

◇ 预防医学研究 ◇

蒙特卡洛模拟对分层整群抽样下样本轮换的评价

傅颖, 高歌, 刘升学, 周云华, 施佳琛

摘要 目的 采用蒙特卡洛模拟(SAS 编程)对分层整群抽样下样本轮换连续调查的方法和公式进行信度、效度和精度评价。方法 以嘉兴市中小学生肺活量的调查结果模拟总体参数,各在抽样比为 10% 和 40% 时应用 SAS 编程进行分层整群抽样下样本轮换的蒙特卡洛模拟,分别模拟 100 个样本并按公式计算获得 100 个总体均数的 95% CI,采用 95% CI 法比较模拟计算出的总体均值估计量与总体均数,以评价该调查方法及其公式的信度、效度。另外,分别在两种抽样比下模拟分层整群抽样下样本轮换连续调查和固定样本连续调查各 100 个样本并比较两种方法下 100 个样本均值的标准差,以评价抽样精度。结果 两种抽样比下模拟总体均数几乎均被包含在总体均值的 95% CI(仅在 10% 的抽样比下有两次不包含);两种抽样比分层整群抽样下样本轮换连续调查的样本均值的标准差较固定样本的均值标准差小。结论 分层整群抽样下样本轮换的方法及其统计公式有较高的信度、效度和精度,今后可应用于医学领域的连续性调查。

关键词 蒙特卡洛模拟;样本轮换;信度;分层整群抽样

中图分类号 R 181.2

文献标志码 A **文章编号** 1000-1492(2014)11-1618-04

科学调研中为了了解总体在不同时间上的水平、变化或累计,常采用连续调查法,如城市住户调查^[1]、营养状况调查^[2]和慢性病哨点监测^[3]等。其中样本轮换下的连续调查因规避了固定样本的样本老化、样本疲劳和全新样本取样成本高等缺点而被广泛应用。流行病学动态队列连续调查中,因调查时期长样本脱落较多,需要设计合理的样本轮换方法并提供正确的统计公式^[4]。为了评估一种调查方法及其公式能否真实可靠地得到所需结果,只对一次调查进行讨论是不够的,需要对相同或近似条件下大量重复调查进行综合分析。鉴于此,该研究采用分层整群抽样下样本轮换的方法连续调查嘉兴

市中小学生的肺活量,并运用蒙特卡洛模拟(SAS 编程)对该调查方法及其公式进行评价。

1 统计方法与调查实例

1.1 分层整群抽样下样本轮换 设将总体分成 L 个层,各层群的个数分别为 $N_1, N_2, \dots, N_L$, $\sum_{k=1}^L N_k = N$ 。各层独立地在整群抽样下进行样本轮换,在第 1 次第 k 层独立抽取 n_k 个群, $\sum_{k=1}^L n_k = n$ 。 m_{hk} 代表 h 次 k 层保留的样本群数, u_{hk} 代表 h 次 k 层轮换的样本群数,对任何 h , $m_{hk} + u_{hk} = n_k$, $k = 1, 2, \dots, L$,把这两部分抽中的群放一起并观测所有的次级单元。 $\bar{Y}_{hk}$ 是 h 次 k 层次级单位的总体均值; $\bar{Y}_{hkm}$ 、 $\bar{Y}_{hku}$ 分别为 h 次 k 层抽样保留部分、轮换部分中次级单位的样本均值; $\bar{Y}_{h-1,km}$ 为 h 次 k 层保留部分在 $h-1$ 次调查时次级单位的样本均值; ρ_{hk} 、 b_{hk} 分别为 k 层 h 次与 $h-1$ 次调查同一批对象在相同指标变量上的相关系数、回归系数; S_{hk}^2 代表 h 次 k 层次级单位的方差。

1.2 统计公式^[5]

1.2.1 总体均值的估计量

第 h 次总体均值的估计量为: $\hat{\bar{Y}}_h = \sum_{k=1}^L W_k \hat{\bar{Y}}_{hk}$ (1)

其中,层权 $W_k = \frac{N_k}{N}$ (2)

$\hat{\bar{Y}}_{hk} = \Phi_{hk} \bar{Y}_{hku} + (1 - \Phi_{hk}) [\bar{Y}_{hkm} + b_{hk} (\hat{\bar{Y}}_{h-1,k} - \bar{Y}_{h-1,km})]$ (3)

1.2.2 总体均值估计量的方差

h 次总体均值估计量的方差:

