On Correlation between Variables of Constant Sum

F. CHAYES

Geophysical Laboratory, Carnegie Institution of Washington
Washington, D. C.

Abstract. Composition data are subject to the condition that the sum of the parent variables in any item is constant. This imposes a linear restraint which suppresses positive and increases negative covariance. Neither the resulting 'spurious' correlation itself nor the difficulty it creates with regard to the interpretation of composition data has been adequately described, and no general remedy has yet been suggested. This note describes some of the more important effects of a constant item-sum on correlation. It also proposes a test against the alternative of 'spurious' correlation arising from interaction between variables of equal variance, and a modification that may prove applicable to arrays characterized by inhomogeneous variance.

Introduction. In many areas of scientific inquiry the initial observations are percentages, or their conversion to percentages is a prerequisite to meaningful comparison of sample items with each other. In either event, the numerical data available for interpretation are subject to the bothersome restriction that the sum of all the variables in any item (including those of no present interest) is a constant. The effect of a constant item-sum on covariance may be both drastic and devious. It can nearly always be ignored with safety if an hypothesis can be tested by means of interrelations between variables whose contribution to the total variance of the array is small. It may also be reduced or, for practical purposes, eliminated, by relating the restricted variables to an 'outside' variable, e.g., time, position, etc., instead of to each other. Often, however, this latter stratagem merely serves to camouflage and complicate rather than eliminate the difficulty.

Neither alternative will usually be available to the petrologist, in whose work observations subject to this restriction are of central importance. Ordinarily he is concerned with interrelations between variables whose contribution to the total variance of the array is large, and only rarely is he able to press into service an 'outside' variable to reduce or mask the bias arising from the constant item-sum. Petrology thus provides excellent examples of the unavoidable use and frequent misuse of data of this sort. Indeed, these are so common that specific documentation of the illustrative material used below is neither necessary nor entirely fair.

On the necessity of negative correlation in a closed table. Letting X_{ij} represent the amount of i in the jth item in a sample of size N, $\bar{x}_i$ the sample average for i, and $x_{ij} = (X_{ij} - \bar{x}_i)$, we define as 'closed' any table containing measurements such that

$$\sum_{i=1}^{M} X_{ij} = \sum_{i=1}^{M} \bar{x}_i = K \qquad (1a)$$

for all j. The most important consequence of (1a) is simply that

$$x_{1j} + x_{2j} + x_{3j} + \cdots + x_{mj} = 0 \qquad (1b)$$

Squaring (1b), summing the squares and cross-products over $1 \leq j \leq N$, and dividing through by $(N - 1)$, we have that

$$\sum_{i=1}^{M} V_i + 2 \sum_{i=1}^{M-1} \sum_{k=i+1}^{M} p_{ik} = 0 \qquad (2)$$

where V_i is the variance of i and p_{ik} is the covariance of i and k.

Substracting x_{ij} from both sides of (1b) and repeating the squaring, summing, and division, we also have

$$\sum_{i=2}^{M} V_i + 2 \sum_{i=2}^{M-1} \sum_{k=i+1}^{M} p_{ik} = V_1 \qquad (3)$$

Next, substracting (3) from (2)

$$V_1 + \sum_{k=2}^{M} p_{1k} = 0 \qquad (4)$$

Since the numbering of variables in (1b) is arbitrary, we have thus shown that

$$V_i + \sum_{\substack{k=1 \\ k \neq i}}^{M} p_{ik} = 0 \quad \text{for} \quad 1 \leq i \leq M \tag{5}$$

i.e., every row of the covariance matrix sums to zero.

V_i being by definition positive, at least one of the $(M - 1)$ covariances attached to each variable must be negative. Now if $p_{ij} < 0$ and the others are positive, we have from (5) that

$$|p_{ij}| \geq V_i \tag{6}$$

But $|p_{ij}| \leq s_i s_j$, where s_i and s_j are standard deviations, so that

$$s_i s_j \geq V_i$$

and, dividing through by s_i,

$$s_j \geq s_i \tag{7}$$

Thus p_{ij} may be the only negative covariance of X_i if and only if $s_j \geq s_i$; otherwise (5) will require more than one negative covariance. From (7), further, if p_{ij} is in fact the only negative covariance of X_i it cannot perform the same function for X_j, which must accordingly have at least two negative covariances. The smallest number of negative covariances which will satisfy (5) for all i is $(M - 1)$. If, for instance, the $(M - 1)$ covariances attached to the variable of maximum variance are negative it is possible, at least in principle, for all other covariances to be nonnegative.

