UNINTENTIONAL RE-SAMPLING OF INDIVIDUALS AND HOUSEHOLDS

Håkan L. Lindström, Statistics Sweden

1 Possible consequences of re-sampling

When several samples are drawn from the same population, and with replacement after each of them, some units will be members of more than one sample. When this is not a consequence of planned full or partial coordination of the samples we will refer to these unintentionally re-sampled units as urits and refer to the occurrence of such urits as unintended re-sampling UIR

The reaction of an individual who is asked to participate in a second survey, maybe within a short time after the first, is not quite predictable. It is feared that some will become non-respondents when they are asked the second time and afterwards. Another possibility is that, while they can be persuaded to participate in the survey, they are reluctant respondents, so that their answers will be incomplete and more likely to contain errors. If there are such effects they may be cumulative and more important every year. Positive effects may also be possible. Some people who intended to refuse but were persuaded to be respondents have explained that the reason why they hesitated was fear of not being able to answer the questions. Afterwards they decided that this was much easier than they had expected.

When repeated participation in surveys is planned, the expected advantages are extended analytic possibilities and gain in the precision of some estimators. When repeated sampling is unintended, none of these advantages is there and the consequences are thought to be mainly bad.

That this issue is not merely a theoretical one has been illustrated by reports from the interviewers of Statistics Sweden. In one case three interviews were made in one household in the same day. A woman was interviewed in the morning for the Labor Force Survey and in the evening for the Survey of Living Conditions. Since the Survey of Living Conditions used a family cluster sample, her husband was interviewed afterwards. In another family, husband and wife were sampled independently of each other in the Labor Force Survey, each with interviews on eight occasions within two years. The household was also sampled in the Survey of Consumer Buying Expectations with interviews on five occasions. The following year the wife was sampled in the Survey of Political Party Preferences with interviews on three occasions. In a third case at least eleven interviews were reported within a few years among five families living on the same small street.

According to the interviewers, such coincidences are not unusual but far too frequent. The interviewers often have to explain to people why they have been re-sampled and have to work hard to convince the re-sampled individuals that they should participate in the surveys. One interviewer said that about 25 percent of the non-respondents declared that previous sampling was the reason for their refusal.

2 The aim of the studies

As Sweden has a comparatively small population and several surveys are performed each year, the problem may be more embarrassing here than in many other countries. It has also been under study for several years at Statistics Sweden.

In order to get an adequate description of the extent of sampling and unintentional re-sampling of individuals and households in Sweden and to be prepared to reduce the consequences of UIR if obvious disturbances were found, a number of studies has been made.

Fear of increasing non-response rates and impaired data quality is not the only reason for these studies. It is also thought that, in fairness, the response burden should be distributed as evenly as possible. An evenly distributed response burden may also be worth aiming at when one explains the concern about confidentiality problems to the sampling population.

Anyhow, similar problems will appear also in countries with large sampling populations, for example if the same primary units are retained in several multistage samples.

3 The level of sampling

The most heavily sampled age groups are those between 15 and 75 years of age, as many surveys center on characteristics that are infrequent among the young and the old. The risk of increased non-response errors, measurement errors and coast are complementary arguments for truncation of the sampling population by age.

The size of this sampling population in Sweden is just above six million people. In the seventies the population register of Statistics Sweden was used as a sampling frame not only for our own surveys, but also for those performed by the leading private sampling institutions. For one year, 1977, it was possible to calculate the total number of individuals or households that were sampled from this sampling frame. It amounted to more than 420,000, and 230,000 of these were for the surveys of Statistics Sweden. As far as we can judge, the number sampled by the same organizations was not much different in the years immediately before 1977 and has not changed much afterwards.

As some surveys are panel studies and range over more than one year, the number of people newly sampled each year is smaller. While there are also surveys using other sampling frames, it was not possible to estimate how many more units were sampled that those included above.

This means that during a year with the total sample size of 1977, at least one in fifteen Swedes is sampled. If the sampling fraction of 1977 is uniform and is maintained in the future, every Swede should expect to have been sampled at least four times by the time he or she is 75 years old.

This outline is of course simplified. Some surveys are stratified, others use varying sampling probabilities and in some cases the sampling population is restricted with respect to region or age of the individuals. In some cases the sampling unit is a household and not an individual. The total sampling probability obviously varies between individuals and cannot be easily calculated. We must be content with approximate results in calculating the number of re-sampled individuals.