$V(\hat{\bar{Y}}_h) = \sum_{k=1}^L W_k^2 V(\hat{\bar{Y}}_{hk})$ (4)

其中, $V(\hat{\bar{Y}}_{hk}) = \Phi_{hk}^2 \frac{S_{hk}^2}{u_{hk}} + (1 - \Phi_{hk})^2 [\frac{S_{hk}^2 (1 - \rho_{hk}^2)}{m_{hk}} + \frac{\rho_{hk}^2 S_{hk}^2}{n_k}]$ (5)

1.2.3 最优权数和最优样本轮换率 在第 k 层采取整群抽样下的样本轮换,以群为观察单位, h 次 k 层最优权数为:

$\Phi_{hk} = \frac{(1 - \rho_{hk}^2) / m_{hk} + \rho_{hk}^2 / n_k}{(1 - \rho_{hk}^2) / m_{hk} + \rho_{hk}^2 / n_k + 1 / u_{hk}}$ (6)

2014-04-09 接收

基金项目:国家自然科学基金(编号:81273188)

作者单位:苏州大学医学部公共卫生学院流行病与卫生统计学教研室,苏州 215123

作者简介:傅颖,女,硕士研究生;

高歌,男,教授,博士生导师,责任作者, E-mail: gaoge@suda.edu.cn

$$h \text{ 次 } k \text{ 层 最优样本轮换率: } \frac{u_{hk}}{n_k} = [1 + \left(\frac{1 - \sqrt{1 - \rho_{hk}^2}}{1 + \sqrt{1 - \rho_{hk}^2}} \right)^{h-1}] / 2 \quad (7)$$

1.3 调查实例 于2010年10月和2011年6月采用分层整群抽样下样本轮换的方法连续调查嘉兴市一所小学和一所中学全体在校学生的肺活量2次,共18个小学班和24个中学班,以小学、中学为两层,以班级为群进行调查。根据样本代表性及人力物力财力的综合考量,2010年第一次($h=1$)调查,按50%的抽样比分别从小学和中学随机抽取9个班和12个班,所抽取班的学生全部进入样本。2011年第二次($h=2$)调查,先进行预调查,抽取少许与第一次调查相同的样本单元以估计相关系数(ρ)和回归系数(b)。由式(7)可得小学层最优样本轮换率为0.72,即应保留3个班轮换6个班,中学层最优样本轮换率为0.62,即应保留4个班轮换8个班,然后实施样本轮换的正式调查。两层两次共调查中小學生1 971人次。2010年调查得中小學生肺活量总体均值估计量、估计方差、95% CI (非样本轮换,按照普通分层整群抽样的公式计算而得)分别为1 907 ml、757 203、(202, 3 612) ml;由公式(1)~(7)得2011年中小學生肺活量总体均值估计量、估计方差、95% CI 分别为:2 220 ml、6 174.4、(2 065, 2 374) ml。

2 蒙特卡洛模拟

2.1 构建模拟总体 以中小學生肺活量第一、二次调查结果作为模拟总体参数,分两层(小学生和中学生)模拟。第一次调查两层的模拟总体均值分别为989、2 596 ml,模拟标准差分别为273.3、403.3 (调查数据直接计算所得);第二次调查两层的总体模拟均值分别为1 315、2 903 ml,模拟标准差分别为138.5、89.9 (按本文样本轮换的公式计算所得)。模拟总体人数为100 000,小学层和中学层的层权分别为0.43、0.57,两层共包含1 000个群。

2.2 模拟分层整群抽样下样本轮换的连续调查 利用模拟总体,第一次分别以10%、40%的抽样比在各层中抽取部分群,第二次保持抽样比不变,在各层中按最优样本轮换率公式(7)计算出轮换班个数和保留班个数后,实施样本轮换,以此模拟两种抽样比下各100个样本。

2.3 效度和信度评价 根据模拟总体计算得第二次调查总体均值为2 222 ml。在抽样比为10%、40%时对模拟总体分别进行100次模拟样本轮换的分层整

群抽样,并对每一样本按本文公式计算,得200个样本均值及总体均值的95% CI ,结果见表1、2。在10%

表1 计算机模拟结果(抽样比为10%)

