Department of Applied Statistics
Johannes Kepler University Linz

IFAS Research Paper Series
2014-67

Simplified Variance Estimation for Three-Stage Random Sampling

Andreas Quatember

October 2014

Simplified Variance Estimation for Three-Stage Random Sampling[1]

Andreas Quatember

IFAS - *Department of Applied Statistics*
Johannes Kepler University (JKU) Linz, Austria

I Introduction to the Problem

When a population U under study is partitioned into C clusters i (i=1,…,C) as primary sampling units (P), each cluster i itself is partitioned into M_i subclusters j (j=1,…,M_i) as secondary sampling units (S), and each subcluster j contains N_{ij} elements k (k=1,…,N_{ij}) as tertiary sampling units (T), the total t of a study variable y (for instance, a certain part of costs per unit) can be written as

$$t = \sum_{i=1}^{C} \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} y_{ijk} \, . \qquad (1)$$

In (1), y_{ijk} denotes the value of y for the k-th element in the j-th subcluster of the i-th cluster. Here and in the following, the basic literature used is Särndal et al. (1992), Ardilly and Tillé (2006), Lohr (2010), and Quatember (2014a).

If a proportion P of t and the total of another variable a (for instance, the overall costs of a unit)

$$A = \sum_{i=1}^{C} \sum_{j=1}^{M_i} \sum_{k=1}^{N_{ij}} a_{ijk} \, ,$$

is of interest, then P is given by

$$P = \frac{t}{A} \, .$$

For the partial costs y_{ijk} per unit, $y_{ijk} = q_{ijk} \cdot a_{ijk}$ applies.

To estimate a total t of a variable y unbiasedly, the general Horvitz-Thompson (HT) estimator is given by

$$t_{HT} = \sum_{k=1}^{n} \frac{y_k}{\pi_k}$$

with π_k, the first-order sample inclusion probabilities of elements k (k=1,…,n). Its variance is given by

$$V(t_{HT}) = \sum_{k=1}^{N} \sum_{l=1}^{N} \Delta_{kl} \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l}$$

[1] This paper is a product of the contract project with the same name (available on [10/16/2014]: http://www.jku.at/content/e263/e16099/e16086/e173791/?view=PROD&feid=198&fp_id=3552)

with the covariance of the sample inclusion indicators I_k and I_l: $\Delta_{kl} = \pi_{kl} - \pi_k \cdot \pi_l$. Therein, π_{kl} denotes the second-order sample inclusion probabilities of two elements k and l of the population. This theoretical variance is estimated unbiasedly by

$$\widehat{V}(t_{HT}) = \sum\nolimits_{k=1}^{n} \sum\nolimits_{l=1}^{n} \frac{\Delta_{kl}}{\pi_{kl}} \cdot \frac{y_k}{\pi_k} \cdot \frac{y_l}{\pi_l}$$

(see, for instance, Quatember 2014, p. 24ff, or Särndal et al., p. 42ff).

Hence, the proportion P is estimated by

$$\widehat{P} = \frac{t_{HT}}{A}.$$

Its variance is given by

$$V(\widehat{P}) = \frac{1}{A^2} \cdot V(t_{HT}),$$

which is estimated unbiasedly by

$$\widehat{V}(\widehat{P}) = \frac{1}{A^2} \cdot \widehat{V}(t_{HT}).$$

II Three-stage element sampling

In the case of a three-stage sampling scheme in a population U partitioned into clusters and subclusters as described at the beginning of Section I, a probability sample of size c is selected at the first stage with cluster inclusion probabilities π_I (i=1,…,C). At the next stage, a probability sample of m_i subclusters j is drawn within each of the c sample clusters i with conditional subcluster inclusion probabilities $\pi_{j|i}$ (j=1,…,M_i). At the third stage, within each of the sampled subclusters j of cluster i of the second stage, a probability sample of size n_{ij} is drawn with conditional inclusion probabilities $\pi_{k|ij}$ for all elements k (k=1,…,N_{ij}). With these terms, the HT estimator of the three-stage process (3st) is given by

$$t_{3st} = \sum\nolimits_{i=1}^{c} \sum\nolimits_{j=1}^{m_i} \sum\nolimits_{k=1}^{n_{ij}} \frac{y_{ijk}}{\pi_{ijk}}.$$

