ESTIMATING THE VARIANCES OF PRICE INDEXES BY HALF SAMPLING: WHY IT WORKS

Phillip S. Kott, Bureau of Labor Statistics

The Bureau of Labor Statistics (BLS) estimates Laspyres price indexes for a multitude of commodity and product classes using complex ratio formulae and, where possible, probability proportionate to size (PPS) sampling designs. Half sampling techniques are employed to estimate the variances of these indexes. There is little proof, however, that these methods produce accurate measures of sampling variances. This paper uses superpopulation models to provide formal analysis and support of BLS variance estimation procedures.

The statistical literature, of course, is not mute on the subject. McCarthy (1966) introduces the concept of balanced half sample variance estimation and shows that it is exact for linear estimators with simple random sampling within strata. A host of empirical papers study the use of McCarthy's techniques for measuring the variance of ratio estimators under the same sampling design. These include Frankel (1971), Kish and Frankel (1974), Lemeshow and Levy (1978) and many others.

A formal analysis of complex estimators is provided by Krewski and Rao (1981), who show that balanced half sample variance estimation has desired properties as the number of strata approaches infinity. Of more practical interest is their analysis when the number of strata is finite. Krewski and Rao use a superpopulation model to analyze the variance estimation of ratio estimators again with simple random sampling within strata (see also 1979). Their superpopulation work differs from the analysis presented here. It is more Bayesian and is limited to the combined ratio estimator with simple random sampling within strata. In addition, Krewski and Rao specify random variables in a way that leads to some misleading results when applied to price index estimators. Nevertheless, a special case of their results are comparable to a special case of the results found here.

1. INTRODUCTION

BLS designs its sampling procedures to facilitate half sample variance calculations. When the price index for a particular population of commodities (or products) is desired, the population is first stratified, usually by geographical region, and then two primary sampling units are chosen from each stratum through some form of PPS sampling. In order to introduce the concept of a superpopulation model in the most straightforward manner, we begin in Section 2 with a price index for a homogenous population of commodities that is estimated based on a without replacement PPS sample of two commodity units.

We extend this simple analysis tp subsampling in Section 3 and to stratified populations in Section 4. Probability samples are drawn with what are assumed to be optimal measures of size in the first four sections. As a result, the index estimators have a linear form not much different from that in the classical half sampling literature (see McCarthy). Nevertheless, since sampling is proportionate to size, half sample variance estimators are not in general unbiased relative to the sampling design. This is where the superpopulation model comes in: these variance estimators are shown to be nearly model-unbiased ("nearly" because a finite population correction factor is omitted).

In Section 5, the assumption of optimal measures of size in PPS sampling is dropped. This forces the index estimator into ratio form. Arbitrary half sample variance estimators are shown to be nearly model-unbiased for a separate ratio estimator if and only if the superpopulation model is extended to include another random variable. A direct estimator of the sampling variance is introduced and shown to be more model-efficient than a "fully balanced" half sample variance estimator. In Section 6, a direct variance estimator (what Krewski and Rao call the "linearization estimator") proves to be model-biased downward for a combined ratio estimator when the number of strata is small or a few strata are relatively large. In addition, the fully balanced half sample variance estimator turns out also to be model--biased downward, but less so and only when the strata are not of equal size.

2. THE HORVITZ-THOMPSON ESTIMATOR AND PPS SAMPLING

The simplest type of price index is a long term relative for a population of N units (commodities or products):

$$R^t = (\Sigma P_i^t Q_i^0)/(\Sigma P_i^0 Q_i^0) = \Sigma b_i x_i,$$

where P_i^s is the price of unit i at time s,

Q_i^0 is the quantity of i at 0 (the base period),

$b_i = P_i^0 Q_i^0 / \Sigma P_i^0 Q_i^0$ is i's base expenditure (revenue) share,

and $x_i = P_i^t / P_i^0$ is i's price trend.

Suppose we wish to estimate this relative based on a (probability) sample, S, of two units. Godambe (1955)

shows that no linear, unbiased estimator, $\hat{R}^t=\sum_i a_i x_i$, where $E(\hat{R}^t)=R^t$, exists with minimum variance for all possible values of the x_i (price trends). How then can the "best" estimation strategy be determined?

