Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 77(1) 2023 27–34

Network Sampling for Estimation of the Size of a Finite Population with Special Features

Manisha Pal[1] and Bikas K. Sinha[2]

[1]Calcutta University, Kolkata (Retired Professor); currently associated with St. Xavier's University, New Town, Kolkata as Senior Professor of Statistics
[2]Indian Statistical Institute, Kolkata (Retired Professor)

Received 26 March 2023; Revised 14 April 2023; Accepted 15 April 2023

SUMMARY

We are interested in unbiased estimation of the unknown size of a finite population. The population units [also called Ultimate Units (UUs)] are not directly accessible in any way. We can only have access to any UU via an appropriate Reference Unit [RU] only if it captures the UU in question. The literature is scanty and the state of knowledge also seems to be imperfect. We aim at providing an overview of the literature in this fascinating area of research and its application.

Keywords: Estimation of size of specially featured population; Reference units; Ultimate units; Probing and no-probing; Bipartite Di-Graph; Networking.

1. INTRODUCTION AND LITERATURE REVIEW

Estimation of the unknown size (N) of a finite population [in different contexts/application areas] has been of interest to the statisticians. The well-known Catch - Mark - Release - Recatch method is talked about frequently. There are available non-trivial and often-needed alternatives to this method as well. We will not enter into discussions on the Catch-Recatch methodology as such but cite a few relevant references at the end. We also refer interested readers to Chaudhuri (2015), who effectively used network sampling and adaptive sampling in tandem to capture sparsely located elements in unknown pockets.

In this paper, we will deal with finite populations having some peculiar / special features and discuss the problem of estimation of the unknown size (N) of the specially featured population/subpopulation. The peculiarity of such a population lies in the fact that none of these population units is directly accessible to the investigator, but there is a well-defined finite population [also called reference population] closely connected to the population of interest, whose units may be accessed directly. For example, consider a general population and, within it, suppose there are individuals (i) having access to food coupons only through distribution centers or service centers, or, (ii) suffering from a rare disease and necessarily visiting one or more diagnostic centers for treatment, and the like. These individuals are otherwise not identifiable in the general population. The only way to identify them is by visiting the appropriate centers meant for such purpose.

It is postulated that there is a well-defined network connecting the reference population to the population of our interest. While the totality of reference units gives the total view of the population network of our interest, a sample of such reference units creates only a partial view of the same. Based on this sample network, it is required to unbiasedly estimate the size (N) of the population of our interest.

Corresponding author: Bikas K Sinha
E-mail address: bikasksinha118@gmail.com

We refer to Seber (1982) for literature review and a host of results in standard formulations of the problem. It needs to be noted that we are dealing with a non- standard set-up and follow a network approach to reach the ultimate units of our interest. The first non-trivial study of the problem of estimation of N based on a partial view of the network seems to be due to Kiranandan (1976). Following this, there are two more studies. Maiti *et al.* (1993) derived interesting results on unbiased estimation of N, being completely unaware of Kiranandan′s work. Later, Sinha *et al.* (2006) took cognizance of Kiranandan′s work and presented an equivalent version, along with a critical appraisal of Kiranandan′s work. These studies address the above problem with/without additional information [which may be made available in terms of what has been termed as probing]. Based on a random sample of reference units drawn from the reference population and the sample network drawn, we need a count of the number of the ultimate population units so captured and their individual incidences within the sampled reference units. In most studies, this alone would constitute the data and we should proceed towards estimation of N from here. This corresponds to the scenario of no probing. However, it seems to be an utterly impossible proposition to go anywhere from here! Kiranandan (1976) succeeded in deriving a novel formula for $(\hat{N})$, an unbiased estimator for N, in terms of the frequency counts of the ultimate population units captured via the sample network. Unfortunately, in most applications, $(\hat{N})$, so computed, may turn out to be a large negative quantity! There are other undesirable features of $(\hat{N})$ as well, like it may turn out to be less than the actual number of ultimate units captured/observed in the process of sampling. The criterion of unbiasedness thus seems to take its toll in this process.

2. HANDLING SPECIALLY FEATURED POPULATIONS

We are referring to finite populations whose units are not directly accessible in any manner. In that respect, apriori, no question of labelling of such units arises. Accessibility of the units occurs only through a finite [identifiable] collection of Reference Units [RUs]. That is why, we refer to the inaccessible units as Ultimate Units [UUs]. And we are also referring to the existence of a Bipartite Di-graph connecting the RUs with the UUs. Our objective in this study is to estimate the unknown number N of the size of the population of UUs.

