| | |
|---|---|
| TITLE: | Journal of the International Association Mathematical Geology |
| VOLUME/ISSUE/PAG | 13 / 2 175-189 |
| DATE: | 1981 |
| AUTHOR OF | Aitchison, J. |
| TITLE OF | A new approach to null correlations of p: |
| ISSN: | 0020-5958 |
| OTHER | OCLC: 1589202 |
| CALL NUMBER: | QE1 .I465 |
| DELIVERY: | FTP-to-Ariel: 129.82.28.195 |
| REPLY: | Mail: |

*Mathematical Geology, Vol. 13, No. 2, 1981*

# A New Approach to Null Correlations of Proportions[1]

**J. Aitchison[2]**

*Much work on the statistical analysis of compositional data has concentrated on the difficulty of interpreting correlations between proportions with an assortment of tests for null-correlations, for independence except for the constraint, F-independence of bounded variables, neutrality in the mean and in the median. This paper questions the appropriateness of characterizing the dependence structure of proportions in terms of such concepts, suggests an alternative method of modeling, develops necessary distribution theory and tests, and illustrates the methodology in applications.*

**KEY WORDS:** closed and open array, compositional data, correlations between proportions, Dirichlet distributions, logistic-normal distributions, petrogenesis, tests for basis independence.

## INTRODUCTORY REVIEW OF THE PROBLEM

Some 20 years ago misinterpretations of correlations between proportions in the analysis of modal or compositional data became a matter of concern almost simultaneously in geology (Chayes, 1960, 1962; Krumbein, 1962; Chayes and Kruskal, 1966) and in biology (Mosimann, 1962, 1963). And still the debate continues (Butler, 1979). The difficulty arises because a basis or open vector of uncorrelated positive quantities $x_1, \ldots, x_{d+1}$ leads to a composition or closed vector of proportions $y_i = x_i/(x_1 + \cdots + x_{d+1})$ $(i = 1, \ldots, d+1)$ which are necessarily correlated. How then are any apparent correlations in compositional data to be interpreted: as indicative of nonzero correlations in the basis or as merely induced through the process of forming a composition from an uncorrelated basis? The early work on this problem concentrated on determining the values of the induced or null correlations under a variety of model assumptions and on suggesting significance tests for comparing computed correlations against the null values. These tests were always presented with some hesitation for three important reasons.

[1]Manuscript received 23 July 1980; revised 6 October 1980.
[2]Department of Statistics, University of Hong Kong, Hong Kong.



0020-5958/81/0400-0175$03.00/0 

1. The distributions of the test statistics are not known (Mosimann, 1962, p. 81; Chayes and Kruskal, 1966, p. 696) and do not fall within the framework of any standard testing approach such as generalized likelihood ratio tests.
2. The tests of null correlations are carried out separately for each pair of proportions. This procedure therefore is open to the same kind of criticism as the application of all $\frac{1}{2}k(k-1)$ $t$-tests of pairwise comparison of $k$ treatments without a preliminary overall $F$-test. The theory lacks the analogue of such an overall test (Chayes and Kruskal, 1966, p. 696).
3. When the tests detect nonnull correlations it is by no means safe (Miesch, 1969) to conclude that the corresponding quantities in the basis are uncorrelated. Thus, despite the fact that the battery of pairwise tests, criticized in point (2), is not designed as an overall test of the hypothesis that all correlations of the basis are zero, this hypothesis is the only one which the battery effectively tests. No satisfactory analysis of the nonnull case is available.

More recent work has largely been an attempt to introduce new concepts of nonassociation for proportions and relevant tests of significance: neutrality of one proportion with respect to another (Connor and Mosimann, 1969), neutrality in the mean (Darroch, 1969; Darroch and Ratcliff, 1970, 1971; Bartlett and Darroch, 1978), $F$-independence (Darroch and James, 1974), and neutrality in the median (Darroch and Ratcliff, 1978). It remains to be seen whether these concepts, naturally more sophisticated than the concept of open correlations, prove straightforward enough for geologists to interpret. There is, however, a more fundamental difficulty. The properties of $F$-independence and neutrality lead almost inevitably to the description of variability through the Dirichlet class of distributions. The fact that, in the words of Darroch and James (1974, p. 479), "the Dirichlet distribution is almost the only one defined for continuous, positive, bounded-sum random variables which is easily handled for inference and descriptive purposes" then leads to the awkward question of how $F$-dependence and nonneutrality are to be analyzed.

This paper suggests that the unsatisfactory features of the above theories can be largely remedied by concentration on three fundamental aspects:

a. a fuller appreciation of the relationship of closed to open variables, or, in the terminology of this paper, of composition to basis, and in particular the extent to which inferences can be made from compositional data to basis models;
b. the introduction, as a consequence of (a), of a form of modeling which more simply, directly, and tractably connects independence and nonassociation of the components of a basis to properties of the corresponding composition; and
c. the identification of a rich enough parametric class of distributions for com-

positional data which allows the description of both nonassociation and association within a single framework.

Since the measures of dependence used are covariances the main thrust of the paper may be seen as an attempt to resolve some of the difficulties of the null-correlation approach. As a bonus the resolution of (c) provides a tool which opens the way to further developments of the other approaches. For a more detailed discussion of some of these developments, see Aitchison (1981a,b).

