NEW SERIES, NO. 176 (VOL. XXVI) DECEMBER, 1931

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION

Formerly the Quarterly Publication of the American Statistical Association

STATISTICAL CORRELATION AND THE THEORY OF CLUSTER TYPES

BY RAGNAR FRISCH, *University of Oslo*, AND BRUCE D. MUDGETT, *University of Minnesota*

1. THE NOTION OF CLUSTER TYPES AND THEIR GEOMETRIC REPRESENTATION IN THREE DIMENSIONS

Let there be given a statistical population composed of N observations each characterized by n variable attributes, $x_1, x_2 \ldots x_n$. Let n be called the dimensionality of the observations. It is proposed to discuss the types of systematic variation that may take place between the several variables: What is the degree of freedom of the system? Which variables can be considered independent? And so on. To simplify, we shall consider only linear relationships between the variables. This will be sufficient to indicate the nature of the possibilities it is proposed to analyze. Most of the argument in the present paper is built upon the theory of cluster types developed by Ragnar Frisch in his paper, "Correlation and Scatter in Statistical Variables," in the *Nordic Statistical Journal*, Vol. 1, 1928, pp. 36–102. To simplify the discussion and to make it possible to build upon a direct geometric intuition of the situation, the case of three variables, x_1, x_2 and x_3 will first be discussed. That is to say, we deal with a population having three attributes measured along three rectilinear axes. In this case, the dimensionality of the population, n, is equal to three. A point in three-dimensional space then represents a given observation, that is, a given individual in the population of N. Each such observation is characterized by a given value of each of the three variables x_1, x_2 and x_3. By assuming that the variables x_1, x_2 and x_3 are measured as deviations from their means, the mean of each x_i will lie in the origin of the coördinate axes. The swarm of N observation points thus

obtained forms a three-dimensional scatter diagram. This scatter diagram may exhibit any one of the following *cluster types.*

A. *The disorganized swarm.* In this case the observations are distributed in a disorganized way in space with no restrictions upon their positions, or no loss of freedom. The number of degrees of freedom in the population will be called its rank, or its unfolding capacity, and will be designated by ρ, so that in the present case $\rho = n = 3$. The rank is now equal to the dimensionality. The case is illustrated by a swarm of bees distributed widely in a given space without concentration at any point or again by the raindrops as they fall through space in a storm.

B. *The plane (the "pancake").* If the swarm is pressed together in a given direction it assumes the shape of a "pancake." This pancake has as its ideal representation a plane passing through the origin. This plane represents the systematic variations while the deviations from the plane, the thickness of the pancake, represent the accidental variations. It may now be said that the observations come close to lying in a plane, the accidental variations producing a thin slab in place of a plane. From the point of view of systematic (as distinguished from accidental) variations, the population has now been subjected to the loss of one degree of freedom. The representative point may move freely within the slab but cannot go beyond it except by "accident." It may be said that the swarm has received a one-dimensional flattening, or, again, that it has been *simply* flattened. The population may also be called *simply collinear* in this instance. If the number of degrees of freedom lost by the population be designated by p, the result now is: $p = 1$, $\rho = 2$, $p + \rho = n$. The rank is now exactly one less than the dimensionality. The case may be illustrated perhaps by a swarm of bees that has alighted on a board which passes through the origin of the coördinate axes, the bees forming a cluster on this board.

C. *The rod.* If the population discussed under B be pressed together so that it come to cluster along a line in the plane, the swarm of scatter points loses one further degree of freedom. The observations become concentrated around a rod through the origin. This case is referred to as *multiply* flattened, or *multiply* collinear (more precisely, two-fold flattened). The rank of the population is now 1, $\rho = 1$; the flattening is $p = 2$. The sum of the rank and the flattening is, of course, still equal to 3.

D. *The point.* Subject the rod to one further degree of flattening, or to the loss of its one remaining degree of freedom and the observations become concentrated around a point or in a tiny ball at the origin. The meaning of this is that from the point of view of systematic variations,

the observed population does not show any variation at all. Whatever variation there has been is in the nature of accidental errors of observation. Now the flattening is $p=3$. That is, the flattening is equal to the dimensionality. All degrees of freedom are lost and the unfolding capacity, ρ, is equal to zero.

These four cases represent the main types of relationships between the variables x_1, x_2 and x_3 of the population when the systematic variation in the x's is linear in character. Proceeding now to a more detailed discussion of these main types, it is necessary to distinguish between certain subtypes. In case (A) there is no systematic relationship existing between the variables. The selection of an arbitrary value of x_1 and an arbitrary value of x_2 leads to no particular expectation with respect to x_3. Geometrically, if a line $x_1=\text{constant}=a_1$, $x_2=\text{constant}=a_2$ is drawn in the coördinate space (x_1, x_2, x_3) there will be no concentration of the observations around any particular value of x_3 on this line. And similarly for comparisons of x_2 with x_1 and x_3 etc.

Case (B) is not so simple. The slab, or plane, may assume various positions so long as it passes through the origin (this requirement being, of course, a matter only of the selection of the origin of the x's). The following subcases are to be distinguished:

B 1. The plane may contain *none of the coördinate axes.* As it passes through the origin it makes an angle not equal to zero with each of the three coördinate axes.

