Background Facts
on
Economic Statistics

2003:3

SAMU
The system for co-ordination of frame populations and samples from the Business Register at Statistics Sweden

Department of Economic Statistics

The series Background facts presents background material for statistics produced by the Department of Economic Statistics at Statistics Sweden. Product descriptions, methodology reports and various statistics compilations are exampels of background material that give an overview and facilitate the use of statistics.

Publications in the series
Background facts on Economic Statistics

2001:1 Offentlig och privat verksamhet – statistik om anordnare av välfärdstjänster 1995, 1997 och 1999

2002:1 Forskar kvinnor mer än män? Resultat från en arbetstidsundersökning riktad till forskande och undervisande personal vid universitet och högskolor år 2000

2002:2 Forskning och utveckling (FoU) i företag med färre än 50 anställda år 2000

2002:3 Företagsenheten i den ekonomiska statistiken

2002:4 Statistik om privatiseringen av välfärdstjänster 1995–2001. En sammanställning från SCB:s statistikkällor

2003:1 Effekter av minskad detaljeringsgrad i varunomenklaturen i Intrastat – från KN8 till KN6

2003:2 Consequences of reduced grade in detail in the nomenclature in Intrastat – from CN8 to CN6

These publications and others can be ordered from:
Statistics Sweden, Publication Services, SE 701 89 ÖREBRO, Sweden
phone +46 19 17 68 00 or fax +46 19 17 64 44.

You can also purchase our publications at our **Statistics Shop**:
Karlavägen 100, Stockholm, Sweden

2003:3

SAMU
The system for co-ordination of frame populations and samples from the Business Register at Statistics Sweden

Statistics Sweden
2003

Producer Statistics Sweden
Department of Economic Statistics, the Methodology Function

Inquiries Annika Lindblom, phone +46 19 17 60 86
e-mail: annika.lindblom@scb.se

ISSN 1650-9447
Printed in Sweden
SCB-Tryck, Örebro 2003.04 MILJÖMÄRKT Trycksak 341590

Table of contents

Summary

The majority of the business surveys at Statistics Sweden use the SAMU to establish a frame population and to draw co-ordinated samples. The main objectives of the SAMU are to obtain comparable statistics, high precision in estimates of change over time (in terms of a relative standard error) and to spread the response burden among the businesses.

The SAMU has recently been implemented in a new user-friendly environment with additional facilities. A main reason for developing the new SAMU was to adapt to Statistics Sweden's new Business Register. The new SAMU can be used to establish a frame population and to draw a co-ordinated sample for a particular survey at any time during the year. All surveys use the same version of the Business Register, the one at the moment available in the SAMU, in order to produce comparable statistics. A new version of the Business Register is implemented into the SAMU at four different occasions every year.

The new SAMU includes, besides the co-ordination, technological support for delimitation, stratification and for allocation of a business survey. This report includes descriptions of methodology and of practice in the new SAMU.

1 Introduction

The majority of the business surveys at Statistics Sweden use the so-called SAMU[1] to obtain comparable statistics by using co-ordinated frame populations and to draw co-ordinated samples. The technique, which is used by the SAMU to draw co-ordinated samples, was developed at Statistics Sweden in the late 1960's and the SAMU was used for the first time in 1972.

The SAMU offers a clever method to generate, from an updated Business Register (BR), simple random samples that are positively co-ordinated (large overlap) between consecutive occasions for repeated surveys. This quality promotes repeated surveys to draw completely new samples more often. Before the SAMU was introduced repeated surveys often used quite old samples in combination with complementary samples among newly started businesses. Maintaining a sample for a long time will always mean difficulties in terms of estimation and less efficient survey design. The SAMU also offers a method to spread the response burden among businesses as well as a way to obtain comparable statistics by the co-ordination of frame populations. A very important feature at the introduction of the SAMU was the positive co-ordination over time (large overlap). During the years, though, the importance of comparable statistics and of spreading the response burden has increased which means that now the three features are equally important.

The SAMU has recently been implemented in a new user-friendly environment with additional facilities. A main reason for developing the new SAMU was to adapt to Statistics Sweden's new BR, especially to the new types of units and the new environment. There was also a need for improvements of the facilities in the SAMU. This report includes a description of the new SAMU: first a methodological part and then a practical part, which describes how the SAMU works in practice. For more detailed information on the methodology see Ohlsson (1992).

[1] SAMordnade Urval in Swedish, co-ordinated samples in English

2 The Business Register at Statistics Sweden

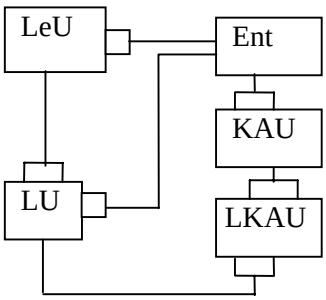
2.1 Different types of units in the Business Register

The frame population used in the SAMU is based on the Business Register (BR) at Statistics Sweden. During the 1990s Statistics Sweden developed a new BR called "Företagsdata-basen" and it was implemented in a new environment (PC-system).

A major difference between the new BR and the old one is that the new BR includes additional unit types.

The following types of units are included in the new BR:
- Legal Unit (LeU)
- Enterprise Unit (Ent)
- Kind of Activity Unit (KAU)
- Local Unit (LU)
- Local Kind of Activity Unit (LKAU)
- Firm
- AST Unit

The Firm and the AST unit are not important for the production of statistics and will therefore not be discussed further in this report.
The units in the new BR, which are of interest for the SAMU:



In November 2002 the BR contained about 1.6 million legal units and about 842 000 of them were considered as active. In the maintenance of the BR there are certain rules, which are applied in order to decide if a legal unit is active or not. Every active legal unit (in contrast to an inactive legal unit) is connected to an enterprise (KAU, LKAU and LU) in the BR. The number of multiple location enterprises were about 8 100.

Table 1 shows the number of enterprises included in the BR in November 2002. The majority of the enterprises in the BR are small (98 percent) in terms of number of employees and annual turnover. A very small number of the enterprises (0.2 percent) are considered as large according to these size measures.