## THE RELATIONSHIPS OF BASES AND COMPOSITIONS

### Algebraic Considerations

From any $(d+1)$-dimensional vector $x$ of positive quantities a $d$-dimensional vector $y$ defined by

$$y_i = x_i/(x_1 + \cdots + x_{d+1}) \qquad (i = 1, \ldots, d)$$

can be formed. We then write $y = C(x)$ and call $y$ the *composition* of the *basis* $x$. We note that $y$ is a vector of bounded sum 1 in the sense of Darroch and James (1974), since $e_d^T y \leqq 1$, where $e_d$ is the $d$-dimensional vector of units. Moreover, it is clear that $y$ adequately describes the proportions of the constituents of $x$ since the proportion of the $(d+1)$th constituent is $y_{d+1} = 1 - y_1 - \cdots - y_d$. The use of $y$ rather than the augmented $\{y, y_{d+1}\}$ is mathematically more convenient since a composition is a $d$-dimensional rather than a $(d+1)$-dimensional entity, belonging to the $d$-dimensional simplex

$$S^d = \left\{ y : y_i > 0 \qquad (i = 1, \ldots, d), \qquad \sum_{i=1}^{d} y_i < 1 \right\}$$

It is well recognized that in many, probably most, geological applications an underlying basis is more conceptual than real, a convenient peg on which to hang the discussion of nonassociation of proportions. What seems less clearly understood is the rather limited nature of the inferences possible about such conceptual bases from information on compositions. Starting at a purely algebraic level we see immediately that, given a composition $y$, there is no way of uniquely reconstructing its basis. For if $x$ is such a basis so that $y = C(x)$ then, since $C(zx) = C(x)$, where $z$ is a constant scalar or a random variable, we see that $zx$ is also a basis. In fact the class of bases leading to a composition $y$ is characterized by this multiplicative property, and the property of common composition defines equivalence classes of bases. In geometric terms for $d = 1$ the bases are points in the positive quadrant and compositions can be represented by points on the line segment from (1, 0) to (0, 1). The equivalence class of bases corresponding to a given composition $y$ is formed by points on the ray from the origin through $y$.

Since a composition is uniquely, and most simply, specified by the values of $y_1, \ldots, y_d$ and hence $y_{d+1} = 1 - y_1 - \cdots - y_d$ the main thrust of the correlation approach to dependence has been in terms of correlations between $y_i$ and $y_j$ and the complications of interpretation. But the composition could equally well be specified in terms of any other $d$-dimensional vector $v$ related to $y$ through a one-to-one transformation. In searching for a sensible such transformation we should surely recognize the ability of a composition to determine a basis only up to a multiplicative factor. This immediately suggests the use of the ratio transformation $v_i = y_i/y_{d+1} (i = 1, \ldots, d)$ or even better, the logratio transformation

$$v_i = \log (y_i/y_{d+1}) = \log y_i - \log y_{d+1} \qquad (i = 1, \ldots, d) \tag{1}$$

since differences are usually simpler to handle than ratios. The inverse transformation of (1), from $v$ to $y$, is the generalized logistic transformation

$$y_i = \exp (v_i) \Big/ \left\{1 + \sum_{j=1}^{d} \exp (v_j)\right\} \qquad (i = 1, \ldots, d) \tag{2}$$

This may seem at first sight an exotic tool for the analysis of compositions and lead to very complicated interpretation problems. The opposite is the case; the transformation provides a natural and simple link between compositions and their equivalent bases and so probes to the root of the proportion correlation problem. The reason for such a simplification is not hard to find. The algebraic difficulty with compositions is their confinement to the simplex $S^d$, a difficult set to handle mathematically, whereas the corresponding space $R^d$ of $v$ vectors, the whole of $d$-dimensional real space, is a simpler space for analysis.

## Expectation Relationships

The relationship between the dependence structure of a basis $x$ and its composition $y$ takes a simple form when we work in terms of the equivalent $v$-specification of the composition. In particular, if $u = \log x = \{\log x_1, \ldots, \log x_{d+1}\}$ then we can find very simple relationships between

$$\lambda = E(u), \qquad \Omega = V(u),$$

and

$$\mu = E(v), \qquad \Sigma = V(v)$$

Note that the dimensions of the vectors $\lambda$ and $\mu$ are $d + 1$ and $d$, and the dimensions of the matrices $\Omega$ and $\Sigma$ are $(d + 1) \times (d + 1)$ and $d \times d$. Let $A$ be the $d \times (d + 1)$ matrix $[I_d, -e_d]$ where $I_d$ and $e_d$ are the identity matrix and vector of units, each of dimension $d$. Then

$$\mu = A\lambda, \qquad \Sigma = A\Omega A^T \tag{3}$$

For a given $\mu$ and $\Sigma$ we cannot find unique $\lambda$ and $\Omega$ since, as we have seen earlier, there is an equivalence class of bases corresponding to a single composition. We can, however, easily identify the class of $\lambda$ and $\Omega$ corresponding to a given $\mu$ and $\Sigma$. For the vector $\{y_1/y_{d+1}, \ldots, y_d/y_{d+1}, 1\}$ forms a basis of $y$ and so the general form of bases with composition $y$ is $x = \{zy_1/y_{d+1}, \ldots, zy_d/y_{d+1}, z\}$, where $z$ is any positive random variable. Since $\lambda = E(\log x)$ and $\Omega = V(\log x)$ we can see that the degree of arbitrariness is 1 for $\lambda$, represented by the arbitrary mean $\alpha$ of $\log z$; and is $d + 1$ for $\Omega$, represented by the covariances $\beta_1, \ldots, \beta_d$ between $\log z$ and $\log (y_1/y_{d+1}), \ldots, \log (y_d/y_{d+1})$ and by the variance $\gamma$ of $\log z$. The appropriate expressions are then

$$\lambda = \begin{bmatrix} \mu + \alpha e_d \\ \alpha \end{bmatrix}, \qquad \Omega = \begin{bmatrix} \Sigma + e_d\beta^T + \beta e_d^T + \gamma U_d & \beta + \gamma e_d \\ \beta^T + \gamma e_d^T & \gamma \end{bmatrix}$$

where $U_d$ is the $d \times d$ matrix of units.

