BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-85-20

A COMPARISON OF TWO ESTIMATORS FOR THE QUARTERLY FINANCIAL REPORT SURVEY

by

David W. Chapman
Statistical Research Division
Bureau of the Census
Room 3524, F.O.B. #3
Washington, D.C. 20233 U.S.A.

Recommended by: Paul Biemer

Report completed: December 1984

A Comparison of Two Estimators for the Quarterly Financial Report Survey

by

David W. Chapman
Paul P. Biemer

Statistical Research Division

December 1984

ACKNOWLEDGEMENTS

This report was prepared by David W. Chapman and Paul P. Biemer of the Statistical Research Division. The research was based on information given at meetings and on documents provided by the following Economic Surveys Division personnel: Mitchell Trager, James Lowerre, Kenneth Sausman, and Edwin Robison. The following persons reviewed the first draft and provided useful comments: Edwin Robison and James Lowerre of the Economic Surveys Division and Kirk Wolter and Lawrence Ernst of the Statistical Research Division. Also, Carma Hogue of the Statistical Research Division participated in the project research.

Table of Contents

A Comparison of Two Estimators for the Quarterly Financial Report Survey

1. Introduction

The Quarterly Financial Report (QFR) Survey is a continuing survey of about 15,000 corporations per quarter. Of these about 6,000 are certainty (self-representing) selections and are always in the sample. The remaining 9,000 corporations are selected on a rotating basis. These corporations remain in the sample for eight quarters with one-eighth of the sample being rotated in each quarter.

One half of the annual noncertainty sample (about 4,500 corporations) is selected during the second quarter of each year and divided into four panels of approximately equal size. These four panels are introduced, one at a time, in the four quarters following the quarter in which they were selected.

The method of selecting the approximately 4,500 noncertainty corporations for the sample each year is a stratified random sample with an approximate optimum allocation to strata based on previous data. The strata are defined by the cross classification of asset size class and type of industry. Both of these classification variables are subject to change for a corporation between the time the corporation is selected and the time it is enumerated in the QFR survey. Some classifications change because of actual changes in the characteristics of the corporation over time. Other changes are due to classification errors at the time of sampling.

The most important estimate derived from the QFR Survey is the estimated total net income for all corporations in the population or in a subpopulation (e.g., for corporations in a specific industry). If y_i is

the net income reported for a specific quarter by the i-th corporation in a subpopulation, and if w_i is the weight assigned to that corporation for estimation purposes, the subpopulation total, Y, would be estimated as follows:

$$\hat{Y} = \sum_i w_i y_i \ .$$

If the weight, w_i, is set equal to the inverse of the probability of selection, $\hat{Y}$ is an unbiased estimator. In most survey applications w_i is initially taken to be this inverse. However, w_i is often adjusted to account for nonresponse. Furthermore, w_i is sometimes adjusted so that the weights add to known or estimated subpopulation totals. This latter adjustment is usually referred to as a ratio adjustment or a post-stratification adjustment. Post-stratification adjustments are used in situations for which the mean square error of an estimator should be reduced, even though the use of such adjustments introduces a bias in the estimator.

With the current method of estimating Y for the QFR Survey, referred to as the "variable-weights" estimator (VWE), the weight w_i is not assigned to a corporation based on the initial probability of selection. Instead, the weights are assigned on the basis of only a post-stratification adjustment. The post-stratification cells are defined by the joint classification of initial asset size class and type of industry at the time of enumeration. Each sample corporation in a specific cell is assigned a weight equal to the ratio of the estimated total number of corporations in the cell at the time of enumeration to the number of sample corporations in the cell at the time of enumeration.

The more typical method of assigning weights to corporations--i.e., that based on initial selection probabilities--is not being considered

for the QFR. The estimator based on these weights is referred to as the "fixed-weights" estimator (FWE). The purpose of this investigation has been to compare the VWE and FWE with respect to bias and variance and to recommend which estimator to use. Since there is no sampling error associated with certainty selections, the study has addressed only the noncertainty portion of the sample.

Section 2 contains a summary of the results. Section 3 includes the notation used in developing and presenting the statistical results. An approximation for the bias of the VWE is given in Section 4. Approximations for the variances of both estimators are given in Section 5. The final section contains conclusions and recommendations.

2. Summary of Results

Since the VWE does not take into account the probabilities of selection, it is a biased estimator of a population total. Furthermore, the bias of the VWE does not decrease with increasing sample size. The level of bias depends on the variation among the initial stratum sampling rates, which appears to be quite high. Since the weight assigned to a corporation for the FWE of a total is equal to the inverse of its selection probability, the FWE is unbiased. Consequently, the FWE has an important advantage over the VWE in terms of bias.

However, for estimating a subpopulation total for all corporations in a specific type of industry, the variance of the VWE is probably lower than the variance of the FWE. This conclusion is based on the fact that for the VWE all the corporations selected from a specific asset size class that are classified in the same type of industry at the time of enumeration are given the same weight. On the other hand, the weights assigned to corporations for the FWE for the same subpopulation total will vary

somewhat because of corporation switches into that specific type of industry between the time of sampling and the time of enumeration. This weight variation will generally add to the variance of the FWE. As discussed in more detail in Section 5, it is also possible, though not evident, that the variance of the VWE is less than that of the FWE for an estimator of a total for the entire population. Consequently, the VWE appears to have an advantage over the FWE in terms of variance.

