CAN PLUS DIGIT SAMPLING GENERATE A PROBABILITY SAMPLE?

Gösta Forsman, Stig Danielsson, Linköping University
Gösta Forsman, Department of Mathematics, Linköping University, S-581 83 Linköping, Sweden

Key Words: Telephone Surveys, Plus Digit Sampling

1. Introduction

In countries with a large part of the population reachable by telephone, telephone interviewing has become an important mode of data collection in interview surveys.

The procedure for sampling individuals or households in telephone surveys is often an area sample or a register sample of individuals or households to which telephone numbers are matched from directories or other sources. However, survey agencies early recognized the possibility of reaching households by sampling telephone numbers, thus saving time and money in the data collection process.

A major sampling problem connected with phone number sampling is the lack of a complete sampling frame that covers all residential phone numbers in the population. Directories cover only a fraction of the telephone household population. It has, e.g., been estimated that in the United States, less than 60 % of the telephone households are listed, while 64 % are listed in the United Kingdom. Unlisted numbers are of two kinds; numbers activated after the directory was printed and secret numbers. By secret numbers we mean that the number is not listed in any directory and the subscriber pays for not having it listed. Studies have shown that households with unlisted numbers differ from households with listed numbers for many characteristics.

Probability sampling methods for phone number sampling have been developed, in particular in the United States and Canada. These methods generate probability samples of the total telephone household population, including both listed and unlisted numbers. The simplest method is to randomly generate the number of digits necessary to form a telephone number (ten digits in the United States and Canada). This method, a simple form of so called Random Digit Dialing (RDD), is, however, inefficient since only a few percent of the generated numbers would be household numbers. To improve the efficency of RDD, information about clustering of residential numbers can be used as auxiliary information in the sampling design. This can increase the hit rate (i.e., the proportion residential numbers in the sample) to over 50 %. RDD can also be combined with autodialing (to exclude not activated numbers from the RDD sample) and matching with lists of business numbers (to exclude business numbers), which further increases the hit rate.

In Sweden, telephone surveys have been used in official statistics on households and individuals as well as in market and opinion research since the early 1960s. A high coverage rate among Swedish households - estimated to about 99% - has contributed to make telephone the today predominant mode of data collection in these areas.

Phone number sampling is used in Sweden only in private sector agencies. The sampling methods used are almost exclusively non-probability sampling methods. Probability sampling methods, such as Random Digit Dialing (RDD), have rarely been used. The main reasons are

a) The good alternative sampling frames. The Swedish Population Register is commercially available as a sampling frame. It offers simple and inexpensive one-stage sampling of individuals and households. Using this register, the inclusion probabilities of the sampling units can be controlled better than in phone number sampling. The disadvantage with using this frame is that the register does not include phone numbers. Therefore, phone numbers must be matched from directories and other sources.

b) Important information about the phone number system is not readily available. For example, Swedish Telecom has so far not supplied information about the number of residential listed numbers in sequences, or banks, of 100 or 1000 numbers in a computer readable form. This is not because of policy reasons but rather a lack of communication between research agencies and Swedish Telecom.

c) The Swedish numbering system is not standardized, like, e.g., that of the United States. There are variable length of numbers and variable length of area codes and prefixes.

d) Because of the above reasons, there is a cultural resistance to RDD among research agencies. RDD is seen as something that is suitable in the United States but not in Sweden. RDD is regarded too expensive, too inefficient and too difficult to administer.

Swedish Telecom offers on a commercial basis systematic samples of phone numbers from an up-to-date file of "listed" household numbers, i.e., the non-secret residential numbers that would have been printed in a directory (if a directory had been printed at the sampling date). The advantage of using this frame is that the unlisted residential numbers almost exclusively are reduced to secret numbers.

In Sweden, one of the most commonly used sampling methods in marketing and opinion polls is *plus digit sampling*. In this procedure, an integer (e.g., the number one) is added to the last digit of each of the phone numbers sampled from a directory, thus constituting a new sample which generally includes both listed and unlisted numbers.

