ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS

HAROLD HOTELLING

Columbia University

(*Continued from September issue.*)

8. DETERMINATION OF PRINCIPAL COMPONENTS FOR INDIVIDUALS

To determine from his test scores the value of a principal component γ for an individual, the formula

$$\gamma = \frac{a_j z_j}{k} \tag{33}$$

(summation over all the n tests indicated by the repetition of j of the last section) would be appropriate if there were no error of measurement, so that the test scores could be taken as identical with the true scores z_j. It is however evident that if the reliabilities of the tests vary widely, the weights of the more reliable tests should be increased relatively to the others. The ordinary rule for combining independent observations is that the weights should be inversely proportional to the variances of the chance errors;[1] this rule must however be modified in the present case, since the tests which constitute the observations are not independent; and indeed, such a weighted mean would be the same for all the γ's, and would have nothing to do with the a's.

We shall estimate the value of γ for an individual as the linear function γ' of his test scores such that the mean value in the population of

$$(\gamma' - \gamma)^2$$

shall be a minimum. This criterion gives the same results as to require the correlation of γ' and γ to be a maximum, except that the coefficients in the linear function γ' may be multiplied by a constant which is determined by the former but not by the latter condition.

If the analysis has been performed upon a matrix of raw correlations, with reliability coefficients in the principal diagonal, the

[1] This is the system of weights which makes the variance of a weighted mean a minimum. It may also be deduced from the work by Truman L. Kelley on pp. 212–213 of *Interpretation of Educational Measurements*, World Book Company, 1927.

simple formula (33) provides immediately the solution of the problem. In what follows, we suppose that the analysis has been performed upon the matrix of correlations corrected for attenuation, as in the example of Section 5.

The variances of the true scores z_i have been taken as unity and those of their chance errors ϵ_i as σ_i^2; consequently the variance of the observed score $x_i = z_i + \epsilon_i$ is $1 + \sigma_i^2$ in these units. But now let us denote by y_i the observed score expressed in standard measure, *i.e.* put

$$y_i = \frac{z_i + \epsilon_i}{\sqrt{1 + \sigma_i^2}} = (z_i + \epsilon_i)\sqrt{r_I} \tag{34}$$

The difference between the sample mean and that of the population we ignore, as being a small quantity of higher order than we are considering, and take both these means to be zero. For the products with which we shall deal we have the following population means, or expectations:

$$Ey_iy_j = r'_{ij}, \tag{35}$$

the observed correlation between the scores, equal to unity if $i = j$, and otherwise given by (32). Also, from (34), since $Ez_iz_j = r_{ij}$, we have

$$Ey_iz_j = \sqrt{r_I}r_{ij}, \tag{36}$$

where the summation convention does not apply, though $I = i$, the use of capital letters as subscripts serving to distinguish such cases; but in formulae such as those below, summation with respect to j is to be understood, as this lower-case letter occurs twice; in each term of the sum, J is to have the same value as j.

From (33) and (36),

$$Ey_i\gamma = \frac{\sqrt{r_I}r_{ij}a_j}{k}.$$

Now by (16),

$$r_{ij}a_j = ka_i;$$

and putting

$$b_i = \sqrt{r_I}\,a_i,$$

we obtain

$$Ey_i\gamma = b_i. \tag{37}$$

Now putting

$$\gamma' = c_iy_i, \tag{38}$$

the quantity to be made a minimum by a proper choice of the c_i is

$$\begin{aligned} T &= E(\gamma' - \gamma)^2 \\ &= E(c_i y_i - \gamma)(c_j y_j - \gamma) \\ &= r'_{ij}c_i c_j - 2b_i c_i + 1, \end{aligned} \tag{39}$$

the last expression following from (35), (37), and the fact that γ is expressed in standard measure, so that $E\gamma^2 = 1$.

Differentiating (39) with respect to b_i $(i = 1, \cdots, n)$, we obtain the n equations

$$r'_{ij}c_j = b_i. \tag{40}$$

By any of the methods used for solving normal equations, (40) may be solved for the c_j.

Let R'_{ih} be the cofactor of r'_{ih} in the determinant of these correlations, divided by this determinant. Then, just as in (4),

$$R'_{ih}r'_{ij} = \delta_{hj}. \tag{41}$$

Multiplying both sides of (40) by R'_{ih}, summing with respect to i, and using (41), we find that the solution of (40) may be written in the form

$$c_h = R'_{ih}b_i. \tag{42}$$

If all or nearly all the principal components are to be expressed numerically in terms of the test scores, it is best to compute the n^2 quantities R'_{ih}, and then for each component to use (42). The R'_{ih} may be found most readily by applying to (41) a method such as that of Doolittle for solving normal equations. The coefficients in these equations are the uncorrected correlations, as in (40); but since the right members are replaced in turn by the columns

$$\begin{array}{lll} 1 & 0 & 0 \ldots 0 \\ 0 & 1 & 0 \ldots 0 \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots\ldots} \\ 0 & 0 & 0 \ldots 1, \end{array}$$

and since in the solution all these columns can be carried along simultaneously, the work is much reduced below that required for a direct solution of (40). This procedure is readily checked throughout by carrying along an additional column, the entry in each row being the sum of all the preceding entries in that row.

For the four tests we have been using as an illustration, the raw correlations are:

1.	.633	.241	.059
	1.	−.055	.065
		1.	.425
			1.

The procedure described above gives for the inverse matrix of the R_{ik}':

1.959	−1.292	−.646	.244
	1.865	.529	−.271
		1.441	−.609
			1.261

These matrices are of course to be thought of as square and symmetrical; a column is read by moving down to the last of the entries in that column as written above, and then to the right. With this understanding, the columns are to be multiplied by those of the matrix of b's below; this is obtained from that at the end of Section 5 by multiplying each row by the square root of the reliability coefficient (given in Section 7) for the corresponding test.

