Journal of Official Statistics
Vol. 3, No. 3, 1987, pp. 251 – 265

Estimating the Variance of a Complex Statistic: A Monte Carlo Study of Some Approximate Techniques

Claes Andersson,[1] Gösta Forsman, and Jan Wretman[2]

Abstract: This paper describes a Monte Carlo study designed to illustrate the performance of four approximate techniques of estimating the sampling variance of a ratio between two estimated consumer price indices. The variance estimation techniques under study are: (1) traditional Taylor linearization, (2) Taylor linearization with a simplified variance estimation formula, (3) repeated random groups, and (4) jackknifing. These techniques are compared on the basis of observed relative bias, observed relative mean square error, and observed confidence interval coverage rate in a series of 1 000 samples drawn from the same population of stores. The sampling design is stratified sampling with PPS sampling within each stratum, and the data that are used are authentic price data from a price measurement study.

Key words: Finite population sampling; variance estimation; complex statistic; consumer price index; Taylor linearization; repeated random groups; jackknife; Monte Carlo study.

1. Introduction

This paper is a report from a Monte Carlo study designed to illustrate the performance of four approximate techniques of estimating the sampling variance of the ratio of two estimated consumer price indices.

The variance estimation problem has the following background: retail price data for the Swedish consumer price index have traditionally been collected through direct observation of *current prices* of specified commodities in a sample of stores. Recently, it was suggested to use *list prices* instead of current prices. The list prices are a sort of "recommended" retail prices decided by the retail organization with which the store is associated. They can be easily obtained from these organizations. This would make it unnecessary to visit the stores and would thus reduce the data collection costs. But the list prices do not exactly agree with the current prices, and the question is whether an index based on list prices can be used as an approximation of an index based on current prices.

We conduct a methodological study to compare these two indices. In our study, both list and current prices are collected for the same set of commodities in the same sample of stores and we calculate two indices. The comparison should then be based on the ratio between these two indices. Since data is collected only from a sample of stores, it becomes important to know the impact of sampling variability on the index ratio.

[1] Senior Statistician, Statistical Research Unit, Statistics Sweden, S–701 89 Örebro, Sweden.

[2] Senior Statisticians, Statistical Research Unit, Statistics Sweden, S–115 81 Stockholm, Sweden.

Hence, the problem of how to estimate the sampling variance of a ratio between two indices emerges. We conducted our Monte Carlo study in order to illustrate this problem.

Since the index ratio is a complicated statistic, there are no simple, exact estimates of its variance. We considered four variance estimation techniques usually referred to as "approximate." Lacking theoretical comparisons between these techniques, we performed a simulation study in which 1 000 independent samples were drawn from a miniature population of 99 stores, for which a complete set of price data (both list and current prices) was available. Computing the four types of variance estimates for each sample made it possible to get an idea of the sampling distributions of the variance estimators and their usefulness for constructing confidence intervals.

2. The Estimation Problem

We describe the estimation problem in standard sampling terminology, rather than index number terminology. Let $U = \{1,\dots, k,\dots, N\}$ be a finite population of N stores. A consumer price index is to be based on a group of p specified commodities. For each store ($k = 1,\dots, N$) and each item ($i = 1,\dots, p$), the following variables are defined:

y_{ki} = (current retail price of item i, in store k, at time t=1) × (total turnover of all items in store k)

x_{ki} = (current retail price of item i, in store k, at time t=0) × (total turnover of all items in store k)

y'_{ki} = (list price of item i, in store k, at time t=1) × (total turnover of all items in store k)

x'_{ki} = (list price of item i, in store k, at time t=0) × (total turnover of all items in store k)

If a store does not carry an item, the corresponding variables are defined as having a value of zero. Population totals are denoted

$$Y_i = \sum_{k=1}^{N} y_{ki},$$

etc. Actually, the definitions of the variables above are somewhat idealized. In practice, the total turnover of an individual store is unknown, and the number of employees in the store is used instead (assuming the turnover to be roughly proportional to the number of employees).

The consumer price index, I, using current prices, is defined in this study as

$$I = \sum_{i=1}^{p} a_i Y_i / X_i,$$

where a_i is the turnover of item i for all stores in the population. The consumer price index, I', using list prices is similarly defined as

$$I' = \sum_{i=1}^{p} a_i Y'_i / X'_i.$$

The index ratio $R = (I/I')100$ will be estimated by

$$\hat{R} = (\hat{I}/\hat{I}')\,100 =$$

$$= \{(\sum_{i=1}^{p} a_i \hat{Y}_i/\hat{X}_i)/(\sum_{i=1}^{p} a_i \hat{Y}'_i/\hat{X}'_i)\}\,100, \quad (2.1)$$

where $\hat{Y}_i$ and so on denote Horvitz-Thompson (1952) estimators of the corresponding population totals Y_i, etc. Our problem is to estimate $V(\hat{R})$, the sampling variance of $\hat{R}$. The sampling design considered is stratified sampling with PPS sampling without replacement in each stratum. This gives

$$\hat{Y}_i = \sum_{h=1}^{H} \sum_{k \varepsilon s_h} y_{ki}/\pi_k,$$

where s_h is the sample of stores from stratum h, H is the number of strata, and π_k is the inclusion probability of store k.

We can summarize our view of the inference problem as follows. We consider a fixed population of stores and a fixed set of commodities. The index ratio R is a population characteristic tied to this population. The statistic $\hat{R}$ is an estimator of R, based on data from a probability sample of stores, and $V(\hat{R})$ is the sampling variance of $\hat{R}$ in repeated sampling of stores from the same population by the same sampling design. The problem is to estimate $V(\hat{R})$ using data from a single sample.

