International Journal of Statistical Sciences *ISSN 1683-5603*
Vol.24(1), March, 2024, pp 85-89
DOI: https://doi.org/10.3329/ijss.v24i1.72023

Some Questions Related to Rao-Blackwellization and Association Rule Mining

T. J. Rao

Retired Professor, Indian Statistical Institute, Kolkata, India
*Correspondence should be addressed to T. J. Rao
(Email: tjrao7@gmail.com)

[Received January 1, 2024; Accepted February 8, 2024]

Abstract

Prof. CR Rao has been awarded the prestigious 2023 International Prize in Statistics**.** The citation reads: *"In his remarkable 1945 paper published in the Bulletin of the Calcutta Mathematical Society, Calyampudi Radhakrishna (C.R.) Rao demonstrated three fundamental results that paved the way for the modern field of Statistics and provided statistical tools heavily used in science today……".* These three results are 'Cramer-Rao Lower Bound' (CRLB), 'Rao- Blackwellization' (RB) and the third one now flourished as 'Information Geometry'. In this paper, we shall discuss two offshoots from his work over the eight decades. Several articles have appeared on his life and work (see for example, T. J. Rao (2019 and 2023a, 2023b) and Kumar (2023)). The first offshoot is based on one of the three breakthrough results, namely, Rao–Blackwell Theorem, first proved by C.R. Rao in 1945, when he was just 25 years old and also by Blackwell later in 1947. The second one is on Association Rule Mining (ARM), which he developed when he was 96 years old. These two papers reveal the transition of statistical methodologies from Fisherian concepts to recent applications of AI and ML. In this paper we shall pose some questions which need further study.

Keywords and Phrases: Rao-Blackwellization, WhatsApp, Message clustering, Smart response utility; Association rule mining, Sampling design, Inclusion probabilities.

AMS Classification: 62D05, 62P25, 68W20.

1. Introduction to Result 1.

According to one report, India has 535.8 million WhatsApp (WA) users in 2023, followed by Brazil with 148 million people using this app, while United States stands in the fourth position.

There have been some non-rigorous attempts to collect information by using WA. Some routine research published the descriptive and demographic statistics. *Statista* among others is one useful source. However, not many analytical studies have appeared, but during the last couple of decades there seems to be an increased interest in this aspect as well, as evidenced by certain publications. Gulacti *et al.* (2016) analysed the usage of WA for communication between Consulting and Emergency Physicians. This is based on a retrospective, observational study in the emergency department of a hospital consisting of 614 consultations.

One of the interesting analyses is by Rosenfeld *et al*. (2018) titled, 'WhatsApp usage patterns and prediction of demographic characteristics without access to message content', *Demographic Research*, 39, Art.22, 647–670, who use around 6 million encrypted WA messages from 111 students between the ages of 18 and 34. The encryption used the HMAC hash function so that the contents of messages were unknown to the authors and thus strict privacy of the content was adhered to. The objective of their study is to build up models to predict WA patterns among single users and groups of users. Earlier studies used message content leading to ethical and privacy questions. Encrypting data, Rosenfeld *et al.* got Internal Review Board approval to proceed further. For detailed analyses and results we refer to Rosenfeld *et al.* (2018) original paper. On the other hand, Mars *et al.* (2019) question the use of Instant Messaging for clinical studies where neither the guidelines are provided nor they are followed even if they do exist in some cases.

1.1. A new Concept

Though Rosenfeld *et al*.'s research (2018 a,b**)** could lead to further such data analytical contributions, as mentioned by them *"more accurate models can be built through studying data from more users, with a wider range of ages and different ethnic backgrounds*" , since their sample size is small, age range is much limited and user category is students only.

