Special Proceedings: ISBN #: 978-81-950383-8-1
27th Annual Conference, 21-23 February 2025; pp 31-41

# Fondly Remembering Two Interesting Collaborations

**A. Goswami**
*School of Mathematical and Computational Science*
*Indian Association for the Cultivation of Science, Jadavpur, Kolkata, India*

Received: 07 June 2025; Revised: 17 June 2025; Accepted: 19 June 2025

**Abstract**

This article is based on the talk delivered by the author at the 27th Annual Conference of the Society of Statistics, Computer and Applicationns, held at NEHU, Shillong in February 2025. The talk was the fourth in the prestigeous Bikas K. Sinha Endowment Lecture Series. It is a fond recollection of the exciting journey of active collaboration with Professor Sinha on two separate occasions. These were on problems being investigated by Professor Sinha and the author was called on by him and was presented with the wonderful opportunity of collaborating with him.

*Key words:* Biodiversity; Species richness and species abundance; Time till full discovery; Social network; Measure of reciprocity; Sequential sampling stategy.

**AMS Subject Classifications:** 60G40, 62L12.

## 1. Introduction

### 1.1. A prologue from the author

It was around 1979-80 that I first met Professor Bikas K. Sinha or "*Bikas-da*" (as Professor Bikas K. Sinha is most popularly known to almost all of his younger colleagues, friends and also a large majority of his ex-students). I was a first year research scholar (Ph.D. student) then in the Stat-Math division at I.S.I., Kolkata, and attended, along with a few others, a research course on Optimal Designs taught by Bikas-da. Had I not left for USA next year to pursue my PhD, there was a distinct possibility that I could have earned the honour of being the first PhD student of Bikasda in ISI. Anyway, it was a little over eight years later when I retuned to ISI to join as a faculty in Stat-Math, that my contact with Bikasda had a restart. And, from then on till today, Bikasda remained my friend, my elder brother, my long time colleague. What I am going to discuss in this article are two interesting problems that we collaborated on. These were problems that originally Bikasda was investigating and at some point, he decided to discuss those with me and get me on board. This is another fascinating thing about Bikasda that when he discusses a problem with someone else, he can instill genuine interest in that person about the problem and, more

Corresponding Author: A.Goswami
Email: alok.goswami@iacs.res.in

often than not, it culminates into a productive collaboration. In fact, I do not personally know of anyone who has had as many collaborators as Bikasda has worked with, and that too from widely different fields of expertise, from pure mathematics to applied statistics, from sociology to biology. After having seen and known Bikasda for a very long time, first as his student and then as his younger colleague, I can only repeat what anyone and everyone, who has come in touch with him in any capacity, will say. He is one of the finest human beings that I have known, always jovial, always in high spirits, always with that reassuring smile on his face that makes a very bad day seem not so bad. During our joint tenure in ISI for a long period of time, there have been many occasions when Bikasda and I travelled together to multiple places to participate in a variety of academic programmes. Each of these occasions presented a wonderful opportunity for me to know Bikasda more closely, outside the framework of our relationship in ISI. They just reinforced my impression of him as an amazingly fascinating person with an unstoppable liveliness that always spread like an infection. Of all these journeys, there was one that deserves a special mention. Sometime during 2017, one day Bikasda asked me casually (at least that's how it seemed to me) '*ki bhaiya, bangladesh jabe naki?*', meaning "brother, interested in a trip to Bangladesh?" I was, of course, very excited and replied with an emphatic 'Yes'. Sure enough, a trip was organized for September that year. The two of us were invited to deliver lectures, first at Rajshahi University for 3-4 days and then at East West University in Dhaka for another 3-4 days. The entire trip was wonderful. The exciting academic interactions with young masters' and PhD students, the fascinating traditional Bangladeshi hospitality, the lazy evenings spent strolling along the serene banks of the river *Padma*, stopping occasionally to enjoy a sip of '*Morich Cha*' (finely brewed tea garnished with a pod of split green pepper to make it just a bit fiery) — all of these made it one of the most enjoyable trips of my life. But the part of the trip that will always remain etched in my memory, happened during our road trip from Rajshahi to Dhaka. We took a slight detour to visit Bikasda's ancestral village. Riding through that village with Bikasda recalling his childhood memories, lots and lots of magical stories about the young twin brothers$\cdots\cdots$I really had a glimpse into a different side of Bikasda that not too many persons I know had perhaps seen$\cdots\cdots$visualizing the twin brothers strolling and running through the muddy roads of the village, occasionally indulging in small mischiefs for fun, that is typical of that age$\cdots\cdots$it was a total surprise that I didn't expect. And Bikasda being Bikasda hadn't informed me in advance of this plan. We finally stopped at his ancestral home where members of their family who stayed back still live. Needless to say we were offered an elaborate lunch, with multiple fish dishes in abundance, before we hit the road again on our way to Dhaka. The detour turned out to be an absolutely memorable surprise indeed! Thank you Bikasda for gifting me this magical experience!!!

