

# Chance Agreement and Significance of the Kappa Statistic

Nobo Komagata

Department of Computer Science  
The College of New Jersey  
PO Box 7718, Ewing, NJ 08628  
komagata@tcnj.edu

## Abstract

Although the  $\kappa$  statistic has been used widely as an indicator of rater agreement, there have been some concerns about the existence of different definitions and some peculiar results involving skewed data. This note evaluates different definitions of  $\kappa$  and also demonstrates that the problem with directly comparing  $\kappa$  values, especially for skewed data, can be avoided by comparing their significance.

## 1 Introduction

The  $\kappa$  statistic seems the most commonly used measure of inter-rater agreement in Computational Linguistics, especially within the discourse/dialog community, e.g., Carletta et al. [1997]. The  $\kappa$  statistic is supposed to provide a means to compare inter-rater agreements of different experiments in a meaningful way. To detect the ‘goodness’ of inter-rater agreement, several proposals have been made regarding acceptable  $\kappa$  values. For example, Landis and Koch [1977, p. 165] considers  $\kappa > 0.8$  “almost perfect” (as well as other labels); Krippendorff [1980, p. 147] considers  $\alpha > 0.8$  (a closely-related measure) for “reporting on variables”; Emam [1999] (based on empirical distribution in Software Engineering) considers  $\kappa > 0.75$  “excellent” (some other approaches are reviewed in Di Eugenio [2000, Sec. 2]). We must note that these proposals often come with warnings such as “clearly arbitrary” Landis and Koch [1977, p. 165], “should not be adopted ad hoc” Krippendorff [1980, p. 147], and a cautious description about the use of  $\kappa$  in Carletta [1996, p. 252].

In spite of these warnings, the above-mentioned threshold values are widely used for judgment, e.g., Carletta [1996, p. 252] and Carletta et al. [1997, p. 25]. In this connection, several issues associated with the use of  $\kappa$  have been raised, most recently by Di Eugenio [2000]. One aspect Di Eugenio [2000, Sec. 2.1.2] points out is about the nature of data, e.g., independence among categories. Two other more technical issues are: (1) existence of different ways of computing chance agreement, an essential component in the  $\kappa$  statistic, and (2) the behavior of  $\kappa$  on skewed data. In this note, we discuss the latter two issues in detail.

As for the point (1), different ways of computing chance agreement has been pointed out in Fleiss [1971, p. 379], Siegel and Castellan [1988, p. 290], and Di Eugenio [2000, Sec. 2.1.1]. However, the effect of the difference has not been investigated in detail. As for the point (2), the effects of skewed data have been pointed out in Kraemer [1979, p. 470], Grove et al. [1981, p. 412], Chu-Carroll and Brown [1997, Sec. 2.2], and Di Eugenio [2000, Sec. 2.1.1]. Feinstein and Cicchetti [1990] seems to be the most detailed account of the situation.<sup>1</sup> The use of  $\kappa$  on skewed data is often considered *problematic* partly because the above-mentioned thresholds do not appear to be applicable. One potential solution to this ‘problem’ is to compute

---

<sup>1</sup>Chu-Carroll and Brown [1997, Sec. 2.2] proposes to “lower” the chance agreement in their computation, which seems to mislead the interpretation of their measures.

the significance of  $\kappa$  (instead of comparing raw  $\kappa$  values) (Di Eugenio et al. [1998, p. 327], and later work). However, such a practice does not seem to be well established in the CL community.

In this note, I echo the warning of Di Eugenio [2000], and provide examples that would illuminate the issues involved in the computation of  $\kappa$  statistic. In Section 2, we compare the  $\kappa$  statistic of Cohen [1960] (limited for two raters) and Fleiss [1971] (applicable to multiple raters, also adopted in a widely-cited text by Siegel and Castellan [1988]). A possible extension of Cohen [1960] to multiple raters is also discussed. In Section 3, we support the practice of Di Eugenio et al. [1998] and emphasize the necessity of computing significance through illustrative examples as a means to overcome difficulty comparing  $\kappa$  values.

