Brit. J. prev. soc. Med. (1966), **20**, 42–43

SOME TAXONOMIC IMPLICATIONS OF A CURIOUS FEATURE OF THE BIVARIATE NORMAL SURFACE

BY

J. H. EDWARDS

Department of Social Medicine, University of Birmingham

Consider a group of individuals on whom two variates are observable, and suppose that there are monotonic transformations of each which are normally distributed, and which are linearly related to one another so that their relationship may be sufficiently described by the correlation coefficient.

Suppose now that these individuals are specified by some arbitrary pair of thresholds into a 2×2 table, and let the classes in this table be classified arbitrarily, where a, b, c, d represent frequencies.

a	b
c	d

It would seem reasonable to suppose that the most direct intuitive appraisal of any interaction would be dependent on a, b, c, and d themselves, rather than any derived sum, as the marginal totals, etc., and that it would be based on some feature independent of scale, as a ratio.

The cross-ratio bc/ad would seem the most likely estimator in assessing the degree of association at a casual or subconscious level. This is not an incontrovertible opinion; it could be refuted by comparing the degrees of association of various pairs of attributes as assessed by the statistically uninformed, and assessing the ranking order with that of various other estimates, as $a/(a+b)-c/(c+d)$; $(ad-bc)/(a+b+c+d)^2$; etc.

Supposing that the cross-ratio were the dominant feature in the intuitive appraisal of association, then it has the remarkable property that its value is almost independent of the 'thresholds' involved in classification provided the proportions it defines are not too small. This is even more remarkable when it is appreciated that adjectives may be regarded as 'thresholds' specifying regions of distributions.

The relative constancy of the cross-ratio may easily be demonstrated from tabulations of the tetrachoric functions. It appears impossible to make a simple and close algebraic approximation, although the identity in the limiting situation $a=d, b=c$, $r \to 0$,

$$z=(1/2)\log_e((1+r)/(1-r))=(\pi/8)\log_e(bc/ad)$$

may be demonstrated (Edwards, 1957), and this remains a close approximation for all but the largest values of r for dichotomies within the range $\pm\sigma$, and a fairly close approximation within the range $\pm 2\sigma$ when $r \leqslant 0{\cdot}5$.

The important taxonomic implication is that, to a close approximation (in fact the robustness of the correlation coefficient to unjustified assumptions of normality or homoscedasticity probably limits the importance of these restraints), the apparent intensity of association of the two variates is largely independent of the opinions, or measuring apparatus, used to specify the dichotomies.

To give an example: a group of investigators sent out with a stick and a piece of string to measure the height and girth of a population would all reach a fairly close agreement about the intensity of the relationship as specified by the ratio (short thick $\times$ tall thin)/(tall thick $\times$ short thin.) This agreement would, within wide limits, be independent of any consistency in the lengths of the sticks and strings. Similarly, the apparent intensity of association of such variates as stupidity in parent and child, disordered electrocardiograms and hypertension, etc., are largely independent of the relative proportions specified by such terms when used by various observers. Estimation procedures which exploit this curious characteristic of the bivariate normal surface, such as that due to Woolf (1954), may be said to show semantic invariance.

The taxonomic importance of this is that it may appear intuitively reasonable to suppose that such constancy is indicative of a true bimodality in at least one of the variates. Indeed, the casually acquired and firmly entrenched opinions which still obstruct various continuous and quasi-continuous

concepts of feeble-mindedness, neuroticism, psychotism, hypertension, etc., would seem strong evidence of this intuitive failing, and of the primary 'feeling' of a bimodality when dealing with partitions of an indistinctly perceived unimodal distribution.

Not only does this constancy of the cross-ratio appear a sufficient reason for this misunderstanding; it is also liable to misinterpretation as implying that there is some special feature in the arbitrary threshold adopted by each individual. In terms of the stick and string analogy it might seem reasonable to infer from the constancy of the results that the sticks and strings were all of the same dimensions, or, at any rate, that they were paired in some particular way. In fact, the results contain almost no information relevant to this. A similar case arises in genetic segregation, where the consequences of a character influenced by a multiplicity of loci, or by only two genes at one locus, may be indistinguishable on quite extensive data (Edwards, 1960).

A similar intuitively surprising relationship exists if two distinct populations exhibit some measurable character, some transformation of which is distributed normally, and if the variance is the same in the two populations. In this case, dichotomizing this variate will lead to a 2×2 table whose cross-ratio will be more or less invariant in spite of variations in the threshold of dichotomy. (For the logistic distribution this cross-ratio is absolutely invariant. In practice, the normal and logistic distribution differ so slightly that the difference can rarely be appreciated on less than a hundred observations.) The cross-ratio will, in fact, be very nearly equal to $4d/(e^{\sqrt{2\pi}})$, where d is the difference between the means in units of standard deviation.

If one reconsidered the previous analogy, one could obtain fairly accurate information on the difference in height between, say, Welshmen and Scotsmen, by defining tallness and shortness with reference to a stick of unknown length. The most interesting implication of this relationship is that it appears to some extent to explain the remarkable way in which quite arbitrary levels of significance appear to lead to consistently satisfactory results in the opinion of many persons using them for practical purposes.

A similar situation arises in the rejection of null hypotheses using significance tests. Consider a trial of a substance which makes some plants k units taller, and that one group of plants under test shows an increase in height of h units. Conventional significance tests involve considering only the two situations $k=0$ and $k=h$, and the expected distribution of the mean on these two hypotheses may be represented as two overlapping normal distributions with means at 0 and h and with the same variance.

Consider now any arbitrary level of statistical significance dichotomizing the distribution of the test statistic on the null hypothesis and also the distribution expected on the other hypothesis, and let the partitions of these distributions be represented by the areas a, b, c, and d.

Then $b/(a+b)$ is the level of significance.

If H_0 is correct the ratio of the risks of accepting H_0 correctly (say R_1) to rejecting it wrongly (say E_1) is a/b, and, if H_h is correct, the ratio of accepting H_h correctly (say R_2) to rejecting it wrongly (say E_2) is d/c.

Now, if the intuitive appraisal of the reliability of statistical tests were to depend on the value of

$$E_1E_2/R_1R_2,$$

which is equal to bc/ad, then this intuitive appraisal will be dependent on a characteristic relatively invariant under different levels of significance, including conventionally ridiculous ones (*i.e.* 50 per cent.). Similarly, any consistent errors in the tabulation of functions related to significance levels need have no adverse practical consequences. It may be added that any prior probabilities regarding the expectation that H_0 and H_h is operative would influence the cross-ratio by a constant amount, and would not influence the possibility of the intuitive consistency of the reliability being relatively uninfluenced by the level of significance.

These arguments are not advanced with any intention of discrediting the practical value of many applications of significance tests or other criteria of credulity or surprise. It is merely intended to suggest that, when one of several methods which might underlie the intuitive appraisal of association is operative, circumstances could arise in which any level of significance would appear equally useful, and that, in such cases, their pragmatic defence on grounds of consistency might be invalid.

SUMMARY

A curious feature of dichotomizing classifications is discussed, and the concept of semantic invariance advanced. An implication of the approximately invariant features considered is the possibility that the long-term consequences of decisions based on significance tests may be relatively invariant to the levels of significance, or to the precision of any tabulated functions used.

REFERENCES

Edwards, J. H. (1957). *Brit. J. prev. soc. Med.*, **11**, 73.
—— (1960). *Acta genet. statist. Med.*, **10**, 63.
Woolf, B. (1955) *Ann. hum. Genet.*, **19**, 251.