An interesting question arises as to what is the analogue of the result given by equation (1), when the signs of all the numerator variates are reversed, that is, they now assume finite negative values while their corresponding U's remain positive as before. The probabilities for each pair remain unchanged. In this situation let $(-1)Z=z$, where Z is positive. Thus the variate value of $\sqrt{z}=\sqrt{-1}\sqrt{Z}$ will be a complex number without a real part, and since this circumstance rules out the use of Cauchy's inequality, we adopt the following stratagem to establish the analogous proposition. If the sign of each z is reversed, then obviously equation (1) can be applied. Now multiplying each side of this inequality by -1 we find:

$$-\mathrm{Cov}\left[U, \frac{Z}{U}\right] \geqq (\sqrt{-1})^2 V(\sqrt{Z}) \qquad (5)$$

In equation (5), as it stands, by factoring the coefficient -1 into the covariance function as a coefficient of Z, and $(\sqrt{-1})^2$ into the variance function as a coefficient of $\sqrt{Z}$, we obtain:

$$\mathrm{Cov}\left[U, \frac{z}{U}\right] \geqq V(\sqrt{z}) \qquad (6)$$

in view of the relation $z=-Z$. It is interesting to note that equation (1) has served as a scaffolding for the derivation of equation (6), where equality is attained if, and only if, $-z=cU^2$.

A statistically meaningful interpretation of the inequality equation (6) hinges on the interpretation of the lower bound of its covariance function, namely:

$$V(\sqrt{z})=V(\sqrt{-1}\sqrt{Z})=-V(\sqrt{Z}) \qquad (7)$$

That is, the variance of the complex variates $\sqrt{z}$, or let us say the abstract variance of the $\sqrt{z}$'s, is equal to the variance of the square root of their corresponding real variates multiplied by -1.

The last property of random variables as manifested by equation (6) may throw some light on the nature of the underlying variables in linear models used in the analysis of variance in circumstances when their estimates of variance are negative. It is hoped to investigate this question.

J. C. Koop

Department of Experimental Statistics,
University of North Carolina at Raleigh.

A Probabilistic Similarity Index

Williams *et al.*[1] have recently proposed that, for similarity indices used for numerical taxonomy, each attribute should be weighted according to the sum of the χ^2 values obtained from 2×2 tables testing the association of this attribute with all others. This relates the measure of similarity to the distance separating the two individuals in a multi-dimensional space where each axis representing an attribute has a scale depending on the importance of that attribute. Though based on χ^2 calculations, the index is not considered to be probabilistic, though the authors adumbrate the possibility of developing probabilistic methods of similarity analysis.

A similarity index based directly on probability theory is in use in this laboratory, which takes full account of ordered and quantitative attributes as well as of those which in their nature are binary. For each pair of individuals in a sample or population, the exact probability is computed for each attribute in turn that a random sample of two will resemble one another not less closely than the two under test. In computing this probability, the null hypothesis is that the two individuals in question are part of the same population as the rest, while the alternative hypothesis is that these two are a sample from another population.

The phrase "resemble one another not less closely" clearly calls for definition; the appropriate definition varies with the type of attribute.

For purely qualitative attributes, those concordances supporting the alternative hypothesis against the null hypothesis which are less probable under the latter are taken to be evidence of closer resemblance (or better evidence of resemblance). Thus, if an attribute can take three values x_1, x_2, and x_3, and the probabilities that, for any particular individual, the attribute will take these values are respectively 0·1, 0·3 and 0·6, agreement in having the first value would be considered better evidence of resemblance than agreement in having the second or third.

For ordered (but non-metrical) attributes, the degree of resemblance is taken as a function of the proportion of individuals in the same classes as the two under test, or in any intermediate class. Thus, if five ordered classes contain the following proportions of the population: 0·1, 0·2, 0·1, 0·4, 0·2, two individuals in the first and third classes are taken as being more similar than two in the fourth and fifth. For metrical attributes, the range is taken as a criterion of similarity, so that two individuals differing by less in the numerical values of this attribute are regarded as more similar than another pair in which the numerical values are more widely separated. A logarithmic scale might, of course, be used where appropriate.

The probability of the observed, or any closer, degree of resemblance having been computed for each attribute, the problem then remains of combining these probabilities, on the assumption of their independence. Where the probabilities arise from a continuous distribution, the solution has been given by Fisher[2]. He computed:

$$-2\sum_{i=1}^{n} \ln p_i$$

and showed that this quantity was a χ^2 variate with $2n$ degrees of freedom. For most types of attributes, this provides an acceptable approximation to the true multinomial distribution. An exception is formed by binary attributes—the most common type—for which the probability of the observed degree of resemblance between the two individuals is limited to three values. Here, Fisher's continuous solution may lead to a large positive bias, and the method adopted instead has been to combine these probabilities in small groups by the exact multinomial method (to apply it to any large number of attributes would be computationally impracticable), and then combine these group probabilities, now quasi-continuous, by Fisher's method.

When this has been done, the resulting combined probability P_t provides a satisfactory index of dissimilarity, and its complement $(1-P_t)$ is an index of similarity ranging (like those commonly used) from zero when the two individuals are completely dissimilar to near unity when they are closely similar. Unity can be attained only when the population on which the probability estimates are based is infinite.

The whole procedure outlined in this paper has been programmed in *Fortran II* for the IBM 1620_3 computer, and copies of the programme are available to enquirers. A full account of the technique is being prepared for publication.

I thank Mr. D. W. G. Moore, officer in charge of the Computing Centre of the University of Western Australia, for the facilities used in this work.

David W. Goodall

C.S.I.R.O. Division of Mathematical Statistics,
Western Australian Regional Laboratory,
Nedlands, Western Australia.

[1] Williams, W. T., Dale, M. B., and Macnaughton-Smith, P., *Nature*, **201**, 426 (1964).

[2] Fisher, R. A., *Statistical Methods for Research Workers*, thirteenth ed. (Oliver and Boyd, Edinburgh and London, 1963).