SOME ROBUSTNESS AND CONVERGENCE PROPERTIES OF THE KAPPA STATISTIC

Claude O. Archer, Brentwood Veterans Administration Hospital
and University of California, Los Angeles

Norman W. Reccius, Brentwood Veterans Administration Hospital

Introduction

During the last several years the statistic that has emerged as the dominant measure of agreement (as a form of reliability) for categorical data is the kappa statistic introduced by Cohen (1960). This special case of association uses the simple or observed proportion of agreement adjusted for occurrence by chance. Later, Cohen (1968) expanded the concept to include a weighted kappa. Others — Fleiss, Cohen, and Everitt (1969), Fleiss (1971), Fleiss and Cohen (1973), and Fleiss (1975) — have described some of the statistical properties of the kappa statistic, including exact and large sample standard errors, and equivalence to the intraclass correlation coefficient. More recently, Landis and Koch (1977a, 1977b) have expanded the concept of kappa-type statistics to a heirarchical variety to deal with the problem of agreement among multiple observers.

As a means of expanding our practical understanding of the kappa statistic beyond the indices of spread, and relation to correlation as mentioned above, we examine the variation of the kappa statistic as a function of the number of categories or scale steps that may be used in a study. This investigation covers four simple discrete distributions, and is carried out using proportion of agreement as the reference point. Knowledge is also developed to increase insight into the number of categories or scale steps that are mathematically optimal, while retaining consistency with earlier studies on reliability. For example, Nunnally (1967) stated that in terms of psychometric theory, the advantage is always with using more, rather than fewer, scale steps. The reliability of rating scales as a monotonically increasing function of the number of steps was further noted by Guilford (1954). Also, Garner (1960) reiterated essentially the same thing in relating the number of scale steps to the information or the amount of discrimination that was inherent in the scale. The comments of these authors were made with no mathematical justification. More recently, Green and Roe (1970) have taken a multidimensional-scaling approach to the problem, and Ramsey (1973) has investigated the precision of the estimation of scale values by using a maximum-likelihood approach while varying the number of categories and the amount of discrimination. Our study adds some mathematical justification to the literature for the agreement problem. It is limited to the case of unweighted kappa as first defined by Cohen (1960).

Discrete Distributions

When investigating the properties of a descriptive statistic, it is necessary to examine various distributions so that one sees the behavior of the statistic under a variety of conditions. This enables us to realize the scope of any inferences that we may make. The kappa statistic is a measure of agreement for categorical or nominal data first defined by Cohen (1960) as

$$\kappa = \frac{p_o - p_c}{1 - p_c} = 1 - \frac{\delta}{1 - p_c}$$

where

p_o = observed proportion of agreement
p_c = expected proportion of agreement

and

$\delta = 1 - p_o$.

Agreement is defined as identical categorization or rating by two individuals, which we visualize as the diagonal elements of a Person 1 by Person 2 categorization matrix. For our study, it is assumed that the observed row and column marginals of this matrix have independent identical distributions and hence, determine p_c. Under these assumptions the value of kappa is computed as a function of the number of categories for the four particular discrete distributions described below. The four distributions are described in terms of k successive proportions for the marginals.

1. Uniform
   1 : 1 : ··· : 1 (k times)
2. Triangular
   1 : 2 : ··· : k
3. Symmetric, Center Peak
   1 : 2 : ··· : (k+1)/2 : ··· : 2 : 1 ; k odd
   1 : 2 : ··· : (k/2) : (k/2) : ··· : 2 : 1 ; k even
4. Symmetric, Center Dip
   (k+1)/2 : ··· : 2 : 1 : 2 : ··· : (k+1)/2 ; k odd
   (k/2) : ··· : 2 : 1 : 1 : 2 : ··· : (k/2) ; k even

The coefficient kappa as a function of δ and k can now be computed for these four distributions. Since k = 1 yields the trivial case of complete agreement, we consider $k \geq 2$.

For the uniform distribution $p_c = 1/k$, hence kappa is

$$\kappa = 1 - \left[\frac{k}{k-1}\right]\delta, \quad k \geq 2$$

Moreover, note that if k is fixed, κ is a simple linear function of δ; also $\kappa \to 1 - \delta = p_o$ as $k \to \infty$.

For the triangular distribution, we sum from one to k as follows:

$$p_c = \frac{\sum j^2}{\left[\sum j\right]^2} = \frac{2}{3}\,\frac{(2k+1)}{k(k+1)}$$

Therefore

$$\kappa = 1 - \left[\frac{3k\,(k+1)}{3k^2 - k - 2}\right]\delta \; ; \; k \geq 2$$

As before, $\kappa \to 1 - \delta = p_o$ as $k \to \infty$.

