Kappa — A Critical Review

Author: Xier Li

Supervisor: Adam Taube

Department of Statistics, Uppsala University

Abstract

The Kappa coefficient is widely used in assessing categorical agreement between two raters or two methods. It can also be extended to more than two raters (methods). When using Kappa, the shortcomings of this coefficient should be not neglected. Bias and prevalence effects lead to paradoxes of Kappa. These problems can be avoided by using some other indexes together, but the solutions of the Kappa problems are not satisfactory. This paper gives a critical survey concerning the Kappa coefficient and gives a real life example. A useful alternative statistical approach, the Rank-invariant method is also introduced, and applied to analyze the disagreement between two raters.

Key words: Kappa coefficient, weighted Kappa, agreement, bias, prevalence, unbalanced situation, Rank-invariant methods

Contents

1.Introduction

1.1 Background

In medical studies, due to different experience of observers and different medical methods or measurements, the results are not precise. In a study, there can be different observers to assign the same subjects into categories, or the investigator lets the observer use different methods or measurements to make judgements. So the study of observer (raters) agreement is very important in many medical applications. For instance, two nurses judge the results of injecting penicillin (allergic or not allergic). Different radiologists assess xeromammograms (normal, benign disease, suspicion of cancer, cancer). Different observers may simply have different perceptions about what the categories mean. Even if there is a common perception, measurement variability can occur. Analyzing the observers' agreement is very meaningful for medical investigations. There are several statistical approaches to assess the agreement between two or more raters. In spite of its shortcomings, the Kappa coefficient is still one of the most popular approaches.

1.2 Aim

The aim of this paper is to give a critical survey concerning the kappa coefficient and draw the attention to some useful alternative statistical approaches.

2 Kappa— a presentation

The kappa statistic first proposed by Cohen(1960), was originally intended to assess agreement between two or more equally skilled observers. In the following, the kappa statistic will be presented.

Table 1. Two observers classify a number of cases according to some finding, say whether a suspicious symptom is present or not, absolute frequencies

Observer A	Observer B		Total
	Yes	No	
Yes	a	b	a+b
No	c	d	c+d
Total	a+c	b+d	n

The first effort for creating a measure of agreement was made by Youden(1950) with the so called Youden Index Y=(a+d)/n. However, this index will have a certain value($Y \neq 0$), even if there is just chance agreement between the observers.

Kappa, is a statistic concerned with the observed agreement on top of chance agreement. The expected proportion of units where the observers give the same results if they are assumed to act independently, is denoted by p_e and can be written:

$$p_e = [(a+b)(a+c)+(b+d)(c+d)]/n^2$$

Let p_o denote the observed proportion of units where the two observers really give identical classifications, $p_o = (a+d)/n$. Then the kappa coefficient is defined as

$$\kappa = (p_0 - p_e)/(1-p_e) \qquad (1)$$

Thus, *Kappa is interpreted as the proportion of agreement between raters after chance agreement has been removed*.

A numerical example is given in Table 2, for diagnosis of 100 cases, where two doctors (or raters) deem whether a surgical operation is needed or not.

Table 2. Diagnosis from two doctors (need operation or not)

Doctor A	Doctor B		Total
	Operation	Not operation	
Operation	76	9	85
Not operation	1	14	15
Total	77	23	100

The proportion of results where diagnosis of Doctor A and Doctor B coincide is $p_o = (76+14)/100 = 0.90$. Assuming that the diagnosis of Doctor A and Doctor B are independent we expect the proportion $p_e = (85\times 77 + 15\times 23)/100^2 = 0.69$ on the diagonal. Hence the kappa is

$$\kappa = (0.90 - 0.69)/(1 - 0.69) = 0.68 .$$

The Kappa is a measure of agreement with the following properties. When $p_o = 1$, it has a maximum value of 1.00 and agreement is perfect. When $p_e = p_o$, it has the value zero, indicating no agreement better than chance. When $p_o = 0$, Kappa has a minimum value of $-p_e/(1-p_e)$ and negative value shows worse than chance agreement.

The maximum Kappa is a modification of Cohen's kappa, obtained by using the maximum possible value of p_o to substitute the value 1 in the denominator of Cohen's calculation of Kappa. [6] Thus,

$$\kappa_{\max} = \frac{p_o - p_e}{\max p_o - p_e} \qquad (2)$$

We change the frequencies according to the marginal distribution. Suppose that b>c, we put (a+c) in the left upper corner, (b-c) in the right upper corner and (d+c) in the right upper corner. In this situation p_o reaches its maximum value.

$$\max p_o = (a+d+2c)/n$$

$$p_o = (a+d)/n$$

$$p_e = [(a+b)(a+c)+(b+d)(c+d)]/n^2$$

Thus, $\kappa_{\max} = \dfrac{(a+d)/n - [(a+b)(a+c)+(b+d)(c+d)]/n^2}{(a+d+2c)/n - [(a+b)(a+c)+(b+d)(c+d)]/n^2}$

Since b>c, then (a+d+2c)/n <1, $\kappa_{\max}$ is always larger than κ. In the example in Table 2, $\kappa_{\max} = (0.900-0.689)/(0.920-0.689) = 0.91$

$\kappa_{\max}$ is the biggest possible Kappa value, given the actual marginal frequencies. It does not solve the actual problem, but a possible approach is to study $\kappa / \kappa_{\max}$.

Landis and Koch(1997) suggested the following table for interpreting kappa values:

Table 3. Recommended labeling of agreement on the basis of numerical kappa values.

Value of kappa	Strength of agreement
<0.20	Poor
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Good
0.81-1.00	Very good

However, these recommendations are just rules of thumb not based on proper scientific reasons. Therefore it is still an open question how the magnitude of Kappa should be judged.

For the example in Table 2, the kappa value is 0.68. We can interpret this as: the two doctors give the same results in 68% of the cases after the coinciding diagnosis' cases due to chance alone have been removed. Furthermore, we can claim that there was "good agreement" between Doctor A and Doctor B, according to Landis and Koch.

Kappa is commonly used in medical studies and its application also extends to psychology and educational research and related fields.

3 The problems of Kappa

When we apply Kappa to investigate agreement, some problems may occur. In order to interpret the problems, some basic concepts are introduced by using Table 3.1

Table 3.1 Two raters and two categories

Observer A	Observer B		Total
	Yes	No	
Yes	a	b	g_1
No	c	d	g_2
Total	f_1	f_2	n

In the above table, f_1 and f_2 are defined as marginal totals of Observer B. g_1 and g_2 are defined as marginal totals of Observer A. When $f_1 = f_2 = g_1 = g_2 = n/2$, the marginal total values are called perfectly balanced. In our practical investigations, when $f_1 \approx f_2$ and $g_1 \approx g_2$ we can normally say they are balanced. If not, the situation is unbalanced.

If Observer A and B have different frequency of occurrence of a condition in an investigation, we say that there is a bias between the observers, and this is not reflected in the Kappa value. Bias Index (BI) is the difference in proportions for the two raters. It can be calculated as:

BI=(a+b)/n-(a+c)/n=(b-c)/n.

We notice that the probabilities b/n and c/n are correlated. However, it is possible to give a confidence interval [BI-2SE(BI), BI+2SE(BI)],

where $SE(BI) = \frac{1}{n}\sqrt{b + c - \frac{(b-c)^2}{n}}$ (See Reference [1] pp. 237)

The prevalence of Observer A for positive ratings is the proportions of "Yes" by him, it can be calculated as (a+b)/n. And for negative ratings is the proportions of "No" , it can be calculated as (c+d)/n. Similarly, the prevalence of Observer B for positive ratings is (a+c)/n, for negative ratings is (b+d)/n.