样本号	样本均值	95%可信下限	95%可信上限	样本号	样本均值	95%可信下限	95%可信上限
1	2 220	2 214	2 226	51	2 222	2 216	2 228
2	2 220	2 214	2 226	52	2 220	2 215	2 225
3	2 223	2 218	2 228	53	2 224	2 218	2 230
4	2 226	2 221	2 231	54	2 222	2 217	2 227
5	2 221	2 215	2 227	55	2 226	2 219	2 233
6	2 220	2 214	2 226	56	2 223	2 217	2 229
7	2 217	2 212	2 222	57	2 220	2 215	2 225
8	2 224	2 218	2 230	58	2 224	2 218	2 230
9	2 227	2 221	2 233	59	2 219	2 213	2 225
10	2 224	2 218	2 230	60	2 222	2 215	2 229
11	2 225	2 219	2 231	61	2 224	2 218	2 230
12	2 227	2 221	2 233	62	2 222	2 216	2 228
13	2 225	2 218	2 232	63	2 220	2 214	2 226
14	2 222	2 216	2 228	64	2 222	2 216	2 228
15	2 226	2 221	2 231	65	2 217	2 212	2 222
16	2 220	2 214	2 226	66	2 223	2 217	2 229
17	2 220	2 215	2 225	67	2 220	2 214	2 226
18	2 221	2 215	2 227	68	2 224	2 218	2 230
19	2 221	2 216	2 226	69	2 221	2 215	2 227
20	2 225	2 220	2 230	70	2 224	2 219	2 229
21	2 229*	2 224	2 234	71	2 220	2 214	2 226
22	2 219	2 213	2 225	72	2 218	2 212	2 224
23	2 219	2 214	2 224	73	2 226	2 220	2 232
24	2 221	2 214	2 228	74	2 224	2 219	2 229
25	2 221	2 216	2 226	75	2 221	2 216	2 226
26	2 223	2 217	2 229	76	2 221	2 215	2 227
27	2 223	2 217	2 229	77	2 225	2 220	2 230
28	2 218	2 213	2 223	78	2 223	2 217	2 229
29	2 229*	2 224	2 234	79	2 224	2 218	2 230
30	2 226	2 220	2 232	80	2 224	2 218	2 230
31	2 222	2 216	2 228	81	2 222	2 217	2 227
32	2 219	2 213	2 225	82	2 219	2 213	2 225
33	2 223	2 217	2 229	83	2 226	2 220	2 232
34	2 220	2 215	2 225	84	2 220	2 214	2 226
35	2 222	2 216	2 228	85	2 222	2 217	2 227
36	2 223	2 217	2 229	86	2 221	2 216	2 226
37	2 223	2 217	2 229	87	2 222	2 217	2 227
38	2 223	2 217	2 229	88	2 226	2 220	2 232
39	2 223	2 218	2 228	89	2 223	2 216	2 230
40	2 223	2 217	2 229	90	2 219	2 213	2 225
41	2 219	2 213	2 225	91	2 225	2 219	2 231
42	2 225	2 219	2 231	92	2 219	2 214	2 224
43	2 220	2 215	2 225	93	2 220	2 214	2 226
44	2 224	2 218	2 230	94	2 221	2 215	2 227
45	2 222	2 216	2 228	95	2 222	2 216	2 228
46	2 225	2 219	2 231	96	2 223	2 217	2 229
47	2 225	2 219	2 231	97	2 223	2 218	2 228
48	2 220	2 214	2 226	98	2 222	2 216	2 228
49	2 221	2 215	2 227	99	2 227	2 221	2 233
50	2 223	2 218	2 228	100	2 220	2 214	2 226

*:总体均值的95% CI 不包含总体均值

表2 计算机模拟结果(抽样比为40%)