The symmetric distribution with a central peak is considered next. In this case, we can take advantage of symmetry and sum from one to k/2, hence when k is even

$$p_c = \frac{2\sum j^2}{\left[2\sum j\right]^2} = \frac{4}{3}\,\frac{k+1}{k\,(k+2)}$$

and after some algebra,

$$\kappa = 1 - \left[\frac{3k^2 + 6k}{3k^2 + 2k - 4}\right]\delta.$$

When k is odd, symmetry can again be used; each summation occurs from one to (k-1)/2, and

$$p_c = \frac{2\sum j^2 + \left[\frac{k}{2}\right]^2}{\left[2\sum j + \frac{k}{2}\right]^2} = \frac{4}{3}\,\frac{(k^3 + 3k^2 - k)}{(k^2 + 2k - 1)^2}$$

hence,

$$\kappa = 1 - \left[\frac{3\,(k^2 + 2k - 1)^2}{3k^4 + 8k^3 - 6k^2 - 8k + 3}\right]\delta.$$

Continuing with the same methodology for the symmetric distribution with a central dip, the same result as for the symmetric distribution with a central peaking point is obtained when k is even. On the other hand, when k is odd, the summations are from one to (k-1)/2, so

$$p_c = \frac{2\sum j^2 - 1}{\left[2\sum j - 1\right]^2} = \frac{4}{3}\,\frac{(k^3 + 6k^2 + 11k - 6)}{(k^4 + 8k^3 + 14k^2 - 8k + 1)}$$

and

$$\kappa = 1 - \left[\frac{3\,(k^4 + 8k^3 + 14k^2 - 8k + 1)}{3k^4 + 20k^3 + 18k^2 - 68k + 27}\right]\delta.$$

Practical Implications

The formulas that were derived above examine the variability of the coefficient kappa as a function of the number of categories, k, for four discrete distributions. The practical implications of these results for psychosocial studies using a categorical data collection are related below.

The concept of reliability, and subsequently the more narrow concept of agreement, evolved out of a practical need for demonstrating how consistent a particular instrument was under varying conditions. The need had arisen out of the recognition that sources of error, such as the instrument being used, the variability of the persons doing the ratings, and the variability of the patients or things being rated were important considerations. The practical implication of this recognition has been to insist upon "high" reliability.

The literature has uniformly dealt with this problem in very loose terms. For example, it is generally felt that a reliability of .9 is great, .8 is good, and .5 is poor, but the means for more understanding is lacking. In this paper we hope to conceive a more solid, meaningful interpretation for the concept of agreement when the kappa statistic is used.

This is done by relating the coefficient kappa to the simpler concept of proportion of simple or observed agreement, that is, the number of times that two people agree out of the total number of possibilities of agreement and nonagreement. The comparisons are done for the aforementioned discrete distributions and values of k = 2, 3, 4, 5, 6, 8, 10, and 20.

Figure 1 shows these comparisons for the case

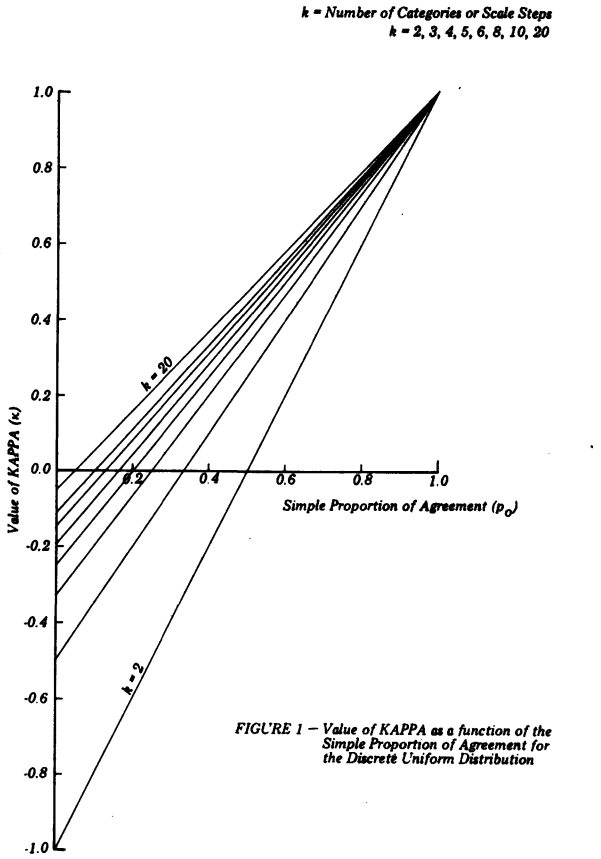

FIGURE 1 – Value of KAPPA as a function of the Simple Proportion of Agreement for the Discrete Uniform Distribution

