WEIGHTED KAPPA FOR THREE RATERS

Y. S. Lin, The Pillsbury Company

I Introduction

An important problem when collecting data involving several raters' rating on objects is whether or not the raters display a reasonable level of agreement. For continuous response data, intraclass correlation coefficient (e.g., Haggard, 1958) can be used as an appropriate agreement measure. However, with non-continuous response data (i.e., nominal or ordinal data), the standard intraclass correlation coefficient is not an appropriate measure of agreement. Goodman and Kruskal (1954) suggested that for the situation when each of the r raters independently assigns N responses (one to each of the N objects) among I categories, a measure of agreement, adjusted for chance, among r raters is needed.

Many coefficients of relative agreement measure have been proposed within the last two decades. Cohen (1960) suggested an agreement coefficient called Kappa (κ) to measure the relative agreement level for two raters. Asymptotic distribution of the maximum likelihood estimator of Kappa was given by Fleiss, Cohen, and Everitt (1969) and also by Bishop, Fienberg, and Holland (1975). Extension of Kappa coefficient to more than two raters situation have been considered by Fleiss (1971), Light (1971), Lin (1975), and others. In the works of Fleiss and Light the extension to more than two raters have been restricted to a special case, namely, an agreement measure based on average pairwise agreements. The extension by Lin is also for a special situation while only the raw complete agreement (e.g., for r = 3, all three raters agree) score is used to define the relative Kappa coefficient. While all these extensions are useful for many practical situations, there exists a need for a more general Kappa-type agreement coefficient for more than two raters.

In this paper we define a new weighted Kappa-type coefficient using both the complete and some high partial agreement scores with weights. For example, in a three raters case, we could define a weighted Kappa based on at least two out of three raters agreeing, with some weight W ($0 \le W \le 1$) attached to the partial agreement where exactly two of the three raters agree. While the general weighted Kappa for $r \ge 3$ can be easily obtained, only three raters situations will be illustrated. The asymptotic distribution for the maximum likelihood estimator of weighted Kappa is obtained. Some potential values of weights are also suggested.

II Weighted Kappa Based on at Least Two of Three Raters Agree

For a three raters case, let P_{ijk} be the probability that an object is assigned to cell (i,j,k) for i,j,k = 1,2,3,---,I, in the IxIxI contingency table and the weight W be a pre-assigned non-negative constant ($0 \le W \le 1$). Then a weighted Kappa-type agreement coefficient based on at least two of the three raters agree can be defined as:

$$\kappa_{3-2} = (A-B)/(1-B) \qquad (1)$$

where $A = \sum_i P_{iii} + \sum_{i \neq j}\sum W(P_{iij} + P_{iji} + P_{jii})$,

$$B = \sum_i P_{i..}P_{.i.}P_{..i} + \sum_{i \neq j}\sum W(P_{i..}P_{.i.}P_{..j} + P_{i..}P_{.j.}P_{..i} + P_{j..}P_{.i.}P_{..i}).$$

III Results

Let $\hat{\kappa}_{3-2}$ be the maximum likelihood estimator of κ_{3-2} with a fixed N, the total number of objects being rated. The following results can be shown:

Theorem

Assuming nultinomial sample model, the asymptotic distribution of $\sqrt{N}(\hat{\kappa}_{3-2} - \kappa_{3-2})$ normal with mean zero and variance

$$\sigma^2_{\hat{\kappa}_{3-2}} = V/(1-B)^4 \qquad (2)$$

where $V = A(1-A)(1-B^2) + (1-A)^2(\theta_4 - \theta_5^2) + 2(1-A)(1-B)(\theta_1\theta_5 - \theta_3)$

where $\theta_3 =$

$$(1-W)\sum_i P_{iii}(P_{.i.}P_{..i} + P_{i..}P_{..i} + P_{i..}P_{.i.}) - W\sum_{i \neq j \neq k}\sum\sum P_{ijk}(P_{.i.}P_{..i} + P_{j..}P_{..j} + P_{k..}P_{.k.})$$

$\theta_4 =$

$$\sum_i\sum_j\sum_k P_{ijk}\{(1-W)(P_{.i.}P_{..i} + P_{j..}P_{..j} + P_{k..}P_{.k.}) - W(1-\delta_{ijk})(P_{.i.}P_{..i} + P_{j..}P_{..j} + P_{k..}P_{.k.})\}$$

$$\Theta_5 = (1-W)\Sigma\Sigma\Sigma P_{ijk}(P_{.i.}P_{..i}+P_{j..}P_{..j}+P_{k..}P_{.k.})-W\Sigma\Sigma\Sigma_{i\neq j\neq k} P_{ijk}(P_{.i.}P_{..i}+P_{j..}P_{..j}+P_{k..}P_{.k.})$$

and $\delta_{ijk} = \begin{cases} 1 & \text{if } i=j=k \\ 0 & \text{if otherwise.} \end{cases}$

With great care in calculations, the proof of the above theorem can be obtained in a straightforward fashion following a result from Goodman and Kruskal (1972). The asymptotic distribution of $\hat{\kappa}_{3-2}$ under the null case (when all three raters are independent) is given in the following corollary.