The comparison of the FWE and the VWE amounts to a bias-variance trade-off: The FWE is unbiased, whereas, the VWE is biased. However, the VWE appears to have a lower variance than the FWE, at least for certain subpopulation estimates. Unfortunately, it is difficult to compare the mean square error (MSE) of the two estimators because the approximate variance expression for the variance of the VWE given in Section 5 is rather complex. Therefore, there is apparently no straightforward algebraic comparison that can be made of the two estimators. However, as discussed in Sections 5 and 6, other types of comparisons are possible.

3. Notation

In an attempt to simplify the analysis without invalidating the comparisons, the sample rotation was not included in the analysis. Instead, it was assumed that the entire sample was selected at a single point in time and that the data were gathered for all sample respondents at a later single point in time. Of course, this simplified model still allows for changes in both a corporation's asset size class and type of industry between the time of sampling and the time of enumeration.

The notation used in the next two sections involves terms with either two or four subscripts. For two subscripts, i and h, the first subscript

refers to the asset size class at the time of sampling while the second refers to the industry classification at the time of sampling. For terms with four subscripts, the first two subscripts represent the same specification as for a two-subscript term. The next two subscripts, j and k, represent the asset size class and the industry classification, respectively, at the time of enumeration.

Specifically, the following notation is used:

N_{ih} = the number of corporations in the population that were in the stratum specified by asset class i and type of industry h at the time of sampling,

n_{ih} = the sample size allocated to stratum i-h,

$f_{ih} = n_{ih}/N_{ih}$ = the initial sampling rate used in stratum i-h,

N_{ihjk} = the number of corporations in the population that were in stratum i-h at the time of sampling and in asset class j and industry type k (i.e., stratum j-k) at the time of enumeration,
[Of course, for i=j and h=k, N_{ihjk} represents the number of corporations in the population that start in stratum i-h and are still there at the time of enumeration.]

n_{ihjk} = the number of the n_{ih} corporations selected from stratum i-h that end up in stratum j-k at the time of enumeration,

Y_{ihjk} = the total of the Y-characteristic among those N_{ihjk} corporations in the population that were in stratum i-h at the time of sampling and in stratum j-k at the time of enumeration,

$\bar{Y}_{ihjk}$ = the mean of the Y-characteristic among those N_{ihjk} corporations in the population that were in stratum i-h at the time of sampling and in stratum j-k at the time of enumeration,

$\bar{y}_{ihjk}$ = the simple unweighted mean of the Y-characteristic among those corporations in the sample that were selected from stratum i-h and ended up in stratum j-k,

S^2_{ihjk} = the variance of the Y-characteristic among those N_{ihjk} corporations in the population that were in stratum i-h at the time of sampling and in stratum j-k at the time of enumeration.

A dot inserted in place of a subscript denotes summing or averaging over the replaced subscript. For example,

$n_{i..k} = \sum_h \sum_j n_{ihjk}$ = the number of corporations in the sample that started in asset class i at the time of sampling and ended in industry type k at the time of enumeration.

As another example,

$\bar{y}_{i.jk} = \sum_h n_{ihjk} \bar{y}_{ihjk} / \sum_h n_{ihjk}$

= the simple unweighted mean of those sample corporations that were in asset class i when sampled and in stratum j-k when enumerated.

As a final example,

$Y_{...k}$ = the sum of the Y-characteristic for all corporations in the population that end up in industry type k at the time of enumeration.

4. Estimated Bias of the Variable-Weights Estimator

The VWE for a population total has been presented in Bureau documentation in terms of $\hat{Y}_{i.jk}$, an estimated total for all corporations in the population that were in asset size class i at the time of sampling and end up in asset class j and industry type k at the time of enumeration. These estimates are summed on i and j to obtain an estimated total for

industry type k and summed on i and k to obtain an estimated total for asset class j. The basic VWE is the following*:

$$\hat{Y}_{i.jk} = \left(\frac{\hat{N}_{i..k}}{n_{i..k}} \right) n_{i.jk} \bar{y}_{i.jk}, \qquad (1)$$

where

$$\hat{N}_{i..k} = \sum_h \sum_j \left(\frac{N_{ih}}{n_{ih}} \right) n_{ihjk} = \sum_h \left(\frac{N_{ih}}{n_{ih}} \right) n_{ih\cdot k}.$$

The terms in equation (1) are defined by the notation scheme given in the previous section.

The "variable weight" in the VWE is the term in parentheses in equation (1). This post-stratum weight is simply the inverse of the estimated time-of-enumeration sampling rate for units that started in asset class i and ended up in industry type k.

By comparison, the fixed weights estimator, $\hat{Y}'_{i.jk}$, of the same total is the following:

$$\hat{Y}'_{i.jk} = \sum_h \left(\frac{N_{ih}}{n_{ih}} \right) n_{ihjk} \bar{y}_{ihjk} = \sum_h \left(\frac{N_{ih}}{n_{ih}} \right) y_{ihjk}. \qquad (2)$$

For the FWE the "fixed weight" is the term in parentheses in equation (2). This term is simply the inverse of the selection probability of the units in stratum i-h at the time of sampling. The FWE is unbiased.

Since the VWE does not take into account the probability of selection, it is biased. An approximate expression for the bias of the VWE is given below:

$$\text{Bias}\,(\hat{Y}_{i.jk}) \doteq \sum_h Y_{ihjk} \left(\frac{f_{ih} N_{i..k}}{\sum_h f_{ih} N_{ih.k}} - 1 \right). \qquad (3)$$

The derivation of equation (3) is given in Appendix 1.

---

*Though the notation varies somewhat from that used here, the basic VWE of a total is given in an Economic Surveys Division memorandum, "[illegible] Estimation: IRS Portion," dated January 5, 1983.

