Indian J. pure appl. Math., 16 (12) : 1407-1419, December 1985

STATISTICAL DECOMPOSITION OF DIRECTED DIVERGENCE DUE TO A NUMBER OF FACTORS AND INTERACTIONS AMONG THEM

J. N. KAPUR, F. N. A.

Indian Institute of Technology, Kanpur, Kanpur 208016

(*Received* 18 *May* 1985; *after revision* 15 *July* 1985)

The directed divergence of one set of data from another set, arising due to a number of factors and the interactions among them has been decomposed into 'between' and 'within' various levels of each factor and each interaction, on the same lines which the corresponding decomposition for a single factor was done by Theil[19,20] and Kapur[10].

1. INTRODUCTION

One of the most important and useful properties of Shannon's[17] measure of entropy is that the total entropy can be decomposed into entropy between groups and entropy within groups. This statistical decomposition principle has been most forcefully explcited by Theil[20] in applications of entropy and divergence measures to the study of income inequalities, racial segregation, industrial concentration, political alignment, combination of statements in balance sheets, brand preferences in marketing and so on. In his applications, Theil[20] also used the statistical decomposition properties of measure of directed divergence due to Kullback and Leibler[12].

Kapur[10] in a recent paper has discussed the statistical decomposition of measures of entropy and directed divergence due to Shannon[17], Havrda and Charvat[3], Renyi[11], Sharma and Teneja[18], Ferrari[2] and Kapur[5-8]. Kapur[10] has also entended this principle of decomposition to measures of inaccuracy due to Kerridge[11] and Kapur[9] and to measures of useful information due to Belis and Guiasu[1]. He has also shown that the decomposition of the total $S\ S$ in the analysis of variance can be interpreted as analysis of directed divergence of order two.

In an alternative approach, Rao[13-15] has developed techniques for decomposition of diversity (DEDIV), apportionment of diversity (APDIV) and analysis of diversity (ANYDIV) motivated by problems of anthropology, genetics, economics, sociology and ecology.

In the present paper, we extend Kapur's[10] discussion to the case of a number of factors and of interactions among them.

2. Statistical Decomposition due to Two-Factors

Let x_{ij} and y_{ij} be the values in the two sets of data corresponding to the ith level of the first factor and jth level of the second factor ($i = 1, 2,\ldots, m; j = 1, 2,\ldots,n$). Let

$$\sum_{j=1}^{n} x_{ij} = T_{i.},\quad \sum_{i=1}^{m} x_{ij} = T_{.j},\quad \sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij} = \sum_{i=1}^{m} T_{i.}$$

$$= \sum_{j=1}^{n} T_{.j} = T_{..} \qquad \ldots(1)$$

$$\sum_{j=1}^{n} y_{ij} = U_{i.},\quad \sum_{i=1}^{m} y_{ij} = U_{.j},\quad \sum_{i=1}^{m}\sum_{j=1}^{n} y_{ij} = \sum_{i=1}^{m} U_{i.} = \sum_{i=1}^{n} U_{.j} = U_{..} \qquad \ldots(2)$$

then

$$D = \sum_{i=1}^{m}\sum_{j=1}^{n}\frac{x_{ij}}{T_{..}} \ln \frac{x_{ij}/T_{..}}{y_{ij}/U_{..}} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{x_{ij}}{T_{..}}$$

$$\times \ln \frac{x_{ij}/T_{i.}\; x_{ij}/T_{.j}\; T_{i.}/T_{..}\; T_{.j}/T_{..}}{y_{ij}/U_{i.}\; y_{ij}/U_{.j} U_{i.}/U_{..}\; U_{.j}/U_{..}}$$

$$= \frac{1}{2}\sum_{i=1}^{n}\frac{T_{i.}}{T_{i.}}\sum_{j=1}^{n}\frac{x_{ij}}{T_{i.}} \ln \frac{x_{ij}/T_{i.}}{y_{ij}/U_{i.}}$$

$$+ \frac{1}{2}\sum_{j=1}^{n}\frac{T_{.j}}{T_{..}}\sum_{i=1}^{n}\frac{x_{ij}}{T_{.j}} \ln \frac{x_{ij}/T_{.j}}{y_{ij}/U_{.j}}$$