For proportional to size without replacement sampling (πPS), the inclusion probabilities at the different stages take on the values

$$\pi_i = \frac{x_i \cdot c}{\sum\nolimits_{i=1}^{C} x_i},$$

at the first sampling stage with auxiliary size variable x,

$$\pi_{j|i} = \frac{u_j \cdot m_i}{\sum\nolimits_{j=1}^{M_i} u_j},$$

at the second sampling stage with auxiliary size variable u, and

$$\pi_{k|ij} = \frac{z_k \cdot n_{ij}}{\sum\nolimits_{k=1}^{N_{ij}} z_k},$$

at the third sampling stage with auxiliary size variable z.

In such a three-stage probability sample with first-order element inclusion probabilities

$$\pi_{ijk} = \pi_i \cdot \pi_{j|i} \cdot \pi_{k|ij},$$

the Horvitz-Thompson (HT) estimator as defined in the previous section yields

$$t_{3st} = \sum\nolimits_{i=1}^{c} \sum\nolimits_{j=1}^{m_i} \sum\nolimits_{k=1}^{n_{ij}} \frac{y_{ijk}}{\pi_{ijk}} = \sum\nolimits_{i=1}^{c} \frac{1}{\pi_i} \cdot \sum\nolimits_{j=1}^{m_i} \frac{1}{\pi_{j|i}} \cdot \underbrace{\sum\nolimits_{k=1}^{n_{ij}} \frac{y_{ijk}}{\pi_{k|ij}}}_{t_{HT,ij}} = \sum\nolimits_{i=1}^{c} \frac{1}{\pi_i} \cdot \underbrace{\sum\nolimits_{j=1}^{m_i} \frac{t_{HT,ij}}{\pi_{j|i}}}_{t_{HT,i}}.$$

Hence, this HT estimator can be written as

$$t_{3st} = \sum\nolimits_{i=1}^{c} \frac{t_{HT,i}}{\pi_i} \qquad (2)$$

with $t_{HT,i}$, the HT estimator of t_i, the total of y in cluster i. The HT estimator $t_{HT,ij}$ estimates unbiasedly the total of y in the j-th subcluster of the i-th cluster.

For

$$x_i = \sum\nolimits_{j=1}^{M_i} \sum\nolimits_{k=1}^{N_{ij}} a_{ijk},$$

the sum of the costs of variable a in the i-th cluster,

$$u_j = \sum\nolimits_{k=1}^{N_{ij}} a_{ijk},$$

the sum of the costs a within the j-th subcluster of the i-th cluster,

$$z_k = a_{ijk},$$

the costs a of the k-th element of the j-th subcluster within the i-th cluster,

$$\pi_{k|ij} = n_{ij} \cdot \frac{a_{ijk}}{\sum\nolimits_{k=1}^{N_{ij}} a_{ijk}},$$

$$\pi_{j|i} = m_i \cdot \frac{\sum\nolimits_{k=1}^{N_{ij}} a_{ijk}}{\sum\nolimits_{j=1}^{M_i} \sum\nolimits_{k=1}^{N_{ij}} a_{ijk}}$$

and

$$\pi_i = c \cdot \frac{\sum\nolimits_{j=1}^{M_i} \sum\nolimits_{k=1}^{N_{ij}} a_{ijk}}{\sum\nolimits_{i=1}^{C} \sum\nolimits_{j=1}^{M_i} \sum\nolimits_{k=1}^{N_{ij}} a_{ijk}} = c \cdot \frac{\sum\nolimits_{j=1}^{M_i} \sum\nolimits_{k=1}^{N_{ij}} a_{ijk}}{A}$$

applies. Hence, for t_{3st},

$$t_{3st} = \frac{1}{c} \cdot \sum\nolimits_{i=1}^{c} \frac{1}{m_i} \cdot \sum\nolimits_{j=1}^{m_i} \frac{1}{n_{ij}} \cdot \sum\nolimits_{k=1}^{n_{ij}} \underbrace{\frac{A}{a_{ijk}}}_{A \cdot q_{ijk}} \cdot y_{ijk}$$

applies.