In terms of the approximate expression in equation (3), the level of bias depends upon the variation of the stratum sampling rates (f_{ih} terms) applied at the time of sampling. In fact, this bias expression equals zero if the sampling rates are all equal since, in that case, $f_{ih} = f$ and $\sum_h f_{ih} N_{ih.k} = f \sum_h N_{ih.k} = f N_{i..k}$. (Actually, it is easy to show that the VWE and the FWE are identical if the stratum sampling weights are all equal.) It should be noted that this bias is not a function of sample size; that is, the bias does not decrease as the n_{ih} terms are increased.

An interesting perspective on the bias of the VWE is obtained by considering the VWE of $Y_{i..k}$. This estimator is simply the sum over j of the $\hat{Y}_{i.jk}$ estimator given in equation (1). It is easy to show that it can be written as follows:

$$\hat{Y}_{i..k} = N_{i..k} \, (y_{i..k}/n_{i..k}) \, .$$

Since, as shown in Appendix 1, $\hat{N}_{i..k}$ is an unbiased estimator of $N_{i..k}$, the bias of $\hat{Y}_{i..k}$ comes primarily from the bias of the unweighted mean $(y_{i..k}/n_{i..k})$ as an estimator of $(Y_{i..k}/N_{i..k})$. An approximate expression for the bias of $\hat{Y}_{i..k}$ is obtained by summing over j the bias expression for $\hat{Y}_{i.jk}$. This amounts to replacing Y_{ihjk} in equation (3) by $Y_{ih.k}$.

It would be worthwhile to substitute values, or approximate values, into equation (3) to obtain approximate bias levels for the VWE. These substitute values would be based on current or prior data from the QFR. For some factors, a range of values might be considered to investigate the sensitivity of the bias expression to differing values of the terms.

For the QFR sample selected in 1984, the initial sampling rates (f_{ih} values) have been enumerated for all 150 strata (i.e., 5 asset size classes by 35 types of industry, minus 25 empty cells). However,

values for the other terms in equation (3) were not available in time to include in the results discussed in this report. Inspection of the distribution of the 150 f_{ih} values, provided in Table 1, still gives some indication of whether or not the bias of the VWE is important.

Table 1. Distribution of Initial Sampling Rates for the 1984 QFR Sample

Initial Sampling Rate	Number of Strata	Percent of Strata
Less than .005	16	10.7
.005 up to .01	23	15.3
.01 up to .02	19	12.7
.02 up to .03	19	12.7
.03 up to .05	15	10.0
.05 up to .07	19	12.7
.07 up to .10	16	10.7
.10 up to .20	17	11.3
.20 up to .30	4	2.7
.30 up to .80	0	0
.80 up to .90	2	1.3
	150	100.1

It is evident from Table 1 that there is considerable variation in the initial stratum sampling rates for this QFR sample. The mean of these rates is 0.06 with a standard deviation of 0.10. This suggests that the bias of the VWE may be large enough for serious concern.

5. Variances of the Estimators

An advantage that the FWE has over the VWE is that it is unbiased, whereas the VWE is biased. However, in addition to biases, it is important to compare the variances of the two estimators. Approximate expressions for the variances of the FWE and VWE are given in equations (4) and (5) below:

$$\mathrm{Var}(\hat{Y}'_{i.jk}) \doteq \sum_h \frac{N_{ih}}{n_{ih}} N_{ihjk} S^2_{ihjk} + \sum_h \frac{N_{ihjk}}{n_{ih}} (N_{ih} - N_{ihjk}) \bar{Z}^2_{ihjk} . \qquad (4$$

$$\mathrm{Var}(\hat{Y}_{i.jk}) \doteq \frac{1}{E^2(n_{i..k})} \Bigg\{ E^2(y_{i.jk}) \sum_h \frac{N_{ih.k}}{n_{ik}} (N_{ih}-N_{ih.k})$$

$$+ \frac{N_{i..k}^2 E^2(y_{i.jk})}{E^2(n_{i..k})} \sum_h f_{ih} \frac{N_{ih.k}}{N_{ih}} (N_{ih} - N_{ih.k})$$

$$+ N_{i..k}^2 \Big[\sum_h f_{ih} N_{ihjk} S_{ihjk}^2 + \sum_h f_{ih} \frac{N_{ihjk}}{N_{ih}} (N_{ih}-N_{ihjk}) \bar{Y}_{ihjk}^2\Big]$$

$$- 2 \frac{E^2(y_{i.jk}) N_{i..k}}{E(n_{i..k})} \Big[N_{i..k} - \sum_h N_{ih.k}^2/N_{ih}\Big]$$

$$+ 2 N_{i..k} E(y_{i.jk}) \Big[Y_{i.jk} - \sum_h \frac{N_{ih.k}}{N_{ih}} Y_{ihjk}\Big]$$

$$- 2 \frac{N_{i..k}^2 E(y_{i.jk})}{E(n_{i..k})} \Big[\sum_h f_{ih} \frac{N_{ihjk}}{N_{ih}} (N_{ih}-N_{ih.k}) \bar{Y}_{ihjk}\Big] \Bigg\}, \qquad (5)$$

where

$$E(y_{i.jk}) = \sum_h f_{ih} Y_{ihjk} \quad \text{and}$$

$$E(n_{i..k}) = \sum_h f_{ih} N_{ih.k} .$$

The derivation of these approximate variance expressions is given in Appendices 2 and 3. In these derivations, finite population corrections have been excluded to simplify the expressions. Also, in deriving the approximate variance expression for the VWE, only the linear terms of the Taylor series expansion for the VWE were retained. Higher order terms were omitted because they would add considerable complexity to the approximate variance expression. In a MSE comparison, such terms would be $O(1/n^2)$ or higher and would be dominated by the $O(1/n)$ terms in the variance expression and the $O(1)$ terms in the bias expression.

