·问题与探讨·

复杂抽样调查数据实例分析

吕筠　何平平　李立明

【导读】 提出复杂抽样调查数据的分析思路和方法以及忽视权重和抽样设计时会出现的问题。文中以 2002 年中国居民营养与健康状况调查数据中高血压患病率的估算为例，分加权和不加权、考虑和不考虑整群设计特征的四种组合情况对数据进行分析。表明忽视权重的设置会影响点估计和标准误的估计，忽视对整群设计特征的考虑不仅会高估结果的精确度，还会得到地区间患病率有差异的假阳性结论。因此使用合理的统计方法分析复杂抽样调查数据非常有必要。

【关键词】 复杂调查数据；权重；抽样误差；假设检验

Data analysis from surveys using complex sampling methods　*LV Jun, HE Ping-ping, LI Li-ming. Department of Epidemiology & Biostatistics, School of Public Health, Peking University Health Science Center, Beijing* 100083, China

Corresponding author: *LI Li-ming*, Email: lmlee@pumc.edu.cn

【**Introduction**】 To present statistical methods on appropriate data analysis from complex surveys and errors arising from ignorance of weights or design of samples. We took Chinese National Nutrition and Health Survey in 2002 as an example to analyze the prevalence of hypertension among population aged 15 and over. We used four combinations of analyses, including with or without weighting or considering sample designs. If weights is omitted, it would result in biased prevalence estimates and also influence the estimates of standard errors. While omitting sample designs would result in underestimating the standard error estimates and then testing the false positive hypothesis. Through appropriate analysis, we found Chinese people in large-sized cities had the highest prevalence of hypertension (28.77%, 95% *CI*: 25.69%-31.84%) while people in the poorest rural area having the lowest prevalence of hypertension (14.21%, 95% *CI*: 12.64%-15.79%). The prevalence of hypertension among people in small and medium-sized cities and other rural areas ranged from 20.48% to 24.37% with statistically insignificant difference. It is necessary to use appropriate methods to analyze data from complex surveys.

【**Key words**】 Complex survey data; Weight; Sampling error; Hypothesis testing

在对除单纯随机抽样以外的概率抽样调查数据进行分析时，研究者至少要考虑两方面的问题。第一，权重的设置。在概率抽样中，总体的每名个体入选概率是已知、非零的，但是不一定相等。如果调查对象的入选概率不等或在某些重要特征上与目标人群的分布不一致，研究者就需要为个体信息赋以适当的权重，以期得到无偏的总体点估计。第二，方差估计。整群抽样设计的抽样误差通常比单纯随机抽样大；而在分层抽样设计中，如果层内的个体变异小于整个总体的变异，则其抽样误差会小于单纯随机抽样。在进行方差估计时，忽视抽样设计，尤其是整群抽样设计，不仅会低估参数的标准误，还会影响假设检验，甚至得出错误的结论[1]。目前，针对复杂抽样调查数据有专门的分析原则和一系列分析方法；遗憾的是，国内学者普遍缺乏对复杂抽样调查数据分析方法的正确认识和掌握，多是直接使用传统的、以单纯随机抽样设计为前提假设的统计分析方法。为此我们将以 2002 年中国居民营养与健康状况调查（营养调查）数据中高血压患病率的估算为例，展示复杂抽样调查数据的分析思路和方法以及忽视权重和抽样设计时会出现哪些问题，从而说明使用正确的方法分析复杂抽样调查数据的必要性。

作者单位：100083 北京大学公共卫生学院流行病与卫生统计学系
通讯作者：李立明，Email：lmlee@pumc.edu.cn

资料与方法

1. 抽样方法：营养调查采用多阶段分层整群随机抽样方法。首先根据经济发展水平及类型进行分层，即将全国各县（市、区）划分为大城市、中小城市和一至四类农村共 6 类地区。第一阶段利用系统抽样方法，从每一类地区中随机抽取 22 个县（市、区）；第二阶段从每个样本县（市、区）中随机抽取 3 个乡镇/街道；第三阶段从每个乡镇/街道中随机抽取 2 个村/居

