{\partial\beta}\right\}^{-T}\right]\Big|_{\beta=\hat{\beta}} = D\hat{V}\{\hat{G}(\beta)\}\Big|_{\beta=\hat{\beta}}D' \quad (9)$$

其中$D=(X_s'WX_s)^{-1}$，X_s是样本中自变量x的矩阵，W是样本权重的对角阵，$\hat{V}\{\hat{G}(\beta)\}\Big|_{\beta=\hat{\beta}}$是加权误差，$\hat{G}(\beta) = \sum_{j=1}^{m} w_j s_j x_j$，$s_j$为拟似然模型的残差。上述计算均须通过计算机迭代过程来实现。此误差估计法又称为sandwich估计法。

3. 再抽样误差估计法的基本原理：常用的再抽样法包括平衡半样本法（balanced repeated replications，BRR）和刀切法（jackknife repeated replications，JRR）。两种方法有不同：①适用条件不同，即BRR只能用于每层均有两个群的复杂样本，而JRR不限；②再抽样的方式不同，设复杂样本来自两阶段有放回抽样的自加权设计，且每层的群含量均为2。BRR的基本思想：设共有H层，从每层中随机有放回地抽出一个群，由H个群构成一个拟样本，重复这样的工作$K(K\leqslant 2^H)$次，获得K个拟样本；JRR的基本思想：从原有样本中的第1层中取出一个群h_1，该层中剩下的群h_2乘以权重2，合并其他$H-1$层形成第一个拟样本；同样的方法获得剩下的拟样本，共形成H个拟样本。通过改变各层中被排除群的顺序即可获得H个类似的互补拟样本。通过多个拟样本，利用适当的方法计算其点估计值和误差估计值[6]。

实例分析

［**实例1**］

1. 资料来源：江苏省2007年宫颈癌筛查资料。该调查采用分层两阶段抽样方法，抽取全省苏南、苏中、苏北3个区域共11个城市（分别包含3、3、5个城市）；人群为≤65周岁已婚妇女共190 982人。选用宫颈某项细胞学检查结果。

抽样调查的特征变量：分层变量（strata，苏南、苏中、苏北）；初级抽样单位变量（city，地级市）；末级抽样单位（person，市民）；抽样权重为抽样比的倒数。

2. 研究目的：估计江苏省≤65周岁已婚妇女的宫颈某项细胞学检查指标的阳性率。

3. 研究结果：如表3所示，传统方法估计江苏省≤65周岁已婚妇女的宫颈癌某细胞学检查指标的总体阳性率为0.3822%，可信区间为0.3546%～0.4099%；而加权线性化法和加权JRR法估计的总体阳性率为0.4113%，不难看出，传统方法所估计的总体率和标准误均偏小，从而导致可信区间较窄，不包含加权估计所得的总体率。此时，如果按照传统方法下结论或是进一步作假设检验，容易得出错误的统计描述或推断。

表3 传统方法和加权法估计江苏省已婚妇女宫颈癌某指标的阳性率（%）

方法	$\bar{x}$(95%*CI*)	$s_{\bar{x}}$	DEFF
传统分析法	0.3822(0.3546～0.4099)	0.000 141	-
加权线性化法	0.4113(0.3534～0.4692)	0.000 251	2.9473
加权JRR法	0.4113(0.3571～0.4655)	0.000 235	2.5824

［**实例2**］

1. 数据来源：美国1976—1980年全国健康和营养状况第二次调查（NHANES Ⅱ）资料。调查采用分层整群按比例抽样的方法，抽取32个城市64个代表性地段，每个城市有2个地段；人群年龄6月龄至74岁，共27 801人。其中16 563人被请求提供血液样本，但仅收集到10 351人含有血标本的资料，采血比例约62%。本文仅提取分析变量和特征参数。分析变量主要有：血铅含量（lead）、是否为黑人（black：是=1，否=0）、性别（female：女=1，男=0）、年龄（age，age2为年龄的二次项）、身高（height）、体重（weight）、是否患有心脏病（heartatk：是=1，否=0）等。抽样调查的特征变量：分层变量（county，城市）；初级抽样单位变量（location，代表性地段）；抽样权重为抽样比的倒数（pweight）。