### 1.2. Two interesting problems

The first problem we discuss here falls in the domain of biodiversity analysis. Two quantities that are regarded as central to the measurement of biodiversity are *species richness* and *species abundance*. In the context of developing appropriate sampling strategies to gain understanding of species richness and species evenness, a conjecture, supported also by empirical observations, that has inherently played a crucial role, is that the species evenness distribution which allows for a minimal sample size is the one, in which, for a fixed size of species richness, the abundance rates are all equal [see Gore and Paranjape (1997), (2001)].

This was our focus of study in one of the two collaborations mentioned above and is discussed in greater detail in Section 2. The main objective of our work was to try and provide an analytical proof of thisd conjecture and, luckily, we succeeded. For a fixed species richness size, we considered the distribution of "effort size" for full discovery, as a function of the underlying species abundance rate vector, and showed analytically that, it is stochastiaclly smallest when all the species are "as equally abundant as possible", a phrase that is going to be clarified in Section 2. The analysis is done separately for the two cases, namely, for infinite population and for finite population. The mathematical formulation of the problem and sketches of the proof of the main results are given in Section 2. The details are skipped here, since complete details are available in Goswami and Sinha (2006).

The other problem discussed in this article is in the context of social networks. Concept of *Reciprocity* in a social network is recognized as an important characteristic for study in order to gain understanding of the network. Rao and Bandyopadhayay (1987) suggested a simple natural measure of Reciprocity. Our focus in this work of collaboration was the choice of optimal sampling strategies for unbiased estimation of the above measure of reciprocity. Being a diehard champion of optimality, Bikasda had a firm conviction that, in this particular context, with some of its inherent special characteristics, the standard (SRSWOR, sample mean) strategy cannot be the last word. The strength of our collaboration, with Professor S. Sengupta from Calcutta University also joining the team, was rooted in that conviction. Finally (and despite challenges thrown by some rather complicated algebra), our work was able to illustrate that certain naturally suggested sequential strategies using specific selection rules coupled with appropriately defined unbiased estimators, fare better than the usual (SRSWOR, sample mean) strategy. Outlines and some crucial steps behind the core ideas are discussed in some detail in Section 3. For complete details and more (including numerical illustrations), the reader may see Goswami, Sinha and Sengupta (1990).

## 2. Analyzing time till discovery of all species

We consider a population consisting of $m$ different species. The number $m$, called the *species richness*, is assumed to be fixed in this analysis. Clearly, only $m \geq 2$ is of any interest. The variable parameter in our analysis is going to be the *abundance rate vector*

$$\mathbf{p} = (p_1, \cdots, p_m), \tag{1}$$

where $p_i$, the abundance rate of the $i$th species, is its proportion in the population (equivalently, the probability that a randomly drawn unit from the population is from $i$th species).

In case of a finite population, say, of known size $N$, the admissible abundance rate vectors will have to be necessarily of the form

$$\mathbf{p} = \left(\tfrac{N_1}{N}, \cdots, \tfrac{N_m}{N}\right), \tag{2}$$

where $N_1, \cdots, N_m$ are positive integers adding up to $N$, while for an infinite population, any set of numbers $p_i \in (0, 1),\ i = 1, \cdots, m$, that add up to 1, will constitute a possible abundance rate vector $\mathbf{p} = (p_1, \cdots, p_m)$.

The focus and aim of our investigation here are as follows. Suppose we keep on drawing units at random from the population until all the $m$ species are "discovered". Denoting $T$

to be the number of draws needed, it is clear that $T$ is a random variable whose distribution depends on the abundance rate vector $\mathbf{p}$, and also on the sampling scheme − specifically, in case of a finite population, whether the sampling is done with or without replacement. We state the main result that we are able to prove.

**Main result**: *In both cases of random sampling from an infinite population as well as random sampling, with or without replacement, from a finite population, the random variable $T$ is "stochastically smallest" when all the $m$ species are "almost" equally abundant.*

In case of any concern about the phrase "almost" equally abundant used above, we want to point out that this is relevant only in finite population case and will be elaborated in Section 2.2. We proceed now to present an outline of the steps through which the above result is proved. The cases of infinite population and finite population need to be handled separately and that is what we do. But at the start, let us state a simple result which will be used frequently in the sequel. Noting that relabelling the $m$ species among themselves does not have any impact on the time $T$ till full discovery, the following result is obvious.

**Theorem 1:** The distribution of $T$, under an abundance rate vector $\mathbf{p}$, remains invariant over permutations of the coordinates of the vector $\mathbf{p}$.