Particular interest has been shown in the past in two related questions

1. What is the extent of the correlation or covariance structure induced by the constraining process of forming a composition from a basis? To what extent are correlations observed in compositions real or just induced by the constraint?
2. How can we recognize from the covariance structure of a composition that it could have been produced from a basis of uncorrelated or independent components?

The strength of the present approach lies in the simplicity of the relationship for the circumstances described. For if a basis has independent components, then the logarithms of the components are also independent and so $\Omega = \text{diag}(\omega_1, \ldots, \omega_{d+1})$. Then the corresponding $\Sigma_0$ takes the form

$$\begin{aligned} \Sigma_0 &= \text{diag}(\omega_1, \ldots, \omega_d) + \omega_{d+1} U_d \\ &= \begin{bmatrix} \omega_1 + \omega_{d+1} & \omega_{d+1} & \cdots & \omega_{d+1} \\ \omega_{d+1} & \omega_2 + \omega_{d+1} & \cdots & \omega_{d+1} \\ \omega_{d+1} & \omega_{d+1} & \cdots & \omega_d + \omega_{d+1} \end{bmatrix} \end{aligned} \tag{4}$$

Since $\omega_1, \ldots, \omega_{d+1}$ are all positive we see that any composition corresponding to an independent basis must have equal positive covariances of the logratios $v$ and that this common covariance must be less than every variance of a logratio $v$. We shall see later how we can construct from standard statistical theory a reasonable test of whether compositional data conform to this structure, that is, whether the hypothesis $H_0$ of basis independence is tenable.

We reemphasize that, even if we know that a composition has this special covariance structure, all we can say is that the basis belongs to the equivalence

class that contains an independent basis. Possible bases then have a covariance matrix of the special form

$$\begin{bmatrix} \omega_1 + 2\beta_1 + \omega_{d+1} & \beta_1 + \beta_2 + \omega_{d+1} & \cdots \beta_1 + \beta_d + \omega_{d+1} & \beta_1 + \omega_{d+1} \\ \beta_1 + \beta_2 + \omega_{d+1} & \omega_2 + 2\beta_2 + \omega_{d+1} & \cdots \beta_2 + \beta_d + \omega_{d+1} & \beta_2 + \omega_{d+1} \\ \beta_1 + \beta_d + \omega_{d+1} & \beta_2 + \beta_d + \omega_{d+1} & \cdots \omega_d + 2\beta_d + \omega_{d+1} & \beta_d + \omega_{d+1} \\ \beta_1 + \omega_{d+1} & \beta_2 + \omega_{d+1} & \cdots \beta_d + \omega_{d+1} & \omega_{d+1} \end{bmatrix}$$

## Distributions for Compositions

For a full analysis of compositional data some parametric class of distributions to describe the pattern of variability would be a clear advantage. Since compositions are elements or vectors in the simplex $S^d$ the modeling problem is to find suitable distributions over this mathematically difficult space. The most, indeed the only, familiar class of distributions over $S^d$ is the Dirichlet class with probability density functions

$$\frac{\Gamma(\alpha_1 + \cdots + \alpha_{d+1})}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_{d+1})} \prod_{i=1}^{d+1} y_i^{\alpha_i - 1}$$

where $\alpha = (\alpha_1, \ldots, \alpha_{d+1})$ is the parameter. The main difficulty with this family is its known mathematical property that such a compositional distribution can *always* arise from an *independent* basis, whose components are independent and gamma-distributed with equal "scale" parameters.

Thus although the Dirichlet class can serve a useful purpose in providing a picture of the unavoidable and spurious correlations that arise simply out of the process of closure of an independent basis it cannot by its very nature be used to study proper covariance between components. For the description of real covariance a class of distributions over the simplex $S^d$ with a much richer covariance structure is required. It is only recently (Aitchison and Shen, 1980) that such a class has been fully identified. A clue as to how to devise such a class and why the class turns out to be tractable in statistical analysis has been touched upon earlier. Since the awkward space $S^d$ and the simple space $R^d$ are related to each other through the logistic and logratio transformations the following question immediately poses itself. If we start with a nice class of distributions in $R^d$, and what can be nicer than the multivariate normal class, and transform this to $S^d$ through the logistic transformation, do we obtain a useable, tractable class in $S^d$, rich enough to describe distributions of real compositional data? We claim that the answer is certainly yes, and have set out the main interesting properties and some simple applications to compositional data in a previous paper (Aitchison and Shen, 1980). Since considerations here are directed toward presenting a new perspective on the problem of correlations in compositional data, only the properties of this logistic-normal class of distributions which are of immediate interest for this particular problem are recorded.

