CLUSTER ANALYSIS

by

Anatol W. Holt

B.A., Harvard University

(1950)

Submitted in Partial Fulfillment of

the Requirements for the

Degree of Master of Science

from the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Author _____

Professor in
Charge of Thesis _____

Chairman, Dept. Comm.
on Graduate Students _____

Math
Thesis
1953

CLUSTER ANALYSIS

by

Anatol W. Holt

Submitted to the Department of Mathematics
on June 25, 1953 in Partial Fulfillment
of the Requirements for the
Degree of Master of Science

Abstract

This paper deals with a problem of data analysis, and as such, with a problem in statistics. On the basis of certain general considerations concerning procedures of observation, a thesis concerning "clusters" of data and their significance is developed. In the light of this, a procedure is described by means of which a set of data may be analyzed for clusters. A small example of such an analysis is provided in appendix I. In appendix II some heuristic remarks are made concerning the possible application of cluster analysis to linguistic data. It is envisioned that the concepts developed in this paper will be most useful in scientific areas where the fundamental invariants are related to observation procedures which require large masses of data for single determinations.

ACKNOWLEDGEMENTS

I am particularly grateful to Noam Chomsky, Junior Fellow in Linguistics at Harvard, who aroused my interest in this problem and whose patient criticisms and helpful suggestions were invaluable. I am similarly indebted to William Turansky, mathematician at the Eckert-Mauchly Computer Division of Remmington Rand.

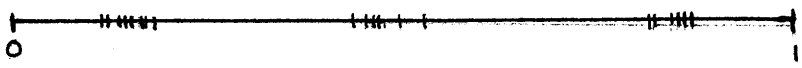
I am much indebted to the Electronics Research Laboratory at M.I.T., and to Remmington Rand, for support in course of the work..

TABLE OF CONTENTS

Introduction

Statement of the Problem

Suppose someone showed you a set of numerical "results" (represented below by figure 1) which he had obtained by repeating a well specified procedure of observation and calculation under stated conditions.

Fig. 1 

The figure is not only supposed to represent the numbers which he actually got, but is also supposed to suggest that the range of numbers which he could possibly have gotten is the interval from 0 to 1. Knowing nothing more specific about the procedure or the conditions under which it was performed, you will undoubtedly remark that the observations fall into three groups - i.e., that they form three clusters. Noticing such a configuration of data is almost inevitably coupled with the expectation that there is an "explanation" for it. If, for example, the data came from a well conducted scientific experiment, you will suspect that there are three discoverable conditions which distinguish between the various repetitions of the experiment.

Our problem is this: we wish to explicate the term "cluster" as it is used in the preceding paragraph. The

explication is to be such that the judgement of "three clusters" attributed to the reader will be obtainable as the result of a well specified calculation.

History of the Problem

So far as I have been able to discover, no attempts have been made to formulate the notion of "clustering" precisely. There are various things in the literature of statistics which are remeniscent of "clustering - i.e., the analysis of variance (with factor analysis as a special variety of this)[1] and so-called discriminant functions which are constructed to enable one to distinguish with maximum reliability between members of what one supposes to be two distinct populations.[2] In addition to this there is a book by Prof. R. C. Tryon,[3] called "Cluster Analysis", which describes certain procedures which are of use in discovering "clusters" of profiles - profiles being a set of results from psychological tests applied to some individual. As with factor analysis, these clusters are supposed to be evidence of essentially independent mental faculties which are fewer in number than the number of tests which define the profiles. Prof. Tryon does not, however, give us a definition of what a "cluster" is, and his procedures of handling the data are imprecise and without justification beyond the "usefulness" of the results of analysis - the latter being, in turn, very difficult to evaluate.

1. See M. G. Kendall, The Advanced Theory of Statistics, Vol II, pp. 175-246, London, 1948
2. ibid., pp. 341-348
3. R. C. Tryon, Cluster Analysis, An Arbor, Mich., 1939

Heuristics

In looking at "clusters", it seems evident from the start that they are groups of points which lie close to each other in comparison with the distances between points belonging to different groups. If, without precise definition, we call the former "internal distances" and the latter "external distances", then it would seem that clusters are groups of points which make internal distances small while making external distances large. One observes immediately however, that these are not independent conditions. We can always find an analysis of the data which makes the internal distances as small as we please, simply by making the number of groups large enough. This will, however, be at the expense of making external distances small also. This raises a question which, in its following form, is, I think, unanswerable, namely: how much should one be willing to increase the internal distances for a given increase in external ones. The procedures of Prof. Tryon depend on some ad hoc decisions as to how much of one is worth how much of the other. I believe that any procedure which is based on some optimization of internal and external distances with respect to each other, is bound to involve arbitrary judgements which, however, are not intrinsic in what is sought for. There are some heuristic arguments which make this latter view plausible. Consider, for example, two two-dimensional "result spaces" (a result space being the

collection of all conceptually possible results of a given procedure of observation) and a set of sixteen results in each, represented below by figures 2 and 3.

