Survey Methodology:
Sampling – Calibration – Estimation

Markus Gintas Šova[1] & Camille Vanderhoeft[2]

Part I: Theory

by Markus Gintas Šova

Київ – Kyiv 2004.XI.15-19

[1]ONS – Office for National Statistics, UK
e-mail: markus.sova@ons.gov.uk

[2]NSI – National Statistical Institute, Belgium
VUB – Free University of Brussels, Belgium
e-mail: Camille.Vanderhoeft@statbel.mineco.fgov.be

1. Basic Design-Based Sampling and Estimation

1.1 Introduction and Principles

The design-based approach to survey inference is based on the principle that each unit in the population has fixed values of its study variables. There are several reasons we may wish to use a sample, instead of a census, for inferences about the population. The main reason is one of cost. A full census can be costly and require a lot of resources. It is also possible for a well-designed sample to be more accurate than a census. This is because it may be possible to deal with data quality problems in a sample, but not in a census.

It is in the nature of sampling that different samples from the same population result in different estimates for population parameters. Given that the population values are fixed, some random process needs to be introduced for the estimates to have statistical properties. Hence we use a probability-based sampling selection method.

This course introduces a number of sampling and estimation methods for the design-based and model-assisted approaches. For a more in-depth coverage of these approaches to survey inference, the user is referred to Särndal *et al.* (1992) and Cochran (1977). These are widely regarded as the classic texts in the field.

1.2 Simple Random Sampling

Simple random sampling (SRS) is the simplest form of probability sampling. For a population of N units we select a single unit such that each unit in the population has a probability of $1/N$ of being selected. We then select a second unit from the remaining population with each having a probability of selection of $1/(N-1)$. Selection continues in this manner until a sample of n units is achieved. Thus each unit has a sample inclusion probability of n/N, and each possible sample of n units has the same probability of being selected.

If we wish to estimate the population mean of some study variable y, we use the sample mean as an estimator:

$$\hat{\bar{y}}_U = \bar{y}_s = \frac{1}{n}\sum_{i \in s} y_i \qquad (1.1)$$

Here U refers to the population, s to the sample, and y_i to the value of the study variable for unit i.

The population total t_y is the product of the population mean and the population size N. We may estimate the population total by substituting the estimated population mean into this relationship:

$$\hat{t}_y = N\hat{\bar{y}}_U = \frac{N}{n}\sum_{i\in s} y_i \tag{1.2}$$

Equation 1.2 is a weighted sum of sampled values, the weight for each unit's value being N/n. This weight is the *design weight* or *sampling weight*. It is a function of the sample selection method.

If we wish to estimate the proportion of the population which has a particular characteristic, we can regard each y_i as an indicator for unit i. In other words, y_i takes the value 1 if unit i has the characteristic, and it takes the value 0 if unit i does not have the characteristic. Estimating the proportion now becomes the task of estimating the population mean of the study variable, and equation 1.1 can be used.

The population variance is defined as:

$$\begin{aligned} S_{yU}^2 &= \frac{1}{N-1}\sum_{i\in U}\left(y_i - \bar{y}_U\right)^2 \\ &= \frac{\sum_{i\in U} y_i^2}{N-1} - \frac{\left(\sum_{i\in U} y_i\right)^2}{N\left(N-1\right)} \end{aligned} \tag{1.3}$$

This can be estimated using the sample variance:

$$\begin{aligned} S_{ys}^2 &= \frac{1}{n-1}\sum_{i\in s}\left(y_i - \bar{y}_s\right)^2 \\ &= \frac{\sum_{i\in s} y_i^2}{n-1} - \frac{\left(\sum_{i\in s} y_i\right)^2}{n\left(n-1\right)} \end{aligned} \tag{1.4}$$

This can be used to estimate the variance of the estimated population total:

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{t}_y\right) = N^2\frac{\left(1-f\right)}{n}S_{ys}^2 \tag{1.5}$$

or the variance of the estimated population mean (or proportion):

$$\hat{\mathrm{V}}\mathrm{ar}\left(\hat{\bar{y}}_U\right) = \frac{\left(1-f\right)}{n}S_{ys}^2 \tag{1.6}$$

Formulae 1.5 and 1.6 both refer to the *sampling fraction* $f=n/N$ which is the proportion of the population which is in the sample.

