CONDITIONAL AGREEMENT SCORE WITH TWO RATERS

Y. S. Lin, The Pillsbury Company

I. INTRODUCTION

When measuring agreement among many raters with categorical data, researchers generally deal with the agreement score {coefficients such as Cohen's Kappa (1960), Guttman's λ (1941), Scott's π (1955), etc.} which have the following common properties:

1. No assumptions were made concerning a "correct" assignment of items to categories. Agreement scores were not measured relative to any "correct" or "standard" assignment of items, but only with respect to the rater's internal consistency.
2. All raters are weighted equally.

Since there are many situations where the first characteristic will not be appropriate, we consider a slightly different kind of agreement problem. In this paper we will discuss the situation where it is reasonable to assume the existence of a "correct" ("expert") or "standard" assignment of items to categories. In this situation we may wish to either (1) compare each non-standard rater's response to that of the standard, or (2) measure the conditional agreement given the response from the standard is in category i (i = 1,2,...,I). This paper deals with the latter approach, illustrated with a two-rater case in a categorical setting. Cohen's Kappa-type conditional agreement score is used. The asymptotic variance and covariance for the maximum likelihood estimate of the conditional agreement score are obtained under multinomial sampling assumption. Asymptotic confidence interval for each conditional agreement score and the difference between two conditional agreement scores can then be calculated. An example will be used for illustration.

II. DEFINITION OF THE CONDITIONAL AGREEMENT SCORE

Assuming that there are two raters categorizing N items among I categories with one of the two raters being considered as "standard", we would like to examine the agreement between the two raters for only those items which the standard (or correct) assignments is in category i (i = 1,2,...,I).

For the IxI contingency table, we propose the following conditional agreement measure, $K_{(R1=i)}$, for those items which the first rater (standard rater or row variable) assigned to the i^{th} category:

$$K_{(R1=i)} = \{(P_{ii}/P_{i.}) - P_{.i}\}/(1-P_{.i}). \quad (1)$$

Here, $P_{ii}/P_{i.}$ is the probability that each of the ($NP_{i.}$) items would be assigned to category i by rater 2 given that rater 1 already assigned them to category i. Since the expected chance agreement for i^{th} category by both raters is $P_{i.}P_{.i}$, the conditional expected chance agreement of rater 2 given that rater 1 is in category i is $P_{i.}$. Then normalize the difference $(P_{ii}/P_{i.}) - P_{.i}$, (the probability of rater 2 agreeing with rater 1 in category i beyond chance agreement given that rater 1 has been known to have assigned the item to category i), by $1 - P_{.i}$. The range of $K(R1=i)$ is between $-P_{.i}/(1-P_{.i})$ and +1. The upper bound of $K(R1=i)$ is achieved if and only if $P_{ii} = P_{i.}$, that is, all the items that rater 1 is known to have assigned to category i are also assigned to the same category by rater 2. When none of the items known to be assigned to category i by rater 1 is assigned to category i by rater 2, i.e., $P_{ii} = 0$, then $K(R1=i) = -P_{.i}/(1-P_{.i})$. The conditional agreement score is equal to zero if, and only if, $P_{ii} = P_{i.}P_{.i}$ (or $P_{ii}/P_{i.} = P_{.i}$), same as in the unconditional situation. The conditional agreement score as defined in (1) is identical to the one suggested by Coleman (1966) in an unpublished paper (see Light, 1971) and is discussed in Bishop, Fienberg, and Holland (1975). To obtain the conditional agreement score for rater 1 given rater 2 is in category i, we simply interchange $P_{i.}$ and $P_{.i}$ in equation (1).

III. ASYMPTOTIC DISTRIBUTION OF $\hat{K}(R1=i)$

Assuming the multinomial sampling model with $X_{i.}$'s (observed number of items assigned to category i by rater 1) fixed, and let $\hat{K}$ (R1=i) denote the maximum likelihood estimator of $K(R1=i)$, then $\hat{K}(R1=i)$ is given by:

$$\hat{K}(R1=i) = (NX_{ii} - X_{i.}X_{.i})/\{X_{i.}(N - X_{.i})\}. \quad (2)$$

By applying the results given by Goodman and Kruskal (1972) we can easliy show that the asymptotic distribution of $(\sqrt{N}\{\hat{K}(R1=i) - K(R1=i)\})$ is normal with mean zero and variance:

$$\mathrm{Var}\{\hat{K}(R1=i)\}$$
$$=((P_{i.}-P_{ii})\{P_{ii}(1-P_{.i}-P_{i.}+P_{ii})$$
$$+(P_{i.}-P_{ii})(P_{i.}P_{.i}-P_{ii})\})/\{P_{i.}^3(1-P_{.i})^3\}. \quad (3)$$

Under the null hypothesis of independence the asymptotic variance given in (3) above is reduced to:

$$\mathrm{Var}_0\{\hat{K}(R1=i)\}=\{P_{.i}(1-P_{i.})\}/\{P_{i.}(1-P_{.i})\}. \quad (4)$$