3. Four Methods of Variance Estimation

Since $\hat{R}$ is a complex statistic (a ratio between two weighted sums of ratios), it is not possible to estimate $V(\hat{R})$ using traditional methods of unbiased variance estimation. Some approximate variance estimation technique must be employed. We considered the following four methods. Although the theoretical support is not very strong, they seem to be used frequently in survey practice.

The *first method* uses a first-order *Taylor approximation* technique, as described by Tepping (1968) and Woodruff (1971), in combination with a variance estimation formula adapted to without replacement sampling, which gives a variance estimator denoted $\hat{V}_{T1}$. Considering $\hat{R}$ as a function of the Horvitz-Thompson estimators ($\hat{Y}_i$ etc.), expanding around their expected values (Y_i etc.) and changing the order of summation, as suggested by Woodruff (1971), we arrive at the approximation

$$\hat{R} \doteq R + 100 \left(\sum_{i=1}^{p} a_i Y'_i/X'_i\right)^{-1} \hat{D}. \qquad (3.1)$$

Where

$$\hat{D} = \sum_{h=1}^{H} \sum_{k \varepsilon s_h} d_k/\pi_k,$$

is the Horvitz-Thompson estimator of the population total

$$D = \sum_{k=1}^{N} d_k,$$

and

$$d_k = \sum_{i=1}^{p} a_i \left[X_i^{-1} \{y_{ki} - (Y_i/X_i)x_{ki}\} - R\, X_i'^{-1} \{y'_{ki} - (Y'_i/X'_i)x'_{ki}\}\right].$$

Assuming that $\hat{R}$ behaves approximately like the random variable (3.1), we get the following approximation to the variance of $\hat{R}$

$$V(\hat{R}) \doteq 100^2 \left(\sum_{i=1}^{p} a_i Y'_i/X'_i\right)^{-2} V(\hat{D}).$$

This is estimated, using the variance estimation formula given by Yates and Grundy (1953) and Sen (1953), and inserting estimated values for unknown population quantities, by

$$\hat{V}_{T1} = 100^2 \left(\sum_{i=1}^{p} a_i \hat{Y}'_i/\hat{X}'_i\right)^{-2} \times \sum_{h=1}^{H} \sum_{\substack{k \varepsilon s_h \\ l \varepsilon s_h \\ k<l}} \sum \left(\frac{\pi_k \pi_l}{\pi_{kl}} - 1\right) \left(\frac{d_k^*}{\pi_k} - \frac{d_l^*}{\pi_l}\right)^2, \qquad (3.2)$$

where

$$d_k^* = \sum_{i=1}^{p} a_i \left[\hat{X}_i^{-1} \{y_{ki} - (\hat{Y}_i / \hat{X}_i) x_{ki}\} - \hat{R}\, \hat{X}_i'^{-1} \{y'_{ki} - (\hat{Y}'_i / \hat{X}'_i) x'_{ki}\}\right].$$

The *second method* uses the same Taylor approximation as above, but with a simplified variance estimation formula that would have been appropriate if the sample had been drawn by PPS sampling *with replacement* in each stratum. To obtain this variance estimator, denoted $\hat{V}_{T2}$, the expression (3.1) is rewritten as

$$\hat{R} \doteq R + 100\, (\sum_{i=1}^{p} a_i\, Y'_i / X'_i)^{-1} \sum_{h=1}^{H} \bar{u}_{s_h},$$

where

$$\bar{u}_{s_h} = \sum_{k \varepsilon s_h} u_k / n_h ,$$

and

$$u_k = n_h d_k / \pi_k , \text{ for } k \varepsilon s_h.$$

Proceeding as if $\bar{u}_{s_h}$ were a Hansen-Hurwitz estimator (associated with Hansen and Hurwitz (1943)) with $p_k = \pi_k / n_h$ for each $k \varepsilon s_h$ and inserting estimated values for unknown population quantities, we arrive at the estimation formula

$$\hat{V}_{T2} = 100^2\, (\sum_{i=1}^{p} a_i\, \hat{Y}'_i / \hat{X}'_i)^{-2} \times \sum_{h=1}^{H} \sum_{k \varepsilon s_h} (u_k^* - \bar{u}^*_{s_h})^2 / n_h\, (n_h - 1), \quad (3.3)$$

where

$$\bar{u}^*_{s_h} = \sum_{k \varepsilon s_h} u_k^* / n_h ,$$

and

$$u_k^* = n_h\, d_k^* / \pi_k , \text{ for each } k \varepsilon s_h.$$

The *third method* uses *repeated random grouping,* as suggested by Norlén and Waller (1979). This involves a series of independent repetitions of a random groups procedure (see, for instance, Hansen et al. (1953)). The resulting variance estimator is denoted $\hat{V}_{RRG}$.

The random grouping is as follows. The sample of stores is divided into G non-overlapping groups so that each group contains stores from all strata in (roughly) the same proportions as in the full sample. For each group g ($g = 1, \ldots, G$), an estimate $\hat{R}_g$ of R is calculated analogously to $\hat{R}$ but using data only from that particular group. We thus obtain $\hat{R}_1, \ldots, \hat{R}_G$ from which we then compute

$$\bar{R} = \sum_{g=1}^{G} \hat{R}_g / G ,$$

and

$$\hat{V}^* = \sum_{g=1}^{G} (\hat{R}_g - \bar{R})^2 / G(G-1) .$$

We consider $\hat{V}^*$ as a first approximate estimator of $V(\hat{R})$ (although, properly speaking, it should be considered an estimator of $V(\bar{R})$).