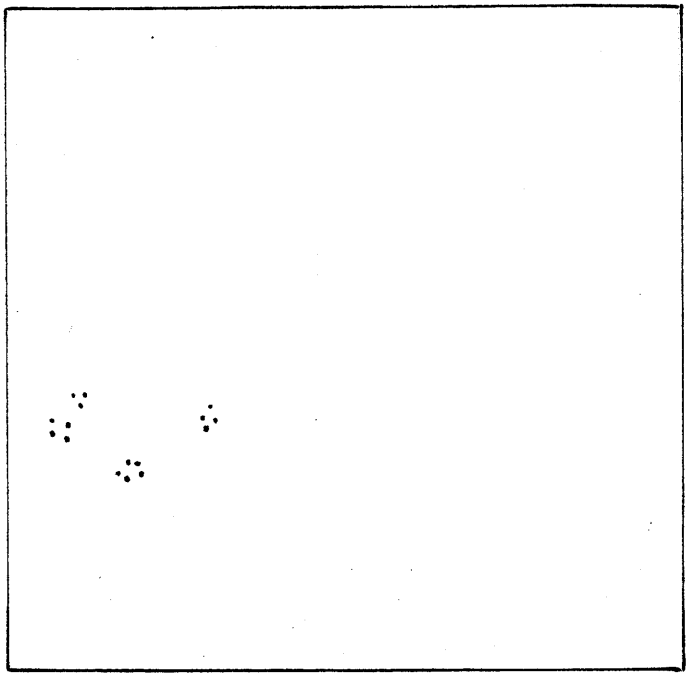


Fig. 2

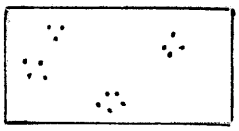


Fig 3

If you consider the data in figure 2, you might first observe that all of the points lie quite close to each other, i.e., form one cluster. Or perhaps you would be more struck by a trichotomy, one of the three resulting groups displaying some sub-structure in turn. In figure 3 the only grouping which seems at all clearly indicated, is a division in four. Measured in absolute units, the distance relations among the points are essentially the same in figure 2 and in figure 3. This example is supposed to point to the fact that how a set of data is clustered, depends not only on the distance relations among the points, but on the entire result space in which this set of data is located. It goes without saying that

any definition which depends upon some relative adjustment of internal and external distances could not accomodate a case which one would naturally regard as a single cluster since, in such a case, there would be no external distances.

Content and Organization of this Paper

Section one leads from a general discussion of data interpretation to a statement of the clustering problem, and to a general sketch of a proposed method of solution. It must be recognized however, that in point of history, the particular method of solution was discovered first and the general framework was investigated later. There are still some gaps between the two which have not been properly filled. Section one is built around an example which the reader is begged to regard as a skeleton for a general discussion, and hence to forgive idealizations to which some of the matter has been submitted; also to recognize that most of the remarks of general significance will be couched in terms of the example, and will not be repeated in "pure" terms.

The second section is a development of the proposed technique for isolating clusters in a set of data.

There are two appendices. The first gives a concrete example of how the method is applied to sets of five points on the unit interval. The second appendix gives a sketch of how the method might be applied to a scientific problem, namely the discovery of grammatical categories in a language by analyzing relations of co-occurance in a large body of data.

Words and Phrases

Procedure of Observation

A procedure of observation, V, is specified if we state the most general class of entities to which it is applicable (such a class of entities will be called a "universe" and denoted by "X"), the space of its conceptually possible results, $\emptyset_V$, (called the "result space"), and through what operations a given entity of X is related to a given result of $\emptyset_V$. Viewed abstractly, V is simply a function whose domain is X and whose range is $\emptyset_V$.

Partition of a Set, S.

A partition of S is a family of subsets of S which are mutually exclusive and whose sum is S.

Partition of a universe, X, induced by a procedure of Observation, V

Since V is a function, the subsets of X such that each is obtained by taking all the inverse images under V of a fixed result, constitute a partition of X. A subset, thought of as a member of a partition, will be called a "part".

Measure induced by V over a subset Y, of X, on $\emptyset_V$

Suppose that there is a probability measure defined on X. Given a subset $Y \subset X$, we obtain a measure m, on $\emptyset_V$, relative to Y, by the following definition.

For any $A \subset \emptyset_V$, $\quad m(A) = Pr(V^{-1}(A) \cap Y)$

The family of procedures derived from V

Every partition, P, of $\emptyset_V$ yields a derived procedure, V', on X, in the following manner. The result space of V' is taken

to be P (so that the results of V' are subsets of $\emptyset_V$), and V' carries an entity of X into that part of $\emptyset_V$ in which lies the image of the entity under V. The family of derived procedures has as many members as there are partitions of $\emptyset_V$, and if one includes the partition which has as many parts as there are points in $\emptyset_V$, then V itself belongs to the family of procedures derived from V.

A thing definable by a stock of procedures

A "thing" is any subset Y, of a universe. Now suppose we have a stock of procedures which have a universe X as their common domain. Consider all the subsets of X which are parts of partitions induced on X by any of these procedures. The subset Y is definable by the given set of procedures if it is obtainable as the boolean sum of products of the parts of X mentioned above.

