·基础理论与方法·

死因监测整群抽样与不等概率抽样设计方案的比较

廉恒丽 徐勇勇 虢玲霞 谭志军 刘丹红 饶克勤

【导读】 为比较不同整群抽样设计方法的抽样误差及设计效应，评价不等概率抽样在死因监测中的应用效果。以陕西省107个县(市、区)作为抽样框架，采用等概率整群抽样和不等概率整群抽样等设计方案抽取样本，用复杂抽样方法计算不同方案样本的抽样误差和设计效应。不同的抽样方案得到不同的抽样误差估计，分层整群抽样的标准误小于完全随机整群抽样；不等概率抽样(πPS抽样)的设计效率虽略逊于等概率的完全随机整群抽样，但扩大了监测范围。结论：对于抽样框架明确的整群抽样调查数据，在统计分析时不应脱离预先设定的抽样设计方案和设计参数。死因监测采用不等概率抽样设计，能增加样本的权重，提高死亡率的地区代表性。

【关键词】 整群抽样；复杂抽样；不等概率抽样；抽样误差

Selection of sentinel sites for death surveillance, using cluster or unequal probability sampling

LIAN Heng-li[1], *XU Yong-yong*[1], *GUO Ling-xia*[2], *TAN Zhi-jun*[1], *LIU Dan-hong*[1], *RAO Ke-qin*[3].
1 *Department of Health Statistics Fourth Military Medical University, Xi'an* 710032, China;
2 *Department of Health of Shaanxi Province*; 3 *Ministry of Health P.R China*
Corresponding author: *XU Yong-yong*, Email: xuyongy@fmmu.edu.cn; *RAO Ke-qin*, Email: raokq@moh.gov.cn

This work was supported by a grant from the Ministry of Science and Technology (the Special Program of Prevention and Control of Significant Infectious Diseases: AIDS, Viral Hepatitis and Otherwise) (No. 2009ZX10002-027)

【Introduction】 To compare the sampling errors from cluster or unequal probability sampling designs and to adopt the unequal probability sampling method to be used for death surveillance. Taking 107 areas from the county level in Shaanxi province as the sampling frame, a set of samples are drawn by equal probability cluster sampling and unequal probability designs methodologies. Sampling error and effect of each design are estimated according to their complex sample plans. Both the sampling errors depend on the sampling plan and the errors of equal probability in stratified cluster sampling appeares to be less than simple cluster sampling. The design effects of unequal probability stratified cluster sampling, such as πPS design, are slightly lower than those of equal probability stratified cluster sampling, but the unequal probability stratified cluster sampling can cover a wider scope of monitoring population. Conclusions: Results from the analysis of sampling data can not be conducted without consideration of the sampling plan when the sampling frame is finite and a given sampling plan and parameters, such as sampling proportion and population weights, are assigned in advance. Unequal probability cluster sampling designs seems to be more appropriate in selecting the national death surveillance sites since more available monitoring data can be obtained and having more weight in estimating the mortality for the whole province or the municipality to be selected.

【Key words】 Cluster sampling; Complex sampling; Unequal probability sampling; Sampling error

死因监测数据是反映一个国家或地区居民健康状况和疾病流行严重程度的直接指标[1]。由于受到社会经济发展条件的制约，我国的死因监测仍采用哨点监测的方法，即以全国2400多个县(区)级单元作为抽样框架[2]，采用整群抽样的方法按一定抽样比例，随机抽取部分县(区)作为监测点。为了提高监测点数据对本地(省、市级)死亡率水平的代表性，本研究以陕西省107个县(市、区)级单元作为抽样框架，比较不同整群抽样设计方法的抽样误差及设计效应，并探讨不等概率抽样在我国死因监测和传染病防治中的应用问题。

基本原理

当抽样对象为某地自然人群时，方便的抽样方

DOI: 10.3760/cma.j.issn.0254-6450.2010.04.023
基金项目：科技部艾滋病和病毒性肝炎等重大传染病防治重大专项(2009ZX10002-027)
作者单位：710032 西安，第四军医大学卫生统计学教研室(廉恒丽、徐勇勇、谭志军、刘丹红)；陕西省卫生厅办公室(虢玲霞)；卫生部统计信息中心(饶克勤)
通信作者：徐勇勇，Email: xuyongy@fmmu.edu.cn；饶克勤，Email: raokq@moh.gov.cn

