Jour. Ind. Soc. Ag. Statistics
46(1), 1994 : 147–155

Variance Estimation Techniques in Categorical Data Analysis for Survey Data*

Anil Rai
I.A.S.R.I., New Delhi

Summary

In case of categorical data analysis of survey data, the modification to the ordinary chi-square or likelihood-ratio test statistic is necessary to draw the above required inferences correctly. These modifications can be improved by estimating the various parameters of interest like, vector of cell proportions and its variance-covariance matrix by taking care of the survey design under which the data has been collected. In this study different variance estimation techniques were compared for a proposed combined ratio estimator of cell proportions and it was found that Taylor linearized variance estimator is most suitable after considering the magnitude and bias of variance estimator.

Key words : Taylor Linearized variance estimator, Jackknife variance estimator, Balanced Repeated Replication, Bias.

Introduction

The analysis of categorical data is very common in large scale surveys. For example investigator may wish to find out the homogeneity of proportion of farmers growing different varieties of various crops with previous surveys in India and abroad. Researcher may like to study the association of the various factors with the help of the test of independence of attributes to get an idea about the structure of the population which will help him considerably in planning of future surveys. Finally, one may be interested to find out that the target population of the farmers of different categories has been covered or not under a developmental project with the help of goodness-of-fit test.

In case of categorical data analysis of survey data, the modification to the ordinary chi-square or likelihood-ratio test statistic is necessary to draw the above required inferences correctly. Further, these modifications can be improved by estimating the various parameters of interest like, vector of cell proportions and its variance-covariance matrix by taking care of the survey design under which the data has been collected.

There has been mainly three methods of sampling in case of categorical data analysis. In the first case total sample size is fixed and units are selected

[*] The paper was presented at the 47th Annual conference of the society at S.V. Agricultural College, Tirupati and adjudged best for "ISAS young Scientist Award, 1993".

directly from the population. In second case, one of the marginal frequencies of the contingency table is fixed and selection of the units is made within the categories as in the case of stratified or cluster sampling. The selected units in both the cases are placed in different cells of contingency table according to their inherent set of characteristics, whereas in third case of controlled comparative trials the units are placed in to different cells according to their levels of different set of treatment applied to selected units. The first two methods are common in large scale surveys whereas third method is mainly applied to animal experiments.

In this study different variance estimation techniques were compared for a proposed combined ratio estimator of cell proportions by closely following the approaches of Krewski and Rao [1], Rao and Wu [4].

2. *Combined Ratio Estimator of Cell Proportions :*

Most common sampling design in survey sampling is stratified multi-stage sampling because of its advantages over the other sampling techniques. Let us assume that population is divided in to L strata, N_h is the total number of primary sampling units (PSU's) in the h-th stratum, M_{ht} is the total number of secondary sampling units (SSU's) in the t-th PSU of the h-th stratum and let $M_h = \sum_{t=1}^{N_h} M_{ht}$ denote total number of SSU's in the h-th stratum out of total M, SSU's in the population. Let n_h PSU's are selected with inclusion probability π_{ht} from h-th stratum. Let m_{ht} denote the total number of SSU's selected from t-th PSU and h-th stratum.

Define,

$$Y_{ihtk} = 1, \text{ if k–th SSU from t–th PSU of h–th stratum falls in i–th category}$$
$$= 0, \text{ otherwise}$$
$$x_{htk} = 1, \text{ if k–th SSU from t–th PSU falls in h–th stratum}$$
$$= 0, \text{ otherwise}$$

Let

$$\hat{Z}_{iht} = \sum_{k=1}^{m_{ht}} \frac{Y_{ihtk}}{\pi_{htk}} \quad h = 1, 2 \ldots L; t = 1, 2, \ldots n_h$$

$$\hat{U}_{ht} = \sum_{k=1}^{m_{ht}} \frac{x_{htk}}{\pi_{htk}} \qquad h = 1, 2. . . L; t = 1, 2, . . . n_h$$

$$\bar{z}_i = \sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{t=1}^{n_h} \frac{\hat{z}_{iht}}{M_h \pi_{ht}}$$

$$\bar{u} = \sum_{h=1}^{L} \frac{W_h}{n_h} \sum_{t=1}^{n_h} \frac{\hat{U}_{ht}}{M_h \pi_{ht}}$$

So, combined ratio estimator of i-th cell is given by

$$\hat{P}_{ic} = \frac{\bar{z}_i}{\bar{u}} = \hat{\theta} \quad \text{(say)} \qquad . . . (1)$$

3. *Comparison of Variance Estimation Techniques*

Many large scale surveys now involve large number of strata with relatively few PSU's selected with in each stratum. The various variance-covariance estimators of the above non-linear statistics like Taylor linearization, jackknifing, Balanced repeated Replication etc. were compared for categorical data analysis under the regularity conditions discussed in Krewski and Rao [1].