$$+ \frac{1}{2}\sum_{i=1}^{m}\frac{T_{i.}}{T_{..}}\sum_{j=1}^{n}\frac{x_{ij}}{T_{i.}} \ln \frac{T_{i.}/T_{..}}{U_{i.}/U_{..}}$$

$$+ \frac{1}{2}\sum_{j=1}^{n}\frac{T_{.j}}{{..T}}\sum_{i=1}^{m}\frac{x_{ij}}{T_{.j}} \ln \frac{T_{.j}/T_{..}}{U_{.j}/U_{..}} = \frac{1}{2}\sum_{i=1}^{m}\frac{T_{i.}}{T_{..}} D_{i.}$$

$$+ \frac{1}{2}\sum_{j=1}^{n}\frac{T_{.j}}{T_{..}} D_{.j} + \frac{1}{2}\sum_{i=1}^{m}\frac{T_{i.}}{T_{..}} \ln \frac{T_{i.}/T_{..}}{U_{i.}/U_{..}}$$

(*equation continued on p.* 1409)

$$+ \frac{1}{2}\sum_{j=1}^{n}\frac{T_{.j}}{T_{..}} \ln \frac{T_{.j}/T_{..}}{U_{.j}/U_{..}} = \frac{1}{2}D_{w_1} + \frac{1}{2}D_{w_2}$$

$$+ \frac{1}{2}D_{B_1} + \frac{1}{2}D_{B_2} \qquad \ldots(3)$$

where

$$D_{i.} = \sum_{j=1}^{n}\frac{x_{ij}}{T_{i.}} \ln \frac{x_{ij}/T_{i.}}{y_{ij}/U_{i.}}, \quad D_{.j} = \sum_{i=1}^{m}\frac{x_{ij}}{T_{.j}} \ln \frac{x_{ij}/T_{.j}}{y_{ij}/U_{.j}} \qquad \ldots(4)$$

$$D_{w1} = \sum_{i=1}^{m}\frac{T_{i.}}{T_{..}} D_{i.}, \quad D_{w2} = \sum_{j=1}^{n}\frac{T_{.j}}{T_{..}} D_{.j} \qquad \ldots(5)$$

$$D_{B1} = \sum_{i=1}^{m}\frac{T_{i.}}{T_{..}} \ln \frac{T_{i.}/T_{..}}{T_{i.}/U_{..}}, \quad D_{B2} = \sum_{j=1}^{n}\frac{T_{.j}}{T_{..}} \ln \frac{T_{.j}/T_{..}}{U_{.j}/U_{..}} \qquad \ldots(6)$$

so that $D_{i.}$, $D_{.j}$ are the directed divergences within the ith and jth levels of the first and second factors, D_{w1} and D_{w2} are the directed divergences within the first and second factors and D_{B_1} and D_{B_2} are the directed divergences between the various levels of the first and second factors. Thus the total directed divergence has been decomposed into four parts, two of which correspond to directed divergence between the various levels of the first and second factors and the remaining two parts correspond to directed divergence within the various levels of the two factors.

In particular, x_{ij} and y_{ij} may represent the yields of jth variety in the ith block at two different times. The differences between x_{ij}'s and y_{ij}'s arise due to :

(i) differences in yield between varieties

(ii) differences in yield between blocks

(iii) differences in yields of different varieties within the same block

(iv) differences in yields of same variety in different blocks.

Our discussion gives the decomposition of the total directed divergence into four components, one due to each of the above four causes.

We may also consider the special case when y_{ij}'s are all equal i. e. we may compare the present yields against the hypothesis that all blocks are equally fertile and all varieties are equally good. Our present data do not agree fully with this hypothesis and our present discussion shows how far the differences are due to differences in varieties and how for these are due to differences in blocks.

