Cross-Classified Sampling for the Consumer Price Index

Esbjörn Ohlsson

SCB

R&D Report
Statistics Sweden
Research - Methods - Development
1992:7

INLEDNING

TILL

R & D report : research, methods, development / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1988-2004. – Nr. 1988:1-2004:2.
Häri ingår Abstracts : sammanfattningar av metodrapporter från SCB med egen numrering.

Föregångare:

Metodinformation : preliminär rapport från Statistiska centralbyrån. – Stockholm : Statistiska centralbyrån. – 1984-1986. – Nr 1984:1-1986:8.

U/ADB / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1986-1987. – Nr E24-E26

R & D report : research, methods, development, U/STM / Statistics Sweden. – Stockholm : Statistiska centralbyrån, 1987. – Nr 29-41.

Efterföljare:

Research and development : methodology reports from Statistics Sweden. – Stockholm : Statistiska centralbyrån. – 2006-. – Nr 2006:1-.

R & D Report 1992:7. Cross-classified sampling for the consumer price index / Esbjörn Ohlsson.
Digitaliserad av Statistiska centralbyrån (SCB) 2016.

urn:nbn:se:scb-1992-X101OP9207

Cross-Classified Sampling for the Consumer Price Index

Esbjörn Ohlsson

R&D Report
Statistics Sweden
Research - Methods - Development
1992:7

Från trycket April 1992
Producent Statistiska centralbyrån, utvecklingsavdelningen
Ansvarig utgivare Åke Lönnqvist
Förfrågningar Esbjörn Ohlsson, 08/783 45 13 alt.
08/16 45 58

ISSN 0283-8680
Garnisonstryckeriet, Stockholm

Cross-Classified Sampling for the Consumer Price Index

by Esbjörn Ohlsson
Department of Enterprise Statistics
Statistics Sweden

ABSTRACT. The Swedish Consumer Price Index utilizes several two-dimensional samples, each of which is the cross-classification of a sample of outlets (shops, etc.) and items (products). Such a sampling procedure is called *Cross-classified sampling* in the paper. We are interested in the problem of deriving the variance of an estimator based on a cross-classified sample; in particular we want a variance formula for the CPI.

In the first part of the paper we give a general decomposition of the variance and some results which simplify variance calculations in cases with stratification and/or sampling with probabilities proportional to size. In the second part, we illustrate how the general results can be applied to derive a variance formula for the CPI.

Keywords: Cross-classified sampling, two-dimensional sampling, pps sampling, variance estimation, Consumer Price Index.

- CONTENTS -

Statistics Sweden
Dept. of Enterprise Statistics
Esbjörn Ohlsson

February 1992

CROSS-CLASSIFIED SAMPLING FOR THE CONSUMER PRICE INDEX

0. INTRODUCTION AND OUTLINE OF THE PAPER

Large parts of the Swedish Consumer Price Index (CPI) are based on price quotations from two-dimensional samples, each of which is the cross-classification of a sample of outlets (shops, restaurants, etc.) and a sample of items (products). In the sequel, such a procedure for sampling from a two-dimensional population will be called *Cross-Classified Sampling* (CCS). In this paper we first deal with the general problem of deriving the variance of an estimator based on a Cross-classified sample. The general results are then used to obtain a formula for the sampling variance of a "sub-index" of the CPI. Such formulas are important both for allocation purposes and for assessing the accuracy of the (estimated) CPI.

Many papers on two-dimensional sampling are concerned with the problem of sampling an area (plane sampling). Quenouille (1949) deals with two-dimensional samples which are obtained by using stratified simple random sampling or equal probability systematic sampling in each dimension at a time; variances are derived under a superpopulation model. In the terminology of Quenouille, CCS can be described as the case with *aligned* samples in both dimensions. For recent references extending the work of Quenouille, see Iachan (1982), who gives a review of papers on two-dimensional systematic sampling, and Ripley (1981), who discusses plane sampling. Other papers on two-dimensional sampling discuss how to sample among the cells in the crossing of two stratifications. This leads to "latin square" type samples, in which each one-dimensional population unit is represented exactly once; for an overview and references, see Cochran (1978). As far as we can see, the papers in the mentioned areas can not be used to solve the CCS variance problem.

In the particular case where the samples in both dimensions of the CCS procedure are drawn by simple random sampling without replacement, the variance of the sample mean can be obtained from the results in Vos (1964). To the best of our knowledge, variance formulas for CCS with more complicated one-dimensional sampling procedures, are not available in the literature.

The problem of estimating the sampling variance of a CPI has received quite some interest during the last years. An early reference is Banerjee (1956), who addresses the problem of optimal allocation of the item sample. Recent papers include Valliant (1991) on the United States CPI; the two-dimensional CPI sample in the US is, however, a two-stage sample and not a CCS. Balk & Kersten (1986) and Biggeri & Giommi (1987) use balanced half-samples and similar methods to estimate the variance due to the use of weights from a household expenditure survey. The (one-dimensional) variance due to the sampling of outlets in the Swedish CPI, conditioning on the item sample and making the samplifying assumption of *with* replacement sampling was computed by Andersson, Forsman & Wretman (1987). For further references on the problem of estimating the CPI variance, we refer to the mentioned papers. The connection between the CPI samples and the two-dimensional sampling procedures discussed in Vos (1964) was first noted by Dalén (1991a). This observation was the starting point for the present work.

The paper is divided into parts A and B. In part A we give some general results on the variance for estimators based on Cross-classified sampling: In section 1 we define Cross-classified sampling and give a basic decomposition of the variance. In section 2 we give a result which simplifies the calculations when we have stratified samples. In section 3 we discuss pps sampling and the Horvitz-Thompson estimator in the CCS case. In part B of the paper, consisting of sections 4-7, we show how the theoretical results of part A can be used to derive a variance formula for an index, such as a CPI. The type of index under consideration is defined in section 4. The variance formula is presented in section 5 and proved in section 6. Section 7 contains a simulation study on CPI data.

PART A.

GENERAL RESULTS ON THE VARIANCE IN CROSS-CLASSIFIED SAMPLING (CCS)

1. DEFINITION OF CCS AND A VARIANCE DECOMPOSITION

We consider a two-dimensional population with M·N units, arranged in a matrix with N rows and M columns. For each population unit we have a value of our target variable y, these values form the matrix $\{y_{ij};\ i=1,2,\ldots,N;\ j=1,2,\ldots,M\}$. A survey is carried out with the object of estimating the population total

$$Y = \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \ . \tag{1.1}$$

To this end, a random sample S^R of rows and a random sample S^C of columns are drawn. The sampling procedures for rows and columns are assumed to be independent; in all other respects S^R and S^C are arbitrary here.

Definition 1.1. A *cross-classified sample* S from a two-dimensional population, indexed by $\{(i,j)\}$ is the cross-classification of S^R and S^C, i.e. $S=\{(i,j):\ i\in S^R,\ j\in S^C\}$, were S^R and S^C are independent samples from the rows and columns, respectively.

On the basis of the y-values for the units in the sample S, we form some estimator $\hat{Y}$ of Y. The problem we shall focus on in this paper is the derivation of an explicit formula for $V(\hat{Y})$.