2.1 Reference Units and Ultimate Units

Suppose there exists a total of M distinct service centers catering to a total of N distinct individual beneficiaries. An individual beneficiary may, however, avail of services from multiple service centers, without any restriction whatsoever. To start with, it is stipulated that the individual beneficiaries are neither directly accessible nor are they identifiable. The only way to learn about them is through visiting one or more of the service centres to which they are affiliated/attached/registered. The M service centres are distinct and identifiable, and form a finite labeled population of Reference Units - abbreviated as RUs. We assume this number M to be known. In case we decide to check/inspect all the service centres, we gain access to all the N beneficiaries, termed as Ultimate Units - abbreviated as UUs, through the underlying network which forms a bipartite digraph, in the language of graph theory. On the other hand, in case M is large so that complete enumeration is cost-prohibitive, we may as well take recourse to sampling of the service centres and end up with an incomplete description of the underlying network involving a part of the population of beneficiaries.

2.2 Bipartite Di-Graph and a Network of Reference and Ultimate Units

Imagine that a sample of RUs of size m, written as $s(m)$, has been selected according to *SRSWOR*(M, m). [Note that this could as well be any other Fixed Size (m), i.e., *FS*(m), sampling scheme with desirable properties]. Since, apriori, the UUs are neither identifiable nor bear any labelling for the total count N [which is itself unknown], we may assign labelling of these UUs in the sample as follows:

Set $RU_{s(m)} = [RU_{i1}, RU_{i2}, RU_{im}\ ;\ 1 \le i_1 < i_2 < \dots < i_m \le M\ ]$.

Note that the RUs are already labelled as (RU_1, RU_2, ..., RU_M) in some convenient order and we may conveniently/deliberately simplify the notations by dropping the prefix RU. So, WOLG, let us set $s(m) = [1, 2, ..., m]$. We pick up the first sampled RU, i.e., RU_1 and find out a listing of UUs contained therein. If this total number is n_1, we assign serial numbers UU_1, UU_2, ..., UU_{n_1} to these n_1 UUs in some manner. Note that there is no hard and fast rule for doing so. Next we check the second sampled RU, i.e., RU_2. If there are some common UUs with RU_1, we simply carry out the identification and do not assign any extra label to such UUs. If there are new UUs, we assign them serial numbers starting from UU_{n_1}+1, and so on until these new UUs [i.e., those not covered by RU_1] are exhausted. We continue this way till we come to the last sampled RU i.e., RU_m. Over here, we list all those UUs, if any, which have not so far been covered by the union of the RUs, viz. RU_1 to RU_m, taken up previously for listing of UUs. Incidentally, in the process, we have generated a Bipartite Digraph involving the two sets: RUs and UUs. Further, while the selected number of RUs is fixed, the number of UUs captured in the process is a random quantity. Let us denote this number by $n(s(m))$. Henceforth, we may use the notation n and the resulting sample of UUs may be denoted by $UUs(n)$.

We must note that the bipartite digraph so generated is based on the sampled RUs and the induced sample of UUs. We have henceforth deliberately used the notations *ru*s and *uu*s to indicate subset selection context.

2.3 Sampling from Reference Units and Creation of a Sample Network

In the above, we have mentioned about the population of M RUs and selection of a fixed size of m RUs, adopting SRSWOR or any other convenient $FS(m)$ sampling design. Based on this random selection of m RUs, conceptually we are capable of generating an underlying network connecting the sampled RUs and the induced UUs. This network will be referred to as a Sample Network. There is an underlying population network which is not accessible to us unless we carry out complete enumeration of all the M RUs. Before proceeding further, let us consider the example of food coupon distribution indicated in Section 1.

Suppose there are $M = 5$ Service Centers [SCs] i.e., RUs with reference to the given population in respect of providing Food Coupons. A beneficiary has free access to any one or two or even more SCs without any restriction whatsoever. We will be using RUs and SCs interchangeably and conveniently in the example.

Ideally, if we are in a position to scan through the records of all the M SCs, we have a perfect idea about the entire and complete Network of ALL SCs crossed with ALL beneficiaries in the reference population. This is the Population Network we are referring to. In that case, there is no need for any further statistical analysis at sampling level. However, most often, it is prohibitive to carry out a thorough investigation at the population level involving ALL the M SCs. That is why we suggest that one may adopt a convenient $FS(m)$ sampling scheme for random selection of m RUs out of M RUs. [Our example deals with only 5 SCs but in reality it could be 20 or even more - so that sampling a subset would be called for.]

