

基于变异系数的最优估计量及性质的研究

徐 晨¹, 常桂松¹, 吴文彬²

¹东北大学理学院数学系, 辽宁 沈阳

²抚顺新钢铁有限责任公司智造中心, 辽宁 抚顺

收稿日期: 2022年10月14日; 录用日期: 2022年11月8日; 发布日期: 2022年11月17日

摘 要

讨论调查变量的最优估计, 一直是抽样调查中的重要研究课题。首先本文在一般简单估计基础上提出广义均值估计的定义, 其中一般简单估计可以看作是广义均值估计的特例, 接着讨论广义均值估计的最优估计量与最优估计量的设计效率, 最后通过一个实例计算广义均值估计的最优估计值。

关键词

抽样调查, 变异系数, 最优估计

Research on Optimal Estimation and Properties Based on Coefficient of Variation

Chen Xu¹, Guisong Chang¹, Wenbin Wu²

¹Department of Mathematics, School of Sciences, Northeastern University, Shenyang Liaoning

²Intelligent Manufacture Center, Fushun New Steel Co. Ltd., Fushun Liaoning

Received: Oct. 14th, 2022; accepted: Nov. 8th, 2022; published: Nov. 17th, 2022

Abstract

Discussing the optimal estimation of survey variables has always been an important research topic in survey sampling. Firstly, a generalized mean estimator based on the general simple estimator is defined in this article, and the general simple estimator is a special case of the generalized mean estimator. Then the optimal estimator and the design efficiency of the generalized mean estimator are discussed. Finally, an example is given to calculate the optimal estimator of generalized mean estimator.

Keywords

Survey Sampling, Coefficient of Variation, Optimal Estimation



1. 引言

在抽样调查中, 对于调查变量 Y , 常用样本均值 \bar{y} 来估计总体均值 \bar{Y} , 称为一般简单估计, 而得到调查变量 Y 的最优估计一直是抽样调查的重要研究课题。Singh 与 Pal [1] 研究应用指数估计法在两种情况下的连续性抽样的最优估计。Singh 与 Priyanka [2] 提出了基于两次抽样结果的可提高估计精度的估计模型。Priyanka [3] 等为了解决无响应问题, 提出了一类单参数插补技术, 并提出了相应的点估计。Nath [4] 在两相抽样中使用折衷插补方法, 并用比率加乘积估计人口平均数, 进一步讨论了所提出估计量的优良性。

当已知一些辅助信息, 并满足一些特定条件时, 例如基于与调查变量 Y 相关性较强的辅助变量 X , 可用比估计 $\hat{Y}_R = \frac{\bar{y}}{\bar{x}} \bar{X}$, 回归估计 $\hat{Y}_L = \bar{y} + b(\bar{X} - \bar{x})$ 来估计总体均值 \bar{Y} , 在一定条件下, 比估计与回归估计的估计量的均方误差小于一般简单估计的方差, 即提高了估计的精度[5] [6] [7]。文献[8]-[16]分别提出了估计总体均值时改进的比估计与回归估计方法, 各种估计结果如表 1 所示。

Table 1. Various estimators of population mean and mean squared error of the estimators