We shall now present a decomposition of $V(\hat{Y})$, which will be useful in the further derivations. Let E^R denote conditional expectation, given the outcome of S^R. By the independence of S^R and S^C, E^R is simply the expectation over column samples. Conditional expectation given S^C is denoted E^C, while E is overall expectation. For variances we analogously define V^R, V^C and V.

Let

$$\hat{Y}^R = E^R(\hat{Y}) \ , \qquad \hat{Y}^C = E^C(\hat{Y}) \ . \tag{1.2}$$

$\hat{Y}^R$ will be called the *row estimator* of Y. The reason for this is that $\hat{Y}^R$ depends solely on the outcome of the row sample: $\hat{Y}^R$ is an estimator which could be used if the columns were completely enumerated. Similarly, $\hat{Y}^C$ is the *column estimator* of Y. If $\hat{Y}$ is unbiased, then so are obviously both the row and column estimator.

Theorem 1.1. The variance of an estimator $\hat{Y}$, which is based on a cross-classified sample, can be decomposed as follows,

$$V(\hat{Y}) = VR + VC + VRC \ , \tag{1.3}$$

where

$$VR = V(\hat{Y}^R) \ , \quad VC = V(\hat{Y}^C) \ , \quad VRC = V(\hat{Y}-\hat{Y}^R-\hat{Y}^C). \tag{1.4}$$

VR will be called the *row variance*, VC the *column variance* and VRC the row and column *interaction variance.* Before proving the theorem we shall illustrate it in a simple example.

Example 1.1. Suppose that S^R is a simple random sample drawn without replacement (srswor) and having size n, and that S^C is an srswor of size m. Let

$$y_{i\cdot} = \sum_{j=1}^{M} y_{ij}, \qquad y_{\cdot j} = \sum_{i=1}^{N} y_{ij}, \qquad y_{\cdot\cdot} = Y = \sum_{i=1}^{M}\sum_{j=1}^{M} y_{ij}. \tag{1.5}$$

$$\bar{y}_{i\cdot} = y_{i\cdot}/M \ , \qquad \bar{y}_{\cdot j} = y_{\cdot j}/N \ , \qquad \bar{y}_{\cdot\cdot} = y_{\cdot\cdot}/(NM) \ . \tag{1.6}$$

The conventional unbiased estimator of Y is given by

$$\hat{Y} = \frac{NM}{nm} \sum_{i\in S^R}\sum_{j\in S^C} y_{ij} \ . \tag{1.7}$$

In this case the row and column estimators are readily found as

$$\hat{Y}^R = \frac{N}{n} \sum_{i\in S^R} y_{i\cdot} \ , \qquad \hat{Y}^C = \frac{M}{m} \sum_{j\in S^C} y_{\cdot j} \ . \tag{1.8}$$

Furthermore, let

$$\sigma_R^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\bar{y}_{i\cdot}-\bar{y}_{\cdot\cdot})^2, \qquad \sigma_C^2 = \frac{1}{M-1} \sum_{j=1}^{M} (\bar{y}_{\cdot j}-\bar{y}_{\cdot\cdot})^2,$$

$$\sigma_{RC}^2 = \frac{1}{N-1}\frac{1}{M-1} \sum_{i=1}^{N}\sum_{j=1}^{M} (y_{ij}-\bar{y}_{i\cdot}-\bar{y}_{\cdot j}+\bar{y}_{\cdot\cdot})^2. \tag{1.9}$$

The variance of $\hat{Y}$ in (1.7) is given by (1.3) and

$$VR = \frac{N^2M^2}{n}\left(1 - \frac{n}{N}\right)\sigma_R^2 , \tag{1.10}$$

$$VC = \frac{N^2M^2}{m}\left(1 - \frac{m}{M}\right)\sigma_C^2 , \tag{1.11}$$

$$VRC = \frac{N^2M^2}{nm}\left(1 - \frac{n}{N}\right)\left(1 - \frac{m}{M}\right)\sigma_{RC}^2 . \tag{1.12}$$

Note the similarity between the variances in (1.9) and the mean sums of squares in a two-way *analysis of variance* table. This is one reason for using the terminology "row, column and interaction variance".

In section 3 it will be indicated how (1.10)-(1.12) can be derived from Theorem 3.1 there. The expression for $V(\hat{Y})$ in the srswor case can alternatively be derived by using the results in Vos (1964, "Method C.1"). Note that Vos (1964) can not be used to cover more general sampling procedures S^R and S^C; the need for a generalization from srswor to pps and stratified samples was the starting point for the work reported here. ■

Proof of Theorem 1.1. Let $Z = \hat{Y}-\hat{Y}^R-\hat{Y}^C$, so that

$$\hat{Y} = \hat{Y}^R + \hat{Y}^C + Z. \tag{1.13}$$

It is sufficient to prove that $\hat{Y}^R$, $\hat{Y}^C$ and Z are mutually uncorrelated. From the independence between S^R and S^C it follows that $C(\hat{Y}^R,\hat{Y}^C)=0$. Furthermore,

$$C(\hat{Y}^R,Z) = C(\hat{Y}^R,\hat{Y}-\hat{Y}^R) - C(\hat{Y}^R,\hat{Y}^C) = C(\hat{Y}^R,\hat{Y}-\hat{Y}^R) . \tag{1.14}$$

Let $\mu = E(\hat{Y})$. From (1.2) we get

$$E^R[(\hat{Y}^R-\mu)(\hat{Y}-\hat{Y}^R)] = (\hat{Y}^R-\mu)\cdot E^R[\hat{Y}-\hat{Y}^R] = 0 . \tag{1.15}$$

By taking expectations in (1.15) we find that $C(\hat{Y}^R,\hat{Y}-\hat{Y}^R)=0$; together with (1.14) this yields $C(\hat{Y}^R,Z)=0$, as desired. By symmetry, we also have $C(\hat{Y}^C,Z)=0$, which concludes the proof. ■

Remark 1.2. From (1.2)-(1.4), the well-known identity

$$V(\hat{Y}) = V[E^R(\hat{Y})] + E[V^R(\hat{Y})] , \tag{1.16}$$

and its counterpart with R replaced by C, we get the following

scheme

$$\begin{aligned} V(E^R(\hat{Y})) &= VR \\ E(V^C(\hat{Y})) &= VR + VRC \\ V(E^C(\hat{Y})) &= VC \\ E(V^R(\hat{Y})) &= VRC + VC \end{aligned} \qquad (1.17)$$

By combining (1.17) and (1.3) we can get other decompositions of $V(\hat{Y})$, e.g.

$$V(\hat{Y}) = E(V^R(\hat{Y})) + E(V^C(\hat{Y})) - VRC \ . \qquad (1.18)$$

In the present paper, however, we will not use any other decomposition than (1.3). ■

2. CROSS-CLASSIFIED SAMPLING AND STRATIFICATION

If we did impose an ordinary two-way stratification on our two-dimensional population, sampling independently in each cell, then the results of the preceding section could, of course, be used inside each cell, and $V(\hat{Y})$ could be found by simply adding the cell variances. In the CPI case (and possibly others), however, the rows and columns are separately stratified. The cross-classification of these stratifications yield cells in which the samples are actually *dependent*. In this section we will discuss how to derive variances with the latter type of stratification.

