International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 7 ǁ 2021 ǁ PP. 23-28

Cohen's Kappa Statistic and newKappaStatistic for Measuring and Interpreting Inter-Rater Agreement

Chaman Lal Sabharwal
Computer Science Department,
Missouri University of Science and Technology,
Rolla, MO 65409 USA

Abstract
The kappa statistic is used to describe inter-rater agreement and reliability. Kappa statistic is applied to interpret data that are the result of a judgement rather than a measurement. Measurement of the extent to which the raters assign the same score to the same variable is called inter-rater reliability. In datamining, it isusually measured as percent agreement, as the number of agreement scores divided by the total number of scores. In 1960, Jacob Cohen reviewed the use of percent agreement critically due to its inability to account for chance agreement. He introduced the Cohen's kappa, developed to account for the possibility that raters actually guess on at least some variables due to uncertainty. This was meant to help healthcare professionals. But the assumption created limitations, paradox. It is explained with numerous examples in this paper. We propose a new metric newKappa that is consistent with intuition, human cognition, observed and random agreements. The percent agreement, Cohen's kappa agreement, and proposednewKappa agreement scores along with interpretation of the scores are compared to show the newKapp is preferable measure for testing inter-rater agreement.

Keywords: *kappa, inter-rater, observed agreement, random agreement, chance agreement*

Date of Submission: 06-07-2021 Date of acceptance: 20-07-2021

I. INTRODUCTION

The kappa statistic is used to describe interrater agreement and reliability. Kappa statistic is applied to interpret data that are the result of a judgement rather than a measurement. Measurement of the extent to which the raters assign the same score to the same variable is called interrater reliability. In datamining, it is usually measured as percent agreement, as the number of agreement scores divided by the total number of scores. In 1960, Jacob Cohen critiqued use of percent agreement due to its inability to account for chance agreement. He introduced the Cohen's kappa, developed to account for the possibility that raters actually guess on at least some variables due to uncertainty[1]. This was meant to help healthcare professionals. But the assumption created limitations, paradox. It is explained with numerous examples.

Kappa statistic is used to interpret data that are the result of a judgement rather than a measurement[2]. A common approach to quantify agreement between judges is called the kappa statistic. Cohen's Kappa Statistic measures the level of inter-rater agreement between two judges who classify items into mutually exclusive categories[3]. Kappa analyses the prevalence of observed agreement between two raters to the probability of expected agreement between them to determine if the ratings are independent. Kappa is used to determine how far the two raters agree when both raters apply a criterion to determine whether or not some condition holds[4]. That is, Cohen's kappa describes strength of inter-rater agreement. Cohen's Kappa attempts to account for inter-rater agreement if purely by chance[5].Cohen's kappa statistic has a paradox that interprets low agreement even where there is high precent observed and random agreement[4],[6], [7]. To correct that ambiguity, we present a new kappa whose classifcation is consistent with human cognition.

The paper is organized as: Section 2 is backgroung information, termintiogy and the problem defintion, Section 3 describes the proposed newKappaclassifer and its comparison with Cohen's kappa, Section 4 is conclusion why newKappa is preferable to Cohn's kappa. Numerous examples are given for this purpose. It ends with references in the Section 5.

II. BACKGROUND

2.1 Terminology

For analysis, the true data values are synonymously used actual or observed data. The modeled values are known as predicted, estimated data. Depending on the data and purpose, the terms evaluator, grader, judge, rater, critic are synonymous in use for the model or the application to create data for analysis. The terms similarity, match, agreement are used for measuring and interpreting the quality of agreement between true and estimated values.

2.2 Problem Definition

For analysis, the true data values are synonymously used actual or observed data. The modeled values are known as predicted, estimated data. Depending on the data and purpose, the terms evaluator, grader, judge, rater, critic are synonymous in use for the model or the application to create data for analysis. The terms similarity, match, agreement are used for measuring and interpreting the quality of agreement between true and estimated values.

