Statistics and Application 统计学与应用, 2019, 8(6), 895-900
Published Online December 2019 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2019.86101

Hans汉斯

Approach to a Generalized Ratio Estimator and the Optimality

Chen Xu, Chuan He

Department of Mathematics, Northeastern University, Shenyang Liaoning
Email: xuchen@mail.neu.neu.cn, hexiaodong9@163.com

Received: Nov. 14th, 2019; accepted: Nov. 27th, 2019; published: Dec. 4th, 2019

Abstract

Ratio estimation of population totals is one of the oldest uses of auxiliary information in survey sampling. A new generalized ratio estimator is proposed in this article, and the usual ratio estimator is a special case of the new generalized ratio estimator. Then this article discusses under what circumstances, the mean squared error is minimum. Also the minimum mean squared error is derived. Finally, a meaningful application is given.

Keywords

Ratio Estimator, Mean Square Error, Survey Sampling

一种广义比估计及其性质的研究

徐　晨，何　川

东北大学理学院数学系，辽宁　沈阳
Email: xuchen@mail.neu.neu.cn, hexiaodong9@163.com

收稿日期：2019年11月14日；录用日期：2019年11月27日；发布日期：2019年12月4日

摘　要

抽样调查中比估计是一种可以提高抽样精度的估计方法，本文在一般比估计基础之上提出一种广义比估计方法，并说明一般比估计方法只是广义比估计方法的特例。接着本文中讨论了广义比估计的最优值及其估计量的优良性，最后本文中讨论了一个实际应用实例。

关键词

比估计，均方误差，抽样调查

文章引用: 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.
DOI: 10.12677/sa.2019.86101

1. 引言

抽样调查中常常利用比估计来提高调查的精度，对于调查的变量 Y，如果有一个与调查变量 Y 相关性较高的辅助变量 X，并且利用已有资料知道总体中辅助变量 X 的总值和均值，即可考虑利用比估计来估计调查变量 Y 的总值或者均值[1]。另外，如果需要调查总体中两个变量 X 与 Y 的比值，也可以利用样本比值给出总体比值的估计。因此比估计是抽样调查中常用的方法，适用面广，常用于简单随机抽样法中，也可以用于分层随机抽样法中。

首先给出利用已知的辅助变量 X 的信息构造比估计，可以提高调查变量 Y 估计值精度的介绍。简单随机抽样下的比估计定义为：总体的两个变量 Y 和 X 的总值或均值的比率为 $R=\frac{\bar{Y}}{\bar{X}}=\frac{Y}{X}$，简单随机样本中两个变量 Y 和 X 的总值或者均值的比率，也就是总体比例的估计值 $\hat{R}=\frac{\bar{y}}{\bar{x}}=\frac{y}{x}$，因此调查变量 Y 的均值的比估计为 $\hat{\bar{Y}}_R=\hat{R}\bar{X}=\frac{\bar{y}}{\bar{x}}\bar{X}$，调查变量 Y 的总值的估计值为 $\hat{Y}_R=\hat{R}X=\frac{\bar{y}}{\bar{x}}X=N\hat{R}\bar{X}$，其中辅助变量 X 的总值或者均值需已知[2] [3]。

由抽样调查中的知识可知比估计有如下性质[4]-[11]：

- 比估计和变量 Y 的均值估计均为有偏估计：$E\left(\hat{R}-R\right)=O\left(\frac{1}{n}\right)$，$E\left(\hat{\bar{Y}}_R-\bar{Y}\right)=O\left(\frac{1}{n}\right)$。
- 比估计和变量 Y 的均值估计的均方误差分别为：

$$MSE\left(\hat{R}\right)=E\left(\hat{R}-R\right)^2=\frac{1-f}{n}\cdot\frac{1}{N-1}\cdot\frac{1}{\bar{X}^2}\cdot\sum_{i=1}^{N}\left(Y_i-RX_i\right)^2+O\left(\frac{1}{n^{3/2}}\right)=O\left(\frac{1}{n}\right),$$

