Special Proceeding of the 22nd Annual Conference of SSCA held at Savitribai Phule Pune University, Ganeshkhind, Pune, during January 02-04, 2000; pp 153-162

On Weighted Distributions and Applications

Saumyadipta Pyne[1,2,3]

[1]*Public Health Dynamics Laboratory, University of Pittsburgh, Pittsburgh, USA*

[2]*Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, USA*

[3]*Health Analytics Network, PA, USA*

Received: 15 August 2020; Revised: 22 August 2020; Accepted: 25 August 2020

Abstract

The birth centenary of C.R. Rao in 2020 presents an occasion to not only celebrate the remarkable life and career of a living legend of statistics, but also remember the worldwide immense development of the field over the past century. Here, I discuss about C.R. Rao's pioneering work on a general theory of weighted distributions presented in 1965, which was followed by significant development in that area. I end with discussion of some recent advances in methodology with applications to environmental data fusion.

Keywords: C.R. Rao; Weighted distribution; Density ratio model; Data fusion.

0. About This Paper

This paper is based on the first lecture delivered during the C.R. Rao Birth Centenary Session of the 22nd Annual Conference of Society of Statistics, Computer and Applications and ISGES 2020, held on January 2, 2020, at the Department of Statistics, Savitribai Phule Pune University, Pune. Three more talks were delivered during this session. The session was Chaired by Professor Vinod K. Gupta.

1. Background

Calyampudi Radhakrishna Rao, popularly known as "C.R. Rao", was born on the 10th of September, 1920, in the Madras Presidency of British India. He is widely regarded as a "living legend" in the field of statistics, and known for Cramer-Rao Bound, Rao-Blackwell Theorem, Rao Score Test, Fisher-Rao distance, Generalized Inverse, Quadratic Entropy, and Orthogonal Arrays, among his numerous path breaking contributions. Along with P.C. Mahalanobis, he played a major role in developing the Indian Statistical Institute (ISI) into the major center of statistical research and education by the 1940s. C.R. Rao's long list of awards includes the Padma Vibhushan (2001) and Padma Bhushan (1968), India Science Award (2009), the United States (US) National Medal of Science (2002), Wilks Memorial Award (1989), the Guy Medal in both Silver (1965) and Gold (2011).

Compared to some of Rao's other breakthroughs in statistics as stated above, the topic of my lecture, weighted distributions, is relatively less well known but has, nonetheless, led to great advances of research in the subsequent decades. It provides a curious counterpoint to the popular refrain that every student of statistics gets used to in her daily practise: "randomly

Corresponding Author: Saumyadipta Pyne
Email: spyne@pitt.edu

drawn samples are assumed to be independently and identically distributed." The topic has been found relevant to many theoretical and applied areas of statistics, and could, in fact, help modern data science to deal with analysis of samples that may not be collected through well-designed experiments. Interestingly, in the arc of Rao's remarkable career as it spanned over the better part of the last century, this topic happens to make a uniquely historic contribution.

In 1911-1914, (later Sir) Henry Wellcome, the British pharmaceutical entrepreneur had led, over four seasons, an archeological expedition at the Jebel Moya site in the southern Gezira plain of Sudan (Addison, 1949). It is the site of the largest pastoralist cemetery in Northeast Africa, currently dated to 5000–500 BC (Brass, 2016). The excavated skeletal remains were shipped to London where these were warehoused poorlyand even suffered from flooding. After the death of Wellcome (in 1936), and by the end of two World Wars, the remains and the excavation records were transferred to the Duckworth Laboratory in the University Museum of Archaeology and Ethnology at Cambridge University. The Wellcome Trust wanted the study to be completed under the curatorship of J.C. Trevor.

