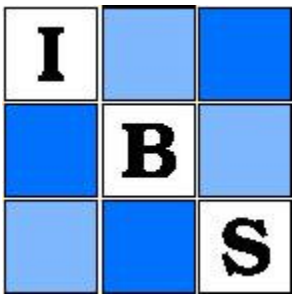


A METHOD FOR CLUSTER ANALYSIS

A. W. F. EDWARDS[1] AND L. L. CAVALLI-SFORZA

International Laboratory of Genetics and Biophysics (Pavia Section), Istituto di Genetica, University of Pavia, Italy.

SUMMARY

A method for investigating the relationships of points in multi-dimensional space is described. Using an analysis of variance technique, the points are divided into the two most-compact clusters, and the process repeated sequentially so that a 'tree' diagram is formed. The application of the method to problems of classification is particularly stressed, and numerical examples are given.

1. INTRODUCTION

A number of the problems with which we have been concerned recently require the consideration of the relationships between points in space. For example, the dispersion of human chromosomes as they appear on a photograph of mitosis involves a two-dimensional array of points; the dispersion of stars involves a three-dimensional array (when information on distance is available); in taxonomy, n species which have each been scored for m characters may be represented by n points in a space of m dimensions, and information on the natural groupings of the species may be obtained from the groupings of these points; and so with patients who exhibit the presence or absence of a number of symptoms.

Indeed, any investigation into the classification of objects, for which taxonomy may be taken as the type example, or any investigation into the dispersion of things, requires the elucidation of spatial groupings. Sometimes, as in taxonomy, the groups and sub-groups must be found, and at other times, as with chromosomes, the problem is more purely statistical—do certain chromosomes tend to cluster together? In this paper we concentrate on the former (and simpler) aspect of the problem, except for a slight digression in Section 8 on the dispersion of chromosomes, a matter of current interest. The latter, or probabilistic, aspect, with its vast literature, is not of immediate concern to us here.

We are aware of several investigations bordering on this subject, especially in the field of numerical taxonomy, where the papers have been reviewed by Sneath and Sokal [1962]. Rao [1952] has described an

[1]Present address: Department of Mathematics, Stanford University, Stanford, California, U.S.A.

intuitive approach to Cluster Analysis (using the generalised distance of Mahalanobis) which is particularly relevant, and will be mentioned below. However, our specific approach does not seem to have been considered before, so that a comprehensive review of previous work would be out of place.

Both Hotelling's [1933] Principal Component Analysis and Fisher's [1936] Discriminant Analysis are cognate with Cluster Analysis. Principal Component Analysis is, in effect, a method of projecting points in multi-dimensional space into a space of fewer dimensions so that the maximum amount of information is retained. The first component can be described as the orthogonal projection into one dimension that maximises the variance of the projected points; removing this dimension and its associated variance, the second component is the projection into one dimension that maximises the variance of the projected new array of points, and so on. Discriminant Analysis, on the other hand, seeks that linear combination of characters (to use taxonomy as an example) which best enables a species to be classified. It provides a sorting procedure into previously-chosen classes, and thus differs from Cluster Analysis, in which no such classes are admitted. However, a close affinity between the two analyses will become evident later.

2. GENERAL CONSIDERATIONS

Cluster Analysis, in the first instance, seeks to identify clusters of points in space. A prerequisite of a successful analysis is, naturally, a method of presenting the results, and we have found a 'tree' diagram useful. In this, as will be seen, each cluster of points is grouped on an individual branch. The peculiar advantage of this representation is that a multi-dimensional tree, reaching out to every point in the space, can be mapped, with all its ramifications, on paper in two dimensions. The analogy with a taxonomic tree will be evident. In particular instances it may be meaningful to assign lengths to the branches, and this will be mentioned below.

Starting at the base of the tree, the first bifurcation will represent the splitting of the points into two clusters; each branch so formed will then split again as the two clusters are themselves each resolved into two more clusters, and the process continues until the individual points are reached. Thus the main problem to be solved is how to divide a cluster of points into two in the most acceptable manner, for the repetition of this process will give the complete analysis.