2. 研究目的：包括估计总体人群的血铅含量、估计总体人群的心脏病患病率和探讨心脏病发病的影响因素。

3. 研究结果：

（1）估计总体人群的血铅含量：如表4所示，传统方法估计所得总体血铅含量为14.3203 μg/dl，加权法估计总体含量为14.3527 μg/dl，两者接近；而对于标准误，传统方法约是加权法的1/3，故传统方法所得可信区间较窄。如根据传统分析方法的结果作统计推断时易得出拒绝H_0的结论，但很可能违背事实。

表4 传统方法和加权法估计总体人群的血铅含量（μg/dl）

方法	$\bar{x}$(95%*CI*)	$s_{\bar{x}}$	DEFF
传统分析法	14.3203(14.1485～14.4922)	0.0877	-
加权线性化法	14.3527(13.8225～14.8828)	0.2599	8.5719
加权BRR法	14.3527(13.8422～14.8631)	0.2503	7.9457
加权JRR法	14.3527(13.8225～14.8828)	0.2600	8.5726

（2）估计总体人群的心脏病患病率：如表5所示，传统方法估计的总体患病率为4.5995%，95%*CI*：4.1958%～5.0031%，而加权估计法所得总体率为2.9738%，不包含在上述可信区间内，可见传统方法高估了总体率。由此可见，此例中传统方法所估计的总体参数或进一步作统计推断其结果同

[实例1]，将会出现误判的结论。

表5　传统方法和加权法估计总体人群的心脏病患病率(%)

方法	$\bar{x}$(95%CI)	$s_{\bar{x}}$	DEFF
传统分析法	4.5995(4.1958～5.0031)	0.002 059	–
加权估计法	2.9738(2.5968～3.3508)	0.001 848	1.2253
加权BRR法	2.9738(2.6022～3.3454)	0.001 822	1.1905
加权JRR法	2.9738(2.5968～3.3508)	0.001 848	1.2253

(3)探讨影响心脏病发病的因素：如表6所示，模型中各系数相差不大，但传统方法所得各系数的误差项均被低估，从而导致了体重对于心脏病危险性的意义和加权估计法的结论不同：传统分析法由于低估了误差项从而得出体重是心脏病的危险因素，而加权估计法所得结论恰好相反。

表6　传统方法和加权法探讨心脏病发病的危险因素(logistic回归)

变量	OR值($s_{\bar{x}}$)			
	传统分析法	加权线性化法	加权BRR法	加权JRR法
height	0.9922(0.0078)	0.9858(0.1241)	0.9858(0.0120)	0.9858(0.0124)
weight	1.0074(0.0038)*	1.0102(0.0058)	1.0102(0.0058)	1.0102(0.0058)
age	1.4056(0.0771)*	1.4524(0.1101)*	1.4524(0.1144)*	1.4524(0.1110)*
age2	0.9977(0.0005)*	0.9975(0.0006)*	0.9975(0.0007)*	0.9975(0.0006)*
female	0.3932(0.0556)*	0.3786(0.0796)*	0.3786(0.0770)*	0.3786(0.0796)*
black	1.0175(0.1655)	0.9487(0.1671)	0.9487(0.2107)	0.9487(0.1677)

注：*回归系数假设检验的$P<0.05$

4. 结论：在多阶段抽样调查资料分析中，如果使用传统分析法，当分析变量为定量变量时，样本量较大时点估计值相差不大，但误差项将明显被低估；当分析变量为定性变量时，传统方法估计所得误差同样被低估，并且点估计也将有偏，高估或者低估均有可能；拟合logistic回归模型时，传统方法也存在低估系数标准误差得到错误推断的风险。由此可见，传统分析法已不再适合于多阶段抽样调查资料的统计分析。

