QÜESTIIÓ, vol. 23, 3, p. 561-571, 1999

GENERALIZATION OF THE KAPPA COEFFICIENT FOR ORDINAL CATEGORICAL DATA, MULTIPLE OBSERVERS AND INCOMPLETE DESIGNS

V. ABRAIRA*
A. PÉREZ DE VARGAS**
Hospital Ramón y Cajal

This paper presents a generalization of the kappa coefficient for multiple observers and incomplete designs. This generalization involves ordinal categorical data and includes weights which permit pondering the severity of disagreement. A generalization for incomplete designs of the kappa coefficient based on explicit definitions of agreement is also proposed. Both generalizations are illustrated with data from a medical diagnosis pilot study.

Keywords: Agreement, kappa, incomplete designs

AMS Classification (MSC 2000): 62P10, Q2B15

* Unit of Clinical Biostatistics. Hospital 'Ramón y Cajal'. Crta. Colmenar km 9,1, planta -2 D. 28034 Madrid (Spain). Tel. +34 91 3368103. Fax. +34 91 3369016. E-mail: victor.abraira@hrc.es.
** Department of Biomathematics. University Complutense of Madrid. Facultad de Biología. Ciudad Universitaria. 28040 Madrid (Spain). Tel. +34 91 3945078. Fax. +34 91 3945051.
E-mail: alpedeva@eucmax.sim.ucm.es.
Address for correspondence: Unidad de Bioestadística Clínica. Hospital 'Ramón y Cajal'. Crta. Colmenar km 9,1, planta -2 D. 28034 Madrid (Spain).

– Received May 1998.
– Accepted July 1999.

1. INTRODUCTION

An important feature of any measurement or classification device is the reproducibility or reliability, which in classification is also referred to as concordance or agreement. From the seminal paper by Cohen [1], introducing the kappa coefficient (κ) to assess concordance between two observers using binary classifications, a great effort has been made to extend this index to more general conditions. Thus, Cohen [2] generalized kappa to weighted kappa in order to encompass ordinal variables incorporating an a priori assignment of weights to each of the cells of the $k \times k$ table of joint nominal scale; Landis and Koch [3] proposed an approach by expressing the quantities which reflect the extent to which the observers agree among themselves as functions of observed proportions obtained from underlying multidimensional contingency tables, using the GSK method [4]; Davies and Fleiss [5] proposed a generalization for multiple observers by the average of pairwise agreement. Although some limitations of kappa index are known such as that its value depends on the balance and symmetry of marginal totals of the table [6, 7] and some alternative methods of evaluating agreement among observers have been proposed [8, 9, 10, 11], the kappa index is still a very frequently used statistic in clinical epidemiology literature (e.g. Elmore *et al.* [12], Jelles *et al.* [13], Pérez *et al.* [14].

This paper generalizes Schouten's [15] and Gross's [16] proposal for multiple observers and incomplete design, as to encompass ordinal variables with the inclusion of weights to enable pondering the severity of disagreement among different categories. Another generalization for incomplete designs is also proposed, based on the explicit definitions of agreement by Landis and Koch [17]. This generalization is approached in a simpler way than the very general method of Koch *et al.* [18].

Both generalizations were motivated by the study shown in section 5. We tried to assess the concordance among several physicians evaluating the current health status of people affected by the Toxic Oil Syndrome. In this study there were some ordinal multicategorical variables as «peripheral neuropathy» and «sclerodermiform changes of the skin», and in order to avoid seeing each patient too many times at short intervals times for the same sign, an incomplete design should be used.

2. GENERALIZATION OF κ INDEX

The κ index, proposed by Cohen, is defined as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{1}$$

where P_o is the proportion of observed agreement and P_e the proportion of expected agreement in the hypothesis of independence between observers. When there are only two observers, the definition of agreement is obvious. However, when there are more than two observers, agreement can be defined in diverse ways [19]. In this paper, we will restrict ourselves to pairwise agreement [5] (section 2.1) and majority agreement [17] (section 2.2).