## 2 Chance Agreement

This section compares the  $\kappa$  statistics of Cohen [1960] and Fleiss [1971], as well as the computation of chance agreement in Krippendorff [1980]. To be consistent, we mainly adopt the notation used in Siegel and Castellan [1988]. Since Cohen is limited to two raters, this section first focuses on that case. After comparing Cohen and Fleiss, we will also explore an extension of Cohen to multiple raters, which could be used as an alternative to Fleiss.

Let us first consider the following data for two raters  $X$  and  $Y$ , two categories  $A$  and  $B$ , and objects 1 through 16 ( $N = 16$ ):

Objects		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Raters	$X$	$A$	$A$	$A$	$A$	$A$	$A$	$A$	$A$	$B$	$B$	$B$	$B$	$B$	$B$	$B$	$B$
	$Y$	$A$	$A$	$A$	$A$	$A$	$A$	$A$	$B$	$A$	$B$	$B$	$B$	$B$	$B$	$B$	$B$

(1)

As a preparation, we consider the following table of joint probabilities to classify the judgments of the two raters.

Raters		$Y$			
		Category	$A$	$B$	$\Sigma$
$X$	$A$		$P(AA)$	$P(AB)$	$P(A_X)$
	$B$		$P(BA)$	$P(BB)$	$P(B_X)$
	$\Sigma$		$P(A_Y)$	$P(B_Y)$	

(2)

Each of the table cell can be computed with the following formulas:

$$P(AA) = \frac{\#(A_X A_Y)}{N}, P(AB) = \frac{\#(A_X B_Y)}{N}, P(BA) = \frac{\#(B_X A_Y)}{N}, P(BB) = \frac{\#(B_X B_Y)}{N}$$

$$P(A_X) = P(AA) + P(AB), P(B_X) = P(BA) + P(BB)$$

$$P(A_Y) = P(AA) + P(BA), P(B_Y) = P(AB) + P(BB)$$

Then, (2) can be filled in as follows:

Raters		$Y$			
		Category	$A$	$B$	$\Sigma$
$X$	$A$		0.44	0.06	0.50
	$B$		0.06	0.44	0.50
	$\Sigma$		0.50	0.50	

The probability of actual agreement between  $A$  and  $B$ , i.e.,  $P(A)$ , can be computed as

$$P(A) = P(AA) + P(BB) = 0.44 + 0.44 = 0.88$$

This formula/value is the same for Cohen [1960] and Fleiss [1971]. That is, the formula of Fleiss degenerates to that of Cohen for two raters.

We then need to adjust the value of  $P(A)$  against the chance (or expected) agreement,  $P(E)$ . According to Cohen,  $P(E)$  is the probability of  $A$  and  $B$  making the same decision, shown as follows:

$$P(E_C) = P(A_X)P(A_Y) + P(B_X)P(B_Y) = 0.50 \times 0.50 + 0.50 \times 0.50 = 0.50$$

According to Fleiss, the expected agreement is as shown below, where  $p_j$  is the probability of the category  $j$ .

$$P(E_F) = \sum p_j^2$$

But we also note the following:

$$p_A = \frac{\#(A_X) + \#(A_Y)}{2N} = \frac{P(A_X) + P(A_Y)}{2}$$

Then, we can compute  $P(E_F)$  as follows, the sum of the average of a rater choosing each category:

$$\begin{aligned} P(E_F) &= \left( \frac{P(A_X) + P(A_Y)}{2} \right)^2 + \left( \frac{P(B_X) + P(B_Y)}{2} \right)^2 \\ &= \left( \frac{0.50 + 0.50}{2} \right)^2 + \left( \frac{0.50 + 0.50}{2} \right)^2 = 0.50 \end{aligned} \quad (3)$$

For the present data,  $P(E_C) = P(E_F)$ . The  $\kappa$  statistic can then be computed as a chance-adjusted value of  $P(A)$  shown below.