The random grouping procedure just described is then independently repeated M times. We thus obtain the sequences $\bar{R}_1, \ldots, \bar{R}_M$ and $\hat{V}_1^*, \ldots, \hat{V}_M^*$, which are then averaged to produce $\hat{R}_{RRG} = \sum_{m=1}^{M} \bar{R}_m / M$, which is an estimator of R (not identical to $\hat{R}$), and

$$\hat{V}_{RRG} = \sum_{m=1}^{M} \hat{V}_m^* / M, \quad (3.4)$$

which is the final variance estimator obtained by this repeated random groups procedure. The idea behind the repetition is, of course, that a presumed lack of stability of $\hat{V}^*$ will be remedied by taking the average of a number of independent replicates.

The *fourth method* is based on a *jackknife* procedure, and gives the variance estimator $\hat{V}_{JK}$. One store at a time is deleted from the sample. Each time an estimate $\hat{R}^{(j)}$ of R is calculated, analogous to $\hat{R}$, but based only on data from the remaining n–1 stores (j=1,..., n). A jackknife estimator of R is then

$$\hat{R}_{JK} = n\,\hat{R} - (n-1)\,\hat{R}^{(\cdot)},$$

where $\hat{R}^{(\cdot)} = \sum_{j=1}^{n} \hat{R}^{(j)}/n$.

The variance estimator produced by the usual jackknife formula is

$$\hat{V}_{JK} = (n-1) \sum_{j=1}^{n} (\hat{R}^{(j)} - \hat{R}^{(\cdot)})^2/n, \qquad (3.5)$$

which we consider as an approximate estimator of $V(\hat{R})$.

4. The Design of the Monte Carlo Study

A Monte Carlo study was designed to illustrate the performance of the four variance estimators $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG}$, and $\hat{V}_{JK}$ described in Section 3. We had at our disposal a miniature population of 99 stores (53 department stores and 46 other stores). For these stores, both current and list prices had been collected with respect to a set of 159 commodities (food and alcoholic beverages) in January 1982 (t = 0), and March 1982 (t = 1). Not all 159 commodities were found in each store, though, mainly because the small stores did not keep as wide a variety of goods as the larger ones. A few commodities were found in less than five stores. The commodities were divided into 10 groups. Table 1 shows the number of commodities in each group as well as the number of these commodities that were actually found in the stores.

Table 1. Commodity Groups and Number of Commodities in the Stores in the Population

Commodity group number	Type of goods	Total number of commodities in the group, by definition	Number of commodities actually found. Average per store
1	Flour, grain and bread	20	16.7
2	Meat	17	7.8
3	Fish and canned fish	16	14.1
4	Milk, cheese and eggs	25	14.1
5	Household fats	14	11.4
6	Roots, vegetables and fruits	21	14.4
7	Coffee, tea and cocoa	12	8.8
8	Other foods	17	10.1
9	Soft drinks and light beer	9	7.3
10	Alcoholic beverages	8	4.2
All groups		159	108.9

From the population of 99 stores, 1000 independent samples of 32 stores were drawn by stratified sampling using PPS sampling without replacement in each stratum[3]. The inclusion probabilities were proportional to the number of employees in the store.

The stratum sizes and the number of sampled stores from each stratum are given in Table 2, where the corresponding data for the regular consumer price index survey are also given. (The regular survey uses five strata of stores, two of which correspond to the two strata used in the Monte Carlo study.)

Table 2. Stratum Size and Number of Sampled Stores

Stratum		Number of stores			
		Monte Carlo study		Regular survey	
		Population	Sample	Population	Sample
1	Department stores	53	10	192	10
2	Other stores	46	22	1 511	22

The sample sizes for the two strata in the Monte Carlo study are the same as in the regular survey. The strata themselves, however, are much smaller, and the relation between stratum sizes is not the same. These facts restrict the possibility to generalize the findings of the Monte Carlo study to a regular survey.

Each sample was obtained by a sequential sampling scheme suggested by Sunter (1977), such that:

i) the sample size, n, is fixed;
ii) each unit in the population has an inclusion probability proportional to its size measure z_k, that is, $\pi_k = n\, z_k / Z$ (where $Z = \sum_{l=1}^{N} z_l$), except for very large units ($z_k > Z/n$), which are selected with probability 1, and very small units, which are assigned a revised common size measure, equal to their average size, and hence become selected with equal probability;
iii) $\pi_{kl} > 0$ for all k and l;
iv) $\pi_k \pi_l - \pi_{kl} \geqslant 0$ for all k and l, ensuring nonnegative variance estimates.

Before sample selection, the stores in the sampling frame were ordered by decreasing size within each stratum.

In the first stratum (with 53 stores), the 15 smallest stores were assigned equal probability, according to Sunter's proposal. The inclusion probabilities in this stratum varied between 0.067 and 0.947.

In the second stratum (with 46 stores), the size distribution was rather skew. The six largest stores were selected with probability 1. The inclusion probabilities of the remaining 40 stores varied between 0.285 and 0.895, with the 29 smallest stores being assigned equal probability.

When the repeated random groups method was used, each sample (of 32 stores) was randomly divided into four groups of equal size (8 stores). These four groups were composed of stores from the two strata as shown in Table 3.

[3] The simulations were carried out on an IBM 360/370. We used the multiplicative-congruential random number algorithm proposed by Lehmer (1951). The algorithm uses modulus 2^{31} with multiplier 7^5 recommended by Lewis et al. (1969) for the IBM 360/370.