If $v$ in $R^d$ is $N_d(\mu, \Sigma)$, that is $d$-dimensional multivariate normal with mean vector $\mu$ and covariance matrix $\Sigma$, then $y$ in $S^d$, related to $v$ through the logistic transformation (2) is said to follow a logistic-normal distribution, written $L_d(\mu, \Sigma)$ with parameters $\mu$ and $\Sigma$. Following a previous line of development we can show that the composition $y$ of a basis $x$ which is multivariate lognormally distributed, say $\Lambda_{d+1}(\lambda, \Omega)$ in the notation of Aitchison and Brown (1957), is $L_d(\mu, \Sigma)$ where $\mu$ and $\Sigma$ are related to $\lambda$ and $\Omega$ through (3). Indeed since multinormal, lognormal, and logistic-normal are all defined in terms of these first- and second-order moments (possibly of logarithms and logratios) all the comments on covariance structure of compositions and independence of bases carry through into this distributional form. In particular a logistic-normal composition $L_d(\mu, \Sigma)$ arising from a lognormal basis with independent components must have $\Sigma$ of the form (4). Thus we see that we have here the possibility of formulating a test of whether compositional data could be regarded as having arisen from independent bases. The question is translated into asking whether the covariance structure of the *logratios* of the proportions is consistent with, or contrary to, the special structure (4) for $\Sigma$.

These two classes of distributions over $S^d$, the Dirichlet and the logistic-normal, are not unrelated. The logistic-normal is by far the richer and provides a stronger tool of statistical analysis and yet at the same time can be used as a substitute for any Dirichlet distribution. For Aitchison and Shen (1980) show that any Dirichlet distribution $D_d(\alpha)$ can be closely approximated by a logistic-normal distribution $L_d(\mu, \Sigma)$ where $\mu_i = \delta(\alpha_i) - \delta(\alpha_{d+1})$, $\sigma_{ii} = \epsilon(\alpha_i) + \epsilon(\alpha_{d+1})$, $\sigma_{ij} = \epsilon(\alpha_{d+1})$ for $i \neq j$ where $\delta$ and $\epsilon$ are the digamma function $\Gamma'(\cdot)/\Gamma(\cdot)$ and the trigamma function $\delta'(\cdot)$, respectively. Closeness is here judged in terms of the Kullback and Leibler (1951) measure of directed divergence of one density function from another. Not surprisingly the $\Sigma$ for this closest logistic-normal distribution takes the basis-independence form (4). Thus we see that any statistical test of this particular covariance structure can be regarded as not only a test of the feasibility of an independent basis but also as a test of whether the Dirichlet, the archetypal independence distribution, or the more general form of logistic-normal distribution is required.

Mosimann (1975a,b) has pointed out a property of lognormal bases which may have prevented previous consideration of the logistic-normal as a serious alternative to Dirichlet distributions. He shows that any *lognormal basis* $x$ with a $\Lambda_{d+1}(\lambda, \Omega)$ distribution and with a composition having additive isometry (Mosimann, 1975b, p. 223) or equivalently proportional invariance (Darroch and James, 1974, p. 476), that is, with a composition independent of $\Sigma_{i=1}^{d+1} x_i$, must be degenerate. At first sight it therefore appears that logistic-normal distributions should be applied only to situations where we are assured that the compositions need not satisfy such additive isometry; and as Darroch and James point out, the use of compositional data often presupposes the satisfaction of proportional invariance. It is perfectly possible, however, to have a nondegener-

ate logistic-normal distribution for a composition with the proportional invariance property satisfied, provided we do not insist on the *basis itself being lognormal.* There seem to be no strong grounds for insisting on such lognormality in the components. For example, it is easy to devise a petrogenetic model in which $\Sigma x_i$ turns out to be lognormal and the associated compositional distribution logistic-normal with additive isometry. More specifically the distribution with density function

$$\frac{\Sigma x_i}{(2\pi)^{(1/2)d} x_1 \ldots x_{d+1} |\Sigma|^{1/2}} \exp\left\{-\tfrac{1}{2}\sum_{i,j}\sigma^{ij}\left(\log\frac{x_i}{x_{d+1}}-\mu_i\right)\left(\log\frac{x_i}{x_{d+1}}-\mu_j\right)-\tfrac{1}{2}(\log\Sigma x_i-\alpha)^2/\omega^2\right\}$$

where $\Sigma^{-1} = [\sigma^{ij}]$, satisfies these requirements. There is no need even to insist on $\Sigma x_i$ having a lognormal distribution. The way therefore seems open for an investigation of the use of logistic-normal distributions in the analysis of compositional data. Aitchison and Shen (1980) have shown an application to discriminant analysis. Here we concentrate on the provision of a test of basis independence.

## AN OVERALL TEST OF BASIS INDEPENDENCE

The analysis of the previous section has led us inevitably to the problem of testing a null hypothesis $H_0$ that the covariance structure of $\Sigma_0$ is $\mathrm{diag}(\omega_1, \ldots, \omega_d) + \omega_{d+1} U_d$, where $\omega_i \geqq 0$ $(i = 1, \ldots, d+1)$, the hypothesis of "basis independence," against the alternative hypothesis that $\Sigma$ takes a general positive definite form. The problem is similar to many considered in psychometric analysis, for example in Mukherjee (1970), except for the constraints on the positivity of the $\omega_i$. The method adopted is the generalized likelihood ratio test, whose computation requires a simple iterative procedure. The simplicity depends on the special structure of the covariance matrix $\Sigma_0$ under the null hypothesis $H_0$. It is easy to show that

$$\det \Sigma_0 = \omega_1 \ldots \omega_{d+1}\left(\frac{1}{\omega_1}+\frac{1}{\omega_2}+\cdots+\frac{1}{\omega_{d+1}}\right)$$

$$\Sigma_0^{-1} = \mathrm{diag}(\tau_1, \ldots, \tau_d) - \left(\sum_{i=1}^{d+1}\tau_i\right)^{-1}\tau\tau^T$$

where $\tau_i = \omega_i^{-1}$ $(i = 1, \ldots, d+1)$ and $\tau$ is the $d$-vector with components $\tau_1, \ldots, \tau_d$.