There are several applications wheredetermining and interpreting the agreement between two judges is necessary for further decision making. Such analysis may be critical in some cases, than in other cases: Two doctors may diagnose whether or not each of a group of patients has disease (heart, diabetes) based on specific symptoms; Two researchers both assess whether each of papers submitted to a conference is acceptable or not, based on some evaluation criteria (innovation, originality, clarity, citations etc); two book reviewers may post book reviews on Amazon; two movie critics may post movie reviews their websites[8],[9]; applications for hashtags high jacking, real or spam.

Since observations are subjective to some degree, it is all the more important in healthcare where judgement is a matter of life and death. The interpretations of diagnostic tests of physical exams, radiographic findings are subjective interpretation by observers to some degree[7]. It is vital that the observation is not a product of guessing.In academic fields, accuracy is the ratio of correct answers to total questions; error is the ratio of wrong answers to total questions. These metrics work well for an ideal data set — which doesn't exist in the real world[10]. Sometimes judgement is interpreted as the term precision (reliability), this is a term that is incorrectly used for accuracy, precision is built in the kappa statistic. Traditionally, it was measured as percent agreement, calculated as the number of agreement scores divided by the total number of scores. In 1960, Jacob Cohen critiqued use of percent agreement due to its inability to account for chance agreement [1], [11].Rather than just calculating the percentage of items that the raters agree on, Cohen's Kappa attempts to account for the fact that the raters may happen to agree on some items purely by chance[3], [5]. That is, when interpreting kappa, the user should keep in mind that the kappa value itself could be due to chance [7], see Table6. Either way, higher observed and expected value can have lower kappa value and lower observed and expected value can have higher kappa value adding to further confusion[4], [6]. See examples, Tables 4,5,6,7.

How is kappa statistic computed. It is best explained with the help of an example. Suppose we have an academic conference ,papers are invited and are distributed to reviewers to review for acceptaning papers to be presented at the conference. The reviewers grade them with some scores and net outcome is accept (yes), reject (no) for each of the papers. Suppose n papers are assigned to each of two reviewers/judges A and B. The reviewer A accepts n_1 papers, rejects n_2 papers, the reviewer B accepts m_1 and rejects m_2 papers. The conference chair, CC, uses this information to determine which papers are accepted for the conference, which are rejected, the remaining are left for poster session.CC forms a confusion matrix and describes the two ratings as follows: accept a papers, reject d papers, the b papers accepted by B and rejected by A; and the c paper accepted by A and rejected by B are allocated as $b+c$ poster sessions papers respectively; depending on the reliable quality of agreement. Since A accepts n_1 papers and rejects n_2; B accepts m_1 papers and rejects m_2, the expected values for acceptance and rejection may turn out to be different from actually accepted or rejected. Let p_o be the observed probability of both accepted and rejected papers,the prevalence (proportion) of observed agreement, $p_o = \frac{a+d}{n}$. The question is how to ascertain that the graders are in satisfactory agreement.

Table 1: Confusion for the reliability of agreement

A B	accept	reject	total
accept	a	b	m_1
reject	c	d	m_2
total	n_1	n_2	n

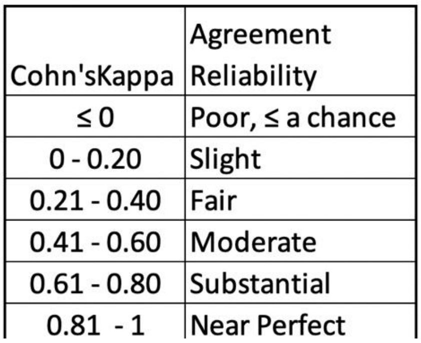
The symbol, p_e, represents the hypothetical prevalence of chance agreement[5]. The expected proportional agreement for accept and reject is random agreement. This is calculated as follows:

$$p_e = \frac{n_1}{n} \bullet \frac{m_1}{n} + \frac{n_2}{n} \bullet \frac{m_2}{n} = \frac{n_1 \bullet m_1}{n^2} + \frac{n_2 \bullet m_2}{n^2} = \frac{n_1 \bullet m_1 + n_2 \bullet m_2}{n^2}.$$

Note the expected value, p_e, is not necessarily the same as observed value, $p_o = \frac{a+d}{n}$.The difference between the observed and expected is p_o - p_e. By scaling this difference, Cohen's kappa is defined as a statisticthat includes agreement by chance as kappa = $\frac{p_o - p_e}{1 - p_e}$ [11], see examples.

