Brit. J. Psychiat. (1971), 118, 659–62

Agreement Among Raters

By A. E. MAXWELL

Introduction

It is frequently the case in investigations in the behaviour sciences that a number of independent raters are asked to rate the same sample of subjects with regard to certain signs, symptoms or characteristics of these subjects, and the question of comparing the results given by the raters then arises.

For a set of *m* variables (signs, symptoms, etc.) and a random sample of *n* subjects drawn from some population, the results given by each rater can be tabulated in an $n \times m$ table in which the entry y_{ij} in the i-th row and j-th column of the table, is the score given by the rater to the i-th subject on the j-th variable. In many instances the scores are ratings on a five- or seven-point rating scale. If they fulfil certain well-known assumptions then a comparison of the results given by two (or more) raters can be made efficiently by a two-way analysis of variance, with interaction, of the data. An appropriate model, which enables possible correlation between the variables to be taken into account, is fully discussed elsewhere (e.g. Greenhouse and Geisser, 1959). The results of the analysis of variance may also be employed to derive coefficients of internal consistency of the data and to provide measures of agreement between the raters (e.g. Maxwell and Pilliner, 1968).

But in investigations of the type in question it is not uncommon for the scores y_{ij} to be restricted to the values '1' and '0', where '1' indicates that a sign or symptom is present and '0' that it is absent. Provided the sample size and the number of variables are both fairly large, an analysis of variance approach might again reasonably be employed to assess possible differences between raters. But more elementary procedures are also available. One of these, which is readily applicable when only two raters are involved, is described in this paper.

Model for Dichotomously-scored Data

In the case of two raters and scores y_{ij} restricted to the values 1 or 0, a simple measure of agreement would be the *proportion* of times on which the raters agreed. However, the amount of agreement might well vary from variable to variable, or from patient to patient, and it would be desirable to test for such variation. If the latter were found to be negligible then one would be satisfied that the *proportion* provided a consistent measure of agreement for the data as a whole.

Let us denote the two raters by the letters *a* and *b* respectively. Let y_{ija} denote the score given by rater *a* to the i-th patient on the j-th variable, and y_{ijb} denote the corresponding score given by rater *b*. Let us now set up a table (see Table I), in which the entries x_{ij} represent agreement (or disagreement) between the two raters. The rules for constructing Table I are as follows:

$$\text{if } y_{ija} = y_{ijb} \quad \text{set } x_{ij} = 1$$
$$\text{and if } y_{ija} \neq y_{ijb} \quad \text{set } x_{ij} = 0.$$

In Table I the total of the i-th row is indicated by X_i and the proportion of agreements by $P_i = X_i/m$. Similar statistics for the columns of the Table are x_j and $p_j = x_j/n$. It is clear that a test of the equality of the proportions P_i would provide a test of whether agreement (or equivalently disagreement) between the two raters varied 'among subjects'. Similarly a test of the equality of the proportions p_j would provide a test of whether agreement between the two raters varied 'among variables'.

TABLE I

Agreements (1) and disagreements (0) between two raters who rate n subjects on m signs: $x_{ij} = 1$ or 0

Subjects	Variables							Totals	Proportions
	1	*2*	*3*	...	*j*	...	*m*	$\underline{X}$	(P_i)
1	x_{11}	x_{12}	.	. .	.	.	x_{1m}	X_1	X_1/m
2	x_{21}				.		.	X_2	.
3	.				.		.	.	.
.	.				.		.	.	.
i	x_{i1}	.	.	. .	x_{ij}	.	x_{im}	X_i	X_i/m
.	.				.		.	.	.
.	.				.		.	.	.
n	x_{n1}	.	.	. .	.	.	x_{nm}	X_n	X_n/m
Totals	x_1	x_2			x_j		x_m	ΣX	
Proportions (p_j)	x_1/n			...	x_j/n		x_m/n		

In each case the test statistic required is similar to that used in Cochran's Q-test (Cochran, 1950). To test whether the Ps differ amongst themselves we calculate (see Table I)

$$\chi^2 = \sum_{j=1}^{m} (x_j - \bar{x})^2 / \sum_{i=1}^{n} P_i Q_i, \qquad (1)$$

and refer the calculated value to the chi-square distribution with $(m - 1)$ degrees of freedom. In equation (1) $\bar{x}$ is the mean of the x_j's and $Q_i = 1 - P_i$. Similarly, to test whether the p's differ amongst themselves we calculate:

$$\chi^2 = \sum_{i=1}^{n} (X_i - \bar{X})^2 / \sum_{j=1}^{m} p_j q_j \qquad (2)$$

based on $(n - 1)$ degrees of freedom. In this instance $\bar{X}$ is the mean of the X_i's, and $q_j = 1 - p_j$. Interpretation of the information provided by these two tests can most easily be appraised by considering a practical example.

AN EXAMPLE

Two psychiatrists independently interviewed a sample of depressed patients and noted the presence or absence of each of a list of symptoms. Four of the symptoms were *i* worrying, *ii* anxiety, *iii* depression and *iv* irritability. The ratings given to the first patient were as follows:

	Symptoms			
	i	*ii*	*iii*	*iv*
1st psychiatrist	1	1	0	1 ,
2nd psychiatrist	1	0	1	0 ,

Hence the first row of the agreement table (Table II) is

$x_{1j} =$	1	0	0	0 .