Since the sign of the covariance fixes the sign of the correlation coefficient, it follows that in any closed table containing M variables,

(*a*) Of the $(M - 1)$ correlations involving each variable, at least one must be negative. For the variable of maximum variance at least two must be negative.

(*b*) Of the $\binom{M}{2}$ total correlations that can be formed from the table, at least $(M - 1)$ must be negative. There is no a priori algebraic requirement that any of the remainder be positive, and it is quite unlikely that they will all be so unless one of the variances is very much larger than the others.

Dividing (5) by s_i gives

$$s_i + \sum_{\substack{k=1 \\ k \neq i}}^{M} r_{ik} s_k = 0 \tag{8}$$

from which, since $/r_{ik} s_k/ \leq s_k$, it is evident that if s_i is greater than the sum of any j of the other standard deviations, at least $(j + 1)$ of the covariances of variable i must be negative. In particular, if $j = (M - 2)$, all the covariances of variable i will be negative. If we are told, for instance, that standard deviations of 0.5, 1.4, 1.5, 3.0, are associated, respectively, with variables X_1, X_2, X_3, X_4 in a 4-variable system, we know at once that all covariances involving X_4 must be negative.

It is easy to show, however, that positive correlation must exist somewhere in the array if some one of the variances, say V_1, is enough larger than the others. If

$$V_1 \geq \sum_{2}^{M} V_i$$

we have, because of (8), that all the covariances of X_1 are negative, and from (3) that one or more of the covariances relating variables X_2, X_3, $X_4 \cdots X_M$ must be positive.

Although

$$V_1 > \sum_{2}^{M} V_i$$

is evidently a sufficient condition for the emergence of positive correlation among variables $2 \leq i \leq M$, I do not believe it can be shown to be necessary except if $M = 3$, for which see below.

Correlation in a closed table with three variables. If $M = 2$ the whole notion of correlation is, of course, trivial, for if $X + Y = \bar{x} + \bar{y} = K$, it is obvious that $V_x = V_y$ and $r_{xy} = -1$. When $M = 3$ the situation is far more complex, as much petrographic experience attests. It is nevertheless true that although each of the three correlation coefficients is now in principle free to vary from -1 to $+1$, any assumed or observed set of variances completely fixes all three coefficients. (Whether we regard the variances as dependent on the covariances, or vice versa, is, to some extent, a matter of taste. In a descriptive science it is always desirable to classify objects before discussing their relations with each other. From this point of view variance appears a more fundamental property than covariance, and

throughout this note, accordingly, it is taken as independent.) We first prove this assertion.

Using (5) to obtain

$$V_i + p_{ij} + p_{ik} = 0 \qquad (11)$$

we readily find, by rotating subscripts in (11) and solving the resulting set of simultaneous equations, that

$$p_{ij} = \tfrac{1}{2}[V_k - (V_i + V_j)] \qquad (12)$$

from which

$$r_{ij} = \frac{1}{2}\left[\frac{V_k - (V_i + V_j)}{s_i s_j}\right] \qquad (13)$$

so that if $M = 3$, r_{ij} is a single valued function of V_i, V_j and V_k.

We note that $r_{ij} > 0$ if and only if $V_k > (V_i + V_j)$. Positive correlation need not occur at all, but if it does appear it is confined to the relation between the variables of least and intermediate variance. Correlation between the variable of maximum variance and each of the others must be negative. If $V_k = V_i + V_j$, variables i and j will appear to be unrelated, in the sense that $r_{ij} = 0$. If no variance is greater than the sum of the other two, all three correlations will be negative.

We do not ordinarily suppose that the choice of mineral or oxide ranges, upon which most petrographic classifications are based, determines whether the variables concerned vary directly or inversely, or how strongly they do either. Yet (13) shows that something very like this must happen if $M = 3$. Whether, in the sialic fraction of a series of granites, for instance, quartz and alkali-feldspar tend to vary directly will depend on how variable plagioclase is. Unless the variance of plagioclase is larger than the sum of the variances of quartz and alkali-feldspar, there is no possibility of positive correlation between the latter, and unless it is considerably larger than this sum the correlation will fail of significance. Now in rocks called 'granite' the permissible range—and hence to a large extent the variance—of plagioclase content will depend on taxonomy and nomenclature. In deciding that classification is not worth worrying about, petrologists have in effect decided that this question is not worth answering.