It is neither observed nor or expected agreement, it simply magnifies the difference. If kappa = 1, then there is a perfect agreement; if k=0, it means that observed and random agreement are equivalent. For other values of kappa statistic, the agreement may be characterized as poor, slight, fair, moderate, substantial, and near perfect Table 2 by diving the range into hypothetical uniform confidence intervals, [0.0, 0.2], [0.2, 0.4], [0.4, 0.6], [0.6, 0.8], [0.8, 1]

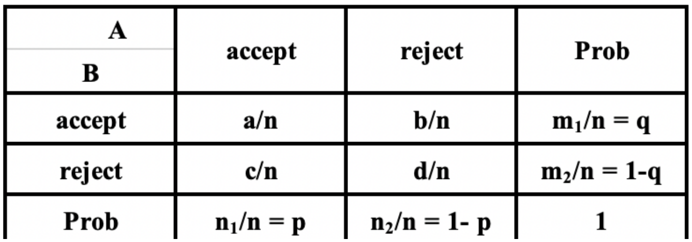
Table2: Kappa values and their Interpretation[5]



Cohn'sKappa	Agreement Reliability
≤ 0	Poor, ≤ a chance
0 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1	Near Perfect

How did it come aboutfor values to quantify agreement between two raters. The standard for a "good" or "acceptable" kappa value is arbitrary [2]. But, this is often considered as a rule of thumb for Cohen's kappa [3], [11]. The prevalence (proportion) of observed disagreement is $\frac{b+c}{n}$, but expected value would be $\frac{n_1 \bullet m_2 + n_2 \bullet m_1}{n^2}$. In terms of probabilities, the Table 2 can be written in terms of proportional matrix in Table 3 as

Table 3: Confusion Matrix for prevalence of observed values

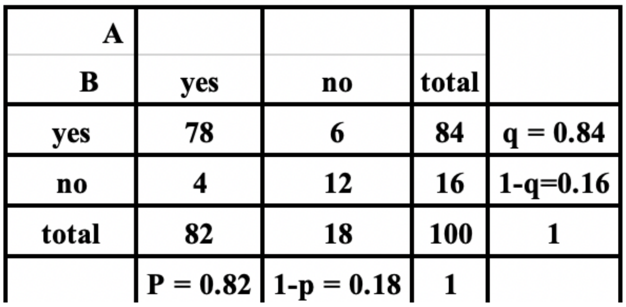


A B	accept	reject	Prob
accept	a/n	b/n	$m_1/n = q$
reject	c/n	d/n	$m_2/n = 1-q$
Prob	$n_1/n = p$	$n_2/n = 1- p$	1

The Table 3 translates into $p_o = \frac{a+d}{n}$and $p_e = pq +(1-p)(1-q)$

Step-by-Step Calculation of Cohen's Kappa [5]. Suppose two museum curators are asked to rate two paintings on whetheror not they're good enough to be hung in a new exhibit,see Table 4:

Table4:Observed rating of paintings for Museum



A B	yes	no	total	
yes	78	6	84	q = 0.84
no	4	12	16	1-q=0.16
total	82	18	100	1
	P = 0.82	1-p = 0.18	1	

From the observed values of both raters yes and both raatersno, we have

$p_o = \frac{78+12}{100}$ = 0.78+0.12 =0.90

From the expected values for both yes and both no, including agreement by chance

p_e = 0.82*0.84+0.18*0.16 = 0.6888 + 0.0288 =0.7176

Clearly the two values P_o and P_e are not equal.
The value of Cohn's kappa is

$$\text{kappa} = \frac{p_o - p_e}{1 - p_e} = \frac{0.9 - 0.7176}{1 - 0.7176} = \frac{0.1824}{0.2824} = 0.6459$$

which is almost 2/3 of direct observed value, p_o. Kappa classifies it as substantial agreement.