The simplicity of this sampling method is very appealing. However, it does not generate a strict probability sample, since the inclusion probabilities for residential numbers following either an unlisted residential number, a non-residential number or a non-working number, are zero. Thus, under the traditional theoretical sampling framework it is not possible to estimate population parameters unbiasedly.

Despite the fact that a plus digit sample is not a strict probability sample, it is often regarded in Sweden as an approximate simple random sample and simple random sampling estimators of population parameters are used. This is done for the sake of convenience and without any theoretical justification.

The purpose of the research presented here is to develop a model for the plus digit sampling situation under which unbiased estimators of population parameters can be derived. Under the model, we also calculate the bias of the plus digit sample mean when it is used as an estimator of the population mean. The basic assumption in the model is that the mix between listed and unlisted numbers is random. This provides a possibility to calculate inclusion probabilities and thus unbiased estimators of population parameters for a plus digit sample.

Tests of randomness on a data set indicate that the assumption of randomly mixed listed and unlisted numbers is realistic within 100-banks.

In Section 2, we describe the Swedish phone number system and in Section 3 our data set. Section 4 introduces plus digit sampling. In Section 5, the model is presented as well as inclusion probabilities for plus digit sampling under the model. Estimation of the population mean under the model is discussed in Section 6. Section 7, finally, deals with model justification including the tests of randomness.

2. The Swedish Phone Number System

In Sweden a phone number consists of two parts, area code and subscriber number. The area code consists of two parts: a national prefix - the digit 0 - and the area number. Also the subscriber number consists of two parts: a series number and a subscriber suffix. The area code can include two, three or four digits. The subscriber number includes five to eight digits of which the number series are 1 to four digits and the subscriber suffix always includes 4 digits. Per July 1, 1995, there exists 264 area codes.

Examples.

0	142	1	1426
0	13	28	1000
national prefix	area number	series number	subscriber suffix

$$\underbrace{\text{national prefix}\quad\text{area number}}_{\text{area code}}\quad\underbrace{\text{series number}\quad\text{subscriber suffix}}_{\text{subscriber number}}$$

For sampling purposes the mix between business numbers and residential numbers is important. This mix varies between 100-banks. Most 100-banks contain a majority of residential numbers, around 70-90 percent, and 5-10 percent business numbers. However, other distributions occur occasionally and a substantial part of the 100-banks are entirely devoted to business numbers. See also Danielsson and Forsman (1997).

There are (per July 1, 1995) 5 595 845 residential phone numbers in Sweden. The secret numbers are 378 760 or 6,77 % of all residential numbers.

The percentage secret phone numbers is considerably higher in the bigger cities in Sweden than in smaller cities and rural areas. In the three largest cities, Stockholm, Gothenburg and Malmö the percentages of secret numbers are 10 %, 12 % and 14 %, respectively.

3. The Data Set

To evaluate plus digit sampling and to validate our model, we used a data set of 20 000 consequtive numbers from the city of Linköping, including information of the phone number status: listed residential, unlisted residential (i.e., in our case secret numbers), business, not activated.

This information about phone numbers is not commercially available. However, systematic samples from the category "listed residential numbers" can be bought from Swedish Telecom.

The distribution of the four categories in each of the 200 hundred-banks are presented in Danielsson and Forsman (1997).

4. Plus Digit Sampling

Generally, plus digit sampling is a list assisted procedure in which a sample is selected from a directory and an integer is added to the last digit of the selected numbers (Lepkowski, 1987). The new number created is the number being included in the sample (the original listed numbers being discarded). The added digit is usually 1 (plus one dialing) but can be any fixed digit, or two or more digit numbers. The digit(s) may also be randomly selected.

The plus digit sample procedure we shall discuss here is plus one sampling with the initial sample being a systematic sample from a population of listed residential phone numbers.