Table 1
Number of enterprises in the BR in November 2002

Size measure: Number of employees	Number of enterprises	Proportion of enterprises, (%)	Proportion of employees, (%)	Proportion of turnover, (%)
Small: < 20				
without employees	627 745	74.6	0	7
with employees	197 117	23.4	20	20
Medium: 20-199	15 289	1.8	20	25
Large: > 199	1 746	0.2	60	48
Total	841 897	100	100	100

The grouping of enterprises into the categories small, medium and large is only for the purpose of this table and it is not official.

2.2 Sources used in the maintenance of the Business Register

The BR is updated weekly from information on registration and deregistration of legal units collected mainly from the National Tax Board. The registration of a legal unit includes information on legal entity, economic activity, number of employees, status and contact variables (like name, address and telephone number). This information is not always completed and, if it is, the quality can be quite uneven. In order to improve the quality of this information the staff at the BR contacts every newly registered legal unit with at least ten employees.

Information on a newly registered active legal unit is used to create the other types of units in the BR linked to the legal unit. Other sources like the register of Enterprise Groups, Profit and Loss Accounts, Balance Sheets, inquires and surveys conducted by Statistics Sweden are used to profile enterprises, to delineate one legal unit into more than one local unit or to delineate an enterprise into more than one kind of activity unit. Direct contact with specific legal units is of course of great importance when profiling and delineating.

The information used to update variable values for all types of units in the BR is mainly obtained from the National Tax Board and from surveys and inquires conducted by Statistics Sweden. The BR is updated weekly from information concerning contact and status variables. But the main part of the information used to update variables like economic activity, number of employees and annual turnover is available to the BR only at few occasions during the year. Therefore the quality of the BR is especially high in close connection to these updates.

For more detailed information on the BR contact the staff at the BR or see www.scb.se/foretagsregistret/innehall/fdballmant.asp (only in Swedish).

2.3 Types of units used in the SAMU

In the SAMU it is possible to establish frame populations based on enterprise units, kind of activity units and local units. These unit types, which can serve as frame population units in the SAMU, can also be used as sampling units, either for censuses or for sample surveys.

The enterprise unit, the kind of activity unit and the local unit are referred to as statistical units in the BR and they are, except for the local unit, mainly created for producing the economic statistics. The local unit did exist in the old BR and is referred to as both a statistical unit and as an administrative unit. The local kind of activity unit is also referred to as a statistical unit but at development time of the new SAMU there were no requests for using this unit as a frame population unit (the frame population unit does not always have to coincide with the observation unit).

Recently, though, it has become clear that there are needs for introducing the legal unit as a frame population unit in the SAMU. This implementation will be finalised during the first half of 2003.

3 Frame population and sample co-ordination

The use of the same version of the BR and the co-ordinated sampling for business surveys has three main objectives:

- to obtain comparable statistics by promoting the use of the same version of the BR for many surveys
- to obtain positive co-ordination of samples over time for the same survey
- to obtain samples that are co-ordinated - negatively or positively - among different surveys

The use of the same version of the BR for many surveys (co-ordination of frame populations) makes it possible to compile comparable statistics, which is important for e.g. the National Accounts. Surveys used by the National Accounts should use similar definitions of frame population units and compatible variables in the survey design. There are several business surveys using the SAMU only to establish a frame population and these surveys use the SAMU only for comparability reasons.

To increase the precision in estimates of change over time, the co-ordination design ensures that consecutive samples for the same survey are overlapping, although each sample is drawn from an up-to-date register. Negative co-ordination between surveys is used to spread the response burden among the businesses. Positive co-ordination between surveys is used to facilitate comparisons of variable values in various surveys.

Co-ordination between surveys and over time is obtained by using the so-called JALES[2] technique, which is based on the use of random numbers permanently associated with the frame population units. The method is used by Statistics Sweden and several other countries. The JALES technique is described further down in this report and more detailed information is available in Ohlsson (1992).

3.1 The SAMU-versions of the Business Register

The SAMU can be used to establish a frame population for a particular survey and to draw a co-ordinated sample at any time during the year. In order to produce comparable statistics all surveys in the SAMU use the same version of the BR, the version at the moment available in the SAMU. A new version of the BR is implemented into the SAMU at four different occasions every year; in March, in May, in August and in November. These occasions are selected to fit the needs from the different surveys as well as the quality of the BR. As mentioned before, in section 2.2, the quality of the BR is especially high in close connection to updates. The SAMU-versions of the BR are derived at these time points.

3.2 Sequential simple random sampling

The SAMU uses a technique called sequential simple random sampling to draw a simple random sample without replacement of size n from a frame population of size N: With each unit in the register there is a random number associated, taken from a set of random numbers uniformly distributed over the interval (0,1). The frame population is then ordered in ascending random number sequence and the n first units on the list are included in the sample. It can be shown (see Ohlsson (1992)) that sequential simple random sampling without replacement is equivalent to simple random sampling without replacement.

[2] This technique was developed at Statistics Sweden in the late 1960' s by Johan Atmer and Lars-Erik Sjöberg from whom 'JALES' is an acronym

This is a very clever method for drawing a simple random sample without replacement. For example, after the sample is drawn, it is easy to adjust the sample size by including more units from the list or by excluding the last units (according to the size of the random numbers) included in the sample. This can, for example, be used to increase the sample size before the questionnaires are sent out or while the survey is running.

3.3 The JALES technique

The basic idea in the JALES technique is to associate a *permanent, independent and unique* random number, uniformly distributed over the interval (0,1), with every unit in the register. For every unit persisting in the register the same random number is used on each sample occasion. Every new unit is assigned a new random number while closed-down units are withdrawn from the register with their random number. On each sample occasion a new sequential simple random sample is drawn, using the permanent random numbers. In this way we always get a simple random sample from the up-dated register. However, a large overlap with the latest sample can be expected since persistants have the same random numbers on both occasions. This enables better precision in estimates of change over time.

By the symmetry of the uniform distribution, we could just as well take the last n units to obtain the simple random sample. We can, in fact, select the first n units to the left, or to the right, of any fixed point in the interval (0,1). If there are not enough units to the right (or left) of our starting point a, we simply continue the selection to the right (or left) of the point 0 (or 1).