Suppose that the data $D$ consist of $n$ compositional vectors $y^{(1)}, \ldots, y^{(n)}$

and that the matrix of corrected cross-products of logratio vectors $v^{(1)}, \ldots, v^{(n)}$ is $V = [v_{ij}]$. The loglikelihood function, already maximized with respect to $\mu$, can be expressed in the following way

$$\log L = -\tfrac{1}{2} n \log |\Sigma| - \tfrac{1}{2} \operatorname{trace}(\Sigma^{-1} V)$$

Under the alternative hypothesis the maximizing $\Sigma$ is $\hat{\Sigma} = (1/n)\, V$. Under the null hypothesis $H_0$ the maximizing process can easily be studied through the use of the standard Newton-Raphson iterative methods, taking precautions to investigate the possibility that the maximizing $\omega$ may be on one of the boundaries. The computational details are set out in the Appendix.

Let $\Lambda_0(D)$ denote the generalized likelihood ratio test statistic so obtained. Although there is little possibility of determining the exact distribution, there is at least recourse to previous work on the asymptotic distribution theory of generalized likelihood ratio test statistics. Since the constraints imposed by $H_0$ involve inequalities, the standard results of Wald (1943) are not directly applicable, but require adjustment along the lines of Chernoff (1954) and Feder (1968). These adjustments are perhaps too complicated for general use and so we have chosen a simpler approach; this allows us to wedge our problem between two other problems for which the standard theory applies.

Consider the two hypotheses

$$H_1 : \Sigma_1 = \operatorname{diag}(\omega_1, \ldots, \omega_d) + \omega_{d+1} U_d$$

without the restriction $\omega_{d+1} \geqq 0$ on $\omega_{d+1}$; and

$$H_2 : \Sigma_2 = \operatorname{diag}(\omega_1, \ldots, \omega_d)$$

that is with $\omega_{d+1} = 0$. It is clear that $H_2$ implies $H_0$ and that $H_0$ implies $H_1$ so that the corresponding generalized likelihood ratio test statistics satisfy

$$\Lambda_1(D) \leqq \Lambda_0(D) \leqq \Lambda_2(D)$$

for all data sets $D$. The hypothesis $H_1$ is a special form of a covariance structure studied by Mukherjee (1970) within the framework of the Wald (1943) theory. In the form considered here it places $\frac{1}{2}d(d-1) - 1$ constraint equations on the elements of $\Sigma$. Similarly $H_2$ falls trivially within general theory placing $\frac{1}{2}d(d-1)$ constraint equations on the elements of $\Sigma$. As already mentioned the appropriate asymptotic theory of $\Lambda_0$ under the null hypothesis is difficult, following the line of argument of Gleser and Olkin (1973) and involving mixtures of $\chi^2\{\frac{1}{2}d(d-1) - 1\}$ and $\chi^2\{\frac{1}{2}d(d-1)\}$ distributions. In order to avoid such complications and obtain a readily applicable test we appeal to the argument of embedding $H_0$ between $H_1$ and $H_2$ and use a playsafe critical value, namely the $\chi^2$ value associated with $\frac{1}{2}d(d-1)$, the greater number of degrees of freedom.

Thus the asymptotic test we advocate computes $\Lambda_0(D)$ along the lines set out in the Appendix and then rejects the hypothesis $H_0$ of basis independence at

significance level at most $\alpha$ when

$$\Lambda_0(D) > \chi^2\{\tfrac{1}{2}d(d-1);\alpha\}$$

where $\chi^2(\nu;\alpha)$ is the upper $\alpha$ point of the $\chi^2(\nu)$ distribution.

A simpler alternative to the above will often be sufficient for the practical purpose of rejecting basis independence. Since $H_0 \subset H_1$, rejection of $H_1$ implies rejecting of $H_0$, and so we may content ourselves by attempting to use the data $D$ to reject the hypothesis $H_1$. The test statistic $\Lambda_1(D)$ is easily obtained by a modification of the $\Lambda_0(D)$ procedure in the Appendix, by removal of the insistence that $\omega_i \geqq 0$ $(i = 1, \ldots, d+1)$.

## SOME APPLICATIONS OF THE TEST

To illustrate the test of basis independence we use a number of applications which have appeared repeatedly in the literature to illustrate tests of null correlation. Because of the pairwise nature of the previous tests these have all tended to be low-dimensional; we therefore also undertake the analysis of some higher-dimensional data sets to emphasize the simplicity of the test procedure.