Unfortunately, the approximate variance expressions given in equations (4) and (5) for the two estimators are difficult to compare, primarily due to the complexity of the variance expression for the VWE. Consequently, mean square error derivations and comparisons have not been made.

Even though a concise algebraic comparison of the two variances does not seem possible, it does appear from examining the form of the VWE and the FWE that the VWE has a lower variance than does the FWE. Specifically, it can be observed from inspection of the VWE of $Y_{i.jk}$ given in equation (1) that the sample responses are all given a single weight. However, referring to the FWE given in equation (2), the sample responses will not generally receive the same weight. The fixed weight is assigned at the time of sampling in accordance with the sampling rate used in the stratum defined by the initial asset size class and industry type. The amount of weight variation depends on how many of the $n_{i.jk}$ corporations moved into industry type k between the time of sampling and the time of enumeration. Weight variation generally causes some increase in variance as compared to a procedure which assigns a single weight.

For an estimate of a total for all corporations in the population, it is not evident that the VWE would have a lower variance than would the FWE. The estimator of a population total consists of sums of $\hat{Y}_{i.jk}$ terms (VWE) or $\hat{Y}'_{i.jk}$ terms (FWE). Since neither the $\hat{Y}_{i.jk}$ terms nor the $\hat{Y}'_{i.jk}$ terms are independent, the variance of the sum of such terms is not equal to the sum of the variances. Consequently, even if $\hat{Y}_{i.jk}$ does have a lower variance than $\hat{Y}'_{i.jk}$ for all i, j, and k, it does not follow that the sum of the $\hat{Y}_{i.jk}$ terms would have a lower variance than the corresponding sum of $\hat{Y}'_{i.jk}$ terms. Therefore, it it not possible to infer from the definitions of the VWE and the FWE which of these two estimators has the lower variance

in the case of estimating a total for the entire population (or for most large subpopulations).

An approach to comparing the variances of the VWE and FWE estimators for $Y_{i.jk}$ would be to estimate the parameters in equations (4) and (5) from data collected recently in the QFR. Deriving and substituting sample estimates for the population parameters in equations (4) and (5) might be tedious, but perhaps this could be done for a subpopulation. If sample estimates were also made for the parameters in the bias formula for the VWE given in equation (3), a comparison of the mean square errors could also be made.

6. Conclusions and Recommendations

As pointed out earlier, the fact that the FWE is unbiased, whereas the VWE is biased, is certainly an advantage of the FWE. However, the variance of the VWE of $Y_{i.jk}$ is very likely to be less than that of the corresponding FWE of $Y_{i.jk}$, because only a single weight is assigned to the corporations for the VWE, whereas the weights assigned to the corporations for the FWE will vary. This variation is due to the variation of sampling rates from stratum to stratum and on the number of corporations that change their industry classification between the time of sampling and the time of enumeration.

Consequently, for estimating $Y_{i.jk}$, there appears to be a bias-variance trade-off involved with the choice between the VWE and the FWE. Ideally, a comparison of the mean square errors (MSEs) of the two estimators would be made since the MSE incorporates both the variance and bias of an estimator. Unfortunately, the bias and variance expressions in equations (3), (4), and (5) may be too complex to allow for a useful algebraic comparison of the MSEs. In light of this difficulty, four

optional approaches have been considered for comparing the VWE and the FWE. Each of these options is discussed below.

(1) A numerical comparison of the MSEs of the two estimators could be made by substituting into equations (3), (4), and (5) sample estimates of the unknown parameters, as suggested in the previous section. This approach would be most applicable if the basic comparison of the MSEs of the two estimators is not affected by the inclusion of the rotation groups in the analysis. It seems likely that the rotation groups would not have an important impact on the comparison. Because of the complexities of the formulas, this approach would be tedious.

(2) If it is felt that the rotation groups would affect the comparisons, they could be introduced into the analysis. If this were done, the bias and variance expressions for the FWE and VWE could become considerably more complex than they already are. This could make any type of algebraic or numerical comparison infeasible.

(3) For the full QFR rotation sample, variance estimators for the FWE and the VWE could be computed from the QFR data and compared. Since the FWE is a stratified random sample estimator, standard formulas could be used to estimate its variance. For the VWE, however, variance estimation is much more difficult due to the assignment of weights after enumeration. (Actually, for the VWE, a variance estimator already exists; however, its performance for this type of comparison may not be acceptable.) A replication or pseudo-replication method--e.g., the jackknife method--may have to be used to obtain an appropriate estimate of variance for the VWE. If this were done, the same type of variance estimator should be used for the FWE for purposes of comparison. To compare the [illegible]

of the bias of the VWE would have to be estimated and added to the estimated variance of the VWE. This bias estimate would be obtained theoretically by developing a bias expression for the VWE for the rotation sample that is analagous to that given in equation (3). Sample estimates of the parameters of the derived bias expression would then be needed. The major problem with this type of comparison would be the existence of biases of unknown magnitude in the variance estimators. These biases might be such that the comparison of these MSE estimators of the VWE and the FWE would be misleading.

(4) The comparison of the VWE and the FWE could be based primarily on bias considerations since the variance of an estimator can be controlled to a large extent by increasing the sample size. This would dictate the selection of the FWE since it is unbiased, whereas the VWE is biased and the bias does not decrease with increasing sample size.