The major theoretical problem with plus digit sampling is that the inclusion probabilities are zero for an unknown percentage of the population of phone numbers. For example: In our Swedish data set of 20000 consequtive phone numbers, 14.5 % of the residential numbers had inclusion probability zero. Among non-secret numbers the percentage was 14.4 while it was 14.8 among the secret numbers.

One problem sometimes mentioned with plus digit sampling is that the hit rate (the proportion residential numbers in the sample) declines the greater the distance is from the initial number. This is not obvious the Swedish case. Simulations on our data set resulted in the hit rate 95,1 % for plus one sampling, 94,6 % for plus five sampling, 94,9 % for plus ten sampling, and 94,0 % for plus 20 sampling.

Another concern with plus digit sampling is that although the technique is effective in bringing in unlisted numbers in the sample, the proportion may be lower than the proportion known to be unlisted. This does not seem to be the case in Sweden. Further simulations on our data set resulted in the proportions 9.94 % for plus one, 9.81 for plus five, 9.83 for plus ten, and 10.16 for plus 20 sampling. The true proportion in the data set was 10.05 %.

5. The Model

We adopt the following view of the sampling situation: We assume that the sampling frame, i.e., the listed residential numbers, is randomly generated within each 100-bank. Then the resulting plus-digit sample, with non-residential numbers excluded, might be viewed as an approximate simple random sample of all (listed and unlisted) telephone households within each 100-bank. More specifically, we assume that within each 100-bank the actual order of telephone numbers is a randomly selected permutation of all possible permutations of phone numbers. We also assume that the unlisted residential numbers are randomly mixed among all residential numbers. This model can be regarded a superpopulation model, see, e.g., model E_{RP} in Cassel, Särndal and Wretman (1977, page 87).

We notice that under this model, a systematic sample is equivalent to a simple random sample, since the frame is randomly generated. We also notice that, under the model, all residential phone numbers have nonzero inclusion probabilities. To calculate the inclusion probabilities, we look at sampling from one 100-bank. Denoting the number of residential telephone numbers by N, the number of listed residential telephone numbers by N^l and the sample size by n, we have (see Danielsson and Forsman, 1997) the inclusion probability for a listed residential number

$$\pi_i = \frac{N^l - 1}{N^l}\frac{n}{N - 1}$$

and for an unlisted residential number

$$\pi_i = \frac{n}{N - 1}.$$

The difference between the inclusion probabilities depends on the fact that the initial sample is selected from a frame of listed numbers which slightly reduces the inclusion probabilities of the listed numbers in the final sample.

We notice that for both listed and unlisted numbers, the inclusion probabilities are close to those of simple random sampling, i.e., $\pi_i = \frac{n}{N}$ for all i.

The joint inclusion probabilities π_{ij} for two residential numbers i and j are

$$\pi_{ij} = \frac{n(n - 1)}{(N - 1)(N - 2)}\frac{N^l - 2}{N^l}$$

if both units are listed,

$$\pi_{ij} = \frac{n(n - 1)}{(N - 1)(N - 2)}$$

if both units are unlisted, and

$$\pi_{ij} = \frac{n(n - 1)}{(N - 1)(N - 2)}\frac{N^l - 1}{N^l}$$

if one unit is listed and the other is unlisted.

We notice that the joint inclusion probabilities are very similar to the corresponding ones [$\frac{n(n - 1)}{N(N - 1)}$] for simple random sampling of n units from N.

Remark: "n" is the size of the final plus digit sample with non-residential numbers excluded. The initial sample size is larger since some of the numbers in the initial sample are not followed by a residential number. Thus, in strict terms, n is a random variable and the calculations here are conditional to a fixed value of n (n>0).

6. Estimation of the Population Mean Under the Model

Recall that the model postulates that the order of the phone numbers in the population is generated randomly within 100-banks. We will first look at estimation within a 100-bank and then at estimation over all 100-banks. Details of the calculations are presented in Danielsson and Forsman (1997).