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.88 - 0.50}{1 - 0.50} = 0.75$$

Here is a summary of the values.

	$P(A)$	$P(E)$	$\kappa$
Cohen	0.88	0.50	0.75
Fleiss	0.88	0.50	0.75

In general, however, Cohen and Fleiss end up with different  $\kappa$  values. Let us next consider the following data:

Objects		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Raters	X	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B
	Y	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B

(4)

Following the same procedure, we can fill in the table (2) and compute  $P(A)$ ,  $P(E)$ , and  $\kappa$  as follows:

Raters		Y		
		Category	A	B
X	A	0.50	0.00	0.50
	B	0.44	0.06	0.50
	$\Sigma$	0.94	0.06	
		$P(A)$	$P(B)$	$\kappa$
Cohen		0.56	0.50	0.13
Fleiss		0.56	0.60	-0.08

In this case, there is a substantial difference in the  $\kappa$  value. In fact, according to Cohen, the actual agreement is better than chance, while according to Fleiss, it is worse than chance (strong disagreement). The discrepancy is due to the skewed distribution of categories between the two raters. Cohen computes chance based on each rater's judgment; Fleiss computes chance by averaging out the probability of all categories for each rater.

More generally, it is possible to compute the difference between the chance agreements of the two definitions:

$$\begin{aligned}
P(E_F) - P(E_C) &= \left( \frac{P(A_X) + P(A_Y)}{2} \right)^2 + \left( \frac{P(B_X) + P(B_Y)}{2} \right)^2 - [P(A_X)P(A_Y) + P(B_X)P(B_Y)] \\
&= \left( \frac{P(A_X) - P(A_Y)}{2} \right)^2 + \left( \frac{P(B_X) - P(B_Y)}{2} \right)^2 \\
&= \left( \frac{0.5 - 0.94}{2} \right)^2 + \left( \frac{0.5 - 0.06}{2} \right)^2 = 0.10
\end{aligned}$$

That is, Cohen and Fleiss agree if and only if  $P(A_X) = P(A_Y)$  and  $P(B_X) = P(B_Y)$ , as seen in (1). The additional assumption made by Fleiss is that the distribution of categories is even for each rater. Note that Siegel and Castellan [1988, p. 291] states that the additional assumption made by Fleiss [1971] is that  $P(R_X)$ , for some rater  $R$  and the category  $X$ , is "the same for all raters". But this statement is too weak because even without different  $P(R_X)$  for two raters, Cohen and Fleiss agree as long as  $P(A_X) = P(A_Y)$  and  $P(B_X) = P(B_Y)$ . We can verify this by modifying (1) as follows: replacing  $B$  with  $A$  for, e.g., objects 14 through 16 for both  $X$  and  $Y$ .

Before proceeding, let us examine yet another way of computing chance agreement discussed in Krippendorff [1980, p. 134]. In this case, chance agreement is based on a combinatorial selection process.

$$P(E_K) = \frac{\#(A_X) + \#(B_X)}{2N} \frac{\#(A_X) + \#(B_X) - 1}{2N - 1} + \frac{\#(A_Y) + \#(B_Y)}{2N} \frac{\#(A_Y) + \#(B_Y) - 1}{2N - 1}$$

Since this approach does not distinguish the source of the available categories, when Cohen and Fleiss disagree significantly, it gives a value closer to, but still different from, Fleiss.