表 1. 各种总体均值估计值与估计值的均方误差

估计值	均方误差
$\frac{\bar{y}}{\bar{x}^a} \bar{X}^a$	$\frac{\bar{Y}^2(1-f)}{n} [C_Y^2 + a^2 C_X^2 - 2a\rho C_X C_Y]$
$\bar{y} \frac{\bar{X} + C_X}{\bar{x} + C_X}$	$\frac{\bar{Y}^2(1-f)}{n} \left[C_Y^2 + \left(\frac{\bar{X}}{\bar{X} + C_X} \right)^2 C_X^2 - 2\rho \frac{\bar{X}}{\bar{X} + C_X} \frac{C_Y}{C_X} \right]$
$\frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x}} \bar{X}$	$\frac{1-f}{n} \left[\frac{\bar{Y}^2}{\bar{X}^2} S_X^2 + S_Y^2(1-\rho^2) \right]$
$\frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x} + \rho} (\bar{X} + \rho)$	$\frac{1-f}{n} \left[\frac{\bar{Y}^2}{(\bar{X} + \rho)^2} S_X^2 + S_Y^2(1-\rho^2) \right]$
$\frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x} + C_X} (\bar{X} + C_X)$	$\frac{1-f}{n} \left[\frac{\bar{Y}^2}{(\bar{X} + C_X)^2} S_X^2 + S_Y^2(1-\rho^2) \right]$
$\frac{\bar{y} + b(\bar{X} - \bar{x})}{C_X \bar{x} + \rho} (C_X \bar{X} + \rho)$	$\frac{1-f}{n} \left[\frac{(C_X \bar{Y})^2}{(C_X \bar{X} + \rho)^2} S_X^2 + S_Y^2(1-\rho^2) \right]$
$\frac{\bar{y} + b(\bar{X} - \bar{x})}{\rho \bar{x} + C_X} (\rho \bar{X} + C_X)$	$\frac{1-f}{n} \left[\frac{(\rho \bar{Y})^2}{(\rho \bar{X} + C_X)^2} S_X^2 + S_Y^2(1-\rho^2) \right]$

其中 a , b 为任意实数, C_Y , C_X 分别为调查变量 Y 与辅助变量 X 的变异系数, ρ 为调查变量 Y 与辅助变量 X 相关系数, 一般简单估计和比估计均可以看作是这些估计值的特例, 文献[8]-[17]均进一步讨

论了这些估计值的性质及最优的估计量。若在抽样框中不已知辅助变量 X 的值, 或者不存在与调查变量相关性较高的辅助变量 X , 基于调查变量 Y 的变异系数 C_Y , 其中 C_Y 是调查变量 Y 离散程度的一个归一化度量值, 本文给出广义均值估计的定义, 而且一般简单估计也可以看作广义均值估计的特例, 进而进一步讨论广义均值估计的最优估计量。

2. 广义均值估计

2.1. 广义均值估计的定义及性质

定义 1. 调查变量 Y 的均值的广义均值估计为 $\bar{y}_k = k\bar{y}$, 其中 \bar{y} 为样本均值, k 为任意实数。

定理 1. 调查变量 Y 的均值的广义均值估计 \bar{y}_k 是总体均值 \bar{Y} 的有偏估计。

证明: $E(\bar{y}_k) = E(k\bar{y}) = k\bar{Y}$, 其中 k 为任意实数。□

可见, 当广义估计 \bar{y}_k 中的 $k = 1$ 时, 广义估计即为简单随机抽样中的一般简单估计, 并且是总体均值 \bar{Y} 的无偏估计。

定理 2. 调查变量 Y 的均值的广义均值估计 \bar{y}_k 的均方误差 $MSE(\bar{y}_k)$ 为

$$MSE(\bar{y}_k) = \bar{Y}^2 \left[k^2 \frac{N-1}{Nn} C_Y^2 + (k-1)^2 \right],$$

其中 N 为总体中个体的个数, n 为样本容量, C_Y 为调查变量 Y 的变异系数。

$$\begin{aligned} MSE(\bar{y}_k) &= E[k\bar{y} - \bar{Y}]^2 \\ &= E[k(\bar{y} - \bar{Y}) - (k-1)\bar{Y}]^2 \\ \text{证明:} \quad &= k^2 \text{Var}(\bar{y}_{\text{SRSWR}}) + (k-1)^2 \bar{Y}^2 + 0 \quad \square \\ &= k^2 \frac{N-1}{Nn} S_Y^2 + (k-1)^2 \bar{Y}^2 \\ &= \bar{Y}^2 \left[k^2 \frac{N-1}{Nn} C_Y^2 + (k-1)^2 \right]. \end{aligned}$$