1.3 Stratified Simple Random Sampling

It is sometimes possible to divide a population into sub-populations according to some criterion (such that each population unit belongs to exactly one sub-population). If the population units within a sub-population are somehow similar, but those in different sub-populations are somehow less similar, we can use this information to refine our sampling and estimation. If sampling in each sub-population is performed independently of the sampling in other sub-populations, we refer to the sub-populations as *strata*. Neyman (1934) describes a framework for stratified simple random sampling (SSRS, that is simple random sampling separately performed within each stratum), rejecting the idea of a "generally representative sample", but promoting the idea of a "representative method of sampling".

Under stratification, the population total can be regarded as the sum of the stratum population totals. Thus under SSRS with H strata, using the subscript h to identify individual strata, we get the following population total estimator using equation (1.2):

$$\hat{t}_y = \sum_{h=1}^{H} \hat{t}_{yh} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i \in s_h} y_i \tag{1.7}$$

Note that the design weight N_h/n_h can be different for each stratum, but is the same for each sampled unit within any given stratum. Independence of sampling between strata gives the following variance estimate based on equation (1.5):

$$\hat{\mathrm{Var}}\left(\hat{t}_y\right) = \sum_{i=1}^{H} \hat{\mathrm{Var}}\left(\hat{t}_{yh}\right) = \sum_{h=1}^{H} N_h^2 \frac{\left(1-f_h\right)}{n_h} S_{ys_h}^2 \tag{1.8}$$

The use of stratification raises the question of how the sample should be divided between the strata. In other words, for a specified overall sample size n, what should be the sample size n_h in each stratum? One simple method is to allocate stratum sample sizes proportionally to the stratum population sizes:

$$\begin{aligned} & n_h \propto N_h \\ & \textit{i.e.} \quad n_h \approx N_h \frac{n}{N} \end{aligned} \tag{1.9}$$

The equality is only approximate because the stratum sample size has to be an integer, whereas the expression on the right hand side of equation (1.0) does not have to be. This is known as *proportional allocation*. Neyman (1934) showed that if the stratum population variances are known, optimal allocation is given by:

$$n_h \propto N_h \sqrt{S^2_{yU_h}}$$

$$i.e. \quad n_h \approx n \frac{N_h \sqrt{S^2_{yU_h}}}{\sum_{k=1}^{H} N_k \sqrt{S^2_{yU_k}}} \tag{1.10}$$

By "optimal" we mean the allocation which gives the smallest variance for the estimator of the population total. This allocation scheme is known as *Neyman allocation*. Although in practice the stratum population variances are unlikely to be known, they are often approximated using data from past surveys.

1.4 Unequal Probability Sampling

We saw in section 1.2 that under SRS, each population unit has the same *selection probability*, that is the same probability of being selected into the sample. In section 1.3 we saw that under SSRS population units may have different inclusion probabilities if they are in different strata. Horvitz & Thompson (1952) extend this idea further by allowing each population unit to have a different inclusion probability, which we denote by π_i for unit i. This is *unequal probability sample* (UPS), and gives rise to the *Horvitz-Thompson estimator* for a population total:

$$\hat{t}_{\pi y} = \sum_{i \in s} \frac{y_i}{\pi_i} \tag{1.11}$$

We note that each sampled unit has a weight equal to the inverse of its inclusion probability. Furthermore, knowledge of the population size N is not strictly necessary, and the sample size n may be a random variable, rather than fixed. It can be seen that the estimators 1.2 under SRS and 1.7 under SSRS are special cases of estimator 1.11. It is easily shown that the Horvitz-Thompson estimator is a design-unbiased estimator for the population total t provided that every unit in the population has an inclusion probability greater than zero. However, that is not to say that it is a good estimator for every sampling scheme, as exemplified by the story of Basu's elephants (Basu, 1971).

