

COMPARISON OF RATIO ESTIMATORS BASED ON INTERPENETRATING SUBSAMPLES WITH OR WITHOUT JACKKNIFING

SUBIR GHOSH

and

ROBERTO GOMEZ

Department of Statistics, University of California, Riverside

(Received : February, 1985)

SUMMARY

A comparison of four ratio estimators based on interpenetrating subsamples and with or without jackknifing is done both theoretically and empirically with respect to bias, variance and mean square error.

Keywords : Bias, Interpenetrating subsampling, Jackknife, Mean Square error, Ratio estimation, Simple random sample, Variance.

Introduction

The method of interpenetrating subsamples (IPS) in large scale sampling as introduced in Mahalanobis [4], [5], is to draw a sample in the form of two or more subsamples under the same probability sampling design so that each sub-sample provides a valid estimate of the parameter of interest. The purpose of IPS is to assess both sampling and nonsampling errors in the estimation of the parameter of interest. The United Nations Subcommittee on statistical sampling [11] has recommended the use of IPS with a suggestion of an alternative name "Replicated Sampling." The method of ratio estimation is common in large scale sample surveys in estimating various ratios and the ratio estimator is biased. The Quenouille-Tukey jackknife (see Miller [7]) gives nonparametric estimators of bias and variance.

Consider a population of N units with y as the variable of interest and

x as an auxiliary variable. Denote the population totals of the variables x and y over N population units by X and Y . The population ratio $R = (Y/X)$ is an unknown parameter of interest. We shall draw inference on R on the basis of k interpenetrating subsamples of size m each, (x_{ij}, y_{ij}) $i = 1, \dots, m, j = 1, \dots, k$. Consider four interpenetrating estimators of R for any probability sampling design, two of which are jackknife interpenetrating subsample estimators (JIPS) and the other two are just interpenetrating subsample (IPS) estimators. Compare these four estimators theoretically and also empirically with respect to bias, variance and mean square error (MSE). This comparison is done in Sections 3 and 4 when the sampling design is simple random sample with replacement.

Let \hat{Y}_j and \hat{X}_j be unbiased estimators of Y and X from the j th interpenetrating subsample ($j = 1, \dots, k$). Denote $\hat{Y} = \left(\sum_{j=1}^k \hat{Y}_j/k \right)$ and $\hat{X} = \left(\sum_{j=1}^k \hat{X}_j/k \right)$. Two IPS estimators of R are given by

$$\hat{R}_1 = \frac{\hat{Y}}{\hat{X}} \text{ and } \hat{R}_2 = \frac{1}{k} \sum_{j=1}^k \frac{\hat{Y}_j}{\hat{X}_j}. \quad (1)$$

Denote $\hat{Y}_{(u)} = \left[\sum_{j=1, j \neq u}^k \hat{Y}_j/(k-1) \right]$, $\hat{X}_{(u)} = \left[\sum_{j=1, j \neq u}^k \hat{X}_j/(k-1) \right]$,

$\hat{R}_{1(u)} = (\hat{Y}_{(u)}/\hat{X}_{(u)})$ and $\hat{R}_{1(\cdot)} = \left(\sum_{u=1}^k \hat{R}_{1(u)}/k \right)$. The first JIPS estimator is

then

$$\hat{R}_3 = k\hat{R}_1 - (k-1)\hat{R}_{1(\cdot)}. \quad (2)$$

Let $\hat{Y}_{j(v)}$ and $\hat{X}_{j(v)}$ be unbiased estimators of Y and X from the j th interpenetrating subsample eliminating the v th unit. Denote $r_j = \hat{Y}_j/\hat{X}_j$,

$r_{j(v)} = \hat{Y}_{j(v)}/\hat{X}_{j(v)}$, $r_{j(\cdot)} = \frac{1}{m} \sum_{v=1}^m r_{j(v)}$, $\hat{R}_{2(v)} = \frac{1}{k} \sum_{j=1}^k r_{j(v)}$ and

$\hat{R}_{2(\cdot)} = \frac{1}{m} \sum_{v=1}^m \hat{R}_{2(v)}$. The second JIPS estimator is then

$$\hat{R}_4 = \frac{1}{k} \sum_{j=1}^k [mr_j - (m-1)r_{j(\cdot)}] = m\hat{R}_2 - (m-1)\hat{R}_{2(\cdot)}. \quad (3)$$

Note that for $k = 2$, $\hat{R}_2 = \hat{R}_{1(\cdot)}$, and hence from (2),

$$\hat{R}_1 = (\hat{R}_2 + \hat{R}_3)/2. \quad (4)$$

For $k = 2$, from theoretical comparison, \hat{R}_1 and \hat{R}_3 are equally good over \hat{R}_2 and \hat{R}_4 . For $k > 2$, both theoretical and empirical comparisons suggest the superiority of \hat{R}_1 .