法是把抽样总体可按照行政区域划分为K个初级单元(PSU),如全国范围抽样,PSU可以是省(市、自治区),也可以是县(区)[K为全国省(市、自治区)或县(区)个数]。整群抽样的方法是按一定的抽样比率f,从K个PSU中随机抽取k个PSU作为观察样本。第i个PSU的入样概率为$\pi_0=f=k/K$。由于抽样比率f是一个给定的常数,该抽样方法也称为等概率抽样。如果把第i个PSU所包含的人口数看作该PSU的N_i个次级单位(SSU),抽样总体中共有N个SSU数(人口总数),则第i个PSU的入样概率为$\pi_i=k(N_i/N)$。由于PSU的入样概率π_i与PSU的容量(MOS)N_i成比例,N_i越大,第i个PSU入样的概率越大,属不等概率抽样。当抽样为有放回时,称为PPS抽样,当抽样为无放回时,称为πPS抽样[3]。在死因监测中,令Y_{ij}表示第i个PSU中第j个SSU的死亡标志,即

$$Y_{ij}=\begin{cases}1 & \text{死亡}\\ 0 & \text{存活}\end{cases},\ i=1,2,\cdots,K,j=1,2,\cdots,N_i$$

对于第i个πPS抽样的PSU样本,令死亡人数为y_i,$i=1,2,\cdots,k$。总体死亡人数的Horvitz Thompson估计值(HT估计)为[4]

$$\tilde{y}_{HT}=\sum_{i=1}^{k}\frac{y_i}{\pi_i},\quad \bar{p}_{\pi ps}=\frac{\tilde{y}_{HT}}{N}$$

$$s_{\tilde{y}_{HT}}=\sqrt{\sum_{i=1}^{k}\sum_{i>j}^{k}\left(\frac{\pi_i\pi_j}{\pi_{ij}}-1\right)\left(\frac{y_i}{\pi_i}-\frac{y_j}{\pi_j}\right)^2}$$

总体死亡人数估计的标准误$s_{\tilde{y}_{HT}}$与死亡率标准误成比例,计算难度主要是构建$k\times k$联合概率矩阵$\pi=(\pi_{ij})$,即先后两次无放回抽样分别抽中第i个PSU和第j个PSU的联合概率。由于计算复杂,必须用编程运算[5,6]。

不等概率分层整群抽样的原理同πPS,不同的是按地域分层后,再在各层内采用无放回不等概率整群抽样。

实例分析

1. 抽样框架:陕西省107个县(市、区)人口学信息(人口数、男性人数、女性人数、户数等)来自2007年陕西省户籍资料。分层因素为地区(陕北、关中、陕南)和城乡,死亡率资料来自2007年陕西省统计报告。抽样框架基本参数见表1。

2. 抽样设计方案:①完全随机抽样(抽样比率$=f$,等概率);②完全随机整群抽样(f=1%、10%、25%,等概率);③分层整群抽样(f=10%,等概率);④不等概率整群抽样(f=10%,与人口数成比例);⑤不等概率分层整群抽样(f=10%,与人口数成比例)。

3. 抽样参数与统计量:

(1)完全随机抽样:以人作为监测单位,样本量为n,N=37 827 142,π=4.82‰。

$$s_p=\sqrt{(1-f)}\cdot\sqrt{\frac{\pi(1-\pi)}{N\cdot f}},\ f=\frac{n}{N}$$

(2)完全随机整群抽样加权估计:以各县(市、区)的人口数作为权数(W),随机抽取k个县(市、区)作为监测点

$$K=107,\ \bar{p}=\frac{\sum_{i=1}^{k}wp}{\sum_{i=1}^{k}w}$$

$$s_p=\sqrt{(1-f)}\cdot\sqrt{\frac{\sum w(p-\bar{p})^2}{k(\sum w-1)}},\ f=\frac{k}{K}$$

(3)分层整群抽样:分层因素以地区作为分层标识,将总体分为陕北、关中、陕南3个地区,分别用1、2、3表示。等概率分别抽取k_h个县(市、区)作为监测点,h=1,2,3

k_1=25,k_2=54,k_3=28,$W_h=\frac{K_h}{K}$,h=1,2,3

$$\bar{p}_{st}=\sum W_h\bar{p}_h,\ s_{p_{st}}=\sqrt{\sum W_h^2 s_{ph}^2}$$

(4)不等概率整群抽样:由$\tilde{y}_{HT}$标准误$s_{\tilde{y}_{HT}}$估计死亡率标准误$s_{\bar{p}_{\pi ps}}$

$$\tilde{y}_{HT}=\sum_{i=1}^{k}\frac{y_i}{\pi_i},\ \bar{p}_{\pi ps}=\frac{\tilde{y}_{HT}}{N}$$

