Special Proceedings: ISBN#: 978-81-950383-2-9
25th Annual Conference, 15–17 February 2023; pp 1–10

On the History of Certain Early Developments in Sampling Theory

T. J. Rao
Retired Professor, Indian Statistical Institute, Calcutta

Received: 30 May 2023; Revised: 06 June 2023; Accepted: 08 June 2023

Abstract

In this paper based on the Bikas Sinha Endowment Lecture, we shall first discuss Rao-Blackwellization of some early estimators obtained in finite population sampling theory along with historical aspects. We note that there are certain lapses with respect to priorities and credits in the literature. Next, we shall briefly sketch the role played by Bikas Sinha related to applications during early days.

Key words: Rao-Blackwellization; Probability proportional to size sampling; Call backs; Randomized response technique; Environmental statistics.

AMS Subject Classification: 62D05

1. Introduction

In this paper, we begin with the technique of *Rao-Blackwellization* in finite population sampling theory, a subject which both Prof. Bikas Sinha (BKS) and I are interested in. Rao-Blackwellization, a term credited to C. R. Rao based on his 1945 'breakthrough' paper published in the *Bulletin of the Calcutta Mathematical Society*, provided improved estimators in conventional as well as adaptive, link-tracing, size-biased sampling techniques. Furthermore, Rao-Blackwellization found applications in statistics and a host of other disciplines including sports, namely Rao-Blackwellized Field Goal percentage estimator (RB-FG%) and possibly social networks such as WhatsApp (RB-WA).

We shall first discuss applications related to improving of estimators in finite population sampling theory relating to Probability Proportional to Size Without Replacement (PPSWOR) selection. We note that proper credit is not given to certain publications and point out certain lapses with respect to (wrt) priorities and credits in the sampling literature.

2. Selection with ppswor scheme

Following Basu (1958), wherein he showed that the 'order statistic' (sample units in ascending order of their labels) is a sufficient statistic, Pathak(1961) while discussing sampling from finite populations, noticed that *'any estimator which is not a function of the order statistic'*, can be uniformly improved upon by the use of Rao - Blackwellization technique. BKS along with Sen (Sinha and Sen,1989) goes beyond variance comparisons and generalizes to convex loss functions. In his book on *Finite Population Sampling* with Hedayat (1991), BKS devotes the maximum number of 58 pages for the chapter on PPS sampling. A large part

Corresponding Author: T.J. Rao
Email: tjrao7@gmail.com
This paper is based on the Bikas Kumar Sinha Endowment Lecture "My Significant Interaction with Bikas Sinha" delivered by the author on 15 February, 2023 during the conference.

of his work on sampling (solo and with co-authors) was on PPS sampling among others. We shall discuss the case of sampling of 2 units from a finite population of size N with the study variate Y taking values Y_i and known auxiliary variate X related to Y, taking values X_i on the units U_i, $i = 1,2,....,N$. Let T_Y and T_X denote the population totals of Y and X respectively.

Let $P_i = X_i / T_X$.

2.1. A recap of unbiased estimators of T_Y

International Statistical Institute held its biennial session in Delhi in 1951 from 5-11 December. A short session was held in Calcutta from 16 to 18 December along with other international societies. A. C. Das of Indian Statistical Institute presented a paper on successive sampling. As a passing note in this paper, Das (1951) discussed PPSWOR scheme as well at the end.

Thus, if (i, j) are the labels of units selected by PPSWOR in that order, then Das's (1951) estimators for the first and second draws of sample selection are, respectively,

$$t_{1\,Da} = y_i / p_i \tag{1}$$

and

$$t_{2\,Da} = \frac{1}{(N-1)p_i}\frac{y_j}{p_j}(1-p_i) \quad . \tag{2}$$

After a gap of 5 years, Des Raj (1956) obtained ordered estimators (for $n = 2$):

$t_1 = y_i / p_i$ for first draw,
$t_2 = y_i + [y_j / \{p_j/(1-p_i)\}]$, based on second draw

and

$$\bar{t} = \{y_i (1 + p_i) / p_i\} + \{y_j (1 - p_i) / p_j\}/2 \text{ based on order } (i, j). \tag{3}$$