### Fossil Pollen Counts

Mosimann (1962) has analyzed data of Clisby and Sears (1955) giving 73 sets of the four proportions of fossil pollen grains of pine, fir, oak, and alder. He tests for null correlations under a compound multinomial hypothesis and suspects that some of the correlations are significant, but emphasizes the rather tentative nature of his test procedures. In so far as the relation of oak and pine are concerned these data are also analyzed through the concept of neutrality in the median by Darroch and Ratcliff (1978). One awkwardness of the logistic-normal analysis is that it cannot be applied to data with zero proportions. This difficulty has been circumvented by the admittedly ad hoc device of replacing zeros by 0.005, and then readjusting the proportions to sum to unity. Application of the overall test of basis independence to this adjusted set of data with $n = 73, d = 3$ leads to comparison of the test quantity $\Lambda_0(D) = 11.01$ against critical $\chi^2(3)$ values of 7.81 at 5% and of 11.34 at 1%. Thus there is sufficient evidence at the 5% significance level to reject the basis independence hypothesis associated with these compositional data.

If we concentrate on pine and oak, amalgamating fir and spruce, following the kind of neutrality investigations of Darroch and Ratcliff (1978), we find that the covariance matrix of log {pine/(fir + spruce)} and log {oak/(fir + spruce)} conforms to the basis independence pattern. Thus if interest is really in the composition of pine, oak, and (fir + spruce) we cannot argue against the feasibility of basis independence. This finding is not contrary to the finding of nonneutrality

in the median by Darroch and Ratcliff (1978). The counterpart of their ideas within the logistic-normal framework is in the idea of conditional subcompositions (Aitchison and Shen, 1980), and basis independence does not in general imply independence of subcompositions.

### Taupo Volcanic Rocks

Darroch and Ratcliff (1970, 1978) and Snow (1975) apply tests of neutrality in the mean and the median to two of the chemical components, $SiO_2$ and $Al_2O_3$, of 45 rock samples of the Taupo volcanic association reported by Steiner (1958). The nearest comparison with these analyses is an application of the test of basis independence to the 45 vectors of three proportions, $SiO_2$, $Al_2O_3$, remainder, so that $d = 2$. The test statistic $\Lambda_0(D) = 44.8$ is to be judged against $\chi^2(1)$ values. There is thus overwhelming evidence against the possibility of an underlying independent basis, a result in agreement with the earlier findings. But there is no need to confine ourselves to two of the components and the consequent lumping together of all the other components, an unnatural action if one raises the question of whether the lumping should be by weight or by volume. We have applied the test of $H_0$ to 44 of the vectors, omitting specimen number 10 because of its missing data, and lumping only the very minor oxides where none or only a trace was recorded. This procedure results in 13 components so that $d = 12$. Here $\Lambda_0(D) = 763$ to be tested against $\chi^2(66)$ values, so that again we have no hesitation in rejecting $H_0$.

### Chemical Variation in the Eocene Lavas of the Isle of Skye

Thompson, Esson, and Duncan (1972) present in their Table 2 chemical analyses, showing 10 components, of 32 basalts. Analysis of these for testing for the possibility of basis independence is readily carried out by the procedure of the Appendix and leads to a test quantity $\Lambda_0(D) = 329$, to be compared against upper $\chi^2(36)$ values. Again there is highly significant evidence against the basis independence hypothesis.

### Sediment Variability

McCammon (1975) sets a problem involving two data sets of (sand, silt, clay) compositions for (i) seven specimens of nearshore sediments, and (ii) ten specimens of offshore specimens. The first set yields a test quantity of $\Lambda_0(D) = 0$, seen to be absolutely reasonable when we note that the covariance matrix of the two logratios $v_1 = \log(x_1/x_3)$, $v_2 = \log(x_2/x_3)$ is

$$\begin{bmatrix} 0.731 & 0.216 \\ 0.216 & 0.494 \end{bmatrix}$$

readily conforming to the pattern

$$\begin{bmatrix} \omega_1 + \omega_3 & \omega_3 \\ \omega_3 & \omega_2 + \omega_3 \end{bmatrix}$$

with $\omega_1, \omega_2, \omega_3$ all positive.

For the offshore specimens the covariance matrix

$$\begin{bmatrix} 2.978 & 0.754 \\ 0.754 & 0.453 \end{bmatrix}$$

does not conform to the above pattern, but still the test quantity $\Lambda_0(D) = 1.10$ is not sufficiently large to allow rejection of $H_0$. Thus for both nearshore and offshore data we cannot refute the hypothesis that the composition has its origin in a basis with independent components.

## DISCUSSION

The test of basis independence developed and applied in this paper has several advantages over previous attempts at analysis. It is based on a simple and natural way of linking the dependence structures of open and closed models, it provides an overall test of the complete structure as opposed to separate pairwise tests, and its critical value is relatively well based in asymptotic test theory compared with the tentative nature of many previous tests. More important, however, is that the development does not stop at the test of basis independence but allows further investigation in the event of rejection of the hypothesis. There are many possibilities of further investigation through logistic-normal distributions. There may be other patterned covariance structures, depicting some special forms of dependence, which could be next investigated, along lines similar to those investigated by psychologists; see, for example, Mukherjee (1970) and Gleser and Olkin (1973). Interest may be directed towards subcompositions of the whole vector in the sense of Aitchison and Shen (1980), and the investigation of whether the subcompositions could have independent bases. Alternatively we may be more concerned with some form of statistical principal component analysis and investigation of the effective dimensionality of the pattern of variability. All of this is in striking contrast to other formulations which fail to provide any statistical framework for the quantitative investigation of truly correlated proportions.