For the estimation of P,

$$\widehat{P} = \frac{t_{3st}}{A}$$

is used. A further improvement with respect to the estimation of P may be achieved by a ratio estimator

$$\hat{P}_{rat} = \frac{t_{3st}}{\hat{A}}$$

with

$$\hat{A} = \sum_{i=1}^{c} \sum_{j=1}^{m_i} \sum_{k=1}^{n_{ij}} \frac{A}{c \cdot m_i \cdot n_{ij}} .$$

The variance of the HT estimator t_{3st} according to (2) in a three-stage probability sample is written by

$$V_{3st} = V_P + V_S + V_T . \tag{3}$$

Obviously, the variance is partitioned into three components reflecting the three stages of sampling as three different sources of variation of t_{3st}. In (3),

$$V_P = \sum_{i=1}^{C} \sum_{i'=1}^{C} \Delta_{ii'} \cdot \frac{t_i}{\pi_i} \cdot \frac{t_{i'}}{\pi_{i'}}$$

with $\Delta_{ii'}$, the covariance of the sample inclusion probability of clusters i and i'. t_i denotes the total of y in cluster i. This is the variation with respect to sampling at the first stage of the process. Furthermore, V_S, the variation of the estimator due to the second sampling stage, is given by

$$V_S = \sum_{i=1}^{C} \frac{V_i}{\pi_i} .$$

The variance V_i denotes the variance of $\sum_{j=1}^{m_i} \frac{t_{ij}}{\pi_{j|i}}$ with t_{ij}, the total of y in subcluster j of cluster i (see Formula (4.4.9) in Särndal et al. 1992, p.148). Eventually, the third-stage contribution to the overall variation of t_{3st} is given by

$$V_T = \sum_{i=1}^{C} \frac{\sum_{j=1}^{M_i} \frac{V_{ij}}{\pi_{j|i}}}{\pi_i} .$$

Therein, V_{ij} denotes the variance of the HT estimator $t_{HT,ij}$ according to (4.4.8) in ibid., p.148. This variance component completes the calculation of V_{3st}.

It is this variance that has to be estimated, when the results of a sample survey are to be presented in form of an approximate confidence. “It is good practice in the reporting of survey results to supply … the point estimates with their estimated standard errors, that is the square root of the estimated variances“ (ibid., p. 150). Variance (3) is unbiasedly estimated by

$$\hat{V}_{3st} = \underbrace{\sum_{i=1}^{c} \sum_{i'=1}^{c} \frac{\Delta_{ii'}}{\pi_{ii'}} \cdot \frac{t_{HT,i}}{\pi_i} \cdot \frac{t_{HT,i'}}{\pi_{i'}}}_{\hat{V}_P} + \sum_{i=1}^{c} \frac{\hat{V}_i}{\pi_i} \tag{4}$$

with

$$\widehat{V}_i = \sum\nolimits_{j=1}^{m_i}\sum\nolimits_{j'=1}^{m_i}\frac{\Delta_{jj'|i}}{\pi_{jj'|i}}\cdot\frac{t_{HT,ij}}{\pi_{j|i}}\cdot\frac{t_{HT,ij'}}{\pi_{j'|i}}+\sum\nolimits_{j=1}^{m_i}\frac{\widehat{V_{ij}}}{\pi_{j|i}}$$

and

$$\widehat{V}_{ij} = \sum\nolimits_{k=1}^{n_i}\sum\nolimits_{k'=1}^{n_i}\frac{\Delta_{kk'|ij}}{\pi_{kk'|ij}}\cdot\frac{y_{ijk}}{\pi_{k|ij}}\cdot\frac{y_{ijk'}}{\pi_{k'|ij}}\;.$$

From these formulae, the statistical properties of a two-stage sampling process can immediately be derived (see, for instance, Quatember 2014a, ch. 6).