The purpose of this study is to investigate whether jackknifing has any positive impact on ratio estimation based on interpenetrating subsamples, in terms of reducing bias and mean square error.

2. Bias, Variance and MSE

Denote the bias of \hat{R}_i to the second degree approximation by B_i ($i = 1, 2, 3, 4$), the variance of \hat{R}_i to the second degree approximation by V_i and

$$B_3^{(1)} = \frac{1}{k} \sum_{u=1}^k \{R \text{Var}(\hat{X}_{(u)}) - \text{Cov}(\hat{X}_{(u)}, \hat{Y}_{(u)})\}, \quad (5)$$

$$B_3^{(2)} = R \text{Var}(\hat{X}) - \text{Cov}(\hat{X}, \hat{Y}),$$

$$B_4^{(1)} = \frac{1}{m} \sum_{j=1}^k \sum_{v=1}^m \{R \text{Var}(\hat{X}_{j(v)}) - \text{Cov}(\hat{X}_{j(v)}, \hat{Y}_{j(v)})\},$$

$$B_4^{(2)} = \sum_{j=1}^k \{R \text{Var}(\hat{X}_j) - \text{Cov}(\hat{X}_j, \hat{Y}_j)\}.$$

It can be seen in Murthy [8] that

$$B_1 = \frac{B_4^{(2)}}{k^2 X^2} \text{ and } B_2 = kB_1, \quad (6)$$

The estimators of B_i , $i = 1, 2, 3, 4$, are as follows

$$\hat{B}_1 = \frac{1}{k-1} (\hat{R}_2 - \hat{R}_1), \quad (7)$$

$$\hat{B}_2 = k\hat{B}_1,$$

$$\hat{B}_3 = (k-1) (\hat{R}_{1(\cdot)} - \hat{R}_1),$$

$$\hat{B}_4 = (m-1) (\hat{R}_{2(\cdot)} - \hat{R}_2).$$

Take the expectations of \hat{B}_3 and \hat{B}_4 assuming $\hat{R}_{1(\cdot)}$ and \hat{R}_i , $i = 1, 2$, as

estimators of R .

$$E(\hat{B}_3) = \frac{k-1}{X^2} [B_3^{(1)} - B_3^{(2)}], \quad (8)$$

$$E(\hat{B}_4) = \frac{m-1}{kX^2} [B_4^{(1)} - B_4^{(2)}].$$

Assume

$$\text{Cov}(\hat{X}_j, \hat{X}_{j'}) = \text{Cov}(\hat{Y}_j, \hat{Y}_{j'}) = \text{Cov}(\hat{X}_j, \hat{Y}_{j'}) = 0 \text{ for } j \neq j'. \quad (9)$$

The assumption (9) means that the estimators based on two different subsamples are uncorrelated.

LEMMA 1. Under (9),

$$E(\hat{B}_3) = B_1. \quad (10)$$

Proof. The proof is clear by observing

$$k(k-1) B_3^{(1)} = k^2 B_3^{(2)} = B_4^{(2)}. \quad (11)$$

Note that \hat{B}_3 may or may not be equal to \hat{B}_1 . It can however be checked that, for $k=2$, $\hat{B}_3 = \hat{B}_1 = \frac{1}{2} \hat{B}_2$.