We have concentrated on one aspect of the value of the new approach, a clearer understanding of the nature of null correlations and a practical tool for their analysis. There are many other clarifications and methodologies which emerge immediately. For example, the relation of basis independence to other forms of non-association throws some interesting light on modeling. We can actually devise a simple test for proportional invariance, size homogeneity, or addi-

tive isometry (Aitchison, 1981b). Finally since the logistic-normal has a limit law similar to the central limit theorem for normal distributions, interesting questions can be raised about the possibility of providing a genetic explanation, along the lines of the lognormal genesis by breakage for particle size distributions, of the occurrence of logistic-normal compositional patterns.

## APPENDIX: THE COMPUTATION OF THE TEST STATISTIC

Let $D$ denote the data set consisting of $n$ compositional vectors $y^{(1)}, \ldots, y^{(n)}$ in $S^d$, with corresponding logratio vectors $v^{(1)}, \ldots, v^{(n)}$. Let $\boldsymbol{v}$ denote the mean vector and $V = \sum_{i=1}^{n} (v^{(i)} - \boldsymbol{v})(v^{(i)} - \boldsymbol{v})^T$ the matrix of cross products. As pointed out in the main text, the likelihood under the logistic-normal model is maximized at $\hat{\mu} = \boldsymbol{v}$, $\hat{\Sigma} = V/n$. The only awkward problem is the maximization under the basis independence hypothesis that the covariance structure is of the form

$$\sigma_{ii} = \omega_i + \omega_{d+1} \qquad (i = 1, \ldots, d)$$

$$\sigma_{ij} = \omega_{d+1} \qquad (i \neq j)$$

with $\omega_i \geqq 0$ $(i = 1, \ldots, d+1)$.

As far as investigating the region $\omega_i \geqq 0$ $(i = 1, \ldots, d+1)$ is concerned we can show, after some tedious algebra, that the $(d+1)$ vector $\pounds'(\omega)$ of derivatives of the loglikelihood $\pounds$ has the $i$th component

$$\pounds_i'(\omega) = \tfrac{1}{2}n(\omega_i - \tau) - \tfrac{1}{2}v_{ii} + \tau \sum_{j=1}^{d+1} v_{ij}/\omega_j - \tfrac{1}{2}\tau^2 \sum_{i=1}^{d+1}\sum_{j=1}^{d+1} v_{ij}/(\omega_i\omega_j)$$

where $\tau^{-1} = \omega_1^{-1} + \ldots + \omega_{d+1}^{-1}$ and $v_{i,d+1} = 0$ $(i = 1, \ldots, d+1)$, and that the $(d+1) \times (d+1)$ information matrix is

$$B(\omega) = \tfrac{1}{2}n\{\operatorname{diag}(\omega_1^2 - 2\tau\omega_1, \ldots, \omega_{d+1}^2 - 2\tau\omega_{d+1}) + \tau^2 U_{d+1}\}$$

The usual iterative procedure leading from the $r$th iterate $\omega^{(r)}$ $(r = 0, 1, \ldots,)$ to the $(r+1)$th iterate $\omega^{(r+1)}$ is

$$\omega^{(r+1)} = \omega^{(r)} + [B(\omega^{(r)})]^{-1}\pounds'(\omega^{(r)})$$

The recommended initial values are

$$\omega_{d+1}^{(0)} = \sum_{i=1}^{d}\sum_{j=i+1}^{d} v_{ij}/\{\tfrac{1}{2}nd(d-1)\},$$

$$\omega_i^{(0)} = v_{ii}/n - \omega_{d+1}^{(0)} \qquad (i = 1, \ldots, d) \tag{5}$$

if these are all nonnegative. Otherwise, if the above $\omega_{d+1}^{(0)} < 0$ set

$$\omega_{d+1}^{(0)} = 0.001, \qquad \omega_i^{(0)} = v_{ii}/n \qquad (i = 1, \ldots, d)$$

and if $\omega_{d+1}^{(0)}$ in (5) is nonnegative and $\omega_j^{(0)}$ is the minimum negative value of $\omega_i^{(0)}$ $(i = 1, \ldots, d)$, set

$$\omega_{d+1}^{(0)} = v_{jj}/n, \qquad \omega_j^{(0)} = 0.001,$$
$$\omega_i^{(0)} = (v_{ii} - 2v_{ij} + v_{jj})/n \qquad (i \neq j, d+1)$$

As each iterative stage is completed it is easy to check whether all $\omega_i^{(r+1)}$ are positive. If not, set any which are negative to 0.001 before the next iterative cycle. If the maximization is on the boundary, that is with some $\omega_i$ $(i = 1, \ldots, d+1)$ zero then the above procedure picks up this fact by the corresponding iterate becoming smaller and smaller. As a check on this, any case where some $\omega_i$ is zero is very simply solved, with for $\omega_{d+1} = 0$

$$\hat{\omega}_i = v_{ii}/n$$

and for $\omega_i$ $(i \neq d+1) = 0$

$$\hat{\omega}_{d+1} = v_{ii}/n, \qquad \omega_j = (v_{ii} - 2v_{ij} + v_{jj})/n$$

A simple program for the test procedure has been written in BASIC and implemented on the Wang 2200S minicomputer system. A listing is available from the author on request.

## REFERENCES

Aitchison, J., 1981a, Distributions on the simplex for the analysis of neutrality, *in* Statistical distributions in scientific work, Taillie, C., Patil, G. P., and Baldessari, B. (eds.): D. Reidel Publishing Company, Dordrecht, Holland, to appear.