The calculation of a variance estimate according to (4) may be hard. In particular, the calculation of the second-order inclusion probabilities of selection units at the different stages of the sampling process can be cumbersome or even impossible for certain sampling procedures applied within the three stages of sampling. In particular, this applies for πPS sampling. One possibility to cope with this problem is the estimation of these probabilities (cf. Berger 2004). But, taking into account the theoretical and practical effort of this approach in three-stage sampling, a simpler variance expression than (4) has to be considered.

III Four Options for a Simplified Variance Estimation

III.1 Option I

The simplified option I, applicable as an estimator of the variance of t_{3st}, uses only the first term $\widehat{V}_P$ of the variance estimator (4):

$$\widehat{V}_I = \sum\nolimits_{i=1}^{c}\sum\nolimits_{i'=1}^{c}\frac{\Delta_{ii'}}{\pi_{ii'}}\cdot\frac{t_{HT,i}}{\pi_i}\cdot\frac{t_{HT,i'}}{\pi_{i'}} \qquad (5)$$

This means that only the covariance of the sample inclusion indicator on cluster level and the cluster second-order inclusion probabilities are needed. In fact, $\widehat{V}_P$ overestimates V_P, but does not cover all other components of V_{3st}. This means that $\widehat{V}_P$ provides a negatively biased estimator of V_{3st}. But, experience shows that in many cases the amount of underestimation is small, especially, when the π_i's are small. A compensation of the negative bias using subsampling from the samples after the first stage was discussed by Srinath and Hidiroglou (1980). But, also this simplified biased variance estimator needs the second-order inclusion probabilities at the cluster level to be calculated, which is cumbersome, for example, for πPS sampling.

For a fixed size first-stage probability sample, $\widehat{V}_P$ can be written as

$$\widehat{V}_I = -\frac{1}{2}\cdot\sum\nolimits_{i=1}^{c}\sum\nolimits_{i'=1}^{c}\frac{\Delta_{ii'}}{\pi_{ii'}}\cdot\left(\frac{t_{HT,i}}{\pi_i}-\frac{t_{HT,i'}}{\pi_{i'}}\right)^2 \qquad (6)$$

(cf. Särndal et al. 1992, p.153).

III.2 Option II

Another option (II) of a simplified variance estimator for without replacement sampling schemes is delivered by adapting the variance estimator that would have been obtained when the clusters would have been selected by a with-replacement sampling design. Usually, this will result in an overestimation of the true variance when the sampling is actually done without replacement.

In a multi-stage sampling design, the specific variance estimator is given by

$$\widehat{V}_{II} = \frac{1}{c \cdot (c-1)} \cdot \sum\nolimits_{i=1}^{c} \left(\frac{t_{HT,i}}{p_i} - t_{3st} \right)^2 \qquad (7)$$

with $p_i = \frac{\pi_i}{c}$, the probability for cluster i to be selected in the next step of the with-replacement selection process (cf. Särndal et al. 1992, p.154). This expression incorporates also the variance due to the 2nd and 3rd stage of sampling by the variance of the weighted HT estimators $t_{HT,i}$ of the cluster totals t_i with weights p_i. Therein, $t_{HT,i}$ is given by

$$t_{HT,i} = \sum\nolimits_{j=1}^{m_i} \frac{1}{\pi_{j|i}} \cdot \sum\nolimits_{k=1}^{n_{ij}} \frac{y_{ijk}}{\pi_{k|ij}}$$

as presented in Section 2.

Systematic probability proportional to size sampling without replacement following from a randomly ordered population is an example of a sample selection method, for which $\widehat{V}_{II}$ overestimates V_{3st}The result is a “conservative” confidence interval that can easily be calculated because no second-order inclusion probabilities from any of the stages are needed. For a small sampling fraction of clusters, c/C, the difference between V_{3st}, and $\widehat{V}_{II}$ will be negligible. Särndal et al. (1992) deliver an example for the actual calculation of $\widehat{V}_{II}$ (ibid., p.152f). Because of the with-replacement sampling, it is possible to get different subsamples from the same cluster, and $\widehat{V}_{II}$ captures both parts of the variance V_{3st}: the one due to the selection of the clusters and the part of V_{3st} arising from the estimation of the cluster totals t_i at the following stages.