In March 1946, Trevor sent a telegram to Mahalanobis asking him to send someone to help with the anthropomorphic analysis of the collection. Incidentally, just prior to that, Mahalanobis had assigned a project on analysis of anthropometric data to Rao, who used the D^2 distance for grouping of Indian populations (Mahalanobis, Majumdar and Rao, 1949). This experience led to his selection to go, along with Ramkrishna Mukherjee, an anthropologist, to England in August 1946 to work as visiting scholars at the Cambridge University Museum of Archaeology and Anthropology. The aim of their study was "to undertake laboratory examination of identifiable and usable adult specimens, to analyse the measurements and observations of the field physical anthropologists, and to determine the relationship between the Jebel Moya inhabitants and other African peoples" (Brass, 2016).

Upon completion of their analysis, the long overdue report of this Wellcome Trust project was finally published in 1955 by Mukherjee, Rao and Trevor as a book titled, '*The Ancient Inhabitants of Jebel Moya* (*Sudan*)'. (Mukherjee, Rao and Trevor, 1955) Unfortunately, the 40-year hiatus between the excavation and the anthropometric data analysis had proved to be catastrophic for the remains, which had "disintegrated beyond hope of repair", according to Trevor. Out of more than 3000 skeletal parts originally excavated, only 98 crania, 139 mandibles and a few post-cranial elements had survived for conducting the anthropomorphic studies by Rao and his co-workers at Duckworth Laboratory.

Rao's task was to estimate the unknown mean cranial capacity and other features of the *original* Jebel Moya population from the damaged remains, many of which had measurements missing due to damage. Towards maximum likelihood estimation, one could write the likelihood function using a multivariate (normal) distribution based on the samples with complete measurements, and the derived marginal distribution for those with incomplete set of measurements. However, such estimation assumes that each skull – with all or some of the measurements – can be considered as part of a *random* sample from the original population of skulls. The key question was, however, was such an assumption valid?

While looking at the samples that survived, one could make a curious observation, "are only small skulls preserved?" (Figure 1). If $w(c)$ is the probability that a skull of capacity c is unbroken, then in archeological recovery, *i.e.*, during the data gathering process, it is known that $w(c)$ is a decreasing function of c. The larger a skull, the greater is its chance of being damaged upon burial and recovery, as Rao noted in his study of another collection (Rao and

Shaw, 1948). This will lead to a larger representation of small skulls among the unbroken cranial remains, and, therefore, the mean of the available measurements of the corresponding random variable C will be an underestimate of the mean cranial capacity of the original population.

While analysis of Jebel Moya remains was his full-time job, Rao, who was also a student in King's College, Cambridge, was "suggested" by his Ph.D. advisor R.A. Fisher to work simultaneously at the latter's genetics laboratory. Fisher had long been interested in the concept of "ascertainment" – a mode of sampling that depends on the outcome that one wanted to analyze as a dependent variable. In a classical paper, Fisher studied how the methods of ascertainment can influence the form of the distribution of recorded observations (Fisher, 1934).One can assume a model that has been adjusted for ascertainment will estimate parameters in the *general* population from which the sample was drawn. In usual statistical practice, it is generally assumed that a random sample from a population to be studied can be observed in data. Not surprisingly, therefore, the key specification of 'what population does a sample represent?' might be taken for granted.

However, in many situations, obtaining a random sample may be practically too difficult, or too costly, or indeed, even less preferable to the non-random data that are actually available, *e.g.*, from field observations or non-experimental data or a survey lacking a suitable sampling frame, as in many 'big data' problems. Sometimes the events may be observed only in modified form, *e.g.*, in damage models. Indeed, certain events may either be unobservable due to adoption of the very method that is used for making observations, and therefore, missed in the record. Or, they might be observable but only with certain probabilities or weights that may depend on the samples' specific characteristics, such as conspicuousness, as well as other unknown parameters.

