连续性抽样中最优样本轮换率的确定

叶桂芳

（暨南大学 经济学院，广州 510632）

摘　要：在使用样本轮换的连续性抽样中，依据相邻两期的相关性可以对拼配样本构造组合估计量，还可以利用辅助变量进一步提高现期估计量的精度。文章在此基础上使用连续两期的辅助信息构造指数形式的比率估计量，运用最优化方法计算出最优样本轮换率和最优权重系数，使得总体均值的回归组合估计量的均方误差最小，最后运用模拟数据验证其可以更大程度上提高连续性抽样的估计精度。

关键词：连续性抽样；最优样本轮换率；指数；均方误差

中图分类号：C811　　**文献标识码**：A　　**文章编号**：1002-6487(2015)10-0004-03

0　引言

在实际社会经济中，为了研究社会经济现象随时间的动态变化趋势，传统的一次性调查已不能满足人们对统计信息的有效需求。为了能够及时反映调查总体的变化和发展，取而代之的是连续性抽样调查，比如美国的现时人口调查、加拿大的劳动力调查、我国的城市住户和农村住户调查等。作为使用最为广泛的样本轮换抽样方法，几乎可适用于所有的长期连续性抽样调查。在连续性样本轮换的抽样调查研究中有两个问题居于中心地位：一个是样本轮换率的确定；另一个是样本轮换模式下估计量的构造。由于影响样本轮换率的因素很多，如调查的目的、调查总体的变化速度、被调查者的心里接受程度、调查费用、调查精度等等。国内外学者在舍弃不可量化的影响因素后，对样本轮换率问题进行了大量的研究。

样本轮换的构想，最早是由美国统计学家杰R. J. Jessen(1942)在收集农场调查数据时提出的。W. G. Cochran(1985)归纳总结了前人的研究结果，讨论了在不考虑影响样本轮换率的一些不可量化的因素下，分别对考虑调查费用和不考虑调查费用的简单随机抽样下的样本轮换率进行了研究。A. R. Sen(1973)利用前期和现期样本拼配部分的辅助信息构造合适的估计量，冯士雍，邹国华(1996)不仅利用前期和现期样本拼配部分的辅助信息还考虑拼配样本以外的样本单元的辅助信息对Sen提出的估计量进行了改进。G. N. Singh(2001)利用现期的辅助信息和前期的样本信息提出两阶段连续性抽样下的估计量，2003年拓展到多阶段连续性抽样下的估计量，充分利用前期辅助信息进一步提高了估计精度。徐国祥，王芳(2011)从调查总体的特征出发，讨论了分层抽样下的最优样本轮换率和轮换效果问题，从实证角度对上海市城镇住房空置率抽样调查数据进行分析。本文在前人研究的基础上，依据连续性抽样中相邻两期样本单元之间的相关性和研究变量的辅助信息构造使得均方误差最小的指数形式的回归组合估计量，在保证估计精度的前提下求得最优样本轮换率，尽可能的节约调查成本，最后通过设定模型参数产生一系列模拟数据进行数值分析。

1　连续两期样本下最优样本轮换率的确定

1.1　回归组合估计量的构造

假定对包含 N 个总体单元的调查总体 $U=(U_1,U_2,...U_N)$ 进行连续性抽样调查，前后两期中抽取的样本单元保持不变，记为 n。前一期调查中保留下来的样本作为拼配样本 $m=n\lambda$，现期调查中被替换成新的样本单元作为非拼配样本 $u=n\mu$，$\lambda+\mu=1$，其中 μ 表示现期抽样调查的样本轮换率。$y_0(y_1)$ 分别表示前期（现期）的研究变量，其总体均值记为 $\bar{Y}_0(\bar{Y}_1)$，总体方差记为 $S^2_{y_0}(S^2_{y_1})$，样本方差记为 $s^2_{y_0}(s^2_{y_1})$，拼配样本的方差记为 $s^2_{y_{0m}}(s^2_{y_{1m}})$。$z$ 是已知总体均值 $\bar{Z}$ 和总体方差 S^2_z 并与研究变量 $y_0(y_1)$ 有一定相关性的辅助变量，这里假设辅助信息比较稳定，前后期保持不变。同时给出如下记号：