$$MSE\left(\hat{\bar{Y}}_R\right)=E\left(\hat{\bar{Y}}_R-\bar{Y}\right)^2=\frac{1-f}{n}\cdot\frac{1}{N-1}\cdot\sum_{i=1}^{N}\left(Y_i-RX_i\right)^2+O\left(\frac{1}{n^{3/2}}\right)=\frac{1-f}{n}\cdot\left(S_Y^2-2R\rho S_XS_Y+R^2S_X^2\right)+O\left(\frac{1}{n^{3/2}}\right)=O\left(\frac{1}{n}\right),$$

其中抽样比 $f=\frac{n}{N}$，相关系数 $\rho=\frac{S_{XY}}{S_XS_Y}$。

- 当满足 $R^2S_X^2\le 2R\rho S_XS_Y$，即 $\rho\ge\frac{C_X}{2C_Y}$ 时，其中变异系数 $C_X=\frac{S_X}{\bar{X}}$，$C_Y=\frac{S_Y}{\bar{Y}}$，比估计方法均值 $\bar{Y}$ 的估计值 $\hat{\bar{Y}}_R$ 的均方误差优于简单估值法 $\hat{\bar{Y}}=\bar{y}$ 的均方误差。

2. 广义比估计定义以及性质

2.1. 广义比估计定义

文[12]中对于超总体模型提出了一种广义差估计方法，文[13]中对于超总体模型提出了一种广义比估计方法，并与文[12]中的广义差估计方法进行了比较，讨论了其优良性。本文中对于总体中的二元变量 X 与 Y 给出一个广义比估计的定义，总体的两个变量总值或均值的广义比率

$$R_a = \frac{\bar{Y}}{\bar{X}^a} = \frac{Y}{N^{1-a}X^a}$$

其中 a 为任意实数，简单随机样本中两个变量总值或者均值的广义比，即总体广义比的估计值 $\hat{R}_a = \frac{\bar{y}}{\bar{x}^a} = \frac{y}{n^{1-a}x^a}$，因此调查变量 Y 的均值的广义比估计值为 $\hat{\bar{Y}}_{Ra} = \hat{R}_a\bar{X} = \frac{\bar{y}}{\bar{x}^a}\bar{X}^a$，同样调查变量 Y 的总值的广义比估计为 $\hat{Y}_{Ra} = N\frac{\bar{y}}{\bar{x}^a}\bar{X}^a$，特殊的当 $a = 1$ 时，调查变量 Y 的均值的广义比估计即为一般的比估计，广义比估计与一般比估计要求一样，需要已知辅助变量 X 的总值或者均值；当 $a = 0$ 时，调查变量 Y 的均值的广义比估计即为一般的简单估计。

2.2. 广义比估计的性质

由广义比估计的定义可以推知广义比估计有如下性质：

定理 1. 调查变量 Y 的均值的广义比估计为有偏估计。

$$E\left(\hat{\bar{Y}}_{Ra} - \bar{Y}\right) = \frac{a\bar{Y}(1-f)C_X}{n}\left[\left(\frac{a+1}{2}\right)C_X - \rho C_Y\right]$$

其中抽样比 $f = \frac{n}{N}$，相关系数 $\rho = \frac{S_{XY}}{S_X S_Y}$，变异系数 $C_X = \frac{S_X}{\bar{X}}$，$C_Y = \frac{S_Y}{\bar{Y}}$。

证明： 记 $\varepsilon_0 = \frac{\bar{x} - \bar{X}}{\bar{X}}$，即 $\bar{x} = (1+\varepsilon_0)\bar{X}$，$\varepsilon_1 = \frac{\bar{y} - \bar{Y}}{\bar{Y}}$，即 $\bar{y} = (1+\varepsilon_1)\bar{Y}$，则可知 $E\varepsilon_0 = 0$，$E\varepsilon_1 = 0$，$E\varepsilon_0^2 = \frac{1-f}{n}C_X^2$，$E\varepsilon_1^2 = \frac{1-f}{n}C_Y^2$，$E\varepsilon_0\varepsilon_1 = \frac{1-f}{n}\rho C_X C_Y$。