Let us assume, then, that the rows are divided into G strata of sizes $N_1, N_2, \ldots, N_G$. As usual, S^R is the union of independent samples from each of the strata. Similarly, the columns are divided into H strata of sizes $M_1, M_2, \ldots, M_H$. S^R and S^C are still assumed to be independent, and Theorem 1.1 is still valid.

By crossing the row-stratification and the column-stratification we get a division of the population into G·H cells. Let Y_{gh} denote the population total of the y's in cell (g,h), for g=1,2,...,G; h=1,2,...,H. Then Y in (1.1) can be rewritten

$$Y = \sum_{g=1}^{G} \sum_{h=1}^{H} Y_{gh} \ . \qquad (2.1)$$

We shall assume that the estimator of Y is composed of some estimators $\hat{Y}_{gh}$ of the cell totals Y_{gh}, i.e.

$$\hat{Y} = \sum_{g=1}^{G} \sum_{h=1}^{H} \hat{Y}_{gh} \ . \tag{2.2}$$

$\hat{Y}_{gh}$ is assumed to be computed from the sampled units in cell (g,h) only. In other respects, the $\hat{Y}_{gh}$'s are arbitrary. Note that a pair of $\hat{Y}_{gh}$'s are *not* independent if they are from the same row or the same column. Hence, $V(\hat{Y})$ can not simply be calculated as the sum of the $V(\hat{Y}_{gh})$'s, as with ordinary two-way stratification. On the contrary, we have to add a number of covariance terms to the sum, making variance estimation very cumbersome in practice. However, invoking the decomposition of Theorem 1.1 makes the situation much simpler, as we shall now see.

Introduce the within-cell row and column estimators

$$\hat{Y}^R_{gh} = E^R(\hat{Y}_{gh}) \ , \qquad \hat{Y}^C_{gh} = E^C(\hat{Y}_{gh}) \ . \tag{2.3}$$

The within row stratum g (column stratum h) estimators are defined as follows

$$\hat{Y}^R_{g\cdot} = \sum_{h=1}^{H} \hat{Y}^R_{gh} \ , \qquad \hat{Y}^C_{\cdot h} = \sum_{g=1}^{G} \hat{Y}^C_{gh} \tag{2.4}$$

Trivially, we have the following relations

$$\hat{Y}^R = \sum_{g=1}^{G} \hat{Y}^R_{g\cdot} = \sum_{g=1}^{G} \sum_{h=1}^{H} \hat{Y}^R_{gh} \ , \qquad \hat{Y}^C = \sum_{h=1}^{H} \hat{Y}^R_{\cdot h} = \sum_{g=1}^{G} \sum_{h=1}^{H} \hat{Y}^C_{gh} \ . \tag{2.5}$$

The within row stratum g (column stratum h) variance is defined as

$$VR_g = V(\hat{Y}^R_{g\cdot}) \ , \qquad VC_h = V(\hat{Y}^C_{\cdot h}) \ . \tag{2.6}$$

We also define the within cell (g,h) interaction as

$$VRC_{gh} = V(\hat{Y}_{gh} - \hat{Y}^R_{gh} - \hat{Y}^C_{gh}) \ , \tag{2.7}$$

Theorem 2.1. Suppose we have a CCS procedure where the row and column samples are separately stratified. Then the variance of an estimator $\hat{Y}$*, based on the CCS sample and structured as in* (2.2)*, is given by*

$$V(\hat{Y}) = VR + VC + VRC \ . \tag{2.8}$$

where VR, VC *and* VRC *are defined in* (1.4) *and can be expanded as*

$$VR = \sum_{g=1}^{G} VR_g \ , \qquad VC = \sum_{h=1}^{H} VC_h \ . \tag{2.9}$$

$$VRC = \sum_{g=1}^{G} \sum_{h=1}^{H} VRC_{gh} \ . \tag{2.10}$$

Since VR and VC in (1.4) do depend only on ordinary "one-dimensional" samples, it should be no surprise that they can be expanded as simply as in (2.9). The significance of the theorem lies in the fact that VRC can be equally simply expanded (without any covariance terms). We find that the decomposition of Theorem 1.1 provides us with means of making the variance calculations with these dependent cells almost as simple as in the independent case.

Proof of Theorem 2.1. (2.8) follows directly from Theorem 1.1. (2.9) is an immediate consequence of the independence between the samples in different row (column) strata, and (1.4), (2.5) and (2.6).

It remains to show that VRC, as defined in (1.4), can be expressed as in (2.10). To this end, let $\mu_{gh}=E(\hat{Y}_{gh})$ and

$$\hat{Z}_{gh} = \hat{Y}_{gh} - \hat{Y}^R_{gh} - \hat{Y}^C_{gh} + \mu_{gh} \ . \tag{2.11}$$

Note that $E(\hat{Z}_{gh})=0$. From (1.4), (2.2) and (2.5) we see that

$$VRC = V\left(\sum_{g=1}^{G} \sum_{h=1}^{H} \hat{Z}_{gh} \right) \ . \tag{2.12}$$

We must show that the right-hand sides in (2.12) and (2.10) are equal. By definition (2.7), $VRC_{gh}=V(\hat{Z}_{gh})$ and it is sufficient to show that the $\hat{Z}_{gh}$'s are mutually uncorrelated. By the independence

of the sampling in different rows and column strata it is immediate that

$$C(\hat{Z}_{gh};\hat{Z}_{g'h'}) = 0 , \qquad g\neq g',\ h\neq h'. \tag{2.13}$$

We next turn to the case when $g=g'$ and $h\neq h'$. By the independence between S^R and S^C, we have $E^C(\hat{Y}^R_{gh})=E(\hat{Y}^R_{gh})=\mu_{gh}$, which together with the second equality in (2.3) yields that $E^C(\hat{Z}_{gh})=0$. By the independence of the samples in the column strata h and h', we have

$$E^C(\hat{Z}_{gh}\hat{Z}_{gh'}) = E^C(\hat{Z}_{gh})\cdot E^C(\hat{Z}_{gh'}) = 0 . \tag{2.14}$$

By taking expectations in (2.14), we conclude that

$$C(\hat{Z}_{gh};\hat{Z}_{gh'}) = E(\hat{Z}_{gh}\hat{Z}_{gh'}) = 0 . \tag{2.15}$$

The case with $g\neq g'$ and $h=h'$ follows by symmetry. The proof of Theorem 2.1 is complete. ∎

3. CROSS-CLASSIFIED SAMPLING AND THE HORVITZ-THOMPSON ESTIMATOR

Here we shall discuss the special case when $\hat{Y}$ is the so called Horvitz-Thompson estimator. In doing so, we go back to the unstratified situation. For the one-dimensional inclusion probabilities we use the following notation, for any i,i',j,j',

$$\begin{aligned} \pi^R_i &= P(i\in S_R) & \pi^R_{ii'} &= P(i,i'\in S_R) \\ \pi^C_j &= P(j\in S_C) & \pi^C_{jj'} &= P(j,j'\in S_C) \\ \Delta^R_{ii'} &= \frac{\pi^R_{ii'}-\pi^R_i\pi^R_{i'}}{\pi^R_i\pi^R_{i'}} & \Delta^C_{jj'} &= \frac{\pi^C_{jj'}-\pi^C_j\pi^C_{j'}}{\pi^C_j\pi^C_{j'}} \end{aligned} \tag{3.1}$$

where P denotes probability.