Of these four optional approaches to comparing the estimators, perhaps the most rigorous comparison is the first one: estimating the parameters in equations (3)-(5) and comparing the MSEs. If this is done and the VWE seems to have a lower MSE than does the FWE, consideration should be given to reducing the variance of the FWE by applying a ratio adjustment to it, based on known or estimated population totals. This would involve the introduction of a bias, but it would decrease with increasing sample size.

If alternative (1) is not practical because of the cost and time required for such an undertaking, we recommend that the FWE be chosen, primarily on the basis of the bias comparison (alternative (4)). Not only is the bias of the VWE unknown and one that does not decrease as the sample size increases, it could be relatively large since the QTR initial stratum sampling rates appear to vary considerably, as was demonstrated in Table 1 of Section 4.

Furthermore, it is important to recognize that the FWE is a standard estimator for the type of sample used for the QFR--a stratified random sample with an approximate optimum allocation to strata. Conversely, the VWE is a nonstandard estimator which does not have a theoretical justification in classical sampling and estimation methods since it ignores the initial sampling rates. Consequently, the VWE should be viewed as an experimental estimator and, as such, should not be chosen over the FWE unless strong evidence of its superiority over the FWE is presented.

Appendix 1. Approximate Bias of the Variable-Weights Estimator

The variable-weights estimator of $Y_{i.jk}$, given in equation (1) in Section 4, can be written as follows:

$$\hat{Y}_{i.jk} = \hat{N}_{i..k} \left(\frac{n_{i.jk}}{n_{i..k}}\right) \bar{y}_{i.jk} . \tag{6}$$

The advantage of writing $\hat{Y}_{i.jk}$ in this form is that the three components of equation (6) appear to be independent. It seems very likely that $\bar{y}_{i.jk}$ is independent of each of the other two since they are both functions of sample sizes. With regard to the independence of the other two terms, the proportion of the $n_{i..k}$ corporations that end up in asset class j would not appear to be related to the estimate of $N_{i..k}$. Consequently,

$$E(\hat{Y}_{i.jk}) \doteq E(\hat{N}_{i..k})\, E\left(\frac{n_{i.jk}}{n_{i..k}}\right)\, E(\bar{y}_{i.jk}) . \tag{7}$$

Expressions for each of the three expected values in equation (7) are derived below and substituted into equation (7).

$$E(\hat{N}_{i..k}) = E \sum_h \sum_j \left(\frac{N_{ih}}{n_{ih}}\right) n_{ihjk} = \sum_h \left(\frac{N_{ih}}{n_{ih}}\right) \sum_j E(n_{ihjk})$$

$$= \sum_h \left(\frac{N_{ih}}{n_{ih}}\right) \sum_j n_{ih} \left(\frac{N_{ihjk}}{N_{ih}}\right) = \sum_h \sum_j N_{ihjk} = N_{i..k} . \tag{8}$$

Since $\bar{y}_{i.jk} = \frac{y_{i.jk}}{n_{i.jk}}$, both the 2nd and 3rd terms in equation (6) are ratios of random variables. For a ratio of random variables, x/y, an approximate expression for the expected value is:

$$E(x/y) \doteq \frac{E(x)}{E(y)} + \frac{E(x)}{E^3(y)} \mathrm{Var}(y) - \frac{1}{E^2(y)} \mathrm{Cov}(x,y). \tag{9}$$

If x and y are positively correlated, the two higher order terms in equation (9) will tend to cancel each other to some extent. Therefore, since the numerator

and denominator are obviously positively correlated for both the 2nd and 3rd terms of equation (6), the ratio of the expected values will be used to approximate both the 2nd and 3rd terms in equation (7).

$$E\left(\frac{n_{i.jk}}{n_{i..k}}\right) \doteq \frac{E(n_{i.jk})}{E(n_{i..k})} . \tag{10}$$

Since the numerator will cancel out upon substitution back into equation (7), only the denominator of equation (10) needs to be derived:

$$E\left(n_{i..k}\right) = \sum_h \sum_j E(n_{ihjk}) = \sum_h \sum_j n_{ih}\left(\frac{N_{ihjk}}{N_{ih}}\right) = \sum_h f_{ih}\, N_{ih.k} . \tag{11}$$

For the 3rd term in equation (7),

$$E\left(\bar{y}_{i.jk}\right) = E\left(\frac{y_{i.jk}}{n_{i.jk}}\right) \doteq \frac{E(y_{i.jk})}{E(n_{i.jk})} . \tag{12}$$

Since the denominator will cancel out upon substitution back into equation (7), only the numerator of equation (12) needs to be derived:

$$E(y_{i.jk}) = \sum_h E\left(n_{ihjk}\, \bar{y}_{ihjk}\right) = \sum_h E(n_{ihjk})\, E(\bar{y}_{ihjk})$$

$$= \sum_h n_{ih}\left(\frac{N_{ihjk}}{N_{ih}}\right) \bar{Y}_{ihjk} = \sum_h f_{ih}\, Y_{ihjk} . \tag{13}$$

Upon substitution back into equation (7),

$$E\left(\hat{Y}_{i.jk}\right) \doteq E\left(N_{i..k}\right)\left(\frac{E(n_{i.jk})}{E(n_{i..k})}\right)\left(\frac{E(y_{i.jk})}{E(n_{i.jk})}\right)$$

$$= N_{i..k}\left(\frac{E(y_{i.jk})}{E(n_{i..k})}\right) = N_{i..k}\,\frac{\sum_h f_{ih}\, Y_{ihjk}}{\sum_h f_{ih}\, N_{ih.k}} . \tag{14}$$

From equation (14) on approximate bias expression for $\hat{Y}_{i.jk}$ is easily derived:

$$\text{Bias}\,(\hat{Y}_{i.jk}) = E(\hat{Y}_{i.jk}) - Y_{i.jk} = E\,(\hat{Y}_{i.jk}) - \sum_h Y_{ihjk}$$

$$= \sum_h \frac{f_{ih}\, N_{i..k}}{\sum_h f_{ih} N_{ih.k}}\, Y_{ihjk} - \sum_h Y_{ihjk}$$

$$= \sum_h Y_{ihjk}\, \left(\frac{f_{ih}\, N_{i..k}}{\sum_h f_{ih}\, N_{ih.k}} - 1\right) .$$

This bias expression is given as equation (3) in Section 4.