Prevalence Index (PI) is the difference between the proportions of "Yes" and the proportion of "No". It can be calculated by using the mean prevalence of the observers : [7]

PI=[(a+b)/n+(a+c)/n]/2-[(c+d)/n+(b+d)/n]/2 =(a-d)/n.

Bias and prevalence affect the Kappa value. If PI=0 and BI=0. There are almost no bias and prevalence effect.

There are two types of unbalanced situations: symmetrical unbalanced situation and asymmetrical unbalanced situation. In the unbalanced situations, some problems happen due to bias and prevalence effect. [17][18]Thus, we can not judge the agreements correctly via the Kappa coefficient only. We will discuss the problems in the following parts.

3.1 A graphical illustration

For the study of the Kappa coefficient, we will use the following model. In the Table 3.1.1, imaging there are some cases concerning cancer or non-cancer, the pathologist assigns the cases into positive or negative groups.

Table 3.1.1 diagnostic result

	Cancer	Non-Cancer
Positive	rNP	(1-s)NQ
Negative	(1-r)NP	sNQ
	NP	NQ

We assume there are totally N cases. The proportion of the Cancer cases is P and the Non-cancer is Q, then Q=1-P. Sensitivity is the proportion of positive cases among

the true cancer cases, and is denoted by r. Specificity is the proportion of negative cases among the true non- cancer. It is denoted by s.

Suppose there are two pathologists, and Pathologist A and B, suppose that they have the same r and s. This means that they are equally skilled. We remember that the Kappa was originally created for the study of agreement between two equally skillful observers. Then for the Cancer cases, we can obtain Table 3.1.2. And for the non-cancer case, we obtain Table 3.1.3.

Table 3.1.2 Cancer Cases

	Positive	Negative	Total
Positive	rrNP	r(1-r)NP	rNP
Negative	r(1-r)NP	(1-r)(1-r)NP	(1-r)NP
Total	rNP	(1-r)NP	NP

Table 3.1.3 Non-Cancer Cases

	Positive	Negative	Total
Positive	(1-s)(1-s)NQ	s(1-s)NQ	(1-s)NQ
Negative	s(1-s)NQ	ssNQ	sNQ
Total	(1-s)NQ	sNQ	NQ

According to Table 3.1.2 and Table 3.1.3, we can obtain Table 3.1.3 for Pathologist A and Pathologists B.

Table 3.1.3 Table for two pathologists.

Pathologists B	Pathologist A		
	Positive	Negative	Total
Positive	$r^2P+(1-s)^2Q$	$r(1-r)P+(1-s)sQ$	$Q-sQ+rP$
Negative	$r(1-r)P+(1-s)sQ$	$(1-r)^2P+s^2Q$	$P+sQ-rP$
Total	$Q-sQ+rP$	$P+sQ-rP$	1

Assume $a=r^2P+(1-s)^2Q$ $b=r(1-r)P+(1-s)sQ$ $c=r(1-r)P+(1-s)sQ$ $d=(1-r)^2P+s^2Q$ For the two pathologists, they have the same marginal distribution, the marginal total $f_1=g_1$ $f_2=g_2$ then the observed proportion is $p_o=a+b=r^2P+(1-s)^2Q+(1-r)^2P+s^2Q$ Then the observed proportion is $p_e=f_1g_1+f_2g_2=(Q-sQ+rP)^2+(P+sQ-rP)^2$ we can obtain

$$\kappa=\frac{p_o-p_e}{1-p_e}=\frac{r^2P+(1-s)^2Q+(1-r)^2P+s^2Q-(Q-sQ+rP)^2-(P+sQ-rP)^2}{1-(Q-sQ+rP)^2-(P+sQ-rP)^2}$$

In order to study the relationship of r, s, P , Q and κ, when r=s, let r=s=0.80 and

r=s=0.95, we obtain the curves in Figure 3.1.1, and we can see both of the curves are symmetrical, and when P=0.5, both of the κ reach their maximum value: 0.36 (r=s=0.80) and 0.81(r=s=0.95). We also notice that both of these situations are balanced situations. In the case with r=s=0.95, we also notice that the Kappa value correspond to "Good" for a wide range of P.

When $r \neq s$, let r=0.90, s=0.70 and r=0.95, s=0.80, we can obtain Figure 3.1.2. The two curves are asymmetrical and when P=0.6, both of the κ reach their maximum value: 0.385(r=0.90, s=0.70) and 0.593 (r=0.95, s=0.80). When P=0.33 and r=0.90, s=0.70 the situation is balanced, and κ is 0.32. When P=0.04 and r=0.95, s=0.80 the situation is balanced, and κ is 0.12.

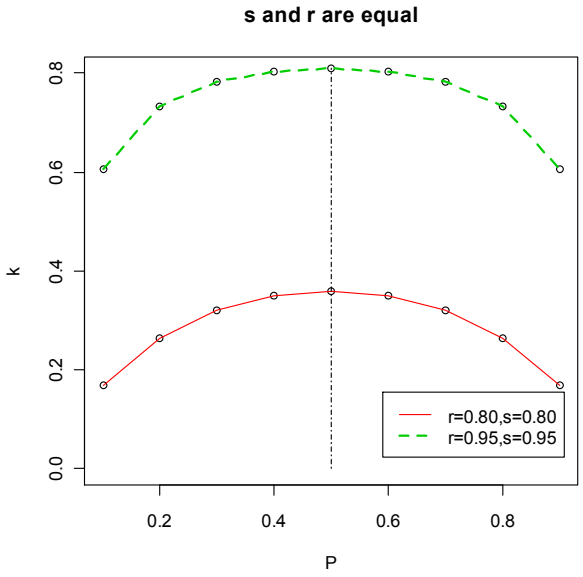

Figure 3.1.1 r=s

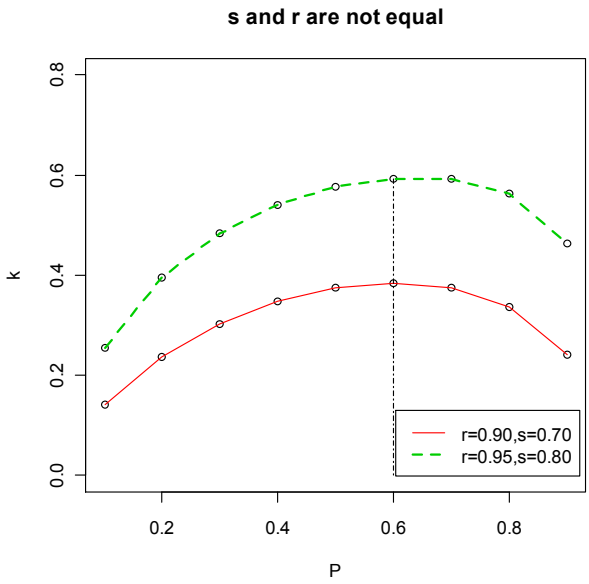

Figure 3.1.2 $r \neq s$

3.2 Problems in symmetrical unbalance

When both $f_1 > f_2$ and $g_1 > g_2$, or both $f_1 < f_2$ and $g_1 < g_2$, we call this situation symmetrically unbalanced. We can imagine the two observers classify a number of slides as cancer or not. In Table 3.2 about half the slides come from cancer patients. In Table 3.3 between 80% and 90% slides cases from cancer patients.

In Table 3.2, f_1=46, f_2=54 and g_1=49 g_2=51, this situation is approximately balanced. While in Table 3.3, f_1=85 f_2=15 and g_1=90 g_2=10, it indicates $f_1 > f_2$ and $g_1 > g_2$. So the situation is symmetrically unbalanced.