3.1 *Taylor Linearized and Jackknifed Variance Estimators :*

Let $\hat{\theta}^{ht} = \dfrac{\bar{z}_i^{ht}}{\bar{u}^{ht}}$ where $\theta^h = \sum_{t=1}^{n_h} \hat{\theta}^{ht}$ and where $\bar{z}_i^{ht}$ is the unbiased estimator $E(z_i) = \bar{Z}$ from the sample after omitting t-th PSU from h-th stratum. Similarly, $\bar{u}^{ht}$ is an unbiased estimator of $\bar{U} = E(\bar{u})$ after omitted t-th PSU from h-th stratum $t = 1, 2, . . ., N_h$; $h = 1, 2, . . . , L$; $i = 1, 2, . . . , I$. We can also express

$$\bar{z}_i^{ht} = \bar{z}_i + W_h (\bar{z}_{ih} - z_{iht})/(n_h - 1) \qquad . . . (2)$$

$$\bar{u}^{ht} = \bar{u} + W_h (\bar{u}_h - u_{ht})/(n_h - 1)$$

Now under the asymptotic frame work by linearization of $\hat{\theta}$ we can see that Taylor linearized variance can be written as

$$v_L = \frac{1}{u^{-2}} \sum_{h=1}^{L} \frac{W_h^2}{n_h} s^{\overset{2}{x_2}}_{e_h} \qquad . . . (3)$$

where, $$\hat{e}_{ht} = (z_{iht} - \bar{z}_{ih}) - \frac{\bar{Z}_i}{\bar{Z}} (z_{ht} - \bar{z}_h)$$

$$(n_h - 1) s^2_{\hat{e}_h} = \sum_{t=1}^{n_h} \hat{e}^2_{ht}$$

Now consider the most commonly used jackknife variance estimator $v_j(\hat{\theta})$ which can be written as

$$v_j(\hat{\theta}) = \sum_{h=1}^{L} \frac{n_h - 1}{n_h} \sum_{t=1}^{n_h} (\hat{\theta}^{ht} - \hat{\theta})^2$$

with the help of the equation (2) and linearization we can find that

$$v_j(\hat{\theta}) = v_L(\hat{\theta}) + \frac{2}{\bar{u}^3} \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h - 1)} s^2_{\hat{e}_{uh}} + O_p(n^{-2.5}) \qquad \ldots (4)$$

where $$(n_h - 1) s^2_{\hat{e}_{uh}} = \sum_{t=1}^{n_h} \hat{e}^2_{ht} (u_{ut} - \bar{u}_h)$$

Further,

$$v_j(\hat{\theta}) = v_L(\hat{\theta}) + \frac{2}{\bar{u}^3} \sum_{h=1}^{L} \frac{W_h^2}{n_h(n_h - 1)} s^2_{\hat{e}_{uh}} + O_p(n^{-2.5})$$

where,

$$(n_h - 1) s^2_{\hat{e}_{uh}} = \sum_{t=1}^{n_h} (\hat{e}_{ht} - \bar{e})^2 (u_{ht} - \bar{u}_h)$$

3.2 *B.R.R. Variance Estimator*

Mccarthy [2] [3] proposed the BRR method when $n_h = 2$ for all h based on a number of half sample formed by deleting one PSU from the sample in each stratum. The set of R-balanced half sample used may be defined by an R xL design matrix $(\delta_h^{(r)})$, $1 \le r \le R$ and $1 \le h \le L$, where δ_h^r is $+1$ or -1 depending upon the r-th half-sample and $\sum_r \delta_h^r = \theta$ for all $h \neq h'$

Define

$$z_{ih}^{(r)} = z_{ih1} \quad \text{if } \delta_h^r = 1$$

$$= z_{ih2} \quad \text{if } \delta_h^r = -1$$

Similarly,

$$u_h^{(r)} = u_{hi} \quad \text{if } \delta_h^r = 1$$

$$= u_{h2} \quad \text{if } \delta_h^r = -1$$

$$z_{ih}^{(r)} = \bar{z}_{ih} + \delta_h^r \, (\Delta z_{ih}) \qquad \ldots (5)$$

where $\Delta z_{ih} = (1/2)\,(z_{ih1} - z_{ih2})$

we can write

$$\bar{z}_i^{(r)} = \sum_{h=1}^{L} W_h \, z_{ih}^{(r)} = \bar{z}_i + \sum_{h=1}^{L} W_h \, \delta_h^r \, \Delta z_{ih}$$

Similarly we can find out

$$\bar{u}^{(r)} = \bar{u} + \sum_{h=1}^{L} W_h \, \delta_h^r \, \Delta u_h$$

The BRR estimator of θ is given as

$$\hat{\theta}_B = (1/R) \sum_{r=1}^{R} \hat{\theta}^{(r)}$$

where
$$\hat{\theta}^{(r)} = \frac{\bar{z}_i^{(r)}}{\bar{u}^{(r)}} \qquad \ldots (6)$$