Values of $b_{i\gamma} = a_{i\gamma}\sqrt{r_i}$

	Component γ			
Test z_i	1	2	3	4
1	779	− 425	−.279	−.232
2	.652	−.589	.271	.217
3	.582	.637	−.360	.179
4	.436	.491	346	−.104

Multiplying the two matrices, we have finally for the c's:

Values of $c_{i\gamma}$

Coefficients of Test Scores in Estimates of Principal Components

	Component γ			
Test z_i	1	2	3	4
1	.414	− 363	−.580	−.876
2	.399	− 345	.582	.827
3	.415	582	−.406	.586
4	.209	.287	.514	−.356

For comparison with these coefficients (which for any particular γ are to be read down a column), we give below the coefficients $A_i = a_{i\gamma}/k$, obtained simply by dividing the columns of the table at the end of Section 5 by their characteristic roots. These are the coefficients which would be used with the true scores, if we knew them.

$$Values\ of\ A_i = \frac{a_{i\gamma}}{k}$$

COEFFICIENTS OF TRUE SCORES IN EQUATIONS FOR PRINCIPAL COMPONENTS

	Component γ			
Test z_i	1	2	3	4
1	.440	− 303	−.557	−1.467
2	.373	− 424	.550	1.394
3	.331	456	−.724	1 139
4	.315	.447	.883	− .824

9. ITERATIVE SOLUTION OF NORMAL EQUATIONS; CONVERGENCE

If the tests are numerous and only a few of the principal components are to be found, it may be better to use an iterative method of solving the normal equations (40). However the advantage is not as overwhelming as that of the iterative method of Section 4 for the determination of the $a_{i\gamma}$. The speediest iterative process now available seems to be that of Kelley and Salisbury.[1] It is a modification, which accelerates convergence, of the following method, which is similar to those of Gauss, Jacobi, and Seidel,[2] differing from that of Seidel only in that the normal equations and the unknowns may here be regarded as transformed so as to make the coefficients correlations. In demonstrating that this method converges, it will follow a fortiori that that of Kelley and Salisbury converges, since in using it, T is diminished even more rapidly than in the following method.

The last member of (39) may be written:

$$T = (r'_{1j}c_j - b_1)^2 + T_1,$$

where T_1 does not involve c_1. Hence if we start with any assumed values whatever for $c_1, c_2, \ldots, c_n$, and then change c_1 so as to

[1] *Journal of the American Statistical Association*, Vol. XXI, 1926, pp. 282–292.

[2] Whittaker and Robinson: "The Calculus of Observations," Sec. 130.

make the expression in parentheses vanish; that is, if we replace c_1 by

$$-r'_{12}c_2 - r'_{13}c_3 - \cdots - r'_{1n}c_n + b_1,$$

we shall thereby *diminish* the value of T, which we are trying to make a minimum; except that if c_1 already had this value, T is unchanged. Then, using the revised values, we replace c_2 by

$$-r'_{21}c_1 - r'_{23}c_3 - \cdots - r'_{2n}c_n + b_2.$$

If this means any change in c_2, T diminishes still further, and this modified value of c_2 makes the squared term vanish. Continuing in this way, each variable in turn is modified so as to diminish T, or to leave it stationary. The condition that T remain stationary through an entire cycle of changes in all the variables is that each of them already has the indicated value. But in this case (40) holds; the solution is already attained. Hence, if the assumed values differ from the true ones, T will actually diminish in a cycle of substitutions, and not remain stationary.

Let us regard the trial values as coordinates of a point in a space of n dimensions. The first modification amounts to moving all such points, representing possible sets of trial values, parallel to the axis of c_1 onto the hyperplane

$$r_{1j}c_j - b_1 = 0. \tag{43}$$

The next modification is equivalent to a projection parallel to the axis of c_2, from this onto a new hyperplane, and so on. All these hyperplanes pass through a point whose coordinates are the desired solution. A complete cycle of the modifications carries the points back to the hyperplane (43), which is thus transformed into itself. The equation

$$T_1 = \alpha, \tag{44}$$

where T_1 is the function of $c_2, \ldots, c_n$ defined above and α is any constant, represents an ellipsoid in (43). This, in the tranformation of the hyperplane into itself, is carried into a new ellipsoid, whose equation may be written

$$T_1' = \alpha, \tag{45}$$

where T_1' is a new quadratic form in $c_2, \ldots, c_n$.

Since each value of T, and therefore of T_1, is diminished in a cycle of the substitutions, each point of (44) is carried into a point *within* (44). Hence the ellipsoid (45) lies entirely within (44), not

even touching it. This geometrical fact supplies us with information regarding the roots of $F(\lambda) = 0$, where $F(\lambda)$ is the discriminant of the second-degree terms of the quadratic form

$$T_1' - \lambda T_1.$$

These roots are invariant under all linear transformations to which both T_1 and T_1' may be subjected. Let us then stretch the space so that $T_1 = \alpha$ becomes the unit sphere. The roots of $F(\lambda) = 0$ are then the squares of the reciprocals of the semiaxes of the transformed ellipsoid $T_1' = \alpha$. Since this will lie entirely within the unit sphere, all these roots must be greater than unity. Let λ' be the least root. Then $\lambda' > 1$. Each value α of T_1 will, in the course of a cycle of substitutions, diminish to a value not exceeding α/λ'. Hence, in m cycles, the reduced value of T_1 will come not to exceed α/λ'^m, and will therefore approach zero. In this way we have a definite proof of the convergence of the Seidel process, which may not have been previously demonstrated.

While the value of T_1 is divided by λ' *at least* in each cycle, it will ordinarily be divided by a greater number, since the trial point will seldom lie on the longest axis of the stretched ellipsoid.