$t_i's$ are defined similarly, $i = 1,2,\ldots,n$ and

$t_n = y_1 + y_2 + \cdots + y_{n-1} + [y_n / (p_n / (1 - p_1 - p_2 - \ldots - p_{n-1})]$ or equivalently,

$t_n = y_1 + y_2 + .. + y_n + [y_n / \{x_n / (T_X - x_1 - x_2 - \ldots - x_{n-1} - x_n)\}]$.

By independence of estimators, it is easily seen that

$$\hat{V}(\bar{t}) = \sum_1^n (t_i - \bar{t})^2 / n(n-1)$$

is a non-negative unbiased estimator of variance.

2.2. Other negative variance estimators

Towards the end of the paper, Das gave an unbiased estimate of variance as well. This estimator received criticism since it *can take negative values*. Horvitz and Thompson (HT, 1952) gave a general homogeneous linear unbiased estimator for T_Y, which had *nice* properties. They also gave an unbiased estimator of variance of their estimator, but it *also takes negative values*. A year earlier, Narain (1951), *independently* obtained the same estimator and published in the *Journal of Indian Society of Agricultural Statistics (JISAS)* but was not mentioned by several authors. Thus, credit goes to Narain and HT, and J. N. K. Rao (1999, 2005) rightly called it as NHT estimator. In a discussion of Rao's 1999 paper, J.K. Ghosh observed that it was "*renamed NHT honouring another pioneer Narain".*

The very next year, Yates and Grundy (YG) in 1953 published an alternative variance estimator which *also takes negative values* (less often than HT's). It is interesting to note that Sen (1953) also published the same estimator as YG's in *JISAS*. It is now termed as SYG estimator, thus crediting all the three Sen, Yates and Grundy by researchers and teachers. However, a careful reading of YG's paper points out that, *unaware of* HT, Yates *also obtained the* HT estimator and *later* Grundy(G) *joined to give the alternative variance estimator.* Perhaps one should rename NHT estimator as NYHT estimator and SYG variance estimator as SG estimator!!

In view of the above discussion, we note that other variance estimators *also take negative values* and Das's estimator is much criticised. We note here that while for Des Raj's estimators, we need the previous Y values to be added to obtain the estimate at a particular draw, for Das's estimator at a particular draw, one need not know the Y values of the previous draws. This property comes in handy when one or more Y values of the previous draws are unavailable due to non-response, non-cooperation, 'not at home's *etc*. In such a situation one has to depend on Das's only and Des Raj's estimator is of no help.

2.3. Basu's concept of 'Face Validity'

More formally, this property can be stated as follows:

'*for estimating population total based on an ordered estimator, it is sufficient to have the Y value at the draw of selection only and the Y-values based on previous draws are not necessary.'*

Borrowing a phrase from Basu (1971), who defines the property of 'face validity', we term the above property as 'order validity'.

Basu (1971) looks at the population total as

$$T_Y = S + S^*,$$

where S is the observed total of Y's and S^* is the unobserved total.

Having observed the Y-values and knowing S, it is now required to estimate S^*.

Now, suppose that the n observed values Y_i / X_i are nearly equal, but y_n / x_n is the largest.

For this situation Desraj's estimator is

$$t_n = y_1 + y_2 + \ldots\ldots + y_{n-1} + (y_n / x_n)(T_X - x_1 - x_2 + \ldots\ldots + x_{n-1})$$

equivalently,

$$t_n = y_1 + y_2 + .. + y_n + [y_n / \{x_n / (T_X - x_1 - x_2 - ..- x_{n-1} - x_n)\}].$$

Basu questions estimating S^* by only one Y value, namely

$$y_n / \{x_n / (T_X - x_1 - x_2 - ..- x_{n-1} - x_n)\}].$$

Hence, Basu claims that it is not unbiasedness, but is hard to define property of 'Face Validity' of an estimate. He claims

$$t_n = \sum_1^n y_i + \{(\sum_1^n y_i / \sum_1^n p_i)\}(1 - \sum_1^n p_i),$$

which uses all n Y-values has a greater face validity.