Procedure V is dependent on a stock, $\mathfrak{S}$, of procedures

Given $\mathfrak{S}$, V is dependent on $\mathfrak{S}$ if all the things definable by V are already definable by the procedures of $\mathfrak{S}$. (It is trivial that if V is dependent on $\mathfrak{S}$, then so is the entire family of procedures derived from V). We shall also have occasion to speak of V as being "near-dependent" on $\mathfrak{S}$. By this we mean that for every thing, Y, definable by V, there is a thing Y' definable by procedures of $\tilde{\mathfrak{S}}$ such that $Pr((Y - Y') + (Y'-Y))$ is small..

Procedure V is an approximation to procedure R

V is an approximation to R, if V and R are mutually near-dependent. If V and R are mutually dependent, we may say that V coincides with R. A result of V and a result of R, whose respective inverse

are identical, or nearly identical (measure of the symmetric difference small), will be called corresponding results.

Section I

A traveller riding through a certain country, C, notices that all of the rooftops are painted in one of two colors - red or green - and asks himself if there is any explanation. Let us consider how the question was brought about. If he had reasoned explicitly it might have been something like this:

S 1.1 "Without any a priori assumptions about the world, there is no reason to suppose that rooftops should be of one color any more than of another, but here only two among many possibilities are realized. This calls for an explanation."

The first sentence of S 1.1 is closely related to a common type of statistical statement, namely:

S 1.2 "Given a distribution which represents my null-hypothesis, I am confronted with a sample which leads me to reject the null-hypothesis."

We shall, in fact, cast the first sentence in just this form, by supplying a distributional interpretation of "Without any a priori assumptions about the world there is no reason to suppose that rooftops should be of one color any more than of another..." in the following manner. We replace the absence of a priori assumptions by an assumption, namely: that if the

results of observation by means of a certain procedure teach us anything that is "noteworthy", it is that the results are not uniformly distributed - i.e., that, within a certain broad frame of reference, the "figure" of statistical structure which we observe, appears against the "ground" of uniform distribution. Stated more precisely: suppose we are given a procedure of observation, V, over a universe X', with result space $\emptyset_V$. We imbed X' in a larger universe, X, with a probability measure which is such that the measure induced by V over X on $\emptyset_V$ is uniform, i.e., that all of the conceptually possible results of V are obtained with equal frequency. (If the result space of V is infinite, then "uniform distribution" must be relative to a pre-given measure on $\emptyset_V$. "Uniformity" then means that subsets of $\emptyset_V$ are obtained from X with probability proportional to their measure). With respect to such a construction, it is now possible to view the probabilities with which various results of V are actually obtained on the basis of our experience, as "atypical". Stated another way, with respect to the artific which we have just introduced, one can view the typical as atypical. As will presently develop, I believe that this is how one comes to ask oneself "why" questions about certain very familiar parts of experience. Throughout the remainder of this paper we shall constantly be assuming that the total domain of procedures of observation have a probability measure such that all of the conceptually possible results of observation are distributed uniformly. It must be heavily emphasized that this is not related to a

metaphysical assumption about the world, nor yet to the conviction that general experience leads one to such a supposition. It is a purely formal device, which serves as a "ground" for certain "figures" in which we are interested - interested because it is just these figures which typically call the scientific "why" question into existence.

And now to return to the traveller.

Suppose that he had pursued his desire for an explanation and had discovered that citizens of country C who own houses and whose annual income is more than some critical amount, all paint their rooftops red, while all others paint theirs' green. Such a discovery could be expressed in the following way. We are given three procedures of observation whose common domain X is the class of all citizens who own houses, and which are specified below:

V. Universe: X ($x \in X$)

Procedure: Observe the color of the rooftop of the house of any x

Result Space: all simple color names of English

R. Universe: X

Procedure: Observe the place of residence of any x

Result space: the names of all possible countries in the world

S. Universe: X

Procedure: Discover the dollar value of the annual income of any x

Result space: integers from 0 to some sufficient maximum.

The traveller's discovery amounts to saying that procedure V,

restricted to the thing defined by the result "country C" of procedure R, is dependent upon procedure S, likewise restricted to this part of X. This is easily seen. Consider S' derived from S by dividing $\emptyset_S$ into two parts; one containing all values less than the critical amount for country C, and the other the remainder. Now within the restricted range, the two things defined by procedure S' exactly coincide with the two things defined by V.

We have already suggested that "why" questions are stimulated by the non-uniformity of results obtained from a procedure of observation under certain restricting conditions. We have now tried to show that the existence of an "explanation" is tantamount to the existence of a dependence relation between one procedure (whose results are being explained) and other procedures - the dependence only existing within a restricted domain. If one simply views a stock of procedures over a universe X abstractly, there is no reason to expect such a link (between non-uniformity and dependence) without some additional assumption of how the things defined by these procedures are related to each other. We do not know how to formulate such an assumption explicitly and, for the present, content ourselves with a label, namely:

S 1.3 A stock of procedures so related that non-uniform distribution of results from one of them over a restricted domain, leads one to expect dependence of the one on the others, within the restricted domain, will be called a stock of "deeply related" procedures.[1]

---

[1] We would like to state the case more strongly than this. There are, perhaps, measures of non-uniformity, which would allow us that the more non-uniform the distribution of results the more certain we feel that there are dependence relations to be found. Measures of non-uniformity suggest themselves in information-theoretic terms, but we shall not pursue the matter further here.