The estimators of V_3 and V_4 are

$$\hat{V}_3 = \frac{k-1}{k} \sum_{u=1}^k (\hat{R}_{1(u)} - \hat{R}_{1(\cdot)})^2, \quad (12)$$

$$\hat{V}_4 = \frac{m-1}{mk^2} \sum_{j=1}^k \sum_{v=1}^m (r_{j(v)} - r_{j(\cdot)})^2.$$

The mean square error in estimating R by \hat{R}_i is denoted by MSE_i . We know

$$\text{MSE}_i = V_i + B_i^2, \quad i = 1, 2, 3, 4. \quad (13)$$

Consider MSE_i here to the second degree approximation. It can be seen in Murthy [8] that

$$\text{MSE}_1 = \text{MSE}_2 = \frac{1}{k^2 X^2} \sum_{j=1}^k \{ \text{Var}(\hat{V}_j) - 2R \text{Cov}(\hat{Y}_j, \hat{X}_j) + R^2 \text{Var}(\hat{X}_j) \}. \quad (14)$$

The estimators of MSE_i , denoted by \hat{MSE}_i , $i = 1, 2, 3, 4$, can be compared under any probability sampling design. Consider, however the simple random sampling with replacement as the sampling design in subsequent sections.

3. Simple Random Sampling

Consider the problem of comparison of \hat{R}_i when the sampling design is simple random sampling with replacement. We have $\hat{X}_j = N \left(\sum_{i=1}^m X_{ij}/m \right)$,

$$Y_j = N \left(\sum_{i=1}^m y_{ij}/m \right), X_{j(v)} = N \left(\sum_{i=1, i \neq v}^m x_{ij}/(m-1) \right) \text{ and}$$

$$\hat{Y}_{j(v)} = N \left(\sum_{i=1, i \neq v}^m y^{ij}/(m-1) \right). \text{ It can be checked that}$$

$$\sum_{v=1}^m \hat{X}_{j(v)} = m \hat{X}_j, \sum_{v=1}^m \hat{Y}_{j(v)} = m \hat{Y}_j, \quad (15)$$

$$m(k-1) \hat{X}_{(u)} = \sum_{j=1}^k \sum_{\substack{v=1 \\ j \neq u}}^m \hat{X}_{j(v)},$$

$$m(k-1) \hat{Y}_{(u)} = \sum_{j=1}^k \sum_{\substack{v=1 \\ j \neq u}}^m \hat{Y}_{j(v)}.$$

3.1. Comparison of $E(\hat{B}_3)$ and $E(\hat{B}_4)$ to the Second Order Approximation

Denote

$$C = \sum_{j=1}^k \sum_{\substack{v=1 \\ v \neq v'}}^m \sum_{v'=1}^m [R \text{Cov}(\hat{X}_{j(v)}, \hat{X}_{j(v')}) - \text{Cov}(\hat{X}_{j(v)}, \hat{Y}_{j(v')})]. \quad (16)$$

THEOREM 1. Under (9)

$$k^2 X^2 m(m-1) E(\hat{B}_3) = k X^2 \frac{m}{m-1} E(\hat{B}_4) + C. \quad (17)$$

Proof. It can be seen from (5) and (15) that

$$m^2 B_4^{(2)} = m B_4^{(1)} + C. \quad (18)$$

It is now easy to check from (6), (8) and (18) that (17) is true. This completes the proof.

Now consider the intra-class model

$$\begin{aligned} \text{Var}(X_{ij}) &= \sigma_x^2, \text{Var}(Y_{ij}) = \sigma_y^2, \\ \text{Cov}(X_{ij}, Y_{ij}) &= \rho_{xy}\sigma_x\sigma_y, \text{Cov}(X_{ij}, X_{i'j}) = \rho_{xx}\sigma_x^2 \\ \text{Cov}(Y_{ij}, Y_{i'j}) &= \rho_{yy}\sigma_y^2, \text{Cov}(X_{ij}, Y_{i'j}) = \rho'_{xy}\sigma_x\sigma_y. \end{aligned} \quad (19)$$

Denote

$$\begin{aligned} C_1 &= R\sigma_x^2\rho_{xx} - \sigma_x\sigma_y\rho'_{xy}, \\ C_2 &= (R\sigma_x^2 - \rho_{xy}\sigma_x\sigma_y) - C_1. \end{aligned} \quad (20)$$

THEOREM 2. Under (9) and (19),

$$kE(\hat{B}_3) = E(\hat{B}_4) + \frac{N^2}{X^2} C_1. \quad (21)$$

Proof. We get from (8) and (16)

$$\begin{aligned} (m-1)C &= N^2km(m-2)C_2 + N^2km(m-1)^2C_1, \\ mX^2E(\hat{B}_4) &= N^2C_2, \\ mX^2kE(\hat{B}_3) &= N^2[C_2 + mC_1], \end{aligned} \quad (22)$$

This completes the proof.