10. TESTS AS SAMPLES OF A LARGER AGGREGATE OF TESTS*

Instead of regarding the analysis of a particular set of tests as our ultimate goal, we may look upon these merely as a sample of a hypothetical larger aggregate of possible tests. Our aim then is to learn something of the situation portrayed by the large aggregate. We are thus brought to a type of sampling theory quite distinct from that which we have heretofore considered. Instead of dealing with the degree of instability of functions of the correlations of the observed tests arising from the smallness of the number of persons tested, regarded as a sample of a larger population of persons, we are now concerned with the degree of instability resulting from the limited number of tests whose correlations enter into our analysis. The theory of this subject is far from complete, but some relevant results will be set forth. These results are based on a concept of the tests as a random sample from a hypothetical infinite population of possible tests. As in other uses of sampling theory, the reservation must be made that if the sampling is not random the results cannot be applied with accuracy. If for example the tests are deliberately chosen so as

*The central importance of this concept came to light in discussions with Professor Clark V. Hull.

to be highly correlated with a particular one of them, or so that a group of them will have low correlations with each other, as might be done in developing a battery of tests designed to estimate some criterion variable with maximum multiple correlation, they cannot be called a random sample.

Our n variables might also be treated as a sample from a larger *finite* population of tests. We might then regard the population of tests as resolved into its principal components $\Gamma_1, \Gamma_2, \ldots$, and ask such questions as to the extent to which γ_1 is likely to be correlated with Γ_1, γ_2 with $\Gamma_2, \ldots$, and γ_n with Γ_n. We should certainly wish to know how well the fractions of total variance of the observed tests contributed by the principal components represent the corresponding fractions of the total variance of the population of tests. In this as in other connections, the concept of sampling from a finite population helps to fix the ideas and make them more tangible, but involves mathematics which is more difficult, and I think less generally applicable, than arises when we go at once to the limit in increasing the population. Using the "infinite population" concept amounts merely to treating our probabilities as independent, and does not involve going into the mathematical theory of aggregates or sets, which have somewhat the same relation to the hypothetical infinite populations of statistics that the pure spaces of analysis situs bear to the metrical spaces of ordinary geometry.

Let the fractions of the total variance contributed by the successive principal components in a sample of n tests be

$$f_1 = \frac{k_1}{n}, \qquad f_2 = \frac{k_2}{n}, \cdots, \qquad f_n = \frac{k_n}{n}.$$

These are the roots of the equation obtained by putting $k = nf$ in the characteristic equation (29), p. 429:

$$f^n - f^{n-1} + \left(\frac{S_2}{n^2}\right) f^{n-2} - \left(\frac{S_3}{n^3}\right) f^{n-3} + \cdots = 0. \tag{46}$$

Here

$$S_2 = \sum_{i>j} D_{ij}, \qquad S_3 = \sum_{i>j>k} D_{ijk}, \cdots$$

where

$$D_{ij} = \begin{vmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{vmatrix}, \qquad D_{ijk} = \begin{vmatrix} 1 & r_{ij} & r_{ik} \\ r_{ji} & 1 & r_{jk} \\ r_{ki} & r_{kj} & 1 \end{vmatrix}, \cdots.$$

It is desirable to show that, by taking a sufficiently large number, n, of tests at random, we can make the probability arbitrarily small that any of a finite set of the roots will differ by more than a fixed amount from definite expected values. This property, if substantiated, makes the roots $f_1, f_2, \ldots$ *consistent statistics*, and justifies the use of the method of principal components in spite of the arbitrariness of metric mentioned in Section 1, for it implies that sets of tests independently devised to study a complex situation may, under the random conditions assumed, be relied upon to give similar results if only they are sufficiently numerous. While this fundamental theorem has not been established with full rigor, its high degree of plausibility follows from the following considerations.

Since each of the correlations r_{ij} has its entire distribution confined between the limits ± 1, all its moments are finite. The same must necessarily be true of every polynomial in the r_{ij}'s. In particular, the determinants D_{ij}, D_{ijk}, . . . , and also the powers and products of these determinants, being polynomials in the r_{ij}'s, must all have definite finite expectations. Now S_k is the sum of

$$C_k^n = \frac{n(n-1)(n-2)\cdots(n-k+1)}{1\cdot 2\cdot 3\cdots k!}$$

of these determinants. If we denote the elementary symmetric functions of the roots by b_1', b_2', . . . , so that b_k' is the sum of the products of the f's taken k at a time, then by (46),

$$b_k' = \frac{S_k}{n^k}.$$

Hence b_k' has a definite expectation, Eb_k', depending on n, but approaching a finite limit β_k as n increases, since C_k^n/n^k approaches a finite limit.

We next show that a value of n may be chosen large enough so that b_k' and Eb_k' will differ arbitrarily little in an arbitrarily great proportion of samples of n. This follows from the Tchebycheff inequality used in establishing various Laws of Great Numbers, as soon as we show that the variance of b_k' approaches zero as n increases.

The variance of b_k' is that of S_k, divided by n^{2k}. The variance of S_k is the sum of the variances of the determinants

$$D_{i_1 i_2 \ldots i_k},$$

and of double the covariances of pairs of these determinants. The number of pairs of such determinants with h common subscript is

$$\frac{1}{2}C_h{}^nC_{k-h}{}^{n-h}C_{k-h}{}^{n-k} = \frac{1}{2}\frac{n!}{h![(k-h)!]^2(n-2k+h)!},$$

a polynomial in n of degree $2k - h$. For $h = 0$, all the covariances vanish, since those determinants which have no common subscripts are entirely independent. The most numerous group of non-vanishing covariances is for $h = 1$, corresponding to determinants with a single common subscript. The number of these terms is given by a polynomial in n of degree $2k - 1$. Since the number of terms in each of the other groups ($h = 2, 3, \cdots, n$) is given by a polynomial of lower degree, and since each term has a fixed value depending only on the population, the total variance of S_k is of degree $2k - 1$ in n. Thus the variance of b_k' vanishes as n^{-1}. Its standard deviation, and therefore, almost always, its deviation from Eb_k', will be of order $n^{-\frac{1}{2}}$ for large values of n.