To fix the ideas, we display the digraph in Fig. 1 below for the case of $M = 5$, $N = 8$, with one-way arrows from the right [RU] to the left [UU].

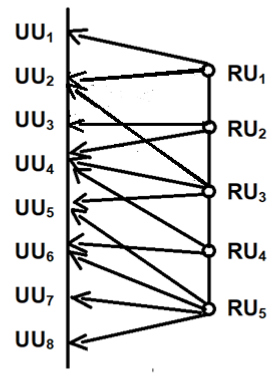

Fig. 1. Network of 5 RUs and 8 UUs

Left Side…UUs & Right Side…..RUs.

Naturally, this incidence pattern can be represented by an Incidence Matrix $I(5 \times 8)$ where $I(i, j) = 1$ if RU_i connects to UU_j ; otherwise, it is 0. The underlying incidence matrix is given in Table 1 below.

Table 1. Incidence Matrix of the Network of 5 RUs and 8 UUs

Reference Units	1	2	3	4	5	6	7	8	N_i
1	1	1	0	0	0	0	0	0	2
2	0	0	1	1	0	0	0	0	2
3	0	1	0	1	1	0	0	0	3
4	0	0	0	1	0	1	0	0	2
5	0	0	0	0	1	1	1	1	4
M_j	1	2	1	3	2	2	1	1	13

We have introduced the notations M_js and N_is, defined respectively as :

$$M_j = \sum_i I(i, j);\ N_i = \sum_j I(i, j).$$

As is evident, these are respectively the column totals and row totals of the incidence matrix. Further,

$$\sum_j M_j = \sum_{i,j} I(i, j) = \sum_i N_i.$$

This is exactly the total number of one-way ties originating from the right side and terminating at the left side. Readers interested in knowing more about such networks may consult Bandyopadhyay *et al* (2011).

For $m = 3$, assume that under *SRSWOR*(5, 3) sampling scheme, the selected sample of 3 RUs is indeed $s(m) = (RU_1, RU_3, RU_4)$. We can also determine from Table 1 that $n(s(m)) = 5$ and, as a matter of fact, the sampled UUs are labelled as $(uu_1, uu_2, ..., uu_5)$. In Figure 2 below we display the induced sample based biograph and Table 2 gives the incidence matrix of the same.

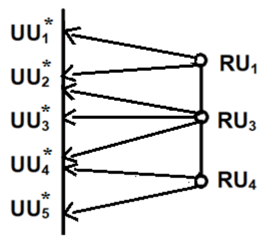

Fig. 2. Network of 3 RUs and 5 induced UUs

Left Side….UU*s & Right Side…..RUs

$UU_1^* = UU_1$; $UU_2^* = UU_2$; $UU_3^* = UU_4$; $UU_4^* = UU_5$; $UU_5^* = UU_6$.

Table 2. Incidence Matrix of the Network of 3 rus and 5 Induced uus

Reference Units	1	2	3	4	5	n_i
1	1	1	0	0	0	2
3	0	1	1	1	0	3
4	0	0	1	0	1	2
m_j	1	2	2	1	1	7

What would be our proposed estimate for N in this case?

It is interesting to note that the entire collection of RUs is available to the statistician and because these are all distinct, one can label them in any arbitrary manner and come up with $RU_1, RU_2, ..., RU_M$, thereby arriving at a finite labelled population of M RUs. It is now natural for the statistician to take recourse to a convenient sampling scheme, say $FS(m)$ sampling scheme with desirable properties of the first and second order inclusion probabilities. Vide Hedayat and Sinha (1991). The statistician is thus naturally drawn to the techniques of finite population sampling and inference. As a particular case, *SRSWOR*(*M, m*) becomes a natural choice.

We have already indicated a method of construction of the sample network based on (m = 3) RUs, *viz.*, RU_1, RU_3, RU_4. It led to n = 5 UUs, with the induced labelling as uu_1 to uu_5. We also have the corresponding incidence matrix of order 3 × 5 in Table 2.

