BIOLOGY

Deviant Index: a New Tool for Numerical Taxonomy

A MEASURE of the extent to which an individual differs from the norm of a population, taking many attributes into account, would often be useful in numerical taxonomy. With metrical attributes, the distance in an attribute space between the point representing the individual and the centroid of the population would serve such a purpose[1]. Similar measures have been proposed for binary attributes, but no fully appropriate way of combining binary and metrical attributes has been proposed, and no suggestion has been made for the inclusion of ordered non-metrical attributes except by allotting arbitrary metrical equivalents to the class values. Significance tests for these distance measures have proved elusive except in the multivariate normal case; but a significance test is often required as a guide to taxonomic action. The index proposed here not only combines information from all types of attributes, but also provides a direct test of significance.

When dealing with a metrical attribute, the natural choice for the norm is the arithmetic mean (unless the distribution is very skew, in which case one might, for example, prefer the geometric mean), and the deviant index for a particular value of such an attribute (as we will call the proposed index of deviation from the norm) is simply the proportion of the population further from the mean than the individual in question.

For ordered but non-metrical attributes, the appropriate choice for a norm seems to be the median value. Here, distance from the median has no meaning, but one may fit classes above and below the median together in accordance with the proportion of the population they contain. Thus, a class is considered to be further removed from the norm than a class on the other side of the median if, together with the tail of the distribution beyond it, it contains a smaller proportion of the population than the class in question with its tail. If the proportions in the successive classes are p_i $(1 \leqslant i \leqslant n)$, with m as the median class, the deviant index for this attribute for an individual with a value x $(< m)$ is thus:

$$D_x = \sum_{i=1}^{x} p_i + \sum_{i=y}^{n} p_i$$

where y is defined by:

$$\sum_{i=y}^{n} p_i \leqslant \sum_{i=1}^{x} p_i < \sum_{i=y-1}^{n} p_i$$

For a purely qualitative attribute, the only possible choice of norm seems to be the most frequent value. No natural ordering of classes around this norm suggests itself, but, as in the similarity index already proposed[4], it seems reasonable to use the proportion of the population in the class for this purpose, the rarer class values being considered to represent wider deviations from the norm than the commoner values. Thus, in this case:

$$D_x = \sum_{i \epsilon S} p_i$$

where:

$$S = \{i : (p_i \leqslant p_x) \wedge (1 \leqslant i \leqslant n)\}$$

When all the D_x for all m attributes have been calculated, they need to be combined to give an overall value $\mathfrak{D}_x$ for the deviation of individual x from the population norm. This can be done by Fisher's[2] method:

$$\chi^2_{2m} = -2 \sum_{i=1}^{m} \ln D_i$$

or the modification with a correction for continuity introduced by Lancaster[3]. $\mathfrak{D}_x$ is then defined as the probability of the χ^2 value obtained, or a greater one.

The analogy between this deviant index and the similarity index described earlier[4] is clear. The latter is the complement of the probability that a pair of individuals with a random selection of attributes would differ as greatly as the pair in question. This amounts to the complement of a deviant index calculated not as above for a single individual but for the deviations from the norm common to a pair. Generalization of this concept to larger sub-sets can be envisaged.

The deviant index proposed here could clearly be used for deciding whether, at a given significance level, a certain individual should or should not be included in a proposed taxon, or for a decision between two possible taxa to which an individual could be ascribed.

Programmes for computing $\mathfrak{D}_i$ for all individuals in a set have been written in *Fortran II* for the IBM 1620_3 computer, and are available on request.

I thank Mr. D. W. G. Moore for facilities in the Computing Centre of the University of Western Australia.

Note. While this communication was under editorial consideration, Hall's note[5] about a proposed 'peculiarity index' was published. He, too, was concerned to express the extent to which different individuals deviated from the norm of a set, and the purposes of the two proposed indices are very similar. The peculiarity index, however, covers binary attributes only, and is unstandardized, so that its magnitude depends on the number of individual (taxa) in the set under consideration. It may be defined as:

$$P = \frac{Nm}{2} - N \sum_{i=1}^{m} q_i$$

where N is the number of individuals (taxa), m the number of attributes, and:

$$q_i = \min\ \{\tfrac{1}{2}, p_i\}$$

p_i being the proportional frequency of the observed value of the ith attribute.

DAVID W. GOODALL

C.S.I.R.O. Division of Mathematical Statistics,
Western Australian Regional Laboratory,
Private Bag, P.O.,
Nedlands, Western Australia.

[1] Sokal, R. R., and Sneath, P. H. A., *Principles of Numerical Taxonomy* (W. H. Freeman and Co., San Francisco and London, 1963).

[2] Fisher, R. A., *Statistical Methods for Research Workers*, thirteenth ed. (Oliver and Boyd, Edinburgh and London, 1963).

[3] Lancaster, H. O., *Biometrics*, **36**, 370 (1949).

[4] Goodall, D. W., *Nature*, **203**, 1098 (1964).

[5] Hall, A. V., *Nature*, **206**, 952 (1965).

Present Status of Known Populations of the Vendace, *Coregonus vandesius* Richardson, in Great Britain

AT least thirteen different populations of *Coregonus* have been recorded in Great Britain and Ireland; four of these were of the Vendace, *Coregonus vandesius* Richardson. The status of these four populations was discussed by Regan[1], who divided them into two sub-species: the Lochmaben Vendace, *C. vandesius vandesius*, from two lochs near Lochmaben in Dumfriesshire; and the Cumberland Vendace, *C. vandesius gracilior*, from two lakes in Cumberland. The most recent revisions of the genus in Europe[2-4] agree that *C. vandesius* in Great Britain is equivalent to *C. albula* (L.) in Europe, though little evidence is brought forward to prove this. Practically nothing original has been published on the species in Great Britain since 1908 (ref. 1), and, in fact, other than very occasional specimens found accidentally at the water's edge, the species has not been taken in any numbers since the beginning of the century. Because of this, and the fact that several unsuccessful attempts have been made in recent years to capture specimens, many workers have assumed that this fish is extinct in Great Britain, or almost