In other words both psychiatrists agreed that this patient had the symptom 'worrying', but they disagreed about the presence or absence of the other three symptoms. For a total of just 10 patients on the four symptoms (to keep the sample simple) the ten vectors of agreement scores are given in Table II. Examination of the results in this Table shows that agreement among raters, where patients are concerned, is perfect for the fourth patient, poor for the first patient and intermediate for the others. Where symptoms are concerned agreement is relatively good for *i* and *iv* but only average for *ii* and *iii*.

The preliminary calculations for the significance tests are shown in Table II; the remaining calculations are as follows:

A Chi-square test 'among patients', using equation 1:

$$\sum p_j q_j = 0{\cdot}87;$$

$$\sum (X_i - \bar{X})^2 = 1^2 + 2^2 + \ldots + 3^2 - 25^2/10 = 6{\cdot}5.$$

$\chi^2 = 6{\cdot}5/0{\cdot}87 = 7{\cdot}47$, d.f. $= 9$, not significant.

https://doi.org/10.1192/bjp.118.547.659 Published online by Cambridge University Press

TABLE II

Agreement table

Patients	Symptoms *i*	*ii*	*iii*	*iv*	Total X_i	Proportions $P_i = X_i/m$	Q_i
1	1	0	0	0	1	0·25	0·75
2	1	0	1	1	3	0·75	0·25
3	0	0	1	1	2	0·50	0·50
4	1	1	1	1	4	1·00	0·00
5	1	1	0	1	3	0·75	0·25
6	1	0	0	1	2	0·50	0·50
7	1	0	1	0	2	0·50	0·50
8	1	1	0	0	2	0·50	0·50
9	1	1	0	1	3	0·75	0·25
10	0	1	1	1	3	0·75	0·25
Total x_j	8	5	5	7	25	$\Sigma P_iQ_i = 1{\cdot}9375$	
$P_j = x_j/n$	0·8	0·5	0·5	0·7			
$q_j = 1 - p_j$	0·2	0·5	0·5	0·3	$\Sigma p_jq_j = 0{\cdot}87$		

B Chi-square test 'among symptoms', using equation 2:

$$\sum P_iQ_i = 1{\cdot}9375;$$
$$\sum(x_j - \bar{x})^2 = 8^2 + \ldots + 7^2 - 25^2/4$$
$$= 6{\cdot}75.$$

$\chi^2 = 6{\cdot}75/1{\cdot}9375 = 3{\cdot}48$, d.f. $= 3$, not significant.

Interpretation of the results for a sample as small as that used above is somewhat unrealistic. It is undertaken simply to underline the basic inferences which may be drawn from analyses of the type in question.

In test (A) the non-significant result indicates that there is insufficient evidence to conclude that agreement (or equivalently, disagreement) between the two psychiatrists varies beyond the limits of chance in their assessment of the symptomatology of the several patients in the sample. By analogy with analysis of variance, the test provides a check on possible *interaction* between psychiatrists and patients. But the analogy is not exact and should not be taken too literally. In a similar sense test B, which in our example also yields a non-significant result, satisfies us that there is no detectable evidence of 'interaction' between psychiatrists and symptoms. In view of these findings it is clear that the proportion of instances in which the psychiatrists agree, namely 25 out of 40, furnishes a reliable 'overall' index of agreement between them.

COMMENT

In situations in which either or both of the significance tests described in this paper give significant results it is unlikely that any single index of agreement which might be derived would have a clear-cut interpretation. Rather than search for such an index it would be preferable to examine the vectors of proportions in Table II and to locate those patients or symptoms concerning which there was marked disagreement. The psychiatrists might then be invited to re-examine their results and try to resolve their differences.

Finally, it is worth noting that in cases of complete agreement or complete disagreement between the psychiatrists the tests of significance given above would break down since both the numerator and the denominator in each of the expressions for χ^2 would be zero. Such cases furnish a salutary warning against the uncritical use of significance tests. In cases in which agreement or disagreement was complete or nearly so, variation in the rows and columns of Table I would be either zero or of negligible magnitude, and this would rule out the application of *any* statistical technique to the data.

Summary

When two raters interview the same sample of subjects and note the presence or absence of a number of characteristics, significance tests are provided for assessing whether agreement between the raters differs from subject to subject, or from characteristic to characteristic. When no such differences exist a simple index of agreement between raters is obtained by calculating the proportion of instances on which their decisions agree.

References

Cochran, W. G. (1950). 'The comparison of percentages in matched samples.' *Biometrika*, **37**, 256–66.

Greenhouse, S. W., and Geisser, S. (1959). 'On methods in the analysis of profile data.' *Psychometrika*, **24**, 95–112.

Maxwell, A. E., and Pilliner, A. E. G. (1968). 'Deriving coefficients of reliability and agreement for ratings.' *Brit. J. math. and stat. Psychol.*, **21**, 105–16.

A. E. Maxwell, M.A., M.Ed., Ph.D., *Professor of Psychological Statistics, Institute of Psychiatry, University of London, De Crespigny Park, S.E.5*

(*Received 4 August 1970*)