Department for Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2004-07

An example for the estimation of the variance of a ratio in a complex survey design

Andreas Quatember

September 2004

An example for the estimation of the variance of a ratio in a complex survey design

Abstract: This paper shows the complexity (or uncomplexity) of the calculation of an estimator for the variance of a ratio of two Horvitz-Thompson-estimators when a complex survey design is used. As basis of the sample design used therin, the most important components of the sample design of the Austrian Microcensus until 2003 are used. This paper served as basis for a direct variance estimation of a ratio in the simulations done by the IFAS-Institute for Applied Statistics of the Johannes Kepler University for the EU-project DACSEIS (Data Quality in Complex Surveys within the European Information Society; IST-2000-26057).

1. Introduction

In sampling theory the term complex survey design means, that the sample is drawn out of a finite universe using a combination of various sample designs. One example of a complex design is the design of the Austrian Microcensus (≡AMC) in the years from 1994 to 2003. This multipurpose, quarterly survey was conducted by the Austrian Statistical Office called "Statistics Austria" in about one percent of the Austrian dwellings stock. In this survey two different sample methods were used after the universe U of all Austrian dwellings was partitioned in two nonoverlapping subuniverses A and B. The dwellings of universe A were mainly urban, whereas the dwellings of the other universe were mainly rural. Then within A and B two different sampling methods were used. In part A the used method was disproportional stratified random sampling of dwellings. In part B there was a disproportional stratified two-stage sampling of dwellings with unrestricted random sampling within both stages (for details see: Quatember, 2002, or Haslinger, 1996).
The Horvitz-Thompson-estimator (in the following: H-T-estimator) for the total t_y of a variable y with

$$t_y = \sum_U y_k \ ,$$

where the sum $\sum_U ...$ is the shorthand for $\sum_{k \in U} ...$ with $k \in U$ meaning all elements k of the universe U and y_k is the value of y for the kth element (we follow mainly the notations of Särndal et al., 1992, p.42ff), is defined as

$$\hat{t}_y = \sum_s \frac{y_k}{\pi_k} \ .$$

The sum $\sum_s ...$ is now the shorthand for $\sum_{k \in s} ....$ π_k is the first order inclusion probability of this element.
For $\hat{t}_y$, the expression

$$\hat{V}(\hat{t}) = \sum\sum_{s} \frac{\Delta_{kl}}{\pi_{kl}} \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l}$$

with Δ_{kl}, the covariance of the sample membership indicators of the k-th and the l-th sampled element, is the well-known unbiased estimator of the variance

$$V(\hat{t}) = \sum\sum_{U} \Delta_{kl} \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l}$$

(see for example: Särndal et al., 1992, p.43). The double sum $\sum\sum_{s}...$ (resp. U) is the shorthand for $\sum_{k\in s}\sum_{l\in s}...$ (resp. U) and π_{kl} the second order inclusion probability of the kth and the lth element.

Because of $s = s_A \cup s_B$; $s_A \cap s_B = \varnothing$ (s_A and s_B being the samples from the two subuniverses) for the complex sampling design, that is under investigation here, we have as H-T-estimator the sum of the two H-T-estimators of parts A and B of the universe:

$$\hat{t}_y = \hat{t}_{y_A} + \hat{t}_{y_B} .$$

Because of the independency of s_A and s_B, the variance estimator $\hat{V}(\hat{t}_y)$ is given by

$$\hat{V}(\hat{t}_y) = \hat{V}(\hat{t}_{y_A}) + \hat{V}(\hat{t}_{y_B})$$

For our complex sample design this gives for subuniverse A

$$\hat{V}(\hat{t}_{y_A}) = \sum_{h=1}^{H_A} N_h^2 \cdot \frac{(1-f_h)}{n_h} \cdot S^2_{ys_h} ,$$

the estimator for the variance of the H-T-estimator for a total in the stratified random sample with unrestricted random sampling of elements within the H_A strata with $S^2_{ys_h}$, the sample variance in stratum h.

$$\hat{V}(\hat{t}_{y_B}) = \sum_{h=1}^{H_B} \left[N_I^2 \cdot \frac{(1-f_I)}{n_I} \cdot S^2_{\hat{t}s_I} + \frac{N_I}{n_I} \cdot \sum_{s_I} N_i^2 \cdot \frac{(1-f_i)}{n_i} \cdot S^2_{ys_i} \right]$$

is the estimator for the variance of this H-T-estimator in the stratified two-stage random sample from universe B with $S^2_{\hat{t}s_I}$, the variance of the estimated cluster totals in the cluster sample s_I, and $S^2_{ys_i}$, the sample variance in cluster s_i.