2.1. Pairwise agreement

A set of N subjects is classified in K ordinal categories by a set G of $J > 1$ observers, with an incomplete design, that is to say, each subject i is only classified by a subset G_i of $J_i \leqslant J$ observers. Let X_{ik} be the number of observers classifying the ith subject into the kth category and w_{lm} the weight corresponding to the agreement-disagreement between categories l and m, obviously with the conditions:

$$w_{mm} = 1; \quad 0 \leqslant w_{lm} < 1 \quad \forall\, l \neq m; \quad w_{lm} = w_{ml}$$

For the ith subject, the number of weighted agreements is:

$$NA_i = \frac{1}{2}\sum_{k=1}^{K} w_{kk} X_{ik}(X_{ik}-1) + \sum_{l=1}^{K}\sum_{m>l}^{K} w_{lm} X_{il} X_{im}$$

and as the number of possible pairs of classifications for each subject i is $\dfrac{J_i(J_i-1)}{2}$, the proportion of weighted agreements for the ith subject is:

$$\frac{\sum_{k=1}^{K} w_{kk} X_{ik}(X_{ik}-1) + 2\sum_{l=1}^{K}\sum_{m>1}^{K} w_{lm} X_{il} X_{im}}{J_i(J_i-1)} = \frac{\sum_{m=1}^{K}\sum_{l=1}^{K} w_{lm} X_{il} X_{im} - J_i}{J_i(J_i-1)} \tag{2}$$

because:

$$\sum_{k=1}^{K} w_{kk} X_{ik} = \sum_{k=1}^{K} X_{ik} = J_i$$

Then, the average proportion of observed agreements for all subjects is the sum of (2) for all subjects divided by the number of subject, so:

$$P_o = \frac{1}{N_c}\sum_{i=1}^{N_c} \frac{\sum_{m=1}^{K}\sum_{l=1}^{K} w_{lm} X_{il} X_{im} - J_i}{J_i(J_i-1)} \tag{3}$$

where N_c is the number of subjects classified by more than one observer.

Let $P_j(k)$ $(j = 1, \ldots, J; k = 1, \ldots, K)$ represent the proportion of times which the jth observer classifies into the kth category. Then, the proportion of expected agreements for the ith subject in the hypothesis of independence between the pair l and m of observers is:

$$\sum_{u=1}^{K} \sum_{k=1}^{K} w_{uk} P_l(u) P_m(k)$$

We note that with incomplete designs the expected agreement is different for each subject because each one is classified by a different subset of observers. Then, the average expected proportion of pairwise agreement in the hypothesis of independence for the ith subject is:

$$\frac{2}{J_i(J_i-1)} \sum_{l=1}^{J_i} \sum_{m>l}^{J_i} \sum_{u=l}^{K} \sum_{k=l}^{K} w_{uk} P_l(u) P_m(k)$$

where, obviously, the sums for m and l are restricted to set G_i of observers which have classified the ith subject. Then, the average expected proportion of pairwise agreement for all the subjects is:

$$P_e = \frac{1}{N_c} \sum_{i=1}^{N_c} \frac{2}{J_i(J_i-1)} \sum_{l=1}^{J_i} \sum_{m>l}^{J_i} \sum_{u=l}^{K} \sum_{k=l}^{K} w_{uk} P_l(u) P_m(k) \tag{4}$$

The κ index is calculated with (1), using (3) and (4). If weights are not included, that is to say $w_{mm} = 1$; $w_{lm} = 0 \ \forall l \neq m$, the expressions (3) and (4) are reduced to expressions given by Schouten [15] and Gross [16].

2.2. Majority agreement

In a study with multiple observers, agreement among observers can be also defined as majority or consensus: there is agreement at an observation if a majority of observers agree; e.g., if there are seven observers, it is possible define agreement when at least five of them agree. Obviously, it is advisable [17] to have a clear majority, e.g., 7-0, 6-1 splits, rather than «tie-breaking» majorities, e.g., 4-3 splits. It is possible to define the following indicator variables z_p, one for each agreement definition [17]:

$$z_{0i} = \begin{cases} 1 & \text{if all observers agree, for the subject } i \\ 0 & \text{otherwise} \end{cases}$$

$$z_{1i} = \begin{cases} 1 & \text{if at least } J-1 \text{ observers agree, for the subject } i \\ 0 & \text{otherwise} \end{cases}$$

$$z_{pi} = \begin{cases} 1 & \text{if at least } J-p \text{ observers agree, for the subject } i \\ 0 & \text{otherwise} \end{cases}$$

for calculating the proportion of observed agreement by means of them as follows:

$$P_{o(p)} = \frac{\sum_{i=1}^{N_c} z_{pi}}{N_c} \tag{5}$$

where N_c is the number of subjects observed whom it is able to observe the defined agreement; that is to say, the number of subjects observed by, at least, $J-p$ observers. In the hypothesis of independence, the proportion of expected agreement for each subject is:

$$\sum_{V \in V_{K,J_i,p}} P_1(V) \cdots P_{J_i}(V)$$

where $V_{K,J_i,p}$ represents the set of permutations with repetition of K elements taken J_i at a time, with at least $J_i - p$ of them remaining equals and $P_j(k)$ $(j = 1, \ldots, J; k = 1, \ldots, K)$, as in section 2.1, the proportion of times which the jth observer classifies into the kth category. The average proportion of expected agreement is its sum for all subjects divided by the number of subjects observed in whom it is possible to observe the defined agreement; that is to say

$$P_{e(p)} = \frac{1}{N_c} \sum_{i=1}^{N_c} \sum_{V \in V_{K,J_i,p}} P_l(V) \cdots P_{J_i}(V) \tag{6}$$

The κ index is calculated with (1), using (5) and (6). The complete design can be considered as a particular case in which $J_i = J \; \forall \; i$ and so $N_c = N$, in this case, (5) and (6) are reduced to formulas given by Landis and Koch [17].

3. INFERENCES ABOUT κ

The kappa statistic is an estimator of the κ parameter for subjects and observers population. In previous formulas, κ is implicitly defined as a function of the probabilities that

are estimated by the proportions P_o and $P_j(k)$. To make inferences about κ we need to compute its standard error. A very general method for this is the jackknife technique [20]. Parr and Tolley [21] have shown that for all real functions (such as kappa) of multinomial proportions, with continuous first and second partial derivatives, in large samples, the jackknife estimator approximately follows a normal distribution and its variance is estimated by the variance of pseudo-value.

Then, a confidence interval for κ is:

$$J(\kappa) - t_{\alpha/2,(N-1)} \frac{S_j}{\sqrt{N}} \leqslant \kappa \leqslant J(\kappa) + t_{\alpha/2,(N-1)} \frac{S_j}{\sqrt{N}}$$

where $J(\kappa)$ is the jackknife estimate and S_j the pseudo-value standard deviation.

4. SOFTWARE

A computer program was written in FORTRAN 77 and runs in PC's under DOS. The program calculates proposed kappa indexes, their jackknife estimates and their standard error, also estimated by same method. It is included in the statistical package PRESTA [22] (PRESTA is a statistical package in Spanish, available on the Internet URL http://www.hrc.es/bioest.html).

5. EXAMPLE

It is a pilot study previous to another study which was made to assess the current health status of people affected by the disease which came to be known as toxic oil syndrome (TOS). The TOS was developed in people who consumed adulterated rapeseed oil sold as cooking oil and it affected more than 20.000 people. A TOS description can be seen in Nadal and Tarkowski [23]. The study to assess the current health status [24], was conducted with all of the 4.015 affected registered in the seven TOS Follow-up Centers of Madrid. Clinical histories and patients' physical examinations were used as data sources. Physical examinations were made by nine different physicians from the Follow-up Centers. The pilot study, shown here, was conducted to assess reliability of the variables potentially most affected by observer subjectivity. The categorical variables included in the study were: peripheral neuropathy, classified in three levels: «no neuropathy», «doubtful neuropathy» and «certain neuropathy»; severity of sclerodermiform changes of the skin, classified in four levels: «no sclerodermia», «fair sclerodermia», «moderate sclerodermia» and «atrophic skin»; and joint contractures classified as «yes» and «no».

5.1. Study Design

Patients: A non random sample of 10 patients affected by TOS chosen to cover all range of clinical degrees of the disease.

Observers: A random sample of 6 physicians chosen from the nine whom later did the current health status study.

Procedure: Before the study, the six physicians participated in a 5-hour workshop, where they were trained in the protocol of variables collection. The workshop included a physical examination of several TOS patients, different from those who would later participate in the study. In order to avoid each patient's being seen too many times for the same sign at short time intervals, a balanced incomplete block design (Fleiss [25]) was selected. In this design, each patient is examined by 3 physicians, each physician examines 5 patients, and all possible pairs of physicians examine the same 2 patients. The examination designation scheme is laid out in Table 1. The efficiency factor [25] for estimating the coefficient of reliability of this design is 0.8, which seem like a reasonable compromise. The order of examinations in each patient was randomly determined using a permutations table. Patients were informed in writing of the purposes of the study and gave their written consent to participate in the study. In order to guarantee the patients' confidentiality no identification data was saved in computer files.