$\bar{y}_{0n}(\bar{y}_{1n})$：前期（现期）研究变量 $y_0(y_1)$ 的样本均值；

$\bar{y}_{0m}(\bar{y}_{0u})$：前期研究变量 y_0 的拼配（非拼配）样本的样本均值；

$\bar{y}_{1m}(\bar{y}_{1u})$：现期研究变量 y_1 的拼配（非拼配）样本的样本均值；

$\bar{z}_m(\bar{z}_u)$：连续两期中辅助变量 z 的拼配（非拼配）样本的样本均值；

$\bar{z}_n$：辅助变量 z 的样本均值；

$\mathrm{Cov}(y_{0m},y_{1m})$：研究变量 $y_0(y_1)$ 中拼配样本的协方差；

$\mathrm{Cov}(y_0,y_1)$：研究变量 $y_0(y_1)$ 的总体协方差；

作者简介：叶桂芳（1989-），女，安徽安庆人，硕士研究生，研究方向：统计调查与分析。

$\rho_{y_0y_1},\rho_{y_1z},\rho_{y_0z}$：表示各研究变量、辅助变量之间的相关系数。

本文的目的在于利用前期的样本信息和总体的辅助信息构造一个合适的估计量并计算出使得现期均方误差最小的样本轮换率 μ。由 Bahl, S. and Tuteja, R. K. (1991)的理论知当研究变量与辅助变量的相关性未知或较低时，利用总体单元的辅助信息对非拼配样本构造如下指数形式的比率估计量比其他形式的估计量更精确：

$$\hat{\bar{Y}}_{1u}=\bar{y}_{1u}\exp\left(\frac{\bar{Z}-\bar{z}_u}{\bar{Z}+\bar{z}_u}\right) \tag{1}$$

对于拼配样本，不仅可以利用辅助变量的信息还可以利用与研究变量 y_1 相关的前期调查信息，构造两阶段的回归估计量：

$$\hat{\bar{Y}}_{1m}=\bar{y}^*_{1m}+b^{(m)}_{y_0y_1}[\bar{y}^*_{0n}-\bar{y}^*_{0m}] \tag{2}$$

其中 $\bar{y}^*_{1m}=\bar{y}_{1m}\exp\left(\frac{\bar{Z}-\bar{z}_m}{\bar{Z}+\bar{z}_m}\right)$，$\bar{y}^*_{0n}=\bar{y}_{0n}\exp\left(\frac{\bar{Z}-\bar{z}_n}{\bar{Z}+\bar{z}_n}\right)$，$\bar{y}^*_{0m}=\bar{y}_{0m}\exp\left(\frac{\bar{Z}-\bar{z}_m}{\bar{Z}+\bar{z}_m}\right)$，$b^{(m)}_{y_0y_1}$ 是基于研究变量 y_0 和 y_1 的样本回归系数。

下面将(1)式和(2)式合并起来可构造样本轮换后现期总体均值 $\bar{Y}_1$ 的回归组合估计量，即非拼配样本的比率估计量和拼配样本的组合估计量的加权平均：

$$\hat{\bar{Y}}_1=\varphi\hat{\bar{Y}}_{1u}+(1-\varphi)\hat{\bar{Y}}_{1m} \tag{3}$$

其中 φ 为待定的权重系数。

1.2 回归组合估计量的均方误差

我们构造的比率估计量 $\hat{\bar{Y}}_{1u}$ 和回归估计量和 $\hat{\bar{Y}}_{1m}$ 往往是有偏估计，因此求得的估计量 $\hat{\bar{Y}}_1$ 也是有偏的。由于估计量的均方误差是方差与其偏倚的平方的和，即 $MSE(\hat{\bar{Y}}_1)=V(\hat{\bar{Y}}_1)+B^2(\hat{\bar{Y}}_1)$，故这里用均方误差的大小去描述估计精度更加准确。为了求得 $\hat{\bar{Y}}_1$ 的均方误差 $M(\hat{\bar{Y}}_1)$，在大样本条件下我们可以做如下处理：