By the independence of S^R and S^C, the two-dimensional inclusion probabilities are just products of the one-dimensional ones. Hence, the well-known, unbiased Horvitz-Thompson estimator takes the following form in the case of Cross-classified sampling,

$$\hat{Y} = \sum_{i\in S^R}\sum_{j\in S^C} \frac{y_{ij}}{\pi^R_i\pi^C_j} . \tag{3.2}$$

The estimators in (1.2) take the form, in the notation of (1.5),

$$\hat{Y}^R = \sum_{i \in S^R} \frac{y_{i\cdot}}{\pi_i^R} \quad , \qquad \hat{Y}^C = \sum_{j \in S^C} \frac{y_{\cdot j}}{\pi_j^C} \quad . \tag{3.3}$$

Note that $\hat{Y}^R$ is the Horvitz-Thompson estimator of Y when the columns are completely enumerated, and vice versa for $\hat{Y}^C$.

Of course, the usual one-dimensional expressions for the variance of a Horvitz-Thompson estimator, in terms of inclusion probabilities, extend immediately to the two-dimensional case. However, we are interested in expressions for VR, VC and VRC in the decomposition of Theorem 1.1. The main reason for this is that we want to use these expressions in conjunction with the results on stratification in Theorem 2.1. The next theorem presents such expressions.

Theorem 3.1. The variance of the Horvitz-Thompson estimator in (3.2)*, based on a cross-classified sample, is given by* (1.3) *and*

$$\mathrm{VR} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \Delta_{ii'}^R \, y_{i\cdot} \, y_{i'\cdot} . \tag{3.4}$$

$$\mathrm{VC} = \sum_{j=1}^{M} \sum_{j'=1}^{M} \Delta_{jj'}^C \, y_{\cdot j} \, y_{\cdot j'} \tag{3.5}$$

$$\mathrm{VRC} = \sum_{i=1}^{N} \sum_{i'=1}^{N} \sum_{j=1}^{M} \sum_{j'=1}^{M} \Delta_{ii'}^R \, \Delta_{jj'}^C \, y_{ij} \, y_{i'j'} \tag{3.6}$$

Proof. The formulas for VR and VC follow from well-known results on the one-dimensional Horvitz-Thompson estimator, see e.g. Brewer & Hanif (1983, p.7), and we turn to VRC. Let I_i^R and I_j^C be indicators of the events that $i \in S^R$ and $i \in S^C$, respectively. Then by (1.1), (1.5), (3.2) and (3.3)

$$\hat{Y} - \hat{Y}^R - \hat{Y}^C + Y = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{I_i^R - \pi_i^R}{\pi_i^R} \, \frac{I_j^C - \pi_j^C}{\pi_j^C} \, y_{ij} \, . \tag{3.7}$$

Furthermore, since $\hat{Y}$, $\hat{Y}^R$ and $\hat{Y}^C$ all have expectation Y in this case, we have

$$\mathrm{VRC} = V(\hat{Y}-\hat{Y}^R-\hat{Y}^C) = E[(\hat{Y}-\hat{Y}^R-\hat{Y}^C+Y)^2] \ . \qquad (3.8)$$

By inserting (3.7) into the right-hand side of (3.8) we find

$$\mathrm{VRC}=E\left[\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{i'=1}^{N}\sum_{j'=1}^{M}\frac{I_i^R-\pi_i^R}{\pi_i^R}\,\frac{I_j^C-\pi_j^C}{\pi_j^C}\,\frac{I_{i'}^R-\pi_{i'}^R}{\pi_{i'}^R}\,\frac{I_{j'}^C-\pi_{j'}^C}{\pi_{j'}^C}\,y_{ij}\,y_{i'j'}\right] \qquad (3.9)$$

Next note that

$$E\left[\frac{I_i^R-\pi_i^R}{\pi_i^R}\,\frac{I_{i'}^R-\pi_{i'}^R}{\pi_{i'}^R}\right] = \Delta_{ii'}^R \ , \quad E\left[\frac{I_j^C-\pi_j^C}{\pi_j^C}\,\frac{I_{j'}^C-\pi_{j'}^C}{\pi_{j'}^C}\right] = \Delta_{jj'}^C \ . \qquad (3.10)$$

By taking term-wise expectations in (3.9), using the independence between S^R and S^C and (3.10) we finally get (3.6), which ends the proof. ∎

Remark 3.1. If we have separate stratification in each dimension, as described in section 2, we can use (3.6) to compute VRC *inside* each cell of the crossing of the two stratifications. By Theorem 2.1 the "overall VRC" is found by simply adding up the within-cell VRC_{gh}'s. VR and VC are one-dimensional, stratified quantities which can be handled as usual. ∎

Example 3.1. Consider the case of example 1.1, where both S^R and S^C are drawn by srswor, of sizes n and m, respectively. Then the inclusion probabilities are well known, yielding

$$\Delta_{ii'}^R = \begin{cases} \dfrac{N-n}{n} & i=i' \\ -\dfrac{1}{N-1}\,\dfrac{N-n}{n} & i\neq i' \end{cases} \qquad \Delta_{jj'}^C = \begin{cases} \dfrac{M-m}{m} & j=j' \\ -\dfrac{1}{M-1}\,\dfrac{M-m}{m} & j\neq j' \end{cases} \qquad (3.11)$$

By inserting (3.11) into (3.4)-(3.6) we obtain, after quite some algebra, the expressions for VR, VC and VRC given in (1.10)-(1.12). ∎

PART B. APPLICATION TO THE CONSUMER PRICE INDEX

4. THE DEFINITION OF THE INDEX AND ITS ESTIMATOR

In this part of the paper we shall illustrate how the results in part A can be applied to a CPI which is based on CCS samples. While the intention of Part A was to work out exact results, we will use some simplifying approximations here in Part B.

We start by defining the (theoretical) index, called I, which is the target for our sampling and estimation procedure. Then we will define its estimator $\hat{I}$. The Swedish CPI is a composition of several "sub-indexes" which use different types of sampling and index calculation. The definition of I ($\hat{I}$) chosen here is a generalization of some of the most important such sub-indexes. Our object is neither to give as general results as possible, nor to give an exact and technically detailed description of the variance calculations in the Swedish CPI, but rather to illustrate how the results in part A can be used.

The population is two-dimensional with outlets (e.g. shops, restaurants) as rows, and items (products) as columns. Both the outlets and the items are stratified (by type of retail trade and item similarity, respectively). The cell (g,h) is the crossing of outlet stratum g and item stratum h. Let v_{gh} be a weight of the cell (in the Swedish CPI, v_{gh} is the turnover for the items in group h traded in the outlets of type g). The weights are normalized so that

$$\sum_{g=1}^{G} \sum_{h=1}^{H} v_{gh} = 1 \ . \qquad (4.1)$$

The overall index I is assumed to be a weighted average of some cell indexes I_{gh},

$$I = \sum_{g=1}^{G} \sum_{h=1}^{H} v_{gh} \, I_{gh} \ . \qquad (4.2)$$

We next give the structure of the I_{gh}'s. Let f_{ij} be some function of the price of item i in outlet j at one or several points in time and let g_{ij} be another such function; cf. Example 4.1 below.