of a uniform distribution. For example, when k = 2 we have a dichotomous distribution with a 50% chance of falling into each of the categories. Similarly, if k = 10 there is a 10% chance of falling into each of the categories. Studying Figure 1 more closely, assume an observed proportion of agreement of 50%. In other words, half of the time the two raters agree as to what they are rating or categorizing. Given a two-point dichotomous scale, kappa is zero, telling us that the agreement is exactly what is expected from pure chance. For a 10-point scale a kappa of approximately 0.45 is obtained; for a 20-point scale under the same situation, we get a kappa of approximately 0.48. Considering Figure 2, which is similar to Figure 1 except that the distributions assumed for the marginals are triangular,

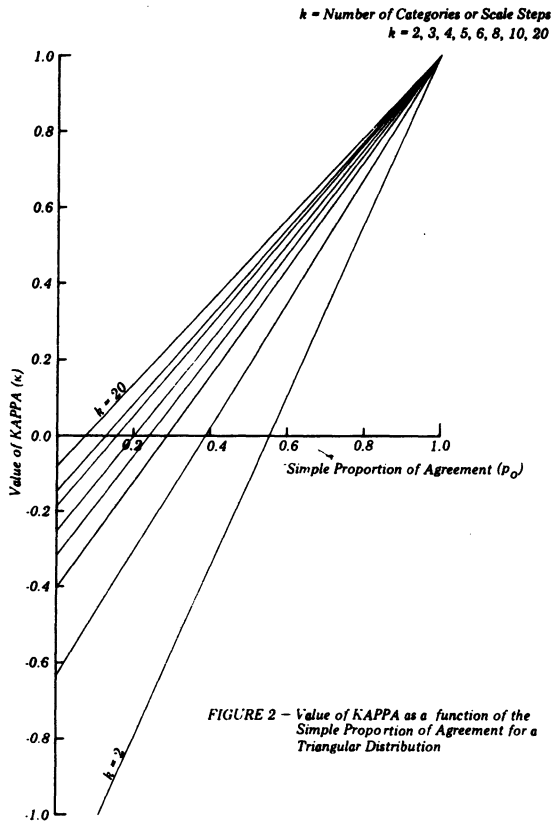

FIGURE 2 – Value of KAPPA as a function of the Simple Proportion of Agreement for a Triangular Distribution

TABLE I

A Partial Tabular Comparison of Kappa and the Simple Proportion of Agreement

Number of Categories k	Simple Agreement p_o	Kappa (κ) Uniform	Kappa (κ) Triangular
3	.5	.250	.182
5	.5	.375	.338
9	.5	.438	.418
20	.5	.474	.465
3	.7	.550	.510
5	.7	.625	.603
9	.7	.663	.651
20	.7	.684	.679
3	.9	.850	.836
5	.9	.875	.868
9	.9	.888	.884
20	.9	.895	.893

we note that when k = 2 and the proportion of simple agreement, p_o, is 0.5, kappa is equal to approximately -0.12. If k = 10 and p_o = 0.5, kappa = .43; if k = 20 we get an approximate value of 0.47 for kappa. For a simple agreement of about 0.9, and more than four categories kappa is between .86 and .90. These results from Figures 1 and 2 imply that the chance of getting a higher coefficient of agreement are better the more points or categories we have, even though the observed proportion of agreement is the same. (Note that we are not taking into account the ability of each person to place things equally well into 2, 6, 10, or 20 categories.) In addition, indications are that the more categories used, the closer the coefficient kappa is to the observed proportion of agreement, p_o.

Further illumination about what kappa means can be obtained by looking at some tabulations of k, p_o, and κ based upon our formulas (or Figures 1 and 2). For p_o = .5, .7, and .9, Table 1 illustrates that for a very good, highly reliable categorization scheme, the number of points does not matter nearly as much. Also, the magnitude of the difference between p_o and kappa is irrelevant for all practical purposes.

These two distributions, the uniform and triangular, have the widest disparity of the four discrete distributions considered, and since this disparity is not very broad the other two examples are not included in the illustrations.