$$\begin{aligned}&\bar{y}_{1u}=\bar{Y}_1(1+e_0)\ ;\ \bar{y}_{1m}=\bar{Y}_1(1+e_1)\ ;\ \bar{y}_{0n}=\bar{Y}_0(1+e_2)\ ;\\&\bar{y}_{0m}=\bar{Y}_0(1+e_3)\ ;\ \bar{z}_u=\bar{Z}(1+e_4)\ ;\ \bar{z}_m=\bar{Z}(1+e_5)\ ;\\&\bar{z}_n=\bar{Z}(1+e_6)\ ;\ \mathrm{Cov}\ (y_{0m},y_{1m})=\mathrm{Cov}(y_0,y_1)(1+e_7)\ ;\\&s^2_{y_{0m}}=s^2_{y_0}(1+e_8)\ ;\end{aligned} \tag{4}$$

其中 $|e_i|<1$，$i=1,2,...8$，e_i 是均值为0的随机误差项。

把(4)式代入(1)式和(2)式可得到：

$$\hat{\bar{Y}}_{1u}=\bar{Y}_1(1+e_0)\exp\left[-\frac{e_4}{2}(1+\frac{e_4}{2})^{-1}\right] \tag{5}$$

$$\begin{aligned}\hat{\bar{Y}}_{1m}=&\bar{Y}_1(1+e_1)\exp\left[-\frac{e_5}{2}(1+\frac{e_5}{2})^{-1}\right]+\bar{Y}_0\beta_{y_0y_1}(1+e_7)(1+e_8)^{-1}\\&\left[(1+e_2)\exp\left\{-\frac{e_6}{2}(1+\frac{e_6}{2})^{-1}\right\}-(1+e_3)\exp\left\{-\frac{e_5}{2}(1+\frac{e_5}{2})^{-1}\right\}\right]\end{aligned} \tag{6}$$

由(3)式可求得估计量 $\hat{\bar{Y}}_1$ 的均方误差：

$$M(\hat{\bar{Y}}_1)=\varphi^2M(\hat{\bar{Y}}_{1u})+(1-\varphi)^2M(\hat{\bar{Y}}_{1m})+2\varphi(1-\varphi)\mathrm{Cov}(\hat{\bar{Y}}_{1u},\hat{\bar{Y}}_{1m}) \tag{7}$$

证明如下：

$$\begin{aligned}M(\hat{\bar{Y}}_1)&=E(\hat{\bar{Y}}_1-\bar{Y})^2=E(\varphi(\hat{\bar{Y}}_{1u}-\bar{Y}_1)+(1-\varphi)(\hat{\bar{Y}}_{1m}-\bar{Y}_1)\\&=\varphi^2M(\hat{\bar{Y}}_{1u})+(1-\varphi)^2M(\hat{\bar{Y}}_{1m})+2\varphi(1-\varphi)\mathrm{Cov}(\hat{\bar{Y}}_{1u},\hat{\bar{Y}}_{1m})\end{aligned}$$

其中 $\mathrm{Cov}(\hat{\bar{Y}}_{1u},\hat{\bar{Y}}_{1m})=E[(\hat{\bar{Y}}_{1u}-\bar{Y}_1)(\hat{\bar{Y}}_{1m}-\bar{Y}_1)]$

由上述各公式可推算出均方误差的具体值为：

$$M(\hat{\bar{Y}}_{1u})=(\frac{1}{u}-\frac{1}{N})(\frac{5}{4}-\rho_{y_1z})s^2_{y_1}$$

$$\begin{aligned}M(\hat{\bar{Y}}_{1m})=&\left[\frac{1}{m}(-\rho^2_{y_0y_1}(\frac{3}{4}+\rho_{y_0z})+\rho_{y_0y_1}(\rho_{y_1z}+\rho_{y_0z}-\frac{1}{2})-\rho_{y_1z}+\right.\\&\left.\frac{1}{n}(\rho_{y_0y_1}(-\frac{3}{4}\rho_{y_0y_1}+\rho_{y_0z}\rho_{y_0y_1}-\rho_{y_0z}+\frac{1}{2}))-\frac{1}{N}(\frac{5}{4}-\rho_{y_1z})\right]s^2_{y_1}\end{aligned}$$

$$\mathrm{Cov}(\hat{\bar{Y}}_{1u},\hat{\bar{Y}}_{1m})=-\frac{1}{N}(\frac{5}{4}-\rho_{y_1z})s^2_{y_1}$$