In practice, we may wish to look at conditional agreement measures simultaneously. For example, we might want to find out whether one rater's agreement with the standard is the same given that responses from the standard are in categories i and ℓ ($i\neq\ell$). For this, it is necessary that we obtain the asymptotic covariance for $\hat{K}(R1=i)$ and $\hat{K}(R1=\ell)$. By using the general δ-method, it can be shown that the asymptotic covariance is given by:

$$\mathrm{Cov}\{\hat{K}(R1=i),\ \hat{K}(R1=\ell)\}$$
$$=\{1/(ND^2E^2)\}(ML(\phi_1-\lambda_1\lambda_2)+DL(\phi_2-\lambda_1\lambda_2)$$
$$+EM(\phi_3-\lambda_1\lambda_2)+DE\{\phi_4+ \quad (5)$$
$$(P_{ii}-\sum_j P_{ij}\Theta_{ij})(P_{\ell\ell}-\sum_j P_{\ell j}\Theta_{\ell j})\}),$$

where $D = P_{i.}(1-P_{.i})$, $E = P_{\ell.}(1-P_{.\ell})$,

$M = P_{ii}-P_{i.}P_{.i}$, $L = P_{\ell\ell}-P_{\ell.}P_{.\ell}$,

$\phi_1 = \sum_j P_{ij}(1-P_{j.}-P_{.i})(1-P_{j.}-P_{.\ell})+$
$\sum_j P_{\ell j}(1-P_{j.}-P_{.i})(1-P_{j.}-P_{.\ell})$,

$\lambda_1 = \sum_j P_{ij}(1-P_{j.}-P_{.i})$, $\Theta_{ij} = P_{j.}+P_{.i}$,

$\lambda_2 = \sum_j P_{\ell j}(1-P_{j.}-P_{.i})$, $\Theta_{\ell j} = P_{j.}+P_{.\ell}$,

$\phi_2 = \sum_j P_{ij}(1-P_{j.}-P_{.\ell})\Theta_{ij}+$
$\sum_j P_{\ell j}\Theta_{ij}(1-P_{j.}-P_{.\ell})$,

$\phi_3 = \sum_j P_{ij}\Theta_{\ell j}(1-P_{j.}-P_{.i})+$
$\sum_j P_{\ell j}\Theta_{\ell j}(1-P_{j.}-P_{.i})$,

$\phi_4 = \sum_j P_{ij}\Theta_{ij}\Theta_{\ell j}+\sum_j P_{\ell j}\Theta_{ij}\Theta_{\ell j}$.

From expressions (3) and (5), the asymptotic variance-covariance matrix can be constructed for $\hat{K}(R1=i)$, $i = 1,2,\ldots,I$. As usual, the estimated large sample variances and covariances are obtained by substituting observed proportions for cell probabilities.

IV. EXAMPLE

The following data with two home economists judging 159 breakfast food items will be used for illustration.

Two judges were asked to rate the values of breakfast foods using a three-category scale (3=good, 2=medium, 1=poor value). The entire test lasts more than one month with at least 30-minute time intervals between tastings. The test results are summarized in the following table. Table 1 gives the actual counts by cell. (Note: The value concept of a product generally is the combination of three elements: usage occasion, quality of the product, and unit price).

Table 1 Cell Frequencies of 159 Breakfast Foods as Rated by Two Judges Regarding Their Value

	Judge 2			
Judge 1	Good	Medium	Poor	Total
Good	63	7	5	75
Medium	7	24	14	45
Poor	4	3	32	39
Total	74	34	51	159

In the example above, estimated Kappa coefficient $\hat{K}_2 = .6077$ and asymptotic standard error $S(\hat{K}_2) = .056$, we concluded that the home economists agreed more often than they would by chance.

By using the conditional measures, $K(R1=i)$, we can attempt to localize the agreement. Suppose that home economist 1 can be assumed as the standard against whose judgment we wish to compare the rating of home economist 2. Using the expression (2) we have:

$$\hat{K}(R1=1) = .701,$$
$$\hat{K}(R1=2) = .406,$$
$$\hat{K}(R1=3) = .736.$$

This shows that home economist 2 agrees most with home economist 1 when the latter assigns the breakfast foods to the good and poor categories. The 95% asymptotic Bonferroni simultaneous confidence limits for each K(R1=i) we get:

K(R1=1) : {.518, .884},

K(R1=2) : {.213, .639},

K(R1=3) : {.519, .954}.

The estimated asymptotic variance-covariance matrix for $\hat{K}$(R1=i), i = 1,2,3, are given by:

.0058	.0006	.0009
	.0064	.0005
		.0076

Using the estimated variance-covariance, the 100(1-α)% confidence limit for the difference between two conditional agreement measures can be constructed, e.g., the 95% asymptotic Bonferroni simultaneous confidence intervals for differences between two conditional Kappas are given by:

$K_{(R1=1)} - K_{(R1=2)}$: {.043, .547},

$K_{(R1=1)} - K_{(R1=3)}$: {-.295, .223},

$K_{(R1=2)} - K_{(R1=3)}$: {.057, .603}.

REFERENCES

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

Guttman, L. "An Outline on the Statistical Theory of Prediction," in P. Horst (Ed.) The Prediction of Personal Adjustment. New York: Social Science Research Council, (1941).

Scott, W. A. "Reliability of Content Analysis: The Case of Nominal Scale Coding," Public Opinion Quarterly. 19, (1955), pp. 321-25.

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

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

Goodman, L. A. and Kruskal, W. "Measures of Association for Cross Classifications, IV: Simplification of Asymptotic Variances," JASA. 67, (June 1972), pp. 415-21.