2. The variance of a ratio

Besides ther estimation of t_y, another subject of interest in the survey may be the estimation of a ratio R of two population totals:

$$R = \frac{t_y}{t_z}$$

For example, y is the number of persons with at least one working hour in the last week and z the size of the whole labour force population. This ratio is estimated by

$$\hat{R} = \frac{\hat{t}_y}{\hat{t}_z}$$

where $\hat{t}_y$ resp. $\hat{t}_z$ are the H-T-estimators of the population totals t_y and t_z.

The question is how to calculate the variance estimator $\hat{V}(\hat{R})$ of the random variate $\hat{R}$. Using Taylor linearization, $\hat{R}$ is approximated by (Särndal et al., 1992, p.178):

$$\hat{R} \approx \hat{R}_0 = R + \frac{1}{t_z} \cdot \sum_s \frac{y_k - R \cdot z_k}{\pi_k}$$

So the variance of $\hat{R}$ may for large samples be approximated by (p. 179):

$$\hat{V}(\hat{R}) = \frac{1}{\hat{t}_z^2} \cdot \left[\hat{V}(\hat{t}_y) + \hat{R}^2 \cdot \hat{V}(\hat{t}_z) - 2 \cdot \hat{R} \cdot \hat{C}(\hat{t}_y, \hat{t}_z) \right] \qquad (1)$$

where $\hat{C}(\hat{t}_y, \hat{t}_z)$ denotes the estimated covariance of $\hat{t}_y$ and $\hat{t}_z$.

The calculation of the estimated variance $\hat{V}(\hat{t}_y)$ resp. $\hat{V}(\hat{t}_z)$ was shown in section 1. So the only unknown quantity of expression (1) is the covariance estimator (p.170)

$$\hat{C}(\hat{t}_y, \hat{t}_z) = \sum\sum_s \frac{\Delta_{kl}}{\pi_{kl}} \cdot \frac{y_k}{\pi_k} \cdot \frac{z_l}{\pi_l}$$

In our case the double sum $\sum\sum_s \ldots$ can be partitioned into $\sum\sum_{s_A} \ldots$, $\sum_{s_A} \sum_{s_B} \ldots$, $\sum_{s_B} \sum_{s_A} \ldots$ and $\sum\sum_{s_B} \ldots$ respectively.

$$\sum\sum_{s_A} \frac{\Delta_{kl}}{\pi_{kl}} \cdot \frac{y_k}{\pi_k} \cdot \frac{z_l}{\pi_l} = \hat{C}_A(\hat{t}_{y_A}, \hat{t}_{z_A})$$

denotes the estimated covariance of $\hat{t}_{y_A}$ and $\hat{t}_{z_A}$ and $\hat{C}_A(\hat{t}_{y_A}, \hat{t}_{z_A})$ of the same in subuniverse B.

Δ_{kl} is zero for $k \in A$ and $l \in B$.

Therefore we may write:

$$\hat{C}(\hat{t}_y, \hat{t}_z) = \hat{C}_A(\hat{t}_{y_A}, \hat{t}_{z_A}) + \hat{C}_B(\hat{t}_{y_B}, \hat{t}_{z_B})$$

Now at first we have to develop $\hat{C}_A(\hat{t}_{y_A}, \hat{t}_{z_A})$, the covariance of the H-T-estimators for part A of the universe under investigation. Because this part is partitioned into H_A strata A_i ($i=1,\ldots H_A$), it follows that

$$\hat{C}_A(\hat{t}_{y_A}, \hat{t}_{z_A}) = \sum_{i=1}^{H_A} \hat{C}_{A_i}(\hat{t}_{y_{A_i}}, \hat{t}_{z_{A_i}})$$

Because of the simple random sample of elements within each stratum, the first order inclusion probability π_k for the k-th element of the stratum A_i is

$$\pi_k = f_{A_i} = \frac{n_{A_i}}{N_{A_i}}$$

In this equation $f_{A_i} = \frac{n_{A_i}}{N_{A_i}}$ is the sampling fraction within stratum i of part A. n_{A_i} and N_{A_i} are the numbers of elements in the sample resp. the population of this stratum. The second order inclusion probability for elements k and l ($k \neq l$) of this stratum is

$$\pi_{kl} = \frac{n_{A_i}}{N_{A_i}} \cdot \frac{n_{A_i} - 1}{N_{A_i} - 1}$$

The covariance Δ_{kl} is

$$\Delta_{kl} = \begin{cases} \pi_k \cdot (1-\pi_k) & \text{if } k = l \\ \pi_{kl} - \pi_k \cdot \pi_l & \text{if } k \neq l \end{cases}$$