3. THE DIVISION OF A CLUSTER

Clearly, the best division of a cluster for our purposes will be that in which the two resulting clusters are as dense as possible, according to

some criterion. We have found that the analysis of variance provides an excellent criterion, which is meaningful in operation, simple in calculation, and useful for testing significance. From the analysis of variance it is known that the sum of the squared distances of points on a line from their mean can be partitioned, when the points are classified into two groups, into two within-group sums of squares and a between-groups sum of squares (see Fisher [1954]). Further, since all the quantities involved are squared distances, it is evident that this is also true for points in any number of dimensions, because these squared distances can all be partitioned into squared distances along the Cartesian axes, so that if the partition is possible along each axis it is possible amongst the points as a whole. Thus we see that when points are divided into two clusters the sum of the squared distances from their mean can be partitioned into the sum of the squared distances of the points of one cluster from *their* mean, the similar sum for the other cluster, and the between-clusters sum of squares; this is but a single-classification analysis of variance conducted in many dimensions. The natural criterion for division is clearly the between-clusters sum of squares, and we will regard the best split as that for which this sum is a maximum (and the within-clusters sum of squares consequently a minimum). Continued splitting according to this criterion will lead to a tree diagram, and with each branching will be associated a between-clusters sum of squares, which will be a measure of the 'importance' of the split. Further, since at the end of this process each cluster contains only one point, there is no within-clusters sum of squares left, and the total of the sums of squares associated with each branching must exactly equal the original sum of squares: all the original variation is accounted for. Table 2, which will be referred to below, exemplifies this type of hierarchical breakdown of the total variation. The natural measure of cluster density is clearly the variance, or the within-cluster sum of squares divided by the number of points (we divide by n rather than $n - 1$ because we are not primarily concerned with estimation).

This criterion of maximising the between-clusters sum of squares is essentially that used in Discriminant Analysis (Fisher [1936]), but whereas we use it to determine the natural clustering of objects from raw observations, Discriminant Analysis uses it to determine what linear function of the observations best discriminates between two *a priori* clusters.

4. THE CALCULATION OF THE SUMS OF SQUARES

This technique is as simple in practice as it is in theory. Since n points can be divided into two clusters in $2^{n-1} - 1$ ways, and the sum of

squares must be calculated in each case (at the present state of our knowledge), it is essential to have a quick method of performing the calculation. In fact it is simpler to minimise the within-clusters sum of squares than to maximise the between-clusters sum of squares. Our method depends upon the fact that the variance of n points on a line is equal to the sum of all the squared distances between points taken two at a time (but each distance being used once only), divided by n^2. Thus the sum of squares about the mean is equal to the sum of the squared distances between points, divided by n. As before, since we are dealing exclusively with squared distances, the result must hold for points in any number of dimensions. Thus all the necessary information for *all* the splits is contained in the half-matrix of the squared pair-distances. This is reasonable, for the half-matrix clearly contains all the information necessary to reconstruct the cluster as far as the relative positions of the points are concerned; what it does not contain is information about the axes—these are, of course, no longer relevant, and have been dropped.

Such a matrix is given in Table 1, the numbers (which are in reality

TABLE 1

THE HALF-MATRIX OF SQUARED DISTANCES (DERIVED FROM THE DATA OF TABLE 3)

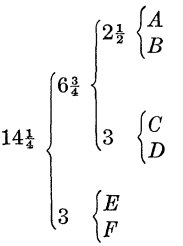
A	*B*	*C*	*D*	*E*	*F*	*Points*
5	11	11	14	14		*A*
	10	6	13	15		*B*
		6	17	21		*C*
			13	15		*D*
				6		*E*
						F

the squares of distances) being taken as integers for the sake of the example: thus the distance between A and B is $\sqrt{5}$ units. This matrix may be taken as an arbitrary numerical example for the moment; in fact it is derived from real data which will be mentioned below. The sum of the numbers is 177 so that the sum of squares is 177/6, or $29\frac{1}{2}$. Investigating the split $ABC : DEF$, the sum of squares for the first cluster is $(5 + 11 + 10)/3$, or $8\frac{2}{3}$, and for the second cluster is $(13 + 15 + 6)/3$, or $11\frac{1}{3}$. The total within-clusters sum of squares is therefore 20. All the other thirty possible splits can be similarly treated, and the best one chosen, the between-clusters sum of squares measuring the 'importance' of the split. In this case the best split is $ABCD : EF$ with a within-clusters sum of squares of $15\frac{1}{4}$, leaving a between-clusters

sum of squares of $14\frac{1}{4}$. Splitting the clusters so found, the final tree is formed; this is given in Table 2, together with the between-clusters sum of squares for each split. The total of these values is, of course, $29\frac{1}{2}$.