Cohen's Kappa may work well to measure agreement sometimes. There is no clear cut rule of thumb for good agreement and it all depends on the nature of data and purpose of the study[3]. It is interesting to note that in some cases it is possible that there is excellent observed agreement, but a poor kappa interpretation, also there is poor observed agreement, but aexcelllent kappa interpretation. Confusion arises when observers get the same over all percentage from various observations, see Table 5.

The question is if there is a better alternative to Kappa statistic? This paper gives more insight into this phenomena why this paradox occurs and what is annother way to define this quantitave measure. In the next section, we define a new metric toquantify the agreement classification that is consistent with human cognition. We base this on two metrics that more in line with human cognition and accuracy relative to observed and expected data.

Some researchers may argue about using percent agreement alone[3]. Some may justify that the percent agreement does not correct for chance agreement, whereas the Kappa statistic corrects for chance agreement. This correction created a paradox. In fact, the question is judgement, not correction. Judgement is made from the observations, not any kind of abstraction.

III. THE NEWKAPPA STATISTIC

3.1 Cohen's kappa statistic

This statistic interprets how well two raters agree with each other, regardless of whether their judgements are right or wrong. If two-raters use a criterion to make the same assessment on the same targets, then their agreement provides evidence for highly reliable rating[3]. If they do not, then either the criterion tool isn't useful or the raters are not well enough trained. Confusion arises when observers get the same over all percentage on different observations[3]. There are cases where (1) there is excellent observed and random agreement, but there is a poor kappa interpretation, also (2) there is poor observed and expected agreement, but there is an excellent kappa interpretation.

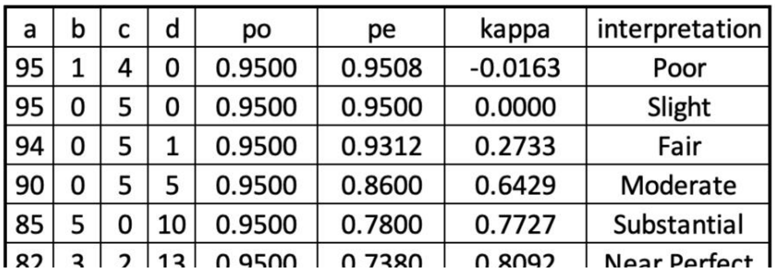
One way around this is to use further statistics of z-statistic for condeence intervals. We do not need this and will not go into it. As stated above, we define a new metric to quantify the agreement classification that is consistent with human cognition. We also define two additional metrics that armore in line with human cognition for accuracy relative to observed and expected data.

It is probable that for the same value of p_o, there is different p_e value leading to different Kappa value creating a confusion. In the following examples of confusion matrices, the observed agreement value, expected agreement value, Kappa statistic and agreement conclusion are described to explain the paradox. In each case below, the observed agreement is the same value, $p_o = 0.9000$. The raters may rate differently, and still get the same value for p_o, same observed agreement percentage. The matrices have different values for off-diagonal terms resulting in sum of main diagonal values as 90, but classification varies from **poor** to **substantial** close to perfect agreement.For Kappa statistic, the confusion is that there is not a clear interpretation of what the kappa value means corresponding to the same observed agreement.

For example, we need a more robust new metric.In the following Table5,a,b,c,d are entries of matrix $\begin{bmatrix} \text{a} & \text{b} \\ \text{c} & \text{d} \end{bmatrix}$, n = a+b+c+d, $p_o = \frac{a+d}{n}$, pe = $\frac{(a+b)(a+c)+(d+b)(d+c)}{n^2}$, kappa = $\frac{p_o - p_e}{1 - p_e}$ are observed agreement, random agreement probabilities and kappa is the Cohen's Kappa.