2.2 广义均值估计的最优估计量及均方误差

定理 2. 设调查变量 Y 的变异系数为 C_Y , 则总体均值的广义均值估计 \bar{y}_k 的最优估计 \bar{y}_k^* 为

$$\bar{y}_k^* = k^* \bar{y} = \frac{\bar{y}}{\frac{N-1}{Nn} C_Y^2 + 1}.$$

证明: 由定理 2 中的结论可知, 当 $\frac{dMSE(\bar{y}_k)}{dk} = 0$ 时, 可解出广义均值估计 \bar{y}_k 均方误差 $MSE(\bar{y}_k)$ 的极值,

$$\begin{aligned} \frac{dMSE(\bar{y}_k)}{dk} &= \bar{Y}^2 \left[2k \frac{N-1}{Nn} C_Y^2 + 2(k-1) \right] \\ \Rightarrow k \left[\frac{N-1}{Nn} C_Y^2 + 1 \right] &= 1 \\ \Rightarrow k &= \frac{1}{\frac{N-1}{Nn} C_Y^2 + 1} = k^*. \end{aligned}$$

$$\text{即 } \bar{y}_k^* = k^* \bar{y} = \frac{\bar{y}}{\frac{N-1}{Nn} C_Y^2 + 1}。 \square$$

推论 1. 调查变量 Y 总体均值的广义均值估计的最优估计 \bar{y}_k^* 的均方误差 $MSE(\bar{y}_k^*)$ 为

$$MSE(\bar{y}_k^*) = \frac{\bar{Y}^2 \frac{N-1}{Nn} C_Y^2}{\frac{N-1}{Nn} C_Y^2 + 1}.$$

$$\begin{aligned} MSE(\bar{y}_k^*) &= \bar{Y}^2 \left[k^{*2} \frac{N-1}{Nn} C_Y^2 + (k^* - 1)^2 \right] \\ &= \bar{Y}^2 \left[\frac{\frac{N-1}{Nn} C_Y^2}{\left(\frac{N-1}{Nn} C_Y^2 + 1 \right)^2} + \frac{\left(\frac{N-1}{Nn} \right)^{2\theta} C_Y^4}{\left(\frac{N-1}{Nn} C_Y^2 + 1 \right)^2} \right] \end{aligned}$$

证明:

$$\begin{aligned} &= \frac{\bar{Y}^2 \frac{N-1}{Nn} C_Y^2}{\left(\frac{N-1}{Nn} C_Y^2 + 1 \right)^2} \left[1 + \frac{N-1}{Nn} C_Y^2 \right] \quad \square \\ &= \frac{\bar{Y}^2 \frac{N-1}{Nn} C_Y^2}{\frac{N-1}{Nn} C_Y^2 + 1}. \end{aligned}$$

可见广义均值估计的最优估计量的设计效率 $Deff$ 为

$$\frac{MSE(\bar{y}_k^*)}{Var(\bar{y})_{SRSWR}} = \frac{\bar{Y}^2 \frac{N-1}{Nn} C_Y^2}{\frac{N-1}{Nn} C_Y^2 + 1} \cdot \frac{1}{Var(\bar{y})_{SRSWR}} = \frac{Var(\bar{y})_{SRSWR}}{\frac{N-1}{Nn} C_Y^2 + 1} \cdot \frac{1}{Var(\bar{y})_{SRSWR}} = \left[\frac{N-1}{Nn} C_Y^2 + 1 \right]^{-1} \leq 1.$$

3. 一个应用例题

若调查某高校学生的月消费情况 Y (单位: 元), 该高校共有在校学生 30,000 人, 现采用简单随机抽样方案, 随机调查 1000 名学生, 得到数据

$$\sum_{i=1}^{1000} y_i = 1135453, \quad \sum_{i=1}^{1000} y_i^2 = 33545539665,$$

- 按照一般简单估计, 估计每位高校学生每月的月消费情况的平均值为

$$\hat{Y} = \bar{y} = \frac{\sum_{i=1}^{1000} y_i}{1000} = \frac{1135453}{1000} \approx 1135.45,$$