### 2.1. Infinite population

In this case, we do not need to distinguish between sampling with or without replacement. As mentioned before, any choice of numbers $p_i \in (0,1)$, $i = 1, \cdots, m$, with $\sum p_i = 1$, will constitute a possible abundance rate vector $\mathbf{p} = (p_1, \cdots, p_m)$. Let us denote

$$\Phi(t, m, \mathbf{p}) = P(T > t \,\big|\, m, \mathbf{p}), \text{ for each } t. \tag{3}$$

The abundance rate vector capturing the "equally abundant" case is denoted $\mathbf{p}_0$, that is. $\mathbf{p}_0 = (\frac{1}{m}, \cdots, \frac{1}{m})$. With these notations, here is our main result.

**Theorem 2:** For an infinite population with $m$ species, where $m \geq 2$,

$$\Phi(t, m, \mathbf{p}) \geq \Phi(t, m, \mathbf{p}_0), \quad \text{for all } t \geq m, \tag{4}$$

with the inequality in (4) being strict, for every $t \geq m$, unless $\mathbf{p} = \mathbf{p}_0$.

**Proof:** (outline) The case $m = 2$ is fairly trivial. For any $\mathbf{p} = (p_1, p_2)$, one has $\Phi(t, 2, \mathbf{p}) = p_1^t + p_2^t$, for all $t \geq 2$, and the right-hand-side, for each $t \geq 2$, can easily be shown to have a unique minimum at $p_1 = p_2 = \frac{1}{2}$, subject to the conditions that $p_1 > 0, p_2 > 0$, $p_1 + p_2 = 1$.

A natural idea now is to complete the proof by induction, but a first step towards that would be to get some relation between $\phi(\cdot, m, \cdot)$ and $\phi(\cdot, m-1, \cdot)$. To do this, we need a notation. For $m > 2$ and for any abundance rate vector $\mathbf{p} = (p_1, \cdots, p_m)$, let us denote $\mathbf{p}^{(i)}$, for $1 \leq i \leq m$, to be the abundance rate vector of size $m - 1$, obtained by removing the $i$th coordinate from $\mathbf{p}$ and normalizing the remaining coordinates, that is, $\mathbf{p}^{(i)} = \left(\frac{p_1}{1-p_i}, \cdots, \frac{p_{i-1}}{1-p_i}, \frac{p_{i+1}}{1-p_i}, \cdots, \frac{p_m}{1-p_i}\right)$. By a fairly straightforward conditioning argument, one can now show that, for any $m > 2$, any $\mathbf{p} = (p_1, \cdots, p_m)$ and any $i \in \{1, \ldots, m\}$,

$$\Phi(t, m, \mathbf{p}) = b(t, p_i; 0) + \sum_{s \geq t-m+2} b(t, p_i; s) + \sum_{s=1}^{t-m+1} b(t, p_i; s)\, \Phi(t-s, m-1, \mathbf{p}^{(i)}), \quad \text{for all } t \geq m, \tag{5}$$

where $b(t, p_i; s)$, $0 \le s \le t$, are the p.m.f. of the Binomial distribution with parameters $(t, p_i)$.

This paves the way for invoking induction. Assuming the assertion (4) to hold for $m-1$ will mean that $\Phi(t', m-1, \mathbf{p}^{(i)}) \ge \Phi\Big(t', m-1, (\frac{1}{m-1}, \cdots, \frac{1}{m-1})\Big)$, for all $t' \ge m-1$, with strict inequality unless $\mathbf{p}^{(i)} = \Big(\frac{1}{m-1}, \cdots, \frac{1}{m-1}\Big)$. Using this in (5) and recombining the terms, one easily gets that, for all $m \ge 3$, all $\mathbf{p} = (p_1, \cdots, p_m)$ and all $i \in \{1, \cdots, m\}$,

$$\Phi(t, m, \mathbf{p}) \ge \Phi\Big(t, m, (\, \tfrac{1-p_i}{m-1}, \cdots, \tfrac{1-p_i}{m-1},\, p_i,\, \tfrac{1-p_i}{m-1}, \cdots, \tfrac{1-p_i}{m-1}\,)\Big), \quad \text{for all } t \ge m, \tag{6}$$

with strict inequality unless the $p_j, j \ne i$ are all equal.

Now, if $\mathbf{p} \ne \mathbf{p}_0$, then there must exist $i$ such that $p_j, j \ne i$ are not all equal. Recalling that $\Phi(t, m, \mathbf{p})$ is invariant under permutations of coordinates of $\mathbf{p}$, we can, without loss of generality, assume $i = 1$. Writing $p_1 = p$, we then have

$$\Phi(t, m, \mathbf{p}) > \Phi\Big(t, m, (p,\, \tfrac{1-p}{m-1}, \cdots, \tfrac{1-p}{m-1}\,)\Big), \quad \text{for all } t \ge m, \tag{7}$$