The Horvitz-Thompson estimator has the following variance:

$$\mathrm{Var}\left(\hat{t}_{\pi y}\right) = \sum_{i \in U} \sum_{j \in U} \frac{\left(\pi_{ij} - \pi_i \pi_j\right) y_i y_j}{\pi_i \pi_j} \tag{1.12}$$

which can be estimated by:

$$\hat{\mathrm{Var}}\left(\hat{t}_{\pi y}\right) = \sum_{i \in s} \sum_{j \in s} \frac{\left(\pi_{ij} - \pi_i \pi_j\right) y_i y_j}{\pi_{ij} \pi_i \pi_j} \tag{1.13}$$

Here, π_{ij} is the *joint inclusion probability* (or *second-order inclusion* probability) of units i and j, the probability that both units are selected into the sample. Unfortunately, the estimator 1.13 can take negative values if sampled units have high inclusion probabilities. The Sen-Yates-Grundy estimator for the variance is usually preferred to 1.13:

$$\hat{\mathrm{Var}}_{SYG}\left(\hat{t}_{\pi y}\right) = -\sum_{i \in s} \sum_{\substack{j \in s \\ j<i}} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j} \right)^2 \tag{1.14}$$

The Sen-Yates-Grundy estimator can also take negative values if $\pi_i \, \pi_{j<} \, \pi_{ij}$ for some $i \neq j$, but in most practical cases this does not occur. It is worth noting that the second-order inclusion probabilities (required for the variance estimators 1.13 and 1.14) can be difficult to estimate for some sampling schemes. Furthermore, the second-order inclusion probabilities have to be greater than zero for every pair of population units. If a sampling scheme has all first- and second-order inclusion probabilities greater than zero, it is referred to as a *measurable* sampling scheme.

Hájek produced an alternative to the Horvitz-Thompson estimator which reduces variance at the cost of bias. Hájek's estimator became known as the "circus estimator" due to it being a response to the story of Basu's elephants:

$$\hat{t}_{Hy} = N \frac{\sum_{i \in s} \frac{y_i}{\pi_i}}{\sum_{i \in s} \frac{1}{\pi_i}} \tag{1.15}$$

A commonly used unequal probability sampling scheme is *probability proportional to size* (PPS). This requires knowledge of some "size" variable for every unit in the population. Each unit is then assigned an inclusion probability proportional to its size value.

2. More Complicated Sampling and Estimation

2.1 Cluster Sampling

There can be circumstances where a sampling frame (that is, a list of all population units) does not exist, but a list of groups of population units is available. Or it may be that the population units are in compact groups, but with the groups scattered over a wide geographical area. In such circumstances it is sensible to implement a sample selection scheme which takes advantage of the information on the groups. One such scheme is *cluster sampling*. Each population unit belongs to exactly one *cluster*, and we have a sampling frame of clusters. From this frame a probability sample of clusters is selected, and every population unit in the selected clusters is surveyed.

As an example of cluster sampling, suppose we wish to conduct a survey on classroom sizes in schools in Ukraine. A list of schools may be available, but a list of classrooms would not be. We could select a sample of schools and then measure the size of every classroom in the selected schools.

Because all population units within the selected clusters are surveyed, the inclusion probability of a population unit is equal to the inclusion probability of the cluster to which it belongs. The Horvitz-Thompson estimator 1.11 can be used to estimate the population total.

2.2 Two-Stage Sampling

Returning to our example of a survey on classroom sizes, suppose that we only measure the size of a random sample of classrooms in each school (in order to reduce the amount of disruption to classes). This is an example of *two-stage sampling*. Under two-stage sampling we have *primary sampling units* (PSUs) instead of clusters. A sample of PSUs is selected according to some probability sampling scheme. This is the first stage of the sampling. Then, for the second stage of the sampling, a sample of *secondary sampling units* (SSUs, under two-stage sampling the SSUs are population units) is selected from each selected PSU according to some random sampling scheme. The first and second stages of sampling occur independently. Thus, the inclusion probability of a population unit is the product of its second stage selection probability and the first stage selection probability of its PSU. As with cluster sampling, the Horvitz-Thompson estimator 1.11 can be used to estimate the

population total. In fact, cluster sampling can be regarded as a special case of two-stage sampling, with all second stage selection probabilities being equal to one.

The principles of two-stage sampling can easily be extended to more than two stages, although care needs to be taken in calculating the second-order inclusion probabilities for variance estimation.

