Statistics & Probability Letters 18 (1993) 345–348
North-Holland

2 December 1993

On the absolute bias ratio of ratio estimators

Xiao-Li Meng

Department of Statistics, University of Chicago, IL, USA

Received January 1993
Revised March 1993

Abstract: The elegant Hartley–Ross inequality on the absolute bias ratio (ABR ≡ |Bias|/S.E.) of a ordinary ratio estimator is here generalized to that of a separate ratio estimator with stratified sampling. It is shown that, as long as the numerators and denominators used to form strata ratios are unbiased estimators, the absolute bias ratio of a separate ratio estimator will never exceed the square root of the sum of squares of the coefficient of variation of the denominators across strata. This provides, at design stages, a simple bound in practice to assess the limit and magnitude of the bias ratio of any separate ratio estimator that shares the same denominators. Exact expressions for biases of separate ratio estimators are also given.

Keywords: Combined ratio estimator; separate ratio estimator; stratified sampling.

1. Biases of ordinary ratio estimators

In sample surveys, ordinary ratio estimators are typically employed to estimate (i) a population total, Y, (ii) a population mean, $\bar{Y}$, or (iii) a population ratio, Y/X. In all of these cases, the ratio estimator has the form

$$r = \frac{\bar{y}}{\bar{x}} Q, \tag{1.1}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of variable x and y, respectively, and Q is a known quantity. In cases (i) and (ii), Q is the population total and mean of variable x, X and $\bar{X}$ respectively, and the ratio estimator r of (1.1) is used to increase the precision in estimating Y and $\bar{Y}$ by taking advantage of the positive correlation between y and x and the known values of X and $\bar{X}$ in the population. In case (iii), $Q = 1$, and the population quantities of variable x need not be known. A comprehensive treatment of ratio estimator (1.1) and other variations can be found in Cochran (1977, Chapter 6).

It is well known that in general, r of (1.1) is biased for $R = (\bar{Y}/\bar{X})Q$. However, this bias is typically unimportant because it is negligible compared to the standard error of r. An elementary but elegant proof of this fact was given in Hartley and Ross (1954), who noticed the following simple identity

$$E(r) - \frac{E(\bar{y})}{E(\bar{x})} Q = -\frac{\mathrm{Cov}(r, \bar{x})}{E(\bar{x})}. \tag{1.2}$$

Thus, if

$$\frac{E(\bar{y})}{E(\bar{x})} = \frac{\bar{Y}}{\bar{X}}, \tag{1.3}$$

Correspondence to: Xiao-Li Meng, Department of Statistics, University of Chicago, 5734, University Av., Chicago, IL 60637, USA.

then the right side of (1.2) gives the exact bias of r. Since $|\mathrm{Cov}(r, \bar{x})| \leqslant \sigma(r)\sigma(\bar{x})$, it then follows immediately that

$$\mathrm{ABR}(r) \equiv \frac{|\mathrm{Bias}(r)|}{\sigma(r)} \leqslant \frac{\sigma(\bar{x})}{E(\bar{x})} = \mathrm{CV}(\bar{x}). \quad (1.4)$$

Since $\mathrm{CV}(\bar{x})$, the coefficient of variation of $\bar{x}$, is typically small in practice (i.e., < 0.2), even with moderate sample size, inequality (1.4) ensures that the bias of r is typically not of practical concern.

The importance of the inequality (1.4) is that it enables us, at the design stage, to assess the limit and magnitude of the bias ratio of r of (1.1) for any sampling variable, y, since the bound in (1.4), $\mathrm{CV}(\bar{x})$, does not depend on y and typically can be calculated from the design and the known population quantities of the auxiliary variable, x. Inequality (1.4) is also very general in the sense that it still holds with $\mathrm{CV}(\bar{x})$ being replaced by the coefficient of variation of the appropriate denominator if $\bar{x}$ and $\bar{y}$ in r are replaced by other unbiased estimators (e.g., $\bar{x}^{(c)}$ and $\bar{y}^{(c)}$ in the next section) for $\bar{X}$ and $\bar{Y}$, respectively. Thus (1.4) can be applied virtually in all practical situations where ordinary ratio estimators are used.

2. Biases of stratified ratio estimators

With stratified sampling, there are commonly two ways to construct a ratio estimator. One obvious way is the *combined* ratio method, which replaces the simple sample means $\bar{y}$ and $\bar{x}$ in the ordinary ratio estimator by the corresponding weighted strata sample means, respectively,

$$r_c = \frac{\sum_{h=1}^{L}(N_h/N)\bar{y}_h}{\sum_{h=1}^{L}(N_h/N)\bar{x}_h} Q \equiv \frac{\bar{y}^{(c)}}{\bar{x}^{(c)}} Q, \quad (2.1)$$

where $\bar{x}_h$, $\bar{y}_h$ are sample strata means of the hth stratum, N_h is the size of the hth stratum, N is the population size, and L is the number of strata. In contrast, the *separate* ratio method first forms an ordinary ratio estimator within each stratum and then appropriately weights them to form a single ratio estimator,

$$r_s = \sum_{h=1}^{L} \frac{\bar{y}_h}{\bar{x}_h} Q_h \equiv \sum_{h=1}^{L} r^{(h)}, \quad (2.2)$$

where Q_h $(h = 1, \ldots, L)$ are known quantities. For example, $Q_h = X_h$ $(h = 1, \ldots, L)$, the strata total of x, if r_s is used to estimate the population total of y.