4 The calculated number of re-sampled individuals and households

The expected number of common elements in two simple random samples from the same population and with replacement after the first sample can be calculated according to the formula $N f_1 f_1$, where N is the population size and f_1 and f_2 are the two sampling fractions.

The number of re-sampled units in a population depends on the number of samples and the size of each. As a first approximation of the total number of urits, let us suppose that a total sample size of n sampling units is divided into k equal-sized simple random samples (with replacement after each sample); this results in an expected number of

$$N\left(\frac{n/k}{N}\right)^2 \left(1-\frac{n/k}{n}\right)^{k-2} \frac{k(k-1)}{2} \doteq \frac{n^2}{N}\,\frac{k-1}{2k}$$

objects sampled exactly two times.

With a population size of six million and a total sample size of 420,000, the number of units sampled twice will be between 7,500 and 15,000. The higher number is approached when k increases. The expected number of units sampled exactly three times will be a few hundred.

As sampling of individuals gives a household a sampling probability that depends on its size, the mean number of individuals in households in which someone is sampled will be higher than the population mean. The mean number of members in such households might be approximately three persons; thus in 1977 around 1.2 million people belonged to households where at least one person was sampled. The expected number of households in which two individuals or more are sampled is not easily calculated but should, if anything, be greater than the number of re-sampled individuals.

It might be more appropriate but obviously more complicated to discuss the response burden on households than on individuals. Many surveys are household surveys. Also, other household members can be exposed to the data collection in several ways even if the sampling object is an individual, for example by proxy interviews.

5 Urits in the surveys at Statistics Sweden

In order to get some objective and at least partially representative data, some samples used in the surveys of Statistics Sweden were studied together. 150,000 individuals and members of sampled households included in samples of 1977 were represented in one register and 90,000 individuals and members of households newly sampled in 1978 in another. The two main objects of the study were to find out the number of urits and the effects of UIR on non-response.

The general characteristics of those surveys included in the study are that they use samples from the same population register, that the data collection method is interview (usually by telephone) and that participation is voluntary. The samples are nation-wide and may be stratified, but cluster sampling and two-stage sampling are not used. The Income Distribution Survey is an exception in that it relies on data collection by mail, but uses telephone follow-up among non-respondents.

The situation under study is in many regards very complex and cannot be described briefly.The results will be approximate to the same degree. In a household survey sometimes one member can answer for all, as in the Survey of Consumer Buying Expectations, and sometimes cooperation from several members is requested, as in the Households' Expenditure Survey. When individuals are the sampling objects they can be sampled in family clusters, as in the Survey of Living Conditions up to 1979. There may also be proxy interviews when the sampled persons are not available. Nor can we be sure that just the one who actually participated in the survey is the only one whose subsequent response behavior was influenced.

The number of those sampled in 1978 who also were in the register of the 1977 samples was 1,530. When approximated with the simple formula $N f_1 f_2$ for all combination of surveys, the calculated number of urits was 1,596. The agreement is surprisingly good, as only one survey used simple random sampling of individuals. Some of the differences between observed and calculated values arise from differences in the age limits of the sampling populations.

In the surveys that relied on interviews for data collection there was no difference in non-response rate in 1978 between those who were in the 1977 sampling register and the others. The individuals could be classified by sex, marital status, age and region. No differences between urits and others in separate population groups could be established. In the Income Distribution Survey, a mail survey, the non-response rate was ten percent higher among the urits than the others. After a telephone follow-up the final non-response rate was almost the same in both groups.

As the group "others" probably included individuals who had been sampled earlier than in 1977 by Statistics Sweden or had been sampled by other agencies up to 1978, one cannot reject the hypothesis that a difference in non-response rates would be possible to establish, if we could discriminate perfectly between the first-time sampled and the formerly sampled.

As part of the investigation a dozen interviewers were instructed to report all cases in which either they could identify a household or individual as formerly sampled or someone spontaneously mentioned that this was the case. The study was made in February, March and April 1979. No regard should be taken to how long time that had passed since the previous survey. The fraction of reported urits was three percent of the sampling units allotted to these interviewers. The percentage reported is very low compared to other estimates of the number of urits. Its level depends on the employment time of the interviewers and on the memory of both parties and very much on the fact that when a household is resampled, the interviewer is not necessarily the same even if the household hasn't moved to another region. It was mentioned that several of those in the sample of the Labor Force Surveys had already been in the sample of the same survey once before. In a few cases the individuals were sampled in more than one survey within the three months of the study.