Estimation Within a 100-bank

The Horvitz-Thompson estimator $\hat{\mu}_y = \frac{1}{N}\sum_{i=1}^{n}\frac{y_i}{\pi_i}$ estimates the population mean μ_y unbiasedly. A more interesting estimator of μ_y from a practical point of view is the sample mean $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ which is unbiased only if all π_i are equal.
The bias of the sample mean when estimating μ_y is

$$\text{Bias}\,(\bar{y}) = E(\bar{y}) - \mu_y = \frac{1}{N - 1}(\mu_y - \mu_y^l)$$

where μ_y^l denotes the mean of the y_i:s in the listed population.

Thus, if the sample mean is used for estimating μ_y, the bias is very likely small and may be regarded as negligible for many practical purposes.

Estimation Over All 100-banks

Let the frame consist of several (L) 100-banks. The 100-banks may be regarded as strata or clusters, depending of the length of the sampling interval in the initial systematic sample. When the sampling interval is considerably less than 100, a systematic sample of phone numbers from the whole frame is obviously equivalent to a proportional stratified random sample under our adopted model. In many situations, e.g., in nationwide surveys, the sampling interval is, however, much larger than 100. Typically in Swedish nationwide opinion polls, the sample size is 1000, in which case the sampling interval is around 5000. Then, under the model, the sampling method is equivalent to two-stage cluster sampling with one selected unit in the second stage.

We will confine the discussion here to the latter case. As a consequence, we will regard a 100-bank as a cluster.

When estimating population parameters over all 100-banks, the above notations and formulas are still valid within a given cluster (k), and formally we should use a subscript "k" to point out this fact. Thus N_k and N_k^l are the numbers of residential phone numbers and listed residential phone numbers, respectively, in cluster k. The plus digit sample size, n_k in cluster k, is 1 for each selected cluster.
We denote the total population size $N_0 = \sum_{k=1}^{L} N_k$ and the total sample size n_0.

Under the adopted model, the systematic sampling scheme is equivalent to two-stage sampling of $n_0 \le L$ 100-banks or clusters with sampling with probabilites (p_k) proportional to the number of listed residential numbers (without replacement) in the first stage, i.e., $p_k = \frac{N_k^l}{N_0^l}$, and random sampling of one unit in the second stage.

In the first stage clusters are drawn with probabilities $n_0 p_k$. In the second stage, the inclusion probabilities of phone numbers within a cluster are given in section 5. Rewritten with the subscript k for cluster and sample size $n_k = 1$, these inclusion probabilities are $\pi_{ki} = \frac{N_k^l - 1}{N_k^l} \frac{1}{N_k - 1}$ for a listed number and $\pi_{ki} = \frac{1}{N_k - 1}$ for an unlisted number.
Un unbiased estimator of the population mean is $\hat{\mu}_y = \frac{1}{N_0} \sum \frac{T_k}{n_0 p_k}$, where T_k is an unbiased estimator of the cluster total, i.e., $T_k = \frac{y_{ki}}{\pi_{ki}}$. This estimator is, however, not interesting for practical purposes, since the π_{ki}:s usually can not be calculated because the N_k:s are unknown. Instead, the sample mean $\bar{y} = \frac{1}{n_0} \sum_{k=1}^{n_0} y_k$ is commonly used as an estimator of the population mean. The bias of $\bar{y}$ can be shown to be

$$Bias(\bar{y}) = \sum_{k=1}^{L} (p_k - w_k)\mu_{yk} + \sum_{k=1}^{L} p_k \frac{1}{N_k - 1}(\mu_{yk} - \mu_{yk}^l),$$

where $w_k = \frac{N_k}{N_0}$. In practice, the first term may be small, since very likely $p_k - w_k = \frac{N_k^l}{N_0^l} - \frac{N_k}{N_0}$ is close to 0. For example: In our data set, $p_k - w_k$ range between -0.00129 and 0.00057 with the arithmetic mean $-1.2 * 10^{-19}$.

