On Hierarchical Sampling, Hierarchical Variances and their connexion with other aspects of Statistical Theory

From a non-homogeneous univariate statistical population we can estimate the mean by a number of alternative methods one of which is a particular type of sampling recently proposed to be extensively used by P. C. Mahalanobis in crop-cutting experiments. Suppose the population in question is such that it could be cut up into a number of district zones (which let us call zones of the first order), each of these could be further cut up into a number of zones (which we call zones of the second order), each of the second order zones could be cut up further into a number of third order zones and so on till finally we get to zones of the $(k-1)$th order, each of which is a statistically homogeneous group. Suppose that the variance 'within' the $(k-1)$th order zones be σ_k^2, 'between' the $(k-1)$th order zones but 'within' the $(k-2)$th order zones be σ^2_{k-1}, 'between' the $(k-2)$th order zones but 'within' the $(k-3)$th order zones be σ^2_{k-2}, and so on till we come to the variance 'between' the 1st order zones which we denote by σ_1^2.

Our statistical population is said to be hierarchical if these variances $\sigma_1^2, \sigma_2^2, \ldots \sigma^2_k$ happen to be of entirely different orders. Now suppose further that from such a population we sample in the following manner: from the 1st order zones, we select at random n_1 zones, from each of these we select at random n_2 second order zones and so on and finally from each selected $(k-1)$th order zone, we pick out at random n_k individuals. We have altogether $n_1, n_2 \ldots n_k$ observations at our disposal; such a sampling has been called nested or hierarchical sampling by P. C. Mahalanobis. The mean of these observations is supposed to estimate the unknown population mean. The variance of the estimated mean could be easily shown (by the *Algebra of Mathematical Expectations*) to be given by

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_1 n_2} + \ldots\ldots\ldots + \frac{\sigma_k^2}{n_1 n_2 \ldots n_k} \qquad (1)$$

If we call our variate x_k and introduce pseudo-variates $x_1, x_2, \ldots x_{k-1}$ such that x_1 numbers, as it were, the first order zones, x_2 the second order zones and so on and finally x_{k-1} numbers the $(k-1)$th order zones and if further we assume a multivariate normal distribution for the variates $(x_1, x_2, \ldots x_k)$, then we have shown that formula (1) could be derived in an elegant manner. This derivation of (1) involves certain restrictions no doubt and thereby robs (1) a little of its generality but at the same time this process of derivation gives a greater insight into certain aspects of the multivariate normal distribution and incidentally gives a neat statistical interpretation of the rectangular statistical co-ordinates introduced by one of us jointly with Mr R. C. Bose in a paper[1] entitled "Normalization of Statistical Variates and the use of Rectangular Co-ordinates in the Theory of Sampling Distributions" published in 1937 in *Sankhya (Indian Journal of Statistics)*.

For a fixed set of values of $(x_1, x_2 \ldots x_{k-1})$, x_k is normally distributed about a mean M_{k-1} which is a function of $(x_1, x_2, \ldots x_{k-1})$ but with a variance which is independent of the particular values of this set. This variance is σ_k^2; again M_{k-1} itself is distributed for a fixed set of values of $(x_1, x_2, \ldots x_{k-2})$ about a mean M_{k-2}, which is a function of $(x_1, x_2, \ldots x_{k-2})$ with a variance which is independent of the particular values of this set; this second variance is σ^2_{k-1}; and thus it goes on till we get to M_1 which is a function of x_1, and which is normally distributed about a constant M_0 with a variance which is σ_1^2.

It is of considerable interest to note that these variances $\sigma_1^2, \sigma_2^2 \ldots \sigma_k^2$ with their direct statistical import could be completely identified with the rectangular (statistical) co-ordinates for the (multivariate-normal) population introduced in the paper referred to above. This furnishes us, with a simple statistical interpretation for those co-ordinates which was lacking in 1936, when these co-ordinates were first introduced as fruitful auxiliaries for certain investigations in Mathematical Statistics.

Statistical Laboratory, S. N. Roy.
Presidency College,
Calcutta, 8-8-1940. Kalishankar Banerjee.

[1] *Sankhya*, 3, 1, 1937.