Behavior Research Methods, Instruments, & Computers
1994, 26 (1), 60-61

PROGRAM ABSTRACTS/ALGORITHMS

Computing Cohen's kappa coefficients using SPSS MATRIX

CLAUDE A. M. VALIQUETTE, ALAIN D. LESAGE, and MIREILLE CYR
University of Montreal, Montreal, Quebec, Canada

and

JEAN TOUPIN
University of Sherbrooke, Sherbrooke, Quebec, Canada

This short paper proposes a general computing strategy to compute Kappa coefficients using the SPSS MATRIX routine. The method is based on the following rationale. If the contingency table is considered as a square matrix, then the observed proportions of agreement lie in the main diagonal's cells, and their sum equals the trace of the matrix, whereas the proportions of agreement expected by chance are the joint product of marginals. The generalization to weighted kappa, which requires an additional square matrix of disagreement weights, both matrices having the same order, becomes possible by the use of the Hadamard product—that is, the elementwise direct product of two matrices.

In reliability assessment, interobserver agreement deserves special attention because it is a *prima facie* requirement. It concerns the intrinsic quality of the elementary datum: if two judges cannot agree, all other reliability considerations, such as internal consistency or temporal stability, become totally irrelevant. This propedeutic attitude is typical of a large class of psychological data relying on human judgment (e.g., behavioral observations, clinical diagnosis, and categorical partitioning), either for laboratory, field studies, or structured interviews. Thus, assessment of this aspect of reliability is crucial and the choice of an appropriate coefficient is of prime importance. Over the years, kappa coefficients have gained wide acceptance among behavioral and social scientists and now appear in standard treatises on data analysis (e.g., Siegel & Castellan, 1988).

The kappa coefficient of agreement was proposed by Cohen (1960, 1968); its use was then limited to the two-judge case only. Several interesting features explain this popularity: the calculation of kappa contains an explicit correction for chance agreement; the exact large sample behavior of this coefficient is known (Fleiss, Nee, & Landis, 1979). This statistic has been investigated from widely different points of view (e.g., the linear assignment model; Hubert, 1987). Moreover, kappa can be viewed within the general framework of analysis of variance as a special case of intraclass correlation (Collis, 1985; Conger & Ward, 1984; Fleiss & Cohen, 1973; Rae, 1984). Several variants have been developed to deal with different data-gathering designs. Two situations need to be distinguished on the basis of whether the team of raters is constant or variable over all subjects. The original coefficient derived by Cohen (1960, 1968) assumes that the same pair of judges evaluates all subjects, as in a conventional factorial one-way analysis of variance. This raw coefficient is applicable with only two judges, with provision for scaled disagreement or partial credit, leading to a weighted kappa (Cohen, 1968). On the other hand, Fleiss (1971) generalized raw kappa to the case where a different (but equal) team of raters evaluates subjects, allowing the use of more than two judges. This case bears some similarity to a nested one-way analysis of variance. Guidelines have been suggested to evaluate the clinical significance of the kappa coefficient of agreement (more than .80 = almost perfect, between .60 and .80 = substantial, etc.; see Landis & Koch, 1977).

Table 1
SPSS MATRIX Program Used to get Raw Kappa on Cohen's Data (1968, p. 214)

```
DATA LIST/V1 to V3 1-9 (2)
*This is the contingency table showing proportion of judgments by two
judges into three categories*
BEGIN DATA
 44 07 09
 05 20 05
 01 03 06
END DATA
MATRIX
GET X /VAR = V1 TO V3
COMPUTE PO = TRACE (X)
COMPUTE C = RSUM (X)
COMPUTE R = CSUM (X)
COMPUTE PE = R * C
COMPUTE NUM = PO - PE
COMPUTE DEN = 1 - PE
COMPUTE KAPPA = NUM/DEN
PRINT KAPPA
END MATRIX
FIN
```

Note—Kappa = .492. The contingency table is from "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit" by Jacob Cohen, 1968, *Psychological Bulletin*, **70**, p. 214. Copyright 1968 by the American Psychological Association. Reprinted by permission.

This research was supported by a grant from the Quebec Health Research Fund (Canada). We thank Christopher Earls, Stéphane Sabourin, and Marc Valiquette for critical and enlightening comments on earlier drafts. Requests for reprints should be sent to Mireille Cyr, Department of Psychology, Université de Montréal, C.P. 6128, Succ. A, Montréal, Québec, Canada H3C 3J7.