Observing three different ways of computing chance agreement, it is clear that we must always identify which  $\kappa$  statistic is being used. Comparing between Cohen [1960] and Fleiss [1971], the stronger assumption of Fleiss does not seem appropriate because the formula (3) inherently includes the irrelevant values  $P(A_X)^2$ ,  $P(A_Y)^2$ ,  $P(B_X)^2$ , and  $P(B_Y)^2$ , as can be seen below.

$$P(E_F) = \frac{P(A_X)^2 + 2P(A_X)P(A_Y) + P(A_Y)^2}{4} + \frac{P(B_X)^2 + 2P(B_X)P(B_Y) + P(B_Y)^2}{4}$$

That is,  $P(E_F)$  is diluted with these terms that correspond to raters' agreement with themselves. Krippendorff [1980, p. 134] too, being closer to Fleiss, has such components in the computation. As a result, for the two-rater case, Cohen [1960] seems to be a more appropriate way of computing  $\kappa$ .

Unfortunately, Cohen [1960] is limited to two raters, and presumably, that is the reason Fleiss [1971] presents an extension. However, Fleiss does not agree with Cohen for the two-rater case due to the stronger assumption on chance agreement; again, the problem is inclusion of 'self agreement'.

Let us now examine the chance agreement of Fleiss for  $k$  raters with  $m$  categories ( $p_j$  for the probability for the  $j$ th category;  $p_{j,r}$  for the probability for the  $j$ th category and the  $r$ th rater):

$$P(E_F) = \sum_{j=1}^m p_j^2 = \sum_{j=1}^m \left( \frac{\sum_{r=1}^k p_{j,r}}{k} \right)^2$$

As in the two-rater case, the computation includes 'self agreement' as can be seen as diagonal elements (in parentheses) in the following table.

	$p_1$	$p_2$	$p_3$	$\dots$	$p_{k-1}$	$p_k$	
$p_1$	$(p_1 p_1)$	$p_1 p_2$	$p_1 p_3$		$p_1 p_{k-1}$	$p_1 p_k$	
$p_2$	$p_2 p_1$	$(p_2 p_2)$	$p_2 p_3$		$p_2 p_{k-1}$	$p_2 p_k$	
$p_3$	$p_3 p_1$	$p_3 p_2$	$(p_3 p_3)$		$p_3 p_{k-1}$	$p_3 p_k$	
$\vdots$				$\ddots$			
$p_{k-1}$	$p_{k-1} p_1$	$p_{k-1} p_2$	$p_{k-1} p_3$		$(p_{k-1} p_{k-1})$	$p_{k-1} p_k$	
$p_k$	$p_k p_1$	$p_k p_2$	$p_k p_3$		$p_k p_{k-1}$	$(p_k p_k)$	

(5)

An extension of Cohen [1960] that degenerates to Cohen for the two-rater case must not include the diagonal elements. The following, called  $P(E_C)$ , satisfies this condition by averaging the lower triangular elements (less diagonal elements):

$$P(E_C) = \sum_{j=1}^m \frac{\sum_{r=1}^{k-1} \sum_{s=j+1}^k p_{j,r} p_{j,s}}{\frac{k(k-1)}{2}}$$
(6)

Note that we can still use the general form of  $P(A)$  in Fleiss [1971].

In order to see the effect of the above formula, let us now apply it to Table 9.15 of Siegel and Castellan [1988, p. 291]. According to Fleiss, the result is  $\kappa = 0.41$  regardless of the actual judgment of the raters. As for the extended Cohen (6), the  $\kappa$  value would vary at least between 0.41 and 0.46. Since  $\kappa = 0.41$  is already significant ( $p < 0.01$ ), this particular case would not affect our judgment. However, this suggests a possibility of large difference in  $\kappa$  and thus a possibility of affecting the evaluation.

In this section, we see that the chance agreement of Fleiss [1971] is less desirable than that of Cohen [1960]. We also see that it is possible to generalize Cohen to multiple raters without the stronger assumption of Fleiss.

### 3 Significance

Although the puzzling  $\kappa$  values for skewed data has been noted and discussed in the literature, few papers actually demonstrate the relation between  $\kappa$  values and their significance. This section examines the relation between  $\kappa$  values and their significance with a few examples.