Presumably this is our data. At this stage, are we in a position to develop adequate and appropriate theory for unbased estimation of N, the total number of UUs in the population? Even under *SRSWOR*(5, 3) sampling scheme, this seems to be an impossible proposition mainly because we are looking for an unbiased estimator for N. Incidentally, the interested readers can easily draw the sample bipartite graph connecting 3 rus with the 5 uus. Further to this, interestingly, we find that we also have data for each captured UU [through our sampling effort] on the number of RUs [out of the sampled 3 RUs] on which it is incident. Naturally, each such incidence number is at least one and does not exceed three. From the population network, we had exact information on N_j for each UU_j. The sample counterpart may be denoted by m_j. It is better to attach indicators to the labels of the UUs since only sampled UUs will possess such information on m_js. Thus, for example, $I(RU_{(s(3))}; (j))$ would provide necessary information where the indicator value is 1 iff UU_j is contained in the collection of UUs based on the sample $RU_{(s(3))}$. Accordingly, we note that $\sum_j I\left(RU_{s(3)}; (j)\right) = n$, the sample number of UUs captured through our sampling effort. We hope we have made the notations clear and meaningful.

As indicated earlier, this impossible proposition [of suggesting an unbiased estimator for N] was made possible by Kiranandan (1976) in his PhD Dissertation at Harvard University. We will take up a detailed presentation of this study in Section 3.2 below. Over there, we will closely follow the work done by Sinha *et al.*(2006). Prior to this, we would like to take up another [relatively easier!] approach for unbiased estimation of N. Here we will closely follow the work of Maiti *et al.*(1993) and this is contained in Section 3.1 below. Both the estimation problems are discussed in Section 3.

3. UNBIASED ESTIMATION OF THE SIZE OF A SPECIALLY FEATURED POPULATION

With our limited understanding about the amount of data on the available ultimate units, we wish to provide an estimate of the size N of the population of UUs. And, that too, an unbiased estimate! That is too much to hope for. There are very few studies under this framework.

3.1 Concepts of Probing and No-Probing

Let us look at the sample incidence matrix and the resulting sample quantities m_js. Is it possible to entertain the idea of 'probing' ? By that we mean we can approach all or a subset of these sampled UUs with the 'proposal / request' to release information on the number of *additional* [unsampled] RUs containing the specific UU whom we are requesting ! Now that we have hypothetically access to the population network, if we approach the sampled UUs and they oblige us by providing / releasing the requested information, then we will be in a position to assess the correctness of the figures and prepare an extended table as follows.

Table 3. Sample of 5UUs induced by SRSWOR(M = 5, m = 3) on 3RUs

Sampled RUs	Induced UUs
RU_1	[(1, 2)]
RU_3	[(2; 3, 4)]
RU_4	[(4; 5)]

Table 4. Induced UUs and their incidence counts

Induced UUs – – ––	Incidence Counts among Sampled RUs	Net Effect AfterProbing
UUs_1	$m_1 = 1$	$M_1 = 1 + 0 = 1$
UUs_2	$m_2 = 2$	$M_2 = 2 + 0 = 2$
UUs_3	$m_3 = 2$	$M_3 = 2 + 1 = 3$
UUs_4	$m_4 = 1$	$M_4 = 1 + 1 = 2$
UUs_5	$m_5 = 1$	$M_5 = 1 + 1 = 2$

In Table 4, the notations are supposed to be self-explanatory. After probing, information on Total Count of RUs in the entire population for each of the induced UUs becomes available. Before probing, this count [for every incident UU] was based on the evidence provided by Sampled RUs. On the other hand, after probing, this count reflects the total count of such incidences in the entire collection of M SCs. However, it is only for the induced UUs in the sample.

In passing, we may note that on final count, information accrued through probing amounts to Y_j for the sampled UU_{s_j} in the frame-work of unified theory of sampling. Maiti *et al.*(1993) followed this approach and used M_j for Y_j and applied sampling techniques to arrive at an unbiased estimate of N, the total number of Ultimate Units in the entire population, based on data under/after probing.

Once more we fix the notations.

We have denoted by M the total number of RUs in the population and by m the sample size for selection of RUs under random sampling. We have denoted by N the total [unknown] number of Ultimate Units in the population. Let us denote by n the sample number of UUs captured through our sampling effort. For the UU_{s_j} in the sample, we have denoted by m_j the number of sampled RUs connected to it before probing. We may rightly denote the corresponding count after probing as M_j. Clearly, for each UU_{s_j} in the sample, These are essentially all the information available to us. Note that we have assumed that each of the n UUs in our induced sample has been asked / requested to adopt probing and extend the count m_j to M_j.

