

放回的 PPS 抽样方案在放回抽样方案类中的可容许性*

邹国华 冯士雍

(中国科学院系统科学研究所, 北京 100080)

关键词 抽样调查 PPS 抽样方案 可容许性

我们知道, 一般而言, 不放回抽样的效率要比放回抽样高些. 但由于放回抽样实施起来简便易行, 特别是其方差的计算十分简单, 所以它往往更受抽样调查工作者的欢迎. 在放回抽样中, 最常用的是采用与单元大小成比例的概率来抽选各个单元(即 PPS 抽样)^[1,2]. 这产生了一个理论问题, 即 PPS 抽样方案是否就是最好的放回抽样方案, 还有没有比 PPS 抽样方案更好的方案呢? 本文的目的就是解决这个问题. 我们获得了充要条件. 所得结果表明: 通常使用的 PPS 抽样方案不但不是最优的(除一个不太可能出现的特例外), 而且有时能被一致地改进. 尽管 Scott 和 Smith^[3] 提供了支持 PPS 抽样方案的证据, 即在一定的条件下, PPS 抽样方案在放回的抽样方案类中具有 minimax 或近似 minimax 性. 出乎意料的是, 我们的结果指出, 当大小为 $N > 2$ 的总体中有一个单元的大小恰好等于其余各个单元的大小之和时, 若只抽样一次, 则 PPS 抽样方案能被改进; 但当抽样次数多于 1 时, PPS 抽样方案却不能被改进.

以下假定总体的大小为 N , 各个单元的指标值为 $Y_1, \dots, Y_N$, 第 i 个单元的大小为 $M_i, i=1, \dots, N$, 记 $M_0 = \sum_{i=1}^N M_i$. 以 $p^* = \{M_i / M_0; i=1, \dots, N\}$ 表示放回的 PPS 抽样方案, 以 $y_1, \dots, y_n$ 表示被抽中的 n 个单元的指标值. 欲估计总体的总和 $Y = \sum_{i=1}^N Y_i$. 采用的估计量是 Hansen-Hurwitz 估计量

$$\hat{Y}_{HH} = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{p_i^*}.$$

我们以估计量的均方误差作为在放回的抽样方案类中比较抽样方案好坏的标准.

本文的主要结果如下:

定理 1 (i) 如果 $\forall i=1, \dots, N$, 均有 $M_i < \frac{M_0}{2}$, 则 PPS 抽样方案 p^* 是可容许的. (ii) 如果存在 i_0 , 使得 $M_{i_0} > \frac{M_0}{2}$, 则 PPS 抽样方案 p^* 不是可容许的. (iii) 如果存在 i_0 , 使得 $M_{i_0} = \frac{M_0}{2}$, 那么当 $N=2$ 时, p^* 可容许; 当 $N > 2$ 时, 若 $n=1$, 则 p^* 不可容许, 若 $n > 1$, 则 p^* 可容许.

证 先证容许性部分. 我们只证(iii)中的断言, (i)类似可得. 首先, 对任何放回的抽样

1994-09-02 收稿

*国家自然科学基金重点资助项目

方案 $p = \{p_i; i=1, \dots, N\}$, $\hat{Y}_{HH}$ 的均方误差是

$$R_p(\hat{Y}_{HH}) = \frac{1}{n} \sum_{i=1}^N p_i \left(\frac{Y_i}{p_i^*} - Y \right)^2 + \frac{n-1}{n} \left(\sum_{i=1}^N \frac{p_i Y_i}{p_i^*} - Y \right)^2, \quad (1)$$

故 $R_p(\hat{Y}_{HH}) \leq R_{p^*}(\hat{Y}_{HH})$ (“ $\leq$ ”改成“ $<$ ”也对,下同)的充要条件是

$$\sum_{i=1}^N p_i \left(\frac{Y_i}{p_i^*} - Y \right)^2 + (n-1) \left(\sum_{i=1}^N \frac{p_i Y_i}{p_i^*} - Y \right)^2 \leq \sum_{i=1}^N p_i^* \left(\frac{Y_i}{p_i^*} - Y \right)^2. \quad (2)$$

由所设知,存在 i_0 , 不防设 $i_0=1$, 使得 $p_1^* = \frac{1}{2}$. 若 $n=1$ 且 $N=2$, 则 $p_2^* = \frac{1}{2}$. 从而由

(1)式知对任何放回的抽样方案 p , 均有 $R_p(\hat{Y}_{HH}) = (Y_1 - Y_2)^2$. 故 p^* 可容许. 以下设 $n > 1$. 假如 $R_p(\hat{Y}_{HH}) \leq R_{p^*}(\hat{Y}_{HH})$, 则在(2)式中取 $Y_R = 1, Y_i = 0 (i \neq k)$ 得

$$(p_k - p_k^*) \cdot (1 - 2p_k^*) + (n-1) \cdot (p_k - p_k^*)^2 \leq 0. \quad (3)$$

