

附录 1

基本符号与相关原理概述

本节首先对不放回样本追加法中所使用的基本符号进行说明，然后对永久随机数抽样技术的基本方法—Poisson 抽样以及条件 Poisson 抽样的相关原理进行概述。

（一）基本符号

一个完整的不放回样本追加法的抽样方案包括对总体的基本抽样和对子总体的追加抽样两部分，以下分别使用 primary 和 complementary 的首字母 p 和 c 作为下标表示基本抽样和追加抽样有关的量。

1. 总体的基本符号。

令有限总体 $U = (1, \dots, N)$ ，目标变量 $Y = (Y_1, Y_2, \dots, Y_N)'$ ，目标变量总体总值为 $Y^* = \sum_{i=1}^N Y_i$ ，辅助变量 $X = (X_1, X_2, \dots, X_N)'$ ，辅助变量总体总值为 $X^* = \sum_{i=1}^N X_i$ ，对应样本 s_p 中的目标变量为 $y = (y_1, y_2, \dots, y_n)'$ ，样本量为 n ，总体中第 i 个单元被抽中的概率为 $p_i = X_i / \left(\sum_{i=1}^N X_i \right)$ ，其中 $p_i > 0$ 且 $\sum_{i=1}^N p_i = 1$ ，第 i 个单元的一阶包含概率为 $\pi_i = np_i$ ，即 π_i 与 X_i 成正比。

2. 子总体的基本符号。

总体基本符号的下标添上 h 即为有限子总体 h 相应的基本符号，如子总体 h 目标变量为 $Y_h = (Y_{h1}, Y_{h2}, \dots, Y_{hN_h})'$ ，故这部分内容不再赘述。将属于子总体 h 且用于对总体推断的基本抽样单元的数量记为 $n_{hp} = \# \{s_p \cap U_h\}$ ，其中 $U_h \subset U$ ，且 $U_h \cap U_{h'} = \emptyset$ ， $\bigcup_{h=1}^H U_h = U$ ，即 $Y^* = \sum_{h=1}^H Y_h^*$ ， $X^* = \sum_{h=1}^H X_h^*$ ， $h, h' = 1, \dots, H$ ， $h \neq h'$ 。

3. 追加样本所对应的子总体部分的基本符号。

令属于子总体 h 且不属于总体基本抽样样本中的单元集合为 $U_{h-s_p} = U_h \setminus s_p$ ，即 U_{h-s_p} 为追加样本所对应的子总体部分， $N_{h-s_p} = \# \{U_h \setminus s_p\}$ ， n_{hc} 为子总体追加抽样的单元数量。总体基本符号的下标添上 $h-s_p$ 即为追加样本所对应的子总体部分相应的基本符号，故这部分内容同样不再赘述。

（二）相关原理概述

1. Poisson 抽样。

Poisson 抽样 (Hájek, 1964) 作为永久随机数抽样技术的基本抽样方法, 是严格的 πps 抽样, 该方法给每个调查单元等概率地分配一个确定的随机数, 使样本单元在各级调查之间相互兼容, 从而提高样本利用率, 具体实施步骤为: 首先, 对总体中的每个单元赋予永久随机数 r_i , 其中 $i=1, \dots, N$, $r_i \sim U[0,1]$ 且相互独立, 给出总体中每个单元的入样概率 π_i , 其中 $\pi_i = np_i$; 其次, 将第 i 个单元的 r_i 与 π_i 进行比较, 如果 $r_i < \pi_i$, 则抽中第 i 个单元; 最后, 按上述步骤判断总体中的每个单元是否入样, 即可得到一个 Poisson 抽样样本, 其实际样本量 m 是一个随机变量, 期望样本量为 n 。

2. 条件 Poisson 抽样。

针对 Poisson 抽样样本量随机所导致的估计精度低、估计结果不稳定等问题, 本文考虑使用条件 Poisson 抽样 (Grafström 和 Matei, 2015) 来得到固定的样本量。条件 Poisson 抽样方法是对 Poisson 抽样的一种修正, 即在 Poisson 抽样基础上设置固定样本量为 n 的条件, 其基本思路是通过不断产生 Poisson 样本, 直到现实样本量 m 与期望样本量 n 相同时样本才被接受。条件 Poisson 抽样过程中的主要问题是其包含概率随样本量 n 的变化而变化, 即当使用包含概率为 $\pi_i = np_i$ 的条件 Poisson 抽样时, 真正的包含概率仅近似等于 π_i 。为解决该问题, Chen 等 (1994) 提出一个递归公式可以快速得到包含概率:

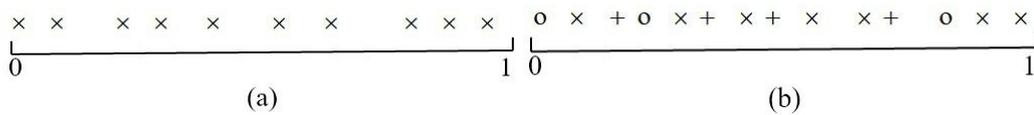
$$\pi_i^{CPS(n)} = n \frac{\frac{\pi_i}{1-\pi_i}(1-\pi_i^{CPS(n-1)})}{\sum_{k=1}^N \frac{\pi_k}{1-\pi_k}(1-\pi_k^{CPS(n-1)})} \quad (A.1)$$

其中, $\pi_i^{CPS(n)}$ 表示样本量为 n 的条件 Poisson 抽样实际所获得的包含概率, 起始值为 $\pi_i^{CPS(0)} = 0$, $i=1, \dots, N$, 其二阶包含概率可用 Deville (2000) 提出的递归公式快速得到:

$$\pi_{ij}^{CPS(n)} = n(n-1) \frac{\frac{\pi_i}{1-\pi_i} \frac{\pi_j}{1-\pi_j} (1-\pi_i^{CPS(n-2)} - \pi_j^{CPS(n-2)} + \pi_{ij}^{CPS(n-2)})}{\sum_{k=1}^N \sum_{l=1, l \neq k}^N \frac{\pi_k}{1-\pi_k} \frac{\pi_l}{1-\pi_l} (1-\pi_k^{CPS(n-2)} - \pi_l^{CPS(n-2)} + \pi_{kl}^{CPS(n-2)})} \quad (A.2)$$

其中, $\pi_{ij}^{CPS(0)} = \pi_{ij}^{CPS(1)} = 0$, $\pi_{ij}^{CPS(2)} = 2 \frac{\frac{\pi_i}{1-\pi_i} \frac{\pi_j}{1-\pi_j}}{\sum_{k=1}^N \sum_{l=1, l \neq k}^N \frac{\pi_k}{1-\pi_k} \frac{\pi_l}{1-\pi_l}}$, $i, j=1, \dots, N$, 且 $i \neq j$ 。

附录 2



附图 1 现有抽样框中单元对应随机数在 $[0,1]$ 区间内分布情况 (a) 以及更新后的分布情况 (b)

其中， \times 表示保留在总体中的单元， o 表示消亡单元， $+$ 表示出生单元。此时，图 (b) 中的 \times 和 $+$ 就是抽样框中实际包含的单元，以使抽样总体和目标总体保持一致，这样就避免了现有方法因使用过时的抽样框而导致样本代表性差、估计结果不稳定等问题。

对子总体的估计方法。

由式 (A.1) 得到子总体 h 的一阶包含概率的递归公式：

$$\pi_{hi}^{LSA-CPS(n_h)} = n_h \frac{\frac{\pi_{hi}}{1-\pi_{hi}}(1-\pi_{hi}^{LSA-CPS(n_h-1)})}{\sum_{k=1}^{N_h} \frac{\pi_{hk}}{1-\pi_{hk}}(1-\pi_{hk}^{LSA-CPS(n_h-1)})} \quad (A.3)$$

继而得到子总体 h 的一阶包含概率，记为 $\pi_{hi}^{LSA-CPS}$ ，根据子总体追加抽样设计方法得到的样本，给出相应子总体总值 Y_h^* 的无偏 HT 估计量为：

$$\hat{Y}_{LSA-CPS h} = \sum_{i=1}^{n_h} \frac{y_{hi}}{\pi_{hi}^{LSA-CPS}}, \quad (A.4)$$