Denoting the right-hand-side as $g(p)$, our final job is to show that the function $g$ on $(0, 1)$ has a unique global minimum at $p = 1/m$ (for all $t \ge m$!). Having thus reduced the problem to one of ascertaining the global minimum of a function of a single variable, one would like to believe that the job should now be easily completed through simple calculus. Unfortunately, that is not to be! The second derivative $g''$ seems intractable. Overcoming this hurdle required a completely different idea. Denoting $f(p) = \frac{1-p}{m-1}$, for $p \in (0, 1)$, the inequality (6) can be paraphrased as saying that, for all $m \ge 3$, all $\mathbf{p} = (p_1, \cdots, p_m)$ and all $i \in \{1, \cdots, m\}$,

$$\Phi(t, m, \mathbf{p}) \ge \Phi\Big(t, m, (f(p_i), \cdots, f(p_i),\, p_i,\, f(p_i), \cdots, f(p_i)\,)\Big), \quad \text{for all } t \ge m, \tag{8}$$

with strict inequality unless the $p_j, j \ne i$ are all equal. Now, if $\mathbf{p} \ne \mathbf{p}_0$, then, as argued above, we may assume that $p_j, j \ne 1$ are not all equal, so that, by denoting $p_1 = p$, we get

$$\Phi(t, m, \mathbf{p}) > \Phi\Big(t, m, (p, f(p), \cdots, f(p))\Big), \quad \text{for all } t \ge m. \tag{9}$$

Now applying inequality (8) with $i = 2$ on the right-hand-side of (9) and then using invariance of $\Phi$ under permutation of coordinates of abundance rate vector will give

$$\Phi\Big(t, m, (p, f(p), \cdots, f(p))\Big) \ge \Phi\Big(t, m, (f(p), f(f(p)), \cdots, f(f(p)))\Big), \quad \text{for all } t \ge m. \tag{10}$$

Repeating this and denoting $f^{(n)}$ to be the $n$th iterate of $f$, one gets the string of inequaities

$$\begin{aligned}\Phi\Big(t, m, (p, f(p), \cdots, f(p))\Big) &\ge \Phi\Big(t, m, (f(p), f(f(p)), \cdots, f(f(p)))\Big) \\ &\ge \cdots\ \cdots\ \cdots\ \cdots \\ &\ge \Phi\Big(t, m, (f^{(n-1)}(p), f^{(n)}(p), \cdots, f^{(n)}(p))\Big) \\ &\ge \cdots\ \cdots\ \cdots\ \cdots\end{aligned}$$

The job is now completed by invoking the Lemma stated below, the proof of which is fairly easy and hence omitted. Using the Lemma and the fact that $\Phi\Big(t, m, (p_1, \cdots, p_m)\Big)$, for every $t \ge m$, is continuous in $p_1, \cdots, p_m$ (an easy exercise!), one can easily put together (9) and the above string of inequalities to complete the proof of assertion (4). □

**Lemma 1:** The function $f(p) = \frac{1-p}{m-1}$, $p \in [0, 1]$, has a unique fixed point at $p_0 = \frac{1}{m}$. Further, $p_0$ is a globally attracting fixed point in the sense that $\lim_n f^{(n)}(p) = p_0$, for all $p \in [0, 1]$.

### 2.2. Finite population

Handling the finite population case follows the same central ideas as in the infinite population case, but, not surprisingly, the actual execution gets somewhat complicated. Recall that in case the population size is finite, say, $N$, the abundance rate vectors are necessarily of the form $\mathbf{p} = \left(\frac{N_1}{N}, \cdots, \frac{N_m}{N}\right)$, where the $N_i$ are positive integers adding up to $N$. If we assume $N$ to be fixed, then the actual "abundance vector" $\mathbf{N} = (N_1, \cdots, N_m)$ can be regarded as an equivalent representation of the underlying parameter, in place of the rate vector $\mathbf{p}$. We will follow this viewpoint and formulate everything in terms of the abundance vector $\mathbf{N}$ as the parameter. Using notations similar to the infinite population case, we denote $\Phi(t, N, m, \mathbf{N})$ to be the probability that for a population of size $N$ with abundance vector $\mathbf{N}$ for $m$ species, more than $t$ draws are needed to discover all the species. This probability, of course, depends on whether the draws are with or without replacement, but we will use the same notation. We ask the same question again, namely, is there a special abundance vector that minimizes $\Phi(t, N, m, \mathbf{N})$, uniformly over all $t \geq m$? The answer we get is 'Yes'; there is a special abundance vector $\mathbf{N}_0$, representing "*almost equal*" abundance for the $m$ species, that does the job, irrespective of whether draws are made with or without replacement.

Let us first explain what we mean by "*almost equal*" abundance and why we need this. Note that all the $m$ species being *exactly* equally abundant would mean that the $N_i$'s must all be equal, which, of course, can happen *only if* the population size $N$ is a multple of $m$. That will certainly be an undesirable restriction on $N$. So, the right thing to do would be to get, for any given $N$, as close as possible to equal abundance and that is what is represented by the special vector $\mathbf{N}_0$, which is formally defined below.