3.4 Positive and negative sample co-ordination

In order to co-ordinate two samples with desired sample sizes n_1 and n_2, choose two arbitrary constants a_1 and a_2 in the interval (0,1). The units with the n_1 random numbers closest from a_1 in one direction (left or right) are included in the first sample. The second sample includes the ones with the n_2 random numbers closest from a_2, in the same or the opposite direction as the first sample.

The maximal positive co-ordination of two surveys is obtained by using the same starting point and direction for both. Negative co-ordination of two surveys can be achieved by choosing different starting points (well apart) and using the same direction. An alternative way is to choose the same starting point (or two close ones) but different directions. There are not always enough units to obtain complete negative co-ordination, but this technique at least reduces the number of units that the surveys have in common.

Negative co-ordination is a very effective tool to spread the response burden among small businesses. It is important to spread the response burden among these businesses because they often do not have the capacity to participate in many surveys. Statistics Sweden also uses additional methods like simplified questionnaires, split questionnaires and administrative data in order to reduce the response burden for small businesses.

There is little room for spreading the response burden among medium sized businesses because these businesses are few, see table 1. They have a proportionately large impact on the estimates in terms of turnover and number of employees and they are therefore often included in samples. Medium sized businesses could meet a heavy burden, especially in industries with few businesses. Thus, for medium sized businesses must additional methods like considering the response burden in the survey design and further use of administrative data be used in order to reduce the response burden.

The small number of large businesses, see table 1, are of great importance for the economic statistics because they have a large impact on the estimates in terms of turnover and number of employees. Therefore it is crucial, with few exceptions, to include all large sized busi-

nesses belonging to the frame population for a specific survey. Otherwise it would be more or less impossible to publish the survey results. At Statistics Sweden the work on response burden regarding large sized businesses is mainly focused on simplifying for the respondents to supply the requested information.

3.5 Co-ordination when surveys are stratified

Stratification of a finite population U means a partitioning of U into H subpopulations, called strata and denoted U_1, ….., U_h, …, U_H. Stratified sampling means that a probability sample s_h is drawn from U_h according to a design $p_h(\cdot)$ and that the selection in one stratum is independent of the selection in all other strata.

Stratified sampling is often used:

- to improve the accuracy in the estimates by dividing the frame population into homogenous sub-populations (strata)
- if estimates of specified precision are wanted for specific subsets of the population (domains). Each domain can be treated as a level in the stratification if domain membership is specified in the frame population.

Economic activities often constitute important domains in businesses surveys. This means that a common stratification in business surveys is a combination of economic activity (industrial strata) and size. Each industrial stratum is divided into homogenous size groups according to a size measure based on, for example, number of employees or annual turnover.

In the SAMU a sequential simple random sample is drawn in each stratum. For a particular survey the same direction and starting point is used in all strata. Suppose that two surveys use different stratification on the same frame population. If the starting points are different then the surveys will be negatively co-ordinated, because a random number which is small (large) in one stratum is also small (large) in another stratum.

3.6 Co-ordination when surveys are based on different unit types

In the SAMU negative or positive co-ordination between surveys based on different types of sampling units is obtained because the units are co-ordinated through their random numbers. In the SAMU the random numbers are assigned to new local kind of activity units, the smallest building brick in the BR. By new is meant that the unit did not exist in the latest SAMU-version of the BR. A single-location enterprise with only one kind of activity unit is given the same random number as its local kind of activity unit. In other words; the local unit, the kind of activity unit and the enterprise are all given the same random number as the local kind of activity unit.

The assignment of random numbers to a multiple-location, or multiple-activity, enterprise is more complicated. In this case the enterprise, kind of activity unit, local unit is given the random number of one of its lower level linked units (the main one according to certain rules). For the majority of the enterprises, the single-location and single-activity enterprises, the co-ordination between the unit types is simple and straightforward. For the multiple-location (and multiple-activity) enterprises the co-ordination is less efficient, because it is only possibly to co-ordinate a multiple-location (or multiple-activity) enterprise with *one* of its lower level linked units. For more information on the assigning of random numbers in the SAMU see Lindblom (2001) (available only in Swedish).

Negative co-ordination between surveys based on different types of units is important for small enterprises, because in respect to them it is vital to spread the response burden. The majority of the small enterprises are single-location and single-activity enterprises and for them the co-ordination between surveys based on different types of units works well. For the multiple-location or multiple-activity enterprises the co-ordination is less efficient, but on the

other hand the majority of these enterprises are large. Large enterprises are almost always included in samples so there are limited opportunities for spreading the response burden among them.

Positive co-ordination between surveys based on different types of units is of course also possible in order to facilitate comparisons between variable values.

Technical note: The random numbers are assigned to the units within the SAMU, which means that there are no random numbers in the BR. The SAMU random numbers should only be used by the surveys in the SAMU so therefore this was the most convenient solution. The assigning of random numbers to new units is done each time a new version of the BR is implemented into the SAMU.

4 Rotation

Due to the positive co-ordination over time, a selected unit may have to participate in a survey for many years. On the other hand, a unit (randomly) not included in the sample has a large probability of not having to participate for many years. As this could be considered as unfair the SAMU has a system for rotation, whereby the selected units rotate out of the sample after a certain number of years. The same method for rotation is applied to all type of units and the rotation is performed once a year (the SAMU-version of the BR derived in November).

The number of years a unit should participate in a survey is a balance between response burden and the decrease in the precision of the estimates of change over time that is acceptable.

Rotation works only if there is room for rotation, i.e. if it is possible for a unit to rotate out of a sample after the certain number of years without immediately rotating into the sample of another survey. Such room is only available among small units where the inclusion probability is small. Units in stratum with larger inclusion probability are rotated as well but it takes longer before the units are rotated out of the sample. How long depends on the size of the inclusion probability. *The main purpose of the rotation in the SAMU is to spread the response burden among small businesses.*

4.1 Methods for rotation

It has been decided that the objective should be to keep the selected units in the sample for five years. By exploring frame populations and inclusion probabilities for different surveys it has become clear that in strata with inclusion probability less than 0.10 there will be room for rotation.

