From Mahalanobis Distance to Fractile Graphs via Sample Survey

Probal Chaudhuri

Indian Statistical Institute, Kolkata

30th Mid-Year Meeting of Indian Academy of Sciences
June 29, 2019

Mahalanobis met Nelson Annandale at the 1920 Nagpur session of the Indian Science Congress. Annandale asked Mahalanobis to analyze anthropometric measurements of Anglo-Indians in Calcutta.

- Mahalanobis, P. C. (1922), 'Anthropological Observations on the Anglo-Indian of Calcutta. Part-1', Analalysis of Male Stature. *Rec. Indian Museum*, **23**, 1–96.
- Mahalanobis, P. C. (1930), 'On Tests and Measures of Group Divergence'. *Jour. Asiatic Soc. Bengal*, **26**, 541–588.

Gilbert Walker was the reviewer of Mahalanobis' 1922 paper, and he appointed him the Chief Meteorologist at Alipore Observatory in Calcutta. Mahalanobis held that part-time position during 1922–26 in addition to his full time job as a physics professor at Presidency College under Calcutta University.

"In 1923 I was working as Meteorologist in Calcutta in addition to teaching physics in the Calcutta University. Fisher was engaged in his researches on the design of experiments at Rothamsted Experimental Station, Harpenden. I had no connexion with agricultural research. By sheer chance, my attention was drawn to the question of 'errors' in some agricultural field experiments, in the form of a series of parallel plots sown with different varieties of rice, repeated in the same order in several blocks. I tried to eliminate, by crude graduation, differences in soil fertility, and published a paper [1925] in an agricultural journal. Fisher saw this paper and immediately sent me reprints of his early papers on the design of experiments and also the paper on the distribution of the ratio of two variances" (Mahalanobis, 1964)

"I also recall that it was at Fisher's suggestion (as I came to know much later) that the newly established Imperial (now Indian) Council of Agricultural Research offered me in 1928 an annual grant of Rs. 2,500 (a princely sum for us in those days) to have a research assistant to take up some work in statistics. This grant led the way to the future development of the integrated programme of theoretical research, training, and applied projects which has been a characteristic feature of the Indian Statistical Institute". (Mahalanobis, 1964)

"In March 1936, in reply to an enquiry from the Secretary, Imperial Council of Agricultural Research, I suggested exploring the possibilities of random sampling methods for estimating the area under different crops in Bengal" (Mahalanobis, 1939)

Contributions of Mahalanobis in the theory and methodology for large scale sample surveys :

- Optimum design of the surveys maximising precision at a given cost or minimising cost for assigned precision.
- Concept of Pilot Surveys.
- Inter-penetrating network of sub-samples (IPNS) for the assessment of sampling variations and the control of non-sampling error.

"... no technique of random samples has, so far as I can find, been developed in the United States or elsewhere, which can compare in accuracy or in economy with that described by Professor Mahalanobis." (Harold Hotelling in his report to Central Jute Committee in 1938)

"The occasional practice of designing a large scale experiment in successive stages may be regarded as a forerunner of sequential analysis. A very interesting example of this type is in the series of sample censuses of jute in Bengal carried out under the direction of P. C. Mahalanobis (1940)." (Abraham Wald)

Jute Survey of Bengal 1940

Mahalanobis' sample survey estimate of jute production was 7540 bales (1 bale = 400 lbs). It was obtained at a cost of 8 lakh rupees with a work force of 600. The plot-to-plot complete enumeration yielded a figure of 6304 bales at an expenditure of 82 lakh rupees and 33,000 employees, which turned out to be an underestimate by 16.6%. This was evident from the customs and trade figure of 7562 bales.

- Survey of rural households by National Sample Survey Office (NSSO) generates data on consumption and expenditure.
- Mahalanobis was looking into per capita total monthly expenditure for a household and the fraction of expenditure on some specific item (e.g., food).

Remarks by Mahalanobis :

- There may be important differences in the consumption patterns in the "rich" and the "poor" sections of the population.
- One has to define "comparable subsets" of the population such as "poor", "rich" and so on.
- Individuals in a subset will have "similar economic status" as defined by "ranks with respect to income".