The BRR variance estimator is given by

$$v_B(\hat{\theta}) = \frac{1}{R} \sum_{r=1}^{R} (\hat{\theta}^{(r)} - \hat{\theta})^2 \qquad \ldots (7)$$

By expanding $\hat{\theta}^{(r)}$ and with the help of equation (5) we can write BRR variance estimator as

$$v_B(\hat{\theta}) = v_L(\hat{\theta}) + A + B + C + O_p(n^{-2.5}) \qquad \ldots (8)$$

where
$$A = \frac{1}{R} \sum_{r=1}^{R} \left[-\frac{2\bar{z}_i^2}{\bar{u}^6} (\Delta \bar{u}^{(r)})^3 + \frac{2\bar{z}_i}{\bar{u}^5} (\Delta \bar{u}^{(r)})^2 (\Delta z_i^{(r)}) \right.$$
$$\left. + \frac{2\bar{z}_i}{\bar{u}^4} (\Delta \bar{u}^{(r)}) (\Delta z_i^{(r)})^2 - \frac{2}{\bar{u}^3} (\Delta \bar{u}^{(r)}) (\Delta z_i^{(r)})^2 \right]$$

$$B = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{\bar{z}_i^2}{\bar{u}^6}\,(\Delta\,\bar{u}^{(r)})^4 - \frac{\bar{z}_i}{\bar{u}^5}\,(\Delta\,\bar{u}^{(r)})^3\,(\Delta\,z_i^{(r)}) - \frac{\bar{z}_i}{\bar{u}^4}\,(\Delta\,\bar{u}^{(r)})^3\,(\Delta\,z_i^{(r)}) + \frac{1}{\bar{u}^4}\,(\Delta\,\bar{u}^{(r)})^2\,(\Delta\,z_i^{(r)})^2\right]$$

$$= O_p\,(n^{-2})$$

$$C = \frac{1}{4R}\sum_{r=1}^{R}\left[\frac{4\bar{z}_i^2}{\bar{u}^6}\,(\Delta\,\bar{u}^{(r)})^4 - \frac{8\bar{z}_i}{\bar{u}^5}\,(\Delta\,\bar{u}^{(r)})^3\,(\Delta\,z_i^{(r)}) + \frac{4}{\bar{u}^4}\,(\Delta\,\bar{u}^{(r)})^2\,(\Delta\,Z_i^{(r)})^2\right]$$

Now it can be proved that

$$E\,(A) = O_p\,(n^{-3}) \qquad \ldots (9)$$

$$E\left[(1/R)\sum_{r=1}^{R}\,(\Delta\,\bar{u}^{(r)})^4\right] = 3\left[\sum_{h=1}^{L}\frac{W_h}{2}\,s_{hu}^2\right]^2 + O_p\,(n^{-3})$$

$$= D_{uuuu} + O_p\,(n^{-3})$$

$$E\left[(1/R)\sum_{r=1}^{R}\,(\Delta\,\bar{u}^{(r)})^3\,(\Delta\,z_i^{(r)})\right] = D_{uuuz_i} + O_p\,(n^{-3})$$

$$E\left[(1/R)\sum_{r=1}^{R}\,(\Delta\,\bar{u}^{(r)})^2\,(\Delta\,z_i^{(r)})^2\right] = D_{uuz_iz_i} + O_p\,(n^{-3})$$

$$E\left[(1/R)\sum_{r=1}^{R}\,(\Delta\,\bar{u}^{(r)})\,(\Delta\,z_i^{(r)})^3\right] = D_{uz_iz_iz_i} + O_p\,(n^{-3}) \qquad \ldots (10)$$

3.3 *Bias of Variance Estimators*

The bias of BRR variance estimators was found for $n_h = 2$. From equation (8) we have

$$v_{BRR} = v_L\,(\hat{\theta}) - \frac{2}{R}\sum_{r=1}^{R}(\Delta\bar{e}^{(r)})^2\,(\Delta\bar{u}^{(r)}) + \frac{3}{R}\sum_{r=1}^{R}\,(\Delta\bar{e}^{(r)})^2\,(\Delta\bar{u}^{(r)})^2 + O_p\,(n^{-2.5})$$