根据(7)式可求得使得 $M(\hat{\bar{Y}}_1)$ 最小的权重系数

$$\begin{aligned}\varphi_{opt}&=\frac{M(\hat{\bar{Y}}_{1m})-\mathrm{Cov}(\hat{\bar{Y}}_{1u},\hat{\bar{Y}}_{1m})}{M(\hat{\bar{Y}}_{1u})+M(\hat{\bar{Y}}_{1m})-2\,\mathrm{Cov}(\hat{\bar{Y}}_{1u},\hat{\bar{Y}}_{1m})}\\&=\frac{\mu_2[(A_1+A_2)-\mu_2A_2]}{[A_3+\mu_2A_7-\mu_2{}^2A_2]}\end{aligned} \tag{8}$$

上式代入(7)式可得到最小均方误差：

$$M(\hat{\bar{Y}}_1)_{opt}=\left[\frac{A_4+\mu A_5+\mu^2A_6}{A_3+\mu A_7-\mu^2A_2}\right]\frac{s^2_{y_1}}{n} \tag{9}$$

其中 $A_1=-\rho^2_{y_0y_1}\left(\frac{3}{4}+\rho_{y_0z}\right)+\rho_{y_0y_1}\left(\rho_{y_1z}+\rho_{y_0z}-\frac{1}{2}\right)-\rho_{y_1z}$；

$A_2=\rho_{y_0y_1}\left(-\frac{3}{4}\rho_{y_0y_1}+\rho_{y_0z}\rho_{y_0y_1}-\rho_{y_0z}+\frac{1}{2}\right)$；$A_3=\frac{5}{4}-\rho_{y_1z}$；$A_4=A_1A_3+A_2A_3-A_3^2f$；

$A_5=-A_1A_3f-A_2A_3-A_2A_3f+A_3^2f$；$A_6=A_2A_3f$；$A_7=A_1+A_2-A_3$；$f=\frac{n}{N}$。

为了求得最优样本轮换率，记作 μ_0，可通过化简(9)式构造使得均方误差最小的方程式：

$$T_1\mu^2+2T_2\mu+T_3=0 \tag{10}$$

其中 $T_1=A_3A_5-A_4A_7$；$T_2=A_3A_6+A_2A_4$；$T_3=A_6A_7+A_2A_5$。

求解得 $\mu=\frac{-T_2\pm\sqrt{T_2^2-T_1T_3}}{T_1}$，只要 $T_2^2-T_1T_3\geq0$，这里求得的就是最优样本轮换率和最小均方误差：

$$\mu_0=\frac{-T_2\pm\sqrt{T_2^2-T_1T_3}}{T_1}$$

$$M(\hat{\bar{Y}}_1)_{opt}=\left[\frac{A_4+\mu_0A_5+\mu_0^2A_6}{A_3+\mu_0A_7-\mu_0^2A_2}\right]\frac{s^2_{y_1}}{n} \tag{11}$$

为验证 $\hat{\bar{Y}}_1$ 估计量的精确性，这里用方差和均方误差的比值(相对有效性)来衡量。我们通常构造的样本均值 $\bar{y}_n$ 和总体均值 $\hat{\bar{Y}}=\varphi\bar{y}_u+(1-\varphi)\bar{y}^*_m$，其中 $\bar{y}^*_m=\bar{y}_m+\beta_{xy}(\bar{x}_n-\bar{x}_m)$ 都是 $\bar{Y}$ 的无偏估计，由 Sukhatme (1984)的结论知：

$$V(\hat{\bar{y}}_n)=(1-f)\frac{s_y^2}{n} \tag{12}$$

$$V(\hat{\bar{Y}})=\left[\frac{1}{2}\left(1+\sqrt{1-\rho_{yx}^2}\right)-f\right]\frac{s_y^2}{n} \tag{13}$$