B 2. The plane may *contain one axis,* say the x_3 axis. It will then be perpendicular to the coördinate plane (x_1, x_2), and will appear as a door hinged to the x_3 axis. It might, of course, equally well have been hinged to any other of the coördinate axes. This is a very important case for what follows.

B 3. The plane of the observations may *contain two of the* coördinate axes, say x_2 and x_3. It now coincides with the (x_2, x_3) coördinate plane and the variable x_1 shows no systematic variation, its variation being only within the thickness of the slab; $\sigma_1=0$ approximately, that is σ_1 would have been zero if it were not for the accidental variations. In the case (B) the plane can, of course, never contain all three axes, so that (B 1–2–3) cover all possible cases.

The rod through the origin, i.e., the main type (C), likewise presents three sub-types.

C 1. The rod does *not lie in any of the coördinate planes.*

C 2. It lies in one coördinate plane, say (x_2, x_3); then $\sigma_1=0$, approximately.

C 3. It lies in two of the coördinate planes, that is it coincides with

one axis. For example, lying in the coördinate planes (x_1, x_3) and (x_2, x_3), it coincides with the x_3 axis and $\sigma_1=\sigma_2=0$, approximately.

The case (D) involves only one situation, the observations lying within a very limited distance (very tiny ball) around the origin. Here σ_1, σ_2, and σ_3 are all approximately equal to zero.

The term "approximately" in the above analysis is used to indicate that the parameters in question, because of the accidental variations, may deviate somewhat from their systematic values. It is because of the accidental variations that the points lie, in one case, in a slab rather than in a plane; again in a rod rather than on a perfect line and finally in a tiny ball rather than rigorously in a point. For brevity hereafter the term "approximately" will as a rule be omitted in the discussion of the various cases.

2. ALGEBRAIC INTERPRETATION OF CLUSTER TYPES IN THREE VARIABLES

The cluster types that have been discussed are basic to an understanding of the nature of the systematic relationships between the variables x_1, x_2 and x_3 and for interpreting the linear regression of any of the variables upon one of, or all, the others. This will be recognized when the algebraic expressions for the various cluster types are brought to mind. We proceed now to discuss the various types in algebraic terms, that is, in terms of the nature of the linear relationships that exist between the variables.

In case (A), the disorganized swarm, or the raindrops, there is complete lack of system in the spatial distribution of the variables and a regression equation between them has no meaning. The only thing to do in this case is to leave the data alone. This case, therefore, may be dismissed from further consideration once a criterion has been established by which the lack of systematic organization can be recognized.

In case (B), the plane, where the observations have lost one degree of freedom, there exists one and only one systematic relationship between the variables of the form:

$$a_1x_1+a_2x_2+a_3x_3=0. \tag{2.1}$$

But this relationship may or may not actually contain all three variables. The case where it does contain them all is the case where the plane contains none of the axes (Case B 1). Here the coefficients a_1, a_2 and a_3 are all different from zero. Then the equation (2.1) may be solved for any x_i in terms of the other x's. That is to say, equation (2.1) may be written in any one of the following three ways:

$$
\begin{aligned}
x_1 &= a_{12\cdot 3}x_2 + a_{13\cdot 2}x_3 \\
x_2 &= a_{21\cdot 3}x_1 + a_{23\cdot 1}x_3 \\
x_3 &= a_{31\cdot 2}x_1 + a_{32\cdot 1}x_2
\end{aligned} \qquad (2.2)
$$

The notation $a_{12\cdot 3}$, etc., is here used instead of the usual $b_{12\cdot 3}$ for the regression coefficients in order to indicate that it is here only a question of solving the equation (2.1) in three different ways. The notation $b_{12\cdot 3}$, etc., for the regression coefficients involves something more than merely different ways of writing the same equation. It refers to different statistical procedures for determining the coefficients, namely different directions in which to make the least squared minimalization.

In the case (B 1) where it is possible to express each variable in terms of the others the variables are said to form a *closed* set.

Case (B 2) is the situation where the observational plane contains the x_3 axis. The set is still collinear but is no longer a closed set; x_3 has now become a superfluous variable. It has no place in the regression. This can be seen easily in the geometric figure. The plane of the observations is perpendicular to the (x_1, x_2) coördinate plane and intersects the latter in a line through the origin. This line is called the trace of the regression plane in the (x_1, x_2) plane. This trace is evidently a locus of points with (x_1, x_2) coördinates. If an arbitrary (x_1, x_2) point be selected, then one of the following two things will happen. Either this (x_1, x_2) point falls on the trace, and then *any* magnitude of x_3 may correspond to the selected (x_1, x_2) point; or the (x_1, x_2) point falls outside the trace, in which case *no* x_3 magnitude will correspond to it. It, therefore, has no meaning now to express x_3 in terms of x_1 and x_2. That which does have a meaning is to express a relation between x_1 and x_2. Selection of any value of x_1 therefore immediately specifies a corresponding value of x_2 as defined by the trace but does not specify any particular value of x_3. Inversely, selection of a particular value of x_3 does not locate a particular value of either x_1 or x_2. The variable x_3 is not a partner in the systematic relationship. This is equivalent to saying that in the equation (2.1) $a_3 = 0$, so that the relationship between the variables is now of the form:

$$a_1x_1 + a_2x_2 = 0$$

where a_1 and a_2 are not equal to zero. In other words, x_1 can be expressed in terms of x_2, thus: $x_1 = a_{12}x_2$; or inversely, $x_2 = a_{21}x_1$. The relationship $x_3 = a_{31\cdot 2}x_1 + a_{32\cdot 1}x_2$ does not exist in the present case; x_1 and x_2 taken by themselves form a closed set, and x_3 is superfluous.