Table 3. Size of Subsamples (= Random Groups) with the Random Grouping Technique

Stratum	Subsample no.				Total
	1	2	3	4	
1	3	3	2	2	10
2	5	5	6	6	22
Total	8	8	8	8	32

To study the effect of the number of repetitions we tried both 10 and 20 repetitions. For one of the commodity groups (group 3), we also tried division of the sample into three groups and five groups, in order to illustrate the effect of group size.

For each of the 1 000 samples, and for each group of goods separately, we calculated:

- Estimates of R, given by $\hat{R}$, $\hat{R}_{RRG}$, and $\hat{R}_{JK}$.
- Estimated variances: $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG}$, and $\hat{V}_{JK}$.
- Confidence intervals: $\hat{R} \pm 1.96\ \hat{V}_j^{1/2}$ (j = T1, T2, RRG, JK).

Summary statistics for the 1 000 samples were then calculated, the most important ones being:

- Bias and variance of the 1 000 observed values of $\hat{R}$.
- Relative bias and relative mean square error of the 1 000 observed values of $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG}$, and $\hat{V}_{JK}$.
- Coverage rate for each series of 1 000 confidence intervals.

These summary statistics were calculated as follows. For a particular group of goods, let the 1 000 realized values of the estimator $\hat{R}$ be denoted

$$\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_{1\,000}.$$

The variance, denoted V, of the observed $\hat{R}$-values was computed as:

$$V = \frac{1}{1\,000} \sum_{l=1}^{1\,000} (\hat{R}_l - \bar{\hat{R}})^2,$$

where $\bar{\hat{R}} = \sum_{l=1}^{1\,000} \hat{R}_l / 1\,000.$

We consider this variance an approximation of the true variance, $V(\hat{R})$, and it will serve as a target value for the different types of variance estimators under study.

Let

$$\hat{V}_1, \hat{V}_2, \ldots, \hat{V}_{1\,000},$$

be the 1 000 realized values of one particular variance estimator $\hat{V}$ (which can be $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG}$, or $\hat{V}_{JK}$). Then, the *observed relative bias* of $\hat{V}$ is defined as

$$\frac{1}{1\,000} \sum_{l=1}^{1\,000} (\hat{V}_l - V)/V,$$

and the *observed relative* MSE of $\hat{V}$ as

$$\frac{1}{1\,000} \sum_{l=1}^{1\,000} (\hat{V}_l - V)^2/V^2.$$

The *observed coverage rate* of confidence intervals is defined as the percentage of the

1 000 intervals

$$\hat{R}_l \pm 1.96\, \hat{V}_l^{1/2} \; (l = 1, \ldots, 1\,000) \,,$$

which actually cover the true value R.

5. Results of the Monte Carlo Study

We first look at some results concerning the estimation of R. The *true values* of R for the ten groups of commodities are given in Table 4, together with the *observed bias* of the 1 000 realizations of the estimators $\hat{R}$, $\hat{R}_{RRG(10)}$, $\hat{R}_{RRG(20)}$, and $\hat{R}_{JK}$, where $\hat{R}_{RRG(10)}$ and $\hat{R}_{RRG(20)}$ denote RRG estimators based on 10 and 20 repetitions. The observed bias is defined as the difference between the mean of the 1 000 obtained estimates and the true value of R. Throughout, we consider $\hat{R}$ the principal estimator of R. The biases of the RRG and JK estimators are shown only for the sake of comparison, and because we obtained their values automatically as a by-product of the variance estimation process.

Table 4. Population Values of R, and Observed Bias of $\hat{R}$, $\hat{R}_{RRG(10)}$, $\hat{R}_{RRG(20)}$, and $\hat{R}_{JK}$ (Stratified Sampling; n=32; 1 000 Replicates)

Group of goods	Population value R	Observed bias of $\hat{R}$	$\hat{R}_{RRG(10)}$	$\hat{R}_{RRG(20)}$	$\hat{R}_{JK}$
1	100.26	.02	.10	.10	–.01
2	99.11	.01	.27	.28	–.08
3	98.53	.25	.71	.71	–.12
4	101.07	.00	–.06	–.06	.05
5	100.54	–.04	.12	.12	–.11
6	100.35	.08	.17	.17	.00
7	101.13	–.05	–.13	–.13	.04
8	99.91	–.23	.05	.05	–.34
9	99.49	.00	.03	.03	.00
10	100.22	–.14	–.03	–.03	–.04

With a few exceptions, the observed bias is smaller than 0.3 per cent of the population value. As for the RRG procedure, the bias in Table 4 is almost identical for 10 and 20 repetitions. Note that if all groups are of exactly equal size (which is not possible with the actual sample sizes), $\hat{R}_{RRG}$ will be identical with $\hat{R}$.

Group 4 is defined in a slightly different way in this study than in the regular survey. Using the original definition, we obtained surprising results for group 4. A closer examination revealed that these incongruities were due to a single commodity, appearing in only one store in the entire population. This store happened to have a high inclusion probability, about 0.4. The commodity had a relatively large weight a_i (see formula (2.1)) and a large ratio $(\hat{Y}_i/\hat{X}_i)/(\hat{Y}'_i/\hat{X}'_i) \doteq 1.2$. We then decided to redefine group 4 by excluding this rare commodity. The ensuing results were much more reasonable.

Fig. 1 below shows the *observed sampling distribution* of the 1 000 realized values of the estimators $\hat{R}$, $\hat{R}_{RRG(10)}$, $\hat{R}_{RRG(20)}$, and $\hat{R}_{JK}$ for commodity group 3.