Note that t_n is nothing but the familiar ratio estimator

$$\hat{Y}_R = \sum_1^n y_i / \sum_1^n x_i)T_X.$$

Following Basu's arguments, one could suggest a concept like face validity as:

'An estimator is said to be 'order-valid' if 'the estimator based on the result of a particular draw does not depend on the Y-values of the previous draws.'

However, this estimator may be inefficient, but in the presence of missing values due to non-response, not-at-home's *etc*., such an estimator may be relevant.

2.4. Lahiri-Murthy unordered estimator

Murthy (1957) concentrated on Des Raj's ordered estimators and discussed how to obtain an unordered (symmetrized) estimator. In a short section of this paper, titled 'unordering of Das's estimators', he briefly mentions Das's estimator, but unorders for another sampling scheme, and not for PPSWOR under consideration. He did not treat Das's estimator the way he did for Des Raj's as described below:

Recalling that for the ordered sample (i, j) the estimator is (3), namely

$$\bar{t}_{ij} = [\{y_i (1 + p_i) / p_i\} + \{y_j (1 - p_i) / p_j\}]/2 \text{ based on order } (i, j) \text{ and}$$

$$\bar{t}_{ji} = [\{y_j (1 + p_j) / p_j\} + \{y_i (1 - p_j) / p_i\}]/2 \text{ based on order } (j, i),$$

Murthy obtained an estimator combining these two by the respective probabilities of the sample as weights, namely $p_i p_j / (1 - p_i)$ and $p_j p_i / (1 - p_j)$, which gave the Unordered (symmetric) Des Raj Estimator:

$$\bar{t}_M = [\{(1 - p_j)(y_i / p_i)\} + \{(1 - p_i)(y_j / p_j)\}] / (2 - p_i - p_j),$$

which is Murthy's (3.17) of his 1957 paper.

A point to be noted here is that Halmos (1946) also mentions symmetrized unbiased estimators. For the last 70 years this is referred to as Murthy's (1957) unordered estimator. In a footnote of his 1957 paper (p. 384), Murthy mentions:

"*Lahiri conjectured that Desraj's estimators can be improved by weighting the different ordered estimators by their respective probabilities and in fact suggested the estimator given by (3.17)".*

So, it may be called Lahiri-Murthy unordered estimator:

$$\bar{t}_{LM} = \quad [\{(1-p_j)\,(y_i / p_i)\} + \{(1-p_i)\,(y_j / p_j)\}] / (2-p_i-p_j)$$

giving credit to Lahiri as well.

Symmetrizing Das's in the same way, we get an interesting symmetrized

$$\hat{Y}_{symm.Das} = (\hat{Y}_{Symm.Desraj} + \hat{Y}_{Midzuno.Lahiri})/2,$$

where $$\hat{Y}_{Midzuno.Lahiri} = \sum_1^n y_i \Big/ \sum_1^n p_i$$

and $\hat{Y}$ is an estimator of T_Y. The readers may like to see Rao(2021b) for details.

2.5. Further unorderings

For obtaining nonnegative SYG (or SG) variance estimators, Brewer (1963) and Durbin (1967) gave simple πps sampling selection procedures based on ordered samples of size 2. Brewer's method consists of selecting the first unit with probability proportional to $p_i(1-p_i)/(1-2p_i)$ and second unit with probability $p_j/(1-p_i)$, $j \neq i$. This gives $\pi_i = 2p_i$ and SYG variance estimator non-negative.

For the same purpose, Durbin's method selects the first unit with probability p_i and the second unit with probability proportional to $p_j[\{1/(1-2p_i)\} + \{1/(1-2p_j)\}], j \neq i$.

We now observe that the selection of units here is based on an order and we unorder these following the above methodology {see Rao (2021b)}. Thus, wherever order is involved in selection of sample, the estimators can be unordered using proper methodology by Rao-Blackwellization.

