PHARE 2002
Multi Beneficiary Statistics Programme
(Lot 1[1])

Quality in Statistics

Task T3: Development of Methods for Quality Assessment

Sampling Issues in Business Surveys

June 2005

prepared by

Jörgen Dalén
Project Expert

[1] Pilot Project 1 of the European Community's Phare 2002 Multi Beneficiary Statistics Programme (Lot 1) is devoted to "Quality Assessment of Statistics" in ten Beneficiary Countries, namely Bulgaria, the Czech Republic, Estonia, Hungary, Latvia, Lithuania, Poland, Romania, Slovakia and Slovenia.

Contents

SAMPLING ISSUES IN BUSINESS SURVEYS 3

Business surveys and target variables 3

Some common sampling methods 4
Simple random sampling and systematic sampling 4
Stratified sampling 5
Probability proportional to size sampling 5
Cutoff sampling 7
Mixed designs 7

Dynamic sampling issues (co-ordinated samples) 8
The permanent random number method 9

Estimators 10
The Horvitz Thompson estimator 10
Ratio and regression estimators 11

Optimum allocation 12
Theoretical results 12
Sample allocation for businesses in practice 13

EU perspective: comparability problems 18

Further reading 20
On πps sampling 20
On business surveys in general 21
On sample rotation 21

Sampling Issues in Business Surveys

Sample surveys conducted by statistical agencies (NSIs), especially those that are of common interest in a harmonised European context, are usually repetitive surveys, carried out at certain frequencies for example monthly, quarterly or annually. This note aims at providing a brief overview over the problems that emerge when designing such surveys in European countries. We will focus on business surveys aimed at estimating economic variables, which are a key part of the European Statistics, where a common framework, high quality and comparability are needed.

We emphasize right at the outset that, given the vast literature on sampling techniques, the purpose of the present note is only to give a brief practical overview over some important sampling issues in official statistics.

Business surveys and target variables

The purpose of economic surveys is to provide an up-to-date picture of the continually changing economy of a country. We want to know the current state of the economy which leads to a demand for *estimates of level* of *variables* such as production, employment, export, import, investment and their distribution between industries and regions. But, and often of even greater importance, we want to know how these variables change over time, which leads to a demand for *estimates of change*. This level/change duality is an important consideration when designing continuing business surveys.

In sampling statistical terms a level usually corresponds to a *target parameter*, which is a sum of variable values over a finite population. Levels for specific industries or regions correspond to sums of variable values over subdomains of this population. Certain kinds of levels such as percentage shares of subdomains and the like also takes the form of ratios between two such sums. In practice the estimates of these ratios will usually be taken from the same sample.

A change is also a ratio but in the time dimension. It refers to the ratio between levels in two periods. An important statistical complication with estimates of change is that the population changes between the periods, to a greater or smaller extent. This means that the same unchanged

sample cannot without some bias represent both periods. In practice, estimates of change are sometimes (typically between consecutive months or quarters) based on the same sample, sometimes based on different samples (typically between years).

A survey has almost always multiple purposes, at least in the sense that a number of different levels (for the whole population as well as subdomains) as well as changes over time (from last month, from same month last year etc.) are required to be estimated. This fact complicates the search for the best designs of a survey, since what is best for one kind of target parameter could be less than optimal for another parameter

The size of the business is a crucial aspect in any business survey. It correlates more or less strongly with most target variables and needs to be taken into account in the sampling or the estimation stage, or both. Different approaches in this regard will be discussed in this note.

Some common sampling methods

Simple random sampling and systematic sampling

In *simple random sampling* (srs) each unit of the population is drawn with the same inclusion probability. Sampling is in practice always without replacement (*srswr*), since we do not want multiple representations of the same unit.

In systematic sampling (*ss*), units in the frame are included at fixed intervals according to the sampling fraction. A random starting point is taken. For example, if the sampling fraction is 0.1 a random starting point from 1 to 10 (the inverse if 0.1) is taken. Say that we obtain 8 as our starting point. The sample will then consist of units number 8, 18, 28, etc. in the frame. If the frame order is effectively random, then ss and srs are in practice equivalent procedures.