In this case, we get

$$D_{B_1} = \sum_{i=1}^{m} \frac{T_{i.}}{T_{..}} \ln \frac{T_{i.}/T_{..}}{1/m} = \sum_{i=1}^{m} \frac{T_{i.}}{T_{..}} \ln \frac{T_{i.}}{T_{..}} + \ln m \qquad \text{...(7)}$$

$$D_{B_2} = \sum_{j=1}^{n} \frac{T_{.j}}{T_{..}} \ln \frac{T_{.j}/T_{..}}{1/n} = \sum_{j=1}^{n} \frac{T_{.j}}{T_{..}} \ln \frac{T_{.j}}{T_{..}} + \frac{1}{2} \ln n \qquad \text{...(8)}$$

$$D_{w1} = \sum_{i=1}^{m} \frac{T_{i.}}{T_{..}} \sum_{j=1}^{n} \frac{x_{ij}}{T_{i.}} \ln \frac{x_{ij}}{T_{i.}} + \ln m \qquad \text{...(9)}$$

$$D_{w_2} = \sum_{j=1}^{n} \frac{T_{.j}}{T_{..}} \sum_{i=1}^{n} \frac{x_{ij}}{T_{.j}} \ln \frac{x_j}{T_{.j}} + \ln n \qquad \text{...(10)}$$

$$D = \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{x_{ij}}{T_{..}} \ln \frac{x_{ij}}{T_{..}} + \ln m\, n \qquad \text{...(11)}$$

so that

$$- \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{x_{ij}}{T_{..}} \ln \frac{x_{ij}}{T_{..}} = - \frac{1}{2} \sum_{i=1}^{m} \frac{T_{i.}}{T_{..}} \ln \frac{T_{i.}}{T_{..}}$$

$$- \frac{1}{2} \sum_{j=1}^{n} \frac{T_{.j}}{T_{..}} \ln \frac{T_{.j}}{T_{..}} \cdot \frac{1}{2} \sum_{i=1}^{m} \frac{T_{i.}}{T_{..}} \sum_{j=1}^{n} \frac{x_{ij}}{T_{i.}} \ln \frac{x_{ij}}{T_{i.}}$$

$$- \frac{1}{2} \sum_{j=1}^{n} \frac{T_{.j}}{T_{..}} \sum_{i=1}^{m} \frac{x_{ij}}{T_{.j}} \ln \frac{x_{ij}}{T_{.j}} \qquad \text{...(12)}$$

or

$$S = \frac{1}{2} S_{B1} + \frac{1}{2} S_{B2} + \frac{1}{2} S_{w1} + \frac{1}{2} S_{w2} \qquad \text{...(13)}$$

where S is the total entropy, S_{B_1} and S_{B_2} are the entropies between the various levels of the two factors and S_{w1} and S_{w2} are the weighted sums of entropies within the various levels of the two factors.

3. Testing Statistical Significance of the Two-Factors

If instead of using Kullback-Leibler[12] measure of directed divergence, we use Havrda and Charvat[3] measure of degree two, and consider directed divergence from the uniform distribution, we find that D_{B_1}, D_{B_2}, D_{w1}, D_{w2} are the sums of squares between and within various levels, of the two factors.

In the usual two-way analysis of variance, we pool the sum of squares within the levels of the two factors to get $D_{w1} + D_{w2}$ as the error sum of squares so that the number of degrees of freedom associated with the error sum of squares $D_{w1} + D_{w2}$ is more than the number of degrees of freedom associated with the sum of squares D_{w1} or D_{w2} alone. We now find the ratios of mean squares corresponding to the ratios.

$$D_{B1}/(D_{w1} + D_{w2}),\ D_{B_2} / (D_{w1} + D_{w2}) \qquad \ldots(14)$$

and use the F test to test the significance of the two factors.

Our decomposition of directed divergence into directed divergence between and within various levels is more general than the usual decomposition of the sum of squares for the following reasons :

(i) We are using Kullback-Leibler[12] measure of directed divergence which is much more 'natural' then Havrda and Charvat[3] measure of degree two

(ii) Even if we want to use Havrda and Charvat measure, we can use this measure for any degree rather than for degree two only (Kapur[10]).

(iii) In the usual analysis of variance, we take the second distribution implicitly as the uniform distribution. Here we can take the second distribution as any general distribution.

(iv) In particular we can consider the directed divergence (d.d) of one set of data from another set of data and decompose this d. d into d. d between and within various levels of the two factors. We can then use

$$D_{B1}/D_{w1},\ D_{B_2}/D_{w2},\ D_{B_1}/(D_{w1} + D_{w2}),\ D_{B_2}/(D_{w1} + D_{w2}) \qquad \ldots(15)$$

as measures of the comparative influences of the two factors in the two sets of data.

