·论著·

SAS统计软件的 SURVEYSELECT 过程在血吸虫病流行病学抽样调查中的应用

党辉 郭家钢 徐志敏 王强 吴晓华 周晓农

【摘要】目的 为了减少第三次全国血吸虫病流行病学抽样调查中人为选择样本带来的偏差。方法 采用 SAS 统计软件的 SURVEYSELECT 过程在计算机上进行抽样。结果 分别用计算机在江苏、江西、安徽、湖南、湖北、云南和四川随机抽取 13、23、18、47、58、12 和 68 个样本点；共抽取 239 个样本村，占未达到传播阻断标准乡镇的所有流行村的 1.36‰。结论 SAS 统计软件的 SURVEYSELECT 过程是计算机和现代统计学结合发展的结果，它极大地丰富了现场流行病学，它为流行病学的现场调查提供了简单而快速的样本选择方法，有着广泛的应用空间。

【关键词】抽样调查； 统计分析系统； 等概率抽样； 按初级单位含量比例的概率抽样

The application of the SURVEYSELECT procedure of Statistical Analysis System at schistosomiasis epidemic sampling survey Dang Hui, Guo Jiagang, Xu Ziming, Wang Qiang, Wu Xiaohua, Zhou Xiaonong. National Institute of Parasitic Diseases, Chinese Center for Disease Control and Prevention, Shanghai 200025, China.

【Abstract】 Objective Sampling by the computer decrease the sampling bias by people at the third nation wide schistosomiasis epidemic sampling survey. Methods Sample been sampled by the SAS 's SURVEYSELECT procedure on the computer. Results Jiangsu, Jiangxi, Anhui, Hunan, Hubei, Yunnan and Sichuan have been separately sampled by the computer 13, 23, 18, 47, 58, 12 and 68 sample sites. Conclusion The SAS 's SURVEYSELECT procedure combine the computer with the modern statistics, it enrich largely field epidemiology, it provide the simple and quick sampling method for the field survey of epidemiology.

【Key words】 Sampling survey, SAS (Statistical Analysis System), Equal Probability Sampling, Probability Proportional To Size Sampling

第三次全国血吸虫病流行病学抽样调查是用分层、整群、随机抽样的方法对湖北、湖南、江西、安徽、江苏、四川和云南七省中未达到传播阻断标准乡镇的所有流行村随机抽取样本进行调查。为了减少人为选择样本带来的偏差，在本此调查中第一次采用 SAS 统计软件的 SURVEYSELECT 过程在计算机上进行抽样。

SAS 软件的 SURVEYSELECT 过程是用样本调查来估计总体信息的一种计算机抽样方法，它首先出现在 SAS 统计软件的第七版中，主要用于针对不同的样本设计抽取样本从而对研究总体进行有效的统计推断，特别是较复杂的样本设计如分层、按比例和整群抽样等。

一、SAS 的 SURVEYSELECT 过程的介绍

SURVEYSELECT 过程能提供多种方法用以选择以概率为基础的随机样本，它不仅能选择单纯随机抽样的样本，而且能根据复杂的多阶段样本设计(分层、整群和不等概率)选择样本。在概率抽样中，假设调查总体的每个单位有一个已知的可供选择的正概率，从而避免选择偏差，在统计理论上用样本对调查总体进行有效的推断。

用 PROC SURVEYSELECT 选择样本，首先调入包含抽样基本内容的 SAS 数据集，列出所有被选择的样本；随后在 SAS 的语句中定义特别的选择方法等参数，如样本大小、抽样比和其它可选择的参数。SURVEYSELECT 过程运算后输出结果数据集，其中包含被选的样本、选择概率和抽样的权重。如果进行多阶段选择样本，必须把每个阶段分离开来，独立地按当前阶段引入数据框架和选择参数。

SURVEYSELECT 过程主要提供方法服务于等概率抽样法(Equal Probability Sampling)和按初

作者单位：200025 上海市，中国疾病预防控制中心寄生虫病预防控制所

级单位含量比例的概率抽样法(Probability Proportional To Size Sampling 简称 PPS)。等概率抽样法中在每个被抽样的范围中或某一层中的任何一个单位样本有相同的选择概率;而在按初级单位含量比例的概率抽样法中一个单位样本的选择概率是由初级单位含量比例决定的,详细的关于概率抽样的方法可参见 Kish (1987 和 1965)、Kalton (1983) 和 Cochran(1977)的文章。