It is highly unusual to use only srswr or ss as the sampling designs in official surveys. This is especially the case in business surveys, where the size of the sampling unit is correlated with the target variable. But srswr is often a component of a composite sampling design such as stratified designs.

Stratified sampling

In stratified sampling the population is divided into non-overlapping subpopulations called strata. In business surveys stratification serves at least three different purposes:

- To ascertain that certain subdomains in the population are adequately represented by a minimum sample size. Subdomains are often industries at a certain level of the standard industrial classification but could also be regions of a country.
- To allow for different sample shares of units of different *size*. Often the very largest businesses are sampled with certainty in a so-called take-all stratum. Below the take-all stratum successively smaller sample shares are used for smaller businesses. If the measure of size is the number of employees, size stratification could look like: i) more than 250, ii) 50-249, iii) 10-49, iv) 0-9 employees. Size is often combined with industry into a matrix stratification according to industry x size. The size stratification does not need to be the same for all industries.

  Within strata srswr is often used. This type of design gives rise to the *optimum allocation* problem. Optimum allocation refers to i) choosing the best number of strata, ii) setting the appropriate stratum boundaries and iii) allocating the total sample (or more generally the total budget) to different strata. This problem is discussed in more detail below.
- To allow for different detailed sampling or measurement designs in different strata. It is often the case that different industries or smaller businesses need to be approached with different questionnaires. They may also need to be drawn from different sampling frames. A stratified design allows for greater flexibility in adjusting the sample design to the practical needs and possibilities at hand.

Probability proportional to size sampling

An alternative to size stratification is πps[2] sampling, where πps stands for probability (of inclusion in the sample) proportional to size. (The term is sometimes used also for designs which only give approximate proportionality to size.) In fact πps sampling is not a single method but rather a class of different methods depending of how the πps mechanism is defined. Πps methods are not used extensively today but they have some advantages, which motivates us to give them a brief mention here and they are actually used in some surveys.

[2] Also called pps sampling

An obvious advantage of πps sampling is that it allows for a continuous relationship between size and inclusion probability. This is in contrast to size stratification where, for example, a company with 50 employees may have the same inclusion probability as one with 249 employees.

A disadvantage of πps sampling, on the other hand, is that it is difficult to achieve a fixed sample size with a relatively simple sampling method. Imagine, for example, the most simple of πps designs (called *Poisson sampling*), where the whole population is run through by a computer program which includes unit j in the sample if a uniform random variable takes on a value x_j which is smaller than the predefined inclusion probability π_j for that unit. Obviously, then, the size of the sample is also random (with expected value $n=\Sigma^N_{j=1}\ \pi_j$). Random sample sizes are clearly undesirable, since they reduce one's control over the sample and result in random cost and random accuracy as well.

An approximately πps sampling design, which allows for fixed sample sizes is *order πps sampling*. Exhibit 1 provides some more detail about this sampling design. Order sampling procedures are not exactly πps, but in samples of sufficient size they can be shown to be approximately πps. Order πps sampling is gaining increased popularity as a practical and efficient sampling design, for example in Swedish price and business surveys.

Exhibit 1: Order πps sampling

Here we will only describe the special case of it (its application to the Swedish CPI). A uniform random number $\boldsymbol{U_i}$ between 0 and 1 and a variable $\boldsymbol{z_i=nx_i/\Sigma_i x_i}$, where $\boldsymbol{x_i}$ is a size measure, are associated with each sampling unit i and a *ranking variable* $\boldsymbol{Q}$ is constructed as a function of $\boldsymbol{U}$ and $\boldsymbol{z}$. The units in the universe are then sorted in ascending order and the $\boldsymbol{n}$ units with the smallest value of the ranking variable are included in the sample. Two important examples of such ranking variables are:

$$Q_i = \frac{U_i}{z_i} \text{ and } \quad Q_i = \frac{U_i(1-z_i)}{z_i(1-U_i)}$$

Units, where $\boldsymbol{z_i \geq 1}$ are first included with certainty and excluded from the frame. The procedure is then repeated until there are no such units in the frame after which the sampling procedure takes place according to (14) or (15).

The second variant of ranking variable $\boldsymbol{Q_i}$ (sometimes referred to as Pareto πps) is a marginally better choice and is therefore normally preferred.