样本号	样本 均值	95%可 信下限	95%可 信上限	样本号	样本 均值	95%可 信下限	95%可 信上限
1	2 221	2 218	2 224	51	2 222	2 219	2 225
2	2 222	2 219	2 225	52	2 222	2 219	2 225
3	2 224	2 221	2 227	53	2 223	2 220	2 226
4	2 221	2 218	2 224	54	2 223	2 220	2 226
5	2 222	2 219	2 225	55	2 219	2 216	2 222
6	2 224	2 221	2 227	56	2 222	2 219	2 225
7	2 223	2 220	2 226	57	2 223	2 220	2 226
8	2 223	2 220	2 226	58	2 222	2 219	2 225
9	2 222	2 219	2 225	59	2 224	2 221	2 227
10	2 223	2 220	2 226	60	2 224	2 221	2 227
11	2 223	2 220	2 226	61	2 221	2 218	2 224
12	2 223	2 220	2 226	62	2 222	2 219	2 225
13	2 221	2 218	2 224	63	2 220	2 217	2 223
14	2 221	2 218	2 224	64	2 223	2 220	2 226
15	2 225	2 222	2 228	65	2 224	2 221	2 227
16	2 222	2 219	2 225	66	2 221	2 218	2 224
17	2 220	2 217	2 223	67	2 225	2 222	2 228
18	2 222	2 219	2 225	68	2 221	2 218	2 224
19	2 221	2 218	2 224	69	2 221	2 218	2 224
20	2 222	2 219	2 225	70	2 222	2 219	2 225
21	2 221	2 218	2 224	71	2 222	2 219	2 225
22	2 222	2 219	2 225	72	2 224	2 221	2 227
23	2 221	2 218	2 224	73	2 223	2 220	2 226
24	2 222	2 219	2 225	74	2 223	2 220	2 226
25	2 223	2 220	2 226	75	2 222	2 219	2 225
26	2 222	2 219	2 225	76	2 225	2 222	2 228
27	2 221	2 218	2 224	77	2 221	2 218	2 224
28	2 222	2 219	2 225	78	2 221	2 218	2 224
29	2 223	2 220	2 226	79	2 222	2 219	2 225
30	2 221	2 218	2 224	80	2 221	2 218	2 224
31	2 222	2 219	2 225	81	2 223	2 220	2 226
32	2 221	2 218	2 224	82	2 222	2 219	2 225
33	2 222	2 219	2 225	83	2 223	2 220	2 226
34	2 221	2 218	2 224	84	2 222	2 219	2 225
35	2 222	2 219	2 225	85	2 221	2 218	2 224
36	2 223	2 220	2 226	86	2 223	2 220	2 226
37	2 221	2 218	2 224	87	2 221	2 218	2 224
38	2 224	2 221	2 227	88	2 222	2 219	2 225
39	2 222	2 219	2 225	89	2 221	2 218	2 224
40	2 222	2 219	2 225	90	2 223	2 220	2 226
41	2 222	2 219	2 225	91	2 222	2 219	2 225
42	2 223	2 220	2 226	92	2 222	2 219	2 225
43	2 223	2 220	2 226	93	2 221	2 218	2 224
44	2 223	2 220	2 226	94	2 222	2 219	2 225
45	2 223	2 220	2 226	95	2 223	2 220	2 226
46	2 223	2 220	2 226	96	2 223	2 220	2 226
47	2 221	2 218	2 224	97	2 222	2 219	2 225
48	2 223	2 220	2 226	98	2 222	2 219	2 225
49	2 224	2 221	2 227	99	2 221	2 218	2 224
50	2 221	2 218	2 224	100	2 222	2 219	2 225

* :总体均值的95% CI 不包含总体均值

的抽样比下,100 次模拟中只有两次(样本 21、29)得到的总体均值 95% CI 不包含总体均值;在 40% 的抽样比下,100 次模拟中所有总体均值 95% CI 均包含总体均值。两种抽样比下,100 个样本均数几乎均接近总体均数(模拟真值),表明分层整群抽样下样本轮换的方法及其参数估计公式均具有较高的效度;又因 100 个样本均值几乎均接近同一数值(总体均数),表明此调查方法及其公式具有较高的信度。

2.4 样本轮换与固定样本抽样精度比较 对于固定样本的抽样方式,使用 SAS 编程,两种抽样比下各抽样 100 次,按算术平均数计算得 100 个样本均值,再计算 100 个样本均值的均值及标准差(均数的标准误);对于样本轮换的抽样方式,利用表 1、2 的均值结果,计算 100 个样本均值的均值及标准差(均数的标准误)。结果如表 3 所示,在 10% 和 40% 抽样比下,分层整群抽样下样本轮换的抽样精度均较固定样本的抽样精度高(样本轮换的 100 个样本均值间的标准差较小)。

表3 分层整群抽样下样本轮换与固定样本抽样精度的比较

抽样方法	10% 抽样比		40% 抽样比	
	均值	标准差	均值	标准差
固定样本	2 223	3.097	2 222	1.241
样本轮换	2 222	2.531	2 222	1.123