Introduce the indicator

$$1_{ij} = \begin{cases} 1 & \text{if } j \text{ is traded in } i \\ 0 & \text{else} \end{cases} \qquad (4.3)$$

For each i, let w_i^R be a (marginal) weight of outlet i, and for each j, let w_j^C be a (marginal) weight of item j. We define I_{gh} as the ratio

$$I_{gh} = \frac{\sum_{i\in g} \sum_{j\in h} 1_{ij} w_i^R w_j^C f_{ij}}{\sum_{i\in g} \sum_{j\in h} 1_{ij} w_i^R w_j^C g_{ij}} \quad . \qquad (4.4)$$

Here $i\in g$ indicates that the summation is restricted to the i's in stratum g, and likewise for $j\in h$. Of course, the functions f_{ij} and g_{ij} may include some weights - the marginal weights w_i^R w_j^C are given explicitly in (4.4) only to simplify some of the following formulas. Next let

$$y_{ij} = 1_{ij} w_i^R w_j^C f_{ij} \quad , \qquad x_{ij} = 1_{ij} w_i^R w_j^C g_{ij} \quad . \qquad (4.5)$$

As in section 2, the cell totals of the y's and x's are denoted by Y_{gh} and X_{gh}, i.e.

$$Y_{gh} = \sum_{i\in g} \sum_{j\in h} 1_{ij} w_i^R w_j^C f_{ij}, \quad X_{gh} = \sum_{i\in g} \sum_{j\in h} 1_{ij} w_i^R w_j^C g_{ij}. \qquad (4.6)$$

Now I_{gh} in (4.4) can be written as

$$I_{gh} = \frac{Y_{gh}}{X_{gh}} \quad . \qquad (4.7)$$

This completes the general definition of $\hat{I}$; we next look at some special cases.

Example 4.1. Suppose the index is a measure changes in prices from time 0 to time 1. Let p_{ij}^t be the price of j in i at time t; t=0,1. Upon putting $f_{ij}=p_{ij}^1$ and $g_{ij}=p_{ij}^0$, I_{gh} becomes a (weighted) ratio of mean prices. If instead we let $f_{ij}=p_{ij}^1/p_{ij}^0$ and $g_{ij}=1$, then I_{gh} becomes a (weighted) mean of price ratios. In the Swedish CPI we put

$$f_{ij} = \frac{p_{ij}^1}{(p_{ij}^0+p_{ij}^1)/2}, \qquad g_{ij} = \frac{p_{ij}^0}{(p_{ij}^0+p_{ij}^1)/2}, \tag{4.8}$$

See Dalén (1991a) for the reasons for using (4.8) and the weighting structure of (4.4).

As indicated by these examples, by choosing f_{ij} and g_{ij} properly, we can make (4.4) or (4.7) cover many common index formulas, with a notable exception for the geometric mean of price ratios. ■

In order to get an estimate $\hat{I}$ of the index defined above, we use price quotations (or rather f_{ij} and g_{ij}) from a cross-classified sample, as described in section 1, with separate stratifications of outlets and items, as described in section 2.

We shall assume that while the prices (f's and g's) and indicators 1_{ij} are known only for the sample, the weights w_i^R and w_j^C are known for the entire population. In outlet stratum g, the sample S_g^R is assumed to be drawn with probabilities proportional to the w_i^R, i.e., for some predetermined sample size n_g

$$\pi_{gi}^R = n_g w_i^R / W_g^R, \tag{4.9}$$

where

$$W_g^R = \sum_{i\in g} w_i^R. \tag{4.10}$$

In item stratum h, the sample S_h^C is drawn with probabilities

$$\pi_{hj}^C = m_h w_j^C / W_h^C, \tag{4.11}$$

where

$$W_h^C = \sum_{j\in h} w_j^C, \tag{4.12}$$

for some sample size m_g. Here we must assume that the quantities on the right-hand side in (4.9) and (4.11) do not exceed 1. In practice this is achieved by forming separate strata for large units, in which one makes a complete enumerations of the units. Taking care of such strata is a straight-forward task, but it makes the formulas rather involved. For simplicity we will assume

that no large unit strata are necessary.

The Horvitz-Thompson estimator of Y_{gh} is, by (3.2), (4.5), (4.9) and (4.11),

$$\hat{Y}_{gh} = \frac{W_g^R W_h^C}{n_g m_h} \sum_{i \in S_g^R} \sum_{j \in S_h^C} 1_{ij} f_{ij} . \qquad (4.13)$$

$\hat{X}_{gh}$ is given by (4.13) with f replaced by g. As an estimator of the ratio in (4.7) we take the ratio of the Horvitz-Thompson estimators,

$$\hat{I}_{gh} = \frac{\hat{Y}_{gh}}{\hat{X}_{gh}} = \frac{\sum_{i \in S_g^R} \sum_{j \in S_h^C} 1_{ij} f_{ij}}{\sum_{i \in S^R} \sum_{j \in S^C} 1_{ij} g_{ij}} . \qquad (4.14)$$

Note that $\hat{I}_{gh}$ is "self-weighting"; this property is lost by the introduction of large unit strata, though. Finally our estimated CPI is, by (4.2),

$$\hat{I} = \sum_{g=1}^{G} \sum_{h=1}^{H} v_{gh} \hat{I}_{gh} . \qquad (4.15)$$

5. THE VARIANCE OF THE INDEX

We search for an expression for $V(\hat{I})$; in doing so we must specify which sampling procedures are used to generate samples with inclusion probabilities as in (4.9) and (4.11). In large parts of the Swedish CPI, *Random systematic sampling* and/or *Sequential Poisson sampling* are used. For a description of Random systematic sampling and ordinary Poisson sampling, see e.g. Brewer & Hanif (1983, p.22 and 82); for a description of Sequential Poisson sampling see Ohlsson (1990). We shall assume that Random systematic (pps) sampling has been used. The necessary alterations to treat a situation with (ordinary) Poisson sampling are indicated in Appendix 1; the formulas for Poisson sampling can be used as approximations in the case with Sequential Poisson sampling, see Ohlsson (1990).

As usual when estimating a ratio, $\hat{I}$ is not in general unbiased. Hence, we rather look for an expression for the mean square error of $\hat{I}$, $MSE(\hat{I})$, than for $V(\hat{I})$. By a standard Taylor series linearization argument, $\hat{I}$ is approximately unbiased though, and the simulation results in section 7 indicate that the bias may be negligible, compared to the variance. Hence we will set $V(\hat{I}) \approx MSE(\hat{I})$ below.