由于 $MSE(\hat{\bar{Y}}_1)=V(\hat{\bar{Y}}_1)+B^2(\hat{\bar{Y}}_1)$，可构造相对有效性如下：

$$E^{(1)}=\frac{V(\hat{\bar{y}}_n)}{M(\hat{\bar{Y}})_{opt}}\times 100 \tag{14}$$

$$E^{(2)}=\frac{V(\hat{\bar{Y}})}{M(\hat{\bar{Y}})_{opt}}\times 100 \tag{15}$$

观察上式可知当 $E^{(1)}$，$E^{(2)}$ 的数值越大，说明 $\hat{\bar{Y}}_1$ 估计量的均方误差相对样本均值和总体均值估计量的方差越小，则说明 $\hat{\bar{Y}}_1$ 估计的效果越好。其次 $E^{(1)}$，$E^{(2)}$ 式子中只剩下各变量间的相关系数和样本与总体的比值，便于计算，下面通过模拟数据进行验算。

2 数值分析

为了对比该抽样估计方法的有效性，现通过设定模拟参数得出一系列模拟数据进行理论验证。假设总体单位 N =5000，样本容量 n =1000，$\rho_{y_0y_1}$，ρ_{y_0z}，ρ_{y_1z} 可分别取0.5，0.7，0.9。由R语言编程可得表1：

表1

ρ_{y_0z}		0.5			0.7			0.9		
$\rho_{y_0y_1}$	ρ_{y_1z}	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.5	0.5	0.3456	170.91	167.69	0.445	178.57	175.20	0.5096	182.13	178.69
	0.7	0.2647	175.68	172.37	0.3352	188.00	184.45	0.3839	194.70	191.03
	0.9	0.2241	212.72	208.70	0.2831	237.65	233.16	0.3240	253.01	248.31
0.7	0.5	0.3435	124.33	111.49	0.3801	125.22	112.29	0.4078	125.03	112.12
	0.7	0.2382	140.22	12575.	0.2582	143.19	128.41	0.2710	144.60	129.67
	0.9	0.1801	188.84	169.34	0.1906	197.65	177.24	0.1946	203.68	182.65
0.9	0.5	0.2652	82.71	61.38	0.2695	82.06	60.90	0.2730	81.31	60.34
	0.7	0.1589	98.84	73.35	0.1583	98.58	73.16	0.1570	98.17	72.86
	0.9	0.0989	143.51	106.50	0.0952	144.45	107.20	0.0910	145.15	107.72

为了比较样本量的多少对估计效果的影响，另外取 N =5000，n =500，$\rho_{y_0y_1}$，ρ_{y_0z}，ρ_{y_1z} 分别取0.5，0.7，0.9。得到表2：

表2

ρ_{y_0z}		0.5			0.7			0.9		
$\rho_{y_0y_1}$	ρ_{y_1z}	μ_0	E_1	E_2	μ_0	E_1	E_2	μ_0	E_1	E_2
0.5	0.5	0.3456	132.98	138.47	0.4415	138.17	143.87	0.5096	140.56	146.36
	0.7	0.2647	141.02	146.84	0.3352	149.89	156.07	0.3839	154.67	161.05
	0.9	0.2241	174.63	181.84	0.2831	193.37	201.35	0.3240	204.80	213.26
0.7	0.5	0.3435	100.14	94.77	0.3801	100.79	95.38	0.4078	100.65	95.25
	0.7	0.2382	114.81	108.65	0.2582	117.04	110.76	0.2710	118.10	111.76
	0.9	0.1801	156.37	147.98	0.1906	163.15	154.39	0.1946	167.76	158.76
0.9	0.5	0.2652	68.78	53.12	0.2695	68.27	52.74	0.2730	67.69	52.28
	0.7	0.1589	82.85	64.00	0.1583	82.64	63.84	0.1570	82.33	63.59
	0.9	0.0989	120.82	93.32	0.0952	121.57	93.90	0.0910	122.13	94.34

由表1知前后期研究变量的相关性 $\rho_{y_0y_1}$ 越大，样本轮换率就越小，这与实际是相符的，即前后期相关性较大时可减少样本轮换的数量，充分利用辅助信息的相关性，减少调查对象；当现期研究变量与辅助变量的相关性 ρ_{y_1z} 越大时，样本轮换率越小，而相对有效性 E_1，E_2 的值却越大，这也是与实际情况相符合的，即辅助信息的应用可在保证估计精度的前提下减少样本轮换的数量达到节约成本的目的。