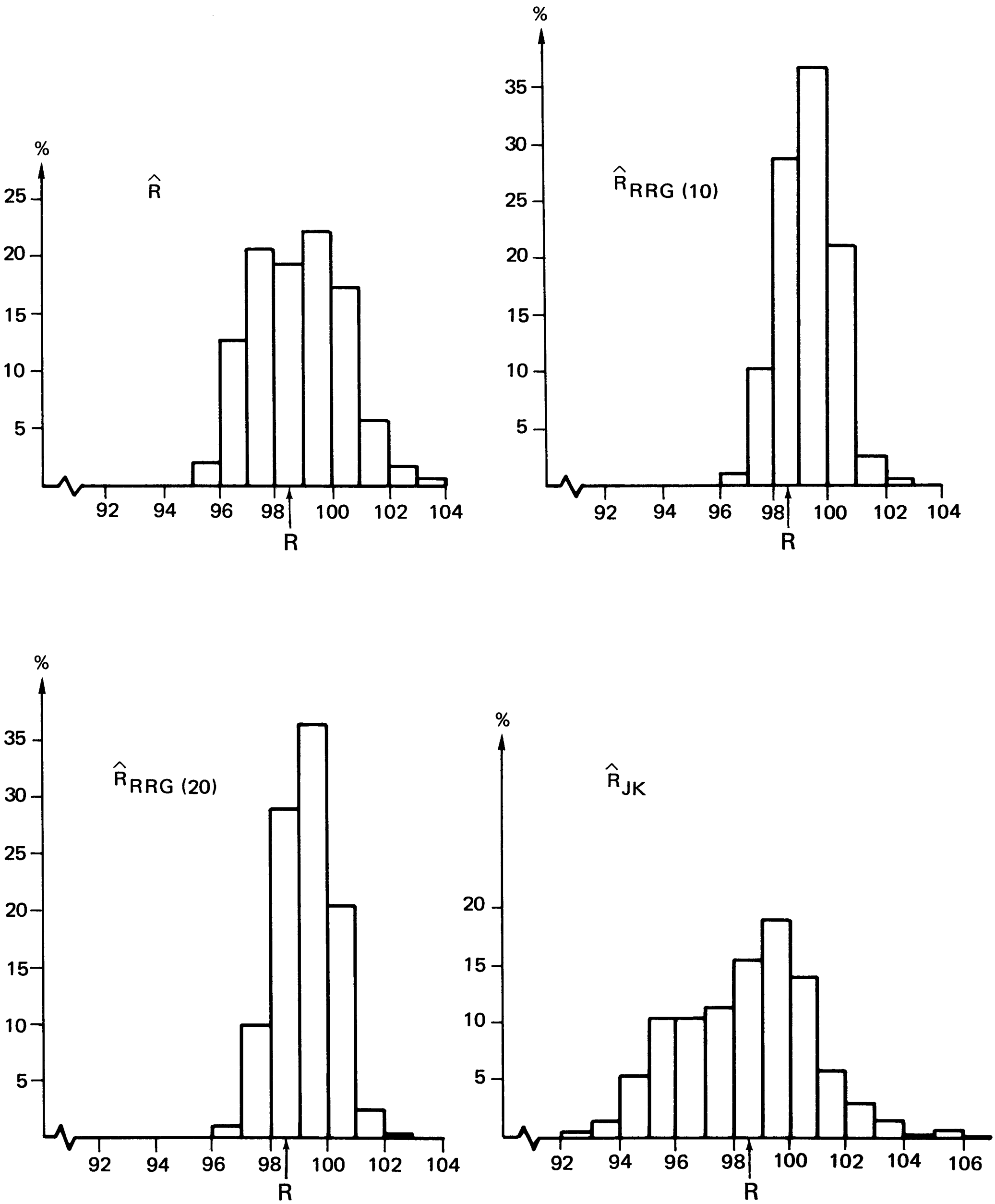

Fig. 1. Observed Sampling Distribution of $\hat{R}$, $\hat{R}_{\mathrm{RRG}(10)}$, $\hat{R}_{\mathrm{RRG}(20)}$, *and* $\hat{R}_{\mathrm{JK}}$ *in Commodity Group 3 (Stratified Sampling;* n=32; 1 000 *Replicates)*

Table 5. Observed Variances of $\hat{R}$, and Observed Relative Bias of Variance Estimates $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG(10)}$, $\hat{V}_{RRG(20)}$, *and* $\hat{V}_{JK}$ *(Stratified Sampling;* $n=32$; 1 000 *Replicates)*

Group of goods	Observed variance V	Observed relative bias of variance estimates				
		$\hat{V}_{T1}$	$\hat{V}_{T2}$	$\hat{V}_{RRG(10)}$	$\hat{V}_{RRG(20)}$	$\hat{V}_{JK}$
1	.13	–.06	.26	.52	.51	.36
2	1.67	–.22	.04	.46	.45	.50
3	2.38	–.63	–.47	–.32	–.32	.87
4	.11	–.15	.17	.58	.57	.96
5	.79	–.05	.27	.50	.51	.37
6	.64	–.13	.18	.31	.30	.50
7	.43	–.43	–.18	.05	.05	.61
8	.71	–.59	–.42	–.07	–.08	.90
9	1.80	–.02	.13	.20	.19	.24
10	.42	–.33	–.13	.26	.26	.60

Each sampling distribution is fairly symmetric. The jackknife procedure gives the most spread-out distribution, while that of the RRG is the most concentrated one. We notice that the sampling distributions of $\hat{R}_{RRG(10)}$ and $\hat{R}_{RRG(20)}$ are similar, which again indicates that using our data, there is no need for more than ten repetitions of the grouping procedure. The sampling distributions for the other groups of goods are not shown here, but the conclusions from Fig. 1 are valid for these other groups as well.