Suppose we are given a stock of deeply related procedures. By means of one of these procedures, V, we obtain a finite set of results from some restricted domain Y. S 1.2 leads us directly to a measure of how seriously one may entertain the assumption that V, over Y, does not induce uniform measure on $\emptyset_V$ and hence, in view of what has been said above, how seriously the "why" question presents itself. The measure in question is the confidence with which one may reject the null-hypothesis of uniform distribution on the basis of the given set of observations. (There are standard statistical techniques available forthe calculation of such confidence measures.)

Again we return to the traveller.

Suppose that the traveller had replaced procedure V with procedure W, described below.

W. Universe: X

Procedure: With a specified wave-length meter and a specified source of illumination, take readings near the surface of a rooftop of any house belonging to an x.

Results: real numbers, from a to b, where these are the limits of the wave-length meter range.

Now the travellers' results would have appeared as in figure 4, below.

Fig. 4

There would have been two clusters which would have caught his attention and for which he would have demanded an explanation.

Let us consider the relation between the result space of procedure V and that of procedure W with its usual metric.

Suppose that $\emptyset_V$ has N members. Then the partition induced by V on X has N parts. Now suppose we wish to find a procedure W', derivable from W, which coincides with V as nearly as possible. A little thought reveals that the partition of $\emptyset_W$ which yields the desired W', is a partition into N sub-intervals (ignoring the fact that purple encompasses part of the red as well as part of the blue range of the spectrum). What is more, the coincidence between W' and V is good (i.e., the probability that an x which yields a certain result in $\emptyset_V$ does not yield the corresponding result in $\emptyset_{W'}$, is small). Beyond this: consider any color discrimination procedure other than V - i.e., a procedure whose result space includes the compound color names of English, or consists of Chinese color names. For each such procedure, we can find a W' which nearly coincides with it, always obtained by partitioning $\emptyset_W$ into sub-intervals. We may also consider what this implies about the measure induced by W over Y on $\emptyset_W$, where Y is a part of X corresponding to a color name. The bulk of the weight will be concentrated on a sub-interval of $\emptyset_W$, and on no sub sub-interval will the measure induced by W over Y be less than the measure induced by W over Z, where Z is a part of X corresponding to a different color name (belonging to the same discrimination procedure as the one which defined Y). All of which shows that there is a special relation between the metric on $\emptyset_W$, and the family of visual color discrimination procedures. Slightly generalized, this leads us to the following definition.

S 1.4 The metric on $\emptyset_R$ of a procedure, R, is "natural" with respect to a given family, F, of procedures, f, if

For every f F there is an R', derivable from R, which nearly coincides with f, and whose result space is a collection of near-spherical sub-regions of $\emptyset_W$.

We are now in a position to state, more or less clearly, when one is lead to look for clusters in a set of results, and to sketch an approach as to how these might be found.

Suppose we are given a stock, $\emptyset$, of deeply related procedures whose common total domain is X. We are also given a procedure R $\emptyset$, with a metric on $\emptyset_R$ which is natural with respect to a family of procedures F $\emptyset$.[1] Now we consider some restricted domain, definable in terms of the procedures in $\emptyset$ F, and named "Y". There may be an f which, when restricted to Y, gives non-uniform results and is therefore of interest. Procedures of F may be difficult to carry out, and they are numerous. We do not apply them directly, but apply R instead, obtaining a sample of results from Y.

Suppose that the measure induced by R over Y on $\emptyset_R$ is uniform. Then there certainly is no f which gives non-uniform results over Y. This is because the metric on $\emptyset_R$ is natural with respect to F, and non-uniformity in results for some f would result in non-uniform weighting of the sub-regions which yield the derived procedure which approximates f. On the other hand, if we find non-uniform weighting of spherical sub-regions of $\emptyset_R$, we may reasonably conjecture the existence of f's which

1. It must be understood that all members of F have finite result spaces. R may well have an infinite result space.

give non-uniform results over Y. The conclusion of the last statement may, ofcourse, not be drawn from the weaker premise that R over Y induces non-uniform distribution on $\emptyset_R$, but we shall proceed as if even this were true. The idea is this: we construct statistics which are designed to discriminate between uniform distribution over $\emptyset_R$ and the sort of distributions which concentrate their weight on spherical sub-regions of R. Given a sample of results in $\emptyset_R$ which came from Y, we calculate sample values for the statistics mentioned above and see which of them (different ones allocating the points of the sample to different regions) gives us the best rejection of the uniform assumption. We then conjecture that some f, whose corresponding subdivision of $\emptyset_R$ conforms with the subdivisions for which the latter statistic tested, also would give us significantly non-uniform results over Y. In support of such an inductive procedure, I offer a quote from Kendall[1]. On the subject of rejecting a null hypothesis on the basis of a sample value of a statistic which, on assumption of the null hypothesis seems very unlikely, he says:

> "....We have seen how it can be justified by confidence-interval or fiducial theory when a parameter is under consideration. When no parameter is specified, the process must, in the present state of our knowledge, rest on more intuitive ideas. My own view is that, in a vague kind of way, we are really considering the range of values of a parameter without realising it. In selecting a statistic to carry out the test, we usually relate it to the sort of effect we are expecting to divert the real state of affairs from those of our hypothesis. For instance if we suspect cyclical effects in a random series we base a test on oscillations in that series. The further the series deviates from randomness, the greater will be the value of our statistic; and consequently, if we could measure deviation from randomness (in the direction of cyclicality), we should have a parameter which could be

1. Kendall, The Advanced Theory of Statistics, Vol II, pp 135, 136

> "located in a range in the manner of confidence intervals. Such a range would exclude the larger values of our statistic if it can be regarded in any sense as estimating the parameter (or, more generally, as increasing with it); and hence the procedure of rejecting the hypothesis if the statistic is among these large values may be justified."

Without having said so, we have, in a general way, described, both the procedure and the meaning of looking for clusters in a set of results. By means of a procedure R, whose result space possesses a natural metric with respect to a family F, of procedures f (themselves deeply related to other procedures), we seek an f which gives significantly non-uniform results over some restricted domain. This is the meaning of looking for clusters.

Given a set of N results in $\emptyset_R$ from Y, we try various partitions of the results which make points belonging to the same part lie close to each other. As a measure for this closeness we have taken the mean-pairwise distance, averaged over all pairs both of whose members belong to the same part. We call this measure the "internal distance" and it is relative to a set of points and a partition of them.

S 1.5 We test alternative partitions of the given set of results from Y to see which of them gives us an internal distance which will permit us to reject the null-hypothesis of uniformity with greatest confidence.

The partition P which gives us the highest confidence of rejection is then interpreted as follows: an f, m of whose results would correspond to m near-spherical regions in $\emptyset_R$ each of which contains one and only one of the parts of P, would also have given us the highest confidence in rejecting the assumption of uniformity on Y (highest with respect to alternative f's)

The determination of a P such as the one above is called finding a "cluster" solution, and the parts of P are called "clusters".

Clusters may themselves be resolved into sub-clusters. This may be thought of in the following manner. Suppose we have found a cluster solution for a set of results from Y X, and with its help have determined and f which gives non-uniform results over Y. Consider the R', derived from R which most nearly approximates this f.. Its results are near-spherical regions of $\emptyset_R$, and if there were m clusters in our solution, then m of these regions will each contain all or most of one cluster. Now let us restrict our attention to one of the clusters and the region in which it lies. Call this region "S". Now we can define a new procedure $\overset{"}{R}$, which is just like R except that its result space is S (and hence its domain is $R^{-1}(S)$). Now we can regard the points of the cluster under consideration as having been obtained by procedure $\overset{"}{R}$ from the domain Y $R^{-1}(S)$, and we can analyze for clusters all over again (This process is intuitively suggested by figures 2 and 3). Doing this is only meaningful if the metric on $\emptyset_R$ is naturally related to a family of procedures, $\overset{"}{F}$. There may well be such a family $\overset{"}{F}$ (namely a subfamily of F) provided that the region S is not too small.. Meaningful or not, we are able to carry out the formalities of cluster analysis within this smaller result space, S.

Thus we see that analyzing a set of results for clusters belongs to a whole class of techniques which resolve some numbers

or functions into components - as with a Fourier analysis of a complex wave form and with various techniques which analyze variance. In the case of clusters what is being analyzed is a quantity which represents the degree to which a distribution differs from uniformity. Successive fractions of this are obtained, these being related to successively smaller frames of reference..

Section II

Development of a Procedure for Cluster Analysis

Given:

1. A bounded metric space, $\emptyset$, with finite measure. (This measure defines what is meant by "uniform distribution" over $\emptyset$, namely, a probability measure which weights every set in proportion to its size)
2. A set of N points of $\emptyset$, $\{x_1, x_2, x_3 \ldots\ldots x_N\}$

For any set of N points, $\{y_1, y_2 \ldots y_N\}$ and a partition of them, P_v with parts $P_{v1}, P_{v2}, \ldots\ldots P_{vr}$, we define a function d_I by the following.

$$d_I\left(\{y_1, y_2 \cdots y_N\}, P_v\right) = \frac{\sum_{k=1}^{r} \sum_{i<j,\ y_i, y_j \in P_k} d(y_i, y_j)}{\sum_{k=1}^{r} \binom{n_k}{2}}$$

where n_k is the size of P_k, and d is the distance of $\emptyset$.
We call this function "internal distance". We also have the function d_E ("external distance") defined by

$$d_E\left(\{y_1, y_2 \cdots y_N\}, P_v\right) = \frac{\sum_{k<l}^{r} \sum_{y_i \in P_k,\ y_j \in P_l} d(y_i, y_j)}{\sum_{k<l}^{r} n_k n_l}$$

d_I is not defined for the partition which has as many parts as there are points, and d_E is not defined for the partition which

which only has one part. We define d_I and d_E to be 0 for the partitions for which they are not automatically defined.

