·问题与探讨·

整群抽样调查数据分析中应正确计算抽样误差

吕筠　何平平　涂文校　李立明

【导读】 为了澄清整群抽样调查数据分析中正确计算抽样误差的必要性，以在某市 15 岁及以上人群中开展的一次两阶段整群抽样调查为例，分别采用适用于单纯随机抽样数据的方法和考虑了复杂抽样设计的方法对数据进行分析。结果显示，忽略对复杂抽样设计的考虑，不恰当的采用适用于单纯随机抽样数据的方法进行数据分析，不仅有可能大大低估样本统计量的抽样误差，在进行假设检验时，甚至会得到错误的结果，故正确分析和报告整群抽样调查数据的抽样误差是非常必要的。

【关键词】 整群抽样；单纯随机抽样；抽样误差；复杂调查数据

Estimation of sampling error on data from cluster sample survey　*LV Jun, HE Ping-ping, TU Wen-xiao, 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 clarify the necessity of applying appropriate statistical methods to calculate sampling error from data of cluster sample survey, we take a two-stage cluster sample survey developed from a population aged 15 and over as example. We use statistical methods based on the assumption of simple random samples and methods considering complex sample design to analyze our data, respectively. Through comparison, we hope to show the potential effects of using improper statistical methods to estimate sampling error on parameter estimation and hypothesis testing. Using standard error algorithms based on the assumption of simple random samples, the standard errors calculated often underestimate sampling error and the hypothesis testing even gets wrong conclusion. When the statistical methods and statistics package for complex survey data are already available, it is necessary for us to use appropriate methods to analyze and report the sampling errors of data from cluster sample survey.

【Key words】 Cluster sampling; Simple random sampling; Sampling error; Complex survey data

传统的统计教学和绝大多数统计书中讲授的统计分析方法都是基于单纯随机抽样的假设。但是，在实际工作中，专业人员很少直接采用单纯随机抽样，更多的是采用整群抽样或整合多种抽样方法的多阶段抽样。在分析这样的复杂抽样调查数据时，不假思索的使用传统的统计分析方法很可能会得到错误结果。适当的分析至少要考虑两个方面的问题：一方面是影响样本统计量点值估计的“加权”问题，另一方面是对样本统计量的抽样误差计算问题。从目前国内医学期刊上发表的大量概率抽样调查报告来看，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。本文着重讨论常用的两阶段整群抽样调查数据分析中抽样误差的计算，分析抽样误差增大的原因、设计效应以及错误计算复杂抽样调查数据的抽样误差会出现哪些问题，从而说明使用正确的方法计算抽样误差的必要性。

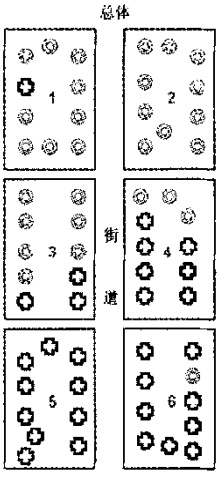
基本原理

在概率抽样调查中，抽样误差的大小受样本量、抽样方法等因素的影响。抽样方法一定时，样本量越大，抽样误差越小。样本量一定时，整群抽样的抽样误差通常要比单纯随机抽样大。抽样误差越大，意味着样本统计量的可信区间越宽，估计值的精确度越低。我们通常用标准误来定量估计样本统计量与总体参数之间的差异，即抽样误差。

1. 整群抽样调查数据抽样误差增大的原因：整群抽样调查数据抽样误差增大主要是因为群内个体在研究的性状上趋向于同质，而群间的变异性加大。以图 1 为例，假定某街道有 6 个小区（如图 1 左侧），每个小区中有 10 幢房屋。我们用圆圈表示普通房屋，十字框表示豪华房屋，可见该街道普通房屋和豪华房屋各占一半，普通房屋和豪华房屋的分布存在

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

明显的聚集性。如果我们想通过整群抽样调查了解街道中普通房屋的比例，以小区为抽样单位，抽取 2 个小区共 20 户进行调查。图 1 右侧的表格中列出了所有可能的抽样组合及计算的样本统计量。已知整个街道中普通房屋的比例为0.50，15 种组合的平均值也为0.50。但是，这 15 种组合之间存在很大的变异，其中 6 个估计值≥0.80或≤0.20（在表格中用 * 标记），误差很大。这个例子反映了当性状在群内趋于同质、而群间变异很大时，整群抽样数据样本统计量的抽样误差会很大。