A practical advantage of order sampling is that it is easy to handle out-of-scope units, which are discovered after the sample is drawn (but before collecting the observations). Such units are just omitted and replaced by the units further down the ordered list so that the sample size remains the intended one.

Cut-off sampling

A very common element of business sampling designs is cut-off sampling. In its pure form it amounts to including all units above a certain size threshold with certainty but none below that threshold. The intuition behind this method is that the variable distributions of companies are often very skew, with a small number of companies accounting for perhaps 80 or 90 percent of the target parameter value. Where an estimate of change is the first priority, the change in the largest units may sometimes adequately represent that of the whole population.

The obvious drawback of this method, however, is that no design-based inference to the whole population is possible and thus not either an objective measure of accuracy.

The definition of cut-off sampling is somewhat blurred by the fact that one could always reduce the target population to those above a certain size and then restrict the inferences made to this reduced population. However, typically it is the case that the population of interest to the users is really “all companies”, i.e., also those below the cut-off threshold.

Mixed designs

The methods mentioned above are often not used in their pure form. For example, we could have

- Industry + size stratification combined with srswr within each stratum or
- Industry stratification + πps sampling within each stratum.

In both cases some strata or units could be drawn with certainty (all units included).

Also, cut-off thresholds are often employed as an ingredient in a composite sampling strategy. One may for example have a stratified design including the largest businesses with certainty, taking a probability sample (srs or πps) of the medium ones and excluding the smallest businesses below a certain cut-off threshold. Such a design makes sense when there are strong reasons to believe that the smallest businesses contribute very little to the target parameter value or where there are great difficulties in obtaining good responses from the smallest businesses.

Dynamic sampling issues (co-ordinated samples)

Most business surveys are repetitive, i.e., they are carried out according to basically the same design over long time. This fact gives rise to dynamic sampling issues such as.

- How often should a new sample be drawn?
- Should the whole sample be replaced at the same time or should it be refreshed gradually, with some overlap?
- Which technique should be used in order to obtain the desired overlap and rotation while at the same time maintaining the basic probability properties of the sample?

Which are the considerations to judge issues like this? We suggest that there are three important criteria here.

Optimise the accuracy of the sample. The older the sample becomes, the less well will it represent the population, where new units are entering and old ones disappearing. An old sample will have a large under- and over coverage.

On the other hand an estimate of change has smaller variance if the same sampling units are included in both time periods. (But also the estimate of change has a bias caused by not accounting for new units entering the population.)

Distribute the response burden. Companies may get worn out by responding and for this reason need to be rotated out after a certain period.

Minimise initiation costs. Responding to a complicated economic survey is a learning process, so that the time required to fill in the questionnaire is longest the first time and then decreases in subsequent response waves. Also the time needed for the statistical agency to initiate a new company into the survey, finding the right contact person etc. is a factor here.

We can immediately see that these considerations run into conflict with each other. The last one provides a reason for never changing the sample, whereas the first two call for rotation at some suitable time intervals. A practical trade-off between them has to be worked out.

We will give one example of an automatic rotation technique here. Other methods are briefly mentioned in ***Further reading***, below:

The permanent random number method

With the permanent random number (*prn*) method, each unit i (company, location etc.) in the sampling frame is assigned a random number X_i drawn independently from the uniform distribution on the interval [0,1]. The frame units are then sorted in ascending order of the X_i. The sample is composed of the first n_h units in the ordered list above a preset value C, which is unique for a certain survey. This procedure is repeated for each stratum h in the survey. It is proved that this technique produces a simple random sample without replacement.

This technique is used by Statistics Sweden in its so called SAMU system for sampling co-ordination and rotation over time. The basic idea is that the random numbers X_i are permanent, i.e. retained over time. On each sampling occasion (normally once a year) the permanent random numbers are used to select a new sample. If there were no changes in sample size or in the sampling frame the samples would be exactly the same on all sampling occasions. In reality, there will be changes due to i) changes in sample size, ii) births and deaths among sampling units and iii) changes in size or kind of activity of a company resulting in a change of stratum. Nevertheless, a high degree of overlap between samples of consecutive years is achieved, which is desirable since there are considerable initiation costs associated with bringing new companies into a survey and since a high degree of overlap results in higher precision for estimates of change.