Case (B 3) where the regression plane coincides with one of the coördinate planes also represents a single relationship of the form (2.1)

between the variables. But in this relationship there are now two coefficients that are zero. If the plane lies in the (x_2, x_3) coördinate plane, $a_2 = a_3 = 0$. The relationship is then represented by

$$a_1 x_1 = 0;\ a_1 = 0;\ \sigma_1 \neq 0.$$

That is, the observations are scattered freely in the $(x_2 x_3)$ coördinate plane, there being no systematic relationship between them in this plane. And x_1 is an *ineffective* variable. So far as the systematic variation is concerned $x_1 = 0$ for all values of x_2 and x_3.

Where the observations are clustered in a rod, *there exist two independent relationships between the variables.* There even exist three relationships, represented by the three equations:

$$\begin{aligned} a_1 x_1 + a_2 x_2 &= 0 \\ a'_1 x_1 + a'_3 x_3 &= 0 \\ a''_2 x_2 + a''_3 x_3 &= 0 \end{aligned} \tag{2.3}$$

But only two of these three relations will be independent. By knowing any two of them, the third can be derived. Any set of two variables taken by themselves now constitutes a two-dimensional collinear set.

In the case (C 1), each of the three two-dimensional collinear sets (2.3) is closed. That is to say, all the coefficients in (2.3) are $\neq 0$. Any of the variables may now be expressed in terms of any *one* of the other variables.

In the case (C 2) the rod lies in one, and only one, of the coördinate planes, say in (x_2, x_3). There is now one closed set of two variables, represented by the equation $a_2 x_2 + a_3 x_3 = 0$, where a_2 and a_3 are both $\neq 0$. Furthermore, there is one ineffective set of one variable, $a_1 x_1 = 0$, $a_1 \neq 0$, $\sigma_1 = 0$. But x_1 cannot be expressed in terms either of x_2 or of x_3 (except in the trivial form $x_1 = 0.x_2 + 0.x_3$).

The case (C 3), where the rod is lying in the x_3 axis involves two ineffective sets of one variable each, namely:

$$\begin{array}{lll} a_1 x_1 = 0 & a_1 \neq 0 & \sigma_1 = 0 \\ a_2 x_2 = 0 & a_2 \neq 0 & \sigma_2 = 0 \end{array} \tag{2.4}$$

that is, there is variation along the x_3 axis but no variation along the x_1 and x_2 axes.

In the case of the point (or tiny ball), (D), there exist three independent relations between the variables. There is no freedom left now in the three variables, all three being ineffective. That is to say, the three independent relations are:

$$\begin{array}{lll} a_1 x_1 = 0 & a_1 \neq 0 & \sigma_1 = 0 \\ a_2 x_2 = 0 & a_2 \neq 0 & \sigma_2 = 0 \\ a_3 x_3 = 0 & a_3 \neq 0 & \sigma_3 = 0 \end{array} \tag{2.5}$$

3. ALGEBRAIC INTERPRETATION OF CLUSTER TYPES IN n VARIABLES

A set of n variables $(x_1, x_2 \ . \ . \ . \ x_n)$ is called a linearly dependent set, or collinear when there exists at least one relation of the form:

$$a_1x_1+a_2x_2+ \ . \ . \ . \ +a_nx_n=0 \tag{3.1}$$

where the coefficients, a_i, are not all equal to zero. A collinear set is said to be flattened exactly p times, or to be p-fold flattened when there exist exactly p independent relationships of the form (3.1) between the n variables. A necessary and sufficient condition for a set to be p-fold flattened is that there exist at least one ρ-dimensional set which is non-collinear, where $\rho=n-p$, while all $(\rho+1)$ and higher dimensional subsets are collinear. In this case there exist exactly p *independent* regressions, each involving not more than $(\rho+1)$ variables, and further being such that the set of these $(\rho+1)$ variables is a simply collinear set.

When $p=0$ the set of n variables is not flattened; there exists no systematic linear relationship between the variables and no regression equation is possible.

When $p=1$ the set is once flattened or is simply collinear. The rank, ρ, of the set is now $(n-1)$ and there exists *one* $(n-1)$-dimensional regression plane.

This regression plane may or may not contain all the variables. The plane will not contain those variables the coördinate axes of which in n-dimensional space lie in the regression plane, and the corresponding coefficients in the regression equation (3.1) will be equal to zero. These variables are superfluous variables in the regression system. If the regression coefficients a_i $(i=1, 2, \ . \ . \ . \ n)$ are all different from zero, all of the n variables are present in the regression equation. The set is then a *closed set.* The $(n-1)$-dimensional regression plane now contains none of the coördinate axes and each x_i can be expressed linearly in terms of the others. In this case there cannot exist any relationship of the form (3.1) involving fewer than n variables.