Since the elementary symmetric functions b_k' have definite expectations about which they cluster with standard deviations of order $n^{-\frac{1}{2}}$, the same is true for the power sums

$$e_k' = \sum_{i=1}^{n} f_i{}^k, \qquad (k = 1, 2, \cdots, n)$$

for these are polynomials in the b_k'. Let

$$\epsilon_k = \lim_{n\to\infty} Ee_k'.$$

If we put

$$\epsilon_k = \sum_{i=1}^{\infty} \phi_i{}^k \qquad (k = 1, 2, \cdots, \infty),$$

then it appears highly plausible that these equations have solutions

$$\phi_1, \phi_2, \cdots,$$

each between 0 and 1, which we take in descending order of magnitude, and which are the population values about which the sample f's cluster. We may expect the ϕ's to satisfy an equation analogous to (46), obtained by dividing by the term of highest degree, replacing each coefficient by its expectation, and letting n increase:

$$1 - \frac{1}{\phi} + \frac{\beta_2}{\phi^2} - \frac{\beta_3}{\phi^3} + \cdots = 0$$

this being now an infinite series. The ϕ's may be described as the fractions of the total variance of the population accounted for by its principal components. The more rapidly the series

$$\phi_1 + \phi_2 + \phi_1 + \cdots,$$

converges to unity, the more definitely do the various abilities tested depend upon a small number of underlying characters. The speed of convergence may be measured for example by

$$\epsilon_2 = \phi_1^2 + \phi_2^2 + \cdots.$$

Each of the ϵ's will be less than that of next lower order, as is the case with the e''s. The e''s and ϵ's are in fact the moments of a frequency distribution in sample and population, respectively, and the question of the extent to which they, or any finite number of them, determine the ϕ's is very similar to the classical moment problems on which so much has been done. This transition from the moments of the distribution of ϕ's to the distribution itself is all that is needed to establish the fundamental theorem of consistency of the f's.

Without going into the interesting mathematical questions raised in this way, we shall be on firm ground in dealing with the ϵ's and their estimation by means of a sample. We shall close this section by proving that e_k', as an estimate of ϵ_k, has a bias of order n^{-1}, and showing how to correct for this bias.

From the relations between the power-sums and the elementary symmetric functions of the roots of (46) we have:

$$e_2' = \sum f^2 = 1 - \frac{2S_2}{n^2},$$

$$e_3' = \sum f^3 = 1 - \frac{3S_2}{n^2} + \frac{3S_3}{n_3},$$

. . . . . .

From the first of these,

$$Ee_2' = 1 - \frac{2ES_2}{n^2} = 1 - \frac{n-1}{n} ED_{ij}.$$

Then

$$\epsilon_2 = \lim Ee_2' = 1 - ED_{ij}.$$

Eliminating ED_{ij},

$$\epsilon_2 = \frac{nEe_2' - 1}{n-1}.$$

Hence if we put

$$e_2 = \frac{ne_2' - 1}{n - 1},$$

e_2 will be an unbiassed estimate of ϵ_2, since $Ee_2 = \epsilon_2$. This bias in e_2' means that the fraction f_1 of the variance contributed by the leading principal component of n tests is an exaggeration, of order n^{-1}, of the corresponding quantity ϕ_1 in the population. For small values of n this correction is very important. In the example of Section 5, in which we took only four tests, e_2' is about .38, so that e_2 is only about .17. If the value of ϵ_2 is .17, ϕ_1 cannot possibly be so great as the value .465 found for f_1 in the example, since the square of this quantity is by itself greater than .17.

The unbiassed estimate of ϵ_3 is found similarly, but more laboriously, to be

$$e_3 = \frac{n^2e_3' - 3ne_2' + 2}{(n - 1)(n - 2)}.$$

For the higher degrees it is simpler to work with the elementary symmetric functions than with the power-sums. The unbiassed estimate of β_k is merely

$$b_k = \frac{n^{k-1}}{(n - 1)(n - 2) \cdots (n - k + 1)} b_k'.$$

11. PRINCIPAL COMPONENTS WITH PERFECT WEIGHTING

Consider a finite or infinite population of variates x_ι, which we shall call *tests*, related to a finite or infinite sequence of variates Γ_i which in an infinite population of *persons* are independently distributed with unit variance. Let the relations be

$$x_i = \alpha_{\iota 1}\Gamma_1 + \alpha_{\iota 2}\Gamma_2 + \cdots, \tag{47}$$

where the coefficients $\alpha_{\iota j}$ do not vary in the population of persons, but vary independently in the population of tests. Let us assume that, for each second subscript j, the $\alpha_{\iota j}$'s have the mean value zero and variance ϕ_j. Denoting the mean value of a quantity in the population of tests by a prefixed E, this means that

$$E\alpha_{\iota j} = 0, \qquad E\alpha^2_{\iota j} = \phi_j, \tag{48}$$

and

$$E\alpha_{ij}\alpha_{kl} = 0 \qquad \text{unless } i = k, \text{ and } j = l. \tag{49}$$

The covariance of x_ι and x_k is

$$p_{\iota k} = \alpha_{\iota 1}\alpha_{k1} + \alpha_{\iota 2}\alpha_{k2} + \cdots = p_{k\iota}, \tag{50}$$

The variance of x_i is

$$p_{ii} = \alpha^2_{i1} + \alpha^2_{i2} + \cdots, \tag{51}$$

and under these hypotheses of independence among the α's cannot be constrained to be unity. Indeed, these assumptions define a *natural unit of measure* for each variate. If values, expressed in these natural units, of a sample of n variates, for each of a sufficient number of persons, are available, we may improve upon the method of analysis into principal components which we have heretofore considered. Under these circumstances, instead of using the matrix of correlations, we should apply the same operations to the matrix of covariances. The results obtained in this way have so much more elegant interpretations for the quantities analogous to the moments of the last section, giving for example exact standard error formulae, that it is worth while to consider their theory in spite of the fact that the absence or indefiniteness of natural units in the tests commonly met with in practice precludes direct applications at present. Exact standard errors for the e_k's of the last section have not been obtained, but some idea of their magnitudes may be based upon the exact results found for the corresponding quantities in the present section.