抽取两个小区（样本中 房屋的比例）

抽取的小区号	房屋的比例
1, 2	19/20=0.95*
2, 3	17/20=0.85*
1, 3	16/20=0.80*
2, 4	13/20=0.65
1, 4	12/20=0.60
2, 6	11/20=0.55
1, 6	10/20=0.50
2, 5	10/20=0.50
3, 4	10/20=0.50
1, 5	9/20=0.45
3, 6	8/20=0.40
3, 5	7/20=0.35
4, 6	4/20=0.20*
4, 5	3/20=0.15*
5, 6	1/20=0.05*

图1 整群抽样调查示意图

2. 设计效应：在公式 1 中，$v(\bar{y})$表示整群抽样时样本统计量的方差；当假定抽样方法为单纯随机抽样时，方差为 $v_{srs}(\bar{y})$。两者的比值 d^2 称为设计效应（design effect）。设计效应反映了整群抽样时样本统计量的精确度损失。

$$d^2 = \frac{v(y)}{v_{srs}(\bar{y})} \quad (1)$$

整群抽样调查抽样误差增大直接影响到调查设计时样本量的估算。例如，根据既往相同人群中开展的同类研究，我们了解到某项指标的整群抽样设计效应为$d^2 = 3.13$。在设计一项新的调查时，当要求的容许误差一定时，如果开展单纯随机抽样需要调查 200 人，则进行整群抽样调查时需要调查$200 \times 3.13 = 626$人，才可以满足设计者对调查精确度的要求。

在流行病学教材中，当介绍整群抽样时，通常会有这样的陈述："群间变异越小，群的规模越小，抽取的群越多，获得样本统计量的精确度就越好。由于抽样误差增大，所以建议整群抽样的样本量比单纯随机抽样增加1/2"。实际上，这种经验的总结也是基于上面介绍的原理。其中建议的整群抽样的样本量增加 1/2，是按整群抽样的设计效应为1.5来考虑的。

3. 错误计算抽样误差对参数估计和假设检验的影响：经整群抽样调查获得的数据，如果仍按基于单纯随机抽样假设的传统统计方法进行分析，在参数估计中，计算得到的标准误很有可能会低估样本统计量的抽样误差，有时差距可高达几十倍，使人们对研究结果的精确度产生错误的认识，影响进一步的研究实践或卫生决策。如果继续进行假设检验，抽样误差的低估会导致检验的 I 类错误概率增大，可能得到假阳性的错误结论。

4. 统计学分析软件：目前，多数分析复杂概率抽样的统计方法已经可以通过统计软件实现，流行病学专业人员在对基本概念有一定的理解后就可方便的使用。例如，Epi Info 中前缀为 Complex Sample 的分析菜单；SAS 中的"Proc Survey"模块；SPSS 中的"Complex Samples"菜单；Stata 中的"Survey data analysis"菜单。除此之外，Sudaan 是专门用来分析复杂抽样调查数据或类似结构数据的统计软件，还有一些其他软件，如 CENVAR、VPLX、WesVar 等。

实例分析

我们在某市 15 岁及以上居民中开展了一次抽样调查，了解居民日常获取健康信息的主要途径。抽样设计为两阶段整群抽样。第一阶段，以村/居委会为抽样单位，共抽取 14 个村/居委会。第二阶段，在抽到的村/居委会中进一步随机抽取一定数量的个体进行调查。最终，调查了 604 人，每个村/居委会调查 22～78 人不等。统计学分析使用 Stata/MP 10.0 for Windows（StataCorp LP, TX 77845 USA）完成。

1. 参数估计：表 1 中展示了按单纯随机抽样和两阶段整群抽样分别得到的样本统计量的标准误。分析的两个变量为居民中收看电视的比例（588/604）和收听广播的比例（137/604）。由表 1 可见：①样本统计量（即两种行为在调查人群中的比例）的点估计不会因为是否考虑抽样方法而发生改变，点值大小只受加权处理的影响。②在正确考虑了整群抽样的设计特征后，两个变量的标准误均增大，尤其是"广播"变量，标准误由0.0171增加到0.0790。从计算的设计效应的大小也可以看出同样的规律，"电视"变量的设计效应为2.29，而"广播"变量的设计效应高达21.46。可见，如果错误的按单纯随机抽样进行分析，会显著低估样本统计量的抽样误差。根据正确的分析，本次调查得到的"广播"变量的 95% 可

信区间非常宽，容许误差高达0.17，相对误差达74%，结果精确度很低，提醒读者在参考这个结果时要谨慎。

为什么整群抽样设计会对“广播”变量的抽样误差有这么大的影响？我们进一步分析了“电视”变量和“广播”变量在各个村/居委会中的比例。由于每个村/居委会中调查的人数很少，得到的比例不能作为该村/居委会人群行为水平的真实反映。但是，该分析可以提供一些线索。由图2和图3可见，“电视”变量的群间变异(88%～100%)不是很大；而“广播”变量的群间变异却是相当明显(4%～100%)。收听广播这个行为在同一村/居委会中有同质的倾向。