Godambe suggests that the x_i be treated as independent, identically distributed random variables from a superpopulation with mean u_x and variance g_x^2. It is then possible to find the linear, sampling-unbiased estimator ($E_s[\hat{R}^t-R^t]=0$) with the least model--expected sampling variance. It is the Horvitz-Thompson (HT) estimator in mean-of-ratios form:

$$\hat{R}^t=\sum b_i x_i s_i/E(s_i)=\sum x_i s_i/2, \qquad (2.1)$$

where s_i is the number of times unit i is in the sample, and

$$E(s_i)=E(s_i^2)=2b_i. \qquad (2.2)$$

(Godambe never said so explicitly, but it is necessary to assume that no unit expenditure share, b_i, is greater than 1/2.)

Equation (2.2) is satisfied by many different "without replacement" PPS sampling designs. Perhaps the simplest to execute is systematic PPS sampling with the b_i as the measures of size. This design coupled with the HT estimator form a strategy that is very popular at BLS. Readers unfamiliar with systematic PPS sampling are referred to Raj (1968), pp. 51-2.

The model-expected sampling variance of the HT estimator under any sampling plan obeying (2.2) is

$$E_x Var_s(\hat{R}^t)=g_x^2(1/2-\sum b_i^2). \qquad (2.3)$$

Unless we want to distinquish between different samples, we will assume that samples consist of the first two enumerated units. Let us suppose further that unit 1 is in half sample A and unit 2 in B. The estimator of variance from either half sample is

$$\hat{Var}_A(\hat{R}^t|s_1=s_2=1)=(x_1-\hat{R}^t)^2$$
$$=(x_1-x_2)^2/4$$
$$=\hat{Var}_B(\hat{R}^t|\cdot).$$

Notice that the model expectation of $\hat{Var}_A(\hat{R}^t|s_1=s_2=1)$ is $g_x^2/2$. Since this is true for every pair of sampled units, we conclude that

$$E_x[\hat{Var}_A(\hat{R}^t)]=g_x^2/2. \qquad (2.4)$$

What is missing form $\hat{Var}_A(\cdot)$ to make it model-unbiased is a finite population correction factor $(1-2\sum b_i^2)$.

Let us examine the assumptions of our superpopulation model to see which are really needed for our results. It is certainly necessary for the x_i to have a common mean. If, in addition, we assumed only that the price trends have a common covariance (c), then one could show that

$$E_x E_s[Var_s(\hat{R}^t)]=\sum(b_i-2b_i^2)(g_x^2-c)/2$$

and

$$E_x E_s[\hat{Var}_A(\hat{R}^t)]=\sum b_i(g_x^2-c)/2.$$

The HT estimator and without replacement PPS sample design may not be optimal in this case. Nevertheless, the model expectation of the variance estimator has nearly no sampling bias (a weaker condition than near model-unbiasedness). The finite population correction factor becomes a complicated expression which is left to the reader to derive.

If the population is large enough and the shares small enough that $\sum b_i^2$ is nearly zero, then we can ignore finite population correction entirely. This is fortunate because the relative sizes of the $g_{x_i}^2$ are unknown. Too keep the notation in this paper as simple as possible, we will assume all $g_{x_i}^2=g_x^2$ from now on. This will also allow us to speak of nearly model-unbiased variance estimators rather than estimators with nearly no model expected sampling biases.

3. SUBSAMPLING

Finite population correction becomes even more complicated when there is a multistage sampling design; in which case, the units discussed above are only primary sampling units (clusters of actual pricing units), and the x_i are themselves HT estimators of price relatives. To simplify the expostion, let us assume that each $x_i=\sum b_{ij} x_{ij}$, $x_{ij}=P_{ij}^t/P_{ij}^0$, $b_{ij}=P_{ij}^0 Q_{ij}^0/\sum_{j=1}^{M_i} P_{ij}^0 Q_{ij}^0$, and M_i is the number of pricing units in cluster i. Suppose that two units are subsampled from each sampled cluster so that

$$\hat{R}^t=\sum b_i \sum b_{ij} x_{ij},$$

and

$$\hat{R}^t=\sum\sum x_{ij} s_{ij} s_i/4,$$

where s_{ij} is 1 if unit ij is in the subsample of cluster i, and $E(s_{ij})=E(s_{ij}^2)=2b_{ij}$.