Wherever weights can be applied to tests which may reasonably be supposed to transform them approximately into the natural units here defined, such weights should be preferred to the equalization of their variances which we have heretofore considered. The considerations of the last section, which indicated that consistent results will be obtained under very general conditions when the variances are arbitrarily equalized if enough tests are used, suggest similarly that any reasonable system of weights will give consistent results.

To make our covariances as much as possible like correlations and thus to obtain results suggestive of those appropriate to the more common situation, let us take the mean value of p_{ii} as unity. Putting $Ep_{ii} = 1$ in (51) and using the second of (48) gives

$$\epsilon_1 = \phi_1 + \phi_2 + \phi_3 + \cdots = 1 \tag{52}$$

These ϕ's resemble but are not identical with those of the last section. The same is true of the quantities ϵ_h defined by

$$\epsilon_h = \sum_{q=1}^{\infty} \phi_q{}^h, \tag{53}$$

and of the other quantities defined below.

In place of (46), the characteristic equation for n tests expressed in natural units is

$$f^n - \left(\frac{S_1}{n}\right)f^{n-1} + \left(\frac{S_2}{n^2}\right)f^{n-2} - \left(\frac{S_3}{n^3}\right)f^{n-3} + \cdots = 0, \qquad (54)$$

where

$$S_1 = \sum_{i=1}^{n} p_{ii}, \qquad S_2 = \sum_{i>j} D_{ij}, \qquad S_3 = \sum_{i>j>k} D_{ijk}, \cdots, \qquad (55)$$

the notation now being

$$D_{ij} = \begin{vmatrix} p_{ii} & p_{ij} \\ p_{ji} & p_{jj} \end{vmatrix}, \qquad D_{ijk} = \begin{vmatrix} p_{ii} & p_{ij} & p_{ik} \\ p_{ji} & p_{jj} & p_{jk} \\ p_{ki} & p_{kj} & p_{kk} \end{vmatrix}, \qquad (56)$$

and so forth. We also put

$$e_h' = \sum_{q=1}^{n} f_q^h, \qquad (57)$$

the f_q's being the roots of (54). Evidently

$$\begin{aligned} e_1' &= \frac{S_1}{n}, \\ e_2' &= \frac{S_1^2 - 2S_2}{n^2}, \\ e_3' &= \frac{S_1^3 - 3S_1S_2 + 3S_3}{n^3}. \end{aligned} \qquad (58)$$

It is now a straightforward matter to find the expectations of the e''s, their variances, and any desired moments of them, in terms of the ϵ's. With the different conditions of the last section this was not possible by any means so far discovered, except for the simplest cases, the first moments of the e''s, which made possible the correction of bias.

The moments of the expressions in (58) may be found with the help of (55) and (56), from the mean values of the powers and products of the p_{ij}. These are obtained by means of (50), (48), (49), and whatever assumption is made regarding the higher moments of the α_{ij}'s. We may assume that the values of the α_{ij} for a fixed second subscript j, are normally distributed; from this it follows that

$$\begin{aligned} E\alpha^3_{ij} = 0, \qquad E\alpha^4_{ij} = 3E\alpha^2_{ij} = 3\phi^2_j; \qquad E\alpha^5_{ij} = 0; \\ E\alpha^6_{ij} = 15\phi^3_j, \cdots. \end{aligned} \qquad (59)$$

From (58) and (55) we have:

$$Ee_1' = \frac{ES_1}{n} = Ep_{ii} = 1;$$

while

$$Ee_2' = \frac{ES_1^2 - 2ES_2}{n^2}. \tag{60}$$

To evaluate the last expression we need ES_1^2 and ES_2.

$$ES^2_1 = \sum_{i=1}^{n} Ep^2_{ii} + 2\sum_{i<j} Ep_{ii}p_{jj}. \tag{61}$$

From (51), (59), (48), (52), and (53),

$$\begin{aligned} Ep^2_{ii} &= E\sum_q \alpha^4_{iq} + 2\sum_{q<r} \alpha^2_{iq}\alpha^2_{ir} \\ &= 3\sum_q \phi^2_q + 2\sum_{q<r} \phi_q\phi_r \\ &= 3\epsilon_2 + (\epsilon^2_1 - \epsilon_2) \\ &= 1 + 2\epsilon_2. \end{aligned}$$

Also, since p_{ii} and p_{jj} are independent when $i \neq j$,

$$Ep_{ii}p_{jj} = (Ep_{ii})(Ep_{jj}) = (Ep_{ii})^2 = 1. \tag{62}$$

Substituting in (61),

$$ES_1^2 = n(1 + 2\epsilon_2) + n(n - 1) = n^2 + 2n\epsilon_2. \tag{63}$$

Now from (50), with $i \neq j$,

$$Ep^2_{ij} = E\Big(\sum_q \alpha_{iq}\alpha_{jq}\Big)^2.$$

Since $E\alpha_{iq}\alpha_{jq}\alpha_{ir}\alpha_{jr} = 0$, when $q \neq r$, this equals

$$E\sum_q \alpha^2_{iq}\alpha^2_{jq} = \sum_q \phi^2_q = \epsilon_2$$

Substituting this result and (62) in the first of (56),

$$ED_{ij} = 1 - \epsilon_2. \tag{64}$$

Hence from (55),

$$ES_2 = \tfrac{1}{2}n(n - 1)(1 - \epsilon_2).$$

Putting this result and (63) in (60) we have

$$Ee_2' = \frac{1 + (n + 1)\epsilon_2}{n}. \tag{65}$$

Consequently an unbiassed estimate of ϵ_2 is

$$e_2 = \frac{ne_2' - 1}{n + 1}. \tag{66}$$

This discounts the sample value e_2', and therefore the fraction of variance attributable to the chief component, even more than does the corresponding result in the preceding section.