TABLE 2

AN EXEMPLARY TREE



This technique is clearly well-suited to electronic computing (which is necessary owing to the number of trials involved), especially as the $2^{n-1} - 1$ splits of n points into two clusters may be characterised by the binary numbers from 1 to $2^{n-1}-1$. In our experience about $(n-1)^2 2^{n-11}$ seconds of computing time are needed to split n points when the squared distances run to two significant figures, so that up to 16 points can be treated in a reasonable time. Thereafter a sequential method is necessary to avoid testing all the splits, and this has also been developed. It is interesting to note that we are limited by computer speed rather than computer capacity. (The above computing time was obtained on an Olivetti Elea 6001, which we are using, and which has an Access Time of 5μ-seconds). A computer is also useful, of course, for calculating the $\frac{1}{2}n(n-1)$ squared distances between points from the coordinates, except in cases of direct measurement on small number of points. We may add that an algorism for finding the best split directly would be very welcome.

It is encouraging to note that when all the points are equidistant one from another, no split can be preferred by our method: if the distance between any two points is x, then splitting n points into clusters of r and $n - r$ will give a within-clusters sum of squares of

$$\tfrac{1}{2}r(r-1)x^2/r + \tfrac{1}{2}(n-r)(n-r-1)x^2/(n-r) = \tfrac{1}{2}(n-2)x^2,$$

which is independent of r. Indeed, for *any* array of points this formula must give the mean within-clusters sum of squares for all the possible splits into clusters of r and $n - r$, if x^2 is taken to be the mean of the squared distances. For, considering all the possibilities, all the squared distances will occur within a cluster with equal frequency. Further,

since the total sum of squares is equal to $\frac{1}{2}(n-1)$ times the mean square distance x^2, the mean within-cluster sum of squares must be $(n-2)/(n-1)$ times the total sum of squares, giving a mean between-clusters sum of squares of $1/(n-1)$ of the total. Since this is independent of r, it must also be the general mean over all possible splits. It can provide a datum from which to gauge the 'importance' of the best split.

5. 'TAXONOMIC' CONSIDERATIONS

In 'taxonomic' situations, in which a number of objects have been scored for the presence or absence of various characters (each accorded equal weight), all the coordinates of the points are 0 or 1, and the calculations become even simpler: since these numbers are unchanged by squaring them, the squared distance between two points is equal to the number of characters in respect of which the two objects differ, In taxonomy itself, this measure of 'distance' is already in use, so that our methods are particularly appropriate; of course, they may also be used with metric characters without difficulty.

In fact the example given above was taken from the bacteriological data of Lysenko and Sneath [1959]. The six organisms were *Escherichia Coli*, *Salmonella*, *Klebsiella*, *Hafnia*, *Proteus vulgaris* and *Morganella*, in that order, and we have omitted the results of the tests 'gas from glucose', 'sucrose', 'dulcitol', 'inositol', 'salicin', 'raffinose' and 'trehalose', which gave variable results in many cases. Information from such tests can in fact be used without difficulty, but we wish here to preserve the simplicity of the example. The consequent table of scores, from which the half-matrix of squared distances was constructed, is given in Table 3.

TABLE 3

MATRIX OF SCORES FOR SIX SPECIES OF BACTERIA

A	111011111000110010001111
B	101011111010010110001110
C	011111111000001111011000
D	101001111010001111011010
E	100001010111110111100001
F	100000000011110001100011

In some instances it may be informative to classify the characters according to the objects, as well as the objects according to the characters: this will give an idea of the independence of the characters, a sub-

ject which will be taken up in Section 7. Finally, having identified several 'taxa' amongst the objects, Discriminant Analysis can be used to determine the most informative characters for the classification of future objects into these 'taxa'. But here a word of warning is necessary, for Discriminant Analysis provides a 'closed system' to which further objects do not contribute any information; it is a synthesis of past experience which does not 'learn' from further cases.

6. SCALE AND DISTANCE

Cluster Analysis is not invariant with respect to changes of scale along the axes, and the user must satisfy himself that his scales are appropriate. Conversely, in 'taxonomic' uses, a character may easily be weighted by changing the scale along the corresponding axis. In these uses the independence of the classifying characters can also be considered, and will be treated in the next section.

From the between-clusters sum of squares (SS) it is easy to calculate the distances from the means of the two clusters to that of the combined cluster, and these distances may be associated with the corresponding segments of the tree. They are given by

$$d_m^2 = \frac{nSS}{m(m+n)} \quad \text{and} \quad d_n^2 = \frac{mSS}{n(m+n)},$$

where the clusters contain m and n points. In this way a metric tree can be found, if it is meaningful for the problem in hand. For fuller discussion of metric trees, especially in evolution, see Edwards and Cavalli–Sforza [1964] and Cavalli–Sforza and Edwards [1964].