6 Recollection of sampling among respondents

In 1980 Statistics Sweden conducted a survey whose main topic was the public attitude to the 1980 census. The questionnaire included some questions concerning experience of sample surveys. The individuals in the sample were specially told that they should disregard censuses and information collected for administrative purposes. The size of the net sample amounted to 807 persons and the non-response rate was 18 percent - half of which were people who refused to participate. Evidently the non-response error cannot be disregarded when the effects of earlier samples are analyzed. Still, some careful conclusions are possible.

Memory of inclusion in a former sample, percent

Sampled	25	
respondents		23
non-respondents		2
Not sampled	73	
Don't remember	2	

Obviously many samplings are forgotten. At a sampling rate of at least seven percent a year, as in 1977, more than 25 percent of the population should have participated in at least one sample. This is all the more obvious as, when a similar question was put in a survey in 1976, about 40 percent of the respondents thought that they had been sampled at least once before.

The proportion of non-respondents among those who said they had been sampled before is less than one tenth, which is below the usual non-response rates (which range between ten and twenty percent in well-managed surveys). This might indicate a higher proportion of urits among the non-respondents of this particular survey. Only the Labor Force Survey and a few others have non-response rates below ten percent.

Eight percent of the respondents reported that some member of their household had been sampled in the previous twelve months. Those who remembered being respondents before reported as many as 14 percent. The difference may depend on the fact that some household surveys ask for the cooperation of two or more household members, but also may arise from a higher degree of recollection among those who have been sampled more often.

Reported period of previous survey, percent

1980 Jan - Sept	8
1979	8
before 1979	9
Not sampled	75

The duration of the recollection of surveys was also studied. The reported percentage "before 1979" is just a bit greater than the percentage of units sampled from the population register in 1977 alone, and this didn't even include all who were sampled that year. The value for "1980 Jan - Sept" is on a reasonable level when compared to the observed level of 1977, and the value for 1979 might also be reasonable.

The data may be subject to both non-response and measurements errors, but even allowing for a high degree of error in the reported year of the previous survey, it is obvious that the duration of the recollection of surveys is not very long. It might also be a reasonable hypothesis that the probability of being a non-respondent in a particular survey on grounds of having been sampled previously is greater the more recently the survey was performed. Consequently non-response bias should be expected to be less important in estimates concerning more distant years.

Of those who said that they had participated earlier, one third also said that they had been sampled twice or more. When the respondents were asked which organization had performed the most recent survey in which they were contacted, one third identified Statistics Sweden, one third private researchers and the rest other researchers or did not remember. The distribution for Statistics Sweden and private survey institutions is almost the same as that calculated for 1977.

The results could be summarized thus: Every year about 10 percent of the adult Swedish population are included in a sample. Many of these forget their participation within a fairly short time. After two years the effect of forgetfulness is large, but there are indications that those effects are weaker in households that have been included in more surveys. About one fourth of the population will remember that he/she was sampled before and less than half of these will remember Statistics Sweden as the data collector.

7 The present considerations

As a course of action to reduce UIR and its assumed harmful effects it was proposed that an individual who once was sampled by Statistics Sweden should be exempted from re-sampling for at least one year after the data collection was completed. A working system of coordination of samples of enterprises can be modified and applied to the population register used as a sampling frame when individuals and households are sampled. The working principle in this system is to assign a random number to every unit in the sampling frame. Each survey is assigned a certain interval of random numbers. If it is desirable that two samples have common units, they are given the same interval; otherwise they are given separate intervals. Stratification with varying sampling fractions within strata can be accomplished by use of subintervals. It should be noted that this sort of coordination cannot be complete, since other sampling frames are used. Neither would it guarantee the exemption from sampling of members of the sampled person's household.

The obvious gains if such measures were taken would be that our interviewers would know that a sampled person had not recently been sampled by Statistics Sweden. They would also be able to guarantee every sampled individual that he/she wouldn't be sampled again for at least one year. The risk of difficulties in data collection would be lessened.

The accomplishment of these ideas has been delayed, however, for various reasons. First of all, there was no clear evidence that UIR had harmful effects, at least none that couldn't be overcome by the interviewers. The total non-response rate has not increased since the studies were planned and has even dropped in a couple of surveys. Still more important is a current study of the feasibility of coordinating the definitions, content and sampling of three important surveys - the Household Expenditure Survey, the Survey of Living Conditions and the Income Distribution Survey. Only when this study is finished it will be meaningful to plan sample coordination between these and other surveys of Statistics Sweden.