又因为

$$\begin{aligned}\hat{\bar{Y}}_{Ra} &= \frac{\bar{y}}{\bar{x}^a}\bar{X}^a = \frac{(1+\varepsilon_1)\bar{Y}}{(1+\varepsilon_0)^a\bar{X}^a}\bar{X}^a = \bar{Y}(1+\varepsilon_1)(1+\varepsilon_0)^{-a} \\ &= \bar{Y}(1+\varepsilon_1)\left(1 - a\varepsilon_0 + \frac{a(a+1)}{2}\varepsilon_0^2 + \cdots\right) \\ &= \bar{Y}\left(1 + \varepsilon_1 - a\varepsilon_0 - a\varepsilon_0\varepsilon_1 + \frac{a(a+1)}{2}\varepsilon_0^2 + \cdots\right)\end{aligned},$$

即

$$\hat{\bar{Y}}_{Ra} - \bar{Y} \approx \bar{Y}\left(\varepsilon_1 - a\varepsilon_0 - a\varepsilon_0\varepsilon_1 + \frac{a(a+1)}{2}\varepsilon_0^2\right),$$

所以

$$\begin{aligned}E\left(\hat{\bar{Y}}_{Ra} - \bar{Y}\right) &\approx \bar{Y}E\left(\varepsilon_1 - a\varepsilon_0 - a\varepsilon_0\varepsilon_1 + \frac{a(a+1)}{2}\varepsilon_0^2\right) \\ &= \bar{Y}\left(0 - 0 - a\frac{1-f}{n}\rho C_X C_Y + \frac{a(a+1)}{2}\frac{1-f}{n}C_X^2\right)。 \\ &= \frac{a\bar{Y}(1-f)C_X}{n}\left[\left(\frac{a+1}{2}\right)C_X - \rho C_Y\right]\end{aligned}$$

定理 1 得证。

定理 2. 调查变量 Y 的均值的广义比估计的均方误差为

$$MES\left(\hat{\bar{Y}}_{Ra}\right)=E\left(\hat{\bar{Y}}_{Ra}-\bar{Y}\right)^2=\frac{\bar{Y}^2(1-f)}{n}\left[C_Y^2+a^2C_X^2-2a\rho C_XC_Y\right]。$$

证明： $MES\left(\hat{\bar{Y}}_{Ra}\right)=E\left(\hat{\bar{Y}}_{Ra}-\bar{Y}\right)^2$

$$\approx\bar{Y}^2E\left(\varepsilon_1-a\varepsilon_0-a\varepsilon_0\varepsilon_1+\frac{a(a+1)}{2}\varepsilon_0^2\right)^2。$$

$$\approx\bar{Y}^2E\left(\varepsilon_1-a\varepsilon_0-a\varepsilon_0\varepsilon_1\right)^2$$

$$=\frac{\bar{Y}^2(1-f)}{n}\left[C_Y^2+a^2C_X^2-2a\rho C_XC_Y\right]$$

定理 2 得证。

定理 3. 调查变量 Y 的均值的广义比估计 $\hat{\bar{Y}}_{Ra}=\frac{\bar{y}}{\bar{x}^a}\bar{X}^a$，当 $a=\rho\frac{C_Y}{C_X}$ 时，估计值 $\hat{\bar{Y}}_{Ra}$ 的均方误差达最小，最小值为 $\min\left\{MES\left(\hat{\bar{Y}}_{Ra}\right)\right\}=\frac{1-f}{n}S_Y^2\left(1-\rho^2\right)$。