3. Nonresponse

So far, we have discussed situations that involved reduction of sampling errors. We shall now move on to the case of non-sampling errors of which non-response due to 'not at homes' and refusal to answer sensitive questions are major contributors. For the first category of 'not at homes', the technique of 'call-backs', while for the other one, 'randomized response technique (RRT) were proposed.

3.1. To call back or not to call back

For this, what is known as Politz-Simmons technique (PST) in the literature, is used to estimate parameters using data on first call itself, thus avoiding 'call-backs,' by asking respondents during the interview a question about their *availability at home (or, otherwise) at the same time during the preceding five week nights.*

However, during the discussion of the paper read by Yates(1946) at the Royal Statistical Society (RSS) Meeting, Hartley(1946) proposed an 'ingenious' and 'decidedly cheaper' alternative to call-backs . Hartley mentions: "*Details of this scheme were given to the War-time Social Survey, but I understand that, owing to pressure of work, an opportunity of trying has, as yet, not arisen".* Soon after, Politz and Simmons (PS,1949) published their work popularly known as Politz-Simmons technique in the *Journal of American Statistical Association* which is on similar lines to the proposed method of Hartley. PS (1949) while acknowledging the work of Hartley, say: *"It has recently been brought to the authors' attention that a somewhat similar plan was proposed independently by H.O. Hartley before the Royal Statistical Society...."*

In the present day context of ever-changing and emerging socioeconomic scenario of the society, it is to be noted that this question itself has become highly sensitive for the respondents who thereby may evade to answer this question truthfully. Rao *et al.* (2016) have applied Warner's (RRT) in this situation and developed a nontrivial randomized response Hartley, Politz, Simmons(HPS) technique.

4. Role of BKS in other early contributions

Hailing from, the then, East Bengal, environmentally rich and ecologically diverse background, it is but natural that BKS turned his attention to 'Environment' and other specialised areas of Statistics (see Rao, 2021a). BKS was appointed as 'Expert on Mission' for United Nations (UN) Statistics Training Programme in 1991based on his early contributions and this led to his serving as a consultant to the United States Environmental Protection Agency (USEPA) in 1993.

At home, he was also appointed a Member of the apex body, National Statistical Commission, Government of India (2006-2009). His expertise involved in social and environment statistics. Other early contributions of BKS include 'Official Statistics in neighbouring developing countries in the Indian sub-continent' (Rao and Sinha (2011). Collaborating with his colleagues JK Ghosh *et al.* (1999), a detailed account of 'Evolution of Statistics in India' was presented. Faculty and research scholars of Sociology and other applied statistics unis of ISI took BKS's and the author's help in organising their surveys rigorously. As an example we cite the design and implementation of an innovative survey of Annual Book Exhibition held in Calcutta Maidan, popularly known as '*boi mela'* wherein random time points are chosen.

5. Rao-Blackwellized WhatsApp

In the earlier sections, we have discussed the application of Rao-Blackwellization for improving the estimators in sampling theory. We have also mentioned its application in sports to obtain improved estimates of Field Goal Percentages (RB-FG%) in Basketball by Daly Grafstein and Bornn (2018). Their interesting analysis could be applied to ratings of sports persons in tennis, cricket and a host of others as well.

The new concept we proposed deals with 'message clustering' and 'smart response utility' while using WhatsApp (WA). Every day, we are flooded with WA messages on our smart phones. Not all users of WA have time to go through all the messages and take suitable action.

The new concept is based on an 'APP' to be constructed which compresses the data and disregards repetitions. An abridged message which is 'sufficient' is composed. For example, 'Good Morning', 'Have a Good day', Happy xxxx (day of the week), *etc*. can be treated as observations repeated with replacement. The App so constructed will have an AI/ML mechanism that recognises the equivalents and exhibits just one or two short lines editing meaningfully and then lists all the users that sent these particular messages, thus solving the problem of 'message clustering'. Now, an individual can quickly choose from the list, to whom the abridged (meaningful) reply can be sent (ignoring some senders) or a single 'Thank you all', if appropriate, thus enabling 'smart response utility'.