Equation 5 is also useful in setting limits to the permissible variation in a three variable classification. Dividing its expansion for $M = 3$ (equation 11 above) by s_i gives

$$s_i + s_j r_{ij} + s_k r_{ik} = 0 \qquad (14)$$

and since $r \geq -1$ it follows immediately that

$$\text{and} \qquad \begin{aligned} s_i &\leq s_j + s_k \,, \\ s_i &\geq |s_j - s_k| \end{aligned} \qquad (15)$$

The force of (15) will perhaps be made a little clearer by the reminder that in the absence of (1) there is no a priori relation between s_i, s_j, and s_k. In any array subject to (1), however, it will always be true that, if $M = 3$, the sum of any two of the standard deviations will equal or exceed the third.

Correlation in a closed table with four variables. We have just noted that if $M = 3$ the correlations are completely fixed by any legitimate choice of variances, so that hypotheses about covariance are no more than disguised hypotheses about variance. The interrelation between variance and covariance is both more complex and less specific if $M = 4$. Writing (1) in the form

$$x_i + x_j = -x_k - x_l$$

squaring, summing, and dividing by $(N - 1)$, as before, we obtain

$$V_i + V_j + 2p_{ij} = V_k + V_l + 2p_{kl} \qquad (16)$$

Bearing in mind that $V_i = s_i^2$ and $|p_{ij}| \leq s_i s_j$, we find, after some rearrangement, that

$$\frac{1}{2s_i s_j}[(s_k - s_l)^2 - (s_i^2 + s_j^2)] \leq r_{ij}$$

$$\leq \frac{1}{2s_i s_j}[(s_k + s_l)^2 - (s_i^2 + s_j^2)] \qquad (17)$$

an inequality which is the 4-variable analogue of (13). It is obviously much less restrictive; if, for instance, the variances are taken as equal, we have from (13) that $r_{ij} = -0.5$ exactly, but from (17) only that $-1 \leq r_{ij} \leq +1$.

Since the latter range is precisely that characteristic of open data, it is tempting to assume that if $M \geq 4$ the 'spurious' correlation characteristic of the closed form becomes small enough to ignore. Unfortunately, this is not so.

Despite our inability to fix the exact value of any correlation from a knowledge of the variances, it is clear from (5) that *at least 3 of the coefficients in any set of 6 must be negative.*

For $M = 4$, eq. 5 becomes

$$V_i + p_{ij} + p_{ik} + p_{il} = 0 \qquad (18)$$

so that

$$s_i + s_j r_{ij} + s_k r_{ik} + s_l r_{il} = 0 \qquad (19)$$

Since $r \geq -1$ it follows that

$$\left.\begin{aligned} s_i &\leq s_j + s_k + s_l \\ |s_i - s_j| &\leq s_k + s_l \end{aligned}\right\} \qquad (20)$$

and

The largest standard deviation must not be greater than the sum of the other three, and the sum of any two must not be less than the difference between the other two. Extension of these results to the multivariate case is immediate. The largest standard deviation will never be larger than the sum of the other $(M - 1)$, and the difference between the largest and the smallest will never be larger than the sum of the other $(M - 2)$.

By forming a new variable, $X_{(k+l)} = X_k + X_l$, we may examine more closely the relation between r_{ij} and r_{kl}. For we then have, from (13), that

$$r_{ij} = \frac{1}{2}\left[\frac{V_{(k+l)} - (V_i + V_j)}{s_i s_j}\right] \qquad (21)$$

which will be positive if and only if $V_{(k+l)} > V_i + V_j$. Expanding the left side of this inequality by the usual rule for the variance of a sum, and rearranging terms, we find that

$r_{ij} > 0$ if and only if

$$r_{kl} > \frac{1}{2s_k s_l}[(V_i + V_j) - (V_k + V_l)] \qquad (22)$$

In the example used earlier, $s_1 = 0.5$, $s_2 = 1.4$, $s_3 = 1.5$, $s_4 = 3.0$, and it was shown that, because of (8), r_{14}, r_{24}, and r_{34} must all be negative. By means of (22) we may now show that r_{12}, r_{13} and r_{23} must all be positive. Thus the signs of all six correlations are fixed by (1) and the sample variances. More generally, we may note from inspection of (22) that:

(*a*) If $(V_k + V_l) - (V_i + V_j) > 2s_k s_l$, r_{ij} will be positive.