In Table5, all examples are different observations with 95% observed agreement. For this observed agreement "*value*", there can be severl different observations. For these observations the expected agreement values are close and consistent. Strangely, he kappa values so staggared that they classify the outcome differentlyfrom poor, slight, fair, moderate, substantial, perfect in this example.

Table5:all six examples are95% observed agreement.
The table contains observed agreement values, random agreement values,
Cohen’s kappa values,classification based on kappa value



a	b	c	d	po	pe	kappa	interpretation
95	1	4	0	0.9500	0.9508	-0.0163	Poor
95	0	5	0	0.9500	0.9500	0.0000	Slight
94	0	5	1	0.9500	0.9312	0.2733	Fair
90	0	5	5	0.9500	0.8600	0.6429	Moderate
85	5	0	10	0.9500	0.7800	0.7727	Substantial
82	3	2	13	0.9500	0.7380	0.8092	Near Perfect

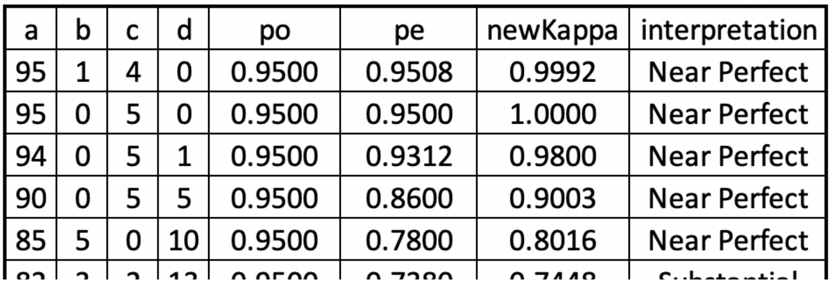
3.2 Proposed newKappa statistic and comparsionwwith Cohen’s kappa

Unlike Cohen’s kappa that included a scale factor from only expect value of agreement, we have created a scale factor that is harmonic mean of observed agreement and expected agreement. The harmonic mean of a and b is $\frac{2}{\frac{1}{a}+\frac{1}{b}} = \frac{2\,a\,b}{a+b}$ which is the product of a and b divided by the mean of a and b.To overcome Cohen’s kappaparadox, we define the newKappa as follows

$$\text{newKappa} = 1 - \frac{(p_o + p_e)\,abs(p_o - p_e)}{2\,p_o p_e}$$

We confirm the effectiness of newKappa, by comparing the agreement classifintion levels betweenoriginal Cohen’s kappa and newKappa on the same examples, using the same criteria for classifying poor, slight, fair, moderate, substantial, perfect for value in the range 0 to 1. In Table6, all examples are 95% observed agreement as in table5. For these observed agreement values, expected agreement values are close and consistent. The newKappa values are consist with classification criteria and human cognition. All case are classified near perctexcpe the last one that is classified as substantial.

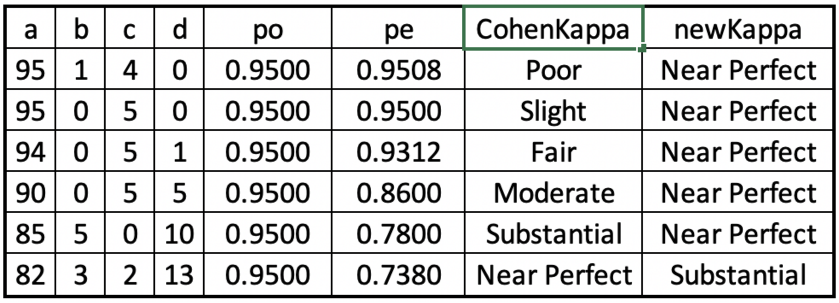
Table 6: all the six examples are 95% observed agreement (column 1).
The table contains observed agreement values, random agreement values,
newKappa values, classification based on newKappa value



a	b	c	d	po	pe	newKappa	interpretation
95	1	4	0	0.9500	0.9508	0.9992	Near Perfect
95	0	5	0	0.9500	0.9500	1.0000	Near Perfect
94	0	5	1	0.9500	0.9312	0.9800	Near Perfect
90	0	5	5	0.9500	0.8600	0.9003	Near Perfect
85	5	0	10	0.9500	0.7800	0.8016	Near Perfect
82	3	2	13	0.9500	0.7380	0.7448	Substantial