委会;第四阶段从每个村/居委会中随机抽取 90 户家庭,抽中家庭的所有常住人口作为调查对象[2]。

2. 分析目的:本次数据分析的主要目的为①估算营养调查 6 类地区 15 岁及以上人群高血压患病率;②对 6 类地区的高血压患病率水平进行比较。

3. 分析思路及方法:为了比较不同分析方法对结果的影响,分为 4 种方式分析数据:方式①不加权、不考虑整群设计特征;方式②不加权、考虑整群设计特征;方式③加权、不考虑整群设计特征;方式④加权、考虑整群设计特征,此为合理的分析。

复杂抽样调查数据的分析区别于传统分析之处就在于分析者必须同时考虑:每条记录所提供的信息权重是多少?抽样过程是否有分层、分阶段?哪些抽样单位属于一个群?是否需要做有限总体校正?这些技术参数在数据库中都要转化为对应的变量,用于数据分析。

(1)权重的设置:本次营养调查中个体的入选概率不等,需要设置权重。理想的做法是逐阶段计算抽样单位的入选概率,入选概率的倒数即为入选权重;然后还可以根据应答率计算无应答权重,根据样本人群中关键变量(如年龄、性别)相对目标人群分布状况的偏离计算事后分层权重(post-stratification weight);最后,所有相关权重的乘积即为每名调查对象的分析权重。由于无法找到第二、三和四阶段更详细的抽样框架和应答信息,对应的人口学资料也很有限,所以我们根据 2000 年全国人口普查资料《中国乡、镇、街道人口资料》分县(市、区)、性别和年龄(分15~64 岁和≥65 岁两组)计算权重[3],即 132 个样本县(市、区)各自分性别和年龄的人口数除以本次调查中各样本县(市、区)分性别和年龄的调查人数。这种算法意味着假定同样本县(市、区)中、同性别、同年龄组的个体具有相同的权重。

(2)有限总体校正:由于较难获得各个阶段中可供抽样的单位总数,所以无法进行有限总体校正(finite population corrections, FPC)。但是,考虑到第一阶段每类地区中仅抽取 22 个县(市、区),相对县(市、区)的总数,抽样比例约在 5% 左右。在抽样比例较小(<5%~10%)的情况下,有限总体校正对结果影响不大,可以忽略[4,5]。

(3)对整群设计特征的考虑:营养调查的四个阶段均为整群随机抽样。其中,第一阶段的抽样单位为县(市、区),通常被称为初级抽样单位(primary sampling units)。由于忽略了 FPC,第一阶段的抽样将被按照有放回(with replacement)抽样来处理;由于第二、三和四阶段中群的信息对方差估计不会有较大的影响,因此分析中将其忽略[5]。

(4)对分层设计特征的考虑:由于分析目的中不涉及估算全国总的高血压患病率,所以不需要按 6 类地区占总人口的比例生成相应的权重。

(5)层间比较的统计学检验:层间比较采用调整的 wald 检验,因为涉及层间多重比较,所以使用 Bonferroni 方法进行调整[5,6]。

(6)设计效应的估算:设计效应(design effect)是当前抽样设计与相同样本量的无放回(without replacement)单纯随机抽样的方差之比,反映了不同抽样设计的效率。本次分析中,假定单纯随机抽样是在各类地区内部进行,而不是全国范围。

4. 分析软件:统计分析使用 Stata/MP 10.0 for Windows(StataCorp LP, TX 77845 USA)完成。以 svy 为前缀的统计命令专门适用于抽样调查数据的分析。标准误的计算方法选择基于一阶 Taylor 线性近似法(first-order Taylor series linear approximation)的线性方差估计(linearized variance estimator)[5]。