It is evident that our technique of Cluster Analysis by sequential splitting is more appropriate to some types of problem than to others, and the same may be said of the resultant tree diagram. The method is well-suited to cases in which the array of points might indeed have arisen by a branching process, such as in evolutionary studies or in some studies on the spatial distribution of organisms; in other cases it is merely a way of reducing the data and detecting clusters. In the former cases the assignment of some point, almost mid-way between two clusters, to a particular cluster indicates the most probable origin of the point on the available information, but in the latter cases there can be no such interpretation, and a tree diagram is correspondingly less informative. Nevertheless, Cluster Analysis is the best method we know for treating such data, and no difficulties should arise as long as its limitations are clearly understood. In particular, in cases in which a prior branching process cannot be postulated, the interpretation of the tree needs special care: two points, one from each of two clusters, which lie at the positions of closest approach of the clusters, may be widely

separated on the tree. It hardly needs adding that the linear arrangement of the points on the tree is largely arbitrary, although this can be rectified to some extent, for example by arranging them according to the first component of a Principal Component Analysis (see e.g. Rao [1952]).

7. INDEPENDENCE, AND A FURTHER EXAMPLE

In many situations to which Cluster Analysis may be applied, each point will represent not one observation, but the mean of many. Thus, in the following example from Rao [1952], each point represents a particular Indian caste or tribe, and was derived from the measurement of many individuals. This will not be so in pure taxonomy, where each organism is scored for the presence or absence of each of many characters, and similar organisms will be scored identically, unless metrical characters are used.

Thus examples may be classified into those with a 'within-points' dispersion matrix, and those without. In most cases with a 'within-points' matrix it will be desirable to eliminate the correlations between classifying characters by a transformation to uncorrelated standardised variables. However, no general rules can be laid down, and each case must be treated on its merits. Indeed, there is no hard and fast distinction between the two types: we may look upon data having a 'within-points' matrix as a series of individual points on which *a priori* clustering has been carried out, and similarly, after clusters of acceptable size have been isolated from data having no 'within-points' matrix, the within-clusters dispersion matrix becomes the 'within-points' matrix of the points representing the cluster means. To make Cluster Analysis fully comparable to Discriminant Analysis it would be necessary, for every possible split, first to transform according to the within-clusters dispersion matrix. The 'most-discriminating' split would then be chosen. We have not pursued this approach.

In simple taxonomic situations, characters will usually be correlated 'between-points'. Such correlations can be eliminated by an orthogonal transformation which rotates the reference-frame of axes bodily, but this leaves the relative positions of the points unchanged, and is therefore irrelevant to Cluster Analysis. However, it may in certain circumstances be thought desirable to standardize the resultant variables for variance: this is tantamount to changing the scales along the new axes so that all the components of the Principal Component Analysis (which gives these axes) have the same variance, and may be accomplished by calculating the values of D^2 (see below) using the 'between-points' dispersion matrix.

The technique of transforming to a set of uncorrelated standardized

variables by using the within-groups dispersion matrix is well-known (for example, see Rao [1952]), and there is no need to describe it here. As has been mentioned above, such a transformation consists of an orthogonal rotation of the axes, followed by changes of scale along the new axes: the old axes thus become oblique to one another, and all the distances between points are affected, in general. Cluster Analysis must be carried out on the new distances. The use of such a transformed distance was first proposed by Mahalanobis, and it is usually referred to as 'Mahalanobis' generalized distance', or 'D'. Rao gives a full numerical example, calculating the values of D^2 between twelve castes and tribes of the United Provinces in India, using nine anthropometric characters. Since these values of D^2 are squared distances in the transformed space they can be used, without modification, in Cluster Analysis. They are given in Table 4.

Rao clusters the castes and tribes using largely intuitive methods, and follows this with a Principal Component Analysis. He comments that 'No formal rules can be laid down for finding the clusters because a cluster is not a well-defined term'. It is thus interesting to compare his results with those obtained by the 'rules' of Cluster Analysis. Of course, with only twelve castes an intuitive solution is not difficult. The clusters that Rao finds are given in Table 5, the figures being the *within-cluster* sums of squares expressed as percentages of the total

TABLE 4

VALUES OF D^2 FOR TWELVE INDIAN CASTES AND TRIBES (AFTER RAO)
Note: These values have been derived from Rao's by multiplying his by a constant so that the largest value becomes 99; in this way the greatest possible accuracy using two figures (as required by our computer program) is attained.