SURVEYSELECT 过程为等概率抽样法提供如下方法:

1. 单纯随机抽样
2. 无限制随机抽样
3. 系统随机抽样
4. 序列随机抽样

SURVEYSELECT 过程为按初级单位含量比例的概率抽样法提供如下方法:

1. 无替代 PPS 抽样
2. 替代 PPS 抽样
3. PPS 系统抽样
4. 每层选择二个单位的 PPS 算法
5. 最少替代的 PPS 序列抽样

SAS 中 SURVEYSELECT 过程为以上过程提供快而有效的算法,从而使它在较大的输入数据集或样本结构中表现优异,实践中此方法常被用于大范围样本调查。

SURVEYSELECT 过程能施行分层抽样,同时可在特别层或调查总体的非重叠亚组之间独立的选择样本;分层就是按某种特征把调查对象分为若干类型、部分或区域的层(Strata),它控制着各层样本大小的分布,此方法被许多调查所采用。例如,分层能保证研究者有兴趣而又实际上相对较小的亚组有足够的样本含量,同时能提高总的估计值的准确性。在使用系统或序列选择方法时,SURVEYSELECT 过程对不太明显分层的影响是通过层之间的控制变量的排序来实现的。

当总样本为相同数据集组成,同时选择方法一致时,SURVEYSELECT 过程能提供重复抽样,它可被用于研究非抽样误差对变量的影响,特别是由不同随访者得到的结果,对于混合样本估计值可用重复来计算标准差。

二、SURVEYSELECT 过程的常用句法

SURVEYSELECT 过程的常用句法有:PROC SURVEYSELECT <选择项>;SIZE 变量;STRATA 变量;CONTROL 变量和 ID 变量。

1. PROC SURVEYSELECT 语句<选择项>中主要定义进出数据集、特别的抽样方法、样本大小、抽样比和其它的抽样参数。

抽样方法包括: ① SRS 单纯随机抽样,② URS 无限制随机抽样, ③ SYS 系统随机抽样, ④ SEQ 序列随机抽样, ⑤ PPS 无替代 PPS 抽样, ⑥ PPS_WR 替代 PPS 抽样, ⑦ PPS_SYS PPS 系统抽样, ⑧ PPS_SEQ PPS 序列抽样, ⑨ PPS_BREWER、PPS_MURTHY、PPS_SAMPFORD。

每层选择二个单位的 PPS 算法,Brewer、Murthy 和 Sampford 法。

2. SIZE 语句定义在按初级单位含量比例的概率抽样法中的包含特定样本大小的变量。

3. STRATA 语句定义一个或多个分层变量。

4. CONTROL 语句主要用在序列抽样方法中,定义特别的一个或多个层间排序的变量。

5. ID 语句定义被选中样本变量,这些变量由输入数据集复制到输出数据集。

三、抽样结果

第三次全国血吸虫病流行病学抽样调查采用分层、整群、随机抽样的方法,分层是按血吸虫病流行特点来分的, 首先把仍有血吸虫病流行的七省(湖北、湖南、江西、安徽、江苏、四川和云南)作为主层;然后根据流行类型划分为 8 个第一亚层:湖沼型流行区包括湖汊亚型、洲滩亚型、洲垸亚型和垸内亚型,水网型流行区的水网亚型,山丘型流行区包括丘陵亚型、高山峡谷亚型和平坝亚型;接着确定第二亚层: 具体做法是在第一亚层的基础上,由流行区县(市、区)血防所(站)根据地理环境、最近一次查病结果、钉螺分布现状以及多年防治经验,将各流行村的居民血吸虫估计感染率分为<1%、1%~、5%~、10%~ 等 4 个层次; 最后每层按整群随机抽样法抽取 1%的行政村作为第三次全国流调调查点进行人、畜血吸虫病调查。具体过程如下:

1. 各省以数据库形式上报‘抽样前流行村基本情况表”,表中包含流行村的基本情况,如村名、人口数、流行类型和估计感染率等,见表 1。

2.核对数据库的正确性,修正数据并按流行类

表 1 某省抽样前流行村基本情况表

县、市国标码	县、市	乡镇	村名	村人口数	村耕牛存栏数	流行类型	村估计感染率
6232900	余干县	梅溪乡	渔业队	471	7	11	4
6232900	余干县	梅溪乡	邹家	1329	209	11	4
6012200	新建县	铁河乡	场属	1365	60	12	1
6012200	新建县	昌邑乡	乡属	1251	35	12	1
6012400	进贤县	三里乡	东岸村	1400	150	12	1
6012400	进贤县	三里乡	前进村	1737	30	12	1
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.