To carry out a rigorous analysis of directed divergence, we have to find the exact distributions of the ratios (15) in random samples and this derivation is not going to be easy.

Even in the special case of analysis of variance, the exact distributions are available only when the parent populations are assumed to be normal. For other population distributions, the F distribution is valid only for large samples. In view of Jaynes[4] entropy concentration theorem, the distribution of the last two ratios in[15] will still be F distribution for large samples. Nevertheless knowledge of the exact distributions of these two ratios, valid for each value of n, would still be highly desirable.

4. Statistical Decomposition for Three or More Factors

The above decomposition can be easily extended to the case of k factors ($k > 2$). The total directed divergence can be decomposed into $2k$ directed divergences of which k will be directed divergences between different levels of the k factors and the other k directed divergence will be weighted sums of directed divergences within the various levels of the different factors. In fact we get

$$D = \frac{1}{k} \sum_{r=1}^{k} D_{Br} + \frac{1}{k} \sum_{r=1}^{k} D_{wr} \qquad \ldots(16)$$

where

D_{Br} = directed divergence between the various levels of the rth factor i. e. it is the directed divergence between two probability distributions obtained from the two sets of data by dividing the total for each level by the total for all levels for that factor and

D_{wr} = weighted sums of the directed divergence within all the levels of the rth factor, then the weights being given by the total of the levels.

We similarly get

$$S = \frac{1}{k} \sum_{r=1}^{k} S_{Br} + \frac{1}{k} \sum_{r=1}^{k} S_{wr} \qquad \ldots(17)$$

and the measures

$$\frac{S_{Br}}{S_{Br} + S_{wr}}, \quad \frac{D_{Br}}{D_{Br} + D_{wr}}, \quad r = 1, 2, \ldots, k. \qquad \ldots(18)$$

5. Statistical Decomposition for both Factors and Interactions

For simplicity, we consider two factors at m and n levels and the interaction between them.

Let x_{ijs}, y_{ijs} be the sth observations for ith and jth levels of the two factors and let a dot suffix denote summation with respect to the variable concerned, then

$$\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{s=1}^{N}\frac{x_{ijs}}{T_{\cdots}}\ln\frac{x_{ijs}/T_{\cdots}}{y_{ijs}/T_{\cdots}}=\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{n}\sum_{s=1}^{N}\frac{x_{ijs}}{T_{\cdots}}$$

$$\ln\frac{\dfrac{x_{ijs}}{T_{\cdots}}\;\dfrac{x_{ijs}}{T_{.j.}}\;\dfrac{x_{ijs}}{T_{ij.}}\;\dfrac{T_{i..}}{T_{\cdots}}\;\dfrac{T_{.j.}}{T_{\cdots}}\;\dfrac{T_{ij.}}{T_{\cdots}}}{\dfrac{y_{ijs}}{U_{i..}}\;\dfrac{y_{ijs}}{U_{.j.}}\;\dfrac{y_{ijs}}{U_{ij.}}\;\dfrac{U_{i..}}{U_{\cdots}}\;\dfrac{U_{.j.}}{U_{\cdots}}\;\dfrac{U_{ij.}}{U_{\cdots}}}$$

$$=\frac{1}{3}\sum_{i=1}^{m}\frac{T_{i..}}{T_{\cdots}}\sum_{j=1}^{n}\sum_{s=1}^{N}\frac{x_{ijs}}{T_{i..}}\ln\frac{x_{ijs}/T_{i..}}{y_{ijs}/U_{i..}}$$

$$+\frac{1}{3}\sum_{j=1}^{n}\frac{T_{.j.}}{T_{\cdots}}\sum_{i=1}^{m}\sum_{s=1}^{N}\frac{x_{ijs}}{T_{.j.}}\ln\frac{x_{ijs}/T_{.j.}}{y_{ijs}/U_{.j.}}$$

$$+\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{T_{ij.}}{T_{\cdots}}\sum_{s=1}^{N}\frac{x_{ijs}}{T_{ij.}}\ln\frac{x_{ijs}/T_{ij.}}{y_{ijs}/U_{ij.}}$$

$$+\frac{1}{3}\sum_{i=1}^{m}\frac{T_{i..}}{T_{\cdots}}\ln\frac{T_{i\cdots}}{T_{\cdots}}+\frac{1}{3}\sum_{j=1}^{n}\frac{T_{.j.}}{T_{\cdots}}\ln\frac{T_{.j.}}{T_{\cdots}}$$

$$+\frac{1}{3}\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{T_{ij.}}{T_{\cdots}}\ln\frac{T_{ij.}/T_{\cdots}}{U_{ij.}/U_{\cdots}}\,. \qquad \ldots(19)$$