B_1	B_2	C_1	C_2	D	Bh	Ch	M	A_1	A_2	A_3	A_4		
	5	65	42	54	84	57	54	22	28	40	62	B_1	Brahmin, Basti
		68	31	53	72	54	49	15	19	28	51	B_2	Other Brahmin
			25	85	96	99	84	50	56	63	79	C_1	Bhatu
				40	46	88	70	24	29	31	54	C_2	Habru
					22	72	46	55	45	43	50	D	Dom
						94	59	48	42	33	42	Bh	Bhil
							8	64	40	51	42	Ch	Chattri
								46	25	27	17	M	Muslim
									6	9	29	A_1	Ahir
										2	11	A_2	Kurmi
											8	A_3	Other Artisan
												A_4	Kahar

TABLE 5

CLUSTERING OF TWELVE INDIAN CASTES AND TRIBES, AS FOUND BY RAO

Caste or tribe	Within-cluster SS (% of total SS)
Ahir, Kurmi, Other Artisan, Kahar	6.48
Chattri, Muslim	1.60
Dom, Bhil	4.39
Brahmin, Basti, Other Brahmin	1.00
Bhatu, Habru	4.99.
	18.46

TABLE 6

'TREE' OF TWELVE INDIAN CASTES AND TRIBES

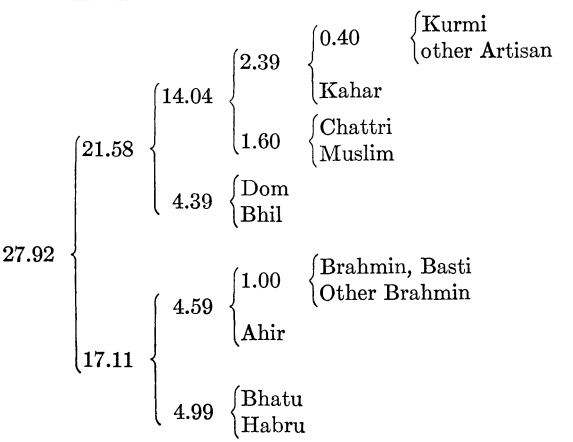


TABLE 7

CLUSTERING OF TWELVE INDIAN CASTES AND TRIBES, AS FOUND BY CLUSTER ANALYSIS

Caste or tribe	Within-cluster SS (% of total SS)
Kurmi, Other Artisan, Kahar	2.79
Chattri, Muslim	1.60
Dom, Bhil	4.39
Brahmin, Basti, Other Brahmin, Ahir	5.59
Bhatu, Habru	4.99
	19.36

sum of squares. The sum of these values is 18.46%, so that 81.54% of the variation has been accounted for by this arrangement, which is evidently very successful. Cluster Analysis, on the other hand, leads to the tree given in Table 6, the between-clusters sums of squares being expressed as percentages of the total. It will be seen that the four major splits account for 80.65% of the total variation; the five clusters so derived are given in Table 7.

As opposed to Rao, we have clustered the Ahir with the Brahmin, but by so doing the percentage of variation accounted for drops from 81.54 to 80.65. But here is no paradox, for there is nothing in our method which ensures that, after several splits, the between-clusters sum of squares shall be an absolute maximum. It is so after one split, but to find the absolute maximum after several splits it would, in general, be necessary to consider all the possible arrangements. Now the number of ways of arranging n different objects into exactly r groups is $n!/r!$ times the coefficient of x^n in the expansion of $(e^x - 1)^r$ (see Whitworth [1901]). When $n = 12$ and $r = 5$, as in the present case, there are thus 1,379,400 ways, or far too many to be tried in a reasonable time. With $r = 3$ groups there are $\frac{1}{2}(3^{n-1} - 2^n + 1)$ ways, or already 28,501 for eleven objects, compared with 1023 splits into

two groups. A systematic investigation of multiple splits is thus seen to be impossible except with very small numbers of objects, and our method avoids it.