(*b*) If $(V_k + V_l) = (V_i + V_j)$, r_{ij} will be positive if and only if k_{kl} is positive.

(*c*) If $(V_i + V_j) - (V_k + V_l) > 2s_k s_l$, r_{ij} will be negative.

As a special case of (*b*), if the variances are equal the number of positive correlations must be either two or zero, and $r_{ij} = r_{kl}$.

Whether or not the variances are equal, an hypothesis that implies some specific relation between any two of the variables implies some specific, though not necessarily identical, relation between the other two; this is evident from (16), so that if r_{kl} is known there is no need to compute r_{ij}. It may be difficult to decide objectively which of these correlations is to be tested for significance, but clearly it will be specious to test both.

Finally, if any two covariances having a common variable are known, the remaining four covariances may be expressed as additive functions of these two and the variances; the third covariance of the variable may be gotten by difference, from (5), and the three from which it is lacking by successive application of (16).

Thus, if we are working with assigned variances—such as are often implied by petrographic classifications—and the hypothesis in question implies specific values for ρ_{ij} and ρ_{ik}, it also implies equally specific expectations for r_{il}, r_{jk}, r_{jl}, and r_{kl}. In a 4-variable closed table there will never be more than two potentially independent correlations.

These essentially nonstatistical controls over covariance greatly complicate the statistical testing and interpretation of correlation in closed arrays. It is fair to add—in fact it is one of the principal objectives of this note to point out—that the difficulty is not eliminated by rejecting statistics in favor of conventional interpretive procedures, whether of the 'genetic' or 'common sense' variety. By thinking solely in terms of geochemical associations, gradual transitions, petrographic affinities, liquid descent lines, etc., we may manage to remain comfortably unaware of the problem; we do not, however, solve it.

Inferences about open variables from observations on closed variables. Enough has been said to indicate that the problem is puzzling and serious. Unless the number of variables is much larger than is common in petrography, or the quantity

$$(V_k + V_j) \ll \sum_1^M V_i$$

the effect of the closed form of statement on r_{kj} cannot be ignored. What can be done about it?

On the assumption that the *parent* variables are in fact subject to no such restraint, and that the percentage form of statement is merely an unavoidable condition of observation, *Sarmanov and Vistelius* [1958] have recently pointed out that its effect can be removed if some one of the parent variables may be presumed constant. The trick is simply to divide each closed variable representing an open variable of nonzero variance by the closed variable representing the open variable of zero variance.

Using α for closed and x for open variables, we define

$$\alpha_i = \frac{x_i}{\sum_1^n x_i}$$

and if, by hypothesis or assumption, $V_1 > 0$, $V_2 = 0$, $V_3 > 0$ for instance, we form new variables

$$\beta_1 = \frac{\alpha_1}{\alpha_2} = \frac{x_1}{x_2}; \qquad \beta_3 = \frac{\alpha_3}{\alpha_2} = \frac{x_3}{x_2}$$

Since x_2 is constant, the correlation of β_1 with β_3 will be exactly that which would be calculated between x_1 and x_3 in the same sample, if the open variables could be measured directly. The authors refer to this as the 'concretionary scheme.'

Sarmanov and Vistelius also show, in the same paper, that if two open variables—say, x_2 and x_4—are independent of each other and of the remaining variables, the correlation coefficient between the ratios

$$\beta_1 = \frac{\alpha_1}{\alpha_2}; \qquad \beta_3 = \frac{\alpha_3}{\alpha_4}$$

will be of the same sign as that between x_1 and x_3, but smaller; this arrangement they call the 'metasomatic scheme.'

Their results can also be obtained as approximations from the *Pearson* [1896–1897] general formula for ratio correlation [*Reed*, 1921; see also *Chayes*, 1949, p. 241, (2)]. For the 'concretionary' scheme the Pearson general formula gives

$$r_{\beta_1\beta_3} = \frac{r_{13}C_1C_3 + C_2^2}{(C_1^2 + C_2^2)^{1/2}(C_3^2 + C_2^2)^{1/2}} = r_{13} \quad (23)$$

since $C_2 = 0$ (C is the coefficient of variation). For the 'metasomatic' scheme Pearson's general formula reduces to

$$r_{\beta_1\beta_3} = \frac{r_{13}C_1C_3}{(C_1^2 + C_2^2)^{1/2}(C_3^2 + C_4^2)^{1/2}} \leq r_{13} \quad (24)$$

since in this arrangement $C_2 > 0$, $C_4 > 0$. Note that the sign of the ratio correlation will be the same as that of r_{13}, as required. Pearson's derivation is an approximation in the sense that it assumes normally distributed 'open' variables with C small enough so that 3d and higher powers of it may be ignored. Sarmanov and Vistelius now show that the solution for the 'concretionary' scheme requires neither assumption, and that the same holds for the 'metasomatic' scheme if no numerical estimate of the ratio correlation is required.)