We now turn to the variance estimation results. The *observed relative bias* of the variance estimators are shown in Table 5.

Suppose that we have decided to estimate R by $\hat{R}$, and that we consider $\hat{V}j$ (j = T1, T2, RRG(10), RRG(20), JK) alternative estimators of the true variance of $\hat{R}$, V($\hat{R}$), which approximately equals V. Which of the suggested variance estimators is the best for this purpose? Table 5 indicates that the T2 estimator might be preferred because of its lower bias. The T1 method leads to downwardly biased and the JK to upwardly biased variance estimates. The RRG procedure overestimates the variance in 8 of the 10 groups of goods. We note that the number of repetitions in the RRG procedure (10 and 20) does not affect the bias of the variance estimate.

For one of the commodity groups, group 3, we also present the *observed sampling distribution* of the 1 000 observed values of $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG(10)}$, $\hat{V}_{RRG(20)}$, and $\hat{V}_{JK}$ in Fig. 2.

The observed sampling distributions of the variance estimates are all rather skew. The variance estimates produced by the jackknife procedure are seen to be much more widely dispersed than those produced by the other procedures.

Table 6 shows the *relative* MSE of the variance estimates $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG(10)}$, $\hat{V}_{RRG(20)}$, and $\hat{V}_{JK}$.

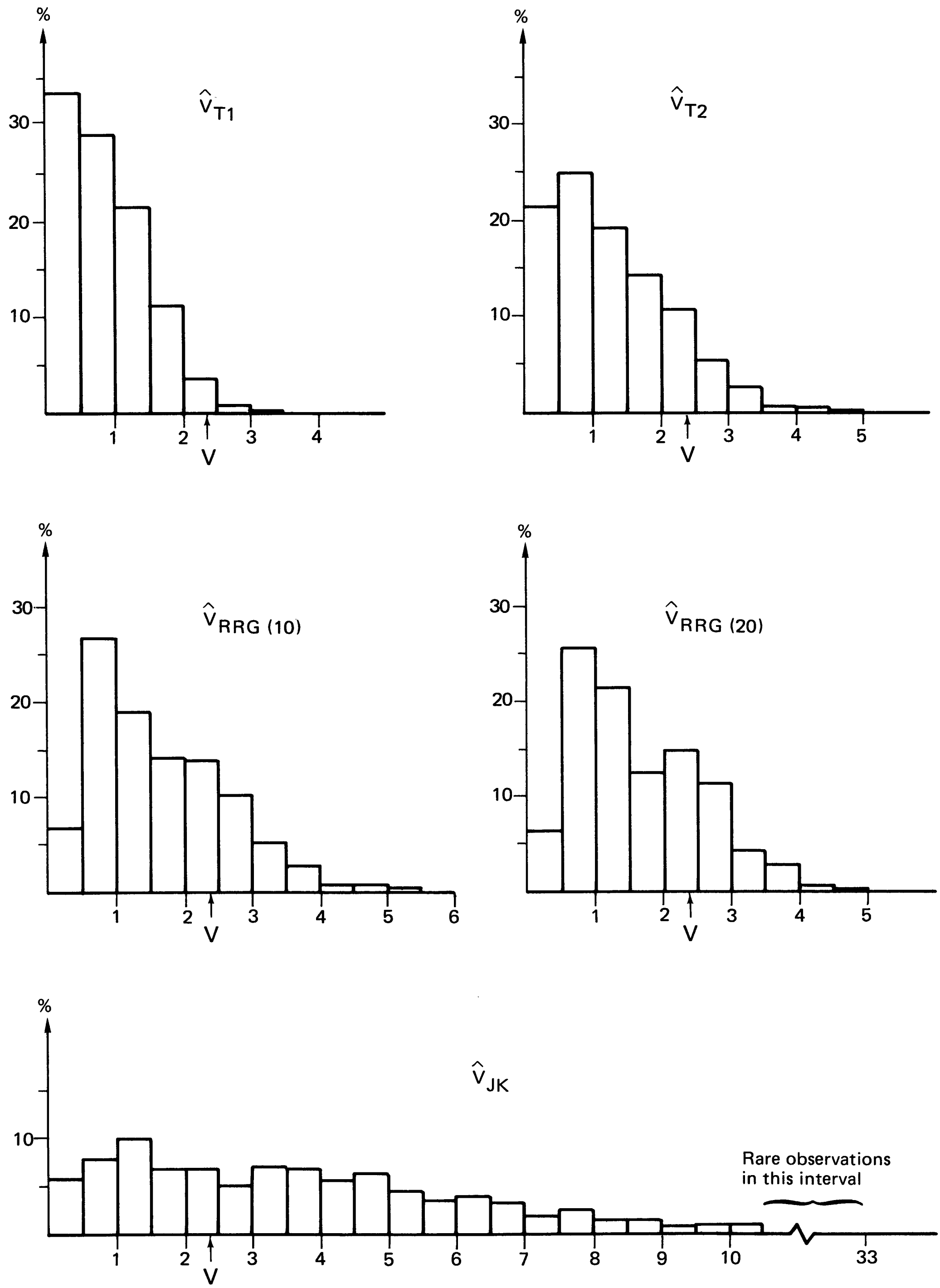

Fig. 2. Observed Sampling Distribution of Variance Estimates $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG(10)}$, $\hat{V}_{RRG(20)}$, *and* $\hat{V}_{JK}$ *in Commodity Group 3 (Stratified Sampling;* n=32; 1 000 *Replicates)*