III.3 Option III

A third option III for the simplified estimation of the variance of t_{3st} uses an estimation of the design effect of the sampling design defined as

$$deff = \frac{V_{3st}}{V_{SI}},$$

where V_{SI} is the variance of the HT estimator of t in simple random sampling without replacement (SI). A biased estimator of deff is given by

$$\widehat{deff} = \frac{\hat{V}_{3st}}{\hat{V}_{SI}},$$

the ratio of two variance estimates. Hence, a biased estimator $\hat{V}_{III}$ of V_{3st} can be defined by

$$\hat{V}_{III} = \widehat{deff} \cdot \hat{V}_{SI}. \tag{8}$$

For equal numbers n_i of elements observed within the sample clusters, SI sampling at the different stages, and large C, the design effect is estimated by

$$\widehat{deff} \approx 1 + \hat{\rho} \cdot (n_i - 1)$$

with $\hat{\rho}$, the estimated intra-class correlation coefficient measuring the homogeneity of units within the same clusters (cf., for instance, Ardilly and Tillé 2006, p.161). For a large population size N compared to the number of clusters C in the population, this measure has a range from zero to one. It reaches the value one for complete homogeneity within the clusters, which is the worst case of sampling with clusters with respect to the variance of the HT estimator (cf. Särndal et al. 1992, p.131). For $\rho \approx 0$, meaning that each cluster has the same heterogeneity with respect to the study variable, the design effect approximately equals one and the variance of the three-stage process can be estimated by the SI variance formula.

Nevertheless, one needs not only an estimate of deff but also an estimate of the variance of the HT estimator with SI sampling. Using data of the three-stage sample to estimate S^2, the variance of y in the population, although the actual sampling was not SI element sampling, delivers a biased estimate of the true S^2. Hence, $\hat{V}_{SI}$ will have a bias of unknown extent.

To take also account of possible unequal inclusion probabilities (like in πPS sampling), an estimator of the overall design effect is calculated by the design effect deffc due to clustering with clusters of unequal sample sizes n_i and a design effect due to the unequal inclusion probabilities deffp. Kish (1987) described an estimator of deffc with the mean value $\overline{n_i}$ of the within-cluster sample sizes, which substitutes the equal sample sizes n_i within clusters in the formula above by

$$\widehat{deffc} \approx 1 + \hat{\rho} \cdot (\overline{n_i} - 1)$$

(cf., for instance, Gabler et al. 1999, or Ganninger et al. 2007). The part of the design effect with respect to unequal inclusion probabilities is estimated by

$$\widehat{deffp} = n \cdot \frac{\sum_{i=1}^{L} w_i^2 \cdot n_i}{\left(\sum_{i=1}^{L} w_i \cdot n_i\right)^2}$$

with w_i, the unique design weights of the weighting class i of L weighting classes (see also Gabler et al. 1999). Then, the variance of the three-stage HT estimator of t is estimated by

$$\hat{V}_{III} = \widehat{deffp} \cdot \widehat{deffc} \cdot \hat{V}_{SI}. \tag{9}$$

For SI sampling within each stage and equal numbers n_i of elements observed within the sample clusters, $\widehat{deffp} = 1$ applies.

III.4 Option IV

Another variance estimation option IV makes use of resampling methods. These computer-intensive methods use computer power instead of heavy calculations. One of these methods is the bootstrap. This resampling procedure was originally developed for i.i.d. situations (Efron 1979). For its application in statistical surveys, different approaches are proposed (see, for instance, Shao and Tu 1995, p.246ff).