In view of the compression and reduction of data and the ability to present 'sufficient' information, we called it the Rao-Blackwellized WhatsApp (RW-WA). In a strict sense, this concept is not like the research of Daly-Grafstein and Bornn (2018). The new App so constructed reduces redundancy, saves time and effort and could even be made premium.

6. Certain lapses in literature and credits

In Sections 2 and 3, we have already mentioned about the credits that were missed out in sampling literature. We shall add here a few more (though not complete) with respect to the early results. The following anecdote may be of interest to the readers who are unaware of the history of the term 'Rao-Blackwellization':

C. R. Rao (1945) established this result and published in the *Bulletin of Calcutta Mathematical Society.*

A couple of years later, Blackwell (1947) obtained the same result in *Annals of Mathematical Statistics.*

Five years later, Scheffe' and Lehmann called it Rao-Blackwell Theorem.

In a 1953 conference, when Berkson named it Blackwellization. C. R. Rao pointed out that he published it in 1945 itself. Berkson replied "Raoization is difficult to say," but later termed it Rao- Blackwellization.

D.V. Lindley, in a book review referred to Blackwell only. When C.R. Rao wrote to him, he replied saying "you have not mentioned it in the introduction of the paper… C. R. Rao replied saying he is unaware that "introduction is written for the benefit of those who only read introduction and not the paper."

In the seminal paper read at the Royal Statistical Society meeting, Neyman derived the optimum allocation of sample size to strata in 1934. It was pointed out to him by Donavan Thompson that Tschprow had already established this result in 1923 Metron paper. Neyman recognized the priority and gave credit to Tchuprow. Thus, one may term this allocation as Tchuprow- Neyman allocation.

Hansen and Hurwitz introduced PPS sampling in their 1942 AMS paper. Mahalanobis in his 1937 paper discusses cumulative totals method for selection with varying probabilities.

Madow and Madow in 1944 discuss systematic sample, while for the selection of sample Anthropometric survey of United Provinces, Mahalanobis, Majumdar and C. R. Rao (1941) used a systematic sample. In the introduction, Mahalanobis points out that for detailed subclassifications, the ultimate sample size would be small giving large errors, a concept echoed in small area estimation.

Later while analysing Bengal Anthropometric data, C. R, Rao recognises that standard tests need to be applied cautiously since the data is based on multi-stage stratification heralding 'Analysis of Complex Surveys'.

Olkin's 1958 Biometrika paper on Multivariate regression estimators was envisaged by B. Ghosh in 1947 in Bulletin of Calcutta Statistical Association.

Murthy in 1964 rediscovers product method of estimation which was attempted using polykays by Robson in 1957 itself.

What we call as Midzuno-Sen (1952) sampling scheme is attributed to Midzuno's student Ikeda (1951), Haj'ek(1949) and Lahiri (1951), now popularly referred to as Lahiri-Midzuno-Sen (LMS) scheme.

Royall's 1970 predictive approach of Biometrika is also attributed to Brewer (1963) for introducing this concept.

It is not clear how one does not find a reference to Kumarappa's 1931 detailed survey of Matar taluka of Gujarat on the advice of Gandhi, which is a medium sized multi subject survey submitted for the attention of the British Raj, while discussing Mahalanobis's surveys of NSS (1950 onwards).

(For details and full references, please see T. J. Rao (2016), *On the History of Certain Early Key Concepts in Sampling Theory and Practice*).

Acknowledgements

The author wishes to thank Prof. Bimal Sinha, Prof. Vinod Gupta and the organisers of the Conference for inviting him to deliver the Bikas Sinha Endowment Lecture. On this occasion we all join in wishing Prof. Bikas Sinha Happy and Healthy years ahead, academically and personally.

References

Basu, D. (1958). On sampling with and without replacement. *Sankhya*, **20**, 287-294.