The calculations required to get the mean values and moments of the power sums e_n' are sufficiently illustrated by the foregoing procedure. The expectations of the powers and products of the $p_{\cdot i}$'s, up to the fourth degree, are tabulated at the end of this section. Those cases not given explicitly in the table are immediately obvious when account is taken of the following principles:

(*a*) The expectation of the product of *independent* quantities is the product of their expectations. Thus

$$Ep^2_{\cdot i}p^2_{kl} = (Ep^2_{\cdot i})\ (Ep^2_{kl}) = (Ep^2_{\cdot i})^2 = \epsilon^2_2,$$

where i, j, k, l, are all different.

(*b*) If the sum of the products of the exponents of the p's by the number of appearances in their respective subscripts of any letter is odd, the expectation is zero. Thus $Ep_{\cdot i}p_{\cdot j}$ and $Ep_{\cdot i}p^3_{\cdot j}$, vanish. The reason for this is that, when the p's are expressed in terms of the α's, each term will contain an odd number of factors with this subscript; their product is independent of the other factors, and will have the expectation zero.

With the help of the table we find by straightforward calculation:

$$Ee_3' = \frac{1 + 3(n + 1)\epsilon_2 + (n^2 + 3n + 5)\epsilon_3}{n^2}. \tag{67}$$

To obtain an unbiassed estimate e_3, of ϵ_3, we solve (65) and (67) for ϵ_3 in terms of Ee_2 and Ee_3, drop the symbol E, and replace ϵ_3 by e_3. This gives

$$e_3 = \frac{n^2e_3' - 3ne_2' + 2}{n^2 + 3n + 5}.$$

Standard errors and higher moments of these quantities can easily be, but have not been, deduced with the help of the table.

If instead of using the moments ϵ_k of the ϕ-distribution, we use the elementary symmetric functions β_k, we find the calculations somewhat simpler. The results are of course equivalent. We find as the unbiassed estimate of β_k,

$$b_k = \frac{S_k}{n(n-1)(n-2)\cdots(n-k+1)}.$$

The variances of b_1 and b_2 are:

$$\sigma^2_{b_1} = \frac{2\epsilon_2}{n},$$

$$\sigma^2_{b_2} = \frac{2}{n-1}(\epsilon_2 - 2\epsilon_3 + \epsilon_4) + \frac{1}{2n(n-1)}(1 - 4\epsilon_2 + 8\epsilon_3 + 6\epsilon^2_2 - 10\epsilon_4).$$

It should of course be remembered that these specific results are all based on the assumed normality of distribution of the coefficients of each component Γ in the population. Any other specific assumption as to these distributions would give analogous but not identical results.

TABLE ASSUMING NORMAL DISTRIBUTION OF THE α_{ij}, i, j, k ARE ALL UNEQUAL

First degree:

$Ep_{ii} = 1.$ $\qquad$ $Ep_{ij} = 0.$

Second degree:

$Ep^2_{ii} = 1 + 2\epsilon_2.$ $\qquad$ $Ep^2_{ij} = \epsilon_2.$

Third degree:

$Ep^3_{ii} = 1 + 6\epsilon_2 + 9\epsilon_3.$ $\qquad$ $Ep_{ii}p^2_{ij} = \epsilon_2 + 2\epsilon_3.$

$Ep^3_{ij} = 0.$ $\qquad$ $Ep_{ij}p_{jk}p_{ki} = \epsilon_3.$

Fourth degree:

$Ep^4_{ii} = 1 + 6\epsilon_2 + 18\epsilon^2_2 + 44\epsilon_3 + 37\epsilon_4.$

$Ep^2_{ii}p^2_{ij} = \epsilon_2 + 2\epsilon^2_2 + 4\epsilon_3 + 8\epsilon_4.$

$Ep_{ii}p_{ij}p^2_{ij} = \epsilon_2 + 4\epsilon_3 + 4\epsilon_4.$

$Ep^4_{ij} = 3\epsilon^2_2 + 6\epsilon_4.$

$Ep^2_{ij}p^2_{ik} = \epsilon^2_2 + 2\epsilon_4.$

$Ep_{ii}p_{ij}p_{ik}p_{jk} = \epsilon_3 + 2\epsilon_4.$

$Ep_{ii}p_{ij}p^2_{ik} = \epsilon_2 + 2\epsilon_3.$

$Ep_{ij}p_{jk}p_{ki}p_{li} = \epsilon_4.$

12. THE "SAND" AND "COBBLESTONE" THEORIES OF THE MIND

If a few mental characters such as general intelligence, cleverness, etc., are sufficient to account for virtually all the variance among individuals in all kinds of performances, we have a radically different situation from that of a large number of independent characters which all make small contributions to the variance. To distinguish between these two conditions, which have sometimes been referred to as the "cobblestone" and the "sand" theories, was one of the original objects in the analysis of mental tests.

One criterion which might be employed to distinguish between these two theories is that of normality of distribution of the scores on a test. If the score is the linear resultant of a large number of independent variates making approximately equal contributions to the variance, the second limit theorem of probability shows that a normal distribution is to be expected. However this criterion is somewhat difficult to apply, for various reasons. Even on the cobblestone theory, we might have a normal distribution, since each of the "cobblestones" might well itself be a normally distributed variate, so that the test is not decisive. Actually many distributions of scores are very far from normal, but this can often be ascribed to the very arbitrary units used; it is then customary deliberately to transform the variate into one of normal distribution.