2.3 Domain Estimation

In a survey, there is often the requirement to estimate some parameter for a subset (or subsets) of the population, known as a *domain*. It is not always known which population units belong to the domain. For example, suppose we are conducting a survey about the heights of SSC employees, and we are interested in domains based on eye colour (blue, green, brown, etc.). The SSC will have a list of employees, but this is unlikely to include information on eye colour. Hence domain membership is identified only after the sample is selected and data collected. Introducing the subscript *d* to identify the domain, the Horvitz-Thompson estimator becomes:

$$\hat{t}_{\pi yd} = \sum_{i \in s_d} \frac{y_i}{\pi_i} \tag{2.1}$$

This requires no knowledge of the domain size N_d, which can be estimated by using each y_i as an indicator for unit i (in other words, y_i takes the value 1 if unit i belongs to domain d, and it takes the value 0 otherwise):

$$\hat{N}_d = \sum_{i \in s_d} \frac{1}{\pi_i} \tag{2.2}$$

To estimate the population domain mean we can use the sample domain mean:

$$\tilde{y}_{s_d} = \frac{\sum_{i \in s_d} \frac{y_i}{\pi_i}}{\sum_{i \in s_d} \frac{1}{\pi_i}} = \frac{\hat{t}_{\pi yd}}{\hat{N}_d} \tag{2.3}$$

As with the Horvitz-Thompson estimator, this requires no knowledge of the domain size. In fact, even if the domain size is known, this estimator is preferred to one which substitutes the true domain size for the estimated domain size. This is due to the positive correlation between the estimated domain total and the estimated domain size. These compensate for each other in the estimator 2.3; if the domain is under-

represented in the sample, both the estimated domain total and the estimated domain size will tend to be underestimated. However, this convenient property is lost if the true domain size is substituted for the estimated domain size. Consequently, when estimating a domain total and the domain size is known, the domain mean multiplied by the true domain total is preferred to the Horvitz-Thompson estimator:

$$\hat{t}_{alt,yd} = N_d \tilde{y}_{s_d} = \frac{N_d}{\hat{N}_d} \hat{t}_{\pi yd} \tag{2.4}$$

The variance of the estimator 2.4 can by approximated by:

$$\mathrm{Var}\left(\hat{t}_{alt,yd}\right) \approx \sum_{i \in U_d} \sum_{j \in U_d} \left(\pi_{ij} - \pi_i \pi_j\right) \left(\frac{y_i - \overline{y}_{U_d}}{\pi_i}\right) \left(\frac{y_j - \overline{y}_{U_d}}{\pi_j}\right) \tag{2.5}$$

which may be estimated by:

$$\hat{\mathrm{Var}}\left(\hat{t}_{alt,yd}\right) = \left(\frac{N_d}{\hat{N}_d}\right)^2 \sum_{i \in s_d} \sum_{j \in s_d} \frac{\pi_{ij} - \pi_i \pi_j}{\pi_{ij}} \left(\frac{y_i - \tilde{y}_{s_d}}{\pi_i}\right) \left(\frac{y_j - \tilde{y}_{s_d}}{\pi_j}\right) \tag{2.6}$$

3. Model-Assisted Estimation

3.1 Introduction

Sometimes a survey can draw on information about data which, although not of direct interest to the survey, may be in some way related to the study variable. A common example is *auxiliary data*. This is data which is available for every population unit. Alternatively, *benchmark data* may be available. This is where summary information (such as a population mean or total) is available for some variable which itself is not of primary interest to the survey. If there appears to be some relationship between the study variable and the auxiliary or benchmark variable, we should be able to use this to improve our estimation. This is achieved by using a model, but in doing so the purpose of the model must be carefully clarified. Särndal *et al.* (1992, p227) provide a good description:

> "The role of the model ξ is to describe the finite population point scatter. We hope that the model ξ fits the population reasonably well. We think that the finite population looks as if it might have been generated in accordance with the model ξ. However, the assumption is never made that the population was really generated by the model ξ."

This last sentence may appear counter-intuitive and in contradiction to the preceding sentences. The reason is that, as stated at the beginning of section 1.1, we have the principle that each unit in the population has fixed values of its study variables. Estimation is assisted by the model, but not based on it, giving the term *model-assisted* estimation.