结 果

1. 数据基本情况:营养调查的 6 类地区148 559 名个体有血压测量值,全部纳入本次分析。这些个体的地区、性别和年龄分布情况如表 1。

表1 6 类地区148 559名有血压测量值的个体地区、性别和年龄分布

地区分类	15 岁~		35 岁~		55 岁~		合计
	人数	%[a]	人数	%[a]	人数	%[a]	
大城市							
男	2 475	21.0	4 687	39.7	4 634	39.3	11 796
女	3 059	21.2	6 031	41.9	5 307	36.9	14 397
中小城市							
男	2 785	26.8	4 530	43.5	3 089	29.7	10 404
女	3 784	29.4	5 789	45.0	3 285	25.5	12 858
一类农村							
男	3 006	26.3	5 317	46.4	3 125	27.3	11 448
女	3 624	27.8	6 319	48.5	3 090	23.7	13 033
二类农村							
男	3 605	31.5	5 182	45.3	2 655	23.2	11 442
女	4 591	34.3	6 215	46.5	2 573	19.2	13 379
三类农村							
男	3 711	31.3	5 200	43.8	2 962	24.9	11 873
女	4 815	34.3	6 361	45.3	2 865	20.4	14 041
四类农村							
男	4 095	37.5	4 383	40.2	2 429	22.3	10 907
女	5 133	39.5	5 268	40.6	2 580	19.9	12 981
合计	44 683	30.1	65 282	43.9	38 594	26.0	148 559

注:[a] 各年龄段人数占同地区、同性别总人数的比例

2. 高血压患病率：由表 2 可见，忽视整群设计特征会大大低估标准误，高估结果的精确度（方式① *vs*. ②，方式③ *vs*. ④）。而忽视权重的设置可以影响点估计，本次分析中影响较大的是大城市和中小城市的患病率点估计；同时也会影响标准误的估计（方式① *vs*. ③，方式② *vs*. ④）。

表2　采用不同方式分析营养调查 6 类地区的高血压患病率

分析方式	地区分类	患病率（%）	$s_{\bar{x}}$	95% *CI*
①	大城市	31.42	0.29	30.86～31.98
	中小城市	24.61	0.28	24.05～25.16
	一类农村	24.12	0.27	23.59～24.66
	二类农村	20.04	0.25	19.54～20.53
	三类农村	22.20	0.26	21.70～22.71
	四类农村	13.44	0.22	13.01～13.87
②	大城市	31.42	1.52	28.41～34.43
	中小城市	24.61	1.25	22.14～27.08
	一类农村	24.12	2.11	19.95～28.30
	二类农村	20.04	1.56	16.94～23.13
	三类农村	22.20	1.36	19.51～24.90
	四类农村	13.44	0.80	11.85～15.03
③	大城市	28.77	0.35	28.09～29.44
	中小城市	20.99	0.45	20.12～21.86
	一类农村	24.37	0.33	23.73～25.01
	二类农村	20.48	0.31	19.88～21.08
	三类农村	22.81	0.31	22.21～23.42
	四类农村	14.21	0.26	13.69～14.73
④	大城市	28.77	1.55	25.69～31.84
	中小城市	20.99	2.45	16.13～25.85
	一类农村	24.37	1.97	20.48～28.26
	二类农村	20.48	1.56	17.39～23.57
	三类农村	22.81	1.29	20.27～25.36
	四类农村	14.21	0.80	12.64～15.79

注：4 种分析方式为①不加权、不考虑整群设计特征；②不加权、考虑整群设计特征；③加权、不考虑整群设计特征；④加权、考虑整群设计特征

由方式④的结果可见，大城市的高血压患病率最高（28.77%，95% *CI*：25.69%～31.84%）；而四类农村地区的患病率最低（14.21%，95% *CI*：12.64%～15.79%）。

3. 地区间高血压患病率的比较：由表 3 可见，如果忽视整群设计特征，标准误的低估会导致假设检验犯一类错误的概率增大，即得到地区间患病率有差异的假阳性结论（方式③ *vs*. ④）。忽视权重的设置则同时通过影响点估计和标准误大小进而影响检验结果（方式② *vs*. ④），但是影响不如前面那么显著。

由方式④的结果可见，除一类农村外，大城市的高血压患病率高于其他地区且差异有统计学意义；中小城市与一、二、三类农村的患病率差异无统计学意义，四类农村的患病率低于其他地区且差异有统计学意义。

4. 设计效应：在方式④的基础上估算 6 类地区内部抽样调查的设计效应。大城市、中小城市、一类农村、二类农村、三类农村、四类农村的设计效应依次为30.9、84.4、51.4、37.1、24.3、12.4。