Both examples have the same p_o=0.85, while in the Table 3.2, κ=0.70. In Table 3.3, κ=0.32, which is much smaller than the kappa value in Table 3.2. For the example in Table 3.3, it has high p_o but its κ is relatively low, which makes it

difficult to assessing agreement via κ. This problem in the symmetrically unbalanced situation was called the "First paradox" by Feinstein and Cicchetti: "*If p_e is large, the chance correction process can convert a relatively high value of p_o into a relatively low value of Kappa.*" [5]

We find p_e in the two examples are 0.50 and 0.78 respectively. From the formula $\kappa = \frac{p_o - p_e}{1 - p_e} = 1 - \frac{1 - p_o}{1 - p_e}$, when p_o is fixed, the increase of p_e, 1- p_e decreases, $(1 - p_o)/(1 - p_e)$ increases, then κ increase.This paradox attributes to high prevalence. For the example in Table 3.3, BI=0.05, PI=0.75.The PI value tells us there is the prevalence effect. The low value of κ (0.32) is due to the prevalence effect.

Table 3.2

Observer A	Observer B		Total
	Yes	No	
Yes	40	9	49
No	6	45	51
Total	46	54	100

Table 3.3

Observer A	Observer B		Total
	Yes	No	
Yes	80	10	90
No	5	5	10
Total	85	15	100

3.3 Problems in asymmetrical unbalance

When f_1 is much lager than f_2, while g_1 is much smaller than g_2, or vice versa, this causes asymmetrical unbalanced marginals. In this situation, for the same p_o, κ will be higher than in the symmetrical imbalance situation. This was called "Second paradox" by Feinstein and Cicchetti: "*Unbalanced marginal totals produce higher values of κ than more balanced total.*"[5]

In order to explain this paradox, two examples in the following Table 3.4 and Table 3.5 are used. In Table 3.4, f_1=70, g_1=60 and f_2=30, g_2=40, f_1 and g_1 are both larger than n/2, f_2 and g_2 are both smaller than n/2, this is a symmetrically unbalanced situation. In Table 3.5, f_1=30, g_1=60 and f_2=70, g_2=40, f_2 and g_1 are both larger than n/2, f_1 and g_2 are both smaller than n/2. f_1 is much smaller than f_2, while g_1 is much lager than g_2, this is asymmetrical unbalance situation. We can see the marginal total is "worse" than in Table 3.4.

Both of the examples have the same p_o (=0.60), in Table 3.4, p_e=0.54, BI=-0.10, PI=0.30 and κ=0.13, while in Table 3.5, p_e=0.46, BI=0.30, PI=-0.1, although the asymmetrical unbalance is apparent, it has a higher value of κ than Table 3.5, which is 0.26. As bias increases, p_e declines and κ increases, prevalence increases, p_e

increases and κ declines. So in Table 3.5, BI is larger and PI is smaller, all of the bias and prevalence cause small value of p_e, finally κ is higher in Table 3.5 than in Table 3.4.

Table 3.4

Observer A	Observer B		Total
	Yes	No	
Yes	45	15	60
No	25	15	40
Total	70	30	100

Table 3.5

Observer A	Observer B		Total
	Yes	No	
Yes	25	35	60
No	5	35	40
Total	30	70	100

3.3 Solution of the two problems

In order to solve these two paradoxes caused by unbalanced marginal totals, Feinstein and Cicchetti suggested that if we want to use kappa, we can also use p_{pos} and p_{neg} as two separate indexes of proportionate agreement in the observers' positive and negative decisions.

For positive agreement, $p_{pos} = a/[(f_1 + g_1)/2] = 2a/(f_1 + g_1)$, and for negative agreement, $p_{neg} = d/[(f_2 + g_2)/2] = 2d/(f_2 + g_2)$.[6] By using the two indexes p_{pos} and p_{neg} together with kappa can only avoid the incorrect judgement by Kappa value, but it cannot solve the problems caused by prevalence and bias effects.

Byrt *et al* use the Prevalence-adjusted bias-adjusted kappa(PABAK) to adjusts Kappa in those imbalance situations.[7] Replace b and c by their average for differences in prevalence, x=(b+c)/2, and also replace a and d by their average for bias between observers, y=(a+d)/2. Then we can get a balanced table (Table 3.6) by adjustment.

Table 3.6

Observer A	Observer B		Total
	Yes	No	
Yes	x	y	x+y
No	y	x	x+y
Total	x+y	x+y	2(x+y)

In this table p_e=0.5, using the formula (1) for Kappa, we can obtain

$$\text{PABAK}=\frac{[x/(x+y)]-0.5}{1-0.5}=2p_o-1$$

Then the prevalence and bias effects are adjusted and it is related to p_o. When x=0,

PABAK has its minimum value -1. When y=0, it reaches its maximum +1. When p_o =0.5, its value becomes 0.

The following formula shows the relationship of Kappa and PABAK [7]

$$\kappa = \frac{PABAK - PI^2 + BI^2}{1 - PI^2 + BI^2}$$

From this formula we can see that, if PI is constant, the larger of the absolute value of BI, the larger is Kappa. If BI is constant, the larger of the absolute value of PI, the smaller is Kappa.

Table 3.7 lists the examples we used, For Table 3.2, κ is 0.70 , p_o is 0.85, then we check the p_{pos} and p_{neg}, both of them are high, and the values of BI and PI are also small. These indexes are identical and show good agreement. For Table 3.3, κ is 0.32, while p_o is 0.85, relatively high. Then we need check the p_{pos} and p_{neg}, p_{pos} is 0.90 and p_{neg} is 0.40. So we can say whether it is suitable to judge the agreement by κ. Then we check the BI and PI, we find PI is high, the prevalence effect occurs. The next step is adjusting κ by using PABAK. The PABAK is 0.70, so from the PABAK, we can judge the agreement is good.

For Table 3.4 and Table 3.5, we use the same approach. The marginal totals in Table 3.4 are symmetrically distributed and in Table 3.5 is asymmetrically distributed, while κ is smaller than Table 3.5 When we use PABAK, the PABAK in Table 3.4 is larger than in Table 3.5. So we can use PABAK to assess the agreement in such imbalanced situations.

Table 3.7

	p_o	p_{pos}	p_{neg}	BI	PI	k	PABAK
Table 3.2	0.85	0.93	0.86	0.03	-0.05	0.70	0.70
Table 3.3	0.85	0.91	0.40	0.05	0.75	0.32	0.70
Table 3.4	0.60	0.69	0.43	-0.10	0.30	0.13	0.20
Table 3.5	0.60	0.56	0.64	0.30	0.10	0.26	0.19

From the discussion above, for the unbalanced situations, we cannot assess the agreement by one index Kappa only, we should also consider other relative indexes, such as p_{pos} and p_{neg}. If p_{pos} and p_{neg} are relatively high, while κ are small, or vice versa. We can check the PI and BI to see the prevalence and bias effect, then use PABAK to access the agreement. Only by using a single index, it will make the judgement of agreement difficult.

From the above discussion, due to bias and prevalence effects, we cannot make right judegements by Kappa value. The bias and prevalence effects are not reflected in the Kappa value. Although some other indexes can be used together with Kappa, the problems seems still not to have been solved satisfactorily.

4 Non dichotomous variables

4.1 Two ordinal classification methods

The kappa coefficient can be extended to the case when the two methods (or raters) produce classifications according to an ordinal scale. The approach will be explained by the following example in Table 4.

Table 4 classification according to two different methods, hypothetical data

	Method B			
Method A	1	2	3	Total
1	14	23	3	40
2	5	20	5	30
3	1	7	22	30
Total	20	50	30	100

The observed proportion is $p_o = (14 + 20 + 22)/100 = 0.56$. Assuming the two methods are independent we expect $p_e = (20 \times 40 + 50 \times 30 + 30 \times 30)/100^2 = 0.35$. Hence the kappa is $\kappa = (0.56 - 0.32)/(1 - 0.32) = 0.35$

This approach can be extended to variables with many ratings in nominal scales. When we change the postion of the category in row or in column, the Kappa value is still the same. This means that the Kappa value does not utilize the information given by the order between the different classification alternatives.