Far more delicate distinctions are possible when only the correlations among the tests, and not the higher moments, are the basis of the calculations. The fundamental consideration then is that on the "sand" theory we should expect low correlations between pairs of tests, while on the "cobblestone" theory we anticipate high ones. This can be made more definite with the help of the ideas used in the last section, though in applying the distinctions we do not find it necessary to assume that the "natural units" there indicated are known, or to base an analysis upon covariances instead of unweighted correlations. The notation used in this section has the same meaning as in Section 11.

The primary question is as to the rapidity of convergence of the series (52), which may have a finite or an infinite number of terms. According to the "sand" theory the series should converge slowly, having no large terms, but a great number with approximately equal magnitudes, and larger than the rest. The "cobblestones" of the other theory may be interpreted as a few large ϕ's at the beginning of the series which account for nearly the whole of its value. There are, of course, infinitely many variations of the compromises between the views thus roughly described.

Taking the special form of the "sand" theory in which the first m of the ϕ's are equal and the rest zero, we may from our data estimate m and infer reasonable upper and lower limits for it. Indeed, from (50) and (51), the correlation of the ith and pth tests is

$$r_{ip} = \frac{\Sigma \alpha_{ij}\alpha_{pj}}{\sqrt{\Sigma \alpha^2_{ij} \Sigma \alpha^2_{pj}}}, \tag{68}$$

each of the summations being with respect to j and extending, under the theory we are now considering, from 1 to m.

Now (68) is also the expression for the correlation coefficient which would be computed from the independent quantities

$$\alpha_{i1} \quad \alpha_{i2} \quad \ldots \quad \alpha_{im}$$

$$\alpha_{p1} \quad \alpha_{p2} \quad \ldots \quad \alpha_{pm}$$

if the usual definition of the correlation coefficient were varied by omitting the requirement that the sample means be eliminated. Without this requirement, the distribution of the correlation coefficient in samples of m is the same as the distribution of the correlation coefficient as usually defined in samples of $m + 1$, provided that, as we have assumed for the α_{ij}'s, the quantities are normally and independently distributed about zero. This is evident from the same geometrical situation which led R. A. Fisher to the discovery of the distribution of the correlation coefficient.[1] The assumption that our tests are taken independently at random from the infinite aggregate implies that the population value of (68) is zero. Specific types of non-random selection of tests could be treated with the help of the distribution of r corresponding to non-vanishing values of the correlation in the population. However, on the assumption of randomness, we take the simplest case of Fisher's distribution, which, on putting $m + 1$ for the sample number, becomes:

$$\frac{1}{\sqrt{\pi}} \frac{\Gamma(\frac{1}{2}m)}{\Gamma[\frac{1}{2}(m-1)]} \quad (1-r^2)^{\frac{1}{2}(m-3)}dr. \tag{69}$$

The usual use of the sampling distribution of r is to determine whether, for a given sample size, the correlation is significantly greater than zero, or to find the greatest or least plausible true value of the correlation corresponding to a given probability of a greater discrepancy. But we shall make a different use of (69). Instead of knowing m and ascertaining whether r is excessive, we now consider that we know r, and wish to find the greatest value of m which, for a given level of probability, can be regarded as plausible. For a random pair of tests, the criterion is that the greatest acceptable value of m shall be that which reduces to some standard probability P, such as .05, the value of the integral of (69) *outside* of symmetrically placed limits which are both equal in absolute value to the sample correlation. The application of this criterion is very simple. For small samples,

[1] *Biometrika*, Vol. X, 1915, p. 507.

we use R. A. Fisher's Table V.A. in his *Statistical Methods for Research Workers*, which gives the values of r beyond which the integral of (69) takes the values .01, .02, .05 and .10, for various sample sizes. In this table, n is less by 2 than the sample size. To obtain m, we therefore, after finding the approximate value of our observed correlation in the column corresponding to our standard probability P (*e.g.* .05), increase the value of n given opposite it by unity.

For sufficiently large values of m, the distribution (69) approximates the normal form. Its variance, easily found by multiplying by r^2 and integrating, is exactly $\sigma^2 = 1/m$. Hence the limiting form of (69) is

$$\sqrt{\frac{m}{2\pi}} e^{-\frac{1}{2}mr^2} dr. \tag{70}$$

Instead of using Fisher's table we may, if m is large enough—that is, if r is small enough—use the fact that the probability .05 of a greater deviation corresponds, for the normal distribution, to the deviation 1.96σ. Putting this equal to the observed correlation r, we have

$$m = \frac{1}{\sigma^2} = \left(\frac{1.96}{r}\right)^2 = \frac{3.84}{r^2}$$

as the greatest acceptable value of m.

A minimum acceptable value of m can be found similarly, by requiring that the integral of (69) *between* the observed correlation and its negative as limits shall have the standard value P. However in this case we cannot use Fisher's table. Neither can we use the assumption of a normal distribution, unless r is very small, because the small values of m which must ordinarily correspond to minimum values make (69) far from normal. The simplest solution of this problem seems to be through trial values of m, expanding (69) in a series of powers of r^2 and integrating term by term. But it is only for correlations numerically less than $P(= .05$, say) that we can infer that m must be as high as 3. For the mental tests we have used for illustration, no corrected correlation so small as .05 occurs.

Between these upper and lower limits, the value of m may be estimated by the principle of maximum likelihood, which amounts to choosing m so to maximize (69). If in place of (69) we maximize its approximate value (70), our estimate is simply

$$m = \frac{1}{r^2}.$$

A more accurate estimate may be made with the help of a number of independent correlations, $r_1, r_2, \ldots, r_p$. Their joint distribution is obtained upon multiplying together expressions like (69). Multiplying together the approximate values (70) instead, taking the logarithm, and omitting terms which do not involve m, we have

$$L = \tfrac{1}{2}p \log m - \tfrac{1}{2}m\sum_{i=1}^{p} r_i^2,$$

which may be defined as the likelihood of m. It is a maximum for

$$m = \frac{p}{\Sigma r^2}. \tag{71}$$

The accuracy of this estimate, at any rate when p is large, is expressed by its variance, which is approximately[1] the negative reciprocal of the second derivative of L; that is

$$\sigma_m^2 = \frac{2m^2}{p}.$$

Now the correlations of any number of variates with a single variate are independent of each other, if all the correlations are based on samples of a given size from a normal distribution. This is geometrically evident when it is recalled that the correlation coefficient is the cosine of the angle between two random lines in hyperspace, and therefore of the angular distance between random points on a hypersphere. If one point is fixed, the distances of the others from it are obviously independent of each other.