The second advantage of the prn method is that the starting value C could be chosen differently for different surveys, resulting in a more fair distribution of response burden among companies. If, e.g., the sampling fraction is 10 % and the difference between the C values of two surveys is 0.5, then the risk for overlap between the two samples would be quite small. This effect is strongest for small companies but normally these companies are also those that are most sensitive to a large response burden.

Yet another way to utilise the prn method is to rotate the sample after some time. This is simply achieved through moving the starting value C for the survey by a certain amount after a certain number of years. This results in a more or less completely new sample, especially for small companies.

A caveat for this method is that neither the distribution of response burden nor the rotation is guaranteed for every single company although achieved for the great majority of especially small companies. Due to births, deaths and stratum changes, it could happen that a certain company could still be included in many surveys at the same time or not be rotated out as intended.

Estimators

In this section we give a very brief overview over some common estimators used in survey sampling.

The Horvitz Thompson estimator

A very general type of estimator is the Horvitz-Thompson (HT) estimator. We assume that we are estimating a population (denoted U) total of a study variable y_k, $t=\Sigma_U y_k$ and start by defining an inclusion probability π_k for each element of the population of N units. If $\pi_k>0$ for all population units then the following HT estimator can be shown to be unbiased with respect to the sampling design[3]:

$$\hat{t} = \sum_s \frac{y_k}{\pi_k}, \text{ where summation is over the sample } s, \text{ containing } n \text{ units.} \qquad (1)$$

Under srswr, $\pi_k=n/N$ for all k and the HT estimator takes the following form:

$$\hat{t} = \frac{N}{n}\sum_s y_k \qquad (2)$$

Under stratified sampling with srs in each stratum h with population size N_h and sample size n_h we have $\pi_k=n_h/N_h$ for all k and the HT estimator becomes

$$\hat{t} = \sum_h \frac{N_h}{n_h}\sum_{s_h} y_k \qquad (3)$$

The factor N_h/n_h, which serves to enlarge the sample sum in order to cover the whole population, is sometimes called the *expansion factor*.

Equations (1) to (3) assume that we have no other information available than the population sizes (and the stratification variable in (3)). But a common situation is that there exists prior information for one or several auxiliary variables, which are correlated with the study variable. We will only discuss the situation with one auxiliary variable, denoted x_k, here, referring the reader to common

[3] The definition of unbiasedness with respect to the sampling design is that the expected value of the estimator over all sample outcomes is equal to the desired population parameter.

textbooks for the situation with several variables. One way of using the auxiliary variable is in πps sampling, where we define the inclusion probability $\pi_k = \frac{nx_k}{\sum_U x_k}$, i.e., proportional to x_k.[4]

Ratio and regression estimators

Πps sampling uses the auxiliary variable in the sampling stage. Another option is to use it in the estimation stage. Two common estimators of this kind are the ratio and the regression estimators. For illustration we provide a simple illustration here of how they work in the case of only one auxiliary variable.

The *ratio estimator* takes the inverse sample fraction of the auxiliary variable as the expansion factor:

$$\hat{t} = \frac{\sum_U x_k}{\sum_s x_k} \sum_s y_k \qquad (4)$$

A more general way of using auxiliary information is through a *regression estimator*. A regression estimator brings in the linear relationship between x and y and relies on a particular regression model for this relationship. For example, in the one-variable case, if $y_k=\beta_1+\beta_2 x_k$, an srswr design is used, and under some other assumptions, the *general regression estimator* takes on the form

$$\hat{t} = \frac{N}{n}\sum_s y_k + \hat{\beta}_2\left(\sum_U x_k - \frac{N}{n}\sum_s x_k\right) \qquad (5)$$

To provide some intuition for this estimator, one could look at it as the simple HT estimator according to (2), adjusted for the fact that we may have obtained a somewhat "biased sample" with regard to the auxiliary variable. The adjustment factor within the brackets is then the difference between the population sum of x_k and the HT estimate of this sum, multiplied with the estimated regression coefficient for x.