(11)

where,

$$\Delta\bar{e}^{(r)} = \sum_{h=1}^{L} W_h \; \delta_h^{(r)} \, \Delta e_h$$

$$\Delta\bar{u}^{(r)} = \sum_{h=1}^{L} W_h \; \delta_h^{(r)} \, \Delta u_h$$

$$\Delta e_h = e_{h1} - e_{h2}$$

$$\Delta u_h = u_{h1} - u_{h2}$$

Let

$$a = \sum_{h=1}^{L} \frac{W_h^2}{n_h^2} S^{2}_{e^{2}_{uh}}$$

$$b = \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S_{xh}^2 \right] \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} \right] \geq 0$$

$$c = \left[\sum_{h=1}^{L} \frac{W_h^2}{n_h} S_{euh}^2 \right]^2 \geq 0$$

where $S_{uh}^2 = E\,(U_{ht} - \bar{U}_h)^2$

$$S_{eh}^2 = E\,(e_{ht} - \bar{E}_h)^2$$

$$S_{euh}^2 = E\,(u_{ht} - \bar{U}_h)\,(e_{ht} - \bar{E}_h)$$

$$S^{2}_{e^{2}_{uh}} = E\,(U_{ht} - \bar{U}_h)\,(e_{ht} - \bar{E}_h)^2$$

$$\bar{E}_h = E\,(e_{ht}) = 0$$

It can be proved that

$$E\left[\frac{1}{R} \sum_{r=1}^{R} (\Delta\,\bar{e}^{(r)})^2 (\Delta\,\bar{u}^{(r)})^2 \right] = b - 2c + O_p\,(n^{-3}). \qquad \ldots (12)$$

From equation (3) we get

$$B\,[v_L\,(\hat{\theta})] = -2a + b + 0\,(n^{-3}) \qquad \ldots (13)$$

Similarly we get

$$B\,[v_B] = -2a + 4b + 6c + 0\,(n^{-3}) \qquad \ldots (14)$$

So from equation (13) and (14) we get

$$B\,[v_B] > B\,[v_L]$$

Now an attempt will be made to find the bias of jackknife variance estimator. From (4) and (13) we have

$$B\,[v_j\,(\hat{\theta})] = B\,\,[v_L\,(\hat{\theta})] + 2\sum_{h=1}^{L} \frac{W_h^2}{n_h^2}\,\frac{n_h - 2}{n_h - 1}\,S^2_{e_{uh}} + 0\,(n^{-3})$$

$$\text{Since, } E\,(s^2_{e_{uh}}) = \frac{n_h - 2}{n_h - 1}\,S^2_{e_{uh}} \quad \text{for } n_h \geq 2$$

so

$$B\,[v_j\,(\hat{\theta})] = B\,[v_L\,(\hat{\theta})] + 2a - 2a' + 0\,(n^{-3})$$

$$= b - 2a' + 0\,(n^{-3})$$

where

$$a' = \sum_{h=1}^{L} \frac{W_h^2}{n_h^2}\,\frac{1}{n_h - 1}\,S^2_{e_{uh}}$$

Further, it was shown that Balanced Repeated Replication estimator is considerably more biased than jackknifing estimator.

4. *Conclusions*

On the basis of this asymptotic method of comparison one can draw the following conclusions ;

(1) For the common two PSU's per stratum design $v_L\,(\hat{\theta})$ and $v_j\,(\hat{\theta})$ are asymptotically equal to higher order terms. This result suggests that choice between $v_L\,(\hat{\theta})$ and $v_j\,(\hat{\theta})$ should depend more on statistical consideration such as computational ease and cost of data collection etc.

(2) BRR variance estimator $v_B\,(\hat{\theta})$ are greater than both linearized variance estimator $v_L\,(\hat{\theta})$ as well as jackknifed variance estimator $v_j\,(\hat{\theta})$.

(3) Bias of the BRR variance estimator $B\,(v_B)$ and bias of jackknifed variance estimator $B\,(v_j)$ are greater than bias of linearized variance estimator $B\,(v_L)$.

(4) Although, theoritically Taylor linearized variance estimator is

considerably better than other two variance estimation techniques in terms of its magnitude and bias of variance estimators but in practical situations it may be inconsistent, because of its localized properties. So, for a practical situations jackknifed and BRR variance estimation techniques are more suitable. If in a given situation BRR variance estimation technique is applicable, it may work efficiently because of its consistent nature as compared to other two variance estimation techniques described above.

REFERENCES

[1] Krewski, D. and Rao, J.N.K., 1981. Inference from stratified samples : Properties of the linearization, jackknife and balanced repeated replication method. *Ann. Statist.*, **6**, No. 5, 1010-1019.

[2] McCarthy, P.J., 1966. Replication : An approach to the analysis of data from complex surveys. Vital and Health Statistics, Ser-2 No. 14 Washington D.C., U.S. Government Printing Office.

[3] McCarthy, P. J., 1969. Pseudoreplication : Half samples. *Rev. Int. Statist. Inst.*, **37**, 239-264.

[4] Rao, J.N.K. and Wu, C.F.J., 1985. Inference from stratified samples : Second order analysis of three methods for non-linear statistics. *J. Am. Statist. Assoc.*, **80**, 620-630.