2. Weighted Distributions

On 15-20 August of 1963, G.P. Patil organized the 'First International Symposium on Classical and Contagious Distributions' at McGill University in Montreal. Patil had visited ISI a decade earlier, where he recollected being advised by Rao to study discrete distributions. In the Montreal symposium, at a 2nd day session chaired by Jerzy Neyman, Rao presented the first paper that formulated and unified weighted distributions in general terms. It was titled, '*On discrete distributions arising out of methods of ascertainment*'. Later, the paper was published by the Statistical Publishing Society, Calcutta, in the Proceedings, 'Classical and Contagious Discrete Distributions', edited by Patil (Rao,1965).

Let X be a random variable (rv) with probability density function (pdf) $g(x;\theta)$ with parameters θ. Traditional statistical analysis assumes that an identically (as well as independently) distributed random sample $X_1, \ldots, X_n$ can always be observed. Weighted distributions arise when $X = x$ enters the sample with a non-zero weight $w(x, \alpha)$ that depends on the observed value x and possibly also on some unknown parameter α. Then x is not an observation on X but on the resulting rvX^w which has the weighted pdf:

$$f(x;\theta,\alpha) = \frac{w(x,\alpha)\cdot g(x;\theta)}{E[w(X,\alpha)]}$$

The denominator $E[w(X;\alpha)]$ is a normalizing constant so that $f(x;\theta,\alpha)$ integrates to 1. The weight $w(x,\alpha)$ could be any non-negative function for which $E[w(X;\alpha)]$ exists.

When x is univariate and non-negative, then the weighted distribution $f(x;\theta) = x \cdot g(x;\theta)/E(X)$ for $w(x,\alpha) = x$ is called *size-biased*; and *length-biased* for $w(x,\alpha) = |x|$ where $|x|$ is some measure of "length" of x. Many length-biased distributions could be shown to belong to the same family as their unweighted versions (Rao, 1965).

Weighted distributions are utilized to modulate the probabilities of the events as they are observed by means of collecting data possibly under less than perfect conditions. In this context, prominent application areas of weighted distributions include truncation and censoring, damage models, size-biased sampling, quadrat sampling, nonresponse in data, "file-drawer" problem in meta-analysis, *etc*. Even mixtures of distributions can be shown to belong to this general formulation (Larose and Dey, 1996). Interestingly, sometimes biased samples available from observational studies may contain more (Fisher) information than their randomly drawn counterparts (Bayarri and DeGroot, 1992). Patil and Rao (1978) studied size biased sampling with applications to wildlife populations and human families. Different applications of weighted distributions were reviewed by Patil and Rao (1977), and in reports and articles written byRao during his Pittsburgh years (Figure 2), *e.g*., Rao (1985), Rao (1988).

As an example of such special cases of weighted distributions, we consider skewed data. Let the p-variate Generalized Skew Elliptical (GSE) distribution of rv $z \in R^p$have pdf of the form $2g(z;\xi,\Omega)\pi(z-\xi)$ where g is an elliptically contoured pdf with location parameter ξ, scale matrix Ω, and skewing function π. An example of a GSE pdf due to Branco and Dey (2001) is $\frac{2}{\sqrt{\Omega}} g(\Omega^{-\frac{1}{2}}(z-\xi)) \cdot \pi(\Omega^{-\frac{1}{2}}(z-\xi))$, which could be formulated as a weighted distribution as follows: $f(z;\theta) = g(\Omega^{-1/2}z)/|\Omega|^{1/2}$, $w(z) = \pi(\Omega^{-1/2}z)$, $E[w(Z)] = 1/2$ (assuming ξ=0). Here, the weight function w distorts the elliptical contours of f *via* generation of asymmetric outliers in the observed sample due toGSE (Genton 2005).

Such formulations demonstrate the capacity of introducing weighting mechanisms to "distort" distributions as required. This allows important applications such as modeling of dynamic patterns as they emerge in data over space and/or time, as in environmental monitoring, statistical ecology, public health, etc. For describing the changes over time in the distribution of an environmentally important variable X, a *propagation function* (PF) models how the frequency of sampling units with the value $X = x$ at one point in time must change in order to produce the distribution that occurs at a later point in time. Thus, PF is a useful tool in long-term monitoring studies since all changes in a distribution can be examined together rather than just changes in single parameters such as the mean.