This is similar to break-up into three factors; the third factor in this case is the interaction between the first two factors. Thus

$$D=\tfrac{1}{3}(D_{w_1}+D_{B_1})+\tfrac{1}{3}(D_{w_2}+D_{B_2})+\tfrac{1}{3}(D_{w_{12}}+D_{B_{12}}) \qquad \ldots(20)$$

and

$$S=\tfrac{1}{3}(S_{w1}+S_{B1})+\tfrac{1}{3}(S_{w2}+S_{B_2})+\tfrac{1}{3}(S_{w12}+S_{B12}). \qquad \ldots(21)$$

6. Statistical Heirarchical Decomposition

Let x_{ijkl} y_{ijkl} be the measurements of the lth individual in the kth population in the jth region of the ith species, then

$$D = \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} \sum_{l=1}^{N} \frac{x_{ijkl}}{T_{....}} \ln \frac{x_{ijkl}/T_{....}}{y_{ijkl}/U_{....}}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} \sum_{l=1}^{N} \frac{T_{ijkl}}{T_{....}}$$

$$\times \ln \frac{(x_{ijkl}/T_{ijk.})\,(T_{ijk.}/T_{ij..})\,(T_{ij..}/T_{i...})\,(T_{i...}/T_{....})}{(y_{ijkl}/U_{ijk.})\,(U_{ijk.}/U_{ij..})\,(U_{ij..}/U_{i...})\,(U_{i...}/U_{....})}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{n} \sum_{k=1}^{p} \frac{T_{ijk.}}{T_{....}} \sum_{l=1}^{N} \frac{x_{ijkl}}{T_{ijk.}} \ln \frac{x_{ijkl}/T_{ijk.}}{y_{ijkl}/U_{ijk.}}$$

$$+ \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{T_{ij..}}{T_{....}} \sum_{k=1}^{p} \frac{T_{ijk.}}{T_{ij..}} \ln \frac{T_{ijk.}/T_{ij..}}{U_{ijk.}/U_{ij..}}$$

$$+ \sum_{i=1}^{m} \frac{T_{i...}}{T_{....}} \sum_{j=1}^{n} \frac{T_{ij..}}{T_{i...}} \ln \frac{T_{ij..}/T_{i...}}{U_{ij..}/U_{i...}} \qquad \text{...(22)}$$

$$+ \sum_{i=1}^{m} \frac{T_{i...}}{T_{....}} \ln \frac{T_{i...}/T_{....}}{U_{i...}/U_{....}}$$

$$= D_{IV} + D_{III} + D_{II} + D_{I} \text{ (say)}, \qquad \text{...(23)}$$

where

D_I	= directed divergence between species
$D_{II} + D_{III} + D_{IV}$	= directed divergence within species
D_{II}	= directed divergence between regions (within species)
$D_{III} + D_{IV}$	= directed divergence within regions (within species)
D_{III}	= directed divergence between populations (within regions, within species)
D_{IV}	= directed divergence within populations (within regions, within species).

We can have a similar heirarchical decomposition of entropies.