There are, at least, two possible methods for rotation:

- Shift the starting points
- Shift the random numbers

If a unit with inclusion probability less than 0.10 is to be rotated out of a sample after five years then the starting point must be shifted by 0.02 a year. The same is obtained by shifting the random number with 0.02 a year. After five years the starting point has shifted 0.1, or the unit has moved 0.1 on the random number line, and should be rotated out of the sample. In the SAMU the random numbers are shifted and the reason for this is the use of the random rotation groups, see section 4.2.

The disadvantage of shifting the random numbers (or starting points) each year with the same length is that the rate of the rotation will vary considerably among strata. For example, in strata with a very small inclusion probability, the majority of the units in the sample will be renewed after a year. This is a disadvantage for the estimates of change over time. To achieve same rotation rate in each stratum there must be individual shifts for each stratum. But to use individual shifts will, in the long run, destroy the positive and negative co-ordination between the surveys. The co-ordination between surveys is maintained only if the same length of the shift is used for all surveys.

4.2 Random rotation groups

In the SAMU the problem with varying rotation rates is solved by grouping all units into five rotation groups. The random numbers (or starting points) for the units are then shifted by 0.10 only in one rotation group each year. That is, all units in rotation group one will shift random number (or starting point) the first year. The second year all units in rotation group two will shift and so on. This method makes the rotation rate in all strata with inclusion

probability less than 0.10 on the average 1/5. This method also ensures the negative and positive co-ordination between surveys.

The last digit in the random number determines which rotation group a unit should be assigned to. This means that the units are randomly assigned to a rotation group. The co-ordination between surveys based on different types of units is also maintained in this way. A single-location or single-activity enterprise and its lower level linked units will be assigned to the same rotation group, because they have the same random number.

A new unit (by new is, in this context, meant that the unit did not exist the latest time the rotation was performed) is not allowed to rotate. It is not allowed even if the unit is assigned to the particular rotation group, which is to rotate the actual year. The reason is the objective to keep a unit in a sample for five years. The next year, though, this unit follows the rest of the units in the same rotation group.

To use the random rotation group method together with the shift of starting points would yield two starting points in each stratum: one initial and one where 1/5 of the units move each year. Two starting points for each stratum are not easy to monitor in a complex system. It is easier to shift the random numbers and therefore this method is used in the SAMU.

5 Co-ordinating in practice

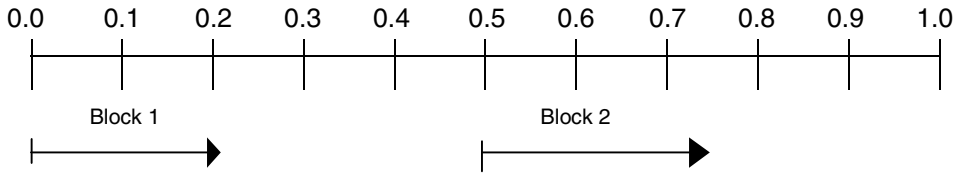
5.1 Blocks

In the SAMU, the same starting point and direction is used for several surveys in order to co-ordinate. Surveys with the same starting point and direction are said to belong to the same block. Consequently, a survey cannot be negatively co-ordinated with *every* other survey. Surveys in the same block are always positively co-ordinated. If that is not desirable, then surveys covering different industries or different size groups should be put together in the same block. Then negative co-ordination is obtained automatically.

5.2 Different alternatives when co-ordinating surveys

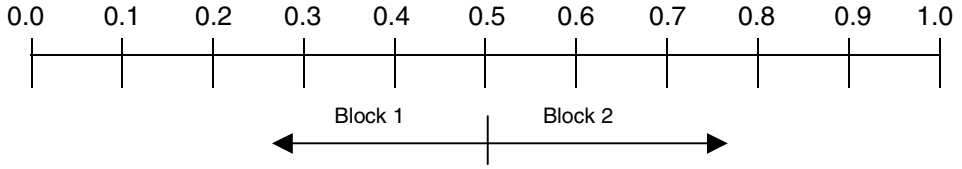
In order to co-ordinate two surveys negatively, they can have separate starting points (well apart) and the sampling direction should be the same, see figure 1.

Figure 1



Another alternative is to give the two surveys the same starting point but opposite sampling directions, see figure 2. The second alternative will give maximum negative co-ordination. But there is a risk that it would not result in an even distribution of the response burden over the entire random number line. Businesses with random numbers well apart from the only starting point will perhaps never be included in any sample.

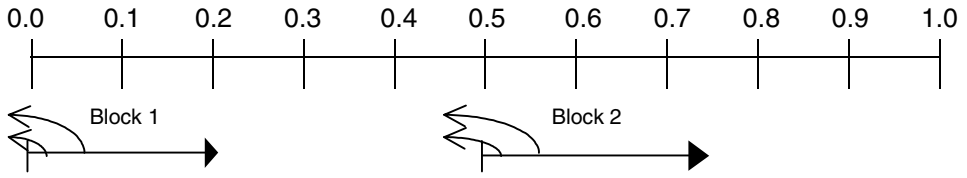
Figure 2



5.3 Sampling direction in combination with rotation

In terms of rotation, it is preferable to have the same sampling direction for all blocks. Units in a block with random numbers nearest to the starting point are included in many samples, probably in the majority of the surveys in the block. That is why the rotation direction should be opposite to the sampling direction in the block, see figure 3.

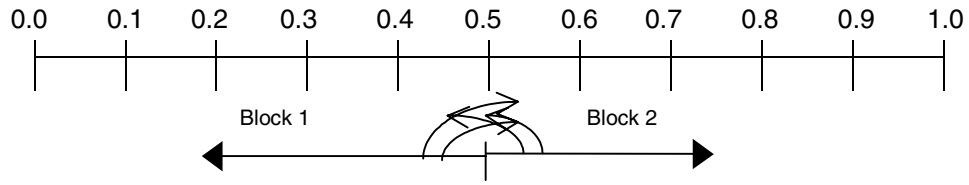
Figure 3



In the case of the same starting point but opposite sampling directions for two blocks there is a problem when rotating the units. If the rotation direction is opposite to the sampling direction, then the units near the starting point in one block will end up near the starting

point in the other block, see figure 4 (“out of the ashes into the fire”). It is possible to solve this problem but it is quite complicated when there are many surveys and several blocks.