Table 6. Observed Relative Mean Square Error (MSE) of Variance Estimates $\hat{V}_{T1}$, $\hat{V}_{T2}$, $\hat{V}_{RRG(10)}$, $\hat{V}_{RRG(20)}$, *and* $\hat{V}_{JK}$ *(Stratified Sampling;* $n=32$; 1 000 *Replicates)*

Group of goods	Observed relative MSE of variance estimates				
	$\hat{V}_{T1}$	$\hat{V}_{T2}$	$\hat{V}_{RRG(10)}$	$\hat{V}_{RRG(20)}$	$\hat{V}_{JK}$
1	.23	.46	.93	.86	.60
2	.28	.38	.83	.73	1.30
3	.46	.36	.26	.25	3.47
4	.28	.49	.90	.79	2.58
5	.34	.67	1.17	1.15	.82
6	.16	.24	.37	.32	.84
7	.34	.31	.31	.27	3.11
8	.46	.39	.23	.20	2.85
9	4.90	5.84	6.43	6.03	7.32
10	.33	.37	.65	.58	2.09

On the average for all groups of goods, the Taylor linearization methods obviously lead to estimates with the smallest relative MSE of the methods studied, while the JK method leads to the largest. The RRG estimates have the largest relative MSE in some groups and the smallest in other groups. Here, it is interesting to see that 20 repetitions give a slightly smaller MSE than 10 repetitions. This difference is almost entirely due to a difference in the variances of $\hat{V}_{RRG(10)}$ and $\hat{V}_{RRG(20)}$.

Table 7 shows the *observed coverage rate* of 1 000 confidence intervals of type $\hat{R} \pm 1.96\ \hat{V}^{1/2}$ for each variance estimation procedure. (Note that all confidence intervals are centered on the same point estimate, $\hat{R}$.)

Table 7. Observed Coverage Rate for 1 000 Confidence Intervals of Type $\hat{R} \pm 1.96\ \hat{V}_j^{1/2}$ (j = T1, T2, RRG(20), JK) *(Stratified Sampling;* $n=32$; 1 000 *Replicates)*

Group of goods	Per cent of 1 000 confidence intervals covering R, when V is estimated by			
	$\hat{V}_{T1}$	$\hat{V}_{T2}$	$\hat{V}_{RRG(20)}$	$\hat{V}_{JK}$
1	90.9	95.4	97.0	96.2
2	88.2	94.9	98.4	97.2
3	65.1	73.4	83.8	92.4
4	88.5	94.3	96.1	97.9
5	91.1	96.0	97.3	96.9
6	90.4	94.6	95.9	96.5
7	78.4	86.6	93.2	91.6
8	71.9	82.7	92.7	98.7
9	78.6	80.1	80.2	80.4
10	79.5	84.4	91.9	91.9

For intervals based on $\hat{V}_{T1}$ and $\hat{V}_{T2}$, the coverage rates throughout tend to be lower than the desired 95%. For intervals based on $\hat{V}_{RRG}$ and $\hat{V}_{JK}$, the coverage rates are lower than 95% in about half of the commodity groups, and higher in the other half.

For the RRG procedure, the question arose about how the behavior of the estimators was affected by the *number of random groups*. Hence, a special study was carried out using only data for commodity group 3. Each sample was now divided alternatively into three and five random groups, and estimates were calculated by the same procedures as before. The results are shown in Table 8, with the results already obtained in the main study, based on four random groups.

Table 8. Comparison of Results for Varying Number of Random Groups with the RRG(20) *Procedure in Commodity Group 3*

Number of random groups	Observed bias of $\hat{R}_{RRG(20)}$	Observed relative bias of variance estimate $\hat{V}_{RRG(20)}$	Observed relative MSE of variance estimate $\hat{V}_{RRG(20)}$	Observed coverage rate (%) of confidence interval $\hat{R} \pm 1.96\ \hat{V}^{1/2}_{RRG(20)}$
3	.62	−.16	.28	86.8
4	.71	−.32	.25	83.8
5	.77	−.41	.28	80.7

The observed bias of the 1 000 observed values of $\hat{R}_{RRG}$ is seen to increase as the number of random groups increases. At the same time, the observed relative bias of the 1 000 variance estimates $\hat{V}_{RRG(20)}$ is getting more and more negative, while the relative MSE remains roughly the same. The coverage rate of the confidence intervals is decreasing. This also holds true for RRG(10), although results for that case are not given here. Thus, with commodity group 3 data, it seems that the use of three random groups would yield somewhat better results than the use of four or five groups. Of course, this conclusion cannot be generalized to other commodity groups.

6. Discussion

We have in this paper studied four methods of estimating the variance of the estimator $\hat{R}$. As is always the case in this type of study, the conditions are unique and the findings cannot be generalized to the set of all survey designs, estimators, characteristics, and populations. Our study is, among other things, characterized by a small sample size and by an incomplete range of commodities in many stores. Despite these limitations, our findings show some similarities with the results of other Monte Carlo comparisons between variance estimation techniques concerning quite different estimators, populations, and sample sizes, as we shall see below.

For the discussion of general aspects on the variance estimation techniques, we shall focus on the accuracy of the variance estimates. Administrative considerations are less relevant in this case since the sample size is small and the processing costs are limited. In a large scale survey, however, the JK and the RRG methods would be expensive since they are based on repeated computations and the cost considerations would influence the choice of variance estimation method.

We have compared the variance estimates in terms of bias, mean square error, and confidence interval coverage probabilities.