表3　营养调查 6 类地区间高血压患病率比较的统计学检验（*P* 值[a]）

分析方式	地区分类	大城市	中小城市	一类农村	二类农村	三类农村
①	中小城市	<0.001				
	一类农村	**<0.001**	1.000			
	二类农村	<0.001	**<0.001**	**<0.001**		
	三类农村	<0.001	**<0.001**	**<0.001**	**<0.001**	
	四类农村	<0.001	<0.001	<0.001	<0.001	<0.001
②	中小城市	0.004				
	一类农村	**0.029**	1.000			
	二类农村	<0.001	0.120	0.610		
	三类农村	<0.001	0.977	1.000	1.000	
	四类农村	<0.001	<0.001	<0.001	0.001	<0.001
③	中小城市	<0.001				
	一类农村	**<0.001**	**<0.001**			
	二类农村	<0.001	1.000	**<0.001**		
	三类农村	<0.001	**0.004**	**0.003**	**<0.001**	
	四类农村	<0.001	<0.001	<0.001	<0.001	<0.001
④	中小城市	0.042				
	一类农村	0.410	1.000			
	二类农村	0.001	1.000	0.618		
	三类农村	0.019	1.000	1.000	1.000	
	四类农村	<0.001	0.048	<0.001	0.003	<0.001

注：同表 2；[a] Bonferroni 调整的 *P* 值，黑体数字的 *P* 值表示假阳性

讨　论

以上的数据分析可见，忽视权重的设置导致城市人群高血压患病率的高估；忽视对整群设计特征的考虑不仅导致结果精确度的高估，还得到地区间患病率有差异的假阳性结论。在做合理的分析数据后，营养调查中 2002 年我国居民的高血压患病率水平以大城市最高，四类农村最低，其他各类地区患病率水平居中且接近。

除此之外，由估算的设计效应可见，这样大地理区域的一次复杂的多阶段随机抽样设计，相对于单纯随机抽样设计来说，结果精确度的损失可高达几十倍；换句话说，如果按基于单纯随机抽样设计的公式估算样本量，还必须扩大几十倍的样本量才可以达到期望的容许误差。传统的建议“整群抽样的样本量比单纯随机抽样增加1/2”，仅仅是按设计效应为1.5来考虑的；对于复杂的多阶段随机抽样设计，这并不是普遍适用的原则；在条件允许的情况下，研

究者最好根据既往相关研究的资料推导设计效应，估算样本量。

由于无法找到更详细的抽样框架和应答信息以及人口学资料，我们对权重的设置做了简化处理，忽略了 FPC。这使我们的分析存在一定的局限性；同时也提示，对权重设置、抽样设计的考虑不只是数据分析阶段的问题，设计阶段对抽样框架和相关人口学资料等基础信息的收集至关重要。

参 考 文 献

[1] 吕筠，何平平，涂文校，等. 整群抽样调查数据分析中应正确计算抽样误差. 中华流行病学杂志，2008，29(1)：78-80.
[2] 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，2005：5.
[3] 国家统计局人口和社会科技统计司. 中国乡、镇、街道人口资料. 北京：中国统计出版社，2002.
[4] Cochran WG. Sampling techniques. 3rd ed. New York: John Wiley & Sons, 1977.
[5] Stata Corp. Stata survey data reference manual. College Station, TX: StataCorp LP, 2007.
[6] Stata Corp. Stata base reference manual. volume 3 (Q-Z). College Station, TX: StataCorp LP, 2007: 461-462.

（收稿日期：2008-02-18）
（本文编辑：张林东）

·疾病控制·

中国澳门艾滋病流行现况

陈文诗

澳门的 HIV 感染者/AIDS 患者记录是从 1984 年开始，1986 年发现首例 HIV 感染者，1989 年发现第 2 例；随后每年均有感染者被检出；检出最多是 1993 年 37 例，其次是 1994 年 35 例，1998 年 31 例。从每年新检出的 HIV 感染者数据看，发病呈“双驼峰”状，从 1989－1993 年检出率迅速上升，随之缓慢下降至 1999 年 9 例；然后再上升到 2004 年的 30 例[1]，到 2007 年的前两季度共发现有感染者 11 例。在首例被发现至今共检出 HIV 感染者 378 例[2,3]；其中本地居民 121 例，非本地居民 257 例；澳门 HIV 流行率为 0.056%，与香港和台湾等地区相附[4]。