Appendix 2. Approximate Variance of the Fixed-Weights Estimator

From equation (2) in Section (4), the fixed-weights estimator of $Y_{i.jk}$ is

$$\hat{Y}'_{i.jk} = \sum_h \left(\frac{N_{ih}}{n_{ih}}\right) n_{ihjk}\, \bar{Y}_{ihjk} \cdot$$

Using the conditional variance theorem,

$$\mathrm{Var}(\hat{Y}'_{i.jk}) = E[\mathrm{Var}(\hat{Y}'_{i.jk}|n_{ihjk})] + \mathrm{Var}[E(\hat{Y}'_{i.jk}|n_{ihjk})] \qquad (15)$$

The two terms in equation (15) are derived separately. Ignoring finite population corrections, the first term of equation (15) is

$$E[\mathrm{Var}(\sum_h \left(\frac{N_{ih}}{n_{ih}}\right) n_{ihjk}\, \bar{y}_{ihjk}|n_{ihjk})] = E[\sum_h \left(\frac{N_{ik}}{n_{ih}}\right)^2 n_{ihjk}^2 \frac{S_{ihjk}^2}{n_{ihjk}}]$$

$$= \sum_h \left(\frac{N_{ih}}{n_{ih}}\right)^2 S_{ihjk}^2\, E(n_{ihjk}) = \sum_h \left(\frac{N_{ih}}{n_{ih}}\right)^2 S_{ihjk}^2\, n_{ik} \left(\frac{N_{ihjk}}{N_{ih}}\right)$$

$$= \sum_h \left(\frac{N_{ih}}{n_{ih}}\right) N_{ihjk}\, S_{ihjk}^2 \cdot \qquad (16)$$

The second term of equation (15) is

$$\mathrm{Var}[E(\sum_h \left(\frac{N_{ih}}{n_{ih}}\right) n_{ihjk}\, \bar{y}_{ihjk}|n_{ihjk})] = \mathrm{Var}(\sum_h \left(\frac{N_{ih}}{n_{ih}}\right) n_{ihjk}\, \bar{Y}_{ihjk})$$

$$= \sum_h \left(\frac{N_{ih}}{n_{ih}}\right)^2 \bar{Y}_{ihjk}^2\, \mathrm{Var}\,(n_{ihjk}) = \sum_h \left(\frac{N_{ih}}{n_{ih}}\right)^2 \bar{Y}_{ihjk}^2\, n_{ih} \left(\frac{N_{ihjk}}{N_{ih}}\right)\left(1-\frac{N_{ihjk}}{N_{ih}}\right)$$

$$= \sum_h \left(\frac{N_{ihjk}}{n_{ih}}\right) (N_{ih} - N_{ihjk})\, \bar{Y}_{ihjk}^2 \cdot \qquad (17)$$

Therefore, ignoring finite population corrections, the variance of the fixed-weights estimator is the sum of equations (16) and (17), as given in equation (4) in Section 5.

Appendix 3. Approximate Variance of the Variable-Weights Estimator

In order to derive an approximate variance expression for the variable-weights estimator, it was written in the following form:

$$\hat{Y}_{i.jk} = \frac{\hat{N}_{i..k}}{n_{i..k}} y_{i.jk} = f(\hat{N}_{i..k}, n_{i..k}, \bar{y}_{i.jk}). \tag{18}$$

Next, $\hat{Y}_{i.jk}$ was expanded about $(N_{i..k}, E(n_{i..k}), E(y_{i.jk}))$ using a Taylor Series expansion, retaining only the linear terms. From this expansion, the following variance expression was derived:

$$\begin{aligned} \mathrm{Var}(\hat{Y}_{i.jk}) \doteq \frac{1}{E^2(n_{i..k})} \Big[& E^2(y_{i.jk})\, \mathrm{Var}(\hat{N}_{i..k}) + \frac{N_{i..k}^2\, E^2(y_{i.jk})}{E^2(n_{i..k})}\, \mathrm{Var}(n_{i..k}) \\ & + N_{i..k}^2\, \mathrm{Var}(y_{i.jk}) - 2\, \frac{N_{i..k}\, E^2(y_{i.jk})}{E(n_{i..k})}\, \mathrm{Cov}(\hat{N}_{i..k}, n_{i..k}) \\ & + 2\, N_{i..k}\, E(y_{i.jk})\, \mathrm{Cov}(\hat{N}_{i..k}, y_{i.jk}) - 2\, \frac{N_{i..k}^2\, E(y_{i.jk})}{E(n_{i..k})}\, \mathrm{Cov}(n_{i..k}, y_{i.jk}) \Big]. \end{aligned} \tag{19}$$

The two expectation terms in equation (19) are given in equations (11) and (13) in Appendix 1. It was also shown there that $E(\hat{N}_{i..k}) = N_{i..k}$. The three variance terms and three covariance terms in equation (19) are derived below, ignoring finite population corrections.