表1　居民中收看电视和收听广播的比例——不同统计分析方法对样本统计量抽样误差大小的影响

方法及变量	比例	s_E	95% *CI*	设计效应
按单纯随机抽样				
收看电视	0.97	0.0065	0.96～0.99	-
收听广播	0.23	0.0171	0.19～0.26	-
按两阶段整群抽样				
收看电视	0.97	0.0099	0.95～0.99	2.29
收听广播	0.23	0.0790	0.06～0.40	21.46

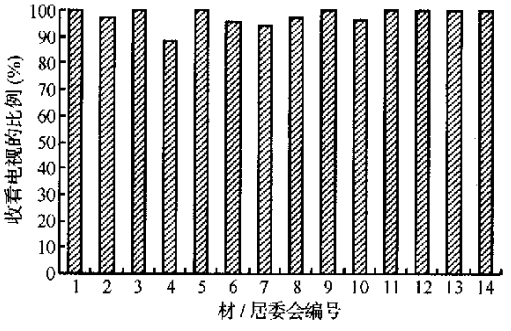


图2　各村/居委会人群中收看电视的比例

2. 假设检验：比较城市和农村居民在3个变量(“是否收听广播”、“是否经常阅读书籍”和“是否会向卫生专业人员咨询健康知识”)分布上的差异是否存在统计学意义。对二分变量的检验方法，基于单纯随机抽样假设，应采用 Pearson χ^2 检验。如果考虑了整群抽样设计，Stata 软件中默认的计算方法是 Rao 和 Scott 校正法，提供了 F 统计量，并给出相应的 P 值。从表2中可以看到，基于不同抽样方法分析得到的 P 值都不同。与单纯随机抽样相比，在正确考虑了整群抽样的设计特征后，P 值一致增大。“向专业人员咨询”这个变量的检验结论甚至发生“质”的变化：考虑了整群抽样的统计分析表明城乡无差异($P=0.27$)，而基于单纯随机抽样的分析却得到了城乡有差异的假阳性结果($P<0.001$)。

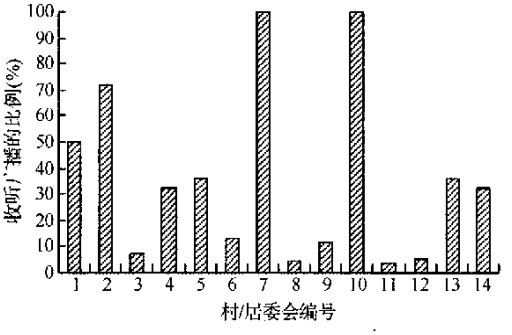


图3　各村/居委会人群中收听广播的比例

结　语

以上实例分析表明，忽略了整群抽样设计，错误的按单纯随机抽样假设的传统方法对数据进行分析，不仅有可能大大低估抽样误差，无法正确认识样本统计量的精确度水平；在进行假设检验时，甚至有可能得到错误的结果。建议研究者在分析复杂抽样调查数据时能够使用正确的统计分析方法；在撰写论文时，清晰介绍调查的抽样设计和使用的统计分析方法，并报告结果的点估计值和可信区间。同时，也建议科技期刊能将此要求纳入审稿的标准中。另外，鼓励研究者能够计算并在投稿论文中报告调查中主要指标的设计效应，今后如果开展类似研究，这个参数可作为研究设计时的重要参考。

表2　城乡居民3个变量分布比较分析——不同统计分析方法对假设检验结果影响

变量	城市/农村		分析方法	检验统计量	*P* 值
	n/N^a	比例			
收听广播	城市：110/300	城市：0.37	单纯	Pearson $\chi^2=66.47$	<0.001
	农村：27/304	农村：0.09	整群	$F(1,13)=7.50$	0.017
阅读书籍	城市：44/299	城市：0.15	单纯	Pearson $\chi^2=15.03$	<0.001
	农村：16/304	农村：0.05	整群	$F(1,13)=5.58$	0.03
向专业人员咨询	城市：211/299	城市：0.71	单纯	Pearson $\chi^2=15.73$	<0.001
	农村：221/261	农村：0.85	整群	$F(1,13)=1.32$	0.27

注：单纯——基于单纯随机抽样假设的常规统计方法；整群——考虑了两阶段整群抽样设计的专门统计方法；“n 为具备某种行为的人数，N 为调查人数

(收稿日期：2007-08-17)

(本文编辑：张林东)