证明： 由定理 2 可知 $MES\left(\hat{\bar{Y}}_{Ra}\right)=\frac{\bar{Y}^2(1-f)}{n}\left[C_Y^2+a^2C_X^2-2a\rho C_XC_Y\right]$，因此对 a 求导可得：

$$\frac{\mathrm{d}\left(MES\left(\hat{\bar{Y}}_{Ra}\right)\right)}{\mathrm{d}a}=\frac{\bar{Y}^2(1-f)}{n}\left[2aC_X^2-2\rho C_XC_Y\right]=0，$$

求得 $a_{\min}=\rho\frac{C_Y}{C_X}$，并且 $\hat{\bar{Y}}_{Ra}$ 的均方误差最小值

$$\min\left\{MES\left(\hat{\bar{Y}}_{Ra}\right)\right\}=\frac{\bar{Y}^2(1-f)}{n}\left[C_Y^2+\left(\rho\frac{C_Y}{C_X}\right)^2C_X^2-2\rho\frac{C_Y}{C_X}\rho C_XC_Y\right]=\frac{1-f}{n}S_Y^2\left(1-\rho^2\right)。$$

定理 3 得证。

由此可见广义比估计方法调查变量 Y 的均值 $\bar{Y}$ 的估计值 $\hat{\bar{Y}}_{Ra}$ 的最小的均方误差优于简单估值法 $\hat{\bar{Y}}=\bar{y}$ 的均方误差 $MES\left(\hat{\bar{Y}}\right)=\frac{1-f}{n}S_Y^2$，并且当调查变量 Y 与辅助变量 X 相关性越高时，广义比估计方法调查变量 Y 的均值 $\bar{Y}$ 的估计值 $\hat{\bar{Y}}_{Ra}$ 的最小均方误差越小。

3. 一个应用例题

调查某一社区居民用于食物的消费的支出，若该社区有居民共 300 户，共 1100 人，现简单随机抽样调查了其中的 35 户居民，调查各户的月食物支出 Y (单位：元)和家庭人口 X，得数据：

$$\sum_{i=1}^{35}x_i=120，\ \sum_{i=1}^{35}y_i=31350，\ \sum_{i=1}^{35}x_i^2=450，\ \sum_{i=1}^{35}y_i^2=29692900，\ \sum_{i=1}^{35}x_iy_i=114440。$$

- 按照简单估值法，估计每户每月用于食物的平均支出的估计值为 $\hat{\bar{Y}}=\bar{y}=\frac{\sum_{i=1}^{35}y_i}{35}=\frac{31350}{35}\approx895.70(元)$，这一估计的均方误差的估计值为：

$$\hat{MES}\left(\hat{\bar{Y}}\right)=\frac{1-f}{n}s_Y^2=\frac{1-f}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2\right]\approx 1196.77$$

- 按照比估计法，以每户人数 X 作辅助变量，记 $R=\frac{\bar{Y}}{\bar{X}}$ 其估计值为 $\hat{R}=\frac{\bar{y}}{\bar{x}}=\frac{y}{x}=\frac{31350}{120}=261.25$，因此每户每月用于食物的平均支出的估计值为 $\hat{\bar{Y}}_R=\hat{R}\bar{X}=261.25\times\frac{1100}{300}\approx 957.92(\text{元})$，这一估计的均方误差的估计值为：

$$\begin{aligned}\hat{MES}\left(\hat{\bar{Y}}_R\right)&=\frac{1-f}{n}\cdot\frac{1}{n-1}\cdot\sum_{i=1}^{n}\left(y_i-\hat{R}x_i\right)^2\\&=\frac{1-f}{n}\frac{1}{n-1}\left[\sum_{i=1}^{n}y_i^2-2\hat{R}\sum_{i=1}^{n}x_iy_i+\hat{R}^2\sum_{i=1}^{n}x_i^2\right]\\&\approx 453.6942\end{aligned}$$