The differences between r_c and r_s have been well studied in the literature (e.g., Cochran, 1977, Chapter 6). As a simple summary, the separate ratio estimator, r_s, typically has a smaller variance but a larger bias than the combined ratio estimator, r_c. The estimator r_s has a smaller variance because its precision is determined by whether the relationship between y and x is a straight line through the origin *within each* stratum, whereas for r_c, the precision depends on a much stronger relationship — whether there is such a *common* straight line *for all* strata. But the bias of r_c is typically negligible compared to its standard error. This can be seen by applying (1.4) to the r_c of (2.1) (Cochran, 1977, p. 166),

$$\mathrm{ABR}(r_c) \leqslant \mathrm{CV}(\bar{x}^{(c)}), \quad (2.3)$$

and typically $\mathrm{CV}(\bar{x}^{(c)})$ is small in practice. In fact, $\mathrm{CV}(\bar{x}^{(c)})$ is typically smaller than $\mathrm{CV}(\bar{x})$ because the stratification reduces the variance (e.g., with simple random sampling and proportional allocation). In contrast, even with large samples, the bias of r_s may not be negligible compared to its standard error if there are many strata and the sample sizes within some strata are small. This is because, as argued in Cochran (1977, p. 165), the bias in r_s is roughly L times that in $r^{(h)}$ (see (2.2)) if the bias has the same sign and similar size in all strata, but the standard error of r_s is only of the order of $L^{1/2}$ times that of $r^{(h)}$. Consequently, the absolute bias ratio of r_s is of the order $L^{1/2} \times \mathrm{ABR}(r^{(h)})$, which could be substantial if L is large and $\mathrm{ABR}(r^{(h)})$ is not too small. It is, therefore, of practical interest to have a simple and general method at the design stage, just as with ordinary ratio estimators, to assess the limit and magnitude of the absolute bias ratio of r_s. The following generalization of (1.4) provides an answer.

In common applications of r_s of (2.2), $\bar{y}_h$ and $\bar{x}_h$ are respectively unbiased estimators for $\overline{Y}_h$ and $\overline{X}_h$, the population strata means of the hth stratum ($h = 1, \ldots, L$). Thus, we can apply (1.2) to each $r^{(h)}$ in (2.2), which yields

$$\begin{aligned}\text{Bias}(r_s) &\equiv E(r_s) - \sum_{h=1}^{L} \frac{\overline{Y}_h}{\overline{X}_h} Q_h = \sum_{h=1}^{L} \text{Bias}(r^{(h)}) \\ &= \sum_{h=1}^{L} - \frac{\text{Cov}(r^{(h)}, \bar{x}_h)}{E(\bar{x}_h)}. \end{aligned} \tag{2.4}$$

It follows, by the Cauchy–Schwarz inequality, that

$$\begin{aligned}|\text{Bias}(r_s)| &\leqslant \sum_{h=1}^{L} \sigma(r^{(h)}) \cdot \text{CV}(\bar{x}_h) \\ &\leqslant \left[\sum_{h=1}^{L} \sigma^2(r^{(h)})\right]^{1/2} \left[\sum_{h=1}^{L} \text{CV}^2(\bar{x}_h)\right]^{1/2}. \end{aligned} \tag{2.5}$$

But, since samples from different strata are independent,

$$\sigma^2(r_s) = \sum_{h=1}^{L} \sigma^2(r^{(h)}). \tag{2.6}$$

Combining (2.5) and (2.6), we obtain

$$\text{ABR}(r_s) = \frac{|\text{Bias}(r_s)|}{\sigma(r_s)} \leqslant \left[\sum_{h=1}^{L} \text{CV}^2(\bar{x}_h)\right]^{1/2}, \tag{2.7}$$

which reduces to (1.4) when $L = 1$.

Although the derivation of (2.7) is almost trivial, we were not able to find such simple arguments in the generally accessible literature. Large-sample approximations to the bias in r_s and standard error of r_s that could lead to (2.7), of course, can be found in many textbooks (e.g., Hansen et al., 1953, Chapter 5). Notice that, if all $\text{CV}(\bar{x}_h)$'s are of similar sizes, then the right side of (2.7) is approximately $L^{1/2}\text{CV}(\bar{x}_h)$, which is the rough bound suggested by Cochran (1977, p. 165) for assessing the magnitude of $\text{ABR}(r_s)$ in practice. Like the Hartley–Ross inequality (1.4), (2.7) allows an assessment of the limit and magnitude of $\text{ABR}(r_s)$ at the design stage for any potential sampling variable y, because the right side of it only requires the knowledge of $\text{CV}(\bar{x}_h) = \sigma(\bar{x}_h)/\overline{X}_h$ ($h = 1, \ldots, L$), which is often available beforehand.