For any abundance vector $\mathbf{N} = (N_1, \cdots, N_m)$, let $d(\mathbf{N}) = \max\{|N_i - N_j| : i \neq j\}$. It is obvious that $d(\mathbf{N})$ is invariant under permutations of coordinates of $\mathbf{N}$ (as is $\Phi(t, N, m, \mathbf{N})$). Also, the larger $d(\mathbf{N})$ is, the farther is $\mathbf{N}$ from "equal abundance". A little reflection will convince the reader that the minimum value of $d(\mathbf{N})$ equals 0 only if $N$ is a multiple of $m$ and is attained by $\mathbf{N}_0 = (\frac{N}{m}, \cdots, \frac{N}{m})$; otherwise, the minimum value is 1, attained by the unique (upto permutations) abundance vector of $\mathbf{N}_0 = \Big(\underbrace{N_0 + 1, \cdots, N_0 + 1}_{k}, \underbrace{N_0, \cdots, N_0}_{m-k}\Big)$, where $N = mN_0 + k,\ 0 < k < m$. In other words, $\mathbf{N}_0$ is always the unique (upto permutations) abundance vector satisfying $d(\mathbf{N}_0) \leq 1$. It is clear that $\mathbf{N}_0$ represents, for any given $N$, an admissible abundance vector where the $m$ species are as equally abundant as possible. Here is our main result for the finite population case.

**Theorem 3:** For a finite population of size $N$, consisting of $m$ $(2 \leq m \leq N)$ different species,

$$\Phi(t, N, m, \mathbf{N}) \geq \Phi(t, N, m, \mathbf{N}_0), \quad \text{for all } t \geq m, \tag{11}$$

irrespective of whether units are drawn with or without replacement. Further, the inequality in (11) is strict (for all $t$, for which the left-hand-side is positive), unless $\mathbf{N} = \mathbf{N}_0$ (upto permutations).

As mentioned earlier, the main ideas of the proof are the same as those in case of infinite population, but the execution of those ideas are far more complicated. That should not be surprising because, in the infinite population case, we were dealing with continuous

variables $p_i$, whereas now we are handling the problem of minimizing a function of positive integer variables $N_i$. The difficulty gets a bit more multiplied by the fact that we also have to treat the cases of sampling WR and WOR differently. Since the details are available in Goswami and Sinha (2006), we just briefly outline the steps here.

As before, the idea is to first prove it for $m = 2$ and then use induction on $m$. For $m = 2$, we can take any abundance vector $(N_1, N_2)$, where (without loss of generality) $N_1 \leq N_2$, and write down explicit formulas for $\Phi\Big(t, N, 2, (N_1, N_2)\Big)$ in both WR and WOR cases. Now, if $N_2 - N_1 \geq 2$, one can, with some work, show that

$$\Phi\Big(t, N, 2, (N_1, N_2)\Big) > \Phi\Big(t, N, 2, (N_1 + 1, N_2 - 1)\Big), \quad \text{for all } t \geq 2.$$

It is now just a matter of repeating this inequality over and over again to finally get (11) for $m = 2$. To proceed with induction now, we need, as before, a formula relating $\Phi(\cdot, \cdot, m, \cdot)$ to $\Phi(\cdot, \cdot, m-1, \cdot)$. This can again be obtained by using a similar kind of conditioning as before. Indeed, for any $m \geq 3$, $N \geq m$ and any abundance vector $\mathbf{N} = (N_1, \cdots, N_m)$, one can get $\Phi(t, N, m, \mathbf{N})$ to be a weighted average of 1 and $\Phi(t-s, N-N_i, m-1, \mathbf{n}_{(i)})$, $1 \leq s \leq t-m+1$, where the weights are p.m.f.s of an appropriate Binomial distribution or Hypergeometric distribution, according as the draws are WR or WOR. Here, $\mathbf{N}_{(i)}$, for any $i$, is the abundance vector of size $(m-1)$, obtained by just removing $N_i$ from $\mathbf{N}$. Using the induction hypothesis now, one can get an inequality analogous to (6), namely,

$$\Phi(t, N, m, \mathbf{N}) \geq \Phi(t, N, m, \mathbf{N}^{(i)}), \quad \text{for all } t \geq m \text{ and each } 1 \leq i \leq m, \tag{12}$$

where $\mathbf{N}^{(i)}$ is the abundance vector whose $i$th coordinate is the same as that of $\mathbf{N}$, that is, $N_i$, while the rest of the coordinates are those of an abundance vector of size $(m-1)$, adding up to $N - N_i$, that represents "as equal abundance as possible" for $m-1$ species in a population of size $N - N_i$. An explicit description is given in Goswami and Sinha (2006). Further, induction hypothesis will also imply strict inequality in (12), unless $d(\mathbf{N}_{(i)}) \leq 1$.