By an "r-set" we shall mean a set of r integers, $m_1, m_2 \ldots m_r$ such that

A partition, P_v, will be said to "conform to the r-set, $\{m_1, m_2, \ldots m_r\}$" if it partitions the points into r parts of those respective sizes.

Strictly speaking, instead of writing "P_v" for the partition of a set of points, $\{y_1, y_2 \ldots y_N\}$, we should have written "$P_v, \{y_1, y_2, \ldots y_N\}$" since P_v is only defined relative to a particular set of points. We shall now provide a means for specifying a partition which is independent of the N points which it partitions.

Given an r-set, v, and any set of points $\{y_1, y_2 \ldots y_N\}$, consider the family F_v of all partitions $P_v, \{y_1, y_2, \ldots y_N\}$ which conform to v. Suppose that F_v contains m members. Now obtain the set of m numbers $d_1, d_2 \ldots d_m$ by calculating d_I for $\{y_1, y_2 \ldots y_N\}$ and each member of F_v. We order the numbers $d_1, d_2 \ldots d_m$ in their natural order. If two of them are identical we nevertheless call one of them less than the other. It does not matter which is made less, as long as some decision is made for each ambiguity. When such an ordering has been made for every set of N points, we can specify a partition for any set of N points in the following manner: given an r-set, v, and an integer w between 1 and m, let $P_{v,w}$ be the partition on any set of N points, which conforms to v and for which d_I is the w'th in order of magnitude among all the partitions

of the N points which conform to v..

The partitions which we shall consider for any set of N points are those, and only those, for which w is 1 - i.e. those partitions which, with respect to the r-set to which they conform, minimize d_I. Hence we can drop the w index for partitions..

We observe that the partition which minimizes d_I relative to the r-set to which it conforms, also maximizes d_E relative to the r-set since, for a fixed r-set, d_I and d_E always add up to a fixed sum.

The uniform distribution over $\emptyset$ determines a distribution f_v over the range of d_I for fixed P_v. Now suppose that we have calculated the distributions, f_v for each v..

We take the points $x_1, x_2, \ldots . x_N$ and calculate d_I for each P_v. The value obtained for a particular P_v will divide the domain of f_v into two intervals. The measure of the interval to the right under f_v is the probability that a sample of N points, partitioned by P_v, will yield a value for d_I greater than the value obtained from $x_1, x_2, \ldots . x_N$, if the N points are taken randomly from $\emptyset$ under uniform distribution. Any P_v which makes this measure greatest is a cluster solution for the set, and its members are clusters.

A special situation arises in the case of the partition with N parts. In this case d_I is always 0 so that f_v has its entire weight on 0. We establish the convention that for any given pattern of points, the d_I value obtained for it (namely 0) splits the weight of the distribution in half - half to the "right" and half to the "left". Therefore the value .50 will always enter the competition for the cluster solution given by the partition with N parts..

