Educational and Psychological Measurement
1984, 44

THE MEASUREMENT OF CLASSIFICATION AGREEMENT: AN ADJUSTMENT TO THE RAND STATISTIC FOR CHANCE AGREEMENT

LESLIE C. MOREY
Yale University

ALAN AGRESTI
University of Florida

Investigators examining empirically derived classifications are often concerned with the replicability of an obtained classification. However, most available statistics which allow replication comparison suffer from various limitations. This paper proposes an adjustment to one of these statistics, the Rand statistic, which will allow comparison across different levels for number of clusters found within a classification. This adjustment permits the user to compare several different classifications with respect to classification agreement, while correcting for the contribution of chance to any observed agreement.

THE problem of measuring classification agreement has been encountered by many investigators in classification research. For example, in research paradigms utilizing cluster analysis techniques, replication designs are frequently used to establish the reliability of a cluster solution. In such instances, a measure is needed to compare the degree of correspondence between a solution and its replication.

The two measures most frequently used by classification researchers are the kappa statistic (Cohen, 1960) and a statistic proposed by Rand (1971). Both have been applied to describe the relative number of agreements between two solutions to the same data set. When used in this manner, the two statistics yield values which correlate fairly well (Milligan, 1981).

Copyright © 1984 Educational and Psychological Measurement

from the SAGE Social Science Collections. All Rights Reserved.

Moreover, both statistics contain some inherent limitations to their utility for classification research. The kappa statistic corrects for chance agreement, a valuable property for this type of application. However, use of kappa is limited to those instances where the number of clusters (k) in the two solutions being compared are identical. Unfortunately, in applied research there are often no clear criteria for determining k, and those criteria which have been proposed will not always yield equivalent values in a replication design. When faced with this situation, the experimenter can use the approach advocated by Rand (1971).

Rand's statistic defines two instances of classification agreement: first, when two solutions agree that two entities are to be assigned to the same cluster; and second, when the solutions agree that the two entities are to be assigned to different clusters. Suppose that in the population of interest, there are k_1 clusters in the first classification, and k_2 clusters in the second. Let P_{ij} be the probability that a randomly selected individual is classified in cluster i in the first solution and cluster j in the second solution. Rand's statistic is defined to be the probability that a randomly selected pair are classified in agreement. This probability equals

$$P_s = \Sigma\Sigma P_{ij}^2 + \Sigma\Sigma P_{ij}(1 - P_{i+} - P_{+j} + P_{ij}) \qquad (1)$$

$$= 1 - \Sigma P_{i+}^2 - \Sigma P_{+j}^2 + 2\Sigma\Sigma P_{ij}^2 \qquad (2)$$

Unlike kappa, this statistic makes no correction for chance agreement. That is, we cannot tell whether a specific value of P_s is "large" or "small", because its value when individuals are classified at random (i.e. $P_{ij} = P_{i+}P_{+j}$) is not zero, and depends on P_{i+} and P_{+j}. This artifact can be a serious disadvantage when the replicability of different classifications are being compared. For example, if there are k_1 equal sized clusters in the first solution and k_2 equal sized clusters in the second, $P_s = (1/k_1k_2) + (1 - (1/k_1))(1 - (1/k_2))$ for random classification and $P_s \uparrow 1.0$ as k_1 and k_2 increase.

Adjusted Statistic

Let P_c represent the probability of chance agreement for a pairing, as calculated using the marginal distributions. For chance agreement, $P_{ij} = P_{i+}P_{+j}$, and thus:

$$P_c = \Sigma\Sigma(P_{i+}P_{+j})^2 + \Sigma\Sigma P_{i+}P_{+j}(1 - P_{i+} - P_{+j} + P_{i+}P_{+j}) \qquad (3)$$

$$= 2\Sigma\Sigma P_{i+}^2P_{+j}^2 + 1 - \Sigma P_{i+}^2 - \Sigma P_{+j}^2. \qquad (4)$$

If we utilize the same correction used with kappa, we obtain the following relationship:

$$\text{adjusted agreement } \Omega = \frac{P_s - P_c}{1 - P_c}. \tag{5}$$

Hence, the adjusted version of Rand's statistic equals

$$\Omega = \frac{2\Sigma\Sigma P_{ij}^2 - 2(\Sigma P_{i+}^2)(\Sigma P_{+j}^2)}{\Sigma P_{i+}^2 + \Sigma P_{+j}^2 - 2(\Sigma P_{i+}^2)(\Sigma P_{+j}^2)}. \tag{6}$$

This statistic equals one for perfect agreement (in which case $P_s = 1$ also), $\Omega = 0$ for chance agreement, and $\Omega < 0$ when agreement is less than expected by chance. Values of Ω greater than zero represent the proportion of the maximum possible difference obtained between the probability of agreement and the probability of chance agreement.