3.2 Estimation of N Under Probing : Theory and Examples

When *SRSWOR*(M, m) is adopted for selection of m RUs out of M RUs, Maiti *et al.*(1993) suggested the formula

$$\hat{N} = (M/m)\sum_j \frac{m_j}{M_j}$$

and it turns out that this estimator is unbiased. In the particular case of $m = 1$, it readily follows that $\hat{N} \geq n$, where n is the sample count of induced UUs.

For the cases of $m > 1$, the estimator given above is unbiased; however, it is not clear if $\hat{N} \geq n$, based on data for each of the all possible C_m^M samples of RUs. At this stage, we skip the proof of unbiasedness of $\hat{N}$.

Maiti *et al.* (1993) started by considering a binary matrix of order $M \times N$, where M stands for the known number of Service Centers and N is the unknown number of Beneficiaries. As the set-up suggests, the entries in the matrix are 0s and 1s. That is what we have considered here. In Sinha *et al.* (2006), the entries have been generalized to any set of real values.

In the general case, it follows that whenever $I(j) = 1$ for some UU_j in the population network, under *SRSWOR*(M, m) sampling,

$E(m_j) = mM_j/M,$

M_j being the total count of RUs in the population network having UU_j incident on each one of them. To avoid triviality, note that $M_j = 0$ simply means that $I(j) = 0$ outright and hence this specific UU does not exist in the existing framework of Reference Units versus Ultimate Units.

Note further that this estimation is possible since, under probing, M_j becomes available whenever $I(j) = 1$ and UU_j is included in the sample as UUs_j so that the investigator gets hold of M_j. Strictly speaking, under $SRSWOR(M, m)$ sampling, for every UU_j for which $I(j) = 1$, m_j has Hypergeometric Distribution with parameters $[M, M_j; m]$ and, hence, $[0 \le m_j \le min(M_j, m)]$. This explains the validity of the formula for $\hat{N}$ in general for $m \ge 1$.

We now take an illustrative example and, for the purpose of illustration of the computations, we may use the same network as shown above.

Example 3.1

In the given network, we have $M = 5$, $m = 3$. Further, while RU_1, RU_3, RU_4 are the selected RUs, these collectively induct a total of 5 UUs. Moreover, with probing applied to each of these 5 inducted UUs, the frequency counts [viz., both m_j and M_j] are shown in Table 4. From this, we derive $\sum_j \frac{m_j}{M_j} = 11/3$.

Therefore, on final count,

$\hat{N} = (M/m) \times \sum_j \frac{m_j}{M_j} = (5/3) \times 11/3 = 55/9 = 6$ [*approx.*]

Remark 3 In case an arbitrary $FS(m)$ sampling design is adopted, the formula changes to

$$N(s(R)) = \sum_i I(i, s(R)) \sum_j I(j|i) / R_j \pi_i$$

Here $I(i, s(R)) = 1(0)$ when RU_i is (is not) in the sample $s(R)$) of RUs. Likewise, $I(j|i)$ is the usual indicator relating to RU_i and UU_j in the incidence network. Further, with reference to the adopted $FS(m)$ sampling design, π's are the first order inclusion probabilities of the RUs.

The above formula is given in Maiti *et al.*(1991). Available multiple estimators due to Sirken (1970), Birnbaum and Sirken (1965) are of the above type.

With this, we close our discussion on unbiased estimation of N under probing.

3.3 Estimation of N Without Probing : Theory and Examples

This time, as before, we start with a random selection of RUs and arrive at the induced collection of UUs. We do not feel encouraged to undertake the task of probing - not even for a subset of the selected and induced UUs. We go back to the example considered in the earlier set-up. Thus, there are altogether 5 RUs and we have selected 3 of them and the data accrued on the induced UUs results in 5 of them with frequency counts (m_i s) as given in Table 3. We reproduce the m_j-values here.

[1, 2, 3, 2, 2]

Without any probing, our data collection effort ends here. It seems quite a daunting task to come up with any clue towards *unbiased* estimation of the total number of UUs i.e., of N in the population. Kiranandan (1976) did that seemingly impossible task and arrived at a fundamental identity showing a representation of N.