When $p>1$ the set is multiply collinear or multiply flattened and there exists a regression manifold of less than $(n-1)$ dimensions.

4. STATISTICAL CRITERIA FOR CLUSTER TYPES

As statistical criteria for these several types of clustering there are here introduced the coefficient of collective alienation and its correlative, the coefficient of collective correlation. Proceeding now to the exact definition of the collective alienation and correlation coefficients, let

$$(\mathrm{R}) = \begin{pmatrix} r_{11}, & r_{12} & \cdots & r_{1n} \\ r_{21}, & r_{22} & \cdots & r_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ r_{n1}, & r_{n2} & \cdots & r_{nn} \end{pmatrix}$$

be the correlation matrix for the n variables $x_1, x_2 \ldots x_n$; r_{ij} is the simple (total) correlation coefficient between x_i and x_j. The determinant value of this matrix is denoted by

$$R = R_{(12 \ldots n)} = \begin{vmatrix} r_{11}, & r_{12} & \cdots & r_{1n} \\ r_{21}, & r_{22} & \cdots & r_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ r_{n1}, & r_{n2} & \cdots & r_{nn} \end{vmatrix}$$

Further, $R_{ij} = R_{ij(12 \ldots n)}$ denotes the element in the i-th row and the j-th column of the adjoint correlation matrix $(\hat{R})$. Each R_{ij} is determined by calculating the determinant value of (R) after crossing out the i-th row and the j-th column and multiplying by $(-1)^{i+j}$. We shall not here enter into a theoretical discussion as to why the collective alienation and correlation coefficients offer plausible criteria of how far a given set of statistical variables deviates from being linearly dependent. This interpretation is discussed at length in the paper by Ragnar Frisch already referred to.[1] Here mention will be made only of the formal hierarchic order that exists between the partial, the multiple and the collective coefficients. Consider the set of variables, $x_1, x_2 \ldots x_n$, and denote the classical partial correlation and alienation coefficients by $r_{ij(12 \ldots n)}$ and $s_{ij(12 \ldots n)}$. For the sake of symmetry, *all* the subscripts, 1, 2 . . . n are written as secondary subscripts, without omitting i and j. In order not to give rise to confusion with the usual notation, where i and j are omitted from the list of secondary subscripts, the secondary subscripts are here enclosed in a parenthesis. Similarly the classical multiple correlation and alienation coefficients are designated by $r_{i(12 \ldots n)}$ and $s_{i(12 \ldots n)}$ respectively. As is well known, these parameters satisfy the equation:

$$r_2 + s_2 = 1 \tag{4.1}$$

where the same set of subscripts is attached to r and to s. The collective correlation and alienation coefficients are also defined so as to satisfy (4.1). But while the partial coefficients depend on two primary subscripts and the multiple coefficients depend on one primary subscript, the collective coefficients have no primary subscripts at all. They are defined with regard to the *set* of variables as such. More precisely, they are defined thus:

[1] In this paper the term *coefficient of scatter* was used instead of *coefficient of alienation*.

(4.2) $s = s_{(1\ 2\ \ldots\ n)} = \sqrt{R_{(1\ 2\ \ldots\ n)}}$ = coefficient of collective alienation in the set $(x_1, x_2 \ldots x_n)$

$r = r_{(1\ 2\ \ldots\ n)} = \sqrt{1 - R_{(1\ 2\ \ldots\ n)}}$ = coefficient of collective correlation in the set $(x_1, x_2 \ldots x_n)$

It is easy to prove that these coefficients satisfy the relations:

$$0 \leqq s \leqq 1 \qquad 0 \leqq r \leqq 1 \tag{4.3}$$

If $n = 2$ the collective alienation coefficient reduces to the simple (total) alienation coefficient and the correlation coefficient (apart from its sign) reduces to the simple (total) correlation coefficient. From the definitions here given it follows that s^2 and r^2 are *polynomials* in the simple correlation coefficients r_{ij}; s^2 and r^2, therefore, never can become of the indeterminate form $\frac{0}{0}$ (unless one or more of the simple correlation coefficients become of this form). This is a fundamental property of the collective alienation and correlation coefficients, which distinguishes these parameters from the corresponding partial and multiple coefficients. The fact that the partial and multiple coefficients may become of the indeterminate form $\frac{0}{0}$ is easily seen from the well known formulae:

$$r_{i(1\ 2\ \ldots\ n)} = +\sqrt{1 - \frac{R}{R_{ii}}} = \text{coefficient of multiple correlation} \tag{4.4}$$

$$r_{ij(1\ 2\ \ldots\ n)} = -\frac{R_{ij}}{\sqrt{R_{ii} \cdot R_{jj}}} = \text{coefficient of partial correlation} \tag{4.5}$$

It is indeed easy to prove that

$$R^2_{ij} \leqq R_{ii} \cdot R_{jj} \tag{4.6}$$

and

$$0 \leqq R \leqq R_{ii} \leqq 1 \tag{4.7}$$

so that if $R_{ii} \to 0$, (4.4) and (4.5) must give rise to indeterminate expressions of the type $\frac{0}{0}$.

