Brit. J. Psychiat. (1983), **143**, 487–489

Coefficients of Agreement

M. E. DEWEY

Summary: Several coefficients have been suggested for the measurement of agreement between raters. Recently Maxwell (1977) has proposed, and Janes (1979) has extended, a new coefficient, which they have termed the *random error* coefficient *(RE).* This paper criticizes their proposal and suggests that Cohen's (1960) coefficient *kappa* (K) is still the measure of choice.

A diagnostic system involves a set of discrete categories, and the accuracy with which a patient can be assigned to a category in such a system is of interest both in its right and as a preliminary to almost any psychiatric research.

It is natural to seek to represent the agreement between two psychiatrists by a single number, by analogy with the well-known coefficient of correlation available for the case of quantitative variables. It might seem most convenient to simply take the proportion of cases on which the psychiatrists agree, ranging form 1 in the situation where they agree on every case to 0 in the situation where they disagree on every case. Unfortunately this otherwise attractive suggestion fails to take account of the fact that some agreement may reflect chance agreement, and attempts have therefore been made to correct the measure for this chance agreement. Fleiss (1975) gives a good review of the work done in this area up to that year.

One method due to Cohen (1960, 1968) has achieved a degree of success and popularity, but it has recently been challenged in this journal by Maxwell (1977) and Janes (1979). It is my purpose here to suggest that the new measure is not an advance on Cohen's coefficient *kappa,* particularly in Janes's extended version.

Derivation of K and RE

If two psychiatrists independently assess each subject in a sample of n patients for the *presence* or *absence* of a particular symptom (no degree of severity of the symptom being considered), their conclusions can be summarized as in Table I. For clarity we have used the same notation as Maxwell: a is the fraction of patients considered by *both* psychiatrists to have the symptom, b is the fraction considered by the first to have it and by the second not to have it, and similarly for c and d.

Hence $a + b + c + d = 1$

The row total p_1 equals $(a + b)$ and is the overall proportion of patients judged by the first psychiatrist to have the symptom; likewise q_1 equals $(c + d)$ and is the proportion of patients judged by the first psychiatrist *not* to have the symptom.

Hence $p_1 + q_1 = 1$

For the second psychiatrist the corresponding variables are:

$p_2 = a + c$
$q_2 = b + d$

Again, $p_2 + q_2 = 1$

The two psychiatrists agree over $(a + d)$ of the patients and disagree over $(b + c)$. The *observed* proportion of agreement $(a + d)$, denoted by p_0, is adjusted by Cohen to take account of chance agreement, as follows:

Let $p\hat{c} = p_1 p_2 + q_1 q_2$

Given that p_1 is the proportion of patients assessed by the *first* psychiatrist as having the symptom and p_2 the proportion assessed by the *second* as having it, $p_1 p_2$ would be the expected rate of agreement on the presence of the symptom if each were assigning symptom-present/symptom-absent scores at random, *but in the observed proportions. Similarly* $q_1\ q_2$ is the rate of agreement on *absence* of symptom that might be expected on the basis of chance, given the observed ratios $p_1 : q_1$ and $p_2 : q_2$. $P\hat{c}$ is therefore the expected value of chance agreement.

Cohen's coefficient *kappa* is then calculated as

$$K = \frac{p\hat{o} - p\hat{c}}{1 - p\hat{c}}$$

The numerator represents the excess of observed over chance agreement and the denominator the maximum possible excess, given the observed marginal ratios $p_1 : q_1$ and $p_2 : q_2$.

The feature critized by Maxwell and by Janes is the

https://doi.org/10.1192/bjp.143.5.487 Published online by Cambridge University Press

TABLE I

Proportions of a sample of n *patients having a particular symptom, as assessed independently by two psychiatrists*

		Second psychiatrist		
		Symtom present	Symptom present	Row totals
First psychiatrist	Symptom present	a	b	p_1
	Symptom absent	c	d	q_1
	Column totals	p_2	q_2	

TABLE II

The data of Table I, expressed in the notation used by Maxwell (1977) in the derivation of the random error coefficient

		Second psychiatrist	
		Symptom present	Symptom absent
First psychiatrist	Symptom present	$P_1 + \frac{1}{2}Q$	$\frac{1}{2}Q$
	Symptom absent	$\frac{1}{2}Q$	$P_0 + \frac{1}{2}Q$

use of the product of the marginal proportions to estimate the proportion of agreement expected on the basis of chance.

In the derivation of their coefficient, named by Maxwell the *random error coefficient* (RE), the starting point is the two proportions P_1 and P_0. P_1 is defined as the proportion of patients in a sample that are assessed by both psychiatrists, *without doubt,* as having the symptom: P_0 is the proportion that are assessed by both psychiatrists, *without doubt,* as *not* having the symptom. The quantity Q, equal to $1 - (P_1 + P_0)$, is then the proportion of patients about whom there is doubt. Maxwell assumes that each psychiatrist makes a decision on each of these patients on a purely random basis, so that they are distributed equally over the four cells of Table I. This gives us Table II. The relationships between a, d, P_1 and P_0 are:

$$a = P_1 + \tfrac{1}{4}[1 - (P_1 + P_0)]$$
$$d = P_0 + \tfrac{1}{4}[1 - (P_1 + P_0)]$$

It is simple to show that

$$P_1 + P_0 = (a + d) - (b + c)$$

The term $(P_1 + P_0)$, which Maxwell selects as the *random error coefficient,* can therefore be quickly calculated from the observed diagnostic decisions, and represents the excess of agreement over disagreement.