注：流行类型— 10.湖沼型 (11.湖汊亚型 12.洲滩亚型 13.洲垸亚型 14.垸内亚型)； 20.水网型； 30.山丘型 (31.平坝亚型 32.高山峡谷亚型 33.丘陵亚型)

估计感染率— 1. <1% 2. 1%~ 3. 5%~ 4. 10%~

型和估计感染率排序。

3.用 SAS 调入数据库，计算出按流行类型和估计感染率分层的样本数，以 1%的比例抽取确定抽样数；原则上 100 个行政村抽取一个点，实际操作中可按四舍五入适当扩大样本，如 150~200 个行政村中，可抽取 2 个样本点。

4. 根据表 2 结果按 SAS 的 SURVEYSELECT 过程编写程序如下：

```
PROC SURVEYSELECT DATA=WORK.DH
OUT=SAMPLE
METHOD=SRS N=(2,1,1,2,2,1,4,4,1);
STRATA 流行类型 估计感染率;
PROC PRINT;
```

表 2 某省按流行类型和估计感染率分层的样本数和抽样数

流行类型		估计感染率			
		<1%	1%	5%	合计
平坝亚型	样本数	150	106	57	313
	抽样数	2	1	1	4
洲滩亚型	样本数	211	236	120	567
	抽样数	2	2	1	5
山丘型丘陵亚型	样本数	448	390	69	894
	抽样数	4	4	1	9
合 计	样本数	809	732	246	1787
	抽样数	8	7	3	18

WORK.DH 为输入数据集及上面的表 1，SAMPLE 为包含被选择样本的输出数据集，方法是单纯随机抽样 (SRS)，N 为各层应抽样本的大小；分层则是先按流行类型再按估计感染率进行；最后运算出结果如表 3。

按以上四个步骤分别随机抽取样本，江苏、江西、安徽、湖南、湖北、云南和四川分别为 13、23、18、47、58、12 和 68 个样本点；共抽取 239 个样本村，占未达到传播阻断标准乡镇的所有流行村的 1.36%；各省详细的分层抽样结果见表 4。

四、讨论

SAS 统计软件的 SURVEYSELECT 过程是计算机和现代统计学结合发展的结果，它极大地丰富

表 3 某省计算机抽样结果

流行类型	估计感染率	国际码	县、市	乡镇	村名	人口数	选择概率	抽样的权重
湖汊亚型	<1%	340826000	宿松县	五里	黎圩	1238	0.0133	75
湖汊亚型	<1%	340827000	望江县	泊湖	团山	788	0.0133	75
湖汊亚型	1%~	341721000	东至县	香隅	张湾	1152	0.0094	106
湖汊亚型	5%~	341702000	贵池区	涓桥镇	桂畈村	2185	0.0175	57
洲滩亚型	<1%	340827000	望江县	泊湖	正兴	486	0.0095	105.5
洲滩亚型	<1%	341721000	东至县	瓦垅	瓦垅	2237	0.0095	105.5
洲滩亚型	1%~	341422000	无为县	白茆镇	同兴	860	0.0085	118
洲滩亚型	1%~	340827000	望江县	杨湾	团结	1005	0.0085	118
洲滩亚型	5%~	340823000	枞阳县	风仪	雁翎	1320	0.0083	120
丘陵亚型	<1%	341802000	宣州区	华阳	东溪	1092	0.0089	112
丘陵亚型	<1%	341823000	泾 县	童疃	花园	1325	0.0089	112
丘陵亚型	<1%	340221000	芜湖县	花桥	东门	2733	0.0089	112
丘陵亚型	<1%	340221000	芜湖县	咸保	北斗	1020	0.0089	112
丘陵亚型	1%~	340881000	桐城市	挂车河	挂镇	1693	0.0103	97.5
丘陵亚型	1%~	340826000	宿松县	长铺	捉马	1815	0.0103	97.5
丘陵亚型	1%~	341802000	宣州区	杨柳	合山	1285	0.0103	97.5
丘陵亚型	1%~	341823000	泾 县	丁家桥	官庄	915	0.0103	97.5
丘陵亚型	5%~	341802000	宣州区	孙埠	正兴	3654	0.0145	69