3.2 Ratio Estimation

A very common model used between the study variable y and the auxiliary variable x is the ratio:

$$y_i = \beta x_i + e_i \quad \text{where } \operatorname{Var}(e_i) \propto x_i \tag{3.1}$$

Under SRS, the specification 3.1 gives the classic ratio estimator:

$$\hat{t}_{Ry} = \frac{\sum_{i \in s} y_i}{\sum_{i \in s} x_i} \sum_{i \in U} x_i \tag{3.2}$$

This is taking the ratio relationship between the study variable and auxiliary variable in the sample, and applying it to the auxiliary population total. Under UPS this generalises to Hájek's π-weighted ratio estimator:

$$\hat{t}_{\pi Ry} = \frac{\sum_{i \in s} y_i / \pi_i}{\sum_{i \in s} x_i / \pi_i} \sum_{i \in U} x_i = \hat{t}_{\pi y} \frac{t_x}{\hat{t}_{\pi x}} \tag{3.3}$$

It now becomes clear that this is a simple form of calibration; an estimate of the study variable total is being multiplied by a *model weight* which is the ratio of the true auxiliary total and an estimate of the auxiliary total. If the model weight is less than 0.5 or greater than 2.0, this is indicative of the sample being extremely unbalanced in terms of auxiliary data.

The combined ratio estimator is a special case of Hájek's π-weighted ratio estimator under SSRS, and assumes that several strata share the same ratio slope.

The ratio estimator is not design-unbiased. However, for large samples it is approximately design-unbiased. A sample size of at least 20 is recommended for this bias to be ignorable.

The variance of Hájek's π-weighted ratio estimator can be approximated by:

$$\mathrm{Var}\left(\hat{t}_{\pi Ry}\right) \approx \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\left(y_i-\beta x_i\right)\left(y_j-\beta x_j\right) \tag{3.4}$$

which can be estimated by:

$$\hat{\mathrm{Var}}\left(\hat{t}_{\pi Ry}\right) = \left(\frac{t_x}{\hat{t}_{\pi x}}\right)^2 \sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right)\left(y_i-\hat{\beta}x_i\right)\left(y_j-\hat{\beta}x_j\right)$$

$$\text{noting that}\quad \hat{\beta} = \frac{\sum_{i\in s} y_i/\pi_i}{\sum_{i\in s} x_i/\pi_i} \tag{3.5}$$

3.3 Linear Regression Estimation

Suppose we wish to assume a linear relationship between the study variable and the auxiliary variable, without requiring the line to pass through the origin:

$$y_i = \alpha + \beta x_i + e_i \quad \text{where } \mathrm{Var}\left(e_i\right) = \sigma^2 \ \forall i \tag{3.6}$$

Under SRS, the specification 3.6 gives the linear regression estimator:

$$\hat{t}_{Ly} = N\left(\bar{y}_s + \hat{\beta}\left(\bar{x}_U - \bar{x}_s\right)\right) \tag{3.7}$$

Under UPS, this generalises to the π-weighted linear regression estimator:

$$\hat{t}_{\pi Ly} = \hat{t}_{\pi y} + \left(t_x - \hat{t}_{\pi x}\right)\hat{\beta}$$

$$\text{where}\quad \hat{\beta} = \frac{\sum_{i\in s}\left(x_i-\tilde{x}_s\right)\left(y_i-\tilde{y}_s\right)/\pi_i}{\sum_{i\in s}\left(x_i-\tilde{x}_s\right)^2/\pi_i} \tag{3.8}$$

The model weight for population unit *i* is:

$$g_i = \left(\frac{N}{\hat{N}}\right)\left(1+\frac{\hat{N}\left(\bar{x}_U-\tilde{x}_s\right)\left(x_i-\tilde{x}_s\right)}{\sum_{j\in s}\left(x_j-\tilde{x}_s\right)^2/\pi_j}\right) \tag{3.9}$$

Note that the model weight is different for every population unit. For household surveys, these model weights are usually between 0.7 and 1.4. For business surveys, they tend to be within 0.3 and 3.0, the wider range being due to the heterogeneity of businesses. It is possible for some model weights to be negative. This requires the

sample to be unbalanced and a population unit to be an outlier in the same direction as the unbalancedness.