3. Two remarks

Remark 1. There is an interesting alternative expression for the bias in r_s. Rewriting r_s of (2.2) by finding the common denominator yields

$$r_s = \frac{\sum_{h=1}^{L} (\bar{y}_h \prod_{l \neq h} \bar{x}_l) Q_h}{\sum_{h=1}^{L} \bar{x}_h} \equiv \frac{a}{b}. \tag{3.1}$$

Since

$$\frac{E(a)}{E(b)} = \sum_{h=1}^{L} \frac{\overline{Y}_h}{\overline{X}_h} Q_h,$$

the population quantity to be estimated, by (1.3), we can apply (1.2) to (3.1) and obtain

$$\text{Bias}(r_s) = - \frac{\text{Cov}(r_s, \prod_{h=1}^{L} \bar{x}_h)}{E(\prod_{h=1}^{L} \bar{x}_h)}. \tag{3.2}$$

Interestingly, applying (1.4) directly to (3.2) would yield

$$\text{ABR}(r_s) \leqslant \text{CV}\left(\prod_{h=1}^{L} \bar{x}_h\right),$$

which, except for $L = 1$, is always less sharp than (2.7) because

$$\begin{aligned}\text{CV}\left(\prod_{h=1}^{L} \bar{x}_h\right) &= \left\{\prod_{h=1}^{L} [1 + \text{CV}^2(\bar{x}_h)] - 1\right\}^{1/2} \\ &> \left[\sum_{h=1}^{L} \text{CV}^2(\bar{x}_h)\right]^{1/2}. \end{aligned}$$

Remark 2. The methods discussed before can also be tried on other types of ratio estimators, although typically with less general results. Take the multivariate ratio estimator (Olkin, 1958)

$$r_{\text{MR}} = \sum_{i=1}^{P} W_i \frac{\bar{y}}{\bar{z}_i} Q_i \equiv \sum_{i=1}^{P} W_i r_i \tag{3.3}$$

as an example, where $(\bar{z}_1, \ldots, \bar{z}_P)$ are the sample means of P auxiliary variables, and $\Sigma W_i = 1$. Just as with (2.4), we can apply (1.2) to each r_i in (3.3), which yields

$$\text{Bias}(r_{\text{MR}}) = \sum_{i=1}^{P} - W_i \frac{\text{Cov}(r_i, \bar{z}_i)}{E(\bar{z}_i)}.$$

It follows that

$$\begin{aligned} &|\text{Bias}(r_{\text{MR}})| \\ &\leqslant \sum_{i=1}^{P} W_i \sigma(r_i) \cdot \text{CV}(\bar{z}_i) \\ &\leqslant \left[\sum_{i=1}^{P} W_i^2 \sigma^2(r_i)\right]^{1/2} \left[\sum_{i=1}^{P} \text{CV}^2(\bar{z}_i)\right]^{1/2}. \end{aligned}$$

Unlike r_s, however, we need impose further assumptions to bound $\text{ABR}(r_{\text{MR}})$ because r_i, $i = 1, \ldots, P$, are typically not independent. For example, if

$$\sigma^2(r_{\text{MR}}) \geqslant \sum_{i=1}^{P} W_i^2 \sigma^2(r_i), \tag{3.4}$$

then

$$\text{ABR}(r_{\text{MR}}) \leqslant \left[\sum_{i=1}^{P} \text{CV}^2(\bar{z}_i)\right]^{1/2}. \tag{3.5}$$

A sufficient condition for (3.4) is $\text{Cov}(r_i, r_j) \geqslant 0$ for all i, j, which does often hold when z_i and z_j are positively correlated (see Olkin, 1958, for calculations of $\text{Cov}(r_i, r_j)$). Extensions of (3.5) to stratified multivariate ratio estimators are straight-forward.

Acknowledgements

The author thanks Steven Pedlow, Donald B. Rubin, and referees for helpful comments. This manuscript was prepared using computer facilities supported in part by the National Science Foundation Grants DMS 89-05292, DMS 87-03942, DMS 86-01732, and DMS 84-04941 awarded to the Department of Statistics at the University of Chicago, and by the University of Chicago Block Fund. The work was supported in part by NSF Grant DMS-92-04504 and in part by the University of Chicago/AMOCO fund.

References

Cochran, W.G. (1977), *Sampling Techniques* (Wiley, New York, 3rd ed.).

Hansen, M.H., W.H. Hurwitz and W.G. Madow (1953), *Sample Survey Methods and Theory, Vol. II* (Wiley, New York).

Hartley, H.O. and A. Ross (1954), Unbiased ratio estimators, *Nature*, **174**, 270–271.

Kish, L. (1965), *Survey Sampling* (Wiley, New York).

Olkin, I. (1958), Multivariate ratio estimation for finite populations, *Biometrika* **45**, 154–165.