and therefore the ratio Δ_{kl}/π_{kl} in the formulas of the covariances $\hat{C}_{A_i}(\hat{t}_y, \hat{t}_z)$ is calculated by:

$$\frac{\Delta_{kl}}{\pi_{kl}} = \begin{cases} \dfrac{f_{A_i} \cdot (1-f_{A_i})}{f_{A_i}} = 1 - f_{A_i} & \text{if } k = l \\ -\dfrac{\dfrac{f_{A_i} \cdot (1-f_{A_i})}{N_{A_i} - 1}}{f_{A_i} \cdot \dfrac{n_{A_i} - 1}{N_{A_i} - 1}} = -\dfrac{1-f_{A_i}}{n_{A_i} - 1} & \text{if } k \neq l \end{cases}$$

Thus we have

$$\begin{aligned} \hat{C}_{A_i}(\hat{t}_{y_{A_i}}, \hat{t}_{z_{A_i}}) &= -\frac{1-f_{A_i}}{{f_{A_i}}^2} \cdot \frac{1}{n_{A_i} - 1} \cdot \sum_{k \neq l} \sum_{s_{A_i}} y_k \cdot z_l + \frac{1-f_{A_i}}{{f_{A_i}}^2} \cdot \sum_{s_{A_i}} y_k \cdot z_k \\ &= -\frac{1-f_{A_i}}{{f_{A_i}}^2} \cdot \frac{1}{n_{A_i} - 1} \cdot \sum_{s_{A_i}} y_k \cdot \sum_{s_{A_i}} z_k + \left(\frac{1-f_{A_i}}{{f_{A_i}}^2} \cdot \frac{1}{n_{A_i} - 1} + \frac{1-f_{A_i}}{{f_{A_i}}^2} \right) \cdot \sum_{s_{A_i}} y_k \cdot z_k \\ &= \frac{1-f_{A_i}}{n_{A_i}} \cdot {N_{A_i}}^2 \cdot \left(\frac{1}{n_{A_i} - 1} \cdot \sum_{s_{A_i}} y_k \cdot z_k - \frac{1}{n_{A_i}} \cdot \frac{1}{n_{A_i} - 1} \cdot \sum_{s_{A_i}} y_k \cdot \sum_{s_{A_i}} z_k \right) \\ &= N_{A_i}^2 \cdot \frac{1-f_{A_i}}{n_{A_i}} \cdot S_{yzs_{A_i}} \end{aligned}$$

where $S_{yzs_{A_i}}$ is the sample covariance of y and z within stratum A_i:

$$\begin{aligned} S_{yzs_{A_i}} &= \frac{1}{n_{A_i} - 1} \cdot \sum_{s_{A_i}} (y_k - \bar{y}_{s_{A_i}}) \cdot (z_k - \bar{z}_{s_{A_i}}) \\ &= \frac{1}{n_{A_i} - 1} \cdot \left(\sum_{s_{A_i}} y_k \cdot z_k - \frac{1}{n_{A_i}} \cdot \sum_{s_{A_i}} y_k \cdot \sum_{s_{A_i}} z_k \right) \end{aligned}$$

Summing up over all H_A strata of part A gives at last:

$$\hat{C}_A(\hat{t}_{y_A}, \hat{t}_{z_A}) = \sum_{i=1}^{H_A} N_{A_i}^2 \cdot \frac{1-f_{A_i}}{n_{A_i}} \cdot S_{yzs_{A_i}}$$

As we can see, to calculate this covariance estimator, we have only to calculate the covariance estimators of the variables y and z within each stratum. ♣

Now we have to look at the covariance $\hat{C}_B(\hat{t}_{y_B}, \hat{t}_{z_B})$ of the H-T-estimators within part B of the universe. This part is partitioned into H_B strata, which gives of course:

$$\hat{C}_B(\hat{t}_{y_B}, \hat{t}_{z_B}) = \sum_{i=1}^{H_B} \hat{C}_{B_i}(\hat{t}_{y_{B_i}}, \hat{t}_{z_{B_i}})$$

Within each stratum B_i, PSUs are selected by simple random sampling and the elements are selected from these PSUs the same way.
For these elements we have the first order inclusion probability π_k given by

$$\pi_k = f_{iI} \cdot f_{ij}$$

with $f_{iI} = n_{iI}/N_{iI}$, the sample fraction of PSUs within B_i (the letter I stands for first stage) and $f_{ij} = n_{ij}/N_{ij}$, the sample fraction of elements within B_{ij}, the j-th selected PSU in stratum B_i. For the second order inclusion probablities π_{kl} we have to distinguish two cases: In the first case, two elements k and l of the sample from B_i belong both to the same PSU B_{ij}. Then we have:

$$\pi_{kl} = \begin{cases} f_{iI} \cdot f_{ij} & \text{if } k = l \\ f_{iI} \cdot f_{ij} \cdot \dfrac{n_{ij} - 1}{N_{ij} - 1} & \text{if } k \neq l \end{cases}$$