并且一般简单估计的均方误差的估计值为

$$\hat{MSE}(\hat{Y}) = \frac{1-f}{n} S_Y^2 = \frac{1-f}{n} \frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right] \approx 176.67^2.$$

- 按照广义均值估计, 首先调查变量 Y 的变异系数 C_Y 的估计值为

$$\hat{C}_Y = \frac{s_Y}{\bar{y}} = \frac{\left[\frac{1}{n-1} \left[\sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 \right] \right]^{\frac{1}{2}}}{\frac{1}{n} \sum_{i=1}^n y_i} \approx 5.004,$$

则应用广义均值估计的最优估计量，来估计每位高校学生每月的月消费情况的平均值为

$$\hat{Y} = \bar{y}_k^* = k^* \bar{y} = \frac{\bar{y}}{\frac{N-1}{Nn} \hat{C}_Y^2 + 1} \approx 1107.71,$$

并且广义均值估计的最优估计量均方误差的估计值为

$$M\hat{S}E(\bar{y}_k^*) = \frac{\hat{Y}^2 \frac{N-1}{Nn} \hat{C}_Y^2}{\frac{N-1}{Nn} \hat{C}_Y^2 + 1} \approx 173.14^2.$$

由此可见广义均值估计的最优估计量与一般简单估计比较，偏差不大，并且广义均值估计的最优估计量的均方误差小。

4. 总结

1) 本文讨论广义均值估计，显然一般简单估计是广义均值估计的特例，并且基于调查变量 Y 的变异系数为 C_Y ，可确定广义均值估计中参数 k 的值，使得广义均值估计的均方误差达最小，即解出广义均值估计的最优估计量。与一般简单估计的抽样误差相比，广义均值估计的最优估计量的设计效率 $Deff \leq 1$ ，因此广义均值估计的最优估计量是优于一般简单估计的。在实际应用中若不知调查变量 Y 的变异系数为 C_Y ，可用样本变异系数来估计总体的变异系数 C_Y 。

2) 若在抽样框中已知与调查变量相关性较高的辅助变量 X 的值，今后工作中还可以进一步结合文献 [8]-[17] 中的结论，讨论广义均值比估计 $\hat{Y}_{kR\alpha} = k \frac{\bar{y}}{\bar{x}^\alpha} \bar{X}^\alpha$ ， $\hat{Y}_{kR} = k\bar{y} \frac{\bar{X} + C_X}{\bar{x} + C_X}$ ，或者广义均值回归估计

$$\hat{Y}_{kL} = k \frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x}} \bar{X}, \quad \hat{Y}_{kL\rho} = k \frac{\bar{y} + b(\bar{X} - \bar{x})}{\bar{x} + \rho} (\bar{X} + \rho) \text{ 等等。}$$

3) 若在抽样框中已知与调查变量相关性较高的两个辅助变量 X_1 与 X_2 的值，结合文献 [17] 中的结论今后工作中还可以进一步讨论广义均值比估计 $\hat{Y}_{kR\alpha} = k_1 \frac{\bar{y}}{\bar{x}_1^{a_1}} \bar{X}^{a_1} + k_2 \frac{\bar{y}}{\bar{x}_2^{a_2}} \bar{X}^{a_2}$ ，

$$\hat{Y}_{kR} = k_1 \bar{y} \frac{\bar{X} + C_{X_1}}{\bar{x} + C_{X_1}} + k_2 \bar{y} \frac{\bar{X} + C_{X_2}}{\bar{x} + C_{X_2}} \text{ 等等。}$$

基金项目

国家自然科学基金青年基金《不定度量量子流形的相关问题研究》，项目批准号：NSFC 1180106。

参考文献

- [1] Singh, H.P. and Pal, S.K. (2017) Search of Good Rotation Patterns Using Exponential Method of Estimation in Two-Occasion Successive Sampling. *Communication in Statistics—Theory and Methods*, **46**, 5466-5486. <https://doi.org/10.1080/03610926.2015.1104351>
- [2] Singh, G.N. and Priyanka, K. (2008) Search of Good Rotation Patterns to Improve the Precision of Estimates at Current Occasion. *Communication in Statistics—Theory and Methods*, **3**, 337-348.