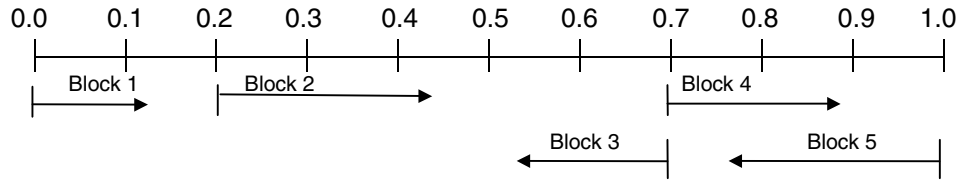
Figure 4



5.4 The co-ordination in the SAMU

There are five blocks in the SAMU, see figure 5. The figure should not be seen as an ideal placing of blocks but rather as the result of 30 years of additions and adjustments. There are two pairs of blocks (block 1, 5 and block 3, 4) with the same starting point but opposite sampling direction. Because of these blocks the rotation scheme in the SAMU does not shift the random numbers with 0.1 for all units in a rotation group. The length of the shifts in the SAMU is adjusted to avoid units near the starting point in one block to end up near the starting point in the other block.

Figure 5



For information on to which block the different surveys in the SAMU are assigned see appendix 1.

The JALES-technique ensures a simple random sample, but does not guarantee a specific unit to be included only in a certain number of surveys. It is difficult to make this kind of guarantees without renouncing the randomness. The same applies to the method for rotation described above which means that it is impossible to guarantee a unit to be rotated out of the sample after a certain number of years. Statistics Sweden focuses on methods like the ones mentioned in section 3.4 in order to reduce the response burden among businesses where spreading and rotation are not completely successful.

5.5 Feed back into the Business Register from sample surveys

It is not advisable to use a lot of feed back from sample surveys into the BR, especially if the technique with permanent random numbers is used. This applies mainly to variables used in the survey design, variables like economic activity, number of employees and annual turn-over. Feed back from surveys would mean that businesses in the BR, which are included in samples, are updated while those not included are not. The large overlap between two consecutive samples means that one receives almost the same sample the next time a sample is drawn and feed back from the survey would mean that this sample is updated. In turn, the estimates will be biased, because the sample is no longer representative for the whole frame population. But from this point of view it is all right to update large businesses with survey feed back because they are, almost always, completely enumerated. In general, feed back from the surveys should rather be used as quality indicators in the maintenance of the BR.

6 The SAMU in practice

At Statistics Sweden there are a few persons working part time with the administration and improvement of the SAMU. Their major responsibilities are to establish the SAMU-versions of the BR, assign the permanent random numbers, to co-ordinate and to rotate the samples. Currently it is only possible to draw a simple random sample without replacement in the SAMU. But there is, at the moment, one survey in the SAMU that uses a πps sample in combination with the JALES-technique to obtain the co-ordination. In this case the actual sample drawing is performed outside the SAMU.

6.1 Suitable time points for the surveys to use the SAMU

The majority of the annual surveys referring to the year (t) draw a new sample in the SAMU in November year (t). The majority of the short-term surveys (quarterly and monthly) in the SAMU have until recently drawn new samples in November year (t-1) for surveys referring to the year (t). But now it has been recommended that the short-term statistics draw new samples twice a year; once in March year (t) and once in August year (t) for surveys referring to the year (t). This improves the accuracy, in terms of coverage, in the estimates produced by the short-term statistics. It also means that the possibility for comparisons between annual and short-term statistics referring to the same year (t) increases because the frame population for both the short-term and the annual statistics are established in the same year (t). This is important for, e.g., the National Accounts.

For surveys conducted only once the most recent SAMU-version of the BR is generally used.

6.2 General information in the SAMU

A new survey in the SAMU is assigned a unique survey number. General information like the name of the survey, contact person, frame population unit and to which block the survey is assigned is stored in the SAMU. There is also information on which allocation variable to use, desired precisions in the allocation and which size groups to completely enumerate (more information on allocation etc. further down). This information can of course be changed between the SAMU occasions.

6.3 Delimiting and stratifying the frame population

The SAMU includes technological support in order to delimit and to stratify the frame population for a specific survey. Currently the responsibility for the technological implementation of the delimitation and stratification is allocated to the persons working with the SAMU. But the objective is to let the users of the SAMU implement the delimitation and stratification. To be able to do this, though, some knowledge of computer programming is needed.

6.3.1 Delimitation

The SAMU-version of the BR includes all units in the selected frame population of a specific survey. One of the first matters regarding a survey is to decide what to include in the frame population. Common variables used by the surveys in the SAMU to delimit the frame population are institutional sector, ownership control, type of legal entity and economic activity. Several surveys in the SAMU use a cut-off limit in terms of, for example, number of employees or annual turnover. The specification of the cut-off limit is also included in the delimitation.

6.3.2 Stratification

The built in stratification variable in the SAMU is prepared for stratification on three levels. The majority of the surveys in the SAMU use economic activity to create the first level in the stratification (industrial strata) because economic activity often constitutes important domains. The industrial strata are then divided into homogenous sub-groups, size groups, according to a specified size measure. The most common variables used as size measure in the SAMU are number of employees, annual turnover and gross wages. For a specific survey it is possible to use different size measures in different industrial strata, to combine different size measures in one industrial stratum or to use the same size measure but different cut-off points for the size groups in the industrial strata.

The majority of the surveys in the SAMU only use two levels in the stratification, economic activity and size. But there are a few surveys using three levels and the third being, for example, region. In fact, it is possible to use more than three levels in the stratification in the SAMU. However, this would imply a slightly more complicated solution compared to the built in stratification.

6.4 Information not available in the Business Register

Several surveys in the SAMU use information not available in the BR to delimit the frame population and/or to stratify the frame population. This can be information collected from administrative registers, previous survey occasions or from other surveys. Information not included in the BR can also be used to earmark small (according to the regular size measure) units of large importance to the survey in order to, for example, completely enumerate them. This kind of information can easily be included in the SAMU and be available in the survey design.