Corollary

With the same assumptions as in above theorem and in addition if $P_{ijk} = P_{i..}P_{.j.}P_{..k}$ for all i,j,k, then the asymptotic distribution of $\sqrt{N}(\hat{\kappa}_{3-2}-\kappa_{3-2})$ is normal with mean zero and variance

$$\sigma^2_{\hat{\kappa}_{0,3-2}} = V_0/(1-B)^4 \qquad (3)$$

where $V_0 = B(1-B)^3+(1-B)^2(\Theta_4-\Theta_5^2)+2(1-B)^2(\Theta_1\Theta_5-\Theta_3)$

and $\Theta_1, \Theta_3, \Theta_4$ and Θ_5 are as defined above.

To construct confidence intervals for κ_{3-2}, the asymptotic standard deviation based on that obtained from the above theorem can be used. For example, 95% asymptotic confidence interval for κ_{3-2} is given by

$$\hat{\kappa}_{3-2} \pm 1.96\hat{\sigma}_{\hat{\kappa}_{3-2}}$$

To test the null hypothesis that all three raters are independent (which implies that κ_{3-2} equals to zero but the reverse is not necessarily true) the asymptotic standard deviation for $\hat{\kappa}_{3-2}$ given in the corollary should be used. That is, the ratio, $z=\hat{\kappa}_{3-2}/\hat{\sigma}_{\hat{\kappa}_{0,3-2}}$ can be used in conjunction with the normal table for level of significance.

IV Example

The following example gives the observed cell proportions based on three food experts' rating of 134 food items in terms of their taste using a three category rating scale (3=good, 2=fair, and 1=poor tasting quality).

Table 1

Judge 1	Judge 2	Judge 3: 1	2	3	$\hat{P}_{ij.}$	$\hat{P}_{i.}$
1	1	.09			.09	.15
	2			.06	.06	
	3					
2	1			.03	.03	.27
	2		.07	.10	.17	
	3	.02	.05		.07	
3	1					.58
	2			.02	.02	
	3		.05	.51	.56	
	$\hat{P}_{..k}$	.11	.17	.72	1.00	1.00

$\hat{P}_{.1.} = .12$, $\hat{P}_{.2.} = .19$, $\hat{P}_{.3.} = .69$

Using equations (1) and (2) with W=1/3, we obtain the estimated weighted Kappa for the above example:

$$\hat{\kappa}_{3-2} = .529$$

$$\hat{\sigma}_{\hat{\kappa}_{3-2}} = .068$$

Thus, the 95% confidence intervals for κ_{3-2} is estimated to be (.396, .662).

Initial simulation results done by Professor E. Chen of the University of Illinois shows that the normality of $\hat{\kappa}_{3-2}$ is easily achieved for sample size N=100 based on the probability model as described in Table 1. Further simulations are being planned for various probability models and different kinds of weights.

REFERENCES

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975), Discrete Multivariate Analysis - Theory and Practices, MIT Press, Cambridge, Mass.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales," Education and Psychological Measurement, 20, 37-46.

Fleiss, J. L. (1971), "Measuring Nominal Scale Agreement Among Many Raters," Psychological Bulletin, 76, No. 5, 378-82.

Fleiss, J. L., Cohen, J. and Everitt, B. S. (1969), "Large Sample Standard Errors for Kappa and Weighted Kappa," Psychological Bulletin, 72, No. 5, 323-27.

Goodman, L. A. and Kruskal, W. H. (1954), "Measures of Association for Cross Classifications," Journal of American Statistical Association, 49, 732-64.

Goodman, L. A. and Kruskal, W. H. (1972), "Measures of Association for Cross Classifications," Journal of American Statistical Association IV, 67, 415-21.

Haggard, E. H. (1958), Intraclass Correlation and the Analysis of Variance, Dryden Press, New York.

Light, R. J. (1971), "Measures of Response Agreement for Qualitative Data: Some Generalizations and Alternations," Psychological Bulletin, 76, No. 5, 365-77.

Lin, Y. S. (1975), "Measurement of Agreement for More Than Two Raters," a paper presented at the Central Regional Meeting of the Institute of Mathematical Statistics, St. Paul, Minnesota, March 1975.