The general property of the collective alienation coefficient which makes this parameter a useful tool in studying cluster types is the following. The collective alienation is equal to zero when, and only when, there exists an exact linear dependency in the set for which the collective alienation is computed, and furthermore this coefficient

increases as the swarm of scatter points takes on a shape that deviates more and more from the shape where a linear dependency exists. When the collective alienation has become equal to 1 (which is its largest possible value) the swarm of scatter points has reached a shape which may be characterized as *perfect unfolding*. The variables have now become orthogonal (uncorrelated), that is to say, all the simple correlation coefficients, r_{ij} are equal to zero. The collective alienation coefficient being equal to unity is the necessary and sufficient condition for orthogonality in the above defined sense.

The three variables x_1, x_2 and x_3 may be taken to illustrate the use of the collective alienation as a tool in determining cluster types. The various cases to be discussed have their ideal pattern in the elementary, well-known algebraic propositions regarding linear forms. What is done here is primarily to translate these propositions from the language of perfect linear dependency to the language of "nearly" linear dependency.

A. The disorganized swarm is characterized by $s_{(123)}$ being near to unity.

B. The plane. One flattening; $p=1$, $\rho=2$, simply collinear. The criterion for this case is that $s_{(123)}$ is near to zero, and furthermore at least one of the three magnitudes, $s_{(23)}=\sqrt{R_{11(123)}}$, $s_{(13)}=\sqrt{R_{22(123)}}$, $s_{(12)}=\sqrt{R_{33(123)}}$ is significantly different from zero.

B 1. Plane contains no coördinate axis. Each of the three magnitudes, $s_{(13)}$, $s_{(23)}$ and $s_{(12)}$ is significantly different from zero. In this case (x_1, x_2, x_3) form a *closed set.*

B 2. Plane contains one coördinate axis (x_3). Now $s_{(12)}=\sqrt{R_{33(123)}}$ is close to zero while $s_{(23)}$ and $s_{(13)}$ are significantly different from zero. In this case the two variables (x_1, x_2) taken by themselves form a *closed set* and x_3 is a superfluous variable.

B 3. Plane contains two coördinate axes, (x_2, x_3): Here $s_{(12)}$ and $s_{(13)}$ are both close to zero, while $s_{(23)}$ is significantly different from zero. In this case x_2 and x_3 form a set of disorganized variables while x_1 is ineffective.

C. The rod. Two-dimensional flattening; $p=2$, $\rho=1$; multiply collinear. The criterion for this case is that $s_{(123)}$ shall be near to zero and, at the same time, that $s_{(12)}$, $s_{(13)}$ and $s_{(23)}$ shall also be close to zero. There now exist two independent relationships between the variables.[1] These two relationships may be written

[1] Strictly speaking, the criteria here considered show only that there exist *at least* two linear relationships. If it be assumed that not all the three variables are ineffective (i.e., that we do not have case D) then the criteria considered show that there are *exactly* two independent relationships.

in different ways. In particular they may by elimination be written in such a way that each of them contains at most two variables. If it be assumed that all the variables are effective then each of the two relations thus obtained must contain exactly two variables. In this case any set of two variables constitutes a closed two-dimensional set.

Criteria for the sub-types (C2) and (C3) cannot be discussed in terms of the collective alienation coefficients only. Or more precisely expressed: The very fact of computing the collective alienation coefficients (which are based on the simple correlation coefficients) involves the assumption that all the variables are effective. But this is also the only assumption made in using the collective alienation coefficients. If none of the variables are ineffective, all the simple (total) correlation coefficients are determinative.

The situation for three variables is now easily generalized to the case of n variables. If a great number of variables $(x_1, x_2 \ldots x_n)$ are observed, and criteria for cluster types are wanted, compute first the collective alienation coefficient $s_{(12\ldots n)}$ for the whole set. If this coefficient is not close to zero, any attempt at studying the variables by means of linear relationships should be abandoned.

If $s_{(12\ldots n)}$ is close to zero, consider all the $(n-1)$-dimensional subsets $(23\ldots n)$, $(13\ldots n)\ldots(12\ldots(n-1))$ that can be formed by leaving out one variable at a time. There are in all n such subsets. Compute the collective alienation coefficient for each such subset. Such a collective alienation coefficient we call an $(n-1)$-dimensional collective alienation coefficient. If there exists *at least one* such $(n-1)$-dimensional coefficient that is significantly different from zero, the set is simply collinear, that is, there exists exactly one linear relation of the form.

$$a_1x_1+a_2x_2+\ldots+a_nx_n=0. \tag{4.8}$$

This relation should be looked upon as actually containing only those variables x_k that are such, that the collective alienation obtained by leaving out x_k, namely,[1] $s_{(12\ldots)k(\ldots n)}$, is significantly different from zero.