The basic assumption for the Taylor linearization method is that terms beyond the linear one in the Taylor series expansion make negligible contribution to the variance of the estimator. For small sample sizes, however, this may not be the case, which can explain why the T1 method gave generally large underestimates of the actual variance in our study. Similar results have been reported from other Monte Carlo studies. In a study concerning the variance of the standard regression estimator of the finite population mean in six small populations and – actually – with sample size $n=32$, Deng and Wu (1984) found that the Taylor linearization estimators tended to be downward biased. Mulry and Wolter (1981) found in a study concerning correlation coefficients on consumer expenditure data for sample sizes $n=60$, 120, and 480 also a negative bias for Taylor linearization estimators, which became smaller as n increased.

That the T2 method tended to give greater variance estimates (on the average) than the T1 method seems to be in line with the theoretical results by Durbin (1953), that, under certain conditions, a variance estimation formula appropriate under sampling with replacement overestimates the variance if applied to sampling without replacement. Our conditions are, however, not exactly the same as those assumed by Durbin.

The overestimation of the Jackknife variance estimator that we found was also confirmed by the Deng and Wu and the Mulry and Wolter studies.

The magnitude of the bias of the RRG method seems to be related to the number of random groups. In Table 8, we saw that the bias increased when the number of random groups increased in commodity group 3. It is interesting to note that Wolter (1985) states that the same holds for the "Random Group" method (RG), in which the grouping is conducted only once (in our notation we could call the method RRG (1)). As the RG and RRG methods are closely related, it is certainly not surprising that they show similar properties.

Rust (1985) also discusses the RG method and explains the increase of the bias by "the larger the number of groups, the greater the departure from true replication." (By "true replication," Rust means the use of independent identically designed subsamples.)

In terms of the mean square error, the Taylor linearization methods generally performed best, while the Jackknife method gave variance estimates with much higher values. Similar conclusions were drawn from the Deng and Wu and the Mulry and Wolter data. In the latter study, the RG method had a mean square error relatively close to that of the Taylor linearization method, which in general holds also for the RRG and T1/T2 methods in our study.

From Table 7, we concluded that the confidence interval coverage rates were too low for the Taylor linearization methods while the coverage rates for the RRG and JK methods generally vary around 95%. Similar results for the Taylor and JK methods were found by Deng and Wu. Mulry and Wolter likewise found higher rates for the JK method. In their study, however, the JK rates were still too low.

If we consider these aspects of accuracy, the obvious conclusion is that the RRG method performed best. It is still somewhat unclear how this method should be designed optimally with respect to the number of groups (G) and the number of repetitions. If G increases, the bias increases, while if G decreases, the variance of the variance estimates increases. The confidence interval coverage rates also depend on G. Obviously, from Table 8, the confidence rates decrease as G increases in commodity group 3 (which can be explained by the increase in bias) and thus $G=3$ produces better coverage rates. For commodity groups 1, 2, and 5, on the other hand, it is likely that $G=5$ would be better than $G=4$, which according to Table 7 yields rates in these groups that are too high.

7. References

Deng, L.Y. and Wu, C.F. (1984): Estimation of Variance of the Regression Estimator. MRC Technical Summary Report No. 2758, University of Wisconsin, Madison.

Durbin, J. (1953): Some Results in Sampling Theory When the Units Are Selected with Unequal Probabilities. Journal of the Royal Statistical Society, B, Vol. 15, pp. 262–269.

Hansen, M.H. and Hurwitz, W.N. (1943): On the Theory of Sampling from Finite Populations. Annals of Mathematical Statistics, Vol. 14, pp. 333–362.

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

Horvitz, D.G. and Thompson, D.J. (1952): A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, Vol. 47, pp. 663–685.

Lehmer, D.H. (1951): Mathematical Methods in Large Scale Computing Units. Proceedings of the Second Symposium on Large Scale Digital Computing Machinery, Harvard University Press, Cambridge, pp. 141–149.

Lewis, P.A.W., Goodman, O.S., and Miller, J.W. (1969): A Pseudo Random Number Generator for the System 360, IBM Systems Journal, Vol. 8, pp. 136–146.

Mulry, M.H. and Wolter, K.M. (1981): The Effect of Fisher's Z-Transformation on Confidence Intervals for the Correlation Coefficient. American Statistical Association, Proceedings of the Section on Survey Research Methods, pp. 601–606.

Norlén, U. and Waller, T. (1979): Estimation in a Complex Survey – Experiences from a Survey of Buildings with Regard to Energy Usage. Statistical Review, Vol. 17, pp. 111–124.

Rust, K. (1985): Variance Estimation for Complex Estimators in Sample Surveys. Journal of Official Statistics, Vol. 1, pp. 381–397.

Sen, A.R. (1953): On the Estimate of the Variance in Sampling with Varying Probabilities. Journal of the Indian Society of Agricultural Statistics, Vol. 5, pp. 119–127.

Sunter, A. (1977): List Sequential Sampling with Equal or Unequal Probabilities without Replacement. Applied Statistics, Vol. 26, pp. 261– 268.

Tepping, B.J. (1968): Variance Estimation in Complex Surveys. American Statistical Association, Proceedings of the Social Statistics Section, pp. 11–18.

Wolter, K.M. (1985): Introduction to Variance Estimation. Springer, New York.

Woodruff, R.S. (1971): A Simple Method for Approximating the Variance of a Complicated Estimate. Journal of the American Statistical Association, Vol. 66, pp. 411– 414.

Yates, F. and Grundy, P.M. (1953): Selection Without Replacement from Within Strata with Probability Proportional to Size. Journal of the Royal Statistical Society, B, Vol. 15, pp. 235–261.

Received June 1985
Revised September 1987