We can now borrow a devise from Scott and Smith (1969) and consider a superpopulation model in which the x_{ij} within each i are independent random variables with mean u_i and variance g_x^2. These means, however, are also independent random variables with a common mean, u_x, and variance, g_T^2. Thus with respect to all the superpopulation variables, the price trends of units from within the same cluster have a correlation coeffient of $g_T^2/(g_x^2+g_T^2)$.

We can then show that the model--expected sampling variance of $\hat{R}^t$ is

$$E_x Var_s(\hat{R}^t)=g_x^2(1/4-\sum b_i^2\sum b_{ij}^2)+$$
$$g_T^2(1/2-\sum b_i^2),$$

while the model-expectation of the half sample variance estimator based on the first stage of sampling alone, given $s_{11}=s_{12}=s_{21}=s_{22}=1$, is

$$E_x \hat{Var}_A(\hat{R}^t|\cdot)=[(x_{11}+x_{12})/2-\hat{R}^t]^2$$
$$=g_x^2/4+g_T^2/2.$$

If we assume that $\sum b_i^2$ is nearly zero, then $\hat{Var}_A(\cdot)$ is nearly model-unbiased. The extension of this result to more complicated subsampling schemes is straightforward.

4. STRATIFIED SAMPLING

The BLS does not estimate many price relative based on only two primary PPS sampling units. Most of the Bureau's index calculations involve complex estimation formulae. In addition, samples are often chosen with suboptimal measures of size. Let us ease into the subject slowly. Suppose we have a population of L strata with N_k units in each each stratum k. Let ki denote the ith unit in the kth stratum. The long term relative of the population is

$$R^t=\Sigma b_k \Sigma b_{ki} x_{ki},$$

where $b_{ki}=P_{ki}^0 Q_{ki}^0/\Sigma P_{ki}^0 Q_{ki}^0$,

and $b_k=\Sigma P_{ki}^0 Q_{ki}^0/\Sigma\Sigma P_{ki}^0 Q_{ki}^0$.

Suppose two units from each stratum are chosen via without replacement PPS sampling. Let the x_{ki} be independent random variables with mean u_k and variance g_k^2. If

$$\hat{R}^t=\Sigma b_k \Sigma x_{ki} s_{ki}/2, \qquad (4.1)$$

then

$$E_x Var_s(\hat{R}^t)=\Sigma b_k^2 (1/2 - \Sigma b_{ki}^2) g_k^2.$$

An arbitrary half sample estimator of the sampling variance is

$$\hat{Var}_A(\hat{R}^t|\cdot)=(\Sigma b_{k1} x_{k1} - \hat{R}^t|\cdot)$$
$$=\Sigma b_k^2 (x_{k1}-x_{k2})^2/4 + \sum_{k\neq l} b_k b_l (x_{k1}-x_{k2})(x_{l1}-x_{l2})/4. \qquad (4.2)$$

While this estimator is nearly model-unbiased when all $\Sigma b_{ki}^2 g_k^2$ are nearly zero, more efficient nearly model-unbiased estimators may be derived by averaging together variance estimators based on different half samples. (In this paper, model efficiency will be measured by (the inverse of) model variance. Alternatives measures, such as $E_x[(\hat{Var}_A(\hat{R}^t|\cdot)-Var_s(\hat{R}^t))^2]$, have been considered by the author. They do not affect the results in the sequel.)

McCarthy (1966) shows that an average of half sample variance estimators is most efficient when the set of half samples on which they are based is orthogonally balanced (pseudoreplication). A set, T, of $L' \geq L$ half samples with this property can be constructed from a L'x L matrix of pluses and minuses with orthogonal columns; a plus (minus) in the jth row and kth column of the matrix would place unit k1 (k2) in half sample j. See McCarthy for details on this procedure.