4.2 Weighted Kappa

One of the undesirable properties of Kappa is that all the disagreements are treated equally. So it is preferable to give different weights to disagreement according to each cell's distance from the diagonal. But the weights can only be given to ordinal data, not to nominal data. The reason is that if we change the position of the category in row or in column, the weights will be different.

The weighted kappa is obtained by giving weights considering disagreement. It was first proposed by Cohen (1968). The weights are given to each cell according to its distance from the diagonal. Suppose that there are k categories, i=1,…,k; j=1,..,k. The weights are denoted by w_{ij}. They are assigned to each cell and their value range is $0 \le w_{ij} \le 1$. The cells in the diagonal (i=j) are given the maximum value, $w_{ii} = 1$. For the other cells' ($i \neq j$), $w_{ij} = w_{ji}$ and $0 \le w_{ij} < 1$.

The observed weighted proportion of agreement are obtained as

$$p_{o(w)} = \sum_{i=1}^{k} \sum_{j-1}^{k} w_{ij} p_{ij}$$

Where p_{ij} is the proportion of the cell in ith row and jth column, then it is given weight w_{ij}. The observed weighted proportion of agreement is the sum of all the

proportion of cells given weights.
Similarly, the chance-expected weighted proportion of agreement is,

$$p_{e(w)} = \sum_{i=1}^{k} \sum_{j-1}^{k} w_{ij} p_{i.} p_{.j}$$

It is the sum of all the expected proportion of the cells given weights.
And weighted kappa is then given by

$$\hat{\kappa}_w = \frac{p_{o(w)} - p_{e(w)}}{1 - p_{e(w)}}$$

It is identical to the Kappa given in formula (1), and the interpretation of the weighted Kappa values is the same, both are according to Landis and Koch in Table 3.
The choice of weights is somewhat arbitrary. Two kinds of weights are normally used, one is suggested by Bartko(1966), where the formula of the weights are:

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2}$$

The other is suggested by Cicchetti and Allison(1971), where the weights are taken as

$$w_{ij} = 1 - \frac{|i-j|}{k-1}$$

We illustrate the weighting procedure by means of data given in Table 5.

Table 5 Two neurologists classify 149 patients in to 4 categories

	Neurologist A					
		1	2	3	4	Total
Neurologist B	1	38	5	0	1	44
	2	33	11	3	0	47
	3	10	14	5	6	35
	4	3	7	3	10	23
	Total	84	37	11	17	149

Calculate the weights suggested by Bartko, we can get $w_{11} = w_{22} = w_{33} = w_{44} = 1$, $w_{12} = w_{21} = 8/9$, $w_{13} = w_{31} = 5/9$, $w_{14} = w_{41} = 0$, $w_{23} = w_{32} = 8/9$ $w_{24} = w_{42} = 5/9$, $w_{34} = w_{43} = 8/9$, then $p_{o(w)}$=0.875, $p_{e(w)}$=0.736, so κ_w=0.53 and κ=0.21. From the value of κ_w, according to Table 3, we can deem the agreement of the two neurologists are "moderate".
Calculate the weights suggested by Cicchetti and Allison, $w_{11} = w_{22} = w_{33} = w_{44} = 1$, $w_{12} = w_{21} = 2/3$, $w_{13} = w_{31} = 1/3$, $w_{14} = w_{41} = 0$, $w_{23} = w_{32} = 2/3$, $w_{24} = w_{42} = 1/3$ $w_{34} = w_{43} = 2/3$, then $p_{o(w)}$=0.754 $p_{e(w)}$=0.603, so κ_w=0.38 and κ=0.21. We can also know there is a "Fair" agreement between the two neurologists.

The value of weighted kappa depends on the choice of weights but the choice of weights is subjective. So the weighted kappa can be different in the same investigation, but the unweighted kappa is not changed. Sometimes an investigator can choose the weights appropriate in a particular situation. Weighted kappa is usually higher than unweighted kappa because disagreements are more likely to be by only one category than by several categories. [1]

4.3 Multiple ratings per subject with different raters

In medical studies, it is very common to have more than 2 raters and ratings. So studying Kappa in multiple ratings per subject with different raters is quite useful.
Suppose that a sample of N subjects has been studied, n is the number of ratings per subject, and k is the number of the categories.(i=1,…,N; j=1,…,k). Define n_{ij} to be the number of raters who assigned the ith subject to the jth category, n is the number of raters, as indicated in Table 4.2. The approach is given by Fleiss (1971). [8]

Table 4.2

	Category					
Subject	1	2	…	j	…	k
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1k}
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2k}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{ik}
. . .	. . .	. . .	. . .	. . .	. . .	. . .
N	n_{N1}	n_{N2}	…	n_{Nj}	…	n_{Nk}

The proportion of all assignments to jth category can be calculated as $p_j = \frac{1}{Nn}\sum_{i=1}^{n} n_{ij}$, where Nn is the total number of all assignments, n_{ij} is the number of ratings for the ith subject assigned into jth category. Since $\sum_j n_{ij} = n$, $\sum_j p_j = 1$

For the ith subject there are totally n(n-1) possible paired assignments, so the proportion of the agreement of the ith subject is

$$P_i = \frac{1}{n(n-1)}\sum_{j=1}^{k} n_{ij}(n_{ij}-1) = \frac{1}{n(n-1)}\left(\sum_{j=1}^{k} n_{ij}^{\,2} - n\right)$$

The overall proportion of agreement can be measured by the mean of P_i

$$\overline{P} = \frac{1}{N}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{i=1}^{k} n_{ij}^{\,2} - Nn\right)$$

The proportion of all assignments to the jth category is p_j. If the raters made their assignments at random, the expect mean proportion of chance agreement is

$$\overline{P}_e = \sum_{j=1}^{k} p_j^{\,2}$$

Analogous to the original Kappa formula, we can obtain the overall agreement, as

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{k} n_{ij}^{\ 2} - Nn[1 + (n-1)\sum_{j-1}^{k} p_j^2]}{Nn(n-1)(1 - \sum_{j=1}^{k} p_j^2}$$

We can also obtain the Kappa coefficient for the category j, denoted by κ_j. For the jth catergory, the proportion of agreement is $\overline{P}_j = \frac{\sum_{i=1}^{N} n_{ij}(n_{ij}-1)}{\sum_{i=1}^{N} n_{ij}(n-1)} = \frac{\sum_{i=1}^{N} n_{ij}^2}{Nn(n-1)p_j}$

After chance agreement in the category is removed, we can obtain the Kappa for the category j, as

$$\kappa_j = \frac{\overline{P}_j - p_j}{1 - p_j} = \frac{\sum_{i=1}^{N} n_{ij}^2 - Nnp_j[1 + (n-1)p_j]}{Nn(n-1)p_j q_j}$$

where $p_j = \frac{1}{Nn}\sum_{i=1}^{n} n_{ij} \quad q_j = 1 - p_j$

In fact, the measure of overall agreement is a weighted average of κ_j.

$$\kappa = \sum_j p_j q_j \kappa_j \,/ \sum_j p_j q_j$$

Fleiss, Nee and Landis (1979) derived the following formulas for the approximate standard errors of κ and κ_j, for testing the hypothesis that the underlying value is zero:[2]

$$\hat{se}(\kappa) = \frac{\sqrt{2}}{\sum_{j=1}^{k} p_j q_j \sqrt{Nn(n-1)}} \times \sqrt{(\sum_{j=1}^{k} p_j q_j)^2 - \sum_{j=1}^{k} p_j q_j (q_j - p_j)} \quad \text{where } q_j = 1 - p_j$$

and $se_o(\kappa_j) = \sqrt{\frac{2}{Nn(n-1)}}$

5 A real life example

Six pathologists in Sweden (1994) had to classify biopsy slides from 14 suspected prostate cancer patients.[9] By an expert panel, the 1 to 7 slides from the patients were found to be “non cancer”, the other 8 to 14 were classified as cancer (the “golden standard”). The slides are classified in 6 categories, 1 to 6, the highest rating indicating cancer. The pathologists had to make their assessment by means of two different methods: [19]

Method I : Fine needle aspiration(FNAB) slides were examined by light microscopy as it is normally done in a routine setting, which means that the examiner has free

access to all magnifications and all parts of the smear.