If all the $(n-1)$-dimensional collective alienation coefficients $s_{(12\ldots)k(\ldots n)}$ $(k=1.2\ldots n)$ are close to zero, the set is at least 2 dimensionally flattened, that is, there exist at least two independent relations of the form (4.8). In order to find out if there exist exactly two or possibly even more than two such relations, consider all the

[1] The inverse parenthesis)(is used to denote "exclusion of."

$(n-2)$ subsets that can be formed by leaving out in all possible ways two of the variables, and compute the collective alienation coefficient for each such subset. Such a coefficient will be called an $(n-2)$-dimensional collective alienation coefficient. If there exists at least one such $(n-2)$-dimensional coefficient that is significantly different from zero, then the flattening is exactly two, that is there exist exactly two independent relations of the form (4.8).

If all the $(n-2)$-dimensional collective alienation coefficients are close to zero, consider the $(n-3)$-dimensional coefficients. If all these should also be close to zero, consider the $(n-4)$-dimensional coefficients, etc. Suppose it be necessary to continue to the ρ-dimensional coefficients before a level is reached where at least one of the collective alienation coefficients is significantly different from zero. That is to say, all the $(\rho+1)$- and higher dimensional collective alienation coefficients are close to zero, while there exists at least one ρ-dimensional coefficient that is significantly different from zero. In this case the rank (*i.e.*, the unfolding capacity) of the set is ρ, and the flattening is $p=n-\rho$. There now exist exactly p independent systematic relations of the form (4.8). By combination and elimination, using these p relations, we may arrive at many sorts of linear relations between the variables. There is in particular one set of relations that is interesting: We may select a certain ρ-dimensional subset which is *not* a collinear set (at least one such set exists by the very definition of ρ). And then we may express each of the *other* p-variables linearly in terms of the selected ρ-dimensional subset.

5. MEANINGLESS RESULTS WHEN LINEAR DEPENDENCIES EXIST

Among the various possible cluster types discussed in the preceding sections there is only one very particular type in which all the orthodox correlation and regression parameters have a sense, namely, the case of a set which is not only collinear but also closed. If the set is not closed a great number of the orthodox correlation parameters lose their meaning. And if the set becomes multiply collinear, each of the orthodox correlation and regression parameters lose their meaning. Consider first the set which is simply collinear but is not a closed set. This is the case where $s_{(12\ \ldots\ n)}$ is close to zero, and one or more, but not all, of the $(n-1)$-dimensional coefficients $s_{(12\ \ldots\ )k(\ \ldots\ n)}$ are also close to zero. Those variables x_k for which $s_{(12\ \ldots\ )k(\ \ldots\ n)}$ is close to zero should simply be looked upon as superfluous variables from the point of view of linear regressions. Under no circumstances must the coefficients of (4.8) be determined by computing the regression coefficients $b_{kj.\ 12\ \ldots\ n}$ of x_k on the other variables. In fact the collective

alienation coefficient $s_{(12\ldots)k(\ldots n)}$ occurs as a denominator in $b_{kj.\,12\ldots n}$. Determining $b_{kj.\,12\ldots n}$ would, therefore, mean forcing a magnitude whose deviation from zero is non-significant into the denominator. The system of regression coefficients thus obtained would consequently be of the form: accidental error divided by accidental error, and would have no sense. This situation is illustrated by the three-dimensional case (B 2), that is the case where there exists exactly one systematic regression plane in (x_1, x_2, x_3), but where this plane contains the x_3-axis. In this case it would obviously have no sense to express x_3 as a linear combination of x_1 and x_2. In this case neither the multiple coefficients of correlation of x_k on the set of the other variables nor the partial coefficient of correlation between x_k and any particular one of the other variables should be computed. All of these parameters will now be without a meaning. In the rigorous case they depend on an expression of the indeterminate $\frac{0}{0}$ form and in the statistical case their values are determined by the ratio between two small quantities due to accidental errors of observation. It is even easy to construct cases where the limiting process $s_{(12\ldots)k(\ldots n)} \to 0$ is carried out in such a way that the value of the partial r may have any value between plus one and minus one (these extreme values included). Or the value of the multiple r may be made to assume any value between zero and one.

And that is not all. The standard error of these correlation parameters will also become meaningless for a similar reason. Neither the multiple nor the partial correlation parameters considered nor the standard error of these will consequently furnish the slightest indication of the cluster type or of the fact that x_k is a superfluous variable.

But in the collective alienation coefficients we have a system of criteria that can never be subject to this kind of meaninglessness, because the collective alienation coefficients, as already mentioned, are polynomials in the simple correlation coefficients. An inspection of the $(n-1)$-dimensional collective alienation coefficients would immediately tell which of the variables should be ousted from the regression system.

Now consider the case where none of the $(n-1)$-dimensional collective alienation coefficients is significantly different from zero, that is, the case where the set is multiply collinear. There now exist at least *two* independent regressions of the form (4.8). The situation is now such that whatever variable in the set $(x_1, x_2 \ldots x_n)$ be selected, the result would always be meaningless if the regression were determined of this variable on the others. Still worse, the standard errors of the regression coefficients computed by the orthodox formulae would also lose

their meaning, so that there would be no warning signal telling us to keep away from this sort of regression.