COROLLARY 1. If $\rho_{xx} = \rho'_{xy} = 0$ (i.e., $C_1 = 0$), then under (9) and (19), we have

$$kE(\hat{B}_3) = E(\hat{B}_4). \quad (23)$$

3.2. Comparison of $E[\hat{V}_3]$ and $E[\hat{V}_4]$ to the First Order Approximation

First a proposition is stated which is very useful in subsequent calculations.

PROPOSITION. Let t_1, \dots, t_k be k random variables with $E(t_i)$, $\text{Var}(t_i)$ and $\text{Cov}(t_i, t_j)$, $i \neq j$, being constants independent of i and j which are equal to $E(t)$, $\text{Var}(t)$, and $\text{Cov}(t, t')$. Then

$$E\left[\sum_{i=1}^k (t_i - \bar{t})^2\right] = (k-1) [\text{Var}(t) - \text{Cov}(t, t')], \quad (24)$$

where $\bar{t} = (t_1 + \dots + t_k)/k$. It follows from (24) that

$$E[\hat{V}_3] = \frac{(k-1)^2}{k} [\text{Var}(\hat{R}_{1(u)}) - \text{Cov}(\hat{R}_{1(u)}, \hat{R}_{1(u')})] \quad (25)$$

$$E[\hat{V}_4] = \frac{(m-1)^2}{mk} \left[\text{Var}\left(\frac{\hat{Y}_{j(v)}}{\hat{X}_{j(v)}}\right) - \text{Cov}\left(\frac{\hat{Y}_{j(v)}}{\hat{X}_{j(v)}}, \frac{\hat{Y}_{j(v')}}{\hat{X}_{j(v')}}\right) \right]$$

where $E[\hat{V}_3]$ and $E[\hat{V}_4]$ are constants independent of (u, u') and (v, v') .

We find to the first order of approximation

$$\begin{aligned} \text{Cov}(\hat{R}_{1(u)}, \hat{R}_{1(u')}) &= \frac{1}{X^2} [\text{Cov}(\hat{Y}_{(u)}, \hat{Y}_{(u')}) - R \text{Cov}(\hat{X}_{(u)}, \hat{Y}_{(u')}) \\ &\quad - R \text{Cov}(\hat{Y}_{(u)}, \hat{X}_{(u')}) + R^2 \text{Cov}(\hat{X}_{(u)}, \hat{X}_{(u')})]. \end{aligned} \quad (26)$$

Considering the model (19) with $\rho_{xx} = \rho_{yy} = \rho'_{xy} = 0$ and denoting $\sigma^2 = \sigma_y^2 - 2R\rho_{xy}\sigma_x\sigma_y + R^2\sigma_x^2$, we get to the first order approximation

$$\begin{aligned} \text{Var}(\hat{R}_{1(u)}) &= \frac{N^2 \sigma^2}{(k-1)mX^2} \quad (27) \\ \text{Cov}(\hat{R}_{1(u)}, \hat{R}_{1(u')}) &= \frac{(k-2)N^2 \sigma^2}{(k-1)^2 m X^2} \\ \text{Var}(r_{j(v)}) &= \frac{N^2 \sigma^2}{X^2(m-1)} \\ \text{Cov}(r_{j(v)}, r_{j(v')}) &= \frac{N^2(m-2)\sigma^2}{X^2(m-1)^2}. \end{aligned}$$

The following result is now easy to verify.

THEOREM 3. *Under (19) with $\rho_{xx} = \rho_{yy} = \rho'_{xy} = 0$, the sampling design being simple random sampling with replacement, and to the first order of approximation, we have*

$$E(\hat{V}_3) = E(\hat{V}_4) = V_1 = V_2. \quad (28)$$

3.3. Conclusion

For $k = 2$, both \hat{R}_1 and \hat{R}_3 are equally good over \hat{R}_2 and \hat{R}_4 . For $k > 2$, \hat{R}_1 and \hat{R}_3 are better than \hat{R}_2 and \hat{R}_4 with respect to bias but \hat{R}_3 is not as good as \hat{R}_1 , \hat{R}_2 and \hat{R}_4 with respect to first order approximation of variance.