The approach that directly mimicks the original idea makes use of the generation of a bootstrap population, from which the resamples are drawn by the original sampling scheme (see, for instance, Quatember 2014b, p.89ff). For this purpose, in a three-stage design, the generation of the bootstrap population from the original sample data has to consider all three stages. Therefore, within the sampled second stage clusters, the sample units k are replicated according to their third stage inclusion probabilities $\pi_{k|ij}$. This results in set-valued estimators of the second-stage sample subclusters j with respect to the interesting variables. Then, these second stage units j have to be replicated according to their second stage inclusion probabilities $\pi_{j|i}$. This results in set-valued estimators of the first-stage sample clusters i. Eventually, by replicating each generated cluster i according to its inclusion probability π_i, the generation of a bootstrap population as a set-valued estimator of the entire population is finished. From this population, which can be called a pseudo-population (Quatember 2014b), a number of B resamples are drawn with the same sampling scheme as the one originally used in the survey and in each of these B resamples the estimator

$$t_{3st,b} = \sum\nolimits_{i=1}^{c} \frac{t_{HT,i}}{\pi_i}$$

is calculated according to (2) (b=1,…,B). Then, the theoretical variance (3) of (2) is estimated by

$$\widehat{V}_{IV} = \frac{1}{B-1} \cdot \sum_{b=1}^{B} (t_{3st,b} - \overline{t_{3st}})^2 \tag{10}$$

with

$$\overline{t_{3st}} = \frac{1}{B} \cdot \sum_{b=1}^{B} t_{3st,b} \ ,$$

the mean value of the estimators $t_{3st,b}$ from the B bootstrap samples. For (10) to be an accurate estimator of the true variance (3), the sample sizes have to be large enough at all three stages because if this is not the case, only a small number of units are replicated at all stages and resamples are drawn from only a small number of different values.

IV Interval Estimation

With one of the options for the estimation of the variance of t_{3st} (or $\widehat{P}$) presented in Section III, one can calculate an approximate (1−α)-confidence interval for the true t by

$$t_{3st} \pm u_{1-\alpha/2} \cdot \sqrt{\widehat{V}_\bullet} \qquad (11)$$

Therein, $\widehat{V}_\bullet$ denotes the used variance estimate and $u_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the standard normal distribution. For this interval to be valid, the central limit theorem must hold and $\widehat{V}_\bullet$ should be a consistent estimator of the true variance $V(t_{3st})$ according to (3).

Considering the computational and technical efforts of the different options discussed in Section III, Option II seems to be of interest for its use as $\widehat{V}_\bullet$ in (11), if the sampling fraction c/C at the first stage is small. In this case, the overestimation of the true variance (3) by the with-replacement variance (7) will be “unimportant” (Särndal et al. 1992, p. 154) and one could as well draw with-replacement samples, so that the variance formula really fits to the applied sampling scheme. The variance estimator itself is design-based.

The quality of Option III (Formula (9)) for its use as $\widehat{V}_\bullet$ in (11), depends mainly on the quality of the estimation of the SI-variance with the data from the observed sample. Therefore, this option is model-based.

If P is estimated by $\widehat{P}$, the approximate confidence interval is given by

$$\widehat{P} \pm u_{1-\alpha/2} \cdot \sqrt{\frac{1}{A^2} \cdot \widehat{V}_\bullet}\ .$$

When the ratio estimator $\widehat{P}_{rat}$ from Section II is used, the interval

$$\widehat{P}_{rat} \pm u_{1-\alpha/2} \cdot \sqrt{\frac{1}{A^2} \cdot \widehat{V}_\bullet}$$

will be conservative, if the model holds that y and a are strongly positively correlated.

V Ways to Improve the Precision of the Estimation of a Total

There are different ways to improve the precision of sample survey results regarding totals or functions of totals such as P (Section I). The first factor is the *sampling scheme* that is used to select the sample elements from the given population. An important component in this direction is the choice of the first-order sample inclusion probabilities π_k for all elements k in the population ($k \in U$). The best of choices for these probabilities is to determine them proportional to the size of the study variable y. In this case, when

$$\pi_k = \frac{y_k}{\sum_{k=1}^{N} y_k} \cdot n\,,$$

every sample even of size n = 1 would provide the perfect Horvitz-Thompson estimator of t because

$$t_{HT} = \frac{y_k}{\pi_k} = \frac{y_k}{\frac{y_k}{\sum_{k=1}^{N} y_k} \cdot 1} = \sum_{k=1}^{N} y_k = t$$

applies. Of course, as the y_k's are unknown in the population, the probabilities cannot be determined in this way. But, the estimator t_{HT} would also have a small variance, when the π_k's can be determined according to a known auxiliary variable a, which is approximately proportional to y.