(5)不等概率分层整群抽样:增加地区分层标识,将总体分为陕北、关中、陕南3个层(计算公式略)。

4. 计算方法:

(1)设计效率(DEFF):

$$\mathrm{Sq}DEFF=\sqrt{\frac{SE_{CS}^2}{SE_{SRS}^2}}=\frac{SE_{CS}}{SE_{SRS}}$$

式中Sq$DEFF$表示$DEFF$的正平方根,SE_{CS}和SE_{SRS}分别为复杂抽样标准误与完全随机抽样标准误的比值,两者效率相等时,比值约等于1。Sq$DEFF$越大,复杂抽样的效率越低[7,8]。

(2)统计软件:使用SPSS 16.0软件复杂抽样设计与分析模块(complex samples)[9]。

5. 抽样结果:见表2。

讨　论

1. 结果评价:

表1　陕西省死因监测县抽样框架基本参数

地区	城市			农村			合计		
	市(区)数	人数	死亡率(‰)	县数	人数	死亡率(‰)	县(市)数	人数	死亡率(‰)
陕北	2	917 700	5.09	23	4 749 267	3.13	25	5 666 967	3.45
关中	18	9 007 648	4.81	36	13 925 805	4.30	54	22 933 453	4.50
陕南	3	2 074 202	5.94	25	7 152 520	6.59	28	9 226 722	6.45
合计	23	11 999 550	5.03	84	25 827 592	4.72	107	37 827 142	4.82

(1)抽样误差比较:完全随机整群抽样(方案B、C、D),随着抽样比例的增高(1%～25%),加权标准误[表2(9)]由2.30‰减少到0.49‰。当抽样比例为100%时(方案A),$SE_{CS}=SE_{SRS}=0$。当抽样比例>10%时,所有整群抽样方案死亡率的加权标准误均<1(0.83‰～0.96‰)。在抽样点相同的情况下($k=11$),分层整群抽样(方案E、G)的标准误小于完全随机整群抽样(方案C)。

(2)监测人口:等概率抽样(方案A、B、C、D、E)监测人口比率与抽样点的比率基本一致[表2(6)],不等概率抽样(方案F、G)监测人口比率大于抽样点的抽样比率(10%),分别占总人口的16.3%(方案F)和13.2%(方案G)。

(3)设计效率:SRS整群抽样(方案B、C、D)0.94～0.99;分层整群抽样(方案E)为0.82;πPS(方案F、G)分别为1.09和1.14[表2(10)]。表2中,还包含以人为抽样对象和以县(区)为抽样对象的两个基本SRS方案,虽然在死因监测中通常不以人为抽样对象,但可以依据其标准误估计整群抽样的设计效率[表2(11)],用设计效率和样本率的标准误估计复杂抽样的标准误:$SE_{CS}\approx \mathrm{Sq}DEFF\times SE_{SRS}$。

2. 结论:依据国家(或省、市)抽样方案获得的整群抽样调查数据,在统计分析时不应脱离预先设定的抽样设计方案和设计参数(如抽样比例和权重[10])。尤其当抽样比例较高(如25%)或分层整群抽样时,基于无限总体抽样的常规统计与基于抽样方案的误差估计会出现较大差异(表2中黑体数字)。

我国许多以自然人口为调查对象的大规模调查,大多采用等概率抽样方法,其缺点是:当以行政区划作为初级抽样单元时,由于每个行政区划的人口数(次级单元)不可能相等,如果采用SRS整群抽样估计方法,虽然在计算样本估计值(如死亡率估计)时,可以用人口数作为权数校正估计值,但抽样点的确定不受人口数的影响。相反,不等概率整群抽样用于选择死因和疾病监测点,虽然设计效率略低于等概率的SRS整群抽样(表2中方案F、G),但显著的优点是:在监测点数目相同、抽样误差不明显增加、人力物力资源基本相同的情况下,用不等概率整群抽样获得的监测点覆盖了更多的人口,提高了总体死亡率估计的样本权重。