4. Empirical Study (to the Second Order Approximation)

The population consists of the countries in 5 states in Mexico, namely Chiapas, Chihuahua, Guerrero, Puebla and Veracruz. Consider three different studies on the same population.

- I : Y = Total number of inhabitants
 X = Total number of households
 II : Y = Total number of literates
 X = Total number of illiterates

III : Y = Total number of persons in primary activities
 X = Total number of economically active people

The data are obtained from the 1960 population general census in Mexico. The five states considered are very similar with respect to $R = (Y/X)$ in Studies I, II and III. Note that in III the data in 2 counties are unavailable. Now draw 1000 times five subsamples of size 30 each by simple random sampling with replacement. In the following table we present the average values of \hat{R}_i , \hat{B}_i , \hat{V}_i , $M\hat{S}E_i$ for three studies.

In studies I, II and III, we find the average estimated bias in \hat{R}_4 is about 5, 4 and 3 times than that in \hat{R}_3 . It is clear that, with respect to bias, \hat{R}_1 and \hat{R}_3 are better than \hat{R}_2 and \hat{R}_4 . In comparison with respect to variance, \hat{R}_2 is superior to \hat{R}_1 , \hat{R}_3 and \hat{R}_4 . In studies I and III, \hat{R}_1 and \hat{R}_2 have smaller $M\hat{S}E$. In study II, \hat{R}_3 has the largest $M\hat{S}E$. In all studies \hat{R}_1 has the smallest $|\hat{B}| / (\hat{V})^{1/2}$.

TABLE A.1—THE VALUES OF N AND R

	I	II	III
N	672	672	670
R	5.39	1.13	0.67

TABLE A.2*—AVERAGE ESTIMATED RATIO, BIAS, VARIANCE AND MSE IN STUDY I

Average	\hat{R}_1	\hat{R}_2	\hat{R}_3	\hat{R}_4
\hat{R}	53914	53954	53905	53905
$(\hat{R} - R)^2$	68.562	78.967	68.321	73.849
\hat{B}	9.809	49.046	9.136	48.948
\hat{V}	90.726	82.200	94.744	111.590
$M\hat{S}E$	91.081	91.081	95.194	113.384
$ \hat{B} / \sqrt{\hat{V}}$	460	3377	535	1136

TABLE A.3*—AVERAGE ESTIMATED RATIO, BIAS, VARIANCE AND MSE IN STUDY II

<i>Average</i>	\hat{R}_1	\hat{R}_2	\hat{R}_3	\hat{R}_4
\hat{R}	11276	11076	11334	11318
$\hat{(R - R)^2}$	408.07	356.02	429.67	436.65
\hat{B}	-50.35	-251.75	-57.30	-242.90
\hat{V}	401.02	375.26	432.59	377.50
\hat{MSE}	402.10	402.10	434.07	389.82
$ \hat{B} /\sqrt{\hat{V}}$	347	1846	379	1150

TABLE A.4*—AVERAGE ESTIMATED RATIO, BIAS, VARIANCE AND MSE IN STUDY III

<i>Average</i>	\hat{R}_1	\hat{R}_2	\hat{R}_3	\hat{R}_4
\hat{R}	6752	7015	6668	6757
$\hat{(R - R)^2}$	41.200	37.867	47.889	46.944
\hat{B}	65.710	328.550	84.867	258.730
\hat{V}	40.464	22.220	52.783	33.782
\hat{MSE}	41.224	41.224	54.350	44.330
$ \hat{B} /\sqrt{\hat{V}}$	977	5281	1051	4188

*The entries in Tables A2 - A4 are multiplied by 10^4 .

5. Miscellaneous

In this section we present an additional JIPS estimator of R and observe that it is in fact identical to an IPS estimator of R , or in other words, the proposed jack knifing does not have any effect on this IPS estimator.