同样使用两种方法给出子总体 h 总值估计量的方差估计量。

第一种方法是使用 Sen-Yates-Grundy 方法得到方差估计量。由式 (A.2) 可以得到子总体的二阶包含概率的递归公式：

$$\pi_{hij}^{LSA-CPS(n_h)} = n_h(n_h-1) \frac{\frac{\pi_{hi}}{1-\pi_{hi}} \frac{\pi_{hj}}{1-\pi_{hj}} (1-\pi_{hi}^{LSA-CPS(n_h-2)} - \pi_{hj}^{LSA-CPS(n_h-2)} + \pi_{hij}^{LSA-CPS(n_h-2)})}{\sum_{k=1}^{N_h} \sum_{l=1, l \neq k}^{N_h} \frac{\pi_{hk}}{1-\pi_{hk}} \frac{\pi_{hl}}{1-\pi_{hl}} (1-\pi_{hk}^{LSA-CPS(n_h-2)} - \pi_{hl}^{LSA-CPS(n_h-2)} + \pi_{hkl}^{LSA-CPS(n_h-2)})} \quad (A.5)$$

其中， $\pi_{hij}^{LSA-CPS(0)} = \pi_{hij}^{LSA-CPS(1)} = 0$ ， $\pi_{hij}^{LSA-CPS(2)} = 2 \frac{\frac{\pi_{hi}}{1-\pi_{hi}} \frac{\pi_{hj}}{1-\pi_{hj}}}{\sum_{k=1}^{N_h} \sum_{l=1, l \neq k}^{N_h} \frac{\pi_{hk}}{1-\pi_{hk}} \frac{\pi_{hl}}{1-\pi_{hl}}}$ ， $i, j = 1, \dots, N_h$ ，且 $i \neq j$ ，

继而得到子总体 h 的二阶包含概率，记为 $\pi_{hij}^{LSA-CPS}$ ，其子总体无偏方差估计量为：

$$\hat{V}_{SYG}(\hat{Y}_{LSA-CPS h}) = \frac{1}{2} \sum_{i=1}^{n_h} \sum_{j=1, j \neq i}^{n_h} \left(\frac{y_{hi}}{\pi_{hi}^{LSA-CPS}} - \frac{y_{hj}}{\pi_{hj}^{LSA-CPS}} \right)^2 \frac{\pi_{hi}^{LSA-CPS} \pi_{hj}^{LSA-CPS} - \pi_{hij}^{LSA-CPS}}{\pi_{hij}^{LSA-CPS}} \quad (A.6)$$

第二种方法是使用 Deville 方法得到子总体 h 的渐近无偏方差估计量：

$$\hat{V}_{Dev}(\hat{Y}_{LSA-CPS h}) = \frac{1}{1 - \sum_{i=1}^{n_h} \left((1-\pi_{hi}^{LSA-CPS}) / \sum_{k=1}^{n_h} (1-\pi_{hk}^{LSA-CPS}) \right)^2} \times \sum_{i=1}^{n_h} (1-\pi_{hi}^{LSA-CPS}) \left(\frac{\frac{y_{hi}}{\pi_{hi}^{LSA-CPS}} - \frac{\sum_{k=1}^{n_h} (1-\pi_{hk}^{LSA-CPS}) y_{hk} / \pi_{hk}^{LSA-CPS}}{\sum_{k=1}^{n_h} (1-\pi_{hk}^{LSA-CPS})}}{\sum_{k=1}^{n_h} (1-\pi_{hk}^{LSA-CPS})} \right)^2 \quad (A.7)$$