Basu, D. (1971). An essay on the logical foundations of survey sampling, part one. In: Godambe, V. P. and Sprott, D. A. (Eds.). *Foundations of Statistical Inference*, Holt, Rinehart and Winston, 203-242.

Blackwell, D. (1947). Conditional expectation and unbiased sequential estimation. *Annals of Mathematical Statistics*, **18**, 105-110.

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, **10**, 213-233.

Daly-Grafstein, D. and Bornn, L. (2019). Rao-Blackwellizing field goal percentage. *Journal of Quantitative Analysis of Sports,* **15**, 85–95.

Das, A. C. (1951). On two phase sampling and sampling with varying probabilities. *Bulletin of International Statistical Institute,* **33**, 105-112.

Des Raj (1956). Some estimators in sampling with varying probabilities without replacement. *Journal of the American Statistical Association,* **51**, 269-284.

Durbin, J. (1967). Design of multistage surveys for estimation of sampling error. *Applied Statistics*, **16**, 152-164.

Ghosh, J. K., Maiti, P., Rao, T. J., and Sinha, B. K. (1999). Evolution of statistics in India. *International Statistical Review,* **61**, 13-34.

Halmos, P. R. (1946). The theory of unbiased estimation. *Annals of Mathematical Statistics,* **17**, 34-43.

Hartley, H. O. (1946). Discussion of paper by F. Yates. *Journal of the Royal Statistical Society*, **109**, 37.

Hedayat, A. S. and Sinha, B. K. (1991). *Finite Population Sampling,* Wiley, New York.

Horvitz, D. G. and Thompson D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya*, **18**, 379-390.

Narain, R. D. (1951). On sampling without replacement with varying probabilities. *Journal of the Indian Society of Agricultural Statistics*, **3**, 169-174.

Pathak, P. K. (1961). Use of 'Order-Statistic' in sampling without replacement. *Sankhya, A,* **23**, 409-414.

Politz, A. N. and Simmons, W. R. (1949). An attempt to get the "not at homes" into the sample without call backs. *Journal of the American Statistical Association,* **44**, 9-31.

Rao, C. R. (1945). Information and the accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Mathematical Society*, **37**, 81-91.

Rao, J. N. K. (1999). Some current trends in sample surveys theory and methods. *Sankhya, B*, **61,** 1-57.

Rao, J. N. K. (2005). Interplay between sample survey theory and practice: An appraisal. *Survey Methodology,* **31,** 117-138.

Rao, T. J. (2016). On the history of certain early key concepts in sampling theory and practice. *Research Report,* **RR 2016-09**, C. R. Rao AIMSCS, Hyderabad, 1-30.

Rao, T. J. (2021a). Environmental statistics-A brief introduction. *Felicitation Volume of International Journal of Statistical Sciences in honour of Bimal Sinha and Bikas Sinha on their 75th Birthday.*

Rao, T. J. (2021b). Unordering of estimators in sampling theory: Revisited. *Journal of Statistical Theory and Practice,* **15**, 2021.

Rao, T. J. and Sinha, B. K. (2011). A brief history of statistics and its development in the Indian sub-continent. *International Journal of Statistical Sciences*, **11** (*Special Issue in memory of Prof. P. C. Mahalanobis).*

Rao, T. J., Sarkar, J., and Sinha, B. K. (2016). Randomized response and new thoughts on Politz-Simmons Technique. In *Handbook of Statistics,* **34** (Eds. C. R. Rao, A. Chaudhuri and T. C. Christophides), Chapter 15, 233-251.

Sen, A. R. (1953). On the estimate of the variance in sampling with varying probabilities. *Journal of the Indian Society of Agricultural Statistics,* **5**, 119-127.

Yates, F. (1946). A review of recent statistical developments in sampling and sample surveys. *Journal of the Royal Statistical Society*, **109**, 12-43.

Yates, F. and Grundy, P. M. (1953). Selection without replacement from within strata with probability proportional to size. *Journal of the Royal Statistical Society, B*, **15**, 253-261.