3 讨论

随着社会经济的发展,样本轮换的应用领域越来越广,几乎可适用于所有长期性连续性抽样调查^[8]。样本轮换的调查方法,是对全新样本和固定样本的折中,既保留了部分原有样本单位,又增加了部分新样本单位,兼有全新样本与固定样本的优点,能在抽样费用与抽样精度之间取得平衡,是减少和控制非抽样误差的重要手段^[9]。

连续性抽样调查具有广阔的应用前景,但目前其理论研究相对不够成熟和完善,在很大程度上限制了其在统计调查领域的推广应用^[8]。目前关于分层整群抽样下样本轮换的理论研究并不缺乏,但对该方法及其公式的实证评价相对少见。

蒙特卡洛模拟也称计算机随机模拟,是基于“随机数”的计算方法,在设定的概率分布内通过随机抽样进行分析的过程,通过随机数发生器采用随

机抽样的方式抽取满足输入变量要求的数据^[8]。对于统计学上难以从理论上进行证明的规律,蒙特卡洛法可以作大量模拟,以无限逼近的方法进行验证;而且其算法复杂度仅仅与模拟次数有关,能够比较直观地描述具有随机性质的事物的特点,程序结构简单,易于编制和调试^[9],是一种值得在医学领域里推广的数据处理方法。

该研究以嘉兴市中小學生肺活量调查为例,使用 SAS 编程对分层整群抽样下样本轮换的连续调查及其参数估计公式进行了评价,表明该方法及其公式具有良好的信度、效度和精度,具有较大的实际应用意义,今后可推广应用于医学领域的连续性抽样调查。

参考文献

- [1] 国家统计局深圳调查队. 深圳住户中高收入专项调查探索 [J]. 中国统计, 2013, 30(1): 11-3.
- [2] 朱跃平, 丁 福. 老年住院患者营养风险筛查及营养支持状况 [J]. 中国老年学杂志, 2013, 33(11): 2609-11.
- [3] 汤 琰. 哨点监测人群中人类免疫缺陷病毒、梅毒、丙肝病毒感染的检测结果分析 [J]. 诊断学理论与实践, 2013, 12(3): 352-4.
- [4] Sen A R. Successive sampling with p auxiliary variables [J]. ANN APPL STAT, 1972, 43: 2031-4.
- [5] 李善斌, 解 燕. 样本轮换在分层整群中的应用 [J]. 苏州大学学报(自然科学版), 2009, 25(4): 10-3.
- [6] 徐国祥, 王 芳. 连续性抽样调查中的样本轮换研究 [J]. 统计研究, 2011, 28(5): 89-96.
- [7] 杜子芳, 刘爱芹. 样本轮换的实施效果分析 [J]. 统计研究, 2002, 19(6): 38-40.
- [8] 陈光慧. 连续性抽样调查理论研究综述 [J]. 数理统计与管理, 2012, 31(3): 426-33.
- [9] 金 胜, 胡福祥. 基于蒙特卡洛模拟的矿山投资风险分析 [J]. 现代矿业, 2013, 29(5): 148-51.
- [10] 杜俏俏, 高 歌. 蒙特卡洛模拟在多分类敏感问题两阶段整群抽样信度与效度评价中的应用 [J]. 中国卫生统计, 2013, 30(2): 217-20.

Evaluation based on Monte Carlo simulation for sample rotation under stratified cluster sampling

Fu Ying, Gao Ge, Liu Shengxue, et al

(Dept of Public Health, Medical College of Soochow University, Suzhou 215123)

Abstract Objective To evaluate the reliability, validity and precision of the successive survey using sample rotation under stratified cluster sampling and its formulae based on Monte Carlo simulation. **Methods** Survey results of some students in Jiaying about vital capacity were taken as simulative population parameters. Monte Carlo simulation was conducted with SAS programming when sampling ratios were 10% and 40% respectively. 100 samples in each sampling ratio were simulated and 95% CI (confidence interval) could be acquired. In order to evaluate reliability and validity of the survey and its formulae, simulated overall mean estimators were compared with population mean using 95% CI. In order to assess the sampling precision, 100 samples were respectively simulated for rotation survey and panel survey at each ratio (10% and 40%) and their standard errors were represented. **Results** Almost all the 95% CI of 100 estimated overall means at each sampling ratio (10% and 40%) included the population mean except for two at 10%. The standard error of rotation survey was lower than panel survey. **Conclusion** Sample rotation under stratified cluster sampling and its formulae are of high reliability, validity and precision, which can be applied to the future medical successive survey.

Key words Monte Carlo simulation; sample rotation; reliability; stratified cluster sampling