In this section, we shall introduce a new concept (floated recently in two Conferences, namely *International Conference on Knowledge discoveries on Statistical innovations & Recent Advances in Optimization (ICONKSRAO)*, December 2022 and *International Conference of the Society of Statistics, Computer and Applications (SSCA) On Significance of Statistical Sciences in the Emerging Scenario (SSSES 2023 –SSCA XXV))*, February 2023**.)**

Every day, we are flooded with WhatsApp (WA) text 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 we have suggested appeals to such users. Even on an ordinary day, an individual receives several 'Good Morning' Messages, photos of Gods/Goddesses, significance of the particular day of the week (In India, each day of the week is linked to a God/Goddess). And if it happens to be a festival, one receives a large number of messages and greetings, some from commercial sources or unknown advertisements. Besides these, there are 'forwards' sent to you with occasional videos.

Next, when a Group of users is formed, such as the one in connection with an international conference with possibly around 300 potential users, the idea is to exchange real time information on the conference. However, some group members post messages like "Thank you sir" or an occasional ''Happy xxxx'' (festival/birthday) or enquiries such as weather conditions at the venue or distances and posts not related to the conference.

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 constructed so 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 devote a fixed time of say 15 mts. around 9 a.m. before going to the office/school/clinic/lab etc. and 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'. It is assumed that for those users who wish to be in regular correspondence with their near family members and or friends that need special attention, such

messages would be separated out first and this App does not concern those. In group messages (say, of a conference), AI /ML would be able to pick up only those messages that do strictly concern the conference. It may be noted that we are not using AI for text mining since that is not our aim. If needed AI could be used partially as required.

In view of the compression and reduction of data and the ability to present 'sufficient' information, we called it the Rao- Blackwellized (C. R. Rao (1945), D. Blackwell (1947)) WhatsApp. We may point out that, in a strict sense, this concept is not like the research of Daly-Grafstein and Bornn (2018) on Basketball Field Goal (FG) Percentages wherein they "Rao-Blackwellized" the FG percentages, conditioning on other relevant available information from trajectory data of the ball. This method could also be applied to ratings of Sports persons in Tennis, Cricket and many others as well, since we now have a lot of data on these sports.

The App planned here saves time and effort, has no redundancy and if costs are involved could be made premium. It can possibly be used for Facebook, X (Twitter) etc. as well, where there is a clutter of unwanted messages or posts in one's account. Some similar APP s seem to be available on the internet which are not yet available in certain brands of cell phones and furthermore, they are of a different nature with a different purpose. The one suggested here is 'user friendly' for regular 'WhatsApp'ers, 'Facebook'ers, etc.

We may recall that Applications of Rao-Blackwellization are abundantly used in addition to improving estimators in conventional sampling theory (see Rao, T. J. (2021)), in Adaptive Sampling, link-tracing, size-biased sampling theories, Dynamic Bayesian Networks, Post Simulation Improvement of Monte Carlo Methods, Cross Validation and Non Parametric Bootstrap, particle filtering , stereology, data compression, Quantum Rao-Blackwell Theorem , Rao-Blackwellized Gaussian Smoothing, Rao-Blackwellized Parts-Constellation Tracker, Rao-Blackwellized Tempered Sampling (RTS), Assessment of California Condor recover, Rao-Blackwellized Field Goal percentage estimator (RB-FG%), recent Physics-based Rao-Blackwellization by Gueken et al. (2023) and a host of others including possibly Rao-Blackwellized WhatsApp (RB-WA), as we have seen above.

Next, we shall present the second problem of C R Rao:

2. Introduction to Result 2.

Search algorithms such as Apriori algorithm were developed for Association Rule Mining towards knowledge discovery (Agrawal *et al.* 1993, among others,) from transaction data, some of which are mostly deterministic. When the rule space for searching is large, such algorithms lead to heavy computations which are time consuming and costly. CR Rao along with Qian, Sun and Wu, developed stochastic search algorithms in Quin *et al.* (2016) to address this problem using idea of the stochastic search via Gibbs Sampling. The applicability of the algorithm has been demonstrated by examples, including one on a real genomic dataset containing a large set of 1067 items.