表 4 七省按流行类型和估计感染率分层的抽样结果

省名	估计感染率	流行类型								
		湖汉亚型	洲滩亚型	洲垸亚型	垸内亚型	水网型	平坝亚型	高山峡谷亚型	丘陵亚型	合计
江苏	<1%		1			5			3	9
	1%~		3			1				4
	5%~									
	10%~									
	小计		4			6			3	13
江西	<1%	1	1	2					1	5
	1%~	1	1	2					4	8
	5%~	1	1	1					2	5
	10%~	1	2	1					1	5
	小计	4	5	6					8	23
安徽	<1%	2	1						5	8
	1%~	1	2						4	7
	5%~	1	1						1	3
	10%~									
	小计	4	4						10	18
湖南	<1%	2		5	5				1	13
	1%~	1		8	5				1	15
	5%~	1		7	2				1	11
	10%~	1	2	4	1					8
	小计	5	2	24	13				3	47
湖北	<1%		3		14				5	22
	1%~		4		12				1	17
	5%~		1		9				1	11
	10%~		1		6				1	8
	小计		9		41				8	58
云南	<1%						2	1		3
	1%~						2	1		3
	5%~						1	2		3
	10%~						1	2		3
	小计						6	6		12
四川	<1%						28	1	10	39
	1%~						6	3	7	16
	5%~						2	1	6	9
	10%~						1	1	2	4
	小计						37	6	25	68
合计		13	24	30	54	6	43	12	57	239

了现场流行病学，它为流行病学的现场调查提供了简单而快速的样本选择方法。以上过程是其在全国血吸虫病流行病学抽样调查中的初步运用，可以大致分为以下几个步骤：1.确定抽样方法，2.完成基本数据库，3.根据设计要求确定各层样本大小，4.在计算机上利用 SURVEYSELECT 过程完成抽样。

SURVEYSELECT 过程是 SAS 公司在其统计软件的第七版中加入的较为复杂的样本选择与计算的过程之一，有着很强的数理统计基础；我们只是尝试部分运用其抽样过程，不足之处谨供商榷。

参 考 文 献

1 Kish, L.(1965), Survey Sampling, New York: John Wiley & Sons, Inc.

2 Kish, L.(1987), Statistical Design for Research,, New York: John Wiley & Sons, Inc.

3 Kalton, G.(1983), Introduction to Survey Sampling, Sage University Paper series on Quantitative Applications in the Social Sciences, series no.07-035, Beverly Hills and London: Sage Publications, Inc.

4 Cochran, W.G.(1977), Sampling Techniques, Third Edition, New York: John Wiley & Sons, Inc.

5 郭祖超主编. 医用数理统计方法. 第 3 版. 北京：人民卫生出版社，1988 年 10 月.

6 金丕焕主编. 医用统计方法. 上海医科大学出版社. 1993 年 2 月.

7 SAS Institute Inc.(2000),SAS procedures Guide, Version 8, Cary, NC: SAS Institute Inc.

8 SAS Institute Inc.(2000),SAS/STAT User's Guide, Version 8, Cary, NC: SAS Institute Inc.

（收稿日期 2006-03-15 编辑 李启扬）

（上接第 86 页）

病学史及仔细的体格检查是减少恙虫病漏诊误诊的关键。夏秋季持续高热，一般抗菌素治疗无效，应常规做外斐氏试验以排除恙虫病。在流行地区，流行季节，不明原因持续高热超过 1 周，可疑恙虫病者，可予强力霉素作诊断性治疗。本组病例肝损害发生率高达 86.6%。均表现为 ALT 及 AST 升高，黄疸少见，经治疗后肝功能多在 2～4 周内恢复，心电图提示心肌损害，心律失常均在体温正常后消失，表明其实质器官损害是一过性的。

参 考 文 献

1 陈香蕊. 恙虫病和恙虫病立克次体. 北京：军事医学科学出版社，2001,76-81.

2 罗红涛. 恙虫病多脏器损害 90 例临床分析. 中国人畜共患杂志，1997,13:80-81.

3 陈文治，曾俊涛. 恙虫病 175 例临床分析. 中国热带医学杂志，2005,5(6):26.

（收稿日期 2006-04-05 编辑 李启扬）