7. Statistical Decomposition of Havrda and Charvat's Entropy

Let p_{ijkl} be the probability of an individual belonging to the ith religion, jth region, kth political party and lth economic status, and $p_{ijk\cdot}$ be the probability of his belonging to ith religion, jth region and kth political party and so on, then

$$\sum_i \sum_j \sum_k \sum_l p^{\alpha}_{ijkl} - 1 = \left(\sum_i p^{\alpha}_{i\ldots} - 1 \right)$$

$$+ \sum_i p^{\alpha}_{i\ldots} \left(\sum_j \left(\frac{p_{ij\cdot}}{p_{i\ldots}} \right)^{\alpha} - 1 \right)$$

$$+ \sum_j \sum_i p^{\alpha}_{ij\cdot\cdot} \left(\sum_k \left(\frac{p_{ijk\cdot}}{p_{ij\cdot\cdot}} \right)^{\alpha} - 1 \right)$$

$$+ \sum_k \sum_j \sum_i p^{\alpha}_{ijk\cdot} \left(\sum_l \left(\frac{p_{ijkl}}{p_{ijk\cdot}} \right) - 1 \right). \qquad \ldots(24)$$

By dividing by $1 - \alpha$, we get a decomposition of Havrda and Charvat's measure of entropy. Here $p_{ij\cdot\cdot}/p_{i\ldots}$ is the probability of a person of the ith religion belonging to the jth region, $p_{ijk\cdot}/p_{ij\cdot\cdot}$ is the probability of a person to belong to the kth political party when it is known that he belongs to the ith religion and jth region and $p_{ijkl}/p_{ijk\cdot}$ gives the probability of a preson belonging to the lth economic status when it is known that he belongs to the ith religion, jth region and kth political party.

We can of course take religious, regions, political parties and economic strate in twenty-four different orders to get twenty-four decompositions.

8. Statistical Decomposition for a Latin Square Design

Let x_{ij}, y_{ij} be the yields from the plot in the ith row, jth column in which kth variety is sown ($i, j, k = 1, 2, \ldots, n$), then

$$D = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_{ij}}{T_{\cdot\cdot}} \ln \frac{x_{ij}/T_{\cdot\cdot}}{y_{ij}/U_{\cdot\cdot}} = \frac{1}{3} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{x_{ij}}{T_{\cdot\cdot}}$$

$$\times \ln \frac{x_{ij}/T_{i\cdot} \; x_{ij}/T_{\cdot j} \; x_{ij}/T_k}{y_{ij}/U_{i\cdot} \; y_{ij}/U_{\cdot j} \; y_{ij}/U_k} \; \frac{T_{i\cdot}/T_{\cdot\cdot} \; T_{ij}/T_{\cdot\cdot} \; T_k/T_{\cdot\cdot}}{U_{i\cdot}/U_{\cdot\cdot} \; U_{ij}/U_{\cdot\cdot} \; U_k/U_{\cdot\cdot}} \qquad \ldots(25)$$

where T_k and U_k are the totals of the yields of kth variety in the two sets of data. Thus

$$D = \frac{1}{3}\sum_{i=1}^{n} \frac{T_{i\cdot}}{T}\sum_{j=1}^{n}\frac{x_{ij}}{T_{i\cdot}} \ln \frac{x_{ij}/T_{i\cdot}}{y_{ij}/U_{i\cdot}}$$

$$+ \frac{1}{3}\sum_{j=1}^{n} \frac{T_{\cdot j}}{T}\sum_{i=1}^{n}\frac{x_{ij}}{T_{\cdot j}} \ln \frac{x_{ij}/T_{\cdot j}}{y_{ij}/U_{\cdot j}} + \frac{1}{3}\sum_{k=1}^{n}\frac{T_k}{T}\Sigma' \frac{x_{ij}}{T_k} \ln \frac{x_{ij}}{T_k}$$

$$+ \frac{1}{3}\sum_{i=1}^{n}\frac{T_{i\cdot}}{T_{\cdot\cdot}} \ln \frac{T_{i\cdot}}{T_{\cdot\cdot}} + \frac{1}{3}\sum_{j=1}^{n}\frac{T_{\cdot j}}{T_{\cdot\cdot}} \ln \frac{T_{\cdot j}}{T_{\cdot\cdot}}$$

$$+ \frac{1}{3}\sum_{k=1}^{n}\frac{T_k}{T_{\cdot\cdot}} \ln \frac{T_k}{T_{\cdot\cdot}} \qquad \text{...(26)}$$

where Σ' denotes summation over all plots which gives the kth variety. This gives the decomposition of total directed divergence into directed divergence between and within rows, columns and varieties.