First, we compute the  $\kappa$  for the following data, where  $\kappa = 0.50$ :

Objects		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Raters	X	A	A	A	A	A	A	A	A	B	B	B	B	B	B	B	B
	Y	A	A	A	A	A	A	B	B	A	A	B	B	B	B	B	B

(7)

If the number of objects is large, following a Central Limit theorem, we can estimate that the distribution of  $\kappa$  be close to ‘normal’. Then, we can adopt the procedure found in Fleiss [1971] (repeated in Siegel and Castellan [1988]), and compute the significance of  $\kappa$  values. Here, we use a special case for two raters, which agrees with the significance computation of Cohen [1960].

$$\text{var}(\kappa) = \frac{P(E)}{N(1 - P(E))} \text{ (for } k = 2\text{)}$$

$$z = \frac{\kappa}{\sqrt{\text{var}(\kappa)}}$$

The results are summarized below.

	$P(A)$	$P(E)$	$\kappa$	$\text{var}(\kappa)$	$z$
Cohen	0.75	0.50	0.50	0.063	2.00
Fleiss	0.75	0.50	0.50	0.063	2.00

We can conclude that the agreement is significant ( $p < .05$ ).

Next, we consider the following data.

Objects		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Raters	X	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B
	Y	A	A	A	A	A	A	A	A	A	A	A	A	A	A	A	B

(8)

The results for this data are summarized as follows.

	$P(A)$	$P(E)$	$\kappa$	$\text{var}(\kappa)$	$z$
Cohen	1.00	0.88	1.00	0.47	1.46
Fleiss	1.00	0.88	1.00	0.47	1.46

We conclude that the agreement is *not* significant ( $p < .05$ ). As we can see, a high  $\kappa$  value (even the perfect agreement) does not guarantee significant agreement because  $P(E)$  is already very high. That is, it is easy to get agreement between the raters.

This situation involving skewed data has been referred to as “problem” Grove et al. [1981, p. 412], “paradox” Feinstein and Cicchetti [1990, p. 543], “highly problematic” Chu-Carroll and Brown [1997, Sec. 2.2], or having “difficulty in the interpretation of  $\kappa$ ” [Berry, 1992]. In addition, Di Eugenio [2000, Sec. 2.1.1] asks “why” and Kraemer [1979, p. 461] writes “its [ $\kappa$  statistic] value as a measure of the quality of an observation in clinical or research contexts is not clear” partly because of this property.

As has been explored, e.g., in Feinstein and Cicchetti [1990], the cause of this property of  $\kappa$  is the high chance agreement due to skewed data. There are some alternatives to  $\kappa$ , e.g., Kraemer [1979]. However, as summarized by Goldman [1992], Shrout et al. [1987, p. 176] and Cicchetti and Feinstein [1990, p. 557] conclude that this property of  $\kappa$  is actually a desirable consequence of the chance agreement adjustment. Thus, when data is skewed, the judges must do better than chance to result in significant agreement; in some cases, the researcher may need to increase the number of objects. As shown in the above example, it is in

general meaningless to directly compare  $\kappa$  values. Instead, by computing  $z$  values, we can still compare the significance of different data.

In Section 1, we introduced an extension of Cohen [1960] that is more desirable than Fleiss [1971]. To be able to compare the significance of  $\kappa$  for multiple raters, we will need a way to compute significance for the extension. Here is a preliminary analysis adapted from Fleiss. Since the numerator of the formula for computing variance in [10] of Fleiss [1971, p. 380] does not depend on  $P(E)$ , we can leave that part as is (except that we use  $\sum p_j^2$  instead of  $P(E)$  because we replace  $P(E)$  with  $P(E_C)$ ). Since  $P(E)$  in the denominator refers to the chance agreement, we can use  $P(E_C)$  in its place. Then, we obtain a formula similar to that of Fleiss, shown in a way similar to (9.30) in Siegel and Castellan [1988].

$$\text{var}(\kappa) = \frac{2}{Nk(k-1)} \frac{\sum p_j^2 - (2k-3) \left[ \sum p_j^2 \right]^2 + 2(k-2) \sum p_j^3}{[1 - P(E_C)]^2}$$

## 4 Conclusion

The first point of this note is that for two raters, the  $\kappa$  statistic of Cohen [1960] is more desirable than that of Fleiss [1971] because the latter requires an unnecessarily strong assumption. Although the original computation of Cohen is limited to two raters, we noted that it can be extended to multiple-rater cases without using the stronger assumption of Fleiss.