When variance estimators based on orthogonally balanced half samples are averaged, the cross terms (the $b_k b_l$, $k \neq l$) of (4.2) vanish:

$$\hat{Var}_O(\hat{R}^t|\cdot)=\Sigma_{i\in T}\hat{Var}_i(\hat{R}^t|\cdot)/L'$$
$$=\Sigma b_k^2 (x_{k1}-x_{k2})^2/4, \qquad (4.3)$$

decreasing the model variance of the resulting orthogonally balanced half sample variance estimator. Notice that in the present context of estimating the sampling variance of $\hat{R}^t$ in (4.1), (4.3) can be calculated directly without an orthogonal matrix.

5. SUBOPTIMAL MEASURES OF SIZE

Quite often BLS does not draw PPS samples with perfect measures of size. In the Producer Price Index (PPI) program, base period revenue shares are frequently approximated using employment data. In the Consumer Price Index (CPI) program, two related but different strata are sometimes collapsed into a single "stratum" for variance calculation purposes. In both programs, short term relatives comparing price levels of adjacent periods in time, say periods t and t-1, are based on samples drawn in the base period (0) rather than the reference period (t-1).

5.1 The Single Stratum Case

Return to the case of two sampled units from a single stratum. Let $B_i=P_i^0 Q_i^0$, and suppose that a set of $D_i \propto B_i$ are used as measures of size to draw a PPS sample. Then

$$R^t=(\Sigma D_i v_i x_i)/(\Sigma D_i v_i),$$

and

$$\hat{R}^t=(\Sigma v_i x_i s_i)/\Sigma v_i s_i),$$

where $v_i=B_i/D_i$,

and $E_s(s_i)=E_s(s_i^2)=2D_i/\Sigma D_i$.

We again let the x_i be independent and identically distributed random variables. Notice that the model expectation of R^t is u_x and its model variance is $\Sigma b_i^2 g_x^2$. Let us assume the model standard deviation, $\sqrt{b_i^2}\, g_x$, is nearly zero so that R^t can be treated as nearly u_x. This assumption will be weakened considerably in Section 6.

Ratio estimators like $\hat{R}^t$ are not sampling-unbiased. We ignore this fact and call what are properly sampling mean squared errors "sampling variances" throughout this paper.

Note that

$$E_x E_s[(\hat{R}^t-u_x)^2]=E_s E_x[(\hat{R}^t-u_x|\cdot)^2].$$

It is a simple matter to show

$$Var_x(\hat{R}^t|s_i=s_j=1)=(1+f_{ij}^2)g_x^2/2, \qquad (5.1)$$

where $f_{ij}=(v_i-v_j)/(v_i+v_j)$.

We will not (yet) take the sampling expectation of the RHS of (5.1). Instead, consider this. Suppose there is a sampling variance estimator, $\hat{V}$, with a model expectation nearly equal to the conditional model variance of $\hat{R}^t$ for every sample; i.e.,

$$E_x(\hat{V}|s_i=s_j=1)=(1+f_{ij}^2)g_x^2/2, \ \forall i \neq j. \qquad (5.2)$$

While $\hat{V}$ might not be a nearly model unbiased estimator of the sampling variance, its model expectation has nearly no sampling bias: $E_s E_x(\hat{V}|\cdot)=E_s E_x[(\hat{R}^t-u_x|\cdot)^2]=E_x E_s[(\hat{R}^t-u_x)^2]$.

We now must find a $\hat{V}$ that obeys (5.2) for all possible samples. $\hat{Var}_A(\cdot)$ doesn't; neither does $\hat{Var}_B(\cdot)$; however, the average of these two does:

$$E_x\hat{Var}_A(\hat{R}^t|\cdot)=E_x[(x_1-\hat{R}^t)^2]$$
$$=(1-2f_{12}+f_{12}^2)g_x^2/2, \qquad (5.3)$$

and

$$E_x\hat{Var}_B(\hat{R}^t|\cdot)=(1+2f_{12}+f_{12}^2)g_x^2/2; \qquad (5.4)$$

but

$$E_x(\hat{Var}_A+\hat{Var}_B)/2=(1+f_{12}^2)g_x^2/2. \qquad (5.5)$$

This last variance estimator can be computed directly:

$$\hat{Var}_D(\hat{R}^t|\cdot)=(1+f_{12}^2)(x_1-x_2)^2/4.$$

An alternative approach to analyzing the three variance estimators in this case involves treating the v_i as independent, identically distributed random variables. Obviously, this treatment is not applicable to "collapsed strata." The specification in such cases and the extension of the subsequent analysis are left to the reader.