With regression estimators, statistical models are brought into play. An important distinction in modern sampling theory is between *design-based inference*, which relies on the inclusion probabilities of the sampling design for deriving the properties of estimators and *model-based inference*, which instead relies on the statistical model (for example the regression model). The general regression estimator is often referred to as a *model-assisted* estimator, which is approxi-

[4] If this expression is larger than 1, we instead set π_k=1, obtaining a take-all stratum. The probabilities are then redefined to refer to the sum of the x_k of the remaining part of the population. This procedure may have to be repeated a number of times until there are no more π_k>1.

mately unbiased under the sampling design but at the same time with good accuracy under the statistical model used.

Optimum allocation

Given the specific circumstances for a certain survey one always wants to choose the best sampling strategy. The relevant circumstances include

- The set of study variables for the survey and the priority between them.
- The set of subdomains for which estimates with given accuracy are needed.
- The priority between estimates of level and estimates of change.
- The sampling frames, including auxiliary variables, which could be used.
- The budget available for the survey.

Unfortunately this general problem has too little structure for a general theoretical analysis. The issue of optimum allocation usually refers to a narrower problem, where the sampling design and estimator are already defined and the problem is how to distribute the sample over the population, or equivalently how to specify the inclusion probabilities for all the population units.

Theoretical results

The classical case for an analysis of optimum allocation is for stratified sampling, where srswr is used for sampling within strata. There are then three issues:

- How many strata?
- Which boundaries between strata?
- How to allocate the total sample (total budget) to the different strata?

Concerning the first question, a first consideration is normally to allow at least one stratum to each domain (subpopulation) for which estimates are desired. Within domains, precision theoretically increases (variance decreases) without bound as the number of strata increases. However, this theoretical gain in precision is very small beyond a certain point. The need for having sufficient sample sizes in each stratum (with consideration to possible non-response and over-coverage) often calls for a minimum stratum size of about five. If this is far from the optimum size then too many strata will instead result in loss of efficiency.

For the second question, the current best answer is still the Dalenius-Hodges (*cum√f*) or Ekman rules. An example of how they work could be found in any sampling textbook, see ***Further reading***, below.

For the third question, the theoretical answer is *Neyman allocation*. Neyman allocation comes in two forms, either with constant cost per sampling unit or with variable cost per unit in different strata. In the first version total sample size is fixed (equal to n) and the sample sizes n_h per stratum are determined by

$$n_h = n\frac{N_h\sigma_h}{\sum N_h\sigma_h}, \qquad (6)$$

where N_h is the stratum population size and σ_h is the stratum standard deviation.

If instead of the total sample size being fixed, the total cost C is fixed at

$$C = c_0 + \sum n_h c_h, \qquad (7)$$

where c_0 is the fixed (overhead) cost, and c_h the unit cost in stratum h, then the optimum sample sizes instead become

$$n_h = n\frac{N_h\sigma_h/\sqrt{c_h}}{\sum N_h\sigma_h/\sqrt{c_h}} \text{ and } n = \frac{(C-c_0)\sum N_h\sigma_h/\sqrt{c_h}}{\sum N_h\sigma_h\sqrt{c_h}} \qquad (8)$$

Sample allocation for businesses in practice

The above classical results are unattainable in practice. The primary reason for this is that the statistical distribution of the study variable y is by definition unknown (y is what we want to estimate!). In practice, one then has to look for an approximately optimal solution. A helpful general result is that optima are normally flat, i.e., small deviations from the exact optima do not result in large losses of precision.

Other complications, when looking for the best allocation are:

- There are often several study variables in a survey and the optimum allocation for one variable may differ from that for another variable.
- There are often several domains of study (subpopulations) for which estimates are desired, in addition for an estimate for the whole population. In the business survey case, such domains are often particular industries, defined by their NACE code. This leads to

requirements for minimum sample sizes for each of these domains of study.

There are a fairly large number of ideas for achieving good allocation solutions in practice, which are described in the literature. Many of them are summarised in the paper by Sigman and Monsour (1995), presented in the section ***Further reading*** below. Like Sigman and Monsour we will next discuss the one-variable problem, the many-variable problem and the many-domain problem each in separate sections.