Let X and Y be the values of an environmental variable on a population unit at two consecutive time-points, with the marginal densities $f_X(\cdot)$ and $g_Y(\cdot)$ (Kaur, *et al*., 1995). The PF is defined as: $w(x) = g_Y(x)/f_X(x)$. This gives a weighted distribution: $g_Y(x) = w(x) \cdot f_X(x)$ such that $E_f[w] \equiv E[w(X)] = 1$. A similar concept of *resource selection function* (RSF) in wildlife habitat modeling was defined as a logistic discriminant function in terms of a ratio of pdf-s for used (f_u) and available (f_a) resources in k habitats: $f_u(x)/f_a(x) = w(x) = \exp(\sum_{i=1}^{k} \beta_i x_i)$, where β_i is the model coefficient for the i^{th} habitat's covariate x_i(McDonald, Gonzalez and Manly,1995).A sampling strategy could be biased in its observation of the used sites, given their notably used features, over the unused sites. Later, the idea of a general density ratio model (DRM) was further developed and applied to case-control studies (Qin, 1998).

3. Recent Extensions to Weighted Systems

Recent studies have extended the concept of weighted distributions to weighted *system* of distributions. This general approach allows us to develop a powerful computational framework for modeling the combined dynamics of environmental samples as collected from multiple sources, *e.g.*, spatially distributed multivariate data streams, automated sensors and surveillance networks (Zhang, Pyne and Kedem, 2020a; Zhang, Pyne and Kedem, 2020b). Often, environmental monitoring stations are not located randomly, may contain built-in redundancies, and are difficult to maintain. Therefore, by systematic fusion of a moderate number of such nearby data sources, we can increase the predictive capability with a combined model.

We assume a reference event distribution g_0, and its possible distortions or "tilt"-ed forms $g_1, \dots, g_m$ due to m different sources. This gives us a weighted system of distributions:

$$
\begin{gathered}
g_1(x) = w_1(x) \cdot g_0(x) \\
\vdots \\
g_m(x) = w_m(x) \cdot g_0(x)
\end{gathered}
$$

Assume that we have data from each of $g_0, g_1, \dots, g_m$. Then, the relationship between a baseline distribution and its distortions or tilts allows us to do inference on the system's behavior based on fusion of data observed from multiple, possibly dependent, sources. In a weighted system using density ratio models, various distributions are "regressed" on a common reference distribution, and our data fusion approach to estimate the parameters, including the densities, uses the entire combined data and not just the reference sample. Tail probabilities of exceeding a pre-specified event threshold can be estimated by using the DRM with variable tilt functions (Zhang, Pyne and Kedem, 2020a).

As an application of our data fusion approach, we combined data from multiple sites on environmental exposures of Radon, a known radioactive carcinogenic gas, using a multi-sample DRM defined as follows:

$$\frac{g_k(x)}{g_0(x)} = \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(x)\right) k = 1, \dots, m$$

where g_0 represents the density of residential radon levels of the county of interest and $g_1, \dots, g_m$ represent the densities of its m nearby sites. Instead of making parametric assumptions on each of these densities, DRM captures the parametric structure of their ratios with a common model. To prevent bias, large standard errors, and loss of power in inference, it is important to properly select the variable tilt functions $\boldsymbol{h}_k$'s.