讨　论

复杂样本存在三个特征：分层效应、群效应和不等概率效应。而传统分析法忽略了这些特征，认为样本是随机的、独立的，往往会低估误差或者得到有偏点估计值，估计准确度达不到预定的水平，甚至作出错误的统计推断。其中，误差估计的偏性可以通过设计效应DEFF来描述，有研究表明$DEFF=1+(n_{cluster}-1)\rho$，($n_{cluster}$为群内样本含量，$\rho$为群内相关系数)。在描述总体特征量时，若DEFF>2则群效应具有考虑价值，表4中血铅含量(lead)的DEFF值分别为8.5719、7.9457、8.5726，表3中细胞学指标的DEFF值分别为2.9473和2.5824，此时使用传统分析法偏差较大；而在回归模型尤其是logistic回归分析中，一般均需考虑设计效应值，即需要使用加权估计的方法作分析。

早在抽样调查方法应用初期，有学者按照调查目的将抽样调查分为描述型调查和分析型调查，现在的调查大多是两者兼顾。两种调查的侧重点不同，描述性调查以描述总体特征量为主要目的；而分析性调查则侧重于探讨变量与变量之间关系，常常通过拟合logistic回归或多重线性回归等模型来探讨变量与变量之间的关系。表3、4、6属于描述型调查的内容，应该用Horvitz-Thompson法和泰勒线性化法估计其总体参数及其误差；而表5则属于分析型调查的内容，此时需要估计回归系数及其标准误，仍然可以使用Horvitz-Thompson法的思想估计其回归系数，而误差估计时有学者认为使用再抽样法如BRR、JRR法效果较泰勒线性法更好[7]。其中BRR法只适合用于每个三水平单位中只有2个二水平抽样单位的资料，因此本文[实例2]符合条件，而[实例1]不符合条件。

综上所述，凡是使用多阶段抽样方案的调查均应采用合适的统计分析方法，如加权估计法，以校正复杂样本的三种效应，得到正确的参数估计及统计推断结果。同时，可以获得各个变量的设计效应，为其他调查提供参考，如为估算有效样本含量(n=DEFF*n)和制定适当的抽样方案等方面提供重要信息，促使在成本效应上达到最佳组合。然而加权估计法也存在着缺陷，其估计效应受到以下因素的影响，如群内相关系数(internal correlation coefficient)、群含量。当群含量很小($n<5$)或者群内相关系数很小，使用加权估计法所得回归系数的估计值及其误差将会有偏，不推荐使用，此时可以考虑使用Pfeffermann等[8]提出的校正权重的方法或者使用加权多水平模型。

参 考 文 献

[1] Demnati A, Rao JNK. Linearization Variance Estimators for Survey Data. Surv Methodol, 2004, 30(1): 4-21.

[2] Tihomir A. General multilevel modeling with sampling weights. Communications in Statistics: Theory and Methods, 2006, 35: 439-460.

[3] 国家"九五"攻关计划糖尿病研究协作组. 中国12个地区中老年人糖尿病患病率调查. 中华内分泌代谢杂志, 2002, 18(4): 280-284.

[4] 孙虹，陈彪，孙菲，等. 北京地区中老年人群原发性震颤的流行病学研究. 中华神经科杂志, 2006, 39(3): 89-92.

[5] 卢伟，刘关霞，李锐，等. 上海15~74岁居民代谢综合征的流行特征. 中华预防医学杂志, 2006, 40(4): 262-268.

[6] Lehtonen R, Pahkinen E. Practical methods for design and analysis of complex surveys. 2nd ed. NewYork: Wiley, 2004.

[7] Rao JNK, Wu CFJ, Yue K. Some recent work on resampling methods for complex surveys. Surv Methodol, 1992, 18: 209-217.

[8] Pfeffermann D, Skinner CJ, Holmes DJ, et al. Weighting for unequal selection probabilities in multilevel models. J R Statist Soc B, 1998, 60: 23-40.

(收稿日期：2009-02-06)
(本文编辑：张林东)