在(3)式中让 $k=1$ 可得 $p_1 = p_1^* = \frac{1}{2}$. 于是当 $N=2$ 时, 有 $p_2 = \frac{1}{2} = p_2^*$. 从而此时 p^* 是可容许的; 当 $N > 2$ 时, 有 $p_k^* < \frac{1}{2} (k=2, \dots, N)$. 又据(3)式知, 对于 $k=2, \dots, N$, 有

$$(p_k - p_k^*) \cdot (1 - 2p_k^*) \leq 0, \quad (4)$$

故 $p_k \leq p_k^* (k=2, \dots, N)$, 因而必有 $p_k = p_k^* (k=2, \dots, N)$, 即有 $p = p^*$. 这说明当 $N > 2$ 时 p^* 也是可容许的.

现证非容许性部分. 由所设条件知, 存在 i_0 , 不防设 $i_0=1$, 使得 $p_1^* \geq \frac{1}{2}$ (对(ii)成立严格不等式, 对(iii)成立等式). 对(2)式作变换 $Z_i = Y_i/p_i^* - Y$ 知 $R_p(\hat{Y}_{HH}) \leq R_{p^*}(\hat{Y}_{HH})$ 的充要条件是

$$\sum_{i=1}^N p_i Z_i^2 + (n-1) \left(\sum_{i=1}^N p_i Z_i \right)^2 \leq \sum_{i=1}^N p_i^* Z_i^2. \quad (5)$$

注意到 $\sum_{i=1}^N p_i^* Z_i = 0$, 故 $Z_1 = -\sum_{i=2}^N p_i^* Z_i / p_1^*$. 将其代入(5)式得到

$$\begin{aligned} & (n-1) \cdot \left[\sum_{i=2}^N (p_1^* p_i - p_i p_1^*) Z_i \right]^2 + (p_1 - p_1^*) \left(\sum_{i=2}^N p_i^* Z_i \right)^2 \\ & \leq p_1^{*2} \cdot \sum_{i=2}^N (p_i^* - p_i) Z_i^2. \end{aligned} \quad (6)$$

若 $n=1$, 则取 $p_1=1, p_i=0 (i=2, \dots, N)$ 可知, 当 $p_1^* = \frac{1}{2}$ 且 $N > 2$ 或 $p_1^* > \frac{1}{2}$ 时, $p = \{p_i; i=1, \dots, N\}$ 优于 p^* . 故(iii)中的断言及(ii)中 $n=1$ 时的结论成立. 因此剩下只需考虑 $n > 1$ 的

情况. 为此, 先取 λ 充分小, 使得 $0 < \lambda < \frac{1}{2}$, 且 $p_1^* < (n-1+\lambda)/(n-1+2\lambda)$. 再取 $a_i (i=2, \dots, N)$, 使其满足

$$\frac{p_1^* + (1-\lambda)(2p_1^* - 1)}{p_1^* [p_1^* - \lambda(2p_1^* - 1)]} \cdot p_i^* \leq a_i \leq \frac{n-1+2\lambda}{\lambda} \cdot p_i^*, \quad (7)$$

并且 $\sum_{i=2}^N a_i = 1$. 这是可以办到的: 考虑函数 $f(x_2 \cdots x_N) = \sum_{i=2}^N x_i$, 其中

$$\frac{p_1^* + (1-\lambda)(2p_1^* - 1)}{p_1^* [p_1^* - \lambda(2p_1^* - 1)]} \cdot p_i^* \leq x_i \leq \frac{n-1+2\lambda}{\lambda} \cdot p_i^*, \quad i=2, \dots, N$$

则易证 $\sum_{i=2}^N a_i$ 的最小值 ≤ 1 , 最大值 > 1 . 故由连续函数的介值定理知, 存在 a_i 满足 (7) 式, 使得 $\sum_{i=2}^N a_i = 1$.

现定义抽样方案 $p = \{p_i; i=1, \dots, N\}$ 如下:

$$\begin{cases} p_1 = \frac{(n-1+2\lambda)p_1^* - \lambda}{n-1}, \\ p_i = \frac{(n-1+2\lambda)p_i^* - \lambda a_i}{n-1}, \quad (i \geq 2) \end{cases} \quad (8)$$

则易验证 p 确是一个抽样方案, 且 $p \neq p^*$. 将 (8) 式代入 (6) 式并加以整理可知 $R_p(\hat{Y}_{HH}) \leq R_{p^*}(\hat{Y}_{HH})$ 等价于

$$\begin{aligned} & \lambda \cdot \left[\sum_{i=2}^N (a_i p_1^* - p_i^*) Z_i \right]^2 + (2p_1^* - 1) \cdot \left(\sum_{i=2}^N p_i^* Z_i \right)^2 \\ & \leq p_1^{*2} \cdot \sum_{i=2}^N (a_i - 2p_i^*) Z_i^2. \end{aligned} \quad (9)$$