Now comes the final step. Recall that, the final step in the infinite population case presented a little hurdle and we needed to use a result like Lemma 1 to complete the proof. Here also, the job is far from over. The passage from (12) to completion of the proof of Therem 3, poses a significant challenge. Luckily, with some effort, we were able to formulate and prove a result that helped us cross this last hard mile. This result played a role very analogous to that played by Lemma 1 in Section 2.1. Unfortunately (but perhaps not surprisingly), neither the statement of the result nor its proof are as straightforward as Lemma 1. For the sake of brevity of this article, we refrain from stating the result here and also skip the details of how the lemma is used to complete the proof of Theorem 3. An interested reader will find all the details in Goswami and Sinha (2006).

## 3. Sampling from a social network: optimal strategies

A social network is a population equipped with an irreflexive binary relation. Denoting the binary relation by $\rightsquigarrow$, we say that two distinct units $i$ and $j$ have some "*tie*" if either $i \rightsquigarrow j$ or $j \rightsquigarrow i$. We say that there is a "*symmetric tie*" or a "*reciprocal tie*" between units $i$ and $j$ if both $i \rightsquigarrow j$ and $j \rightsquigarrow i$ hold. Given a social network, one among several quantities of interest for sociologists, is the extent of reciprocity in the network. Many different measures

have been proposed for this, among which we consider here the simplest one, proposed by Rao and Bandyopadhayay (1987). It is defined as

$$\overline{\theta} = \frac{1}{N(N-1)} \sum\sum_{\substack{i,j \in \Omega \\ i \neq j}} \theta(i,j), \tag{13}$$

where $\Omega$ denotes the population, $N$ the size of the population and $\theta(i,j)$ is defined to be 1 if both $i \rightsquigarrow j$ and $j \rightsquigarrow i$ hold and defined to be 0 otherwise.

The problem that we investigated in Goswami, Sinha and Sengupta (1990) is that of estimating the population parameter $\overline{\theta}$ unbiasedly on the basis of "data available" from a sample of size $n$ drawn from the population. Note that if SRSWOR is used to draw a sample $s$ of size $n$, then one can easily see that the usual sample mean

$$\hat{\overline{\theta}}(s) = \frac{1}{n(n-1)} \sum\sum_{\substack{i,j \in s \\ i \neq j}} \theta(i,j) \tag{14}$$

is an unbiased estimator of $\overline{\theta}$. The question we asked is whether one can do better (in the sense of reducing the variance) by adapting an appropriate sequential sampling scheme, coupled with an appropriate unbiased estimator. To keep the question alive, it is very important to turn our attention to a clear understanding of "data available from the sample".

To clarify our point, let us denote $\overline{\theta}_i$, for each $i \in \Omega$, to be the average number of symmetric ties in which $i$ is involved, that is, $\overline{\theta}_i = \frac{1}{N-1} \sum\limits_j \theta(i,j)$. Clearly then, $\overline{\theta}$ is the population mean of the $\overline{\theta}_i$ and therefore, if one stipulates that from a sample drawn from the population, the value of $\overline{\theta}_i$ will be "available" for each sample unit $i$, then our question stops there. This is because, in that case, the admissibility of the (SRSWOR, sample mean) strategy is a classical result. However, a large number of practitioners in this field are strongly opposed to the stipulation that $\overline{\theta}_i$ for sample units $i$ are "observable". So, let us place here the wide consensus on what is "observable" and what is not.

For each population unit $i$, the "out-set" and "in-set" of $i$ are defined respectively as $\mathcal{O}(i) = \{j : i \rightsquigarrow j\}$ and $\mathcal{I}(i) = \{j : j \rightsquigarrow i\}$, with their cardinalities, denoted by $d_i$ and $e_i$ respectively, being called the "out-degree" and "in-degree" of $i$. It then follows that, $\overline{\theta}_i = \frac{1}{N-1} \Big| \mathcal{O}(i) \cap \mathcal{I}(i) \Big|$, for any $i \in \Omega$, and so information on both $\mathcal{O}(i)$ and $\mathcal{I}(i)$ is required to know $\overline{\theta}_i$. The widely held opinion of practitioners is that, while somewhat reliable information on $\mathcal{O}(i)$ may be available from a sample unit $i$, but information on $\mathcal{I}(i)$ is highly unreliable.

We accept this premise, namely, that the only information "available" from a sample unit $i$, that is reliable and useful, is its out-set $\mathcal{O}(i)$. This would mean that the admissibility of (SRSWOR, sample mean) strategy is no longer guaranteed and therefore, our search for a better strategy becomes valid and meaningful.

Indeed, what we were able to achieve in Goswami, Sinha and Sengupta (1990) is to stitch up an alternative strategy that performs **uniformly better** than the (SRSWOR, sample mean) strategy. The main idea behind our proposed strategy essentially orginated from a detailed examination of the case with sample size $n = 2$. So, we are going to describe

that special case in detail here and follow it up by just giving an outline of the scheme for general sample size $n$.