5.2. Results

Proportions of observed and expected agreement, kappa index, its jackknife estimate and its standard error for all variables are shown in tables 2, 3 and 4. Weighted kappa not was used in joint contractures variable because it has only two categories, squared error weights were used for the other variables [26]. The indexes found indicate a fair to moderate agreement according the benchmark of Landis and Koch [3], which obliged us to repeat observer training before conducting the current health status study. Although the sample size is small, big differences between sample estimation and jackknife estimation of kappa are not observed; which leads us to have confidence in the jackknife estimation of standard error. The differences between pairwise and weighted pairwise indexes, in tables 3 and 4, illustrate that the greatest disagreement occurs between contiguous categories; the differences between pairwise and majority kappa suggest that at least one observer classifies differently from the others. Marginal frequencies of peripheral neuropathy are shown in table 5, where it is seen that «physician 2» is clearly different, as he classified a proportion of 0.6 into the «doubtful» category and 0 in «certain». If analysis is repeated without this observer, all indexes increase considerably (0.7439 with SE=0.1727 for pairwise agreement; 0.8888 with SE=0.0832 for weighted pairwise agreement and 0.7379 with SE=0.2844 for majority agreement).

6. CONCLUSIONS

In the assessment of reliability among multiple observers, unbalanced designs often appear, either by design as in the presented example, or due to missing data. In this paper, we have proposed a simple modification of previous kappa indexes to include unbalanced designs in weighted kappa for ordinal variables and kappa for majority. We have also illustrated their use with real data and, in the example, we have shown how differences among several indexes (pairwise, weighted pairwise and majority) permit identification of the sources of disagreement, which is the main aim of this kind of studies.

Table 1. Balanced incomplete block design used in the TOS study

Patient	*Physi. 1*	*Physi. 2*	*Physi. 3*	*Physi. 4*	*Physi. 5*	*Physi. 6*
1	x			x		x
2			x	x	x	
3			x	x		x
4	x		x		x	
5	x				x	x
6	x	x	x			
7		x	x			x
8		x			x	x
9	x	x		x		
10		x		x	x	

Table 2. Joint contractures (2 categories)

Agreement	P_o	P_e	κ	$J(\kappa)$	$SE(\kappa)$
pairwise	0.6667	0.4827	0.3557	0.3827	0.2267
majority of 3	0.5000	0,2240	0.3557	0.3827	0.2267

P_o: proportion of observed agreement
P_e: proportion of expected agreement
κ: kappa index
$J(\kappa)$: jackknife estimate of kappa index
$SE(\kappa)$: jackknife standard error

Table 3. Peripheral neuropathy (3 categories)

Agreement	P_o	P_e	κ	$J(\kappa)$	$SE(\kappa)$
pairwise	0.6667	0.3387	0.4960	0.4995	0.1387
pairwise we*	0.8667	0.6607	0.6071	0.6095	0.1738
majority of 3	0.5000	0.1176	0.4334	0.4373	0.1622

$*$ Weighted kappa with quadratic weights
P_o: proportion of observed agreement
P_e: proportion of expected agreement
κ: kappa index
$J(\kappa)$: jackknife estimate of kappa index
$SE(\kappa)$: jackknife standard error

Table 4. Sclerodermiform changes of the skin (4 categories)

Agreement	P_o	P_e	κ	$J(\kappa)$	$SE(\kappa)$
pairwise	0.6667	0.2507	0.5552	0.5757	0.1343
pairwise we*	0.9407	0.6868	0.8108	0.8401	0.1062
majority of 3	0.5000	0.0656	0.4649	0.4825	0.1679

$*$ Weighted kappa with quadratic weights
P_o: proportion of observed agreement
P_e: proportion of expected agreement
κ: kappa index
$J(\kappa)$: jackknife estimate of kappa index
$SE(\kappa)$: jackknife standard error

Table 5. Marginal frequencies of peripheral neuropathy

	No	*Doubtful*	*Certain*
Physician 1	0.400	0.200	0.400
Physician 2	0.400	0.600	0.000
Physician 3	0.600	0.200	0.200
Physician 4	0.400	0.200	0.400
Physician 5	0.400	0.200	0.400
Physician 6	0.400	0.400	0.200
Mean	0.433	0.300	0.267

ACKNOWLEDGEMENT

This work was supported in part by FIS grant 96/0421. The authors would like to thank Kathleen Seley for her help in correcting this manuscript.