It is easy to see that under the new model the f_{ij} have mean zero and identical variances, g_f^2, but are not independent. The model-expected sampling variance of $\hat{R}^t$ is $(1+g_f^2)g_x^2/2$, and the three half sample estimators above are x,v-unbiased. The presence of an extra $2f_{ij}$ term in (5.3) and (5.4) make the model variances of $\hat{Var}_A(\cdot)$ and $\hat{Var}_B(\cdot)$ larger than that of $\hat{Var}_D(\cdot)$. (The model variances need no longer be conditional on the sample, because the v_i are identically distributed).

5.2 The Separate Ratio Estimator

Suppose the population is divided into L strata as in Section 4, but now PPS samples of two units from each stratum are drawn using some D_{ki} in place of $B_{ki}=P_{ki}^0 Q_{ki}^0$. In this subsection, we assume that while at least some of the unit base period expenditures (the B_{ki}) are unknown, all the strata base period expenditures, the $B_k=\sum B_{ki}$, are known.

The long term relative can be expressed as

$$R^t=\sum b_k(\sum P_{ki}^t Q_{ki}^0)/(\sum P_{ki}^0 Q_{ki}^0)=\sum b_k R_k^t,$$

where $b_k=B_k/\sum B_k$ as in Section 4.

The separate ratio estimator of R^t based on the sampling scheme discussed above is

$$\hat{R}_S^t=\sum b_k(\sum x_{ki}v_{ki}s_{ki})/(\sum v_{ki}s_{ki}),$$

where $v_{ki}=B_{ki}/D_{ki}$,

$s_{ki}=1$ if ki is in the sample of units from stratum k, and

$E_s(s_{ki})=E(s_{ki}^2)=2D_{ki}/\sum^{N_k} D_{ki}$.

For our superpopulation model, let the x_{ki} and the v_{ki} be independent random variables with repective means and variances u_{vk}, g_x^2 and u_{vk}, $u_{vk}g_v^2$. This is by no means the most general model that can be used to analyze half sample variance estimation techniques. It is possible to allow each stratum to exhibit its own variances (i.e., g_{xk}^2 and g_{vk}^2) and retain the bulk of the subsequent analysis. Nevertheless, this simple model avoids much cumbersome and expositionally useless notation. The extensions to more general forms will be left to the interested reader.

As in the last section, we will assume that for each stratum k, $\sqrt{\sum b_{ki}^2}\,g_x$ is nearly zero ($b_{ki}=B_{ki}/B_k$). Consequently, R^t is nearly $\sum d_k u_{xk}$.

The model-expected sampling variance of $\hat{R}_S^t$ is

$$E_{x,v}\,Var_s(\hat{R}_S^t)=\sum b_k^2(1+g_{f_k}^2)g_x^2/2$$

where $f_k=(v_{k1}-v_{k2})/(v_{k1}+v_{k2})$.

The estimator of this variance based on half sample A, the half sample with all the x_{k1}, is

$$\hat{Var}_A(\hat{R}_S^t|\cdot)=(\sum b_k x_{k1}-\hat{R}_S^t)^2$$
$$=\sum b_k^2(1-f_k)^2(x_{k1}-x_{k2})^2/4$$
$$+\sum_{k\neq l} b_k b_l(1-f_k)(1-f_l)(x_{k1}-x_{k2})(x_{l1}-x_{l2})/4,$$

It is easy to see that this estimator is model unbiased. Nevertheless, other model unbiased half sample estimators exist with less model variance.

An average of fully balanced half sample estimators, i.e., not only are the half samples orthogonally balanced but each sampled unit is in exactly half of the half samples, will result in the following estimator:

$$\hat{Var}_F(\hat{R}_S^t|\cdot)=\sum b_k^2(1+f_k^2)(x_{k1}-x_{k2})^2/4 +$$
$$\sum_{k\neq l} b_k b_l f_k f_l(x_{k1}-x_{k2})(x_{l1}-x_{l2})/4.$$

For a set of L' samples to be fully balanced, L' must exceed L. When L' is a multiple of 4, it is always possible to construct a fully balanced orthogonal matrix.