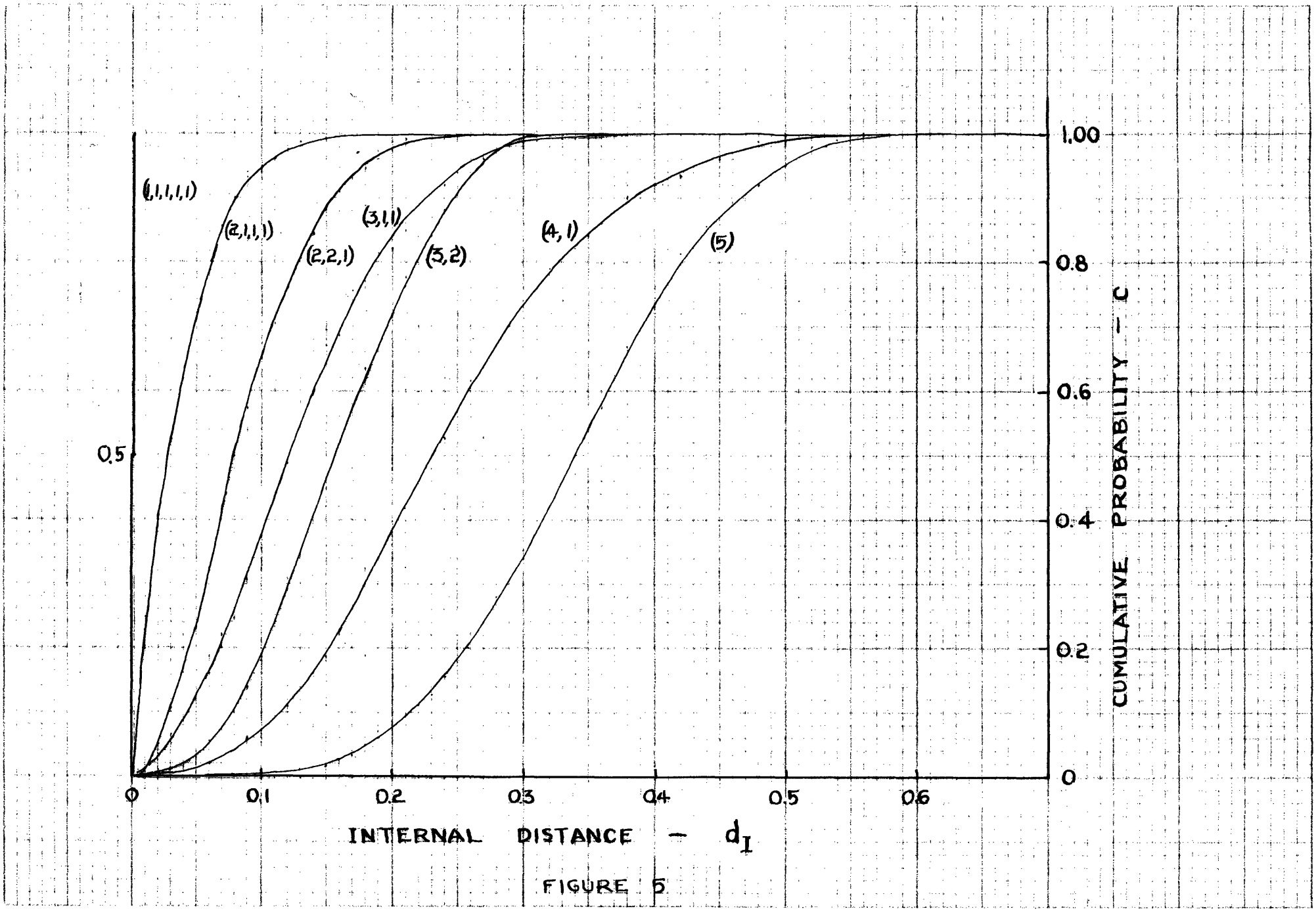
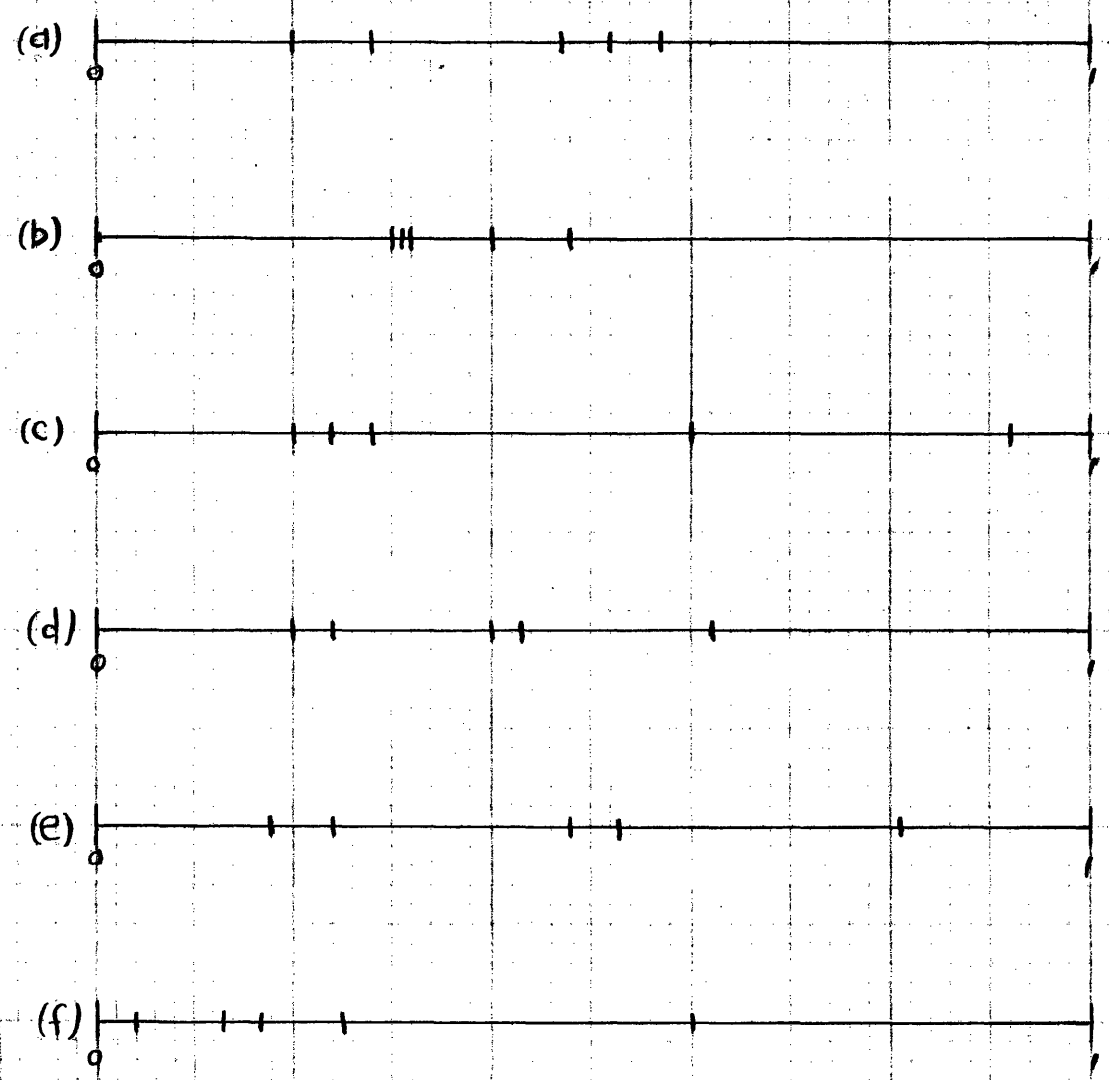