Before presenting the formula for $V(\hat{I})$, we introduce some further notation. For $i \in g$ and $j \in h$, let

$$e_{ij}^{gh} = 1_{ij}(f_{ij} - I_{gh}g_{ij}) \ . \tag{5.1}$$

and put

$$\bar{e}_{i\cdot}^{gh} = \frac{1}{W_h^C} \sum_{j \in h} w_j^C e_{ij} \ , \tag{5.2}$$

$$\bar{e}_{\cdot j}^{gh} = \frac{1}{W_g^R} \sum_{i \in g} w_i^R e_{ij} \ , \tag{5.3}$$

Recalling the definition of X_{gh} in (4.6), set

$$\bar{X}_{gh} = X_{gh}/(W_g^R W_h^C) \ . \tag{5.4}$$

Finally, let

$$p_i^R = \frac{w_i^R}{W_g^R} \ , \qquad p_j^C = \frac{w_j^C}{W_h^C} \ . \tag{5.5}$$

Proposition 5.1. The variance of the index $\hat{I}$ can be approximated as follows

$$V(\hat{I}) \approx MSE(\hat{I}) \approx VR + VC + VRC \ , \tag{5.6}$$

where

$$VR = \sum_{g=1}^{G} \frac{1}{n_g} \sum_{i \in g} \left(1 - (n_g - 1)p_i^R\right) \left(\sum_{h=1}^{H} \frac{v_{gh}}{\bar{X}_{gh}} \bar{e}_{i\cdot}^{gh}\right)^2 p_i^R \ , \tag{5.7}$$

and

$$VC = \sum_{h=1}^{H} \frac{1}{m_h} \sum_{j \in h} \left(1 - (m_h - 1)p_j^C\right) \left(\sum_{g=1}^{G} \frac{v_{gh}}{\bar{X}_{gh}} \bar{e}_{\cdot j}^{gh}\right)^2 p_j^C \ , \tag{5.8}$$

and

$$VRC = \sum_{g=1}^{G} \sum_{h=1}^{H} \frac{1}{n_g m_h} \frac{v_{gh}^2}{\bar{X}_{gh}^2} \times$$

$$\times \left[\sum_{i \in g} \sum_{j \in h} \left(1 - (n_g - 1)p_i^R\right)\left(1 - (m_g - 1)p_j^C\right) \left(e_{ij}^{gh}\right)^2 p_i^R p_j^C \right.$$

$$\left. - \sum_{i \in g} \left(1 - (n_g - 1)p_i^R\right) \left(\bar{e}_{i\cdot}^{gh}\right)^2 p_i^R - \sum_{j \in h} \left(1 - (m_h - 1)p_j^C\right) \left(\bar{e}_{\cdot j}^{gh}\right)^2 p_j^C \right] . \qquad (5.9)$$

Proposition 5.1 is proved in section 6.

The second approximation in (5.6) is partly due to another use of the standard type linearization for the ratio estimator, see (6.1) below, and partly due to the use of approximations for the second order inclusion probabilities π_{ij} of Random systematic sampling, see (6.5). VR might be called the "outlet variance", VC the "item variance" while VRC is "outlet and item interaction".

Variance estimators based on slightly altered versions of (5.7)-(5.9) have been used to evaluate the precision of the Swedish CPI, see Dalén (1991b). Another important issue is the allocation of data capture resources to the outlet and item sample and between the strata in each dimension. Though explicit formulas for <u>optimal</u> allocation are hard to obtain, the variance formulas in Proposition 5.1 have been used for substantial <u>improvement</u> of the allocation of the Swedish CPI.

The term VRC in (5.9) does not have the interaction structure one might expect from (1.9) and the fact that srswor is a particular case of Random systematic sampling. Presumably, this is caused by the rough approximation of the π_{ij}'s. Suppose, however, that the finite population corrections $1 - (n_g - 1)p_i^R$ and $1 - (m_g - 1)p_j^C$ are all close to 1, and hence can be omitted. Then it is readily seen that VRC can be rewritten as

$$V_{int} = \sum_{g=1}^{G} \sum_{h=1}^{H} \frac{1}{n_g m_h} \frac{v_{gh}^2}{\bar{X}_{gh}^2} \sum_{i\in g} \sum_{j\in h} \left(e_{ij}^{gh} - \bar{e}_{i\cdot}^{gh} - \bar{e}_{\cdot j}^{gh} + \bar{e}_{\cdot\cdot}^{gh} \right)^2 p_i^R p_j^C \qquad (5.10)$$

Here

$$\bar{e}_{\cdot\cdot} = \frac{1}{W_h^C} \frac{1}{W_g^R} \sum_{i\in g} \sum_{j\in h} w_i^R w_j^C e_{ij}^{gh} = 0 \ . \qquad (5.11)$$

$\bar{e}_{\cdot\cdot}$ is inserted into (5.10) only to reveal the similarities between (5.10) and (1.9).

6. PROOF OF PROPOSITION 5.1

In the proof of Proposition 5.1, we shall use the linear approximation

$$MSE(\hat{I}) = E[(\sum_{g=1}^{G} \sum_{h=1}^{H} v_{gh} (\hat{I}_{gh} - I_{gh})]^2 =$$

$$= V(\sum_{g=1}^{G} \sum_{h=1}^{H} v_{gh} \frac{\hat{Y}_{gh} - I_{gh}\hat{X}_{gh}}{\hat{X}_{gh}}) \approx V(\sum_{g=1}^{G} \sum_{h=1}^{H} v_{gh} \frac{\hat{Y}_{gh} - I_{gh}\hat{X}_{gh}}{X_{gh}}) \qquad (6.1)$$

which is of a type frequently used in the literature (see e.g. Cochran 1977, p. 31). (6.1) can be shown to yield the same approximation as a first-order Taylor series approximation; we omit the proof of this fact.

Set

$$z_{ij} = v_{gh} \cdot 1_{ij} \cdot w_i^R \cdot w_j^C \, (f_{ij} - I_{gh} g_{ij}) / X_{gh} \ , \qquad (6.2)$$

and note that, by (4.7) and (4.8) the cell total of the z_{ij}'s is

$$Z_{gh} = \sum_{i\in g} \sum_{j\in h} z_{ij} = v_{gh} (Y_{gh} - I_{gh} X_{gh}) / X_{gh} = 0 \ . \qquad (6.3)$$

Let $\hat{Z}_{gh}$ be the Horvitz-Thompson estimator of Z_{gh}, defined as in (3.2), but with y_{ij} replaced by z_{ij}. By (4.9), (4.11), (5.1) and (5.4)

$$\hat{Z}_{gh} = \frac{W_g^R W_h^C}{n_g m_h} \sum_{i \in S_g^R} \sum_{j \in S_h^C} \frac{v_{gh}}{\bar{X}_{gh}} \cdot 1_{ij} (f_{ij} - I_{gh} g_{ij}) =$$

$$= \frac{1}{n_g m_h} \sum_{i \in S_g^R} \sum_{j \in S_h^C} \frac{v_{gh}}{\bar{X}_{gh}} e_{ij}^{gh} \ . \qquad (6.4)$$

$\hat{Z}$ is defined from $\hat{Z}_{gh}$ as in (2.2). From (4.13) and its analogue for X, we see that (6.1) can be rewritten

$$MSE(\hat{I}) \approx V\Big(\sum_{g=1}^{G} \sum_{h=1}^{H} \hat{Z}_{gh} \Big) = V(\hat{Z}) \ . \qquad (6.5)$$

As a further preparation for the proof of Proposition 5.1 we shall specialize the results of section 3 to the case where the row and column samples are both drawn by Random systematic (pps) sampling. In doing so, we return to the unstratified case for a while. Connor (1966) supplied exact expressions for the second-order inclusion probabilities for this procedure. These expressions are, however, unmanageable in practice. We shall use an approximation due to Hartley & Rao (1962), motivated by an asymptotic result (where $N \to \infty$ in such a way that $(n/N) \to 0$, according to Brewer & Hanif, 1983, p.14). For simplicity, we omit all terms except the first one in this approximation, and get

$$\pi_{ii'}^R \approx \frac{n-1}{n} \pi_i^R \pi_{i'}^R, \ i \neq i' ; \qquad \pi_{jj'}^C \approx \frac{m-1}{m} \pi_j^C \pi_{j'}^C, \ j \neq j' . \qquad (6.6)$$