Some further information about the split into two clusters can be obtained by considering the ten best splits, rather than just the single best one. A computer takes little additional time to record these, and it is interesting to compare the difference in the within-clusters sum of squares between the best split and the second best, and the mean difference between the second and third, third and fourth, ... ninth and tenth splits. If the former difference exceeds the latter by an appreciable factor, some confidence may be had in the best split. In the present example the difference between the first two splits is 2.48 (in arbitrary units), and the mean difference amongst the others 0.88. The first three splits all group the Ahir with the Brahmin, Bhatu and Habru, the second adding the Kurmi as well, and the third the other Artisans. The difficulty of knowing to which cluster the Ahir should be ascribed is well-illustrated in Rao's Principal Component chart (his Figure 2 [1952] p. 370). Had there been little to choose between the first few splits, the consequences of each could have been followed through without difficulty, for subsequent splits take a negligible time compared with the first.

It seems to us that both the power and some of the limitations of Cluster Analysis are evident in this example.

In passing, it should be noted that Rao uses 'average within-cluster D^2' as a measure of cluster density: this is $2n/(n-1)$ times our variance; and that he uses the square root of the average between-clusters D^2 as a measure of distance, whereas we prefer the 'real' distance between the means of the clusters, defined by

$$d^2 = \frac{m+n}{mn} \text{ (between-clusters } SS),$$

which is independent of the variances of the two clusters.

8. TESTING A NULL HYPOTHESIS

As was mentioned in the introduction, in some problems we wish to test for the tendency of points, chosen *a priori*, to lie close together. For example, with human chromosomes we might wish to test whether homologous chromosomes tend to lie closer together (on a photograph) than chance would lead us to expect. Statistical procedures have been suggested by Barton and David [1962] and Schneiderman and Smith [1962], the former authors actually using the ratio of twice the within-homologues sum of squares to the total, the moments of whose sampling

TABLE 8

ANALYSIS OF VARIANCE OF CHROMOSOME DISPERSION

	df	*SS*	Variance
Between homologues	22	322.04	14.64
Within homologues	23	184.04	8.00
Total	45	506.08	

$F = 1.83$, $10\% > P > 5\%$.

distribution they investigate. But it seems to us that a straightforward analysis of variance (as Barton and David indeed suggest in their final paragraph, but do not use) is appropriate. Using their example, we obtain the analysis given in Table 8. It will be noted that significance at the 5% level is not quite attained, although it is in Barton and David's test. We might add that, when considering the dispersion of real objects, the fact that two or more objects cannot occupy precisely the same point must introduce a bias into the test.

9. EPILOGUE

Apart from its use in investigating the spatial relationships of real objects, Cluster Analysis lends itself to any problem involving classification. It searches for hidden similarities, and sorts items into abstract groups. Such problems occur in nearly all scientific disciplines, and we feel that our approach may provide a fruitful point of view.

ACKNOWLEDGMENT

The work on which this paper is based was supported by a grant from the U. S. Atomic Energy Commission.

REFERENCES

Barton, D. E. and David, F. N. [1962]. The analysis of chromosome patterns in the normal cell. *Ann. Hum. Genet., Lond. 25,* 323–29.

Cavalli-Sforza, L. L. and Edwards, A. W. F. [1964]. Analysis of human evolution. *Proc. XI Int. Congr. Genetics,* The Hague, 1963.

Edwards, A. W. F. and Cavalli-Sforza, L. L. [1964]. Reconstruction of evolutionary trees. In *Systematics Association Publ. No. 6: Phenetic and Phylogenetic Classification.*

Fisher, R. A. [1936]. The use of multiple measurements in taxonomic problems. *Ann. Eugen., Lond. 7,* 179–88.

Fisher, R. A. [1954]. *Statistical Methods for Research Workers.* 12th edition. Edinburgh: Oliver and Boyd.

Hotelling, H. [1933]. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych. 24,* 417. (Quoted by Kendall [1955])

Kendall, M. G. [1955]. *The Advanced Theory of Statistics.* 2nd ed. London: Griffin and Company.

Lysenko, O. and Sneath, P. H. A. [1959]. The use of models in bacterial classification. *J. gen. Microbiol. 20,* 284–90.

Rao, C. R. [1952]. *Advanced Statistical Methods in Biometric Research.* New York: John Wiley and Sons.

Schneiderman, L. J. and Smith, C. A. B. [1962]. Non-random distribution of certain homologous pairs of normal human chromosomes in metaphase. *Nature, Lond. 195,* 1229–30.

Sneath, P. H. A. and Sokal, R. R. [1962]. Numerical taxonomy. *Nature, Lond. 193,* 855–60.

Whitworth, W. A. [1901]. *Choice and Chance.* 5th edition. Reprinted 1951: New York: Hafner Publishing Company.