(1,1,1,1,1)
(2,1,1,1)
(2,2,1)
(3,1,1)
(3,2)
(4,1)
(5)
0.5
0
0.1
0.2
0.3
0.4
0.5
0.6
1.00
0.8
0.6
0.4
0.2
0
INTERNAL DISTANCE — d_I
CUMULATIVE PROBABILITY — C

FIGURE 5



	(5)	(4,1)	(3,2)	(3,1,1)	(2,2,1)	(2,1,1,1)	(1,1,1,1,1)
d_I	.200	.150	.067	.070	.075	.050	.000
C	.075	.190	.061	.210	.450	.710	.500
d_I	.090	.050	.030	.013	.045	.010	.000
C	.002	.012	.012	.018	.200	.210	.500
d_I	.360	.208	.120	.053	.180	.040	.000
C	.580	.400	.250	.140	.950	.620	.500
d_I	.206	.141	.115	.133	.035	.030	.000
C	.090	.170	.270	.570	.130	.530	.500
d_I	.309	.212	.174	.200	.055	.450	.000
C	.370	.425	.590	.840	.260	.660	.500
d_I	.248	.111	.153	.087	.085	.040	.000
C	.170	.090	.480	.290	.540	.620	.500

TYPE OF CLUSTER

FIGURE 6

Appendix I

I have obtained an approximation to each of the six distributions which are necessary in order to test for clusters for five points on the unit interval. There is one distribution for each of the r-sets (5), (4, 1), (3,2), (3,1,1), (2,2,1), and (2,1,1,1). The approximations were obtained by a Monte Carlo technique from 300 sets of 5 points, randomly distributed on the interval. Figure 5 shows the six cumulative distributions thus obtained. The distribution for (1,1,1,1,1) is also represented.

Figure 6 is a presentation of 6 examples, with an adjoining table which shows the seven d_I values, and the corresponding c values, read from the cumulative distributions. Since c is the complement to 1 of the measure of confidence, the partition which has the smallest c value is the one which represents the cluster solution. We have underlined the smallest c value for each case.

Appendix II

In modern linguistics onehas long conjectured the possibility of discovering form classes (e.g., word classes or morpheme classes) on the basis of the "distributions" of the forms in question. The idea is as follows: suppose a linguistic form occurs in an utterance. One may restate this fact by saying that the form has occurred in a certain "utterance environment", the latter being all parts of the utterance exclusive of the form in question. The usefulness of looking at it in this way comes from the fact that many forms occur in the same utterance environment. For example: "John goes to school" may be regarded as an occurrance of John, in the environment "_____goes to school.", but many other forms, such as "Jack" or "The engineer", also occur in the same environment.

The collection of all environments in which a given form occurs, is called the "distribution" of the form.

Now consider a pair of forms which are synonymous. Virtually any utterance environment in which one of them occurs, the other does also - i.e. their distributions are nearly identical. Consider next the class of all forms which refer to a piece of furniture on which one sits. It is clear that the distributions of all such forms will "heavily" overlap each other. Suppose that one regards obtaining the distribution of a form as a procedure of observation. One might now suspect that one could introduce a distance between distributions (which would, among other things, fulfill the condition that the greater

the overlap between two distributions, the closer they are) which would be naturaly related to a certain family, F, of procedures of observation called "meaning distinction procedures". As we know from traditional approaches to grammatical categegories, these were treated as if they were the partitioning induced on the universe of linguistic forms by certain members of F (i.e., "A noun is the name of a place, person or thing", "A verb is the name of an action", etc.)

Now let us imagine the universe of linguistic forms extended in such a way that every conceivable distribution is obtained with equal frequency (this is meaningful if we consider utterances to be of some maximum finite length). If, with respect to this imagined universe, we restricted our observations to that portion of it which is observable in course of speech communication, we might well expect clusters of distributions to emerge in finite collections of observations. Because of the natural metric, we would expect these clusters to be related to meaning distinction procedures which, on our "restricted" domain give non-uniform results.

Appendix II was written in excessive haste, and the reader is asked to regard it as merely suggestive - not as a coherent presentation.

BIBLIOGRAPHY

1. Hoel, Paul G, Introduction to Mathematical Statistics New York, John Wiley & Sons, Inc., 1949

2. Kendall, Maurice G., The Advanced Theory of Statistics, Vol. II, 2nd edition, London, Charles Griffin & Co., 1948

Mood, Alexander M., Introduction to the Theory of Statistics, 2nd ed., New York, McGraw Hill, 1950

4. Shannon, Claude E, and Warren Weaver, The Mathematical Theory of Communication, Urbana Ill., University Press, 1949

5. Tryon, Robert Choate, Cluster Analysis, Ann Arbor, Mich. Edward Bros., Inc., 1939