Method Π: The slides were examined as digitalized images, each slide represented by three images, each at different magnification. This means that only a limited fraction of all smeared cells were available for judgement. The images were chosen to be representative of a negative or a positive diagnosis of cancer.

Now, we want to assess the agreement of the 6 pathologists by using the two methods. The following Table 5 shows the ratings of the 14 slides given by 6 pathologists using Method Ⅰ.

Table 5 The results by using Method Ⅰ

6 ratings on each of 14 subjects into one of 6 categories						
Subject	1	2	3	4	5	6
1	1	4	1	0	0	0
2	3	2	0	1	0	0
3	2	3	1	0	0	0
4	0	2	0	0	4	0
5	1	1	1	1	2	0
6	5	1	0	0	0	0
7	0	3	3	0	0	0
8	0	1	2	2	1	0
9	0	0	0	1	0	5
10	0	0	0	1	0	5
11	0	0	0	0	0	6
12	0	0	0	0	3	3
13	0	0	1	1	1	3
14	0	0	0	1	1	4
Total	12	17	9	8	12	26
κ_j	0.37	0.16	0.08	-0.05	0.22	0.60

By using the approach given by Fleiss, in Table 5, there are k=6 rating are given to N=14 slides by n=6 pathologists. Then the proportion of all assignment to each 6 category: $p_1 = 12/(14\times 6) = 0.143$ $p_2 = 17/(14\times 6) = 0.202$ $p_3 = 9/(14\times 6) = 0.107$ $p_4 = 9/(14\times 6) = 0.107$ $p_5 = 12/(14\times 6) = 0.142$ $p_6 = 26/(14\times 6) = 0.309$

The proportion of overall agreement in this example is

$$\overline{P} = \frac{1}{n(n-1)}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{i=1}^{k} n_{ij}^{\ 2} - Nn\right) = \frac{1}{14\times 6\times 5}(262-14\times 6) = 0.438$$

The mean proportion of chance agreement in this example is

$$\overline{P}_e = \sum_{j=1}^{k} p_j^{\ 2} = 0.143^2 + \ldots + 0.309^2 = 0.198$$

Then we can obtain the overall kappa,

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1-\overline{P}_e} = \frac{0.438-0.198}{1-0.198} = 0.30$$

The standard error of the overall Kappa of the six pathologists by Method I is $s\hat{e}_0(\kappa) = 0.03$ And $z = \kappa / s\hat{e}_0(\kappa) = 10.00$ which indicates the overall Kappa value is significant. The standard error of the Kappa in each category is $s\hat{e}_0(\kappa_j) = 0.08$

By using Method Ⅰ, from the value $\overline{P}$=0.424, we can know that if a slide be selected randomly and rating by a randomly selected pathologist , the rating of the second slide would agree with the first over 42.4% of the time. The last row of Table 5 shows the agreement of the 6 ratings by 6 pathologists. Kappa values of the rating 4 has the smallest value -0.05, which means poor agreement, while for rating 6, Kappa is 0.60, which means good agreement. The kappa in rating 1 has the value of 0.37, which is the second highest value. It seems the lowest rating and highest rating have "better" agreement among the 6 pathologists than other ratings. All of the individual Kappa values are not significantly different from zero. For the 6 pathologists, the overall Kappa is 0.28, and it is significantly different from zero. According to Table 3, it means they have "fair" agreement by using Method Ⅰ.

The following Table 6 displays the results by using Method Ⅱ.

Table 6 The results by using Method Ⅱ

6 ratings on each of 14 subjects into one of 6 categories						
Subject	1	2	3	4	5	6
1	1	3	0	1	1	0
2	1	1	3	0	0	1
3	2	1	3	0	0	0
4	3	0	2	1	0	0
5	1	0	1	0	4	0
6	1	1	2	0	1	1
7	1	1	2	0	1	1
8	3	1	1	0	0	1
9	0	0	3	2	0	1
10	1	1	1	0	2	1
11	1	0	2	0	1	2
12	1	0	2	1	1	1
13	0	4	0	0	2	0
14	0	1	3	1	1	0
Total	16	14	25	6	14	9
κ_j	-0.22	0.11	-0.04	-0.01	0.07	-0.07

By using the approach given by Fleiss, in Table 6, there are k=6 rating are given to N=14 slides by n=6 pathologists. Then $p_1 = 12/(14\times 6) = 0.190$ $p_2 = 14/(14\times 6) = 0.167$ $p_3 = 25/(14\times 6) = 0.298$ $p_4 = 6/(14\times 6) = 0.071$ $p_5 = 14/(14\times 6) = 0.167$ $p_6 = 9/(14\times 6) = 0.107$

The proportion of overall agreement in this example is

$$\overline{P} = \frac{1}{n(n-1)}\sum_{i=1}^{N} P_i = \frac{1}{Nn(n-1)}\left(\sum_{i=1}^{N}\sum_{i=1}^{k} n_{ij}^{\ 2} - Nn\right) = \frac{1}{14\times 6\times 5}(170-14\times 6) = 0.205$$

The mean proportion of chance agreement in this example is

$$\overline{P}_e = \sum_{j=1}^{k} p_j^{\ 2} = 0.197$$

Then we can obtain the overall kappa,

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} = \frac{0.205 - 0.197}{1 - 0.197} = 0.01$$

The standard error of the overall kappa of the six pathologists by Method II is $s\hat{e}_0(\kappa) = 0.17$. And $z = \kappa / s\hat{e}_0(\kappa) = 0.06$ indicates the overall kappa value is not significant from zero. The standard error in each category is $s\hat{e}_0(\kappa_j) = 0.08$

Similarly, for Method Π, the value $\overline{P}$ =0.205, which means if a slide be selected randomly and rating by a randomly selected pathologist , the rating of the second slide would agree with the first over 20.5% of the time. Kappa in each categories are very small and in overall is 0.01, which in reality here means no agreement among the six pathologists. All of the individual Kappa values except rating 1 are not significantly different from zero and the others are significantly different from zero. For the overall kappa, it indicates poor agreement among the six pathologists and it is not significant different from zero.

We can also obtain the kappa coefficient of the six pathologists by the two methods, respectively. The values of kappa can be seen in the following Table 7. All of the kappa values are small and statistically insignificant, they just tell us there is no agreement between the two methods.

	Pathologist					
	A	B	C	D	E	F
κ	0.01	-0.19	0.05	-0.08	0.06	0.01

Comparing the overall Kappa values of the each method, we can only know that the 6 pathologist have better agreement by using Method I than Method Π. There is no agreement between the two methods. But we can not say which method is better than the other via the Kappa value.

In order to find the systematic interobserver difference, we use a special kind of Receiver Operating Characteristic (ROC) curve, which will be introduced next.