The fully balanced half sample variance estimator has less model variance that any other combination of half sample estimators. Nevertheless, there is a nearly model unbiased sampling variance estimator with less model variance:

$$\hat{Var}_D(\hat{R}_S^t|\cdot)=\sum b_k^2(1+f_k^2)(x_{k1}-x_{k2})^2/4. \quad (5.6)$$

We will call this the direct half sample variance estimator of $\hat{R}_S^t$.

5.3 An Introduction to the Combined Ratio Estimator

In the previous subsection, we assume that the stratum base period expenditures, the b_k, are known and can be used to estimate the long term relative. Quite often this is not the case. If the b_k were estimated using data from the sample,

$$\hat{b}_k=D_k\sum v_{ki}s_{ki}/(\sum D_k\sum v_{ki}s_{ki}),$$

the long term relative estimator would be

$$\hat{R}_C^t=\sum \hat{b}_k\hat{R}_k^t$$
$$=(\sum D_k\sum x_{ki}v_{ki}s_{ki})/(\sum D_k\sum v_{ki}s_{ki}) \quad (5.7)$$

Equation (5.7) defines the combined ratio estimator in its purest form.

In actual practice, many price index estimators employ the principles of both separate and combined ratio estimation. While we will not deal with such estimators here, we can be confident that the properties of variance estimation discussed in this paper extend to more complex index estimators.

6. THE COMBINED RATIO ESTIMATOR

The combined ratio estimator in (5.7) can be re-expressed as

$$\hat{R}_C^t=(\sum d_k\sum x_{ki}v_{ki}^* s_{ki})/(\sum d_k\sum v_{ki}^* s_{ki}), \quad (6.1)$$

where $v_{ki}^*=v_{ki}/u_{vk}$,

and $d_k=D_k u_{vk}/\sum D_k u_{vk}$.

Let the x_{ki} and the v_{ki} be random variables with the same properties as in Section 5. In addition, let the stratum price trend means, the u_{xk}, be independent random variables with mean u_x and variance g_T^2.

To help detemine when a sampling variance estimator is nearly model--unbiased, we will employ three smallness assumptions. These are listed below:

A1. $\Sigma d_{ki}^2(g_x^2+g_T^2)(1+g_{v*}^2)\doteq 0$ for all k;

A2. $(g_x^2+g_T^2)E_v[(\Sigma d_k[v_{ki}^*-1])^j]\doteq 0$, $j>2$;

A3. $g_{v*}^2(g_x^2+g_T^2)\Sigma d_k^3\doteq 0$.

Assumption A1, a rewording of our standard nearness assumption, is sufficient to ignore finite population correction. Assumption A2 allows us to expand ratios by a Taylor series and then truncate higher order terms. When $d_i=1$, Assumptions A1 and A2 imply $E_v[\Sigma b_{1i}^2 g_x^2]\doteq 0$, the weaker nearness assumption promised last section. Assumption A3 will be dropped in Subsection 6.2. Nevertheless, when the number of strata is large and each stratum share is relatively small, the assumption is very reasonable.

Notice that when all the d_k are equal, Assumption A3 truncates terms of order L^{-2}. When A3 is dropped, however, Assumption A2 does not simply allow the inclusion all L^{-2} terms while truncating higher order terms. It also eliminates expressions like $E_v[(v_{ki}^*-1)^3]/L^{-2}$.

6.1 The Case When Assumption A3 is Valid

Let us begin this subsection by showing that the relative R^t is not nearly u_x under Assumption A1. While $E_{x,v,u}(R^t)=u_x$, $Var_{x,v,u}(R^t)\geq\Sigma d_k^2 g_T^2$, a value potentially too large to ignore.

On the other hand, if we take the expectation and variance of R^t with respect to the x_{ki} and v_{ki} only, we have $E_{x,v}(R^t)=\Sigma d_k u_{xk}$, while $E_{x,v}[(R^t-\Sigma d_k u_{xk})^2]$, and $E_{x,v,s}[(\hat{R}_c^t-R^t)(R^t-\Sigma d_k u_{xk})]$ are both nearly zero under A1 and A2.