Heuristically, $N = \sum_j I(j)$, where $I(j)$ assumes the value 1 whenever UU_j is a valid entity in the collection of UUs. But the verification comes only through the RUs which are M in number. We may set $I(j|i) = 1$ if and only if UU_j is induced by RU_i. Moreover, $I(j|i)I(j|k) = 1$ if and only if both the indicators asssume value 1, and so on. We may now state below one useful identity:

$$I(j) = \sum_i I(j|i) - \Sigma \sum_{i<k} I(j|i) I(j|k) + \Sigma\Sigma \sum_{i<k<t} I(j|i) I(j|k) I(j|t) - \ldots + (-1)^{M-1} I(j|1) I(j|2) \ldots I(j|M).$$

The proof of this identity is very elementary in nature. It follows from the counting argument and use of inclusion-exclusion principle. Imagine UU_j is inducted by some s RUs and WOLG, we may assume : $I(j|1) = I(j|2) = \ldots = I(j|s) = 1$. Then the LHS is unity and the RHS is

$$f(s) = s - C_2^s + C_3^s - C_4^s + \ldots$$

It is easy to argue that $f(s) = 1$ for each $s = 1, 2, \ldots$.

We may now state the Fundamental Identity or, the Fundamental Representation Theorem :

Theorem 3.1

$$\boldsymbol{N} = \sum_j I(j) = \sum_j \sum_i I(j|i) - \sum_j \Sigma \sum_{i<k} I(j|i) I(j|k) + \sum_j \Sigma\Sigma \sum_{i<k<t} I(j|i) I(j|k) I(j|t) - \ldots + (-1)^{M-1} \sum_j I(j|1) I(j|2) \ldots I(j|M).$$

$$= \sum_i T_i - \Sigma \sum_{i<k} T_{i,k} + \Sigma\Sigma \sum_{i<k<t} T_{i,k,t} + \ldots$$

where T_i = Number of UUs induced by RU_i, $T_{i,k}$ = number of UUs simultaneously induced by both RU_i and RU_k, and so on.

Proof It is a matter of summing over all *j* for which *I*(*j*) = 1. The rest follows upon simplification.

Remark 4: To find an unbiased estimator for *N*, we now go for term by term estimation. If the sample size is $m < M$, term-by-term estimation will terminate after *m* terms. So we need a condition on the nature of the population network : No Ultimate unit is incident on more than *m* Reference Units. Our choice of the sample size *m* for random selection of RUs must be determined by this pre-condition. In applications, this is not hard to meet with by an appropriate choice of *m*. This point is already mentioned above. Here we have reemphasized it.

Those of us familiar with the unified theory of sampling and inference will find it a routine exercise to go for term-by-term unbiased estimation by using what are called 'joint inclusion probabilities'. When dealing with a term involving exactly one RU, say RU_i, we will consider dividing by π_i; for two RUs, say RU_i and RU_k, we will divide by $\pi_{i,k}$ and so on. So technically, we also need one necessary condition and this is to ensure positivity of all joint inclusion probabilities, upto and including the m^{th} order ! Note that sampling is taking place involving only the RUs and hence all these joint inclusion probabilities refer to the *M* RUs only. In order that all *m*-ples involving the RUs do possess positive joint inclusion probabilities, it is usually assumed that we are undertaking *SRSWOR*(*M, m*) sampling scheme on the population of *M* RUs ! Of course, there are many other choices and also we have varying probability sampling schemes, such as Midzuno Sampling Schemes, for example. We refer to Hedayat and Sinha (1991) for this aspect of sampling and inference.

Under *SRSWOR*(*M*, *m*) sampling scheme, we know

$\pi_i = m/M$; $\pi_{i,k} = m(m-1)/M(M-1) = m^{(2)}/M^{(2)}$; $\pi_{i,k,t} = m(m-1)(m-2)/M(M-1)(M-2) = m^{(3)}/M^{(3)}$; . . .

Under *SRSWOR*(*M*, *m*) sampling scheme, an expression for $\hat{N}$ can now be deduced slowly and carefully. We refer to Sinha *et al.* (2006) for this task and simply reproduce the final expression here.

$$\hat{N} = \sum_f n(f) A(f);$$

$A(f) = 1 - (-1)^f \, [(M-m+f-1)!(m-f)!]/[(M-m-1)!m!]$, $f = 1, 2, \ldots$.

where $n(f)$ = number of UUs, each with incidence number *f* in the sample network on *m* RUs and the induced UUs. The underlying simplification is not at all obvious. The students / researchers have to work around the formula quite carefully and with much attention.