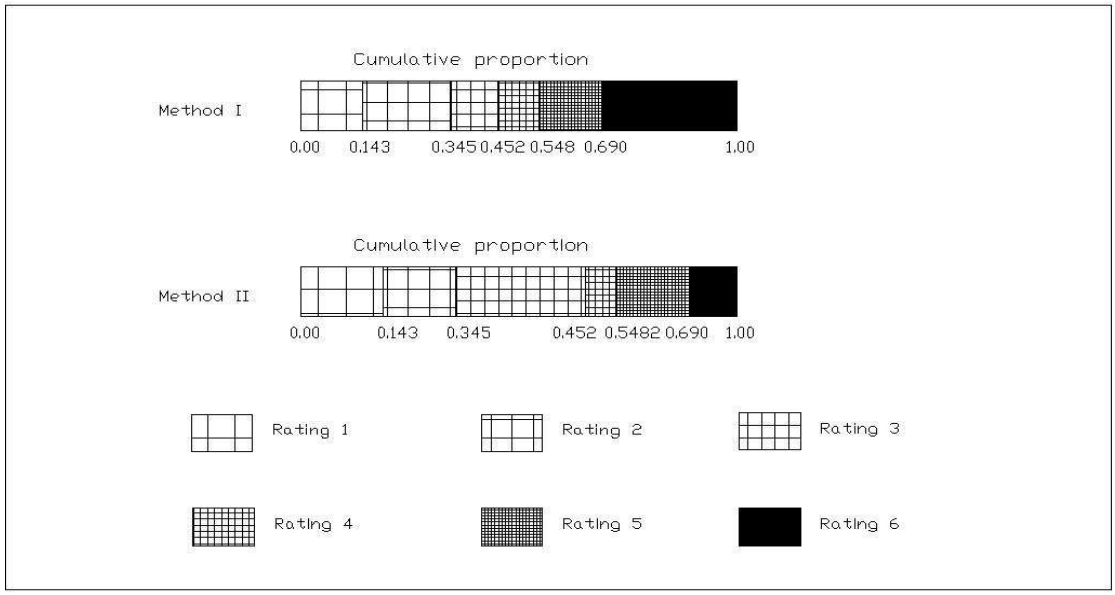

Figure 5.1 The cumulative proportion of Method I and Method II

Let the cumulative frequencies of Method I be the x-coordinates and the cumulative frequencies of Method II be the y-coordinates. If the two distributions are identical, the ROC curve will coincide with the diagonal. From the cumulative proportion in Figure 5.1, we can plot the ROC curve of the two methods, which is shown in the following Figure 5.2.

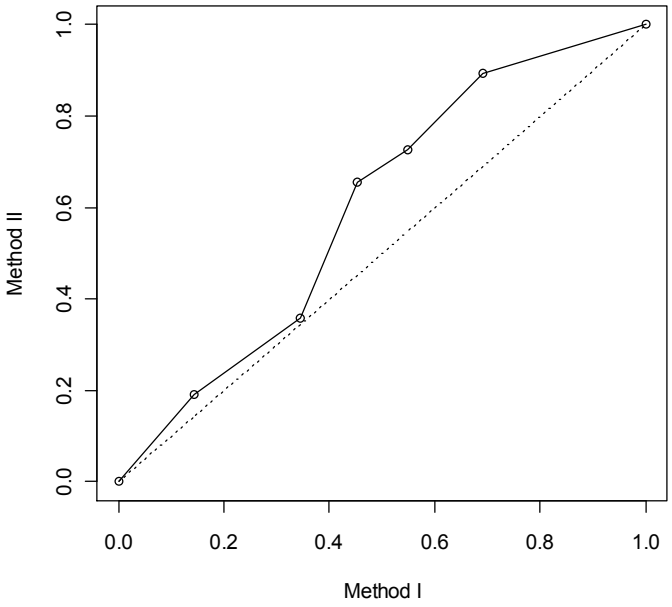

Figure 5.2 ROC curve for systematic disagreement between two methods

From Figure 5.2, we can see the ROC curve falls into the upper left area, it indicates a systematic difference. We can also know the six pathologists classified more slides into low ratings by Method II than by Method I.

6 An example of application of the Rank-invariant method

6.1 Introduction to the Rank-invariant method

The Rank-invariant method was proposed by Svensson(1993). [10] It provides a new way to deeper the study of the agreement of two raters using ordinal scales. By plotting the cumulative proportions of the two marginal distributions, we can get the Relative Operating Characteristic (ROC) curves.[10] Imaging there are two groups of 20 patients, and two doctors classify each group of 20 patients into 4 categories, the results are shown in the Table 6.1 and Table 6.2.

Table 6.1 Diagnosis results of Group A

	Doctor A				
Doctor B	1	2	3	4	cumulative proportion
1	0	0	4	2	0.3
2	0	4	0	0	0.5
3	4	2	0	0	0.8
4	0	2	2	0	1.0
cumulative proportion	0.2	0.6	0.9	1.0	

Table 6.1 Diagnosis results of Group B

	Doctor A				
Doctor B	1	2	3	4	cumulative proportion
1	0	2	0	2	0.1
2	0	2	0	2	0.3
3	0	4	2	0	0.6
4	8	0	0	0	1.0
cumulative proportion	0.4	0.7	0.9	1.0	

Then we can plot the ROC curve in Figure 6.1

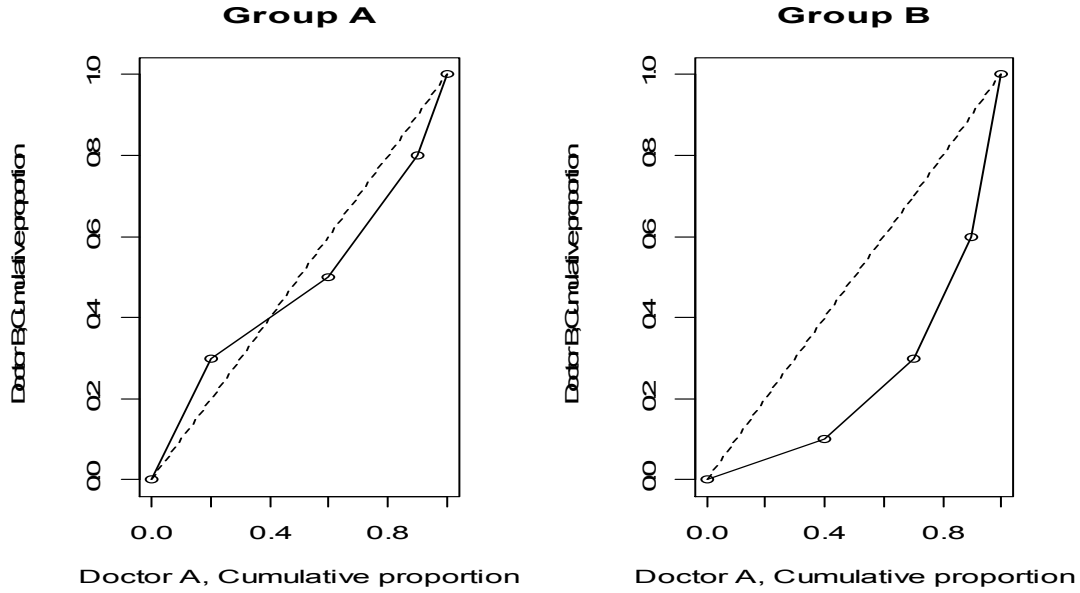

Figure 6.1 ROC curves of two groups

Figure 6.1 shows different shapes of ROC curves. The shapes of ROC curves can tell us something about the systematic change. The systematic difference between the two raters means different marginal distributions. There are two types of systematic disagreement: systematic disagreement in position and systematic disagreement in concentration. If the ROC curve falls into the upper left triangle area or the lower right triangle area, it indicates systematic disagreement in position on the scale. If it is an S-shape ROC curve, it indicates systematic disagreement in concentration of the categories. If the ROC curve falls into the diagonal, it indicates there is no systematic disagreement in position or in concentration. [10] The ROC curve in the example in

Table 6.1 indicates systematic disagreement in concentration and in Table 6.2 it indicates systematic disagreement in position.