6.5 Neyman allocation in the SAMU

In the SAMU, there is a built in module, for Neyman (optimum) allocation in order to give some guidance to the users when deciding sample size in each stratum for a particular survey. Neyman allocation is a common method to use in business surveys but it is very important that the variable used in the calculations is correlated with the study variable(s) (the variable(s) to be measured in the survey). Otherwise, it might be better to use another method for the allocation, for example proportional allocation, or to combine the Neyman allocation with other strategies. The Neyman allocation in the SAMU is optional and the users of the SAMU are free to work out the allocation according to other strategies.

The objectives of the allocation in the SAMU, for a particular survey, are to:

- estimate required sample size in each domain in order to obtain a pre-specified precision in terms of a relative standard error (under the condition that Neyman allocation will be used)
- distribute the required sample size in each domain over the size groups in order to minimize the variance in the domain

It is very important for a survey to have some knowledge about required sample size in each domain in order to obtain a desired precision but due to economic constraints there is, almost always, an upper limit on the total sample size for a survey. This means that the final total sample size for a survey often is a balance between economic constraints and accuracy in terms of relative standard errors. It is also important not to burden businesses through unnecessary large samples.

The Neyman allocation in the SAMU bases the calculations on a variable obtainable for the whole frame population. At the moment the user can chose between using number of employees, annual turnover and gross wages depending on which variable that is most correlated with the study variable (s). It is possible to include more allocation variables in the

SAMU and they would mainly be collected from different administrative registers. For technical details and formulas regarding the Neyman allocation in the SAMU see appendix 2.

6.5.1 The same information used for stratification and for allocation

The given sample sizes are quite dependent of the information used for the stratification and for the allocation. If the same information is used both for stratification and for allocation it should be observed that the given sample sizes normally are underestimated because the variation within strata is to small compared to if the calculations had been based on the study variable (one of the study variables).

6.5.2 Domains in the allocation

In the SAMU it is possible to specify domains consisting of arbitrary whole strata (domains cutting across strata cannot be specified). For one specified precision the Neyman allocation uses the same precision for all domains and each time you run the Neyman allocation, required sample sizes are calculated for ten pre-specified precisions. The Neyman allocation in the SAMU can easily be performed several times if the user would like to see given sample sizes for more than ten precisions.

6.5.3 Using the Neyman allocation in the SAMU

In business surveys large units (in terms of number of employees/annual turnover or important units in another sense) often have a large impact on the estimates and are therefore often decided to be completely enumerated *before* the actual allocation is performed. Size groups to be completely enumerated (decided in advance) must be specified among the general information, see section 6.2.

The Neyman allocation gives sample sizes based on minimized variance regarding the allocation variable in each domain. In practice, it is advisable not to trust the given sample sizes entirely because the allocation variable and the study variable (s) are not one hundred percent correlated. In business surveys it is also common to compensate for non-response in a stratum with the average based on the respondents (at least for small businesses). If too few units are selected in a stratum there is a risk that to few responds. To ensure a sufficiently large sample size in each stratum a minimal sample size can be specified before the allocation is performed. But the response burden should also be considered when specifying a minimal sample size.

The Neyman allocation can initially give sample sizes larger than the total number of units in a stratum. If this situation occurs this stratum (these strata) is (are) considered as completely enumerated and the Neyman allocation recalculates the required sample sizes in the remaining surveyed strata.

6.5.4 Output from the Neyman allocation module

The result of the Neyman allocation is a report available for the users of the SAMU. This report includes information on total number of units in the strata, given sample sizes from the allocation in each stratum (for each of the ten pre-specified precisions regarding the domains), the population total and the variance regarding the allocation variable. This report can be used as guidance for the users when deciding sample sizes in each stratum. There is also an additional report which is available for the SAMU users including more summarized information, like the total required sample size for the survey to achieve the pre-specified precision in each domain.

6.6 Sample drawing and available information

When the definitive sample size in each stratum has been worked out for a survey the SAMU includes a module into which the users can place the sample sizes. Thereafter the co-ordinated sample is drawn and the frame population and sample are stored in the SAMU-database. This information, and other information, from the SAMU-versions of the BR are available to the users. The users collect the information they need from the SAMU,

especially information linked to the survey design and sample occasion, like unit identification, stratum identification, inclusion probability, number of employees, annual turnover and so on for each unit in the sample (or for the whole frame population). Information like name, current status, address and telephone number are not stored in the SAMU-database because this information has to be up-to-date. Therefore it is collected directly from the up-to-date BR.

6.7 Use of previous SAMU-versions

When working with e.g. improvement of the design for a particular survey it is essential to have information available from previous SAMU-versions regarding this survey. Therefore information from every SAMU-version is saved. A particular survey can easily collect information on the frame populations and on the samples from previous SAMU occasions. The complete SAMU-versions of the BR are also saved so that frame populations can be established from previous SAMU-versions of the BR. The random numbers and the information not included in the BR regarding a SAMU-version of the BR are saved as well.

7 Information on response burden

An information system on response burden was developed some years ago based on the surveys included in the SAMU. The objective is to include all business surveys at Statistics Sweden in the information system but it is currently quite complicated to collect needed information from the surveys not included in the SAMU in an efficient way. Another problem is that the sampling unit does not always coincide with the observation unit. The SAMU users sometimes have to up-date units in the sample due to mergers, split-offs, break-ups and take-overs. These events can occur during the time elapsed between the sample occasion and the time of questionnaire send out. It is very important that the information on response burden is correct and up-to-date in this system. At the moment, there are discussions at Statistics Sweden on whether the information system should be based on the SAMU or not. An alternative is to base the system directly on information collected from all business surveys at Statistics Sweden.

7.1 Measuring the response burden

The response burden is always calculated regarding the enterprise unit level independently of what type of unit the survey is based on. An enterprise, with more than one of its lower level linked units included in the same survey, is considered once in terms of number of surveys but the responding time is accumulated.

The response burden in the system is measured in terms of number of surveys the enterprise is included in, perhaps the most straightforward way. The system is also prepared for measuring the burden in terms of average time the enterprise spend filling in questionnaires. This is not quite finalised, mainly due to lack of information on average amount of time enterprises spend filling in the questionnaires.