- 按照广义比估计法，以每户人数 X 作辅助变量，记 $R_a=\frac{\bar{Y}}{\bar{X}^a}$，其中 $a=\rho\frac{C_Y}{C_X}=\frac{S_{XY}}{S_XS_Y}\frac{S_Y/\bar{Y}}{S_X/\bar{X}}=\frac{S_{XY}}{S_X^2}\frac{\bar{X}}{\bar{Y}}\approx 0.0607$ 时，估计值 $\hat{\bar{Y}}_{Ra}$ 的均方误差达最小。因此每户每月用于食物的平均支出的估计值为 $\hat{\bar{Y}}_{Ra}=\hat{R}_a\bar{X}=\frac{\bar{y}}{\bar{x}^a}\bar{X}^a=\frac{31350/35}{(120/35)^{0.0607}}\left(\frac{1100}{300}\right)^{0.0607}\approx 899.37209$，这一估计的均方误差的估计值为：$\hat{MES}\left(\hat{\bar{Y}}_{Ra}\right)=\frac{1-f}{n}S_Y^2\left(1-\rho^2\right)=1196.77\times\left(1-0.8819^2\right)\approx 265.98499$。

  由此可见广义比估计法的估计值相对而言比较适中，并且其估计值的均方误差最小。

4. 总结

本文中讨论了一种广义比估计方法，显然一般比估计方法只是广义比估计方法的特例，广义比估计法中有一个参数 a，可以先根据具体问题确定参数 a 的值使得广义比估计法的估计值的均方误差达最小，并且本文证明了广义比估计方法得到的估计值的均方误差是小于简单估值法估计的均方误差。另外，一般比估计方法要求辅助变量 X 与调查变量 Y 有强相关性，但是广义比估计法没有此要求，当然如果辅助变量 X 与调查变量 Y 有较强相关性，则广义比估计法的估计值的均方误差会更小，效果则更优。

基金项目

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

参考文献

[1] 李金昌. 应用抽样技术[M]. 北京: 科学出版社, 2010: 98-122.

[2] 冯士雍, 施锡铨. 抽样调查-理论, 方法与实践[M]. 上海: 上海科学技术出版社, 1994: 20-30.

[3] 孙山泽. 抽样调查[M]. 北京: 北京大学出版社, 2004: 13-50.

[4] Cochran, W.G. (1978) Contributions to Survey Sampling and Applied Statistics. Academic Press Inc., New York, 3-10. https://doi.org/10.1016/B978-0-12-204750-3.50008-3

[5] Fuller, W.A. (2009) Sampling Statistics. John Wiley & Sons Inc., Hoboken, NJ, 96-110.

[6] Chaudhuri, A. and Stenger, H. (2005) Survey Sampling Theory and Methods. Second Edition, Taylor & Francis Group, New York, 48-60. https://doi.org/10.1201/9781420028638

[7] Foreman, E.K. (1991) Survey Sampling Principles. Marcel Dekker Inc., New York, 48-60.

[8] 倪加勋(主译), 孙山泽(校译). 抽样调查[M]. 北京: 中国统计出版社, 1997: 229-242.

[9] Page, C., Kreling, D. and Matsumura, E.M. (1993) Comparison of the Mean Per Unit and Ratio Estimators under a Simple Applications-Motivated Model. *Statistics & Probability Letters*, **17**, 97-104. https://doi.org/10.1016/0167-7152(93)90003-2

[10] Nassiuma, D.K. (2001) Survey Sampling: Theory and Methods. Nairobi University Press, Nairobi, 50-85.

[11] Sarndal, C.-E., Bengt, S. and Jan, W. (1992) Model Assisted Survey Sampling. Springer-Verlag, New York, 31-60.

[12] Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1977) Foundation of Inference in Survey Sampling. John Wiley, New York, 15-30.

[13] 邹国华, 冯士雍. 广义比估计与广义差估计及其优良性[J]. 系统科学与数学, 1998, 18(3): 359-365.