由 (7) 式知 $a_i \geq p_i^* / p_1^*$. 故根据 Cauchy-Schwarz 不等式得

$$\begin{aligned} (9) \text{ 式左边} & \leq \lambda \cdot \sum_{i=2}^N (a_i p_1^* - p_i^*) \cdot \sum_{i=2}^N (a_i p_1^* - p_i^*) Z_i^2 + (2p_1^* - 1) \cdot \sum_{i=2}^N p_i^* \cdot \sum_{i=2}^N p_i^* Z_i^2 \\ & = (2p_1^* - 1) \cdot \sum_{i=2}^N [\lambda(a_i p_1^* - p_i^*) + (1 - p_i^*) \cdot p_i^*] Z_i^2. \end{aligned} \quad (10)$$

而由 a_i 的选取易知 (10) 式右边不大于 (9) 式右边. 故 (9) 式成立. 又容易看出存在 $Z_i^0 (i=2, \dots, N)$ 使得 (9) 式成立严格不等式. 因此 p 优于 p^* . 证毕.

注 1 由上述定理可知, 如果某个单元的大小超过其余各单元大小的总和, 则 PPS 抽样方案 p^* 能被一致地改进. 我们的解释是, 因为此时有一个单元 (设大小为 M_i) 很大, 所以应以比 M_i/M_0 更大的概率将其抽出. 但是当总体 ($N > 2$) 中有某个单元的大小正好等于其余各单元的大小之和时, p^* 的可容许性却与抽样次数 n 有关. 如果 $n=1$, 即只抽一次, 那么自然是最好能将大单元抽出, 这与上面的说明是一致的. 然而当 $n > 1$ 也即抽样次数多于 1 时, p^* 却不能被改进. 这是有些出人意外的. 这个性质说明, 作为条件 " $\forall i, \text{ 有 } M_i < \frac{M_0}{2}$ " 与条件 " $\exists i_0$,

使得 $M_{i_0} > \frac{M_0}{2}$ 的边界情况, 在条件 “ $\exists i_0$, 使得 $M_{i_0} = \frac{M_0}{2}$ ” 下, PPS 抽样方案当 $n=1$ 时显示

的是在后一条件下的性质, 而当 $n>1$ 时却显示了在前一条件下的性质. 这一现象是有趣的.

最后, 我们给出关于 PPS 抽样方案最优性的结论.

定理 2 PPS 抽样方案 p^* 是最优的充要条件是 $N=2$ 以及 $M_1=M_2$.

证 首先由定理 1 的 (ii) 知, 当 $N=2$ 但 $M_1 \neq M_2$ 时, p^* 不是可容许的, 从而更不是最优的. 其次说明当 $N>2$ 时 p^* 不是最优的. 取 $p_1=1, p_i=0 (i=2, \dots, N)$. 则 p^* 不能一致地优于 $p = \{p_i; i=1, \dots, N\}$. 这是因为由 (2) 式知 $R_{p^*}(\hat{Y}_{HH}) \leq R_p(\hat{Y}_{HH})$ 等价于 $\sum_{i=1}^N p_i^* \left(\frac{Y_i}{p_i^*} - Y \right)^2 \leq n \cdot \left(\frac{Y_1}{p_1^*} - Y \right)^2$. 如果取 $Y_1 = p_1^* Y$, 且使得 $Y_i/p_i^* (i=2, \dots, N)$ 不全相等, 则前式不成立. 最后证明当 $N=2$ 且 $M_1=M_2$ 时 p^* 的最优性. 事实上, 由 (1) 式知, 对任何放回的抽样方案 p , 有

$$\begin{aligned} R_p(\hat{Y}_{HH}) &= \frac{1}{n} (Y_1 - Y_2)^2 + \frac{n-1}{n} (2p_1 Y_1 + 2p_2 Y_2 - Y)^2 \\ &\geq \frac{1}{n} (Y_1 - Y_2)^2 = R_{p^*}(\hat{Y}_{HH}), \end{aligned}$$

故 p^* 是最优的. 证毕.

注 2 由定理 1 知, 对于放回的等概率抽样情形, 即 $M_1 = \dots = M_N$, $p_1^* = \dots = p_N^* = \frac{1}{N}$, p^* 是可容许的. 这说明放回的不等概率抽样不能改进放回的简单随机抽样. 进一步由定理 2 知, 放回的简单随机抽样在 $N=2$ 时还是最优的. 而且由定理 2 的证明过程可以看出, 当 $n>1$ 时, 它是唯一最优的.

参 考 文 献

- 1 冯士雍, 施锡铨. 抽样调查——理论、方法与实践. 上海: 上海科学技术出版社, 1995
- 2 Cochran W G. Sampling Techniques. 3rd ed. New York: John Wiley & Sons, 1977
- 3 Scott A J, Smith T M F. Minimax designs for sample surveys. Biometrika, 1975, 62: 353 ~ 357