7.2 Information in the system

The system includes information on:

- Response burden for one specific enterprise
- Response burden among enterprises in different kind of economic activities and size groups

Information on one particular enterprise is mainly used in contacts with this enterprise. Information on groups of enterprises can be used in order to expose categories of enterprises with a heavy burden. This information can in turn be used as an indicator on the need for considering the response burden in the survey design.

Table 4
Enterprises burdened from surveys in the SAMU during the year 2001

	Number of enterprises included in samples distributed by size class (in terms of number of employees)								
Number of surveys	0	1-4	5-9	10-19	20-49	50-99	100-199	200-	Total
1	1 809	12 316	5 684	5 384	2 860	361	39	15	28 468
2	584	322	955	3 010	2 396	546	138	8	7 959
3	16	20	98	764	1 180	675	262	73	3 088
4	0	0	6	233	798	480	199	55	1 771
5	0	0	0	80	429	338	176	135	1 158
6	0	0	1	28	257	259	175	107	827
7	0	0	0	7	87	238	115	112	559
8	0	0	0	1	43	135	139	150	468
9	0	0	0	0	14	82	100	161	357
10	0	0	0	0	2	47	88	75	212
11	0	0	0	0	2	5	42	47	96
12	0	0	0	0	0	3	23	84	110
13	0	0	0	0	0	2	8	344	354
Total number of enterprises included in samples	2 409	12 658	6 744	9 507	8 068	3 171	1 504	1 366	45 427
Total number of enterprises in November 2001[3]	*613 144*	*147 755*	*33 415*	*17 763*	*10 381*	*3 150*	*1 463*	*1 716*	*828 787*

Table 4 shows enterprises burdened from surveys in the SAMU where information from the SAMU occasions Mars 2001, May 2001, August 2001 and November 2001 is used. The table shows for example that there are 1 809 enterprises without employees included in one survey, 584 enterprises without employees included in two surveys, 12 316 enterprises with between one and four employees included in one survey and so on.

The figures in table 4 clearly indicate the desired pattern. Small enterprises (in terms of number of employees) are included in few surveys and the small number of small enterprises included in two (or even three) surveys is probably due to positive co-ordination.

The information in table 4 is of course very dependent on surveys that use the SAMU during one particular year. Therefore the same information regarding the year 2000 (the SAMU occasions Mars 2000, May 2000, August 2000 and November 2000) is enclosed in table 5. An important difference between the years is the decrease in number of enterprises without employees included in one survey. This decrease between the years is due to the fact that one short-term survey (including small enterprises in the sample) has followed the recommendation to draw a new sample in Mars year (t) instead of November year (t-1). This means that this survey used the SAMU in November the year 2000 but did not use the SAMU during the year 2001.

[3] The public sector is included in the SAMU-version of the BR. Enterprises in this sector are not included in many surveys despite the fact that they have many employees.

Table 5
Enterprises burdened from surveys in the SAMU during the year 2000

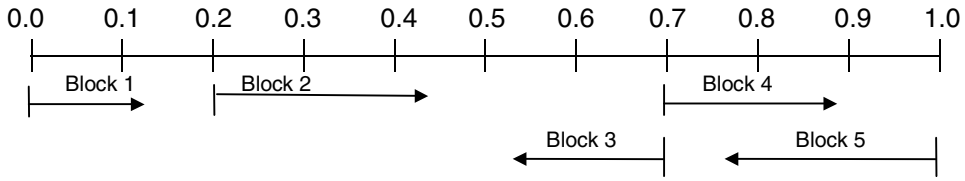
	Number of enterprises included in samples distributed by size class (in terms of number of employees)								
Number of surveys	0	1-4	5-9	10-19	20-49	50-99	100-199	200-	Total
1	4 560	14 765	5 057	6 560	2 392	253	38	76	33 701
2	130	548	1 505	2 565	2 882	483	97	24	8 234
3	11	7	45	565	1 165	719	206	83	2 801
4	0	0	1	78	581	591	205	82	1 538
5	0	0	0	8	121	463	252	110	954
6	0	0	0	1	14	235	256	134	640
7	0	0	0	0	0	127	196	215	538
8	0	0	0	0	0	57	90	303	450
9	0	0	0	0	0	6	25	184	215
10	0	0	0	0	0	0	7	117	124
11	0	0	0	0	0	0	0	21	21
12	0	0	0	0	0	0	0	5	5
13	0	0	0	0	0	0	0	2	2
14	0	0	0	0	0	0	0	1	1
Total number of enterprises included in samples	4 701	15 320	6 608	9 777	7 155	2 934	1 372	1 357	49 224
Total number of enterprises in November 2000[3]	*601 337*	*146 464*	*33 048*	*17 426*	*9 910*	*2 964*	*1 402*	*1 695*	*814 246*

References

Ohlsson, E. 1992 SAMU – The System for Co-ordination of Samples from the Business Register at Statistics Sweden. R&D Report, Statistics Sweden, 1992:18

Lindblom, A. 2001 Slumptal i SAMU. ES-metod nummer 43, Statistiska Centralbyrån, 2001:43

