Interview with Professor Calyampudi Radhakrishna Rao*

Frank Nielsen†

December 2016

Abstract

C. R. Rao has contributed to facets of modern statistics such as differential-geometric methods in statistics, score test, quadratic entropy, orthogonal arrays, multivariate analysis, and generalized inverse of a matrix (singular or not) and its applications. Frank Nielsen — a professor of computer science at École Polytechnique, Palaiseau, France, and a senior researcher at Sony Computer Science Laboratories, Inc. — interviewed Rao this past year to learn more about his life and work. What follows is what he discovered.

Can you briefly tell us about your family and education in India?

I was born on September 10, 1920, in a small town in Madras Presidency (under British rule known as Hadagali). I am the eighth child out of 10 (four girls and six boys) to my parents.

One of my sisters was a Telugu (my mother tongue) poet. Another sister was a business woman selling cars imported from Britain. The seventh child was a boy who had phenomenal memory. He received a gold medal on his anatomy exam for remembering the names of all the bones and other organs of the human body. All of us had all education from primary school to college in India. None of us has any foreign college degrees. I have two MA degrees, one in mathematics and another in statistics from universities in India. However, I was awarded 38 honorary doctorates by universities in 19 countries from six continents.

You have been interested from the very beginning of your research in studying the interplay of statistical methods with differential geometry. Can you please hint at what drove you in the direction of geometric statistics?

I joined the Indian Statistical Institute (ISI) in 1943 as a research scholar. Professor Mahalanobis, head of ISI, asked me to analyze anthropometric measurements made on different castes and tribes in an Indian state by an anthropologist using what is known as Mahalanobis Distance (MD) between two populations (probability distributions).

It occurred to me that MD is appropriate only to multivariate normal distributions and [I] tried to get a general formula for any distribution. I was inspired by the metric developed by Einstein to calculate distances and thought this was an appropriate approach. This led me to the metric based on the Fisher Information matrix.

This paper generated the technical terms Cramr-Rao inequality or bound (CRB), Rao-Blackwellization, Fisher-Rao Metric, and Rao Distance. It received attention by a number of mathematicians and quantum physicists whose work generated technical terms: Complexified and Intrinsic CRB, Quantum CRB, Rao Measure, and Cramr-Rao functional. Rao distance also played an important role in solving problems in econometrics.

Cambridge University (CU) sent an expedition to Jebel Moya in Africa to dig out ancient graves and bring the skeletons to the anthropology department (AD) of CU. The department wanted to analyze the measurements on skeletons to determine the relationship of the people who lived there with people currently living there and in nearby areas.

Figure 1: C. R. Rao

At the invitation of Dr. Trevor of the AD to ISI, I was deputed to go to CU to analyze the data on skeletons. I spent two years (1946–1948) as a paid visiting scholar in the AD of CU analyzing the data on skeletons. The problem involved the computation of the distances of Jebel Moya skeletons with skeletons at other places. A complete map of distances of skeletal data from nearby areas to Jebel Moya data was worked out. The results were reported in the book Ancient Inhabitants of Jebel Moya, published by Cambridge University Press in 1954.

The new methods developed by me to analyze skeletal data were published in the Journal of the Royal Statistical Society and Biometrika during the forties of the last century. They constitute the basic methodology of multivariate analysis that appears in text books on statistics. For my work at CU, I received the PhD degree of CU. A few years later, in 1965, I received the senior doctorate ScD of CU and life membership of Kings College in Cambridge.

You have investigated the role of statistical distances and measures of diversity in statistics (Riemmanian distance, Burbea-Rao divergence, etc.). Can you please briefly explain why it is important to study those various measures?