Let r_1, \dots, r_k be k estimators of R on the basis of k interpenetrating subsamples of size m each and under the same probability sampling design. Then $\bar{r} = (r_1 + \dots + r_k)/k$ is an IPS estimator of R . Denote $\bar{r}_{(u)} = (r_1 + \dots + r_{u-1} + r_{u+1} + \dots + r_k)/(k-1)$, $u = 1, \dots, k$ and $\bar{r}_{(\cdot)} = (\bar{r}_{(1)} + \dots + \bar{r}_{(k)})/k$. Then $\hat{R}_5 = k\bar{r} - (k-1)\bar{r}_{(\cdot)}$ is another JIPS estimator of R , different from \hat{R}_3 and \hat{R}_4 . Clearly, $\bar{r}_{(\cdot)} = \bar{r} = \hat{R}_5$, or, in other words, the IPS estimator is identical with the JIPS estimator

\hat{R}_5 . Now present an observation on unbiased estimator of $\text{Var}(\bar{r})$. It is clear that

$$\begin{aligned} \text{Var}(\hat{R}_5) &= \frac{k-1}{k} \sum_{u=1}^k (\bar{r}_{(u)} - \bar{r}_{(\cdot)})^2 = \frac{1}{k(k-1)} \sum_{u=1}^k (r_u - \bar{r})^2 \\ &= \widehat{\text{Var}}(\bar{r}). \end{aligned} \quad (29)$$

Denote

$$\overline{\text{Cov}} = \frac{\sum_{u=1}^k \sum_{u' \neq u}^k \text{Cov}(r_u, r_{u'})}{k(k-1)} \quad (30)$$

= the average of $k(k-1)$ covariances,

$\text{Cov}(r_u, r_{u'})$, $u \neq u'$, u, u' are in $\{1, \dots, k\}$.

Under the assumption $E(r_1) = \dots = E(r_k)$, it follows that

$$E[\widehat{\text{Var}}(\bar{r})] = \text{Var}(\bar{r}) - \overline{\text{Cov}}. \quad (31)$$

If $\overline{\text{Cov}} = 0$, then $\widehat{\text{Var}}(\bar{r})$ is an unbiased estimator of $\text{Var}(\bar{r})$. If $\overline{\text{Cov}} > 0$ (< 0), then $\widehat{\text{Var}}(\bar{r})$ underestimates (overestimates) $\text{Var}(\bar{r})$.

ACKNOWLEDGEMENTS

The authors would like to thank Professors Jim Press (UC, Riverside), Rupert Miller and Lincoln Moses (Stanford University) for their various help during the preparation and writing of the manuscript, and to Karola Declve for typing the paper with great care.

REFERENCES

- [1] Cochran, W. G. (1977) : *Sampling Techniques*, Third Edition, John Wiley & Sons, New York.
- [2] Efron, B. and Stein, C. (1981) : The Jackknife estimate of variance, *The Annals of Statistics*, 9 : 586-596.
- [3] Krewski, D. and Chakrabarty, R. P. (1981) : On the stability of the Jackknife variance estimator in ratio estimation, *Journal of Statistical Planning and Inference*, 5 : 71-78.
- [4] Mahalanobis, P. C. (1944) : On large-scale sample surveys, *Phil. Trans. Roy. Soc. London*, B231 : 329-451.
- [5] Mahalanobis, P. C. (1946) : Recent experiments in statistical sampling in the Indian Statistical Institute, *Jour. Roy. Stat. Soc.*, 109 : 325-370.
- [6] Mexico Secretaria de Industria Y Comercio. Direccion General de Estadistica. "VIII Censo General de Poblacion 1960". Mexico, D. F. (1963).
- [7] Miller, R. G. (1974) : The Jackknife—a review, *Biometrika*, 61 : 1-15.

- [8] Murthy, M. N. (1967) : *Sampling Theory and Methods*. Statistical Publishing Society, Calcutta, India.
- [9] Rao, P. S. R. S. and Rao, J. N. K. (1971) : Small sample results for ratio estimators, *Biometrika*, **58** : 625-630.
- [10] Royall, R. M. and Cumberland, W. G. (1981) : An empirical study of the ratio estimator and estimator of its variance, *Journal of the American Statistical Association*, **76** : 66-77.
- [11] United Nations (1949) : Preparation of sample survey reports, C(1), New York : Statistical Office of the United Nations.