1. HIV 感染者检出场所[2,3]：2007 年前两季度的 11 例中 6 例为本地居民，其中 1 例是在结核病防治中心检查时被发现，另 1 例是戒毒机构的药物成瘾者，其余 4 例是在医疗机构中被检出的。非本地居民有 5 例，3 例是监狱服刑的囚犯，1 例为娱乐场所人员，1 例是在医疗机构中被发现。

2. 感染途径：本地居民主要以性接触为感染 HIV 的方式，共有 46 例(38%)，其中异性性接触感染 33 例(27%)，同性性接触感染 13 例(11%)；通过静脉注射药瘾而感染的为第二位，有 34 例(28%)；其余的 41 例(33%)感染途径不明。据 1997－2007 年的数据显示，通过异性性接触被感染者每年约 1～3 例；男男性接触感染 1998、2002 和 2005 年分别有 1 例，2004 年为 2 例和 2006 年有 3 例；但 2004 年通过静脉注射药瘾感染 HIV 者大幅上升达到 15 例，之后 2005 年 7 例和 2006 年 4 例又迅速回落；感染途径不详的例数基本保持在 1～6 例间。与中国香港和台湾地区主要传播 HIV 途径相比[4]，性传播、静脉吸毒和感染途径不详的排位顺序基本一致，但澳门的特点是性传播的百分率相对较少，静脉吸毒者居中，途径不详是三地之中最高的。

3. 高危人群筛检：近年澳门的初级卫生保健、预防监测及医疗私隐保密情况日趋完善，许多慢性病患者(肺结核)、要求戒除药瘾者、孕妇保健、捐血质量监控和匿名者均要求并进行了 HIV 检测。从 1997 年开始以人群分类为检测总体和统计，发现捐血者血液(0.002%)、孕妇(0.01%)和匿名者(0.017%)都保持在低的感染率(<0.1%)；其次为结核病患者(0.12%)、娱乐场所外籍劳工(0.25%)、临床怀疑个案(0.34%)和囚犯(0.53%)存在较高感染率(0.1%～0.6%)；最高 HIV 检出率的是自愿戒毒者(1.65%)。

简而言之，澳门仍是属 HIV 低发病率地区[5]，流行情况和邻近的香港及台湾地区相近。过去病例以非本地人居多，但近年本地居民感染率有上升的趋势。在本地人中，静脉注射和异性性接触均为主要传播途径；静脉注射者的感染自 2004 年有快速上升的趋势，但至 2006 年渐趋平稳，情况有受控制迹象；经男男性接触感染的个案在 2005－2006 年稍有上升。非本地人主要通过娱乐场所外籍性工作者及监狱囚犯的筛查发现，以异性性接触为传播途径。因此，建议加强“健康性生活，安全性行为”活动的推广。

参 考 文 献

[1] 卫生局. 2006 年度统计年刊. 澳门：域思广告制作有限公司，2007：115.
[2] 卫生局. 澳门人类免疫缺陷病毒(艾滋病病毒)/艾滋病感染情况季度统计资料(2007 年 1－3 月)[DB/OL]. http://www.ssm.gov.mo/design/NEWS/c_news_fs.htm.
[3] 卫生局. 澳门人类免疫缺陷病毒(艾滋病病毒)/艾滋病感染情况季度统计资料(2007 年 4－6 月)[DB/OL]. http://www.ssm.gov.mo/design/NEWS/c_news_fs.htm.
[4] Virtual AIDS Office. HIV/AIDS Situation in Hong Kong [2006] [DB/OL]. http://www.info.gov.hk/aids/english/new2007/nm16.htm.
[5] UNAIDS, WHO. 07 AIDS epidemic update. Geneva: WHO Library Cataloguing-in-Publication, 2007: 7.

（收稿日期：2008-01-21）
（本文编辑：尹康）

作者单位：中国澳门镜湖护理学院