从两表中还可看出当相邻两期的研究指标相关性较高且辅助变量选择合适时，最大的样本轮换率为0.5096，意味着只要轮换约50%的样本就可以较精确地估计出总体的均值。所以不管是从成本节约还是从精度要求角度看，该方法可以大大减轻基层数据调查的负担，有着广泛的应用前景。

从表1和表2知最优样本轮换率与样本数量的多少没有绝对的关系，即两表中的样本轮换率保持不变，只是估计的效果与样本和总体的比值 $f=\frac{n}{N}$ 有关，且样本量越大估计效果越好。

本文假定各研究变量之间的相关性均大于0.5，此时有比较好的估计效果，当相关性较小时，此时样本轮换率会很大。

3 结论

根据本文的理论和模拟数据分析，对连续性抽样调查进行部分样本轮换，由于样本存在老化现象，新样本的加入可消除此类问题。同时文章根据相邻两期的相关性，利用前期保留下来的样本信息和与前后期总体相关的辅助信息，构造指数形式的回归组合估计量，以最优化方法求得最优样本轮换率和最优权重系数，使得估计量的均方误差最小，无论对于保证估计量的精度还是对于节约调查成本，都是一个不错的选择。

最后应该指出的是本文还可以在此基础上进行更加深入的研究。可以研究连续两个以上不同时间的抽样估计问题，充分利用前期的样本信息，还可考虑添加多个辅助变量进一步提高抽样估计的精度。本文仅采用模拟数据进行分析，可进一步采用实际生活中抽样数据进行理论验证和应用。

参考文献：

[1]Jessen R J.Statistical Investigation of a Farm Survey for Obtaining Farm Facts[J].Iowa Agricultural Station Research Bulletin,1942,(3).

[2]Patterson H D.Sampling on Successive Occasions with Partial Replacement of Units[J].Journal of the Royal Statistical Society,1950,(2).

[3]科克伦.抽样技术[M].北京:中国统计出版社,1985.

[4]Wolter K M.Composite Estimation in Finite Populations[J].Journal of the American Statistical Association,1979,(4).

[5]Sen A R.Successive Sampling with two Auxiliary Variables[J].Sankhya Ser.1971 ,(33).

[6]SinghV K , Singh. G N Chain-type Regression Estimators with two Auxiliary Variables Under Double Sampling Scheme[J].Metron,1991,

(49).

[7]Singh G N , Singh V K On the Use of Auxiliary Information in Successive Sampling [J].Indian Soc.Agric.Stat,2001,54(1).

[8]Biradar R S ,Singh. H P Successive Sampling Using Auxiliary Information on both Occasions[J].Calcutta Stat.Assoc.Bull,2001,51(23).

[9]Singh G N , Priyanka K On the use of Auxiliary Information in Search of Good Rotation Patterns on Successive Occasions[J].Bull.Stat.Econ, 2007,1(7).

[10]Singh G N ,Karna J P .Search of Efficient Rotation Patterns in Presence of Auxiliary Information in Successive Sampling Over two Occasions[J].StatTransition New Ser,2009,10(1).

[11]Singh G N , Prasad S .Some Estimators of Population Mean in two-occasion Rotation Patterns[J].Modeling Simulation Techniques Enterprises,2010,12(1).

[12]冯士雍,邹国华.有辅助信息可利用时的样本轮换方法[J].统计研究,1996,(3).

[13]马树才,杨旭东.分层部分样本轮换抽样下的混合估计[J].辽宁大学学报(自然科学版),1997,24(3).

[14]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(5).

[15]Ball S , Tuteja R K. Ratio and Product Type Exponential Estimator [J]. Information and Optimization Science,1991,12.

[16]Singh G N ,Homa F .Effective Rotation Patterns in Successive Sampling over two Occasions[J].Journal of statistical theory and practice, 2012,(7).

[17]Sukhatme P V , Sukhatme B V , Ashok C .Sampling theory of Surveys with Applications[M].3rd ed.Ames,IA,Iowa State University Press,1984.

（责任编辑/亦　民）