Aitchison, J., 1981b, Testing for additive isometry and proportional invariance: Biometrics, v. 37, to appear.

Aitchison, J. and Brown, J. A. C., 1957, The lognormal distribution: Cambridge University Press, New York.

Aitchison, J. and Shen, S. M., 1980, Logistic-normal distributions: some properties and uses: Biometrika, v. 67, p. 261–272.

Bartlett, N. R. and Darroch, J. N., 1978, Regression and correlation of bounded-sum variables: Vistelius Commemoration Volume.

Butler, J. C., 1979, Trends in ternary petrologic variation diagrams–fact or fantasy?: Amer. Mineralogist, v. 64, p. 1115–1121.

Chayes, F., 1960, On correlation between variables of constant sum: Jour. Geophys. Res., v. 65, p. 4185–4193.

Chayes, F., 1962, Numerical correlation and petrographic variation: Jour. Geol., v. 70, p. 440–452.

Chayes, F. and Kruskal, W., 1966, An approximate statistical test for correlations between proportions: Jour. Geol., v. 74, p. 692–702.

Chernoff, H., 1954, On the distribution of the likelihood ratio: Ann. Math. Stat., v. 25, p. 573–578.

Clisby, K. H. and Sears, P. B., 1955, Palynology in southern North America. Part III. Microfossil profiles under Mexico City correlated with sedimentary profiles: Geol. Soc. America Bull., v. 66, no. 5, p. 511–520.

Connor, J. R. and Mosimann, J. E., 1969, Concepts of independence for proportions with a generalization of the Dirichlet distribution: Jour. Amer. Statist. Assoc., v. 64, p. 194–206.

Darroch, J. N., 1969, Null correlations for proportions: Jour. Internat. Assoc. Math. Geol., v. 1, no. 2, p. 221–227.

Darroch, J. N. and James, I. R., 1974, F-independence and null correlations of continuous, bounded-sum, positive variables: Jour. Roy. Stat. Soc., Ser. B, v. 36, no. 3, p. 467–483.

Darroch, J. N. and Ratcliff, D., 1970, Null correlations for proportions. II.: Jour. Internat. Assoc. Math. Geol., v. 2, p. 307–312.

Darroch, J. N. and Ratcliff, D., 1971, A characterization of the Dirichlet distribution: Jour. Amer. Stat. Assoc., v. 66, p. 641–643.

Darroch, J. N. and Ratcliff, D., 1978, No-association of proportions: Jour. Internat. Assoc. Math. Geol., v. 10, p. 361–368.

Feder, P. I., 1968, On the distribution of the log likelihood ratio test statistic when the true parameter is 'near' the boundaries of the hypothesis regions: Ann. Math. Stat., v. 39, p. 2044–2055.

Gleser, L. J. and Olkin, I., 1973, Multivariate statistical inference under marginal structure: Brit. Jour. Math. Stat. Psychol., v. 26, p. 98–123.

Krumbein, W. C., 1962, Open and closed number systems stratigraphic mapping: Bull. Amer. Assoc. Petrol. Geologists, v. 46, p. 2229–2245.

Kullback, S. and Leibler, R. A., 1951, On information and sufficiency: Ann. Math. Stat., v. 22, p. 525–540.

McCammon, R. B., 1975, Concepts in geostatistics; John Wiley & Sons, New York.

Miesch, A. T., 1969, The constant sum problem in geochemistry, *in* Computer applications in the earth sciences, Merriam, D. F. (ed.): Plenum Press, New York, p. 161–177.

Mosimann, J. E., 1962, On the compound multinomial distribution, the multivariate $\beta$-distribution and correlations among proportions: Biometrika, v. 49, p. 65–82.

Mosimann, J. E., 1963, On the compound negative multinomial distribution and correlations among inversely sampled pollen counts: Biometrika, v. 50, p. 47–54.

Mosimann, J. E., 1975a, Statistical problems of size and shape. I. Biological applications and basic theorems, *in* Statistical distributions in scientific work, v. 2, Patil, G. P., Kotz, S., and Ord, J. K. (eds.): D. Reidel Publishing Company, Dordrecht, Holland, p. 187–217.

Mosimann, J. E., 1975b, Statistical problems of size and shape. II. Characterizations of the lognormal, gamma, and Dirichlet distributions, *in* Statistical distributions in scientific work, v. 2, Patil, G. P., Kotz, S., and Ord, J. K. (eds.): D. Reidel Publishing Company, Dordrecht, Holland, p. 219–239.

Mukherjee, B. N., 1970, Likelihood ratio tests on statistical hypotheses associated with patterned covariance matrices in psychology: Brit. Jour. Math. Stat. Psychol., v. 23, p. 89–120.

Snow, J. W., 1975, Association of proportions: Jour. Internat. Assoc. Math. Geol., v. 7, no. 1, p. 63–73.

Steiner, A., 1958, Petrographic implications of the 1954 Ngauruhoe lava and its xenoliths: New Zealand Jour. Geol. Geophys., v. 1, p. 325–363.

Thompson, R. N., Esson, J., and Duncan, A. C., 1972, Major element chemical variation in the Eocene lavas of the Isle of Skye, Scotland: Jour. Petrology, v. 13, p. 219–253.

Wald, A., 1943, Tests of statistical hypotheses concerning several parameters when the number of observations is large: Trans. Amer. Math. Soc., v. 54, p. 426–482.