本研究是以陕西省107个县(市、区)作为抽样框架,最少人口数为33 389人(佛坪县),最多人口数为979 189人(安康市汉滨区),相差约30倍。在传染病防治现场研究中,用于估计患病率或死亡率的PSU往往更基层(如镇、乡、村),人口数相差会更悬殊,为了以较大概率抽中人口数较多的PSU作为研究现场,不等概率整群抽样设计是一个可供选择的备选方案。

表2 不同抽样设计方案的抽样误差与设计效率比较

设计方案(代码)	抽样比例(%)	抽样对象	PSU(k)	SSU($\sum N_i$)	实际抽样比例(%)	死亡率(‰)	标准误(‰)	加权标准误(‰)	Sq $DEFF^a$	Sq $DEFF^b$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
全面调查(A)	100	县	107		100	4.82	0.22	0.00		43.78
		人		37 827 142	100	4.82	0.00	0.00		1.00
完全随机(B)	1	县	1		0.91	8.09	2.31	2.30	0.99	16.67
		人		250 049	0.66	4.82	0.14	0.14		1.00
完全随机(C)	10	县	11		10.30	6.33	0.96	0.96	0.98	68.49
		人		4 465 237	11.80	4.82	0.01	0.01		1.00
完全随机(D)	25	县	27		25.20	5.21	**0.51**	**0.49**	0.94	22.43
		人		9 388 115	24.80	4.82	0.02	0.02		1.00
分层(E)	10	县	11		10.30	4.76	**1.09**	**0.83**	0.82	66.25
		人		3 708 531	9.80	4.82	0.02	0.02		1.00
πPS(F)	10	县	11		10.30	6.01	**0.94**	**0.88**	1.09	33.77
		人		6 175 717	16.30	4.82	0.03	0.03		1.00
分层πPS(G)	10	县	11		10.30	4.62	**0.75**	**0.83**	1.14	24.39
		人		5 008 621	13.20	4.82	0.03	0.03		1.00

注:[a]SRS以县(区)为抽样对象,SPSS CSDESCRIPTIVES 计算结果;[b]SRS以人为抽样对象

参 考 文 献

[1] Liu F, Liu L, Guo XR, et al. Statistics analysis of the death causes of residents at disease surveillance points in Shannxi province, 2006. Dis Surveil, 2008, 23(9): 589-592. (in Chinese)
刘峰,刘岭,郭晓荣,等. 2006年陕西省疾病监测点居民死亡原因统计分析. 疾病监测, 2008, 23(9): 589-592.

[2] Rao KQ, Chen YD, Chen XZ, et al. Selection of a national area sample for acquisition of national health information in China. Chin J Health Stat, 1992, 9(3): 1-6. (in Chinese)
饶克勤,陈育德,陈晓章,等. 国家卫生统计与专题调查样本地区的实验设计. 中国卫生统计, 1992, 9(3): 1-6.

[3] Shi XQ. The theory and method of sample survey. Shanghai: Shanghai University of Finance & Economics Press, 1999: 134-142, 209-259. (in Chinese)
施锡铨. 抽样调查的理论和方法. 上海:上海财经大学出版社, 1999: 134-142, 209-259.

[4] Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat Assoc, 1952, 47: 663-685.

[5] Hidiroglou MA, Gray GB, Algorithm AS. Construction of joint probability of selection for systematic P.P.S. sampling. J Royal Stat Society, Series C (Applied Statistics), 1980, 29(1): 107-112.

[6] Xue HS, Yang GH. Weights in Horvitz-Thompson statistic for complex samples. J Hyg Res, 2000, 29(1): 61-63. (in Chinese)
薛禾生,杨功焕. 复杂样本Horvitz-Thompson估计量的权数计算. 卫生研究, 2000, 29(1): 61-63.

[7] Fan HK. Sample survey techniques. Tianjin: Nankai University Press, 1995: 10-11. (in Chinese)
樊鸿康. 抽样调查技术. 天津:南开大学出版社, 1995: 10-11.

[8] Liu JH, Jin SG. Estimation of population quantities and their variances in complex sample survey. Chin J Health Stat, 2008, 25(4): 323-334. (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): 323-334.

[9] SPSS Complex Samples™ 16.0 [DB/OL]. http://www.si.uevora.pt/spss/pdf/manual_spss_16/SPSS%20Complex%20Samples%2016.0.pdf.

[10] Lv J, He PP, Li LM. Data analysis from surveys using complex sampling methods. Chin J Epidemiol, 2008, 29(8): 832-834. (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): 832-834.

(收稿日期:2009-09-20)
(本文编辑:张林东)