Looking back on the various cases discussed it is clear that the trouble always comes in those cases where *there exists* (rigorously or approximately) *a linear dependency between those variables that are written in the right member of the orthodox regression equation*, that is, between those variables that are considered as independent in the least square fitting procedure.

In order to get back to a basis where the regressions have a meaning, it is necessary to find those subsets that are *simply* collinear, and then to treat each of these subsets separately by reducing it to a closed set and to determine the linear regression in it. A general scheme for performing this analysis by means of the collective alienation coefficients is the following: Determine the rank p as above explained. Then select that ρ-dimensional subset for which the ρ-dimensional collective alienation coefficient is largest. Call this set the *basis set.* This is the ρ-dimensional subset that comes closest to being an uncorrelated (orthogonal) set. Then form p subsets by combining the basis set with each of the remaining variables. Each of these $(\rho+1)$ sets may be considered as simply collinear and *treated separately.* That is to say, each such $(\rho+1)$-dimensional set should be reduced to a closed set by omitting these variables x_k that are superfluous in the set according to the collective alienation coefficient criterion, as applied to this subset.

The above analysis may be illustrated by an artificially constructed problem. Ten observations were selected on four variables, x_1, x_2, x_3 and x_4, the values being written down arbitrarily except for the requirement that the sum of the variables x_2, x_3 and x_4 for each observation should equal one hundred. (When measured as deviations from their means, therefore, $x_2+x_3+x_4=0$). For convenience the set of ten values for each variable was made to total to even hundreds so that means and deviations from means would not involve decimal values. Using capital X to denote the absolute value of a variable and small x to denote the corresponding deviation from the mean, the data, as above defined, are:

X_1	X_2	X_3	X_4
25	18	29	53
14	27	43	30
37	31	51	18
17	19	34	47
24	12	46	42
29	15	35	50
41	21	28	51

39	28	41	31
52	17	57	26
22	12	36	52
300	200	400	400

The correlation coefficients for these observations which enter into the correlation matrix $(R_{(1234)})$ are:

$$r_{12}=+.169678 \qquad r_{13}=+.408675 \qquad r_{14}=-.392790$$
$$r_{23}=+.218931 \qquad r_{24}=-.689427 \qquad r_{34}=-.857719$$
$$r_{11}=r_{22}=r_{33}=r_{44}=1.00$$

That x_2, x_3 and x_4 form a closed set is shown by the values of the coefficients $s_{(234)}, s_{(23)}, s_{(24)}$ and $s_{(34)}$ for $s_{(234)}=0$, $s_{(23)}=.98$, $s_{(24)}=.72$ and $s_{(34)}=.51$. None of the last three quantities are close to zero, hence the conclusion that (x_2, x_3, x_4) form a *closed* set. The partial correlation coefficients in this set are all equal to minus one and multiple correlation coefficients are all equal to plus one. This is easily checked, as has been done, by actual computation. So far everything is all right. When, however, variable x_1 is included in the system trouble arises. If one tries to compute the partial correlations, $r_{12\cdot34}$, $r_{13\cdot24}$ or $r_{14\cdot23}$ it is found that they are all of the form $\frac{0}{0}$ and similar results hold for the multiple coefficient $r_{1\cdot234}$ and for the regression coefficients $b_{12\cdot34}$, $b_{13\cdot24}$ and $b_{14\cdot23}$.

Mr. H. I. Richards, writing in the March, 1931, number of this JOURNAL on "Analysis of the Spurious Effect of High Intercorrelation of Independent Variables on Regression and Correlation Coefficients," says "that accurate coefficients of multiple correlation and regression can be obtained when the independent variables are perfectly intercorrelated, if there are no errors in the calculations." This is fundamentally wrong and is due to certain slips in Mr. Richards' mathematics. He quotes the formulae for multiple and partial correlation, as given by Yule, as:

$$(7.1) \qquad 1-R^2_{1\cdot234\ldots n}=(1-r^2_{12})(1-r^2_{13\cdot2})(r^2_{14\cdot23})\ldots(1-r^2_{1n\cdot23\ldots\overline{n-1}})$$

$$(7.2) \qquad r_{14\cdot23}=\frac{r_{14}-r_{1\cdot23}r_{4\cdot23}}{(1-r^2_{1\cdot23})^{\frac{1}{2}}(1-r^2_{4\cdot23})^{\frac{1}{2}}}.$$

The latter is not correct, the correct formula being:

$$(7.3) \qquad r_{14\cdot23}=\frac{r_{14\cdot3}-r_{12\cdot3}r_{24\cdot3}}{(1-r^2_{12\cdot3})^{\frac{1}{2}}(1-r^2_{24\cdot3})^{\frac{1}{2}}}.$$

However, this is of minor importance in this connection. Mr. Rich-

ards' fundamental error is of a different nature and independent of whether we start from (7.2) or (7.3). He maintains that if x_2, x_3 and x_4 are perfectly correlated, for instance, by the fact that their sum is equal to one hundred, the result of computing $R_{1.234}$ by (7.1) would be the same whether all three variables x_2, x_3 and x_4 are included or one of them omitted; for example, x_4. This must be so, he claims, because

A. Factors $(1-r^2_{12})$ and $(1-r^2_{13.2})$ are the same in each case.

B. Factor $(1-r^2_{14.23})=1$ whenever $r_{4.23}=1$.

His attempt to prove proposition B is as follows:

1. Factor $(1-r^2_{4.23})$ in the denominator of (7.2) equals zero whenever there is perfect correlation between x_2, x_3 and x_4. This makes the denominator of (7.2) equal zero.
2. The numerator in (7.2) also equals zero since:
   (a) $r_{4.23}=1$
   (b) $r_{1.23}=r_{14}$
3. The form $r_{14.23}=\frac{0}{0}$ must however be equal to zero since the limit of $r_{14.23}$ is zero when $r_{4.23}$ tends toward unity.