Janes extends Maxwell's argument to situations in which symptoms are assessed on a scale rather than as simply present/absent, thus producing a table larger than 2 × 2.

Numerical considerations

It is generally considered desirable for measures of association to lie between 0 and +1 (or −1 and +1 if they are directional). In the case of complete disagreement the lower bound of RE is not −1 but $-1/(k-1)$, where k^2 is the number of cells in the matrix corresponding to Table I. This lower bound equals −1 if and only if $k = 2$. The corresponding result for *kappa* is less simple, as it depends on the marginals: however, if the marginal proportions are correlated (as they will be in practice), the lower bound approaches −1 as k increases (Cohen, 1960). A measure like RE, with a lower bound close to zero for large values of k, compresses the whole range of disagreement into a very narrow band. Moreover, the comparison of tables of different sizes is made more hazardous, because the range of possible values for RE varies in tables of different sizes.

It is only fair to point out at this stage that the practice of norming measures of association to lie in a certain range has been criticized. While describing the aim of comparing tables of varying sizes as 'worthy', Bishop, Fienberg and Holland (1975) wrote: 'we know of no evidence that shows norming can really accomplish it.'

One problem with the argument behind the derivation of RE is that the quantities P_0 and P_1, which can be calculated individually from the values of a, b, c, and d, may take on negative values. For example, if $a = 0.04$, $b = c = 0.16$ and $d = 0.64$, then $P_1 = -0.12$ and $P_0 = +0.48$ (Maxwell's own example). No

https://doi.org/10.1192/bjp.143.5.487 Published online by Cambridge University Press

meaning can be attached to this value of P_1: it does not denote certain agreement in some proportion (−0.12) of n patients. Maxwell states that such a negative value would 'cast doubt on the validity of the judgements' and that 'this would be valuable information to have.' This seems a curious line to take: P_1 and P_0 were originally defined as estimators of probabilities, and these always have values in the range 0 to 1. What we are doing is building a mathematical model of a real-life situation. It is usually accepted in statistics that if the data do not fit the model, and in Maxwell's example they do not, then we should reject the model, not 'cast doubt' on the data. If an estimator of a parameter which is known to be non-negative gives us a negative estimate, then that casts doubt not on the validity of psychiatrists' judgements, but on the validity of our model of their behaviour.

No similar problem occurs with κ.

Problems of interpretation

Moving to questions of a deeper logical nature, Maxwell's major criticism of Cohen is his use of marginal proportions in the calculation of *kappa*. This criticism is implicit rather than explicit, although he does state, that RE is attractive because it does not involve 'reference to prior probabilities of symptom incidence'. In fact, κ does not rely on prior probabilities but on estimates derived from the observed marginal proportions.

The value of a measure like κ which takes into account information from the marginal proportions, as opposed to RE which does not, becomes clearer if we look at the extreme case of a rater who shows no discrimination at all. Taking Maxwell's own example, where $a = 0.04$ $b = c = 0.16$, and $d = 0.64$, we obtain a value for κ of 0 and for RE of 0.36. Now suppose that the second psychiatrist failed to recognise the symptom at all, while the first made the same judgements as before. This would give $a = c = 0$, $b = 0.20$ and $d = 0.80$, κ again has the value 0 but RE now has the value 0.60, indicating a higher degree of agreement. The usefulness of such an indication is difficult to uphold: the second rater is not showing any discrimination at all, and there is no value in agreement between someone who does not discriminate and someone who does. This defect does not occur in the use of κ, which is always zero if one rater uses only one category.

A corollary of this line of argument is that raters can artificially inflate the reliability of their judgements (as measured by RE) by denying the existence of any category which they know to occur with low frequency in a given series. This could create serious problems in community studies where the *a priori* probability of any particular symptom occurring is low: Mellor (1970) showed that even in a series of schizophrenic admissions no first-rank symptoms was present in more than 27 per cent of his cases. Given the pressures to achieve reliability (Robinson *et al*, 1982) the dangers are obvious.

Conclusion

Like the related problem of measurement of association, the measurement of agreement is a difficult one, particularly in tables where the variables of interest have more than two possible values. Nonetheless, it is possible to distinguish between good and less good measures of agreement. The preceding sections attempt to show that Maxwell's proposed new measure (RE) has no advantage over Cohen's κ and in fact has some serious disadvantages, particularly when applied to tables with more than four cells.

References

BISHOP, Y. M. M., FIENBERG, S. E. & HOLLAND, P. W. (1975) Discrete multivariate analysis. Cambridge, Mass.: MIT Press.

COHEN, J. (1960) A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, **20,** 37–46.

—— (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70,** 213–20.

FLEISS, J. (1975) Measuring agreement between two judges on the presence or absence of a trait. *Biometrics*, **31,** 651–9.

JANES, C. (1979) An extension of the random error coefficient of agreement to N×N tables. *British Journal of Psychiatry*, **134,** 617–19.

MAXWELL, A. E. (1977) Coefficients of agreement between observers and their interpretation. *British Journal of Psychiatry*, **130,** 79–83.

MELLOR, C. (1970) First rank symptoms of schizophrenia. *British Journal of Psychiatry*, **117,** 15–23.

ROBINSON, M., COOKSON, I. & WHITE, K. (1982) The 'consentiam' effect: Are your joint ratings really independent? *British Journal of Medical Psychology*, **55,** 285–6.

M. E. Dewey, B.A., *Lecturer in Psychological Statistics, Department of Psychiatry, University of Liverpool, P.O. Box 147, Liverpool L69 3BX*

(Received 14 January; revised 1 February 1983)

https://doi.org/10.1192/bjp.143.5.487 Published online by Cambridge University Press