In the Sarmanov-Vistelius approach, as in Pearson's original discussion of index correlation, interest centers on hypothetical parent variables *not* subject to (1). From hypothesis or prior knowledge we are able to make reasonable assumptions about certain of the open parent variables; these assumptions enable us to make valid inferences about relations between other open variables, from data that can be obtained only in closed form.

In many petrographic problems, however, we have no way of deducing the necessary a priori relations between the open variables. In still others, including some of the commonest and most important, we have no particular reason to suppose the open variables exist. Equation (1) and the 'spurious' correlation to which it gives rise are always encountered in percentage data, whether or not we are able to regard the percentage form of statement as a mere condition of observation. For the analysis of closed data without assumption about the nature or even the existence of underlying open variables, we obviously require a technique quite different from that proposed by Sarmanov and Vistelius.

On the value of ρ indicative of unrelatedness in a closed array. For the open variables of ordinary experience we use $\rho = 0$ as a criterion of independence, concluding that if the quantity $|r_{12} - 0|$ is not sufficiently large the sample offers no reason for supposing $\rho_{12} \neq 0$. This convention is useful not only because of the intuitively appealing relation it establishes between regression and correlation, but also because in samples drawn from a normal population the distribution of r about $\rho = 0$ is known, so that a simple significance test is possible.

The whole arrangement presumes that the

variables might be statistically independent, however, and we have already seen that the variables of a closed array cannot be independent. Indeed, if the number of variables is reasonably small—about that normally encountered in petrology—the signs and sometimes even the relative sizes of many of the correlations can be established from an examination of the variances alone, with no reference to, and possibly also with little bearing on, the causal nexus often inferred from these associations. The problem of deciding when a particular correlation is strong enough to warrant the inference of nonrandom association remains. We require some working definition of 'unrelatedness' in a closed array, and, based on this definition, a numerical value of ρ to replace the zero of the conventional null hypothesis.

From (2) we note that in any closed array

$$\sum_{i=1}^{M-1} \sum_{j=i+1}^{M} p_{ij} = -\frac{1}{2} \sum_{1}^{M} V_i \qquad (25)$$

If as an entry to the problem, we suppose the variances equal, the right side of (25) is simply $-M\sigma^2/2$. Dividing this by the number of items on the left side of the equation, we obtain for the average, or expected value, of the covariance

$$E(p_{ij}) = -\frac{M\sigma^2}{2\binom{M}{2}} = \frac{\sigma^2}{1 - M} \qquad (26)$$

and, dividing through by $\sigma_i\sigma_j$,

$$\rho = (1 - M)^{-1} \qquad (27)$$

Thus, if the only relation between a set of variables of equal variance is the restraint imposed by (1)—so that they are, from the petrographic point of view, unrelated—the expected value of the sample correlation coefficient is $(1 - M)^{-1}$. If our objective is to detect departures from randomness in the association of any two of the variables, we should test r not against $\rho = 0$, as in the conventional null hypothesis, but against $\rho = (1 - M)^{-1}$, by means of the Fisher z transformation[1].

[1] Although the emphasis throughout this paper is on correlation, it is obvious that equation 5 also imposes severe restrictions on regression coefficients. Dividing (5) by V_i we have at once that $\sum b_{ki} = -1$, so that $E(b_{ki}) = (1 - M)^{-1}$, just as for ρ, though in the case of b no assumption about variance homogeneity is required. In the traditional 'variation diagram' of petrology the other six major oxides are plotted against SiO_2, and the small negative slope of many of these indicated regressions is well known. The possibility that they are to be regarded as estimates of $(1 - M)^{-1}$ in the null case is well worth exploring; I hope to discuss it later in another place.