The one-variable problem

For the one-variable problem, the normal situation is that we have access to an auxiliary variable, which is more or less correlated with the study variable. If we allocate according to the standard deviation of this auxiliary variable, we disregard the additional variance which comes from the less than perfect relationship between the study variable and the auxiliary variable.

This additional variance usually has the effect of reducing the difference between the stratum variances. If this is true, the optimum sampling allocation with respect to the study variable would imply bigger sampling fractions among smaller companies than a "Neyman allocation" according to the auxiliary variable.

We will now briefly go through some of the possible allocation strategies and approximations which could be used in practice for business surveys.

Neyman allocation according to the σ_h of the target variable for the last period. With this approach in a periodical survey, the sample is reallocated for each survey round according to the estimated σ_h for the previous period. There are several weaknesses of such an approach that explain why it is rarely used. Firstly, the estimated σ_h are often unstable and the estimate for last year is not necessarily a better estimate of this year's variance than the estimate for two years ago. Secondly, the approach would lead to stratum sample sizes jumping up and down for no good reason, which causes problems for sample co-ordination over time. *We therefore advise against this strategy.*

Neyman allocation according to average σ_h of the target variable over a number of previous periods. (Instead of averaging the σ_h, one could average their squares, the stratum variances.) This

approach is an improvement over the former under the assumption that the real σ_h move fairly slowly over time and that averages over several years are therefore better estimates because of larger underlying sample sizes. Still, it would be advisable not to revise the allocation every year but at planned time intervals such as every third or fifth year. In this way reallocation can also be combined with other considerations, such as sample co-ordination. (Of course, some minor adjustments of the sample every year due to changes in the stratum population sizes may still be necessary. In particular, the take-all strata need to cover all businesses that are currently above the take-all threshold.) *Under the right circumstances we would recommend this approach but note that a number of previous survey periods are needed before it can be applied.*

Neyman allocation according to the stratification variable. This means that the σ_h are calculated for the stratification variable instead of the target variable. The advantage is that the stratification variable is known for the whole population so no estimate is needed. If further there is a strong correlation between the stratification and the target variables, then the allocation will be close to optimal. In practice, however, the correlation is often not so strong and then this allocation method could be far away from optimum and, as mentioned above, it would lead to underallocating small and medium companies. It must also be warned against using the stratification variable variances for estimating the final estimator variances. They will normally be much higher due to the large spread of the target variable within strata compared with the stratification variable itself. *For these reasons we advise against this approach.*

X-proportional allocation. This allocation calls for stratum sample sizes to be proportional to the stratum sum of a measure of size X_{hj}, often the stratification variable itself. This is a similar approach as the previous one. If stratum boundaries are relatively close, the σ_h for the stratification variable may become artificially low and be very poor predictors of the σ_h for the target variable. The stratum sum of X could then be a better alternative. *If very little is known about the distribution of the study variable, for example because it is the first time the survey is done, then this approach may be the best that could be achieved.*

N-proportional allocation. This allocation method simply takes the stratum sample sizes to be proportional to the stratum population sizes. This is clearly inappropriate between size-determined strata. However, between industries or regions it could be a reasonable option, in case the importance of estimates for industries/regions is related to their size.

√*N-proportional allocation.* In case there are also distinct precision requirements for small industries/ regions, there is a need for overrepresenting them in the sample. Allocation proportional to the square root of the stratum sizes is a quick method to achieve this end if the precision requirements are fairly vague.

There are also some other considerations, when determining an allocation scheme.

Minimum sample sizes. In order to estimate a stratum mean (or total) there must be at least one unit in the sample and in order to estimate a stratum variance there needs to be at least two units. Also, for small businesses the non-response (and over coverage) problem is often great. For minimising the risk of obtaining empty strata or strata with only one sampling unit, a minimum size for the initial sample needs to be set. It is often advisable not to have smaller initial samples than five. (This recommendation also puts a limit on the number of strata, since too many strata with a minimum size of five units may lead to overallocating small units.)

Outliers. Due to imperfections in the sampling frame (such as errors in the size measures) but also to dynamic developments in the population, it is fairly typical that some "small" sampling units will be found to have large values for the target variable. This leads to poor precision in the estimates and is a reason to be cautious in overallocating large units when outliers can be expected to occur. (Of course, the best practice would be to try to foresee the potential outliers through expert knowledge and move them to the take-all strata in advance but this may prove to be difficult.)