Example 5 In the example that we have considered above, we have $M = 5$, $m = 3$ and [RU_1, RU_3, RU_4] are the sampled RUs and these altogether lead to a total of 5 induced UUs. Their frequencies are again reproduced below—without probing.

(1, 2, 2, 1, 1)

We find

$f = 1$, $n(1) = 3$, $A(1) = 1 + 2!2!/3! = 5/3$;

$f = 2$, $n(2) = 2$, $A(2) = 1 - 3!1!/1!3! = 1 - 1 = 0$.

Therefore,

$$\hat{N} = \sum_f n(f) A(f) = 3 \times 5/3 - 2 \times 0 = 5.$$

Remark 5 Sinha *et al.*(2006) considered different parameter values of *M* and *N* and an underlying network as a Bipartite Di-Graph. Next a value of *m* was taken and *SRSWOR*(*M, m*) was adopted. For all possible samples, the estimated value of $\hat{N}$ was computed, according to the formula shown above. It was hopelessly observed that for some samples, negative values for the estimate of unknown size (*N*) were realized and for some other samples, abnormally low values for the estimate were achieved !

There are primarily two reasons for this. We are taking recourse to indirect viz., network sampling via the selection and use of RUs. And, further, we require the estimator to be unbiased. Sinha and Padmawar (2015) have made a systematic and detailed study of such conseqences. They considered an example of a population network involving $M = 6$, $m = 3$) and computed numerical values of all possible estimators of N based on a totality of 20 possible samples, each of size 3 under $SRSWOR(6, 3)$ sampling. They observed that though the estimator is unbiased, 6 of the estimators were low/abnormally low. As against the actual value of $N = 10$, there are estimates as low as 1, 2 ! One can also end up with negative values !!

These are realities while we are dealing with unbiased estimation of N . That is why, Sinha and Padmawar (2015) made a systematic study of the problem in considerable details - with the whole objective of generating meaningful estimates. We will not enter into details of the heuristic principle developed by them.

ACKNOWLEDGEMENTS

The authors thank the anonymous reviewer for a careful reading of the manuscript, and for suggesting a related book as reference.

REFERENCES

Bandyopadhyay, S., Rao, A.R. and Sinha, Bikas K. (2011). Models for Social Networks with Statistical Applications. Sage Publications.

Boswell, M.T., Burnham, K.P., Patil, G.P. (1988). Role and use of composite sampling and capture-recapture sampling in ecological studies, Handbook of Statistics, Vol. **6**, Eds. P.R. Krishnaiah and C.R. Rao, North Holland, Amsterdam, 469-488.

Chaudhuri, Arijit (2015) Network and Adaptive Sampling. CRC Press, Taylor & Francis Boca Raton, Florida, USA.

Darroch, J.N. (1958). The multiple recapture census: I Estimation of a closed population, Biometrika, **45**, 343-359.

De, M. and Sengupta, S. (2003). Bernoulli Sequential Estimation of the size of a finite population. Sequential Analysis, **22**, 95-106.

Goodman, L.A. (1953). Sequential sampling tagging for population size problems. Ann. Math. Statist., **24**, 56-69.

Hedayat, A.S. and Sinha, Bikas K. (1991). Design and Inference in Finite Population Sampling. Wiley, New York.

Kiranandan, S.S. (1976). Imperfect Frames in Sample Surveys. Unpublished Ph.D. Thesis,

Harvard University.

Maiti, P., Pal, M. and Sinha, Bikas K. (1993). Estimating Unknown Dimension(s) of a Binary Matrix with Application to Estimation of the Size of a Mobile Population. In ”Probability and Statistics”, Eds. S.K. Basu and Bimal K. Sinha, Narosa Publishing House, New Delhi, 220-233.

Seber, G.A.F. (1982). The Estimation of Animal Abundance and Related Parameters, 2nd Edition, Macmillan, New York.

Sengupta, S. and De, M. (1997). On the estimation of size of a finite population. Sankhya, Series B, **9**, 66 -75.

Sinha, Bikas K., Maiti, P. and Sengupta, S. (2006). Unbiased Estimation in a Finite Population: Case of Multiple Indirect Identiability of Population Units. J. Stat. Theory and Applications. **5**(1), 81 - 90.

Sinha, Bikas K. and Padmawar, V.R. (2015). Estimating the size of a population comprising of individuals with indirect accessibility. Jour. Indian Stat Assoc. **53**(1-2), 175-200.