The variance of the π-weighted linear regression estimator can be approximated by:

$$\mathrm{Var}\left(\hat{t}_{\pi Ly}\right) \approx \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\left(y_i-\alpha-\beta x_i\right)\left(y_j-\alpha-\beta x_j\right) \tag{3.10}$$

which can be estimated by:

$$\hat{\mathrm{Var}}\left(\hat{t}_{\pi Ly}\right) = \sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right) g_i g_j \left(y_i-\hat{\alpha}-\hat{\beta}x_i\right)\left(y_j-\hat{\alpha}-\hat{\beta}x_j\right) \tag{3.11}$$

noting that $\hat{\alpha} = \tilde{y}_s - \hat{\beta}\tilde{x}_s$

3.4 General Multivariate Regression (GREG) Estimation

Suppose we wish to assume a multivariate linear relationship between the study variable y_i and an auxiliary vector $\mathbf{x}_i$:

$$y_i = \mathbf{x}_i'\mathbf{B} + e_i \quad \text{where } \mathrm{Var}\left(e_i\right) = \sigma_i^2 \tag{3.12}$$

Under UPS, the specification 3.12 gives the general multivariate regression estimator, which is also known as the generalised regression estimator (GREG estimator):

$$\begin{aligned} &\hat{t}_{\pi My} = \hat{t}_{\pi y} + \left(\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\pi x}\right)' \hat{\mathbf{B}} \\ &\text{where } \hat{\mathbf{B}} = \left(\sum_{i\in s}\frac{\mathbf{x}_i\mathbf{x}_i'}{\sigma_i^2\pi_i}\right)^{-1}\sum_{i\in s}\frac{\mathbf{x}_i y_i}{\sigma_i^2\pi_i} \end{aligned} \tag{3.13}$$

The model weight for population unit *i* is:

$$g_i = 1 + \left(\mathbf{t}_x - \hat{\mathbf{t}}_{\pi x}\right)'\left(\sum_{j\in s}\frac{\mathbf{x}_j\mathbf{x}_j'}{\sigma_j^2\pi_j}\right)^{-1}\frac{\mathbf{x}_i}{\sigma_i^2} \tag{3.14}$$

The variance of this estimator can be approximated by:

$$\mathrm{Var}\left(\hat{t}_{\pi My}\right) \approx \sum_{i\in U}\sum_{j\in U}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_i\pi_j}\right)\left(y_i-\mathbf{x}_i'\mathbf{B}\right)\left(y_j-\mathbf{x}_j'\mathbf{B}\right) \tag{3.15}$$

which can be estimated by:

$$\hat{\mathrm{Var}}\left(\hat{t}_{\pi My}\right) = \sum_{i\in s}\sum_{j\in s}\left(\frac{\pi_{ij}-\pi_i\pi_j}{\pi_{ij}\pi_i\pi_j}\right) g_i g_j \left(y_i-\mathbf{x}_i'\hat{\mathbf{B}}\right)\left(y_j-\mathbf{x}_j'\hat{\mathbf{B}}\right) \tag{3.16}$$

4. Calibration

4.1 The GREG Estimator as an Example of Calibration

Estimation using calibration is based on two principles. The first is the desire for the model weights to be as close as possible to 1.0 (and in particular to be positive), meaning that the product of weights applied to a population unit should be similar to its design weight alone. The second principle is that the weighted sum of auxiliary vectors should be equal to the population auxiliary vector total. This requirement is the definition of the calibration constraints and can be written algebraically:

$$\sum_{i \in s} \frac{g_i}{\pi_i} \mathbf{x}_i = \sum_{i \in U} \mathbf{x}_i \qquad (4.1)$$

The motivation for these calibration constraints is explained by Deville & Särndal (1992):

> "the calibrated weights must give perfect estimates when applied to each auxiliary variable. That is a consistency check that appeals to many practitioners, because a strong correlation between the auxiliary variables and the study variable means that the weights that perform well for the auxiliary variable also should perform well for the study variable."