A measurement called relative position (RP) was introduced by Svensson. RP is the difference between the probability of the classifications being shifted towards higher categories, and the probability of the classifications being shifted towards lower categories given the actual marginal frequencies. Relative concentration (RC) is the difference between the probability of the marginal distribution of Observer A being concentrated relative to Observer B and vice versa. [10] The empirical formulae of RP and RC, are given by Svensson and here are shown in Appendix 1.1. [12] Both RP and RC values range from -1 to +1.

For the example in Table 6.1, the probability of the classifications being shifted towards higher categories and lower categories are:

$p_{xy} = (4\times 4 + 6\times 12 + 4\times 18)/20^2 = 0.4 \qquad p_{yx} = (8\times 6 + 6\times 10 + 2\times 16)/20^2 = 0.35$

then the difference between the probabilities is RP= $p_{xy} - p_{yx} = 0.4 - 0.35 = 0.05$.

The probabilities of being concentrate are:

$p_{xyx} = \frac{1}{20^3}\times 272 = 0.034 \quad p_{yxy} = \frac{1}{20^3}\times 720 = 0.09$

M=min [($p_{xy} - p_{xy}^{\ 2}$),($p_{yx} - p_{yx}^{\ 2}$)]=0.2275 then the systematic change in concentration can be calculated as $RC = \frac{1}{M}(p_{xyx} - p_{yxy}) = -0.25$

RP=0.05and RC=-0.25, indicate there are systematic differences both in position and concentration in Table 6.1.

For the example in Table 6.2, RP=0.5, RC=0, it indicates there is systematic difference in position.

There are two kinds of interobserver disagreement: systematic disagreement and random disagreement. The random disagreement can be measured by Relative Rank Variance (RV). Ranking is to place an individual to his position in the population in relation to other individuals. [10] Svensson applies a special ranking procedure ("augmented" ranking), based on both two variables. (See Reference [10]) Each individual is given to two rank values, and when the values are different, it indicates individual dispersion of change from the common systematic group changes. The empirical formulae of RV given by Svensson are shown in Appendix 1.2. RV value ranges between 0 and 1. In Table 6.1, for the cell (3,2), the mean ranks are

$\overline{R}_{3,2}^{\ (x)} = 4+4+0.5(2+1)=9.5 \qquad \overline{R}_{3,2}^{\ (y)} = 4+2+4+4+0.5(2+1) = 15.5$

And $RV = \frac{6}{n^3}\sum_{i=1}^{m}\sum_{j=1}^{m} x_{ij}[\overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)}] = \frac{6\times 1064}{20^3} = 0.798$, which indicates that there is random disagreement.

For the absolute values of RP, RC and RV, the higher of the values are, the stronger is, the contribution of the specific disagreement to the observed disagreement.

6.2 An Application

Let us consider the data from diagnosis of multiple sclerosis reported in Westlund and Kurland (1953). There are 149 patients from Winnipeg and 69 patients from New Orleans. They were examined by two neurologists, Neurologist A and Neurologist B. Each of neurologists was requested to classify the patients in Winnipeg and New Orleans in to one of the following diagnostic class:

1. Certain multiple sclerosis
2. Probable multiple sclerosis
3. Possible multiple sclerosis
4. Doubtful, unlikely, or definitely not multiple sclerosis

The results of diagnosis are present in Table 6.1.

Table 6.1 Diagnostic classification regarding multiple sclerosis

Winnipeg	Neurologist A					
		1	2	3	4	Total
Neurologist B	1	38	5	0	1	44
	2	33	11	3	0	47
	3	10	14	5	6	35
	4	3	7	3	10	23
	Total	84	37	11	17	149

New Orleans	Neurologist A					
		1	2	3	4	Total
Neurologist B	1	5	3	0	0	8
	2	33	11	3	0	18
	3	10	14	5	6	22
	4	3	7	3	10	21
	Total	11	29	11	18	69

The distribution of the 149 patients in Winnipeg and 69 patients in New Orleans on the 4 classes are shown in Figure 6.1.Then according to the cumulative proportions, we can draw the ROC curves, which are shown in Figure 6.2.

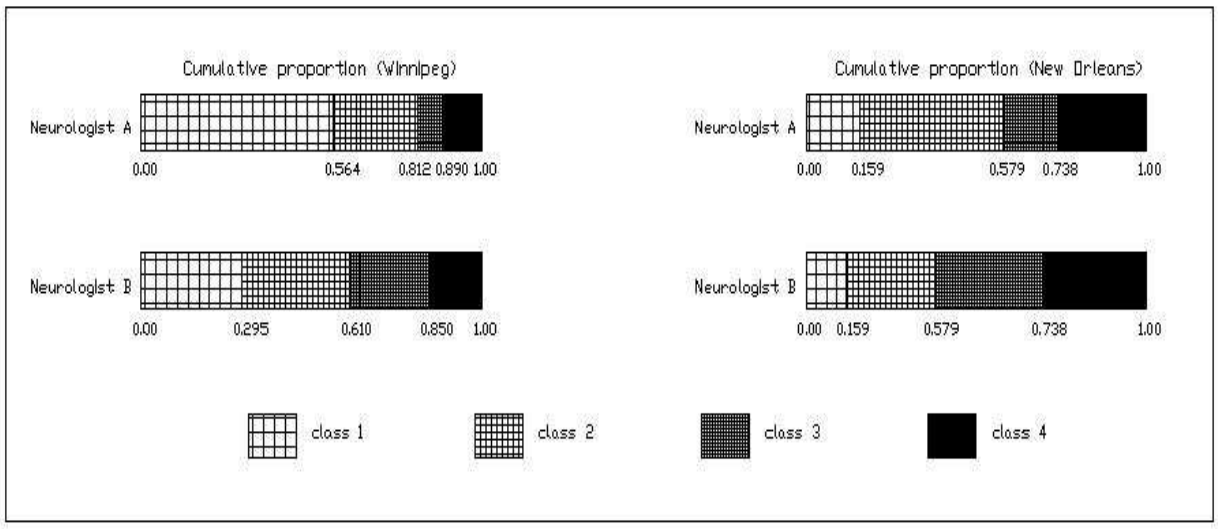

Figure 6.1 The cumulative relative frequency

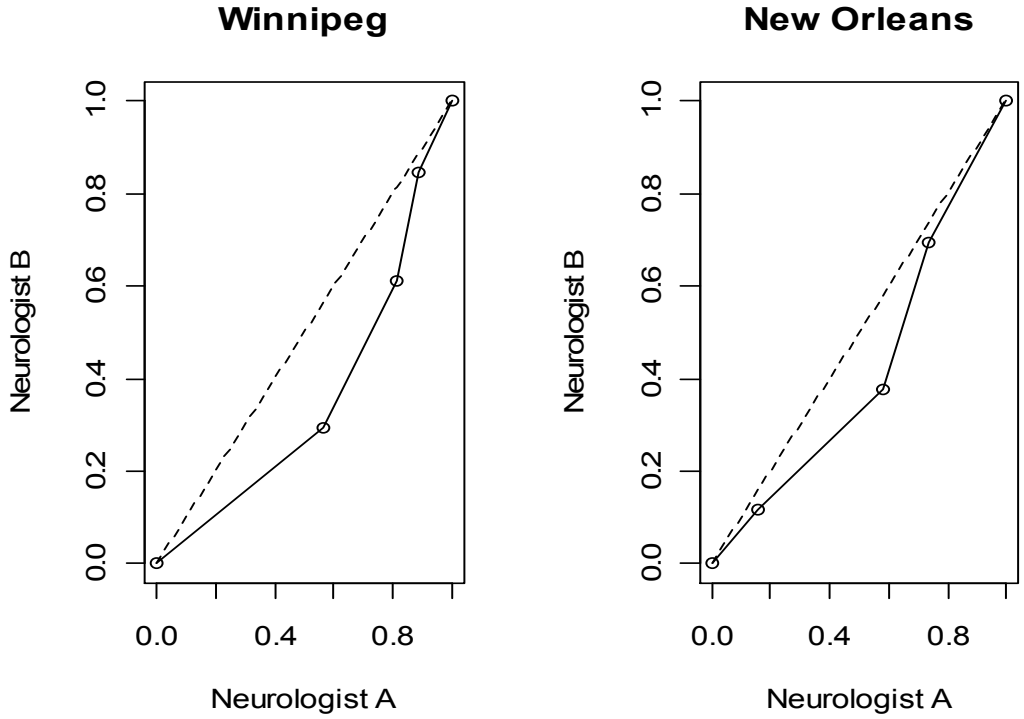

Figure 6.2 ROC curves for systematic disagreement between neurologists

From the above Figure 6.2, both of the ROC curves fall into the lower right triangle, which means systematic agreement in position. Neurologist A classified more patients into lower categories than Neurologist B. The measures of the random and systematic interobserver differences are shown in Table 6.2

Table 6.2 The interobserver reliability measures of systematic and random differences between neurologist.