REFERENCES

[1] Cohen J. (1960). «A coefficient of agreement for nominal scales», *Educational and Psychological Measurement*, 20, 37-46.

[2] Cohen J. (1968). «Weighted kappa: Nominal scale agreement with provision for scaled disagreement or parcial credit», *Psychological Bulletin*, 70, 213-220.

[3] Landis J.R. and Koch G.G. (1977). «The measurement of observer agreement for categorical data», *Biometrics*, 33, 159-174.

[4] Grizzle J.E., Starmer C.F. and Koch G.G. (1969). «Analysis of categorical data by linear models», *Biometrics*, 25, 489-504.

[5] Davies M. and Fleiss J.L. (1982). «Measuring agreement for multinomial data», *Biometrics*, 38, 1047-1051.

[6] Feinstein A.R. and Cicchetti D.V. (1990). «High agreement but low kappa: I. The problems of two paradoxes», *Journal of Clinical Epidemiology*, 43, 543-549.

[7] Guggenmoos-Holzmann I. (1993). «How reliable are chance-corrected measures of agreement?», *Statistics in Medicine*, 12, 2191-2205.

[8] Cicchetti D.V. and Feinstein A.R. (1990). «High agreement but low kappa: II. Resolving the paradoxes», *Journal of Clinical Epidemiology*, 43, 551-558.

[9] Rosner B. (1982). «Statistical methods in ophthalmology: An adjustment for the intraclass correlation between eyes», *Biometrics*, 38, 105-114.

[10] Donner A. and Donald A. (1988). «The statistical analysis of multiple binary measurements», *Journal of Clinical Epidemiology*, 41, 899-905.

[11] Graham P. and Jackson R. (1993). «The analysis of ordinal agreement data: beyond weighted kappa», *Journal of Clinical Epidemiology*, 46, 1055-1062.

[12] Elmore J.G., Wells C.K., Lee C.H., Howard D.H. and Feinstein A.R. (1994). «Variability in radiologist's interpretations of mammograms», *New England Journal of Medicine*, 331, 1493-1499.

[13] Jelles F., Van Bennekom C.A.M., Lankhorst G.F., Sibbel C.J.P. and Bouter L.M. (1995). «Inter- and intra-rater agreement of the rehabilitation activities profile», *Journal of Clinical Epidemiology*, 48, 407-416.

[14] Pérez B., Abraira V., Núñez M., Boixeda P., Pérez Corral F. and Ledo A. (1997). «Evaluation of agreement among dermatologists in the assessment of the color of

Port Wine Stains and their clearance after treatment with the Flaslamp-Pumped Dye Laser», *Dermatology*, 194, 127-130.

[15] Schouten H.J.A. (1986). «Nominal scale agreement among observers», *Psychometrika*, 51, 453-466.

[16] Gross S.T. (1986). «The kappa coefficient of agreement for multiple observers when the number of subjects is small», *Biometrics*, 42, 883-893.

[17] Landis J.R. and Koch G.G. (1977). «An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers», *Biometrics*, 33, 363-374.

[18] Koch G.G., Imrey P.B. and Reinfurt D.W. (1972). «Linear model analysis of categorical data with incomplete response vectors», *Biometrics*, 28, 663-692.

[19] Abraira V. (1997). *Precisión de las clasificaciones clínicas*. Doctoral Thesis. Universidad Computense de Madrid.

[20] Efron B. and Gong G. (1983). «A leisurely look at the bootstrap, the jackknife and cross-validation», *The American Statistician*, 37, 36-48.

[21] Parr W.C. and Tolley H.D. (1982). «Jackknifing in categorical data analysis», *The Australian Journal of Statistics*, 24, 67-79.

[22] Abraira V. and Zaplana J. (1984). «PRESTA, un paquete de procesamientos estadísticos», *Proceeding de la Conferencia Iberoamericana de Bioingeniería*. 100, Gijón.

[23] Nadal J. and Tarkowski S. (1992). «Toxic oil syndrome. Current knowledge and future perspectives. World Health Organization». *Regional Publications European Series*. Nº. 42, Copenhagen.

[24] Gómez de la Cámara A., Posada M., Abaitua I., Barainca M.T., Abraira V., Diez M. and Terracini B. (1998). «Health status measurements in Toxic Oil Syndrome», *Journal of Clinical Epidemiology*, 51, 867-873.

[25] Fleiss J.L. (1986). *The design and analysis of clinical experiments*. John Wiley & Sons, New York.

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