FromTable5, Table6, we see thatthe last column in Table 6 is consitentantobserved agreement and cognition, where each row represents the agreement confusion matrix with observed agreeement, and computed expected agreement , newKappavalue & its classification. The matrices describe that the observed agreement values are 95%, alsorandom agreement values are close to observed values, the newKappa values confirm that judgement. Noticein Table7, we show side by side comparison of Cohen’ kappa and new newKappainterpreations, the newKappa interpretations are consistent with cognition, observed agreement values and random agreement values, where as Cohen’s kappa values and interpretations, intable5,are staggared between levels,poor to perfect.

Table 7:Comparison of Cohen’s kappa and newKappaobserved agreement classifications



a	b	c	d	po	pe	CohenKappa	newKappa
95	1	4	0	0.9500	0.9508	Poor	Near Perfect
95	0	5	0	0.9500	0.9500	Slight	Near Perfect
94	0	5	1	0.9500	0.9312	Fair	Near Perfect
90	0	5	5	0.9500	0.8600	Moderate	Near Perfect
85	5	0	10	0.9500	0.7800	Substantial	Near Perfect
82	3	2	13	0.9500	0.7380	Near Perfect	Substantial

IV. CONCLUSION

We have describedCohen’s kappa classication statistic, its strengths and weaknesseswith explicit examples. To overcome these shortcomnings we defined a new statistic newKappa to clarify and simplify the inter-judge agreement reliably. The newKappavalues and their inter pretationis consistent with human cognition for observed agreement and random agreement. The new metric also takes into account the agreement by chance, and more consistent with human cognition. We have shown with tables how the two statistic values and their classifications differ. For comparison between Cohen’s kappa and newKappa, we used the same data and same classification scale. It is confirmed that newKappa isstable, robust and preferable to Cohen’s kappa.There is no paradox such as Cohen’sParadox.We hope thatthis new newKappa will be more useful to the data mining community.

REFERENCES

[1]. Mary L. McHugh,Interrater reliability: the kappa statistic,Biochem Med (Zagreb). 2012 Oct; 22(3):276–282.

[2]. Yinglin Xia, Kappa Statistics, Progress in Molecular Biology and Translational Science, 2020, https://www.sciencedirect.com/topics/medicine-and-dentistry/kappa-statistics

[3]. Audrey Schnell; What is Kappa and How Does It Measure Inter-rater Reliability? https://www.theanalysisfactor.com/author/audreys/

[4]. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. The problems of two paradoxes. J Clin Epidemiol 1990;43:543-9.

[5]. ZACH, accessed July 2021 posted Feb 2021https://www.statology.org/cohens-kappa-statistic/

[6]. Cicchetti DV, Feinstein AR. High agreement but low kappa: II. Resolving the paradoxes. J Clin Epidemiol 1990;43:551-8.

[7]. Anthony J. Viera, MD; Joanne M. Garrett, PhD, Understanding Interobserver Agreement:The Kappa Statistic, Fam Med 2005;37(5):360-3Johnson Clinical Scholars Program, Univ of North Carolina http://web2.cs.columbia.edu/~julia/courses/CS6998/Interrater_agreement.Kappa_statistic.pdf

[8]. Alan Agresti (2013): Categorical data analysis, 3rd Edition, John Wiley & Sons, Inc., New Jersey, SpringerLink link.springer.com

[9]. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159-74.

[10]. Blake Samaha, Measuring Agreement with Cohen’s Kappa Statistic, Towards Data Science, https://towardsdatascience.com/measuring-agreement-with-cohens-kappa-statistic-9930e90386aa

[11]. Cohen J. A coefficient of agreement for nominal scales. Educationaland Psychological Measurement 1960;20:37-46.