To understand the technique, they give an example of Market Basket Analysis from Agrawal *et al.* (1993). It states that the Rule could be *"90% of all customers who buy bread and butter also buy milk*". Such information is useful for super market's storing and shelving. Furthermore, they identify the most important association rules in a transaction data set.

2.1. Association rules and inclusion probabilities

In this section, we shall describe similarities and dissimilarities between Association Rules and Sampling Designs and pose certain questions on their relation for further study.

To draw a parallel, let **U** = (U_1, U_2, …, U_N) be a collection of N units(items) called Population (item space) and **S** = (s_1, s_2, …, s_M) be a list of samples (transactions) where each sample(transaction) in **S** is a subset of units (items) in **U**, i.e. $s_i \subset$ **U.**

An association rule is defined as an implication of the type u → v where u and v $\in$ **U** and u $\cap$ v = null. The set of units (items) u are called 'antecedent and v are called 'consequent' of the rule.

Support (u) is defined as the proportion of samples in **S** which contain u and Confidence (u → v) is defined as Support (u combined with v) /Support (u). Support measures its commonness and confidence measures its association strength.

It may be noted that association of u with v is not the same as association of v with u.

Consider the Example: U = (1, 2, 3, 4)
S = {s(1, 2,3); s(1,2,4); s(1,3,4);s(2,3,4)} (2.1.1)

Support $\varphi_1 = \varphi_2 = \varphi_3 = \varphi_4 = 3/4$ and Confidence (1→ 3) = Conf. ((1→ 4) = 2/3 etc. Notice that φ's are same as inclusion probabilities π's for SRS WOR design. Also note that Confidence is nothing but the Conditional Probability.

Apriori algorithm measures all rules satisfying minimum (φ_i) or maximum (φ_{ij}) (thresholds).

In general, for a design with given {p(s)}, φ's are not equal to π's, though in SRS it happens so. For unequal probability sampling φ's may still be equal and behave like SRS (N, n) so that φ's and π's may have same values. Another question of interest is to check while $p_i < p_j$, is it also true that $\varphi_i < \varphi_j$? (see T. J. Rao *et al.* (1988)). Sengupta (1979) worked on fixed size unequal probability sampling schemes providing constant π's, taking clue from Sinha (1976). Next consider the following Example

P{s(1, 2,3)} = 0.4; P{s(1,2,4)} = 0.2 and P{s(3,4)} = 0.4 . (2.1.2)

Here $\varphi_1 = \varphi_2 = \varphi_3 = \varphi_4 = 2/3$ (but π's are different in this case). Here $\sum_1^4 \varphi_i = 8/3$.

Furthermore, $\varphi_{12} = (2/3)/(2/3) = 1$, φ_{13} = other $\varphi_{ij} = (1/3)/(2/3) = ½$., and $\sum \varphi_{ij} = 3.5$.

The interpretation of φ's in this case, as opposed to π's needs to be investigated. Also note that $\varphi_{12} > \varphi_1$ as well as φ_2. Thus, whether the inequalities like those for π's and π_{ij}'s are true only for SRS is to be further explored.

However, since algorithms search for rules which have a threshold on φ's, the concept of thresholds on π_i's or π_{ij}'s is implied by them. Papers by Hedayat, Rao and Stufken (1988) on the choice of designs with small or zero π_{ij}'s, Goodman and Kish (1950) and Avadhani and Sukhatme (1973) on Controlled sampling may be of interest. For a variety of probability measures used for Association Rules one may refer to Hahsler (2015). We now pose two other important questions to be explored.

Since the idea is to minimize search for transactions, would OA's play any role? It is very likely that they do. While we are on the subject of Rao- Blackwellization in the first part, is it relevant here in ARM as well?

Acknowledgement. The author wishes to thank Prof. Bikas Sinha for some comments on Result2.