The two principles are combined by minimising a distance function with respect to the model weight, subject to the calibration constraints 4.1. A commonly used distance function is:

$$\sum_{i \in s} \frac{\left(g_i / \pi_i - 1 / \pi_i \right)^2 \sigma_i^2}{1 / \pi_i} \qquad (4.2)$$

The use of the distance function 4.2 combined with calibration constraints 4.1 gives a unique set of model weights which are identical to those of the GREG estimator (equation 3.14). Thus although no model has been explicitly stated, calibration using the distance function 4.2 implies the general multivariate regression model 3.12

4.2 Calibration Estimators

In section 4.1 it was explained how the GREG estimator is an example of a calibration estimator. More generally, calibration can be performed using distance functions other than 4.2. An explicit model (such as 3.12) does not have to be stated, although one is

implied given proper regularity conditions (see Deville & Särndal, 1992, for details of these conditions and examples of distance functions). Calibration can be further refined by allowing the product of the weights to be further constrained by upper and/or lower limits:

$$l_i \leq \frac{g_i}{\pi_i} \leq u_i \qquad (4.3)$$

These limits may be different for each population unit. However, it may be that there is no solution which minimises a particular distance function, subject to the calibration constraints 4.1 and bounded by 4.3

4.3 Raking (Iterative Proportional Scaling)

Raking (also known as "iterative proportional scaling") is a method for estimating the value of cells in a (possibly multi-dimensional) table, subject to pre-defined margins and a set of starting values in each cell. The two-dimensional algorithm is as follows:

1. Each cell in the first row of the table is multiplied by the ratio of the required margin for the row to the current row total. This means that the row now sums to the required margin.
2. Each remaining cell is re-scaled in a similar way, so that each row now sums to its required margin.
3. Each cell is now re-scaled in a similar way, so that each column now sums to its required margin.
4. The process is repeated from step 1 until convergence.

The multi-dimensional algorithm works in the same way, with each dimension "raked" in sequence until convergence. It should be noted that the solution eventually reached depends on the starting values. However, raking is regarded as a good method as long as there are no zero starting values in the cells.

Rather than use the algorithm described above, raking can be implemented using calibration estimation.

5. Non-Response and Response Homogeneity Groups

In practice, surveys are usually subject to non-response. That is, there are some population units which have been selected into the sample, but from whom no data is received. Some surveys treat each population unit as having the same probability of responding, given that the unit is selected. In this case, the response set (the set of sampled units from whom we receive data) is similar to a two-stage sample. The first stage being sample selection, and the second stage being receiving a response. The second stage inclusion probabilities are equal for all units.

In reality, one might expect the probability of responding to be related to the study variable y. For example, in a survey on foreign travel, people may be less likely to respond if they have visited certain countries. This will lead to a serious estimation bias if equal response probabilities are assumed. Some model for the non-response mechanism is required. The difficulty is the lack of study variable data from the non-responders with which to fit a model. However, a simple but useful model is that of *response homogeneity groups* (RHG). Under RHG, the sample is partitioned into a number of groups. Units within a group are assumed to have the same response probability, but this response probability can be different for different groups. Again, this is similar to two-stage sampling, the PSUs being the groups and the SSUs being the units, although with RHGs the groups can be defined after sample selection has taken place. We therefore need to estimate the response probability θ for each group z:

$$\theta_z = \frac{\text{number of responding units in group } z}{\text{number of selected units in group } z} \tag{5.1}$$

This allows us to estimate a combined probability π^* for all responding units:

$$\pi_i^* = \pi_i \theta_z \tag{5.2}$$

Finally, π^* can be substituted for π in the estimation equations.

References

Basu P. (1971) An Essay on the Logical Foundations of Survey Sampling, part one (with discussion), in *Foundations of Statistical Inference* (eds. Godambe, V.P. & Sprott, D.A.). Holt, Rinehart & Winston, Toronto, 203-242.

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

Deville, J.-C. & Särndal, C.-E. (1992) Calibration Estimators in Survey Sampling, *Journal of the American Statistical Association*, **87**, 376-382.

Horvitz, D.G. & Thompson, D.J. (1952) A Generalization of Sampling Without Replacement from a Finite Universe, *Journal of the American Statistical Association*, **47**, 663-685.

Neyman, J. (1934) On the Two Different Aspects of the Representative Method: the Method of Stratified Sampling and the Method of Purposive Selection, *Journal of the Royal Statistical Society*, **97**, 558-606.

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