Just to keep things simple, we make an additional assumption that $d_i > 0$, for each $i$. We assure the reader that this is not at all indispensable and can be easily done away with. From the underlying stipulation that information on $\mathcal{O}(i)$ is available from each sample unit $i$, it is easy to see that the value of $\theta(i,j)$ will be known for each **pair** of units $(i,j)$ in the sample. Now, in case of sample size $n = 2$, the unbiased estimator that the (SRSWOR, sample mean) strategy proposes boils down to

$$\hat{\bar{\theta}}_1(s) = \theta(i,j), \quad \text{if } s = \{i,j\}. \tag{15}$$

What we propose in our alternative strategy is the estimator

$$\hat{\bar{\theta}}_2(s) = \frac{d_i}{N-1}\theta(i,j), \quad \text{if } s \text{ is the } \textbf{ORDERED} \text{ sample } (i,j). \tag{16}$$

It is fairly easy to simplify $E\Big(\hat{\bar{\theta}}_2\Big) = \sum\limits_{\substack{i,j\in\Omega \\ i\neq j}}\sum P\Big(s=(i,j)\Big)\frac{d_i}{N-1}\theta(i,j)$ and hence show that $\hat{\bar{\theta}}_2$ is unbiased. What is really important is that one can do a little computation to get the second moments of the unbiased estimators (15) and (16) and deduce that

$$E\Big(\hat{\bar{\theta}}_1^2\Big) - E\Big(\hat{\bar{\theta}}_2^2\Big) = \frac{1}{N(N-1)}\sum_{\substack{i,j\in\Omega \\ i\neq j}}\sum\Big(1 - \frac{d_i}{N-1}\Big)\,\theta(i,j) \geq 0. \tag{17}$$

An immediate consequence of (17) is that $\hat{\bar{\theta}}_2$ performs uniformly better than $\hat{\bar{\theta}}_1$, in the sense of reducing variance. Further, the right-hand-side of (17) also shows that $\hat{\bar{\theta}}_2$ has **strictly smaller variance** except in the extreme case when $d_i = N-1$ for **all** $i$ with $\sum_j \theta(i,j) > 0$. Of course, noting that our proposed estimator $\hat{\bar{\theta}}_2$ is "order dependent", an initiated reader will immediately see an opportunity of further improving on it by using the classical idea of what is widely known as "*Blackwellization*". By "averaging over order", one gets an even more improved estimator given by

$$\hat{\bar{\theta}}_3(s) = \frac{2d_id_j}{(N-1)(d_i+d_j)}\,\theta(i,j), \quad \text{if } s = \{i,j\}. \tag{18}$$

Having thus described our improved strategy in detail for the case $n = 2$, it is quite natural now to try and extend this idea for a general sample size $n$. This (and much more) was indeed done and reported in detail in Goswami, Sinha and Sengupta (1990). In particular, we were able to exhibit a sequential strategy $(p_0, e_0)$, for any sample size $n$, that performs uniformly better than a size $n$ (SRSWOR, sample mean) strategy. An initiated reader would surely recall that a sampling stategy consists of a pair (sampling scheme, estimator). Our proposed sequential strategy $(p_0, e_0)$ is described below, from which it will be clear to the reader that it is a generalization of what was done for $n = 2$.

**Sequential strategy** $(p_0, e_0)$: For distinct population units $i_1, \cdots, i_k$, let us denote

$$\mathcal{O}(i_1, \cdots, i_k) = \Big(\mathcal{O}(i_1) \cup \cdots \cup \mathcal{O}(i_k)\Big) \setminus \{i_1, \cdots, i_k\} \text{ and } d(i_1, \cdots, i_k) = |\mathcal{O}(i_1, \cdots, i_k)|.$$

- $p_0$ : First draw a SRSWOR of size $(n-1)$, say, $\{i_1, \cdots, i_{n-1}\}$
- $p_0$ : Now draw a random unit, say, $i_n$ from $\mathcal{O}(i_1, \cdots, i_{n-1})$
- $e_0$ : For $s = \{i_1, \cdots, i_{n-1}; i_n\}$ define

$$e_0(s) = \frac{1}{n(n-1)} \sum_{\substack{1 \le k,l \le n-1 \\ k \ne l}} \sum \theta(i_k, i_l) + \frac{2d(i_1, \cdots, i_{n-1})}{n(n-1)(N-n+1)} \sum_{k=1}^{n-1} \theta(i_k, i_n) \qquad (19)$$

The following theorem captures one of our main results in the context of this investigation.

**Theorem 4:** For every $n \ge 2$, the sequential sampling strategy $(p_0, e_0)$ of sample size $n$ perfoms better than the size $n$ (SRSWOR, sample mean) strategy uniformly, in the sense of having smaller variance.