Further improvement of the precision of the estimation of t can be achieved by a more efficient *estimator* compared to $\widehat{P}$ (Section I). One possibility is the ratio estimator $\widehat{P}_{rat}$ (Section II).

These factors for a better performance of an estimator are already implemented in the process described in the sections above. A third factor is the *structuring* of the population. On the one hand, for given n, *clustering* is often a source for a decrease of precision. Nevertheless, different aspects such as travel costs may certainly indicate its use. One the other hand, certain variants of *stratification* at different stages of a three-stage process may increase the precision of the estimator t_{HT}. Stratified simple random sampling with proportional allocation of the sample size n on the strata, for instance, is more efficient than an unrestricted simple random sample as long as the within-stratum mean values of the study variable y differ. The optimum allocation is achieved when the sample could be allocated on the strata proportional to the standard deviations of y in the strata. Hence, more auxiliary information would be necessary (cf. Särndal et al. 1992, Section 3.7.3).

Another possibility that has the potential to improve the accuracy of a given estimator is *post-stratification* (cf. ibid., Section 7.6). As an example, it may happen that, after a simple random sample is drawn from the population without stratification and the variable y of interest is observed therein, it turns out that the mean values of y differ between certain groups. For instance, in a survey on income, the sample means would differ between men and women included in the simple random sample. This means that regarding to the efficiency of the estimation of the overall mean it would have been better to stratify the sample proportional to the sizes of these two groups in the population already in the design-stage of the survey. Post-stratification means the implementation of this idea in the estimation-stage of the survey process. If too many men are randomly selected for the simple random sample, lower weights should be assigned to them. If there are too few women, higher weights should be assigned to them to increase the importance of this too small sample group. Obviously, the efficiency of post-stratification of an unrestricted simple random sample lies somewhere between a stratified and an unrestricted simple random sample.

For a three-stage process as described in Sections I and II, with proportional to size without replacement sampling and clustering at all three stages, proportional and optimum allocation and post-stratification are difficult to implement. The most probable way to include their effects in the estimation of the variance of the estimator would be based on simulation (Section III.4).

Hence, the most effective instrument of improving the precision of the estimation of a parameter t or P in the given three-stage process would be to increase the *sample size.*

References

Ardilly, P., and Tille, Y. (2006). *Sampling Methods: Exercises and Solutions*. Springer, New York.

Berger, Y.G. (2004). A simple variance estimator for unequal probability sampling without replacement. *Journal of Applied Statistics* 31(3), 305-315.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* 7, 1-26.

Gabler S. Hädler, S., and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology* 25(1), 105-106.

Ganninger, M., Häder, S., and Gabler, S. (2007). Design Effects and Interviewer Effects in the European Social Survey: Where are we now and where do we want to go tomorrow? Mannheim. Working paper.

Kish, L. (1987). Weighting in Deft^2. *The Survey Statistician*. The Newsletter of the International Association of Survey Statisticians, June.

Lohr, S. (2010). *Sampling: Design and Analysis*. 2nd edition. Brooks/Cole, Boston.

Quatember, A. (2014a). *Datenqualität in Stichprobenerhebungen*. Springer Spektrum, Berlin.

Quatember, A. (2014b). *Pseudo-Populations - A Basic Concept in Statistical Surveys* (submitted).

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

Shao, J, and Tu, D. (1995). *The Jackknife and Bootstrap*. Springer, New York.

Srinath, K.P., and Hidiroglou, M.H. (1980). Estimation of variance in multi-stage sampling. *Metrika* 27, 121-125.