<https://doi.org/10.1080/03610920701653052>

- [3] Singh, G.N., Priyanka, K., Kim, J.M., *et al.* (2010) Estimation of Population Mean Using Imputation Techniques in Sample Surveys. *Journal of the Korean Statistical Society*, **39**, 67-74. <https://doi.org/10.1016/j.jkss.2009.04.002>
- [4] Nath, K. and Singh, B.K. (2018) Population Mean Estimation Using Ratio-Cum Product Compromised-Method of Imputation in Two-Phase Sampling Scheme. *Asian Journal of Mathematics & Statistics*, **11**, 27-39. <https://doi.org/10.3923/ajms.2018.27.39>
- [5] 李金昌. 应用抽样技术[M]. 北京: 科学出版社, 2010: 22-81.
- [6] 孙山泽. 抽样调查[M]. 北京: 北京大学出版社, 2004: 13-50.
- [7] Kish, L. 抽样调查[M]. 倪加勋(主译), 孙山泽(校译). 北京: 中国统计出版社, 1997: 229-242.
- [8] 徐晨, 何川. 一种广义比估计及其性质的研究[J]. 统计学与应用, 2019, 8(6): 895-900.
- [9] Shashi, B. and Praveen, K.M. (2017) A Family of Unbiased Estimators of Population Mean Using an Auxiliary Variable. *Advances in Computational Sciences and Technology*, **10**, 129-137.
- [10] Gupta, R.K. and Yadav, S.K. (2017) New Efficient Estimators of Population Mean Using Non-Traditional Measures of Dispersion. *Open Journal of Statistics*, **7**, 394-404. <https://doi.org/10.4236/ojs.2017.73028>
- [11] Uraivan, J. and Nuanpan, L. (2019) A Combined Family of Ratio Estimators for Population Mean Using an Auxiliary Variable in Simple Random Sampling. *Journal of Mathematical and Fundamental Sciences*, **51**, 1-12. <https://doi.org/10.5614/j.math.fund.sci.2019.51.1.1>
- [12] Mir, S., Showkat, M., *et al.* (2018) Efficient Estimators of Population Mean Using Auxiliary Information under Simple Random Sampling. *Statistics in Transition New Series*, **19**, 219-238. <https://doi.org/10.21307/stattrans-2018-013>
- [13] Kalim, U., Zavar, H., *et al.* (2022) Estimation of Finite Population Mean in Simple and Stratified Random Sampling by Utilizing the Auxiliary, Ranks, and Square of the Auxiliary Information. *Mathematical Problems in Engineering*, **2022**, Article ID: 5263492. <https://doi.org/10.1155/2022/5263492>
- [14] Anum, I., Hongbo, S., *et al.* (2022) Efficient Estimators of Finite Population Mean Based on Extreme Values in Simple Random Sampling. *Mathematical Problems in Engineering*, **2022**, 15-24. <https://doi.org/10.1155/2022/5866085>
- [15] Pal, S.K., *et al.* (2018) A Family of Efficient Estimators of The Finite Population Mean in Simple Random Sampling. *Journal of Statistical Computation and Simulation*, **88**, 920-934. <https://doi.org/10.1080/00949655.2017.1408808>
- [16] Irfan, M., Javed, M. and Lin, Z. (2018) Efficient Ratio-Type Estimators of Finite Population Mean Based on Correlation Coefficient. *International Journal of Science and Technology*, **25**, 2361-2372.
- [17] 卢静莉. 有限总体均值估计的新方法[J]. 统计与决策, 2021, 37(10): 20-23.