The second point is that the so-called ‘problem’ with skewed data can be avoided by comparing the significance of  $\kappa$  values. This is also possible for the extended Cohen for multiple raters, introduced in Section 2.

In spite of the popularity of the  $\kappa$  statistic, a number of issues are being reported in the literature. In some cases, researchers do not pay sufficient attention to the ‘meaning’ of different  $\kappa$  statistics. In some other cases, researchers use arbitrary thresholds for evaluation with little justification. I hope that this note presents materials that are useful for delineating some of these issues.

Finally, the MS Excel file used for computing various values in this note is available at: “<http://www.tcnj.edu/~komagata/pub/kappa.xls>”.

## Bibliography

- Charles C. Berry. 1992. The Kappa Statistic (letter to the editor). *Journal of American Medical Association*, 268(18):2513.
- Jean Carletta et al. 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics*, 23(1):13–31.
- Jean Carletta. 1996. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, 22(2):249–254.
- Jenifer Chu-Carroll and Michael K. Brown. 1997. Tracking Initiative in Collaborative Dialog Interactions. In *Proceedings of the 35th Annual Meeting/Eighth Conference of the European Chapter of the Association for Computational Linguistics (ACL35/EACL8), Madrid, Spain, July 1997*, pages 262–270.
- Domenic V. Cicchetti and Alvan R. Feinstein. 1990. High Agreement But Low Kappa: II. Resolving the Paradoxes. *Journal of Clinical Epidemiology*, 43(6):551–558.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

- Barbara Di Eugenio, Pamela Jordan, Richmond Thomason, and Johanna Moore. 1998. An empirical investigation of collaborative dialogues. In *Proceedings of the International Conference on Computational Linguistics/37th Annual Meeting of the Association for Computational Linguistics (COLING-ACL98), Montreal, Québec, August 1998*, pages 325–329.
- Barbara Di Eugenio. 2000. On the usage of Kappa to evaluate agreement on coding tasks. In *LREC2000, Proceedings of the Second International Conference on Language Resources and Evaluation, Athen, Greece*, pages 441–444.
- Khaled El Emam. 1999. Benchmarking Kappa: Interrater agreement in software process assessment. *Empirical Software Engineering*, 4:113–133.
- Alvan R. Feinstein and Domenic V. Cicchetti. 1990. High Agreement But Low Kappa: I. The Problems of Two Paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549.
- Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.
- Ronald L. Goldman. 1992. The Kappa Statistic (reply to a letter to the editor). *Journal of American Medical Association*, 268(18):2513–2514.
- William M. Grove et al. 1981. Reliability Studies of Psychiatric Diagnosis. *Archives of General Psychiatry*, 38:408–413.
- Helena Chmura Kraemer. 1979. Ramifications of a Population Model for Kappa as a Coefficient of Reliability. *Psychometrika*, 44(4):461–472.
- Klaus Krippendorff. 1980. *Content Analysis*. Beverly Hills, CA: Sage Publications.
- J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:159–174.
- Patrick E. Shrout, Robert L. Spitzer, and Joseph L. Fleiss. 1987. Quantification of Agreement in Psychiatric Diagnosis Revisited. *Archive of General Psychiatry*, 44:172–177.
- Sidney Siegel and N. John Castellan, Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences, 2nd ed.* New York: McGraw-Hill.