Let $\boldsymbol{X}_0, \dots, \boldsymbol{X}_m$ be the samples from the area of interest and its m nearby sites with sample sizes $n_0, \dots, n_m$ respectively. The sample $\boldsymbol{X}_0$ is referred to as the reference sample and let G denote the corresponding reference cumulative distribution function (CDF). The fused sample is defined as $\boldsymbol{t} = (\boldsymbol{X}_0^T, \dots \boldsymbol{X}_m^T)^T$, with size $n = \sum_{k=0}^{m} n_k$. Inference can be based on the following empirical likelihood obtained from the fused sample $\boldsymbol{t}$:

$$L(\boldsymbol{\alpha}, \boldsymbol{\beta}, G) = \prod_{i=1}^{n} p_i \prod_{k=1}^{m} \prod_{j=1}^{n_k} \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(X_{kj})\right)$$

where $p_i = d(G(t_i))$, and the estimates $\tilde{\boldsymbol{\alpha}}$, $\tilde{\boldsymbol{\beta}}$ and hence $\tilde{p}_i$'s are obtained by maximizing the above likelihood with constraints:

$$\sum_{i=1}^{n} p_i = 1 \qquad \sum_{i=1}^{n} p_i \exp\left(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{h}_k(t_i)\right) = 1 \qquad k = 1, \ldots, m.$$

This gives the estimated reference CDF $\tilde{G}(t) = \sum_{i=1}^{n} \tilde{p}_i \, I[t_i \leq t]$ and the asymptotic result

$$\sqrt{n}(\tilde{G}(t) - G(t)) \xrightarrow{d} N(0, \sigma(t)),$$

as $n \to \infty$. The expression of $\sigma(t)$ and other details regarding estimation and asymptotic result can be found in our recent paper (Zhang, Pyne and Kedem, 2020a). Based on the above result, a 95% confidence interval of the tail probability $1 - G(T)$ for a given threshold $T$is given by

$$(1 - \tilde{G}(T) - z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}, 1 - \tilde{G}(T) + z_{0.025}\sqrt{\frac{\tilde{\sigma}(t)}{n}}).$$

Finally, an optimal choice of tilt function may be made based on a criterion that ensures better specification of the density ratio structure. For instance, such selection can be made using Akaike Information Criterion (AIC) given by $-2\log L\left(\tilde{\alpha}, \tilde{\beta}, \tilde{G}\right) + 2q$ where q is the number of free parameters in the model (Zhang, Pyne and Kedem, 2020b).

4. Conclusion

In statistical inference, Rao noted, "wrong specification may lead to wrong inference, which is sometimes called the third kind of error in statistical parlance. The problem of specification is not a simple one." (Rao, 1988) The aim of my lecture, therefore, was to draw the attention of students and researchers to the fascinating area of weighted distributions that is all the more relevant in the current age of data science. Patil and Rao (1977) remarked, "although the situations that involve weighted distributions seem to occur frequently in various fields, the underlying concept of weighted distributions as a major stochastic concept does not seem to have been widely recognized." As more datasets of value seemingly lacking any rigorous design are collected, and possibly shared through unconventional and occasionally biased sources, the need for working with such less than ideal yet practically useful observations will have to be addressed by statistical pedagogy.

Personally, I had the good fortune to have had Professor Rao as my colleague, mentor and collaborator while serving as the P.C. Mahalanobis Chair Professor (sponsored by the Ministry of Statistics and Program Implementation, Government of India) and Head of Bioinformatics at the CR Rao Advanced Institute of Mathematics, Statistics and Computer Science (AIMSCS) in Hyderabad during 2012-2015. Starting in 2009, we had interactions at the early stages of planning the creation of the institute. We worked together on multiple projects that remain close to my heart, including the organization of the 'International Year of Statistics 2013' (STAT 2013) Conference held at CR Rao AIMSCS on December 28-31, 2013, under my convenorship; inauguration of the 'Data Science Laboratory for Environmental and Health Sciences' at my initiative at CR Rao AIMSCS by Professor Amartya Sen on 19 December, 2013 (Figure 3); co-editing a 2-volume 'Handbook of Statistics' on Disease Modeling and Public Health' (Rao, Pyne and Rao, 2017); founding in 2014 of a Special Interest Group in Computer Society of India on 'Big Data Analytics' with myself as the founding chair. Notably, in STAT 2013, we hosted an early meeting of international experts on 'big data' in India (Pyne, Rao and Rao, 2016).