If $k \in B_{ij}$, $l \in B_{ij'}$ $(j \neq j')$, on the other hand this inclusion probability becomes

$$\pi_{kl} = f_{iI} \cdot \frac{n_{iI} - 1}{N_{iI} - 1} \cdot f_{ij} \cdot f_{ij'}$$

Then in

$$\hat{C}_{B_i}(\hat{t}_{y_{B_i}}, \hat{t}_{z_{B_i}}) = \sum\sum_{s_{B_i}} \frac{\Delta_{kl}}{\pi_{kl}} \cdot \frac{y_k}{\pi_k} \cdot \frac{z_l}{\pi_l}$$

the double sum can be partitioned into:

$$\sum\sum_{s_{B_i}} \ldots = \sum_{j=1}^{n_{iI}} \underbrace{\left(\sum\sum_{s_{B_{ij}}} \ldots \right)}_{\equiv I} + \sum_{j \neq j'} \underbrace{\left(\sum_{s_{B_{ij}}} \sum_{s_{B_{ij'}}} \ldots \right)}_{\equiv II}$$

Expression I gives:

$$I = \sum_{k \in s_{B_{ij}}} \frac{f_{iI} \cdot f_{ij} - (f_{iI} \cdot f_{ij})^2}{f_{iI} \cdot f_{ij}} \cdot \frac{y_k}{f_{iI} \cdot f_{ij}} \cdot \frac{z_k}{f_{iI} \cdot f_{ij}} +$$

$$+ \sum_{k \neq l} \sum_{s_{B_{ij}}} \frac{f_{iI} \cdot f_{ij} \cdot \dfrac{n_{ij} - 1}{N_{ij} - 1} - (f_{iI} \cdot f_{ij})^2}{f_{iI} \cdot f_{ij} \cdot \dfrac{n_{ij} - 1}{N_{ij} - 1}} \cdot \frac{y_k}{f_{iI} \cdot f_{ij}} \cdot \frac{z_l}{f_{iI} \cdot f_{ij}} =$$

$$= \sum_{s_{B_{ij}}} \frac{1 - f_{iI} \cdot f_{ij}}{f_{iI}^2 \cdot f_{ij}^2} \cdot y_k \cdot z_k + \sum_{k \neq l} \sum_{s_{B_{ij}}} \frac{\dfrac{n_{ij} - 1}{N_{ij} - 1} - f_{iI} \cdot f_{ij}}{f_{iI}^2 \cdot f_{ij}^2 \cdot \dfrac{n_{ij} - 1}{N_{ij} - 1}} \cdot y_k \cdot z_l =$$

$$= \frac{\dfrac{n_{ij} - 1}{N_{ij} - 1} - f_{iI} \cdot f_{ij}}{f_{iI}^2 \cdot f_{ij}^2 \cdot \dfrac{n_{ij} - 1}{N_{ij} - 1}} \cdot \sum_{s_{B_{ij}}} y_k \cdot \sum_{s_{B_{ij}}} z_k - \left(\frac{\dfrac{n_{ij} - 1}{N_{ij} - 1} - f_{iI} \cdot f_{ij}}{f_{iI}^2 \cdot f_{ij}^2 \cdot \dfrac{n_{ij} - 1}{N_{ij} - 1}} - \frac{1 - f_{iI} \cdot f_{ij}}{f_{iI}^2 \cdot f_{ij}^2} \right) \cdot \sum_{s_{B_{ij}}} y_k \cdot z_k$$

The first ratio in the last line gives

$$\frac{\frac{n_{ij}-1}{N_{ij}-1}-f_{iI}\cdot f_{ij}}{f_{iI}^{\,2}\cdot f_{ij}^{\,2}\cdot\frac{n_{ij}-1}{N_{ij}-1}}=\frac{n_{ij}-1-f_{iI}\cdot f_{ij}\cdot(N_{ij}-1)}{f_{iI}^{\,2}\cdot f_{ij}^{\,2}\cdot(n_{ij}-1)}$$