Proposition (A) is obviously correct, and (B 1), which is the same as (B 2 a) is also correct as is seen from the formula

$$r_{4.23}=+\sqrt{1-\frac{R_{(234)}}{R_{(23)}}}=1$$

since, in the case under consideration, $R_{(234)}=0$, and it may be assumed that $R_{(23)}\neq 0$ inasmuch as the case where x_2 and x_3 are perfectly correlated is of no interest in the present connection.

But proposition (B 2 b) is not correct. The correlation $r_{1.23}$ is the correlation between x_1 and the value of the variable x_1 calculated from the regression:

$$(7.4) \qquad b_{12.3}x_2+b_{13.2}x_3$$

where the coefficients are determined so as to make the correlation *a maximum*. On the other hand, if $x_2+x_3+x_4=0$ (the x's here being deviations from means) then r_{14} can be looked upon as the correlation between x_1 and the variable:

$$(7.5) \qquad -x_2-x_3.$$

It is quite obvious that the correlation between x_1 and (7.4) need not be the same as the correlation between x_1 and (7.5). The only thing that can be said is that

$$r^2_{14}\lesseqgtr r^2_{1.23}.$$

The proposition (B 3), namely that in the case considered $r_{14.23}=0$, is not correct either. There is even a double reason for its falsity. First of all, where $x_2+x_3+x_4=0$ *there is no question of a limiting process at all.* The value of $r_{4.23}$ is not approaching as a limit the value unity, but is exactly equal to unity all the time and can equal nothing else. In the second place, even if there were a valid case for considering a limiting process at all, it is possible to show that this limiting process can be carried out in such a way as to obtain for the limiting value of $r_{14.23}$ any result whatever between $+1$ and -1. The error which Mr. Richards makes on this point is that he lets $r_{4.23}$ tend toward zero while all the other parameters involved, r_{14}, $r_{1.23}$, etc., are kept constant. This is, however, only a *very special way* of performing the limiting process that has as its final result a situation where $x_2+x_3+x_4=0$ (or where there exists some other exact linear relation between these variables). In general, such a limiting process can be performed by varying certain observational values in x_2 or x_3 or x_4. These observational values must be considered as the *independent variables* during the limiting process. If that is done, we see that the numerator and the denominator in the ratio defined by (7.2), (or by the correct formula (7.3)), *are not functions of a single variable but of several* and the limiting value of a ratio whose numerator and denominator depend on more than one variable, depends not only on the final situation toward which the system tends *but also on the path followed* in order to reach the final situation. We can illustrate this by the case of a ratio $\frac{f(x,y)}{g(x,y)}$ between two functions of two independent variables x and y. Suppose that both f and g vanish at the origin; i.e., $f(0,0)=g(0,0)=0$. What is the limiting value of the ratio $\frac{f(x,y)}{g(x,y)}$ when x and y tend toward zero? That will depend on the path which the representative point in (x,y) coördinates follows on its way toward the origin. For simplicity it may be assumed that f and g have continuous partial deviates in the vicinity of origin. The ratio considered will, therefore, in the vicinity of origin be equal to

$$\frac{f_x dx+f_y dy}{g_x dx+g_y dy} \tag{7.5}$$

where f_x, f_y, g_x, g_y denote the partial deviations at the origin and dx, dy denote the coördinates of the path along which the representative point tends toward the origin. It is quite obvious that in general the limiting value of (7.5) will depend on the ratio $\frac{dy}{dx}$, that is, it will depend on the *direction* from which the point (x,y) tends toward origin.

What Mr. Richards has proved by his limiting process is, therefore, only that it is *possible* to approach to a linear dependency between x_2, x_3 and x_4 in such a particular way that the limiting value for $r_{14.23}$ becomes 0. But he has by no means proved that this limiting value *must* be zero because there exists a linear dependency between x_2, x_3 and x_4. In Frisch's paper, already referred to, there are given examples of limiting processes that will bring the partial correlation coefficients in three variables as close as may be desired to any magnitude between -1 and $+1$ (where the limiting process tends toward representing a linear dependency between two of the variables) and it would not be difficult to give similar examples with one more variable, which is the situation here discussed.

One further comment might be to the point: Mr. Richards gives some numerical computations intended to verify the theoretical proof of his contention that one will get the same value of $R_{1.234}$ whether x_4 is included or not. However, these numerical computations cannot contain any such "verification." Either the computations must be wrong, or he must in some point or another in the computations have *made use of* that fact which should be proved, namely that $r_{14.23}=0$.