Mahalanobis' Fractile Graph

- Mahalanobis divided the dataset into "fractile groups" according to per capita total monthly expenditure of different households.
- Then he plotted the average values of the fraction of expenditure on a specific item (e.g., food) for each fractile group.
- Finally, he joined the consecutive points in the plane by straight lines.
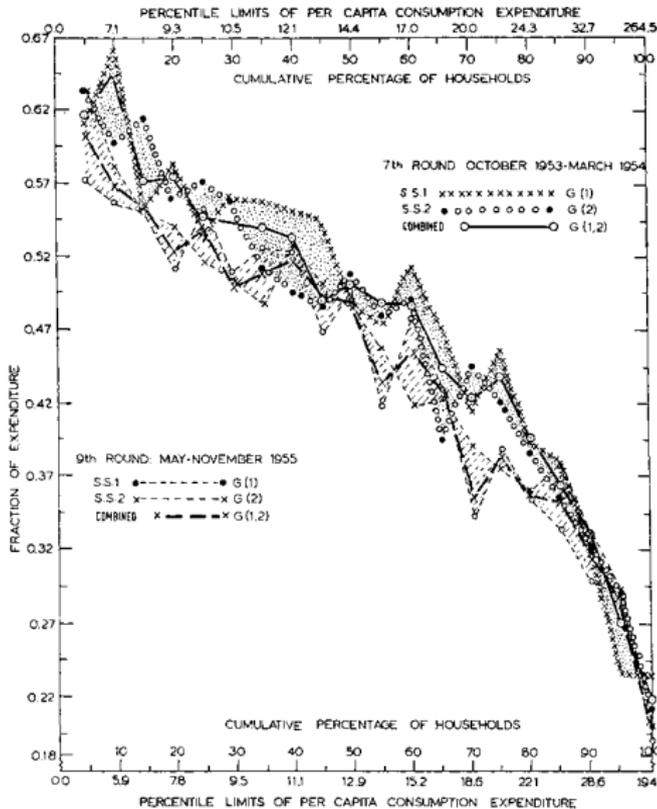
FIGURE 2

National sample survey: all India, rural.
capita expenditure on food grains as fraction of total expenditure p

ECONOMETRICA

JOURNAL OF THE ECONOMETRIC SOCIETY

*An International Society for the Advancement of Economic
Theory in its Relation to Statistics and Mathematics*

AN ISSUE IN HONOR OF

RAGNAR FRISCH

IN THE YEAR OF HIS SIXTY-FIFTH BIRTHDAY

CONTRIBUTORS:

K. ARROW	P. N. RASMUSSEN
J. S. CHIPMAN	D. A. ROWE
A. COWLES	P. A. SAMUELSON
N. GEORGESCU-ROEGEN	E. SCHNEIDER
H. S. HOUTHAKKER	R. M. SOLOW
L. JOHANSEN	J. R. N. STONE
L. R. KLEIN	R. H. STROTZ
T. C. KOOPMANS	H. THEIL
O. LANGE	J. TINBERGEN
P. C. MAHALANOBIS	G. TINTNER
M. MORISHIMA	H. O. A. WOLD

EDITED BY R. H. STROTZ AND E. MALINVAUD

Fractile Transformation of Covariate :

- Let F be the probability distribution of the covariate X.
- In Fractile Regression, we regress Y on $F(X)$, instead of regressing Y on X.
- $F(X)$ has uniform distribution on $[0, 1]$. So, in Fractile Regression, the covariate distribution is always uniform on $[0, 1]$.

the p coefficient ... associated with a c... gestational age in e... Z-score is not equiv... Russell, who estimat... the birthweight distr... tion to multivariate ...

Each model was ... dichotomous term fo... cans, 0 for whites); t... ($X_2 = 1$ for middle e... = 1 for high educat... product terms for ra... ($X_1 \cdot X_2$)]. We calcula... appropriate logits us... achieve the contrast... of African-American... educational stratum ... African-Americans t... tion but no more). ... in the standard way, ... covariances for each...

Example Results
Infant mortality rate... the crossover in mo... babies born in the e... whereas those born a... than European-Ame... graph, by itself, obsc... majority of births occ... American babies are ... 1b displays the distr... among all livebirths... long tail on the le... number of early bir... evident. Without ...
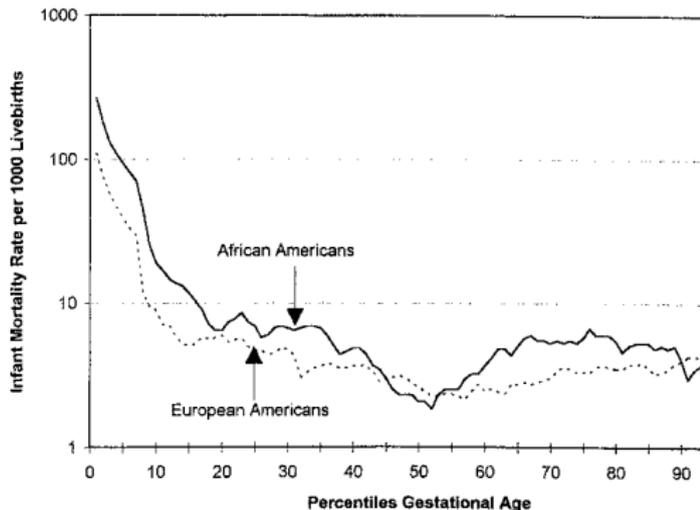
FIGURE 2. Log infant mortality rates plotted against

A conversation between Mahalanobis and noted sociologist Ramakrishna Mukherjee.

- Mahalanobis : "Ramakrishna, after my death what will happen to ISI? Would anything survive of what I created?"
- Ramakrishna Mukherjee : "Professor, your large scale sample survey will survive, D-square will live, fractile graphical analysis, though I don't quite understand it, will survive if it is useful, and some students of yours will spread your message.
- Mahalanobis : "Rabindranath used to say this is a land drained by a river and its tributaries, nothing survives in this climate for too long".