Table 2
SPSS MATRIX Program Used to get Weighted Kappa on Cohen's Data (1968, p. 214)

```
DATA LIST / V1 TO V3 1-9 (2) W1 TO W3 11-16
*The last three columns contain the weights of the disagreement matrix*
BEGIN DATA
44 07 09 0 1 3
05 20 05 1 0 6
01 03 06 3 6 0
END DATA
MATRIX
GET X /VAR = V1 TO V3
GET W /VAR = W1 to W3
*The symbol &* denotes the Hadamard product*
COMPUTE A = X & * W
COMPUTE POM = MSUM (A)
COMPUTE C = RSUM (X)
COMPUTE R = CSUM (X)
COMPUTE P = C * R
COMPUTE PI = T (P)
COMPUTE B = PI & * W
COMPUTE PEM = MSUM (B)
COMPUTE Q = POM / PEM
COMPUTE WKAPPA = 1 - Q
PRINT WKAPPA
END MATRIX
FIN
```

Note—Kappa = .348. The matrix of disagreement weights is from "Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit" by Jacob Cohen, 1968, *Psychological Bulletin*, **70**, p. 214. Copyright 1968 by the American Psychological Association. Reprinted by permission.

At least a dozen different programs are currently available to compute kappa. Raw kappa (Antonak, 1977; Berk & Campbell, 1976; Burns & Cavallaro, 1982; Watkins & Larimer, 1980; Wixon, 1979), weighted kappa with fixed euclidean linear weighting for (dis)agreement (Chan, 1987) or modified linear weights (Bloor, 1983), and ordinary weighted kappa (Cicchetti, Showalter, & McCarthy, 1990) exist on main frame systems and microcomputers. Finally, the last version of SPSS, Release 4 (Norusis, 1990a) provides kappa coefficient by means of the CROSSTABS procedure. This latter procedure, however, presents some inconveniences. First, only the raw (unweighted) version of the kappa is available. Also, a pretabulated data table cannot be directly submitted to the program as a $c \times c$ matrix (c = number of categories) but as c^2 lines, each line containing codes for rows and columns and, finally, the cell frequency. For example, entering crosstabulated data with two judges and five categories requires 25 lines, each containing codes for row (judges), column (categories), and the cell frequency.

The goal of the present paper is to suggest a different way to compute raw and weighted kappa for any disagreement weights, symmetric or asymmetric. The underlying idea is to create two matrices—one for the contingency table, and the other one for the disagreement weights by means of matrix algebra. To do so, we shall use the MATRIX routine, available in SPSS, Release 4 (Norusis, 1990b).

The SPSS MATRIX procedure is a convenient tool, enabling researchers to write programs for a wide range of statistical methods, especially multivariate ones. In Tables 1 and 2, we use Cohen's (1968, p. 214) data to illustrate how to compute raw and weighted kappa by using SPSS MATRIX.

REFERENCES

Antonak, R. F. (1977). A computer program to compute measures of response agreement for nominal scale data obtained from two judges. *Behavior Research Methods & Instrumentation*, **9**, 553.

Berk, R. A., & Campbell, K. L. (1976). A FORTRAN program for Cohen's kappa coefficient of observer agreement. *Behavior Research Methods & Instrumentation*, **8**, 396.

Bloor, R. N. (1983). A computer program to determine interrater reliability for dichotomous-ordinal rating scales. *Behavior Research Methods & Instrumentation*, **15**, 615.

Burns, E., & Cavallaro, C. (1982). A computer program to determine interobserver reliability statistics. *Behavior Research Methods & Instrumentation*, **14**, 42.

Chan, T. S. C. (1987). A DBase III program that performs significance testing for the kappa coefficient. *Behavior Research Methods, Instruments, & Computers*, **19**, 53-54.

Cicchetti, D. V., Showalter, D., & McCarthy, P. (1990). A computer program for calculating subject-by-subject kappa or weighted kappa coefficients. *Educational & Psychological Measurement*, **50**, 153-158.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement*, **20**, 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, **70**, 213-220.

Collis, G. M. (1985). Kappa, measures of marginal symmetry and intraclass correlations. *Educational & Psychological Measurement*, **45**, 55-62.

Conger, A. J., & Ward, D. G. (1984). Agreement among 2×2 agreement indices. *Educational & Psychological Measurement*, **44**, 301-313.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, **76**, 378-382.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational & Psychological Measurement*, **33**, 613-619.

Fleiss, J. L., Nee, J. C., & Landis, J. R. (1979). Large sample variance of kappa in the case of different sets of raters. *Psychological Bulletin*, **86**, 974-977.

Hubert, L. J. (1987). *Assignment methods in combinatorial data analysis*. New York: Marcel Dekker.

Landis, J., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159-174.

Norusis, M. J. (1990a). *SPSS base system user's guide*. Chicago: SPSS Inc.

Norusis, M. J. (1990b). *SPSS advanced statistics user's guide*. Chicago: SPSS Inc.

Rae, G. (1984). On measuring agreement among several judges on the presence or absence of a trait. *Educational & Psychological Measurement*, **44**, 247-253.

Siegel, S., & Castellan, N. J., Jr. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.) New York: McGraw Hill.

Watkins, M. W., & Larimer, L. D. (1980). Interrater agreement statistics with the microcomputer. *Behavior Research Methods & Instrumentation*, **12**, 466.

Wixon, D. R. (1979). Cohen's kappa coefficient of observer agreement: A BASIC program for minicomputers. *Behavior Research Methods & Instrumentation*, **11**, 602.

(Manuscript received December 22, 1992; revision accepted for publication June 15, 1993.)