Did we bump into this sequential strategy by some magic or chance? The answer to that is an emphatic 'No'. In fact, what we were able to do is to describe in detail the string of main ideas that essentially leads one to not just this particular sequential strategy $(p_0, e_0)$, but to a whole class of possible sequential strategies, each of which beats the (SRSWOR, sample mean) strategy uniformly. The strategy $(p_0, e_0)$ is just a special case. The story behind the closed doors is that we did not arrive at our sequential strategies at one go. We did it in two steps. To briefly describe it, let us fix $n > 2$ and denote the (SRSWOR, sample mean) strategy of size $n$ by $(p, e)$. In our first step, we construct a variable sample size strategy $(p^*, e^*)$, with sample size varying between $n$ and $n-1$, which is equivalent to $(p, e)$, in the sense that $E_{p^*}(e^*) = E_p(e)$ and $E_{p^*}(e^{*2}) = E_p(e^2)$. Then, in our next step, we construct a sequential strategy $(p^{**}, e^{**})$ of sample size $n$ and show that it performs uniformly better than $(p^*, e^*)$ and hence uniformly better than $(p, e)$. The important point is that in this last step, we actually prescribe not just one sequential strategy but **a whole class** of possible sequential strategies $(p^{**}, e^{**})$ of sample size $n$, each of which performs uniformly better than $(p^*, e^*)$ (and hence, better than $(p, e)$). It will be too much to give the complete descriptions of all of these and the corresponding proofs here. For complete details, we refer to Goswami, Sinha and Sengupta (1990).

## 4. Some concluding remarks

Section 2.1 :

It is well-known that an arbitrary $\mathbf{p}$ vector is "majorized" by the vector $\mathbf{p}_0$ and hence the results on Schur concave functions will directly apply, provided one can establish Schur concavity of $\Phi(t, m, \mathbf{p})$ as a function of $\mathbf{p}$. This is worth exploring (see Marshall and Olkin (1979)). Another question that occured to the author while taking a fresh relook at the paper just before the talk, is that $\mathbf{p}_0$ is known to have the maximum Shannon entropy among all probability vectors $\mathbf{p}$ of size $m$. It is worth investigating whether that has any role to play. Cracking this may lead to formulating a large number of more general problems and getting interesting answers to those.

Section 2.2 :

Our proposed sequential strategy $(p_0, e_0)$ (and, more generally, $(p^{**}, e^{**})$) leaves several questions unanswered. Firstly, the estimator $e_0$ is order dependent and so, it leads one to possibly think of improving it further by using "Blackwellization" (that is, symmetrization). We tried it, but the variance of the symmetrized version seemed rather intractable. Another more significant limitation with $(p_0, e_0)$ is that the sampling design $p_0$ differs very little from SRSWOR, in the sense that the sequential nature of $p_0$ appears only while drawing the last unit. This would mean that the improvent wouldn't mean much when $n$ is large (as is evidenced by the expression for variance). So, a natural question is whether we can devise a strategy that brings in the sequential nature much earlier, thereby hoping to make more significant improvement over $(p, e)$. Some efforts in this direction were undertaken for sample size $n = 3$. The results we obtained are reported in Goswami, Sinha and Sengupta (1990). However, the issue remains wide open for general $n$ and is certainly worth pursuing.

**Acknowledgements**

I am indeed grateful to the Society of Statistics, Computer and Applications (SSCA) and, in particular, to Professor V.K.Gupta, President (SSCA), for inviting me to deliver the prestigeous Bikas K. Sinha Endowment Lecture during the 27th annual conference of the SSCA held in February 2025 at NEHU, Shillong. Many thanks to my teacher and friend '*Bimal-da*' (Professor Bimal K. Sinha) for proposing and endorsing my name for delivering this 4th lecture in the series. I would also like to thank the organizers, specially Professor Bishal Gurung and my good friend '*Tapan*' (Professor T. K. Chakrabarty), for their wonderful hospitality and efforts in ensuring smooth organization of this big event. Finally, I must again thank Professor V.K.Gupta, for encouraging me to submit this article based on my lecture, for possible inclusion in the Special Proceedings of the 27th Conference.

**References**

Gore, A. P. and Paranjape, S. A. (1997). Effort needed to measure biodiversity. *International Journal of Ecology and Environmental Sciences*, **23**, 173-183.

Gore, A. P. and Paranjape, S. A. (2001). *A Course in Mathematical and Statistical Ecology.* Kluwer, Budapest.

Goswami, A., Sinha, Bikas K., and Sengupta, S. (1990). Optimal strategies in sampling from a social network. *Sequential Analysis*, **9**, 1-18.

Goswami, A. and Sinha, Bikas K. (2006). Some probabilistic aspects in the discovery of species. *Sequential Analysis*, **25**, 103-115.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization with Applications.* Academic Press, New York.

Rao, A. R. and Bandyopadhyay, S. (1987). Measures of reciprocity in a social network. *Sankhya Series A*, **49**, 141-188.