The many-variable problem

There is no universally agreed best method for optimum allocation when estimates for several variables are required. At the same time, the practical problem does not need to be enormous. The following approaches could be relevant.

- *The target variables are correlated.* If the correlation between them is as strong as between each of them and the auxiliary variable, then it may not even be possible to distinguish between the best allocation for each one of them and a simple practical approach according to the previous discussion could be chosen.

- *One target variable is clearly the most important one.* Then, allocating according to that variable and accepting the resulting accuracy for the others would be appropriate. If this leads to an unacceptable result for another variable, maybe one could save a small part of the allowed sample size for augmenting the sample according to the needs for this other variable.
- *Define a weight function for the variables and allocate optimally according to the weighted variance*. Sigman and Monsour show how this could be done, provided that we know the stratum variances for each of the target variables. Another problem is that it is not easy to specify meaningful weights.
- *Specify an upper bound for the estimator variance of each of the target variables and solve the resulting optimisation problem.* Again, Sigman and Monsour provide the details for this method based on known stratum variances for all the variables.

However, the problem of not knowing the stratum variance exactly for any of the target variables results in approximation issues for the last two methods, which are not easily handled.

The many-domain problem

From a theoretical point of view, one could treat the many-domain problem as a special case of the many-variable problem. This is because a survey value of interest for a specific domain is obtained by setting the value of the associated variable to zero in all other domains. However the variables defined this way would be negatively correlated so that the approaches mentioned above are not really relevant.

From a more practical point of view we could reason as follows. Normally there is a defined set of domains for which estimates with a certain minimum level of precision is required. Users are usually not able to state the precision requirements explicitly and in this situation the process of determining them would have to be based on an analysis of the user needs, consultation with users, and negotiations about the available budget. In any case, the end result of this process should be agreed precision requirements by domains. Domains in economic statistics are usually based on *industry* according to some classification (NACE in Europe) and sometimes also, especially for big countries, on a geographical subdivision (*state, region etc.).*

EU perspective: comparability problems

In the EU perspective an important issue is whether and to which extent sampling designs, estimators, allocations etc. need to be harmonised over EU Member States in order to obtain comparable results. This section is intended as a brief introduction to this important problem area.

Sampling frames. High quality business registers to use for sampling frames serve to minimise the coverage errors of surveys. Coverage errors could lead to large biases and are thus a potentially important factor both for the quality and the comparability between business surveys. Important quality aspects of registers are a correct coding of the activities (NACE codes) of the businesses and a precise measure of their size A system also needs to be in place for timely updating of these kinds of data.

Frame updating and the survey reference period. The frequency and timeliness of updating a business register will strongly influence the coverage properties of the samples drawn from the register. Obviously, both over- and undercoverage increase as the lag between births and deaths of businesses and their subsequent entering into the register increases. The time lapse from the last occasion of updating the register until the reference period of the survey adds another sample is drawn and adds to the lag and the potential coverage errors of the survey.

For example, assume that a business register is updated each June with data from the previous year. In October the same year, a sample is drawn and used for asking about information about the next year. The effective time lag from the period that the frame information represents and the period for which this information will be used would then be 2 years.

Parameters. By the term parameter, we are referring to the objective function to be estimated. This includes the target population, the precise definition of the variables involved and the functional form combining these variables (sometimes, but not always a trivial aspect). For statistical domains, where EU comparability is particularly important, these matters will typically be subject to regulations limiting, but not eliminating, the scope for national differences. There are many subtle ways that differences between countries could arise. For economic variables, there could be different accounting rules or practices.

Sampling designs and estimators. Where the parameters are well defined, the potential for non-comparability due to different sampling and allocation methods or different estimators is much smaller. This is true at least where approximately unbiased probability sampling methods are used and estimator variances (coefficients of variation) are not large. This is because the expected value of the estimator will be essentially the same under such sampling designs.

Where there are non-probability elements of the sampling design, the potential for non-comparability is much greater. Especially the setting of cut-off thresholds at different levels and different estimation of the portion of the population below the threshold could result in biased and non-comparable results.