In practice, of course, the sample variances will nearly always be unequal. The very fact that the variance of an accessory or minor constituent must be small while that of a major one may be large practically assures variance inhomogeneity in any sample of reasonable size. The use of $(1 - M)^{-1}$ as a criterion of unrelatedness provides more protection than the null hypothesis against errors of the first kind in the testing of negative correlations, and against those of the second kind in testing positive ones. But if the variances are markedly inhomogeneous, the protection it affords will certainly vary from comparison to comparison.

Expanding (5) for V_1 and V_2 and subtracting the second expansion from the first, we have, after eliminating $p_{12}(=p_{21})$ and rearranging terms, that

$$\sum_{3}^{M} (p_{2j} - p_{1j}) = V_1 - V_2 \qquad (28)$$

so that

$$E(p_{2j} - p_{1j}) = \frac{V_1 - V_2}{M - 2} \quad \text{for} \quad j \geq 3 \qquad (29)$$

and if $V_1 \geq V_2$ the expected covariance p_{2j} will necessarily be a smaller negative number than the expected covariance p_{1j}. (It is obvious from (5) that the expected covariance $E(p_{jk}) \leq 0$ for any $j \neq k$.) Since $\sigma_1 > \sigma_2$ this also follows for the correlation coefficients formed from these covariances unless $p_{2j}/p_{1j} < \sigma_2/\sigma_1$. Except in this latter circumstance, then, it is to be expected that in the absence of nonrandom associations between the variables of a closed array the larger negative correlations will be found among the variables of larger variance.

In view of this it is reasonable to inquire how much sense it makes to test every correlation in an array against $(1 - M)^{-1}$. The answer, I believe, is that the procedure is quite sound if the ratio of maximum to minimum variance is small, and little superior to the null hypothesis if this ratio is large. Correlation analysis of a

rather extensive collection of modal analyses (nearly all those published by Y. Suzuki and myself, together with considerable unpublished material of my own) indicates that in granitic rocks the range of variances is large enough so that routine use of $(1 - M)^{-1}$ as a criterion of unrelatedness would often be inefficient and potentially misleading.

A summary of these computations is given in Table 1, in which the quantity M' denotes the variance rank of the least variable member of the group whose average correlation is shown in the column headed $r(\bar{z})$. Numbering the variances in an array in order of decreasing size, the $r(\bar{z})$ entry in the row headed $M' = 4$, for instance, is the r corresponding to

$$\bar{z} = \tfrac{1}{6}(z_{12} + z_{13} + z_{14} + z_{23} + z_{24} + z_{34})$$

and so forth. Comparison of the columns headed $r(\bar{z})$ and $(1 - M')^{-1}$ indicates that in this particular body of data the quantity $(1 - M')^{-1}$ is a rather good estimator of average correlation between variables of variance rank $\leq M'$, for $M' > 2$. How general this relation may be this author is not yet able to say. To propose its use in a routine significance test would certainly be premature; it nevertheless seems strong enough to warrant further careful study.

Numerical experimentation. Although (27) is essentially algebraic rather than statistical, it is such a curious result that some test of it seems desirable. Fortunately, it is not difficult to construct a suitable test in any high-speed calculator of reasonable capacity. The procedure is to generate in the machine an $M \times N$ array of uncorrelated positive variates of mean μ and variance σ^2, divide each entry in the jth row by

$$\sum_{i=1}^{M} X_{ij}$$

for $1 \leq j \leq N$, calculate the covariance matrix for the resulting $M \times N$ closed array, compute the average Fisher z for the $(M)(M - 1)/2$ covariances, and transform $\bar{z}$ to $r(\bar{z})$. (On a machine of the capacity of the IBM 704 this requires considerably less than 30 seconds if $M \leq 6$ and $N \leq 400$.)

ρ between each pair of columns of the original open array is zero. The expected mean of each column of the closed array is $1/M$. Ignoring a bias which decreases rapidly with increase in M, and is in any event the same for each column of any particular array, the expected variance per column of the closed array is $\sigma^2\mu^{-2}M^{-3}(M - 1)$. If (27) is correct, $r(\bar{z})$ calculated from an array generated in this fashion should be such as might be expected from a parent having $\rho = (1 - M)^{-1}$. The results of a test of this kind, made with two runs of $N = 400$ at each value of M, are shown in Table 2. Further testing of (27) would appear to be superfluous.

TABLE 1. Average 'Effective Variance' or 'Variance Rank' Correlation for 543 Modal Analyses[1] For definition of M', see text.