Before we turn our attention to Figure 3, note that

$$\kappa = 1 - C_k \delta$$

where C_k is the quotient of two different polynomials in k for each of the discrete distributions introduced. Moreover, when $C_k = 1$, then $\kappa = 1 - \delta = p_o$, the simple proportion of agreement. Therefore, graphs of C_k as a function of the number of categories k and the discrete distribution considered are of interest. Figure 3 illustrates how rapidly C_k converges to one, and therefore how rapidly the kappa coefficient converges to the simple proportion of agreement. Only the two most dissimilar of the four discrete distributions are plotted here, since the other two distributions fell between these. The small differences between the curves give a strong indication of the robustness of kappa under the conditions considered. On Figure 3 note the very rapid change for small values of k up to 8 or 9, then a more gradual change to the end of the graph. Past k = 20, values of C_k for both distributions are very slowly asymptotic to one. Beyond k = 12, the practical difference of C_k and 1 is nil for all distributions considered. For example, for k = 12, simple agreement (p_o) on the order of .9 yields δ = .1 and kappa is about .89 for both the uniform and the triangular distributions; the only differences occurring in the third decimal place. The differences beyond k = 12 are even smaller. A further inference drawn from these results is that an optimal number of scale steps appears to be about eight or nine.

Conclusions

The agreement statistic kappa as a function of number of categories and the observed or simple proportion of agreement for the discrete

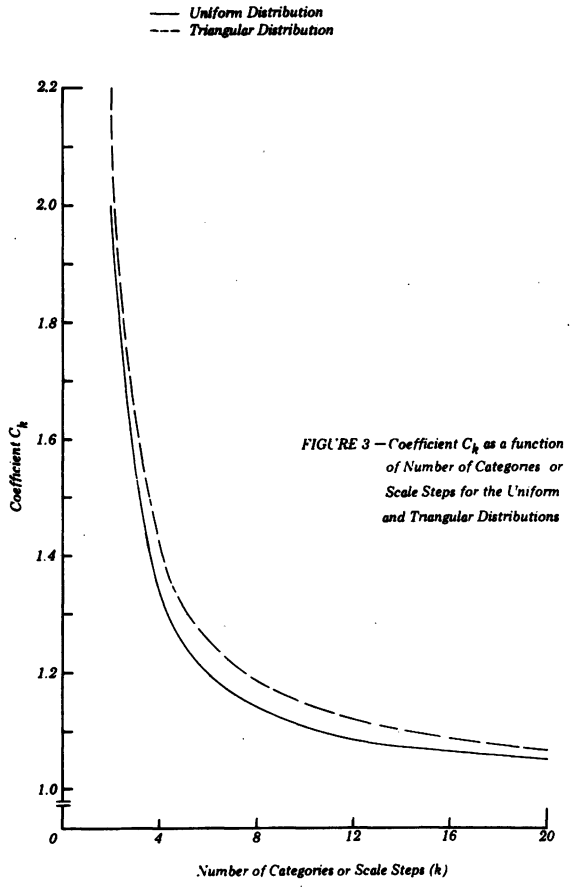

FIGURE 3 — Coefficient C_k as a function of Number of Categories or Scale Steps for the Uniform and Triangular Distributions

uniform, triangular, and symmetric with either center peak or center dip distributions has been studied. Findings indicate that for k moderately large (say $k \geq 8$), there is no practical difference between kappa and the simple or observed proportion of agreement. Also, for practical purposes, the differences between C_k for the distributions considered is negligible, indicating that kappa is a fairly robust indicator of agreement. We have also demonstrated empirically that $\kappa \to p_o$ monotonically as $k \to \infty$, hence a higher value of kappa is obtained with larger values of k, for a fixed amount of simple agreement.

References

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

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

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

Fleiss, J.L. Measuring agreement between two judges on the presence or absence of a trait. *Biometrics, 31,* 651-659, 1975.

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

Fleiss, J.L., Cohen, J., and Everitt, B.S. Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin, 72,* 323-327, 1969.

Garner, W.R. Rating scales, descriminability and information transmission. *Psychological Review, 67,* 343-352, 1960.

Green, P.E. and Rao, V.R. Rating scales and information recovery — how many scales and response categories to use? *Journal of Marketing, 34,* 33-39, 1970.

Guilford, J.P. *Psychometric Models.* (New York: McGraw-Hill) 1954.

Landis, J.R. and Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174, 1977a.

Landis, J.R. and Koch, G.G. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics, 33,* 363-374, 1977b.

Nunnally, J.C. *Psychometric Theory.* (New York: McGraw-Hill) 1967.

Ramsay, J.O. The effect of number of categories in rating scales on precision of estimation of scale values. *Psychometrica, 38,* 513-532, 1973.