One can now employ all three assumptions to determine the model-expected sampling variance of the combined ratio estimator:

$$E_{x,v,u}\,Var_s(\hat{R}_c^t)\doteq\Sigma d_k^2(1+g_{v*}^2)g_x^2/2 + \Sigma d_k^2 g_{v*}^2 g_T^2/2. \qquad (6.2)$$

An arbitary half sample estimator is nearly model-unbiased under Assumptions A1-A3. The average of orthogonally balanced half samples produces a variance estimator with the least model variance among all possible averages of half sample estimators:

$$\hat{Var}_0(\hat{R}_c^t|\cdot)\doteq\Sigma d_k^2[v_{k1}^*(x_{k1}-\Sigma d_k u_{xk}) - v_{k2}^*(x_{k2}-\Sigma d_k u_{xk})]^2/4. \qquad (6.3)$$

There is a more direct way to estimate the sampling variance of the combined ratio estimator that is nearly model-unbiased under A1-A3 and has as little model variance as $\hat{Var}_0(\cdot)$. Replace $\Sigma d_k u_{vk}$ in (6.3) by $\hat{R}_c^t$, the v_{ki}^* by v_{ki}/u_{vk}, and the d_k by $D_k u_{vk}/[\Sigma D_k(v_{k1}+v_{k2})/2]$:

$$\hat{Var}_D(\hat{R}_c^t|\cdot)=(\Sigma D_k^2[v_{k1}(x_{k1}-\hat{R}_c^t) - v_{k2}(x_{k2}-\hat{R}_c^t)]^2/4) \times (\Sigma D_k[v_{k1}+v_{k2}/2])^{-2}. \qquad (6.4)$$

The direct variance estimator in (6.4) can be shown to be identical to what Krewski and Rao call the "linearization estimator" and what Frankel calls the "Taylor series estimator." The near model-unbiasedness of this estimator depends very much on Assumption A3, as we shall see in the next subsection.

6.2 The General Case

In this subsection we will abandon Assumption A3 and discover that while arbitrary half sample estimators and balanced half sample estimators remain nearly model-unbiased when the strata shares are equal, the direct variance estimator in (6.4) is biased downward.

After some tedious calculations, one can express the model-expected sampling variance of $\hat{R}_c^t$ as

$$E_{x,v,u}Var_s(\hat{R}_c^t)\doteq[\Sigma d_k^2-2\Sigma d_k^3+(\Sigma d_k^2)^2]g_{v*}^2 g_T^2/2 + [\Sigma d_k^2/2(1+g_{v*}^2)-\Sigma d_k^3 g_{v*}^2+3/4(\Sigma d_k^2)^2 g_{v*}^2]g_x^2.$$

Also after some work, it can be shown that

$$E_{x,v,u}\,\hat{Var}_A(\hat{R}_c^t|\cdot)\doteq E_{x,v,u}Var_s(\hat{R}_c^t) + ((\Sigma d_k^2)^2-\Sigma d_k^3)g_x^2 g_{v*}^2. \qquad (6.5)$$

From (6.5) and Assumption A1, we can conclude that an arbitrary half sample variance estimator is nearly model--unbiased (model-biased downward) if all the strata shares, the b_k, are (not) equal.

Under Assumptions A1-A3, the average of orthogonally balanced half sample estimators minimizes the model variance. When A3 is dropped, however, full balance produces an estimator with a smaller model variance.

It remains to show that $\hat{Var}_D(\cdot)$ is not model-unbiased. In fact,

$$E_{x,v,u}\,\hat{Var}_D(\hat{R}_c^t|\cdot)\doteq E_{x,v,u}Var_s(\hat{R}_c^t) + -[4\Sigma d_k^3-(\Sigma d_k^2)^2]g_{v*}^2 g_x^2/4].$$

This sampling variance estimator is model-biased downward even when all the strata shares are equal.

6.3 The Krewski-Rao Result

Krewski and Rao (1979 and 1981) restrict their attention to simple random sampling with replacement. In addition, they assume that the v_{ki} have Gamma distributions, and the x_{ki} have variances of $g_T^2+g_x^2 v_{ki}^{t-2}$, $0\leq t\leq 2$. Consequently, the v_{ki} and the x_{ki} are not independent when $t\neq 2$. Constrast this with our analysis, in which the x_{ki} and the v_{ki} are always assumed to be independent. Krewski and Rao's exact results conicide with our approximations only when the x_{ki} have variances equal to $g_x^2+g_T^2$ and the strata shares are all equal.