M'	$\bar{z}(M')$	$r(\bar{z})$	$(1 - M')^{-1}$	Number of Groups	Number of Analyses
2	−0.9175	−0.7247	−1.0000	33	543
3	−0.5113	−0.4710	−0.5000	33	543
4	−0.3461	−0.3329	−0.3333	33	543
5	−0.2494	−0.2444	−0.2500	32	526
6	−0.1929	−0.1905	−0.2000	27	476
7	−0.1710	−0.1694	−0.1667	2	41

[1] Data of Suzuki and Chayes.

A similar test of the quantity $(1 - M')^{-1}$, which we may refer to as the 'variance rank correlation,' is more difficult because if either the means or the variances of the open variables differ, the relation between open and closed variances becomes rather complicated. The effect illustrated in Table 1 can hardly be as general as that just discussed; indeed, it is a rather reasonable guess that it will not hold unless the range of (closed) means and variances is quite considerable. A useful test thus requires the generation of an open array whose transformation will yield a closed array with variables characterized by means and variances at least roughly comparable with those encountered in practice.

TABLE 2. Test of Equation 27

M	$\rho = (1 - M)^{-1}$	$r(\bar{z})$
3	−.5000	−.5005
4	−.3333	−.3336
5	−.2500	−.2506
6	−.2000	−.2005

It is nearly self-evident that

$$E\left(\frac{X_{i,j}}{T_j}\right) = \frac{\mu_i}{\tau} \qquad (30)$$

where

$$\tau = \sum_{i=1}^{M} \mu_i, \quad \rho_{ik} = 0, \quad T_j = \sum_{i=1}^{M} X_{i,j},$$

and Greek letters refer to population parameters of the open variables.

Using a theorem of *Fieller* [1939] for the variance of a ratio, and the well-known formula for the part-whole correlation [see *Snedecor*, 1956, p. 189], it may be shown that

$$\mathrm{Var}\left(\frac{X_i}{T}\right) = \left(\frac{\mu_i}{\tau}\right)^2\left[\frac{\sigma_i^2}{\mu_i^2} + \frac{\sigma_t^2}{\tau^2} - 2\frac{\sigma_i^2}{\mu_i \tau}\right] \qquad (31)$$

where, in addition to the previous conditions, we also require that

$$\sigma_t^2 = \sum_{i=1}^{M} \sigma_i^2$$

which will of course be true if $\rho_{ik} = 0$ for all i and k, $i \neq k$.

TABLE 3. Bellingham Granite Data ($N = 15$) Compared with Statistics Computed from a Numerical Model ($N = 400$) Generated from Uncorrelated Open Variables

A. Means and Standard Deviations

Mineral	Averages		Standard Deviations	
	Observed	Computed	Observed	Computed
Quartz	29.4	29.5	4.42	4.63
K-feldspar	34.2	33.9	5.98	5.89
Plagioclase	29.9	30.1	4.54	4.38
Biotite	4.5	4.5	2.49	2.45
Muscovite	2.1	2.0	1.01	1.04

B. Average Correlations

M'	$\bar{r}(M')$		$(1 - M')^{-1}$
	Observed	Computed	
2	−.65	−.57	
3	−.47	−.46	−.50
4	−.34	−.31	−.33
5	−.22	−.21	−.25

The right side of (30) is the expected mean of the ith closed variable, and the right side of (31) is its expected variance. Although the relation between open and closed variances is very far from simple—it will be noted from (31), for instance, that the closed variance cannot be zero even if the corresponding open variable is a constant—it is possible, by repeated application of (31), to design open arrays which, on transformation, yield closed variables characterized by realistic means and variances. An example is shown in Table 3; the 'real data' are calculated from unpublished modes of 15 specimens of the Bellingham, Minn., granite. The means and variances of the 'numerical model,' generated from uncorrelated open variables, agree fairly well with those of the actual measurements, and the agreement could be improved by further trial and error applications of (31). The average values of $\bar{r}(M')$ obviously show the same general trend in real modes and numerical model. It is clear that in this array a significance test based on $\rho = 0$ is completely unwarranted, a test based on $\rho = (1 - M)^{-1}$ is considerably more realistic, and one based on $\rho = (1 - M')^{-1}$ even more so. How often, or under what circumstances, $(1 - M')^{-1}$ is preferable to $(1 - M)^{-1}$ as a criterion of unrelatedness I do not yet know. The treatment of this aspect of the problem is still scarely more than preliminary.