Appendix 1: Surveys in the SAMU november 2002



Block 1 (Right 0.0) includes	Cut Off point	Take all point
Structural Business Statistics, production sector, NACE 01-05, 10-45	- (< 50 empl. via adm. data)	50 employees
Investment inquiries, NACE 70.2	10 million in real–estate assesm.	200 million in real-estate assesm.
Investment inquiries, NACE 45, 50-52, 60, 62-65.1, 66, 70.1+3, 71-73, 74.202, 74.3	10 (20 some NACE) employees	200 employees
Investment inquiries, NACE 10-40	20 employees (5 in NACE 40)	200 employees
Annual Industrial Statistics: consumption of fuels and energy, NACE 10-37	10 employees	50 employees
Production Statistics, manufacturing NACE 10-37	20 (10 some NACE) employees	20 (10 some NACE) employees
Input Costs for Services and Goods	50 employees	50 employees
Consumer Price Index, NACE 50, 52, 55, 71, 80, 92, 93	-	-
Research and Development in the Business Sector, NACE 10-41, 50-95	50 employees	200 employees
Foreign Trade in Services, NACE 01-99	Different limits (turnover[4])	Different limits (turnover[4])
Block 2 (Right 0.2) includes		
Labour Cost Survey, NACE 10-74	10 employees	500 employees
Structure of Earnings Statistics for the Private Sector, NACE 01-74, 80-93	1 employee	500 employees
Use of Information and Communication Technology in Enterprises NACE 15-74	10 employees	200 employees
Waste and Returnable Raw Materials from the Industry, NACE 13-36	20 employees	200 employees
Block 3 (Left 0.7) includes		
Short Term Employment Statistics in the Private Sector, NACE 01-74, 80-93	1 employee	200 employees
Short Term Employment Statistics in the Public Sector, NACE 01-74, 80-93	1 employee	200 employees
Job Openings and Unmet Labour Demand, NACE 01-74, 80-93	1 employee	200 employees
Wages and Salaries in the Private Sector, NACE 10-99	5 employees	500 employees
Block 4 (Right 0.7) includes		
Structural Business Statistics, service sector NACE 50-95	- (< 50 employees via adm. data)	50 employees
Periodic Business Statistics, service sector	0 empl and turno[4]< 0.5 million	50 employees
Business Statistics on Transports, NACE 60.24, 63.11-12, 63.21, 63.4, 64.12	0 empl and turno[4]< 0.5 million	50 employees
Domestic Trade, turnover, service sector, NACE 50-55, 71-72, 74, 90, 92-93	0 employees and without wages	Turnover[4]>100/200/300 million
Domestic Trade, stocks, service sector, NACE 50-52, 64, 72, 73, 74, 92	Different limits (turnover[4])	Turnover[4]>100/200/300 million
Community Innovation Survey 3, NACE 51-74	10 employees	250 employees
Block 5 (Left 1.0) includes		
Industrial Short Term Indicators, NACE 10-37	10 employees	200 employees
Environmental Protection Expenditures, NACE 10-36, 40-41	20 employees	500 employees
Community Innovation Survey 3, NACE 10-41	10 employees	250 employees

[4] Annual turnover (SEK) collected from the National Tax Board

Appendix 2: Formulas used in the Neyman allocation

Stratification of a finite population $U = \{1,\ldots,k,\ldots N\}$ means a partitioning of U into H subpopulations, called strata and denoted $U_1,..U_h,..,U_H$ where

$U_h = \{$ k: k belongs to stratum U_h $\}$. By stratified sampling is meant that a probability sample s_h is drawn from U_h according to a design $p_h(\cdot)$ (h=1, .., H) and that the selection in one stratum is independent of the selection in all other strata.

The population total (t_{y_h}) in stratum U_h is given by:

$$t_{y_h} = \sum_{U_h} y_k$$

where y_k denotes the value of the allocation variable on unit number k in the population.

The stratum variance regarding the population total is given by:

$$S^2_{yU_h} = \frac{1}{N_h - 1}\sum_{U_h}\left(y_k - \bar{y}_{U_h}\right)^2$$

where y_k denotes the value of the allocation variable on unit number k in the population, N_h total number of units included in the population in stratum U_h and $\bar{y}_{U_h}$ means the population mean in stratum U_h.

Domains

Estimates for domains, i.e. subsets of the population, are often required. In the following we will only consider domains that are aggregates of strata. In other words, we will not consider domains that cut across strata. Let *D* be a domain, $D \in U$.

Suppose that *D* contains *d* strata. $D = \{U_{h_1(D)}, U_{h_2(D)},...,U_{h_d(D)}\}$. Let y_k be the value of the allocation variable for unit number k in the population. The population total $(t_{y(D)})$ in domain *D* is given by $t_{y(D)} = \sum_{k\in D} y_k = \sum_{U_h \in D}\sum_{k\in U_h} y_k$.

The relative standard error (coefficient of variation) regarding $t_{y(D)}$ is given by:

$$\frac{\sqrt{V(t_{Y(D)})}}{t_{y(D)}}$$

where $t_{y(D)}$ = the population total in domain *D*

$V(t_{y(D)})$ = the variance of t in domain *D*

If the relative standard error (coefficient of variation) regarding $t_{y(D)}$ is to be equal to or less than a selected precision α then the following expression should be fulfilled:

$$\frac{\sqrt{V(t_{Y(D)})}}{t_{y(D)}} \leq \alpha_D, \qquad \text{where } \alpha_D = \text{the selected precision in domain } D$$

If the Neyman allocation is used, then $V\left(t_{y(D)}\right)$ is given by:

$$V(t_{y(D)}) = \frac{N_D{}^2\left(\sum_{U_h \in D} W_h S_{yU_h}\right)^2}{n_D} - N_D \sum_{U_h \in D} W_h S_{yU_h}^2$$, where $S_{yU_h}^2$ = the stratum variance in stratum U_h

n_D = total sample size in domain D

N_D = total number in the population in domain D

$$W_h = \frac{N_h}{N_D}$$

This means that if the Neyman allocation is to be used, then the relative standard error in domain D is given by:

$$\frac{\sqrt{\frac{N_D{}^2\left(\sum_{U_h \in D} W_h S_{yU_h}\right)^2}{n_D} - N_D \sum_{U_h \in D} W_h S_{yU_h}^2}}{t_{y(D)}} \leq \alpha_D$$

Solve the equation with respect to n_D:

$$n_D = \frac{N^2{}_D\left(\sum_{U_h \in D} W_h S_{yU_h}\right)^2}{\alpha_D{}^2 t_{Y(D)}{}^2 + N_D \sum_{U_{h \in D}} W_h S_{yU_h}^2}$$

The total sample size, n_D, is calculated and the corresponding Neyman allocation over size strata is given by:

$$n_h = \frac{n_D W_h S_{yU_h}}{\sum_{U_h \in D} W_h S_{yU_h}}$$

ISSN 1650-9447

Statistical publications can be ordered from Statistics Sweden, Publication Services, SE-701 89 ÖREBRO, Sweden (phone: +46 19 17 68 00, fax: +46 19 17 64 44, e-mail: publ@scb.se). If you do not find the data you need in the publications, please contact Statistics Sweden, Library and Information, Box 24300, SE-104 51 STOCKHOLM, Sweden (e-mail: information@scb.se, phone: +46 8 506 948 01, fax: +46 8 506 948 99).