Calculating Actual Agreement

Let n_{ij} equal the number of entities observed in cell cluster i in the first solution and cluster j in the second solution, and let $n = \Sigma\Sigma n_{ij}$. The total number of possible pairings is equal to the combination $\binom{n}{2}$, or $n(n - 1)/2$. We may calculate the number of observed agreements (N_s) and the number of chance agreements (N_c) as follows:

$$N_s = \Sigma\Sigma n_{ij}^2 - 1/2(\Sigma n_{i+}^2) - 1/2(\Sigma n_{+j}^2) + n(n - 1)/2, \tag{7}$$

$$N_c = (\Sigma\Sigma n_{i+}^2 n_{+j}^2)/n^2 - 1/2(\Sigma n_{i+}^2) - 1/2(\Sigma n_{+j}^2) + n(n - 1)/2. \tag{8}$$

Thus, our estimate for adjusted agreement, $\tilde{\Omega}$, may be calculated using

$$\tilde{\Omega} = \frac{N_s - N_c}{\dfrac{n(n - 1)}{2} - N_c} = \frac{\Sigma\Sigma n_{ij}^2 - (\Sigma n_{i+}^2 \Sigma n_{+j}^2)/n^2}{1/2(\Sigma n_{i+}^2) + 1/2(\Sigma n_{+j}^2) - (\Sigma n_{i+}^2 \Sigma n_{+j}^2)/n^2}. \tag{9}$$

If a random sample of people are clustered according to two procedures, with fixed numbers of clusters k_1 and k_2, then $\tilde{\Omega}$ is approximately normally distributed around Ω. We may calculate the asymptotic variance for $\tilde{\Omega}$ using the delta method (Goodman and Kruskal, 1972). Let

$$\Phi_{ij} = 4(P_{ij}(\Sigma P_{i+}{}^2 + \Sigma P_{+j}{}^2 - 2(\Sigma P_{i+}{}^2 \Sigma P_{ij}{}^2)) + P_{i+}(2(\Sigma P_{+j}{}^2)(\Sigma\Sigma P_{ij}{}^2) - (\Sigma P_{+j}{}^2)^2 - \Sigma\Sigma P_{ij}{}^2) + P_{+j}(2(\Sigma P_{i+}{}^2)(\Sigma\Sigma P_{ij}{}^2) - (\Sigma P_{i+}{}^2)^2 - \Sigma\Sigma P_{ij}{}^2)). \tag{10}$$

Then, the asymptotic variance of $\tilde{\Omega}$ is σ^2_{Ω}/n, where

$$\sigma^2_{\Omega} = \frac{(\Sigma\Sigma P_{ij}\Phi^2{}_{ij} - (\Sigma\Sigma P_{ij}\Phi_{ij})^2)}{(\Sigma P_{i+}{}^2 + \Sigma P_{+j}{}^2 - 2(\Sigma P_{i+}{}^2 \Sigma P_{+j}{}^2))^4}. \tag{11}$$

Let us return to the original Rand (1971) article for an example. The data which he presented may be expressed in the contingency table found in Table 1. The number of observed agreements equals

$$N_s = 10 - 9 - 7 + 15 = 9. \tag{12}$$

Rand's statistic is equal to this number divided by the number of possible pairings, which in this case equals 6(5)/2 = 15. Thus, Rand obtained a value of .60 for his example. However, this figure does not account for chance agreement. The expected number of agreements due to chance equals

$$N_c = (18 \times 14)/36 - 9 - 7 + 15 = 6. \tag{13}$$

Thus, by chance alone we would expect to obtain a Rand value of 6/15 = .40. Incorporating the adjustment for chance, we obtain

$$\tilde{\Omega} = (9-6)/(15-6) = .333, \tag{14}$$

which represents the corrected agreement between the classifica-

TABLE 1
Classification Comparison Example from Rand (1971)

Classification One		Classification Two	
Entity	Cluster	Entity	Cluster
A	1	A	1
B	1	B	1
C	1	C	2
D	2	D	2
E	2	E	2
F	2	F	3

		Cluster for Classification Two			
		1	2	3	Total
Cluster For	1	2	1	0	3
Classification One	2	0	2	1	3
	Total	2	3	1	6

tions. In other words, the difference between the number of observed agreements and chance agreements is 33.3% of the maximum possible difference. A calculation of the variance yields:

$$\hat{\sigma}^2/n = .4444/6 = .074. \tag{15}$$

The comparison of classifications of the same data set is an important problem for cluster analysis users. Rand's statistic has a basic appeal to cluster analysis researchers, since it was derived specifically to address the problem of evaluating cluster analysis methods. Hence, this measure has been widely used in Monte Carlo studies of clustering efficiency. In such studies, experimenters may hold constant the value for k, and as such this will not be a complicating factor. However, in an applied study where k may be expected to be quite variable, the Rand statistic loses its value as a comparative metric, since it is highly dependent upon chance agreement as a function of k. The aim of this paper has been to derive an adjustment by which the Rand statistic may be used as an informative comparison measure.

REFERENCES

Cohen, J. (1960). A coefficient of agreement for nominal scales. EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, 20, 37–46.

Goodman, L. A. and Kruskal, W. H. (1972). Measures of association for cross-classifications, IV: Simplification of asymptotic variances. *Journal of the American Statistical Association*, 67, 415–421.

Milligan, G. W. (1981). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 46, 187–200.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66, 846–850.