The second term of the bias expression is also very likely to be small, since N_k is large compared to $\mu_{yk} - \mu_{yk}^l$. For example, if y takes the values 0 and 1 and we assume the extreme situation where $\mu_{yk} - \mu_{yk}^l$ is 1 for all k, the second term would still become as small as (roughly) $\frac{n_0}{N_0^l}$ which in nationwide Swedish surveys would be around 1/4500.

Thus, if the model assumptions are realistic, the bias when using $\bar{y}$ for estimating μ may be regarded negligible in practice.

Remark. It should be noted that the calculation of p_k is based on the assumption that the systematic sampling procedure is cyclic in the clusters. That is, if the last number in a 100-bank, ending with 99, is drawn in the initial sample, the resulting sample unit in the plus one sample is the number ending with 00 in the same 100-bank. In Swedish practise, the plus one sampling unit would be the number ending with 00 in the following 100-bank. Taking this into account would have a slight effect on the p_k:s.

7. Model Justification

Test of Randomness

We tested whether the three categories of numbers (i) listed residential numbers, (ii) unlisted residential numbers, and (iii) other numbers, are randomly mixed within 100-banks. The null hypothesis was that the three categories are randomly mixed while the alternative hypothesis was simply that they are not randomly mixed.

We performed a run test of randomness by using the number of runs (u) among the three kinds of elements.

Standard textbooks give the following expressions for E[u] and V[u] under H_0 (see, e.g., Brownlee, 1965), denoting the number of phone numbers of the above three categories with a, b, and c, respectively.

$$E(u) = \frac{2(ab + ac + bc)}{a + b + c} + 1, \text{ and}$$

$$V(u) = \frac{(2(ab + ac + bc))^2}{(a + b + c)^2 (a + b + c - 1)} - \frac{2(ab + ac + bc) + 6abc}{(a + b + c)(a + b + c - 1)}.$$

The standardized test variable $\frac{u - E(u)}{V(u)}$ is assumed to be approximately distributed $N(0,1)$.

Since 50 of the 200 hundred-banks in the data set had zero residential numbers, the test was conducted in the remaining 150 hundred-banks. The result was that H_0 was rejected in 9 of the hundred-banks, i.e., 6 %. See Danielsson and Forsman (1997) for details.

The test results indicate that the assumption of random mix between listed and unlisted residential numbers is a reasonably correct description of the real situation.

Procedure For Allotting Numbers to Subscribers

The procedure for allotting telephone numbers to subscribers is assumed to be random in our model. In reality the procedure is as follows: A new subscriber can choose between the four available numbers that have been available the longest. Occasionally, in say 5 % of the cases, the new subscriber may get a fifth number (or even more numbers) to consider. Although there may exist preferences for digit combinations that are common many many people, we feel confident that the model assumption of random allocation of numbers is fairly realistic with this procedure. This fact in combination with the results of the tests of randomness give some support for the model.

References

Brownlee, K.A. (1965). *Statistical Theory and Methodology in Science and Engineering*. 2:nd edition. John Wiley & Sons, Inc., New York.

Cassel, C.M., Särndal, C., and Wretman, J.H. (1977). *Foundations of Inference in Survey Sampling*. John Wiley & Sons, Inc., New York.

Danielsson, S. and Forsman, G. (1997). *Plus Digit Sampling. A Model-Based Approach With Some Applications to Swedish Data.* Research Report. Institute of Mathematics. Linköping university. (To appear)

Lepkowski, J.M. (1987). *Telephone Sampling Methods in the United States.* In: Groves et al.: Telephone Survey Methodology. John Wiley & Sons, Inc., New York.

Acknowledgements

The authors wish to thank Mr Lars Malmquist, Swedish Telecom, for preparing the data set used in this research, and Professor Sven Berg, University of Lund who provided valuable comments to an earlier version of this paper.

Support for this research was provided by the Swedish Research Council for the Humanities and the Social Sciences.