$$\mathrm{Var}(\hat{N}_{i..k}) = \mathrm{Var}\Big(\sum_h \frac{N_{ih}}{n_{ih}}\, n_{ih.k}\Big) = \sum_h \frac{N_{ih}^2}{n_{ih}^2}\, \mathrm{Var}(n_{ih.k})$$

$$= \sum_h \frac{N_{ih}^2}{n_{ih}^2}\, n_{ih}\, \frac{N_{ih.k}}{N_{ih}} \left(1 - \frac{N_{ih.k}}{N_{ih}}\right) = \sum_h \frac{N_{ih.k}}{n_{ih}} (N_{ih} - N_{ih.k}) . \tag{20}$$

$$\mathrm{Var}(n_{i..k}) = \mathrm{Var}\Big(\sum_h n_{ih.k}\Big) = \sum_h \mathrm{Var}(n_{ih.k})$$

$$= \sum_h n_{ih}\, \frac{N_{ih.k}}{N_{ih}} \left(1 - \frac{N_{ih.k}}{N_{ih}}\right) = \sum_h f_{ih}\, \frac{N_{ih.k}}{N_{ih}} (N_{ih} - N_{ih.k}). \tag{21}$$

Using the conditional variance formula and ignoring finite population corrections,

$$\mathrm{Var}(y_{i.jk}) = \mathrm{Var}\Big(\sum_h n_{ihjk}\,\bar{y}_{ihjk}\Big) = \sum_h \mathrm{Var}(n_{ihjk}\,\bar{y}_{ihjk}). \tag{22}$$

$$\mathrm{Var}(n_{ihjk}\,\bar{y}_{ihjk}) = E[\mathrm{Var}(n_{ihjk}\,\bar{y}_{ihjk}|n_{ihjk})] + \mathrm{Var}[E(n_{ihjk}\,\bar{y}_{ihjk}|n_{ihjk})]$$

$$= E\Big(n_{ihjk}^2 \frac{S_{ihjk}^2}{n_{ihjk}}\Big) + \mathrm{Var}(n_{ihjk}\,\bar{Y}_{ihjk})$$

$$= S_{ihjk}^2\,E(n_{ihjk}) + \bar{Y}_{ihjk}^2\,\mathrm{Var}(n_{ihjk})$$

$$= S_{ihjk}^2\,n_{ih}\frac{N_{ihjk}}{N_{ih}} + \bar{Y}_{ihjk}^2\,n_{ih}\frac{N_{ihjk}}{N_{ih}}\Big(1-\frac{N_{ihjk}}{N_{ih}}\Big)$$

$$= f_{ih}\,N_{ihjk}\,S_{ihjk}^2 + f_{ih}\frac{N_{ihjk}}{N_{ih}}(N_{ih} - N_{ihjk})\,\bar{Y}_{ihjk}^2. \tag{23}$$

Substituting equation (23) into equation (22),

$$\mathrm{Var}\,(y_{i.jk}) = \sum_h f_{ih}\,N_{ihjk}\,S_{ihjk}^2 + \sum_h f_{ih}\,\frac{N_{ihjk}}{N_{ih}}\,(N_{ih}-N_{ihjk})\,\bar{Y}_{ihjk}^2. \tag{24}$$

$$\mathrm{Cov}\,(\hat{N}_{i..k}, n_{i..k}) = \mathrm{Cov}\Big(\sum_h\sum_j \Big(\frac{N_{ih}}{n_{ih}}\Big) n_{ihjk}, \sum_h\sum_j n_{ihjk}\Big)$$

$$= \mathrm{Cov}\Big(\sum_h \frac{N_{ih}}{n_{ih}}\,n_{ih.k}, \sum_h n_{ih.k}\Big) = \sum_h \frac{N_{ih}}{n_{ih}}\,\mathrm{Var}(n_{ih.k})$$

$$= \sum_h \frac{N_{ih}}{n_{ih}}\,n_{ih}\frac{N_{ih.k}}{N_{ih}}\Big(1-\frac{N_{ih.k}}{N_{ih}}\Big) = N_{i..k} - \sum_h \frac{N_{ih.k}^2}{N_{ih}}. \tag{25}$$

In order to derive $\mathrm{Cov}(\hat{N}_{i..k}, y_{i.jk})$, two special results are used. First, if x, y, and z are random variables such that z is independent of both x and y, $\mathrm{Cov}\,(x, yz) = E(z)\,\mathrm{Cov}\,(x,y)$. Second, suppose that a population of N units

is partitioned into k categories with frequencies $N_1, N_2, \ldots, N_k$. Suppose further that a simple random sample of n units is selected from N, without replacement, and that the observed sample frequencies for the k categories are $n_1, n_2, \ldots, n_k$. Then, $\text{Cov}(n_i, n_j) = -n\left(\frac{N_i}{N}\right)\left(\frac{N_j}{N}\right)$ for $i \neq j$.