Questionnaire and other measurement effects. Different measurement practices as well as different ways that respondents interpret various questions could well lead to significant non-comparability effects.

Non-response and its treatment. Different rates of non-response are obvious sources of bias and therefore also of non-comparability. But also within similar rates of non-response, there may be potential for non-comparability resulting from differences in biases. It would be necessary to study the various subgroups contributing to non-response, both with respect to their general characteristics and their causes for not responding (refusal, no contact etc.) in order to understand the risks for bias and its likely direction.

Treatment of outliers. Alternative methods of handling outliers could lead to vastly different results. What occasionally happens is that in a stratum with a small sampling fraction a large value of a variable occurs. The traditional Horvitz-Thompson estimator calls for expanding this already large value by the inverse of the sampling fraction, which often appears intuitively unreasonable. A judgement then needs to be made by the statistician on what to do. The two simplest decisions that are often made are:

1) to remain with the estimator as decided and inflate the value accordingly or
2) to “move” the unit to a take-all stratum, where it “should have been” and not inflate the value at all.

Clearly, treatment according to 1 versus 2 could lead to vastly different results and are not comparable at all. Various estimators (e.g. Winsorisation) are proposed in the literature, which lead to results in between these two extremes. Our purpose here is only to emphasize the need for harmonisation with respect to outlier treatment.

■

The bottom line of this review of possible sources of non-comparability is that the only one that is found "not guilty" are different sampling strategies (design, allocation and estimation) as long as they belong within the family of approximately unbiased designs. The other problems mentioned are all likely to lead to potentially large comparability problems and it is a matter of judgement in each survey how big the problems are likely to be.

Further reading

On survey sampling in general.

Cochran, W.G. (1977): Sampling Techniques. Wiley. This book is still very useful for learning about the basics of survey sampling. It includes two long chapters on stratified sampling and the optimum allocation problems occurring for this design. It also gives variance formulas for the simple forms of ratio and regression estimators.

Särndal, C-E., Swensson, B. and Wretman, J. (1992): Model Assisted Survey Sampling. Springer. This book covers the modern theory of survey sampling. It is more mathematically demanding than Cochran. It defines and explains the general regression estimator and has a special chapter on optimal sampling designs. Several chapters deal with non-sampling error problems like sampling frames, non-response and measurement errors.

On πps sampling

Brewer, K.R.W and Hanif, M., (1983): Sampling with unequal probabilities. Springer. This is a classical book on πps sampling describing some 50 different methods in this area.

Särndal, Swensson and Wretman (1992, above) also devotes several sections to πps methods.

Rosén, B. (1997a), Asymptotic theory for order sampling, J Stat Plan Inf., 62 135-158 and

Rosén, B. (1997b), On sampling with probability proportional to size, J Stat Plan Inf., 62 159-191. These are two papers which provide the theory behind the novel πps technique called order sampling, described above. There are also a number of less accessible papers that deal with more applied issues concerning this method.

On business surveys in general

Cox, B.G. et al (1995): Business Survey Methods, Wiley. This is a collection of conference papers on a large number of topics in business surveys.

On sample allocation.

Both *Cochran* (1977, above) and *Särndal, Swensson and Wretman* (1992, above) include discussions about the allocation problem. Cochran's treatment is more detailed with respect to the traditional situation in stratified sampling with the HT estimator, whereas Särndal et al takes a broader view including model-based considerations.

Sigman; Richard S. and Monsour, Nash J.: Selecting samples from list frames of businesses. In Cox et al (1995, above). Chapter 8.2 and 8.3 discuss model-based approaches to the allocation problem and the situation with many variables or many subdomains of estimation. Their paper also includes a long reference list with more reading on this topic.

On sample rotation

Several chapters in *Cox et al* (1995, above) deal with the sample rotation problem. Chapter 9 by Esbjörn Ohlsson gives a detailed presentation of the prn technique presented in this note. Chapter 10 by Srinath and Carpenter present a number of other methods called the rotation group method, repeated collocated sampling and modified collocated sampling. Chapter 8.5 by Sigman and Monsour (above) also discuss rotation problems. The reference lists in all these three chapters provide further reading about this issue.