I, therefore, felt immensely proud and privileged to deliver the first lecture in the Professor C.R. Rao Centenary Lecture Session on January 2, 2020, of the International Conference ISGES 2020 organized by the Society of Statistics, Computer and Applications at the Department of Statistics, Savitribai Phule Pune University. Incidentally, the University is also my alma mater, which doubled my pleasure to speak both at this occasion as well as the venue. For this, I thank the organizers of the conference, and, in particular, my dear friend and former ICAR-IASRI National Professor, Professor Vinod K. Gupta, for kindly inviting me to deliver this lecture.

The citation with the National Medal of Science, the highest award in the US in a scientific field, honored Rao "as a prophet of new age for his pioneering contributions to the foundations of statistical theory and multivariate statistical methodology and their applications, enriching the physical, biological, mathematical, economic and engineering sciences."Indeed, to truly appreciate Rao's "putting chance to work", I encourage students to read his popular writings (Rao 1997). It has been a high honour for me to have worked alongside this living legend and received his guidance and blessings. In August 2019, I visited him (and his daughter Teja Rao) at his family home in Buffalo, New York, and expressed my deepest gratitude. I offer my heartiest congratulations to Professor C.R. Rao on his birth centenary year, and wish him a longer, healthy life.

References

Addison, F. (1949). *The Wellcome Excavations in the Sudan: Vol. I, Jebel Moya, 1910-1914*. Oxford University Press, Oxford.

Bayarri, M. J. and DeGroot, M. H. (1992) A "BAD" view of weighted distributions and selection models. In Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (Eds.), *Bayesian Statistics*,**4**, Oxford: University Press, 17-33 (with discussion).

Branco, M. D. andDey, D.K. (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis*, **79**, 99-113.

Brass, M. J. (2016). *Reinterpreting Chronology and Society at the Mortuary Complex of Jebel Moya* (*Sudan*). Archaeopress Publishing Ltd., Oxford.

Fisher, R. A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, **6**, 13–25.

Genton, M. G. (2005). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Chapman and Hall/CRC, Boca Raton.

Kaur, A., Di Consiglio, L., Patil, G. P. et al. (1995). Propagation functions for monitoring distributional changes. *Environmental and Ecological Statistics*, **2**, 239–269.

Larose, D.T. andDey, D. K. (1996). Weighted distributions viewed in the context of model selection: A Bayesian perspective. *Test*, **5**, 227–246.

Mahalanobis P.C., Majumdar D.N. and Rao C.R. (1949). Anthropometric survey of the United Provinces: a statistical study. *Sankhya*, **9**, 89–324.

McDonald, L. L., Gonzalez, L. and Manly, B. F. J. (1995). Using selection functions to describe changes in environmental variables. *Environmental and Ecological Statistics*, **2**, 225-237.

Mukherjee, R., Rao, C. R. and Trevor, J. C. (1955). *The Ancient Inhabitants of Jebel Moya* (*Sudan*). Cambridge University Press, Cambridge.

Patil, G. P. and Rao, C. R. (1977). The weighted distributions: A survey of their applications. In P.R. Krishnaiah (Ed.), *Applications of Statistics*, North Holland, Amsterdam, 383–405.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179–189.

Pyne, S., Prakasa Rao, B. L. S., Rao, S. B. (Eds.) (2016). *Big Data Analytics*: *Methods and Applications*. Springer, New Delhi.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, **85**(**3**), 619-630.

Rao, A.S., Pyne, S. and Rao, C.R. (Eds.) (2017). *Handbook of Statistics*: *Disease Modelling and Public Heath*, *Part A and Part B*. Elsevier, Amsterdam.