In this paper, we made a number of assumptions about the x_{ki} and the v_{ki} in order to simplify the calculations and the terminology. The greater model bias (or model-expected sampling bias) and increased model efficiency of the direct variance estimator relative to the fully balanced variance estimator do not depend on most of these assumptions, only on the independence of the x_{ki} and v_{ki}. The relative bias and efficiency of the two variance estimators are not invariant under Krewski and Rao's model, however, because of a hypothesized dependency between the v_{ki} and x_{ki}. This relationship between the random variables, while reasonable for many applications, is inappropriate for the study of price index estimators.

7. SUMMARY

Let us review the major results of this paper:

1. Under appropriate superpopulation models, half sample techniques produce nearly model-unbiased estimators of the sampling variance of price indexes; "nearly," because these estimators are subject to finite population correction.
2. The finite population correction factor depends on the stipulations of the superpopulation model. Even under the simplest of models, the factor is ambiguous in the presence of subsampling. When samples are drawn with suboptimal measures of size, it is convenient to ignore finite population correction entirely.
3. For most price relative (index) estimates, it is possible to produce a model efficient, nearly model-unbiased sampling variance estimator directly without using an orthogonal matrix. The exception is when the stratum expenditure shares are unknown, and either the number of strata is small or a few stratum shares are relatively large (i.e., when Σb_k^2 can not be ignored).
4. If balanced sampling is necessary to produce a variance estimator that is less model-biased than the direct estimator, then full balance, where each sampled unit is in half of the half samples is more efficient than simple orthogonal balance.
5. The fully balanced half sample variance estimator is nearly model-unbiased for the combined ratio estimator when the strata expenditure shares are equal; otherwise, it has a tendency to be model-biased downward, but less so than the direct half sample variance estimator.

ACKNOWLEDGEMENTS

I would like to thank John Schwemberger, Michael P. Cohen, and Fritz Scheuren for their invaluable contributions to this endevor.

REFERENCES

FRANKEL, M. R. (1971), Inference from Survey Sampling: An Empirical Investigation, Ann Arbor: Institute of Social Research.

CASSEL, C. M., SARNDAL C. E., and WRETMAN, J. H. (1977), Foundations of Inference in Survey Sampling, New York: John Wiley and Sons.

GODAMBE, V. P. (1955), "A Unified Theory of Sampling from Finite Populations," Journal of the Royal Statistical Society, B 17, 269-78.

HORVITZ, D. G. and THOMPSON, D. J. (1952), "A Generalization of Sampling without Replacement from a Finite Universe," Journal of the American Statistical Association, 47, 663-85.

KISH, L. and FRANKEL, M. R. (1974), "Inference from Complex Samples," Journal of the Royal Statistical Society, B 36, 1-37.

KREWSKI, D. and RAO, J. N. K. (1979), "Small Sample Properties of the Linearization, Jacknife and Balanced Half-Sample Methods for Ratio Estimation in Stratified Samples," American Statistical Association, Proceedings of the Section on Survey Research Methods.

____ (1981), "Inference from Stratified Samples: Properties of the Linearization, Jacknife, and Balanced Repeated Repli-cation Methods," Annals of Statistics, 9, 1010-1019.

LEMESHOW, S. and LEVY, P. (1978), "Estimating the Variance of Ratio Estimates in Complex Samples Surveys with Two Primary Units Per Stratum - a Comparison of Balanced Replication and Jacknife Techniques," Journal of Statistical Computation and Simulation, 8, 191-205.

MCCARTHY, P. J. (1966), Replication: An Approach to the Analysis of Data from Complex Surveys, Wahsington DC: National Center of Health Statistics, Series 2, No. 14.

___ (1969), Pseudoreplication - Further Evaluation and Applications of the Balanced Half-Sample Technique, National Center for Health Statistics, Series 2 No. 31.

RAJ, D. (1968), Sampling Theory, New York: McGraw-Hill Book Company.

SCOTT, A. and SMITH, T. M. F. (1969), "Estimation in Multi-Stage Surveys," Journal of the American Statistical Association, 64, 830-40.