$$\text{Cov}(\hat{N}_{i..k}, \bar{y}_{i\cdot jk}) = \text{Cov}\left(\sum_h \sum_j \frac{N_{ih}}{n_{ih}} n_{ihjk}, \sum_h n_{ihjk}\bar{y}_{ihjk}\right)$$

$$= \text{Cov}\left(\sum_h \frac{N_{ih}}{n_{ih}} n_{ih\cdot k}, \sum_h n_{ihjk}\bar{y}_{ihjk}\right) = \sum_h \frac{N_{ih}}{n_{ih}} \text{Cov}(n_{ih\cdot k}, n_{ihjk}\bar{y}_{ihjk})$$

$$= \sum_h \frac{N_{ih}}{n_{ih}} E(\bar{y}_{ihjk}) \text{Cov}(n_{ih\cdot k}, n_{ihjk}) = \sum_h \frac{N_{ih}}{n_{ih}} \bar{Y}_{ihjk} \text{Cov}(n_{ih\cdot k}, n_{ihjk}). \quad (26)$$

$$\text{Cov}(n_{ih\cdot k}, n_{ihjk}) = \text{Cov}\left(\sum_j n_{ihjk}, n_{ihjk}\right) = \text{Var}(n_{ihjk}) + \sum_{j' \neq j} \text{Cov}(n_{ihjk}, n_{ihj'k})$$

$$= n_{ih} \frac{N_{ihjk}}{N_{ih}}\left(1 - \frac{N_{ihjk}}{N_{ih}}\right) + \sum_{j' \neq j} -n_{ih} \frac{N_{ihjk}}{N_{ih}} \frac{N_{ihj'k}}{N_{ih}}$$

$$= f_{ih} \frac{N_{ihjk}}{N_{ih}}(N_{ih} - N_{ihjk}) - f_{ih} \frac{N_{ihjk}}{N_{ih}} \sum_{j' \neq j} N_{ihj'k}$$

$$= f_{ih} \frac{N_{ihjk}}{N_{ih}}(N_{ih} - N_{ihjk}) - f_{ih} \frac{N_{ihjk}}{N_{ih}}(N_{ih\cdot k} - N_{ihjk}) = f_{ih} \frac{N_{ihjk}}{N_{ih}}(N_{ih} - N_{ih\cdot k}). \quad (27)$$

Substituting equation (27) into equation (26):

$$\text{Cov}(\hat{N}_{i..k}, \bar{y}_{i\cdot jk}) = \sum_h \frac{N_{ih}}{n_{ih}} \bar{Y}_{ihjk} f_{ih} \frac{N_{ihjk}}{N_{ih}}(N_{ih} - N_{ih\cdot k})$$

$$= \sum_h \frac{N_{ihjk}}{N_{ih}}(N_{ih} - N_{ih\cdot k}) \bar{Y}_{ihjk} = \bar{Y}_{i\cdot jk} - \sum_h \frac{N_{ih\cdot k}}{N_{ih}} \bar{Y}_{ihjk}. \quad (28)$$

The derivation of $\text{Cov}(n_{i..k}, \bar{y}_{i\cdot jk})$ is similar to the derivation of $\text{Cov}(\hat{N}_{i..k}, \bar{y}_{i\cdot jk})$:

$$\text{Cov}(n_{i..k}, y_{i.jk}) = \text{Cov}\left(\sum_h n_{ih.k}, \sum_h n_{ihjk}\,\bar{y}_{ihjk}\right) = \sum_h \text{Cov}(n_{ih.k}, n_{ihjk}\,\bar{y}_{ihjk})$$

$$= \sum_h \bar{Y}_{ihjk}\,\text{Cov}(n_{ih.k}, n_{ihjk}) = \sum_h \bar{Y}_{ihjk}\, f_{ih}\,\frac{N_{ihjk}}{N_{ih}}\,(N_{ih}-N_{ih.k})$$

$$\Rightarrow \sum_h f_{ih}\,\frac{N_{ihjk}}{N_{ih}}\,(N_{ih}-N_{ih.k})\,\bar{Y}_{ihjk}. \qquad (29)$$

Substituting the variance and covariance expressions in equations (20), (21), (24), (25), (28), and (29) into equation (19), the variance expression given in euqation (5) of Section (5) is obtained. If expressions for $E(n_{i..k})$ and $E(y_{i.jk})$ from Appendix 1 are also substituted into equation (19), the following variance expression is obtained:

$$\text{Var}(\hat{Y}_{i.jk}) \doteq \frac{1}{\left(\sum_h f_{ih}\,N_{ih.k}\right)^2}\left(\sum_h f_{ih}Y_{ihjk}\right)^2 \sum_h \frac{N_{ih.k}}{n_{ih}}\,(N_{ih}-N_{ih.k})$$

$$+\frac{N_{i..k}^2\left(\sum_h f_{ih}\,Y_{ihjk}\right)^2}{\left(\sum_h f_{ih}N_{ih.k}\right)^2}\sum_h f_{ih}\,\frac{N_{ih.k}}{N_{ih}}\,(N_{ih}-N_{ih.k})$$

$$+ N_{i..k}^2\left[\sum_h f_{ih}N_{ihjk}\,S_{ihjk}^2 + \sum_h f_{ih}\left(\frac{N_{ihjk}}{N_{ih}}\right)(N_{ih}-N_{ihjk})\,\bar{Y}_{ihjk}^2\right]$$

$$- 2\,\frac{N_{i..k}\left(\sum_h f_{ih}\,Y_{ihjk}\right)^2}{\sum_h f_{ih}N_{ih.k}}\left(N_{i..k} - \sum_h \frac{N_{ih.k}^2}{N_{ih}}\right)$$

$$+ 2\,N_{i..k}\left(\sum_h f_{ih}\,Y_{ihjk}\right)\left(Y_{i.jk} - \sum_h \frac{N_{ih.k}}{N_{ih}}\,Y_{ihjk}\right)$$

$$- 2\,\frac{N_{i..k}^2\left(\sum_h f_{ih}\,Y_{ihjk}\right)}{\sum_h f_{ih}\,N_{ih.k}}\left[\sum_h f_{ih}\left(\frac{N_{ihjk}}{N_{ih}}\right)(N_{ih}-N_{ih.k})\,\bar{Y}_{ihjk}\right]. \qquad (30)$$