9. Interpretations for a Contingency Table

In an $m \times n$ contingency table, let O_{ij} be the observed frequency in the cell in the ith row and jth column and let a_i, b_j be the total of these frequencies in the ith row and jth column respectively, then on the basis of independence hypothesis, the expected frequency in the i-jth cell is $a_i\, b_j/T$, where T is the grand total of the table. The total directed divergence, between the observed frequency and the expected frequency table can then be decomposed according to (3) as follows :

$$\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{O_{ij}}{T} \ln \frac{O_{ij}/T}{a_i b_j/T^2} = \frac{1}{2}\sum_{i=1}^{m}\frac{a_i}{T} \ln \frac{a_i/T}{a_i/T}$$

$$+ \frac{1}{2}\sum_{j=1}^{n}\frac{b_j}{T} \ln \frac{b_j/T}{b_j/T} + \frac{1}{2}\sum_{i=1}^{m}\frac{a_i}{T}\sum_{j=1}^{n}\frac{O_{ij}}{a_i} \ln \frac{O_{ij}/a_i}{b_j/T}$$

$$+ \frac{1}{2}\sum_{j=1}^{n}\frac{b_j}{T}\sum_{i=1}^{m}\frac{O_{ij}}{b_j} \ln \frac{O_{ij}/b_j}{a_i/T}. \qquad \text{...(27)}$$

Since the row and column totals are the same in the two tables, there is no contribution from the directed divergences between rows and between columns and the total directed divergences is due to k directed divergences within rows and within columns.

Similarly (12) gives the decomposition of the total entropy in the observed frequency table viz.,

$$\sum_{j=1}^{n}\sum_{i=1}^{m}\frac{O_{ij}}{T}\ln\frac{O_{ij}}{T}=\frac{1}{2}\sum_{i=1}^{m}\frac{a_i}{T}\ln\frac{a_i}{T}+\frac{1}{2}\sum_{j=1}^{n}\frac{b_j}{T.}\ln\frac{b_j}{T}$$

$$+\frac{1}{2}\sum_{i=1}^{m}\frac{a_i}{T}\sum_{j=1}^{n}\frac{O_{ij}}{a_i}\ln\frac{O_{ij}}{a_i}$$

$$+\frac{1}{2}\sum_{j=1}^{n}\frac{b_j}{T}\sum_{i=1}^{m}\frac{O_{ij}}{b_j}\ln\frac{O_{ij}}{b_j}. \qquad \ldots(28)$$

10. Decomposition of Vector Observations Applications

Let X_{ij}, Y_{ij}, $T_{i.}$, $T_{.j}$, $T_{..}$, U_i, $U_{.j}$, $U_{..}$ be vectors with same number of components, then the directed divergence is also a vector and each component can be decomposed as above.

Then the components of X_{ij} can be

(i) yield of jth variety in the ith block

(ii) fertilizer given to the jth variety in the ith block

(iii) water given to the jth variety in the ith block

(iv) sale price of jth variety in the ith block

All these will be highly correlated.

It may therefore be worth while to investigate the relationships between the components of the directed divergence and entropy vectors.

We may also use the normalised measure of directed divergence between two probability distribution viz.

$$\left[\sum_{i=1}^{n} p_i \ln p_i/q_i\right]/\ln\frac{1}{q_{\min}} \qquad \ldots(29)$$

where $q_{\min}=\min(q_1, q_2, \ldots, q_n)$ ...(30)

so that each component of the directed divergence vector lies between 0 and 1. Let the components of the directed divergence vector be

$$(d_1, d_2, \ldots, d_n) \qquad \ldots(31)$$

then we form the probability distribution $p_i = d_i / \sum_{i=1}^{n} d_i$ and find the normailized entropy measure

$$-\sum_{i=1}^{n} p_i \ln p_i / \ln n. \qquad \ldots(32)$$

The nearer this value is to unity, the greater is the relationship between the various factors.

As another example, let $\bar{x}_{ij}, \bar{y}_{ij}$ be the mean marks of students in two schools in the ith subject in the jth class, then (3) will give the decomposition of the total directed divergence in marks of first school from second school in terms of directed divergence between classes, within classes, between subjects and within subjects.

Those are a large number of similar applications of the principle of decomposition given in the present paper to ecology, sociology, economics, anthropology genetics etc.