Here n is the (fixed) size of the row sample S^R and m is the size of the column sample S^C. The Δ quantities in (3.1) become

$$\Delta_{ii'}^R \approx \begin{cases} (1-\pi_i^R)/\pi_i^R & i=i' \\ -\frac{1}{n} & i \neq i' \end{cases} \qquad \Delta_{jj'}^C \approx \begin{cases} (1-\pi_j^C)/\pi_j^C & j=j' \\ -\frac{1}{m} & j \neq j' \end{cases} \qquad (6.7)$$

Lemma 6.1. Let both S^R *and* S^C *be drawn by Random systematic (pps) sampling. Let* $\hat{Z}$ *be the Horvitz-Thompson estimator in (3.2) involving a variable* z *with population total* Z=0. *Then an approximation for* $V(\hat{Z})$ *is given by* $V(\hat{Z})$ = VR+VC+VRC, *where*

$$VR \approx \sum_{i=1}^{N} \frac{1}{\pi_i^R} \left(1 - \frac{n-1}{n} \pi_i^R\right) z_{i\cdot}^2 \qquad (6.8)$$

$$VC \approx \sum_{j=1}^{M} \frac{1}{\pi_j^C} \left(1 - \frac{m-1}{m} \pi_j^C\right) z_{\cdot j}^2 \qquad (6.9)$$

$$VRC \approx \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1}{\pi_i^R \pi_j^C} \left(1 - \frac{n-1}{n} \pi_i^R\right)\left(1 - \frac{m-1}{m} \pi_j^C\right) z_{ij}^2$$

$$- \sum_{i=1}^{N} \frac{1}{m \cdot \pi_i^R} \left(1 - \frac{n-1}{n} \pi_i^R\right) z_{i\cdot}^2 - \sum_{j=1}^{M} \frac{1}{n \cdot \pi_j^C} \left(1 - \frac{m-1}{m} \pi_j^C\right) z_{\cdot j}^2 . \qquad (6.10)$$

This lemma is applicable to any ratio estimator, not just $\hat{I}$, after the usual linearization. Note that (6.8) and (6.9) are basically two versions of the well-known variance formula for (one-dimensional) Random systematic sampling, see Brewer & Hanif (1983, formula 1.8.4). Lemma 6.1 is proved by inserting (6.7) into (3.4)-(3.6) and using the fact that Z=0. We omit the proof, which is straight-forward, but quite involved.

In Appendix 1 we give the analogue of Lemma 6.1 in the case where the row and column samples are drawn by Poisson sampling. We are now prepared to complete the proof of Proposition 5.1.

Proof of Proposition 5.1. Our starting point is (6.5). By Theorem 2.1, $V(\hat{Z})$ = VR + VC + VRC, with VR and VC as in (2.9). By (2.10), VRC is found by adding up the VRC_{gh}'s computed inside each cell. The "unstratified" Lemma 6.1 is applicable inside each cell. Note that, by (6.2) and (5.1)-(5.5)

$$z_{ij} = \frac{v_{gh}}{\bar{X}_{gh}} \cdot e_{ij}^{gh}\, p_i^R p_j^C , \qquad z_{i\cdot} = \frac{v_{gh}}{\bar{X}_{gh}} \cdot e_{i\cdot}^{gh}\, p_i^R , \qquad z_{\cdot j} = \frac{v_{gh}}{\bar{X}_{gh}} \cdot e_{\cdot j}^{gh}\, p_j^C . \tag{6.11}$$

By inserting these expressions into (6.10), and using (4.9), (4.11) and (5.5), we arrive at the desired formula (5.9) for VRC. We next turn to VR. By taking expectation over the column sample in (6.4) we get

$$E^R(\hat{Z}_{gh}) = \frac{1}{n_g} \sum_{i \in S_g^R} \sum_{j=1}^{M} \frac{v_{gh}}{\bar{X}_{gh}} e_{ij}^{gh}\, p_j^C = \frac{1}{n_g} \sum_{i \in S_g^R} \frac{v_{gh}}{\bar{X}_{gh}} e_{i\cdot}^{gh} \tag{6.12}$$

and, in a notation analogous to (2.4) and (2.3)

$$\hat{Z}_{g\cdot}^R = \frac{1}{n_g} \sum_{i \in S_g^R} \sum_{h=1}^{H} \frac{v_{gh}}{\bar{X}_{gh}} e_{i\cdot}^{gh} . \tag{6.13}$$

Let

$$\tilde{z}_i = \sum_{h=1}^{H} \frac{v_{gh}}{\bar{X}_{gh}} e_{i\cdot}^{gh} p_i^R . \tag{6.14}$$

Next note that $\hat{Z}_{g\cdot}^R$ is the Horvitz-Thompson estimator of the $\tilde{z}$-total over row stratum g; note also that this total equals 0. Now, we can find $V(\hat{Z}_{g\cdot}^R)$ by applying the well-known approximation formula for the variance in one-dimensional Random systematic sampling, see e.g. formula (1.8.4) in Brewer & Hanif (1983). Equivalently we can use (6.8) with z replaced by $\tilde{z}$. By using the following version of (2.9)

$$VR = \sum_{g=1}^{G} V(\hat{Z}_{g\cdot}^R) , \tag{6.15}$$

and recalling (4.9) and (5.5), we get (5.7).

Finally, formula (5.8) for VC follows by symmetry. This completes the proof of Proposition 5.1. ■

7. A SIMULATION STUDY

In this section we report on a simulation study which was carried out in order to get some indication of the accuracy of the approximations involved in Proposition 5.1. Specifically, the approximations are

(a) Assuming $\hat{I}$ to be approximately unbiased; this is the first $\approx$ of (5.6) and (6.1).

(b) The linearization of the ratio in (6.1).

(c) The approximation of the second-order inclusion probabilities in (6.6).

In the results in Table 7.2 below, the approximation in (a) and the combined effect of (b) and (c), can be checked, for two particular populations.

For the sake of simplicity, the study is restricted to the case without stratification. The data were price quotations for December 1989 and December 1990 from the Swedish CPI. The items considered were different kinds of meat. The outlets were two populations of supermarkets, corresponding to two different chains of retailers. The weights were the actual weights in the CPI, i.e. outlet weights from the business register and item weights from retailers lists. It should be noted that we actually used the CPI sample of oulets and items as our sampling frame in the study. Further specifications for the populations are given in Table 7.1.

Table 7.1. Population specifications.

Popul. No.	Outlets N	Items M	Cells N×M	Non-empty cells	Sample sizes n	m	No. Iter
I	26	45	1170	504	8	9	5403
II	12	41	492	234	4	9	5876

These are the net populations after a few units have been excluded in order to make all desired probabilities less than 1, cf. the discussion following (4.11). An empty cell corresponds to a pair (outlet,item) for which there is no price to be observed ($1_{ij}=0$).