其中， $h = 1, \dots, H$ 。

附录 4

附表 1 二次模型下子总体的各类估计量 MSE、MAPE、MVE 和 VV 的比较

	n	100			400		
	n_1	20	40	60	20	40	60
MSE	LSA-CPS	4.60×10^3	2.10×10^3	1.28×10^3	4.57×10^3	2.04×10^3	1.33×10^3
	PS	4.01×10^4	9.47×10^4	6.74×10^4	4.95×10^3	2.30×10^3	1.32×10^3
	SRS	3.50×10^4	1.43×10^4	7.99×10^3	3.41×10^4	1.31×10^4	7.72×10^3
MAPE	LSA-CPS	0.0185	0.0125	0.0098	0.0184	0.0125	0.0098
	PS	0.0433	0.0853	0.0702	0.0193	0.0129	0.0100
	SRS	0.0514	0.0330	0.0246	0.0508	0.0314	0.0242
MVE	LSA-CPS _{SYG}	4.84×10^3	2.14×10^3	1.25×10^3	4.79×10^3	2.14×10^3	1.24×10^3
	LSA-CPS _{DEV}	4.83×10^3	2.14×10^3	1.25×10^3	4.79×10^3	2.14×10^3	1.24×10^3
	PS	3.80×10^5	1.67×10^5	9.62×10^4	3.81×10^5	1.68×10^5	9.64×10^4
	SRS	3.16×10^4	1.39×10^4	8.08×10^3	3.13×10^4	1.40×10^4	8.12×10^3
VV	LSA-CPS _{SYG}	2.25×10^6	2.44×10^5	5.70×10^4	2.26×10^6	2.25×10^5	5.86×10^4
	LSA-CPS _{DEV}	2.24×10^6	2.40×10^5	5.55×10^4	2.25×10^6	2.21×10^5	5.71×10^4
	PS	1.10×10^9	3.80×10^8	8.94×10^7	5.10×10^8	7.16×10^7	1.80×10^7
	SRS	5.41×10^7	5.09×10^6	1.05×10^6	5.16×10^7	5.23×10^6	1.05×10^6

附表 2 组合模型下子总体的各类估计量 MSE、MAPE、MVE 和 VV 的比较

	n	100			400		
	n_1	20	40	60	20	40	60
MSE	LSA-CPS	1.14×10^5	4.78×10^4	3.09×10^4	1.15×10^5	5.01×10^4	2.78×10^4
	PS	2.74×10^5	5.58×10^5	3.89×10^5	9.91×10^4	5.10×10^4	2.89×10^4
	SRS	3.61×10^5	1.57×10^5	8.88×10^4	3.80×10^5	1.66×10^5	1.05×10^5
MAPE	LSA-CPS	0.0416	0.0270	0.0215	0.0416	0.0276	0.0203
	PS	0.0600	0.0907	0.0757	0.0388	0.0274	0.0205
	SRS	0.0733	0.0484	0.0369	0.0766	0.0500	0.0394
MVE	LSA-CPS _{SYG}	1.15×10^5	5.05×10^4	2.89×10^4	1.15×10^5	5.07×10^4	2.90×10^4
	LSA-CPS _{DEV}	1.15×10^5	5.05×10^4	2.88×10^4	1.15×10^5	5.07×10^4	2.90×10^4
	PS	2.00×10^6	8.73×10^5	4.95×10^5	1.99×10^6	8.75×10^5	4.99×10^5
	SRS	3.62×10^5	1.62×10^5	9.46×10^4	3.66×10^5	1.62×10^5	9.52×10^4
VV	LSA-CPS _{SYG}	8.62×10^8	7.91×10^7	1.75×10^7	8.36×10^8	8.14×10^7	1.85×10^7
	LSA-CPS _{DEV}	8.63×10^8	7.94×10^7	1.74×10^7	8.38×10^8	8.19×10^7	1.86×10^7
	PS	6.59×10^{10}	1.51×10^{10}	3.34×10^9	4.72×10^{10}	5.64×10^9	1.44×10^9
	SRS	9.19×10^9	9.15×10^8	1.76×10^8	9.35×10^9	8.99×10^8	2.04×10^8

参考文献

[1] Chen X H, Dempster A P, Liu J S. Weighted Finite Population Sampling to Maximize Entropy[J]. Biometrika, 1994, 81(3): 457-469.

- [2] Deville J C. Note on the Algorithm of Chen, Dempster and Liu[R]. Unpublished working paper, 2000.
- [3] Grafström A, Matei A. Coordination of Conditional Poisson Samples[J]. Journal of Official Statistics, 2015, 31(4): 649-672.
- [4] Hájek J. Asymptotic Theory of Rejective Sampling with Varying Probabilities from a Finite Population[J]. The Annals of Mathematical Statistics, 1964,35(4): 1491-1523.