Rao, C. R. and Shaw, D. C. (1948). On a formula for the prediction of cranial capacity. *Biometrics*, **4**, 247–253.

Rao, C. R. (1965). On discrete distributions arising out of methods of ascertainment. In G.P. Patil (Ed.), *Classical and Contagious Discrete Distributions*, Calcutta: Statistical Publishing Society, 320–333. Reprinted in *Sankhya*, **A27**, 311–324.

Rao, C. R. (1985). Weighted distributions arising out of methods of ascertainment: What population does a sample represent? In Atkinson A.C., Fienberg S.E. (Eds.), *A Celebration of Statistics*: *The ISI Centenary Volume*. Springer, New York. An earlier version appeared as the Technical Report No. 84-38, July 1984, Center for Multivariate Analysis, University of Pittsburgh, Pittsburgh, PA 15260.

Rao, C. R. (1988). *Weighted and Clouded Distributions*. Technical Report No. 88-01, Center for Multivariate Analysis, University of Pittsburgh, Pittsburgh, PA 15260.

Zhang, X, Pyne, S. and Kedem, B. (2020a). Estimation of residential radon concentration in Pennsylvania counties by data fusion. *Applied Stochastic Models in Business and Industry*, 1-17.

Zhang, X., Pyne, S., Kedem, B. (2020b). Model selection in radon data fusion. *Statistics in Transition*, *New Series*, Special Issue, August, 167-174.

ANNEXURE

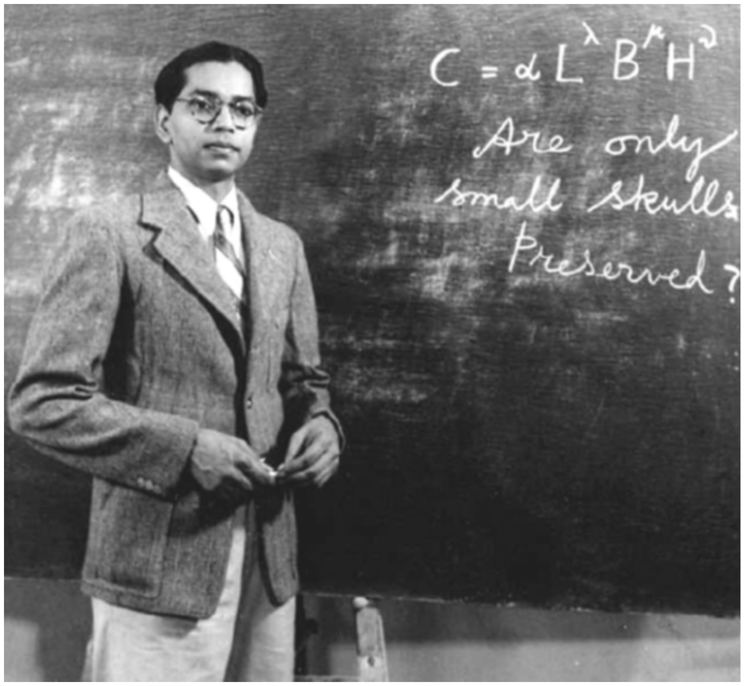

Figure 1: C.R. Rao in Cambridge during the 1940s. For discussion on the cranial capacity formula shown in the picture, see Rao and Shaw (1948).

Figure 2: C.R. Rao (standing 4th from the left) with family, friends, students in Pittsburgh, USA, during the 1980s. With P.R. Krishnaiah (standing 2nd from the left), Rao established in 1982 a unique Center for Multivariate Analysis at the University of Pittsburgh.

Figure 3:Inauguration of Data Science Laboratory, CR Rao AIMSCS, Hyderabad (December 19, 2013). From left to right: (late) Bhargavi Rao, C.R. Rao, Amartya Sen, Saumyadipta Pyne, S.B. Rao, Teja Rao (Inset: foundation plaque).