From (71), applied to the correlations of the first with the other three variates given at the beginning of Section 5, we have as an estimate of m,

$$\frac{3}{(.698^2 + .264^2 + .081^2)} = 5.5.$$

To obtain an estimate of m based on the whole set of observed correlations, and not merely independent ones, we must consider the simultaneous distribution of all the correlations among n variates in samples of $N(= m + 1)$ from a normal population in which all correlations are zero. One way to obtain this is by the use of the partial correlations.

[1] Fisher, R. A.: On the Mathematical Foundations of Theoretical Statistics. *Phil. Trans.*, Vol. CCXXIIA, 1922, p. 328; H. Hotelling: The Consistency and Ultimate Distribution of Optimum Statistics. *Trans. Amer. Math. Soc.*, Vol. XXXII, 1930, pp. 847-859.

$$r_{23.1},\ r_{24.1},\ \ldots,\ r_{2n.1},$$
$$r_{34.12},\ r_{35.12},\ \ldots,\ r_{3n.12},$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$r_{n-1,\,n\cdot 123\ldots n-2},$$

which are cosines of angles between planes and hyperplanes determined by the random lines mentioned above, and are independent of each other and of $r_{12}, r_{13}, \ldots, r_{1n}$. Each of these has the distribution (69), with the value of m diminished for each partial correlation by the number of variates eliminated. Upon multiplying all these together, an expression is obtained whose general form, suggested for $n = 3$ may be established by mathematical induction. Putting

$$\omega = \begin{vmatrix} 1 & r_{12} & \ldots & r_{1n} \\ r_{12} & 1 & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{1n} & r_{2n} & \ldots & 1 \end{vmatrix}, \tag{72}$$

the simultaneous distribution of the correlations is

$$\frac{1}{\pi^{n(n-1)/4}} \frac{\Gamma^{n-1}\left(\frac{N-1}{2}\right)}{\prod_{i=2}^{n} \Gamma\left(\frac{N-i}{2}\right)} \omega^{\frac{N-n-2}{2}} dr_{12} \cdots dr_{n-1n} \tag{73}$$

The integral of (73) over all possible values of the r's must of course be unity. This will be true also if N is replaced by $N + 2k$. A ready means is thus found for evaluating the integral of the product of (73) by ω^k. In this way the kth moment of ω, measured from the origin (not the mean), may easily be shown to be

$$M_k = \frac{\Gamma^{n-1}\left(\frac{N-1}{2}\right) \prod_{i=2}^{n} \Gamma\left(\frac{N-i}{2} + k\right)}{\Gamma^{n-1}\left(\frac{N-1}{2} + k\right) \prod_{i=2}^{n} \Gamma\left(\frac{N-i}{2}\right)}. \tag{74}$$

This last result was reached in a different manner by S. S. Wilks,[1] who proceeded to derive the exact distribution of ω, which for $n = 3$ is expressed as a multiple of a hypergeometric function, and for higher values of n as a complicated multiple integral.

The result (73) may also be reached by starting from J. Wishart's distribution[2] of the sample covariances a_{ij}, putting $a_{ij} = r_{ij}s_is_j$, and integrating with respect to all the s's from 0 to infinity.

[1] *Biometrika*, Vol. XXIV, 1932, p. 492.

[2] *Biometrika*, Vol. XXA, 1928, p. 38, formula (9).

If the observed correlations have been fully resolved into their principal components, the value

$$\omega = k_1 k_2 \cdot \cdot \cdot k_n$$

is easily computed. On the "sand" theory, the correlations should tend to vanish, and ω should therefore be near unity. (ω necessarily lies between 0 and 1.) A value near zero might arise with high correlations. A zero value would imply a number of independent components less than the number of tests, some of the k's then vanishing.

The product of the k's for the example of Section 5 is .235. We may inquire what value of $m(= N - 1)$, with this value of ω, will make (73) a maximum. Since $n = 4$, we are to maximize

$$\frac{\Gamma^3\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m-1}{2}\right)\Gamma\left(\frac{m-2}{2}\right)\Gamma\left(\frac{m-3}{2}\right)}\omega^{\frac{m-5}{2}}.$$

An approximate solution is obtained by equating the values taken by this expression when $m - 1$ and $m + 1$ are put for m. This leads to the cubic

$$(m - 2)(m - 3)(m - 4) = 8\omega,$$

which for $\omega = .235$ has the single real root 6.50. The actual maximizing value found by trial calculations from Legendre's table of the gamma function, after interpolating parabolically from the values $m = 6, 7, 8$, is about 6.8.

The hypothesis of m or more principal components in the population of tests, making equal contributions to the total variance, while the remaining components are comparatively negligible, can be tested accurately with the help of ω, provided a way can be found to integrate the distribution discovered by Wilks, from 0 to ω. If the value of this integral is very small, we should reject the assumed value of m in favor of a smaller one. Likewise we could set an upper limit to plausible values of m by integrating from the observed value of ω to 1. Since no expression for the integral in manageable form, and no tables, are now available, we must in such questions be content for the present either with such information as can be obtained from the moments (74), or from the use of Fisher's table with individual correlations, or from our two kinds of maximum likelihood estimates, corresponding respectively to independent and to complete sets of correlations.