Some petrological considerations. Although petrologists generally prefer some form of graphical evaluation to formal statistical analysis of their data, the underlying assumptions of the two procedures are essentially identical. By 'lack of relation' the petrologist usually means just about what is implied by a sample correlation which fails of significance against the alternative that $\rho = 0$. In refraining from a formal test he sacrifices efficiency, accepts the risk of failing to detect minor subjective differences between observers, and devotes to drafting the time he might otherwise spend in calculating. But his preference for graphical procedures, however wasteful is not essentially unsound provided only that the assumptions underlying correlation analysis are in fact valid in the context of his work.

It is the chief burden of this note that these assumptions are decidedly *invalid* in studies of relations between the major constituents of a closed array. Such studies are of course a mainstay of chemical and theoretical petrology.

By means of one type or another of variation diagram we incessantly ask what relations hold between the variables, without bothering to inquire whether they are related by more than the circumstance that they have certain means and variances and exist in the same body at the same time. With a little ingenuity it is usually possible to devise some projection in which the points will be close enough to a simple enough curve that a new magmatic or metasomatic 'trend' can be announced, or an old one confirmed.

Between major constituents the great majority of such correlations are negative, and with regard to 'genetic' interpretations of these, the moral of the preceding discussion is painfully obvious. In small samples—even in samples considerably larger than those we often use—a negative correlation must be very strong indeed to be significant; the proper criterion of unrelatedness is much more likely to be in the vicinity of -0.5 than of the zero which is explicitly assumed in the null hypothesis and tacitly assumed in graphical procedures. This may be annoying but is not, after all, particularly bewildering; and the remedy is as obvious as the difficulty. Either we must confine our attention to exceedingly strong negative correlations or we must materially enlarge our samples.

The interpretation of small positive sample correlations, or indeed of those which fall between zero and the appropriate negative criterion of 'unrelatedness,' on the other hand, seems a first-class puzzle. A sample correlation of zero, for instance, may indicate a highly significant departure from random association in the direction of what we ordinarily regard as positive correlation. Yet the correlated variance of whichever variable is taken as dependent will be negligible, information about either variable will permit no reliable inference about the other, and in a scatter diagram the data points will distribute themselves in the fashion we have come to regard as indicative of randomness or nonrelation. In short, the usual elegant relation between the geometrical and analytical aspects of product-moment correlation vanishes completely. So, too, does the possibility of making sense of the data by the customary empathic appreciation of variation diagrams.

An example of this conflict between interpretations of 'open' and 'closed' correlations occurs in the data for Bellingham. Since plagioclase and quartz have, respectively, the variance ranks 2 and 3, our argument suggests that in the absence of nonrandom effects the correlation coefficient for the pair should be in the neighborhood of -0.5, and, as shown in Table 3, the average correlation of the three most variable constituents is very close to this. The observed correlation between quartz and plagioclase, however, is $+.0086$. A sample value of $r = +.0086$ as between a pair of open variables would certainly be taken to indicate the absence of nonrandom factors governing the association. Given the information that the variables are members of a 5-variable closed array, and have variance ranks 2 and 3, we should have to announce an exactly opposite conclusion.

Acknowledgments. I am indebted to J. M. Cameron for much helpful discussion, particularly of the relation between closed and open variances, and to Mrs. R. W. Varner, who supervised all and did some of the programming. The calculations leading to Tables 1-3 were run at the National Bureau of Standards.

References

Chayes, F., A petrographic criterion for the possible replacement origin of rocks, *Am. J. Sci., 246,* 413–425, 1948.

Chayes, F., Ratio correlation in petrography, *J. Geol., 57,* 239–254, 1949.

Fieller, E. C. The distribution of the index in a normal bivariate population, *Biometrika, 24,* 428–440, 1932.

Pearson, K., On a form of spurious correlation, etc., *Proc. Roy. Soc.* (London), *60,* 489–502, 1896–1897.

Reed, L. J., On the correlation between any two functions, etc., *Wash. Acad. Sci. J., 11,* 449–455, 1921.

Sarmanov, O. V., and A. B. Vistelius, On the correlation of percentage values, *Doklady Akad. Nauk SSSR, 126,* 22–25, 1958. (In Russian.)

Snedecor, G. W., *Statistical Methods,* 5th ed., Iowa State College Press, 1956.

(Manuscript received September 27, 1960.)