and the expression in the bracket gives

$$\frac{\frac{n_{ij}-1}{N_{ij}-1}-f_{iI}\cdot f_{ij}}{f_{iI}^{\,2}\cdot f_{ij}^{\,2}\cdot\frac{n_{ij}-1}{N_{ij}-1}}-\frac{1-f_{iI}\cdot f_{ij}}{f_{iI}^{\,2}\cdot f_{ij}^{\,2}}=\frac{\frac{n_{ij}-1}{N_{ij}-1}-f_{iI}\cdot f_{ij}-\frac{n_{ij}-1}{N_{ij}-1}+\frac{n_{ij}-1}{N_{ij}-1}\cdot f_{iI}\cdot f_{ij}}{f_{iI}^{\,2}\cdot f_{ij}^{\,2}\cdot\frac{n_{ij}-1}{N_{ij}-1}}=$$

$$=\frac{-N_{ij}+n_{ij}}{f_{iI}\cdot f_{ij}\cdot(n_{ij}-1)}$$

So expression I from above can be written as:

$$I=\frac{1}{f_{iI}\cdot f_{ij}\cdot(n_{ij}-1)}\cdot\left[(N_{ij}-n_{ij})\cdot\sum\nolimits_{s_{B_{ij}}}y_k\cdot z_k-\left(N_{ij}-1-\frac{n_{ij}-1}{f_{iI}\cdot f_{ij}}\right)\cdot\sum\nolimits_{s_{B_{ij}}}y_k\cdot\sum\nolimits_{s_{B_{ij}}}z_k\right]$$

Now we turn to expression II, where we have to sum up over the elements from different sample-PSUs of stratum B_i:

$$II=\sum_{j\neq j'}\sum\nolimits_{k\in s_{B_{ij}}}\sum\nolimits_{l\in s_{B_{ij'}}}\frac{f_{iI}\cdot\frac{n_{iI}-1}{N_{iI}-1}\cdot f_{ij}\cdot f_{ij'}-f_{iI}\cdot f_{ij}\cdot f_{iI}\cdot f_{ij'}}{f_{iI}\cdot\frac{n_{iI}-1}{N_{iI}-1}\cdot f_{ij}\cdot f_{ij'}}\cdot\frac{y_k}{f_{iI}\cdot f_{ij}}\cdot\frac{z_l}{f_{iI}\cdot f_{ij'}}$$

$$=\left(1-\frac{f_{iI}\cdot(N_{iI}-1)}{n_{iI}-1}\right)\cdot\frac{1}{f_{iI}^2\cdot f_{ij}\cdot f_{ij'}}\cdot\sum\nolimits_{s_{B_{ij}}}y_k\cdot\sum\nolimits_{s_{B_{ij'}}}z_k$$

At the end we can sum and get:

$$\hat{C}_B(\hat{t}_{y_B},\hat{t}_{z_B})=$$

$$=\sum_{i=1}^{H_B}\left\{\sum_{j=1}^{n_{iI}}\frac{1}{f_{iI}\cdot f_{ij}\cdot(n_{ij}-1)}\cdot\right.$$

$$\cdot\left[(N_{ij}-n_{ij})\cdot\sum\nolimits_{s_{B_{ij}}}y_k\cdot z_k-\left(N_{ij}-1-\frac{n_{ij}-1}{f_{iI}\cdot f_{ij}}\right)\cdot\sum\nolimits_{s_{B_{ij}}}y_k\cdot\sum\nolimits_{s_{B_{ij}}}z_k\right]+$$

$$\left.+\sum_{j\neq j'}\left(1-\frac{f_{iI}\cdot(N_{iI}-1)}{n_{iI}-1}\right)\cdot\frac{1}{f_{iI}^2\cdot f_{ij}\cdot f_{ij'}}\cdot\sum\nolimits_{s_{B_{ij}}}y_k\cdot\sum\nolimits_{s_{B_{ij'}}}z_k\right\}$$

The covariance estimator for part B therefore consists of one term similar to the sum of covariances of y and z within PSUs and another, where we have to use the product of the sample sums of y and z in different PSUs. ♣

At last we have to sum up the expressions for $\hat{C}_A(\hat{t}_{y_A},\hat{t}_{z_A})$ resp. $\hat{C}_B(\hat{t}_{y_B},\hat{t}_{z_B})$ to get $\hat{C}(\hat{t}_y,\hat{t}_z)$ and use this to calculate $\hat{V}(\hat{R})$ in (1).

References

Quatember, A. (main responsibility) (2002). *Analysis of National Surveys*. Deliverables 2.1 and 2.2 of the DACSEIS project.

Haslinger, A. (1996). Stichprobenplan des Mikrozensus ab 1994. *Statistische Nachrichten*, 4/1996, 312-324.

Särndal, C.-E., Swensson, B., Wretman, J. (1992). Model Assisted Survey Sampling. Springer. New York.