R. A. Fisher developed analysis of variance (ANOVA) of quantitative measurements made on individuals of a population to test the number of subpopulations and their inter-relationships. We have two-way, three-way, and multi-way ANOVA, depending on the number of sub-populations for testing such hypotheses. Can we do such analysis with other measures of variation or diversity functionsor entropy as we named itsuch as mean deviation, range, Gini coefficient of concentration, and various types of entropy functions used by ecologists? In a series of papers, Burbea and Rao developed the necessary mathematical conditions termed as the degree of convexity of a diversity function, which places a limit on the analysis of multiway classified data. We use the acronym ANODIV (analysis of diversity) for decomposition of a diversity measure into a number of components with assignable causes, as in ANOVA. We have shown that Havdra and Chravat entropy, paired Shannon entropy, Rényi entropy, Gini Simpson index, and other well-known entropy functions can be used for two-way ANODIV, but not for higher-order ANODIV. It is shown that quadratic entropy (QE), introduced by me, is a perfect diversity measure that can be used to carry out ANODIV of any order. QE is defined for both quantitative and qualitative variables. Variance is a special case of QE. For details, reference may be made to the paper, "Quadratic Entropy and Analysis of Diversity," Sankhya, 72A, 70–80.

Could you please describe three of your major contributions to statistics?

I may mention orthogonal arrays (OAS), score statistic, and QE, which have received wide applications.

My paper on OAS was accepted as a fresh and original piece of work and published in Proceedings of Edinburgh Mathematics Society in 1949. The Statistical Quality Control (SQC) division of the ISI started using OAS to conduct experiments to determine the optimum combination of several factors to produce a high-quality product and control its quality during production.

An article in Forbes Magazine (March 11, 1969) refers to OAS as "a new mantra" in a variety of industrial

establishments in the USA. A full-length book on OAS by Hedayat, Sloane, and Stufken gives various applications of OAS.

The score test was introduced in a paper titled, "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Applications to Problems of Estimation," published in the Proceedings of the Cambridge Philosophical Society in 1948. It generated the technical term score statistic, which appears in textbooks about econometrics. The score test was developed as an alternative to Pearson Chi-square and Walds test, together known as the Holy Trinity.

Finally, Professor Rao, let me ask you, "What is the place of statistics in data science?"

"A theory can be proved by an experiment, but no path leads from experiment to theory." — Einstein

Computer scientists are using a number of researchers to work on systems that move from data to knowledge to theories, and then can reason, rather than merely recognize patterns and correlations and make predictions. Very little progress has been made in this endeavor. However, some progress has been made through methods of data science, a new interdisciplinary domain that aims at providing data-driven solutions to difficult problems that are ill-posed for precise algorithmic solutions. Such problems abound in computer vision, natural language processing, autonomous vehicle navigation, and image and video processing.

Data science provides solutions to problems by using probabilistic and machine learning algorithms. Often, multiple solutions to a problem are provided and a degree of confidence is associated with each solution. The data science approach closely reflects the way humans solve problems that are difficult to characterize algorithmicallyhandwriting recognition, word-sense disambiguation, and recognizing objects in images.

The emergence of Big Data is the genesis of the data science discipline. Big Data enables scientists to overcome problems associated with small data samples. With big enough data, certain assumptions of theoretical models can be relaxed, over-fitting of predictive models to training data can be avoided, noisy data can be effectively dealt with, and models can be validated with ample test data.

Like computer science, data science is a scientific discipline. It is a scientific discipline because it uses an experiment-oriented scientific approach. Based on empirical evidence, a hypothesis is formulated, and evidence is gathered to perform the hypothesis testing.

This is where cognitive science plays a natural role in complementing data science. Data science deals with data issues, experimental design, and hypothesis testing, whereas cognitive science provides theories, methods, and tools to model cognitive tasks. More specifically, cognitive science provides frameworks to describe various models of human cognition, including how information is represented and processed by the brain.

A longer interview[1] between Rao and Anil K. Bera [1]. See also [2] for an introductory article of the pioneering work of Rao in information geometry.

References

[1] Anil K Bera. THE ET INTERVIEW: PROFESSOR CR RAO: Interviewed by Anil K. Bera University of Illinois at Urbana-Champaign. *Econometric Theory*, 19(02):331–400, 2003.

[2] Frank Nielsen. Cramér-Rao lower bound and information geometry. *arXiv preprint arXiv:1301.3578*, 2013. Published in "Connected at Infinity II: A Selection of Mathematics by Indians", Rajendra Bhatia, C. S. Rajan, and Ajit Iqbal Singh (Editors).

---

[1] http://korora.econ.yale.edu/et/interview/rao.pdf