References (For Result 1):

[1] Blackwell, D. (1947): Conditional Expectation and Unbiased Sequential Estimation. Ann. Math. Statist. 18 (1) 105 – 110.
[2] Daly-Grafstein and Bornn (2019): Rao-Blackwellizing field goal percentage, J. Quant. Anal. Sports 2019; 15(2), 85–95.
[3] Geuken, G-L, Mosler, J and Kurzega, P. (2023): Incorporating sufficient physical information into Artificial Neural Networks: A guaranteed improvement via physics-based Rao-Blackwellization. ArXiv: 2311.06147v1/cs,CE/, 1-31.
[4] Gulacti, U., Lok, U., Hatipoglu, S., and Polat, H. (2016): An analysis of WhatsApp usage for communication between consulting and emergency physicians. J. Medical Systems 40(130), 1-7.
[5] Kumar, T. K. (2023): Dr. Calyampudi Radhakrishna Rao-The Centenarian legend of his centuries in statistics, IISA News Letter, Spring, 2023, 11-16.
[6] Mars , M, Christopher, M., Scott , R.E. (2019): 2019(9), WhatsApp guidelines - what guidelines? A literature review, J. Telemed Telecare, 524-529.
[7] Rao, C. R. (1945): Information and the Accuracy Attainable in the Estimation of Statistical Parameters. Bull.Cal.Math.Soc., 37, 81-91.
[8] Rao, T. J. (2019): Calyampudi Radhakrishna Rao-A Living Legend in Statistical Science: Developer of Statistics as an Independent Discipline, Technical Report., Research Gate.
[9] Rao, T. J. (2021): Unordering of Estimators in Sampling Theory: Revisited., J. Stat. Theory and Pract.,12.
[10] Rao, T. J. (2023a): Calyampudi Radhakrishna Rao-A Living Legend in Statistical Science, IISA News Letter, Spring, 2023,6-11.
[11] Rao, T. J. (2023b): Calyampudi Radhakrishna Rao (1920-2023). Current Science, 793-794.
[12] Rosenfeld, A., Sina, S., Sarne, D., Avidov, O. and Kraus, S. (2018a): WhatsApp usage patterns and prediction of demographic characteristics without access to message content, Demographic Research, 39, Art.22, 647–670.
[13] ---------- (2018b): A Study of WhatsApp Usage Patterns and Prediction Models without Message, arXiv:1802.03393, 1-24.

References (For Result 2):

[1] Agrawal, R., Imielinski, T and Swami, A. (1993): Mining association rules between sets of items in large databases. ACM SIGMOD Rec 22, 207–216.
[2] Avadhani, M.S. and Sukhatme, B. V. (1973): Controlled sampling with equal probabilities and without replacement, Inter. Stat.Rev.,41,175-182.
[3] Goodman, L. A. and Kish, L. (1950): Controlled selection- a technique in probability sampling, Jour.Amer.Stat.Assoc., 45, 350-372.
[4] Hahsler, M. (2015): A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules, 2015, URL: http://michael.hahsler.net/research/association_rules/measures.html
[5] Hedayat, A. S., Rao, C. R. and Stufken, J. (1988): Sampling plans excluding contiguous units. Jour.Stat.Plann and Inf.,19, 462-491.
[6] Qian, G., Rao, C. R., Sun, X. and Wu, Y. (2016): Boosting association rule mining in large datasets via Gibbs sampling. Proc.Nat.Acad.Sc.(Physical Sciences), 113 (18) 4958-4963.
[7] Rao, T. J., Sengupta, S. and Sinha, B. K. (1990-91): Some probability inequalities for PPSWOR sampling scheme, Metrika, 38, 335-343.
[8] Sengupta, S. (1979): Construction of some non-invariant balanced sampling designs, Cal.Stat.Assoc.Bull. , 31, 165-185.
[9] Sinha, B. K. (1976): On balanced sampling schemes, Cal.Stat.Assoc.Bull. 25,129-138.