Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

PRINCIPLES OF COMPOSITIONAL DATA ANALYSIS

BY JOHN AITCHISON

University of Virginia

Compositional data consisting of vectors of positive components subject to a unit-sum constraint arise in many disciplines, for example in geology as major-oxide compositions of rocks, in economics as budget share patterns of household expenditures, in medicine as compositions of renal calculi, in psychology as activity patterns of subjects. 'Standard' multivariate techniques, designed for unconstrained data, are wholly inappropriate and uninterpretable for such data and yet are still being commonly misapplied. Recognition that the study of compositions must satisfy simple principles has led recently to the advocacy of new forms of analysis of compositional data. The nature of the absurdities arising from applying traditional multivariate techniques to compositions is briefly highlighted and a description of the essential aspects and the advantages of the new methodology is provided.

1. Introduction: the Nature of Compositions. An alternative title for this paper could have been *Compositional Data Analysis is Easy*, though the history of the subject would hardly support this view. Almost a century ago Pearson (1897) warned us to beware of naive interpretations of correlations of his product-moment correlation $\operatorname{corr}(u_1, u_2)$, when u_1, u_2 are of the form $(u_1, u_2) = (x_1, x_2)/(x_1 + x_2 + x_3)$, that is when u_1, u_2 are essentially components of a composition. Statisticians and non-statisticians alike have largely disregarded the warning. A recent statistician-created disregard is in the software package Execustat Student Edition (1991) where the introductory tutorial unfortunately uses compositional data consisting of proportions of sand, silt and clay in sediments and refers to correlation coefficients of such proportions. For non-statistician examples the reader has only to browse through geological research journals abounding in arguments which depend on the interpretation of such uninterpretable correlation coefficients. For a detailed account of the sad history of compositional data see Aitchison (1986, Chapter 3).

Compositional data consisting of vectors of positive components subject

AMS 1980 Subject Classifications: Primary 62H, Secondary 62P.

Key words and phrases: Logratio analysis, logistic-normal distributions, perturbation, scale invariance, simplex sample space and subcomposition.