11. Concluding Remarks

(i) In the last section, we stated that the nearer the expression (32) is to unity, the greater will be relationship between the various factors. The question naturally arises as to how near to unity expression (32) should be. In other words, we need the distribution of $S/S_{\max}$ or of $S_{\max} - S$ where $S_{\max}$ is the maximum value of the entropy S. The answer is given by Jaynes[4] concentration theorem which states that in $2N$ trials, $2N\,(S_{\max} - S)$ is distributed as chi--square with $k = n - m + 1$ degrees of freedom where m is the number of constraints other than the natural constraint. However for using the result of this theorem, we need repeated sets of observations. If we have only a single set of observations, then (32) can give only an intuitive measure of the degree of relationship between the various factors.

(ii) Equations (3), (13) can be generalised to

$$D = \lambda\, D_{B_1} + \lambda_{B_2} + (1 - \lambda)\, D_{w_1} + (1 - \lambda)\, D_{w_2};\ 0 < \lambda < 1 \qquad \ldots(33)$$

$$S = \lambda\, S_{B_1} + \lambda\, S_{B_2} + (1 - \lambda)\, S_{w_1} + (1 - \lambda)\, S_{w_2};\ 0 < \lambda < 1. \qquad \ldots(34)$$

Similarly equations (20) and (21) can be generalised to

$$D = \lambda\,(D_{w_1} + D_{B_1}) + \mu\,(D_{w_2} + D_{B_2}) + \nu\,(D_{w_{12}} + D_{w_{12}}) \qquad \text{...(35)}$$

$$S = \lambda\,(S_{w_1} + S_{B_1}) + \mu\,(S_{w_2} + S_{B_2}) + \nu\,(S_{w_{12}} + S_{D_{12}}) \qquad \text{...(36)}$$

where

$$0 < \lambda < 1,\ 0 < \mu < 1,\ 0 < \nu - 1,\ \lambda + \mu + \nu = 1. \qquad \text{...(37)}$$

Similar generalisations can be made for equations (26), (27) and (28).

REFERENCES

1. M. Belis, and S. Guiasu, *IEER Trans. Inf. Th.* *IT*-4 (1968) 592–94.
2. C. Ferrari, *Statistica (Bologona)* **40** (2) (1980), 155–68.
3. J. Havrda, and F. Charvat, *Kybernatica* 3 (1967), 30–33.
4. E. T. Jaynes, *What is the Problem?* Proceedings First ASSP workshop on Spectral Estimation, McMaster University, (edited by S. Haykin and J. Cadrow), 1981, § 1.11 to § 1. 1. 10.
5. J. N. Kapur, *Math. Seminar* **4** (1967), 78–94.
6. J. N. Kapur, *J. Inf. Opt. Sci.* 3 (1983), 207–32.
7. J. N. Kapur, *Management Studies*, 3 (1984), 1–16.
8. J. N. Kapur, *IIT/K Res. Rep.* (1984).
9. J. N. Kapur, *IIT/K Res. Rep.* (1985).
10. J. N. Kapur, On the statistical decomposition of measure of entropy, inaccuracy, directed divergence and useful information, *Proc. Adm. Sci. Conf. Canada Conference Montreal*, 1985.
11. D. E. Kerridge, *J. R. Stat. Soc.* **23B** (1961), 184–94.
12. S. Kullback, and R. A. Leibler, *Ann. Math. Stat.* **22** (1951), 79–86.
13. C. R. Rao, *Sankhya*, **44A** (1982), 1–22.
14. C. R. Rao, *Theoretical Population Biology* **21** (1982), 24–43.
15. C. R. Rao, "Analysis of Diversity. A unified approach". *In Statistical Decision Theory and Related Topics* Vol. III 233 250, Academic Press, New York, 1982.
16. A. Renyi, *Proc. 4th Berkely Symp. Prob. Stat.* 1 (1961), 547–61.
17. C. E. Shannon, *Bell Syst. Tech. J.* 27 (1948), 379–423.
18. B. D. Sharma, and I. D. Taneja, *Informatic Verbet and Kybanates*, **13** (1977), 419–33.
19. H. Theil, *Economics and Information Theory*, North Holland, Amsterdam, 1967.
20. H. Theil, *Statistical Decomposition Analysis*, North Holland, Amsterdam, 1974.