The index formula used was (4.4) in conjunction with (4.8). The simulations were made with the SAS system, version 6.04, on an IBM PC. In particular, we used the random number generator inherent in the SAS system. The number of iterations was considered sufficient when, at most, the last digit in the results below was effected by adding 500 more iterations. In the following table the indexes have been multiplied by 100 (variances by 100^2).

Table 7.2. Results of the simulation.

	Populaton 1	Population 2
1. Index I from (4.1)	98.79	108.12
2. $E(\hat{I})$ from simulation	97.76	107.81
3. $Bias^2$ from 1. and 2.	1.06	0.10
4. Approx $V(\hat{I})$ fr Prop. 5.1	12.65	21.02
5. $V(\hat{I})$ from simulation	13.41	19.23

The first conclusion is that the bias is negligible for population 2 and almost negligeable for population 1. Considering the small sample sizes in these cases, as compared with the real CPI case, these results supports the belief that the bias of $\hat{I}$ can be neglected. Secondly, the approximate variances computed from Proposition 5.1 perform quite well in this case.

APPENDIX 1. CROSS CLASSIFIED POISSON SAMPLES

Samples for the Swedish CPI are drawn either by Random systematic sampling or by an alteration of Poisson sampling called Sequential Poisson sampling, see Ohlsson (1990). For ordinary Poisson sampling, the derivations of the variance of $\hat{I}$ in (4.14) are quite similar to those made for Random systematic in section 6. We shall only give the required analogue of Lemma 6.1 here.

If the row and column samples are both drawn by Poisson sampling, we have

$$\pi^R_{ii'} = \pi^R_i \pi^R_{i'} \ , \ i \neq i' ; \quad \pi^C_{jj'} = \pi^C_j \pi^C_{j'} \ , \ j \neq j' , \tag{A1.1}$$

yielding

$$\Delta^R_{ii'} = \begin{cases} (1-\pi^R_i)/\pi^R_i & i=i' \\ 0 & i \neq i' \end{cases} \qquad \Delta^C_{jj'} = \begin{cases} (1-\pi^C_j)/\pi^C_j & j=j' \\ 0 & j \neq j' \end{cases} \tag{A1.2}$$

By inserting (A1.2) into (3.4)-(3.6) we readily get the following result.

Lemma A.1. Let $\hat{Z}$ be as in Lemma 6.1. When both S^R and S^C are drawn with Poisson sampling, then $V(\hat{Z}) = VR + VC + VRC$, *where*

$$VR = \sum_{i=1}^{N} \frac{1-\pi^R_i}{\pi^R_i} z^2_{i\cdot} \ , \tag{A1.3}$$

$$VC = \sum_{j=1}^{M} \frac{1-\pi^C_j}{\pi^C_j} z^2_{\cdot j} \ , \tag{A1.4}$$

$$VRC = \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{1-\pi^R_i}{\pi^R_i} \frac{1-\pi^C_j}{\pi^C_j} z^2_{ij} \ . \tag{A1.5}$$

REFERENCES

Andersson, C., Forsman, G. & Wretman, J. (1987). On the measurement of errors in the Swedish Consumer Price Index. *Bull. Int. Stat. Inst.* **52**, Book 3, 155-171.

Balk, B.M. & Kersten, H.M.P. (1986). On the precision of Consumer Price Indices caused by the sampling variability of budget surveys. *J. Economic & Social measurm.* **14**, 19-35.

Bannerjee, K.S. (1956). A note on the optimal allocation of consumption items in the construction of a cost of living index. *Econometrica* **24**, 294-295.

Biggeri, L. & Giommi, A. (1987). On the accuracy and precision of the Consumer Price Indices. Methods and applications to evaluate the influence of the sampling of households. *Bull. Int. Stat. Inst.* **52**, Book 3, 137-154.

Brewer, K.R.W. & Hanif, M. (1983). *Sampling with unequal probabilities.* Springer, New York.

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

Cochran, R.S. (1978). Sampling in two or more dimensions. In *Contributions to Survey Sampling and Applied Statistics*, Ed. H.A. David, pp. 113-129. Academic Press, New York.

Connor, W.S. (1966). An exact formula for the probability that two specified sample units will occur in a sample drawn with unequal probabilities and without replacement. *J. Amer. Statist. Assoc.* **61**, 384-390.

Dalén, J. (1991a). Computing elementary aggregates in the Swedish consumer price index. *R & D Report, Statistics Sweden.*

Dalén, J. (1991b). Felmodeller och felkalkyler i Konsumentprisindex. (In Swedish). *Memo, Statistics Sweden.*

Hartley, H.O. & Rao, J.N.K. (1962). Sampling with unequal probabilities and without replacement. *Ann. Math. Statist.* **33**, 350-374.

Iachan, R. (1982). Systematic sampling: A critical review. *Int. Stat. Rev.* **50**, 293-303.

Ohlsson, E. (1990). Sequential Poisson sampling from a business register and its application to the Swedish Consumer Price Index. *R & D Report, Statistics Sweden.*

Quenouille, M.H. (1949). Problems in plane sampling. *Ann. Math. Statist.* **20**, 335-375.

Ripley, B.D. (1981). *Spatial Statistics.* Wiley, New York.

Valliant, R. (1991). Variance estimation for price indexes from a two-stage sample with rotating panels. *J. Business & Economic Stat.* **9**, 409-422.

Vos, J.W.E. (1964). Sampling in space and time. *Int. Stat. Rev.* **32**, 226-241.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to Mr. Jörgen Dalén, for introducing me to the problem and for providing many valuable comments during the course of this work. I am also grateful to Mr. Patrik Öhagen, who assisted in the computer programming.

R & D Reports är en för U/ADB och U/STM gemensam publikationsserie, som fr o m 1988-01-01 ersätter de tidigare "gula" och "gröna" serierna. I serien ingår även **Abstracts** (sammanfattning av metodrapporter från SCB).

R & D Reports Statistics Sweden are published by the Department of Research & Development within Statistics Sweden. Reports dealing with statistical methods have green (grön) covers. Reports dealing with EDP methods have yellow (gul) covers. In addition, abstracts are published three times a year (light brown/beige covers).

Reports published during 1992:

1992:1 (grön)	Industrins konkurrenskraft och produktivitet i fokus - en utvärdering av statistiken **(Margareta Ringquist)**
1992:2 (grön)	Automated Coding of Survey Responses: An International Review **(Lars Lyberg and Pat Dean)**
1992:3 (grön)	TABELLER „... TABELLER „... TABELLER „... - Variation och Förnyelse **(Per Nilsson)**
1992:4 (grön)	Basurval vid SCB? Studier av reskostnadseffekter vid övergång till basurval **(Elisabet Berglund)**
1992:5 (beige)	Abstracts I - sammanfattning av metodrapporter från SCB
1992:6 (grön)	Utvärdering av framskrivningsförfarande för UVAV-statistik **(Kerstin Forssén & Bengt Rosén)**

Kvarvarande **beige** och **gröna** exemplar av ovanstående promemorior kan rekvireras från Inga-Lill Pettersson, U/LEDN, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 49 56.

Kvarvarande **gula** exemplar kan rekvireras från Ingvar Andersson, U/ADB, SCB, 115 81 STOCKHOLM, eller per telefon 08-783 41 47.