Measures of difference	Winnpeg	New Orleans
Random difference:		
Relative rank variance, RV	0.037(0.050)	0.042(0.010)
Systematic difference:		
Relative position, RP	0.290(0.018)	0.161(0.033)
Relative concentration, RC	0.117(0.023)	0.083(0.046)
Coefficient of agreement, kappa	0.21	0.25
Percentage agreement, PA	43%	48%

Jackknife technique of standard error in brackets.(see Appendix 1.3)

According to Table 6.2, the main part of the unreliability can be explained by the systematic difference. For the Winnipeg group, the systematic difference in position between the Neurologist A and Neurologist B was 0.290 (SE=0.010) and the value is significant, which can be also proved by the ROC curve. It means that the neurologists disagreed concerning the cut-off points. The systematic difference in concentration is also significant. Apart from the systematic difference, the significant values of RV can also reflect the random difference. The measures of RV is small (0.037, SE=0.050), but it is not negligible. For the New Orleans group, the main reason for the systematic disagreement is the systematic difference in position (PR=0.161, SE=0.033). The RC value (0.083, SE=0.046) is negligible. The RV value (0.042, SE=0.010) also reflect random difference, which means there is a sign of

individual dispersion of change from the common systematic group change, but the level of the random difference is small. Both values of Kappa in the groups are small (0.21 and 0.25), which mean “Fair” agreement between the two neurologists in the two patients groups.

By using the rank-invariant method, we can know that the main reason for lack of reliability between the two neurologists in classification the patients may be their interpretation of category description and their clinical experience. It can be considerably reduced by specifying the category description and by training the neurologists.

Landis and Koch (1977) analyzed the same data by kappa-type statistics. [14] They also found there is significant interobserver difference between the two neurologists in their overall usage of the diagnostic classification scale. The neurologists have significant disagreement between diagnoses in cut-off points in both patients groups.

Although different statistical approaches were applied in the same data, the results are accordant concerning the same problem.

7 Conclusions

Although Kappa is quite popular and widely used in many research projectes now, unsatisfactory features are still there, which makes the application of Kappa reduced. When we use Kappa in assessing agreement, we should pay attention to the marginal distributions. In fact, it requires similar marginal distributions. When the marginal distribution is not similar, the two paradoxes occur. Prevalence and bias influence the values of Kappa. When the number of categories is changed, the Kappa value will be changed. For weighted Kappa, the choice of weights is subjective and weighted Kappa will have different values by choosing different weights. For the cases such as assessing many raters in different methods of the same subject, the Kappa can be compared by the values, but cannot tell us which method is better. If the subjects are different, the Kappa value can not be compared. In addition, how the magnitude of Kappa should be judged, is still not based on proper scientific reason.

Some other statistical approaches can be also applied in assessing agreement, such as the Spearman’s rank correlation coefficient, [15] and the unit-spaced scores in log-linear models.[16] The rank-invariant method identifies and measures the level of change in ordered categorical responses attributed to a group separately from the level of individual variability within the group.[12] However, this method can only be used for pair-wise data, or comparing all raters with the “golden standard” rater.

Apart from Kappa coefficient, the alternative statistical approaches can also have some undesirable properties and limitations of applications. When we need to assess the agreement, we should choose a suitable approach, according to the aim.

Appendix

All of the formula are introduced by Svensson [10][12]

1.1 Systematic disagreement (From Reference [12])

The parameter of the systematic disagreement in position between pairs of variables (X_k, Y_k) and (X_l, Y_l) is defined by $\gamma = P(X_l < Y_k) - P(Y_l < X_k)$. The parameter of the systematic disagreement in concentration is defined by $\delta = P(X_{l_1} < Y_k < X_{l_2}) - P(Y_{l_1} < X_k < Y_{l_2})$ where X and Y are two sets of marginal distributions, $k \neq l$ and $k, l = 1,..., m.$

Let x_i, y_i denote the ith category frequencies of marginal distributions X and Y, $C(X)_i$ and $C(Y)_i$ denote the ith category cumulative frequencies, the empirical measures of relative position (RP) and of relative concentration (RC) can be calculated as:

The probability of Y being classified toward higher categories than X, which means P(Y<X) is estimated by

$$p_{xy} = \frac{1}{n^2}\sum_{i=1}^{m}[y_i \times C(X)_{i-1}]$$

The probability of X being classified toward higher categories than Y, which means P(X<Y) is estimated by

$$p_{yx} = \frac{1}{n^2}\sum_{i=1}^{m}[x_i \times C(Y)_{i-1}]$$

Then the measure of systematic change in position is $RP = p_{xy} - p_{yx}$

The probability of Y being concentrated between the marginal distribution of X, $P(X_l < Y_k < X_o)$ is estimated by

$$p_{xyx} = \frac{1}{n^3}\sum_{i=1}^{m}\{y_i \times C(X)_{i-1}[n - C(X)_i]\}$$

The probability of X being concentrated between the marginal distribution of Y, $P(Y_l < X_k < Y_o)$ is estimated by

$$p_{yxy} = \frac{1}{n^3}\sum_{i=1}^{m}\{x_i \times C(Y)_{i-1}[n - C(Y)_i]\}$$

Then, the measure of systematic change in concentration is $RC = \frac{1}{M}(p_{xyx} - p_{yxy})$

where M=min [($p_{xy} - p_{xy}^2$),($p_{yx} - p_{yx}^2$)] [12]

1.2 Random disagreement (From Reference [12])

In some cases, random part of the disagreement cannot be explained by the systematic difference, if $\overline{R}_{ij}^{(x)} \neq \overline{R}_{ij}^{(y)}$.

x_{ij} is the (i,j)th cell frequency, where i and $j = 1,...,m$.

$\overline{R}_{ij}^{(x)}$ and $\overline{R}_{ij}^{(y)}$, the mean ranks of the observations in the (i, j) th in the table

n is the number of individuals

The augment ranking procedure means that the mean ranks for observations in the (i,j)th differ from the means ranks in (i,j+1)th cell, $\overline{R}_{ij}^{(x)} < \overline{R}_{i+1,j}^{(x)}$

The observations judged to the ijth cell will be given ranks ranging from

$$\sum_{i1=1}^{i-1} x_{i_1\bullet} + \sum_{j1=1}^{j-1} x_{ij_1} + 1 \qquad \text{to} \qquad \sum_{i1=1}^{i-1} x_{i_1\bullet} + \sum_{j1=1}^{j} x_{ij_1} + x_{ij}$$

Which gives the mean rank according to X of the observations in the ijth cell

$$\overline{R}_{ij}^{(X)} = \sum_{i1=1}^{i-1} x_{i_1\bullet} + \sum_{j1=1}^{j-1} x_{ij_1} + 0.5(1 + x_{ij})$$

In the same way, the mean rank according to Y of the observation in the ijth cell is

$$\overline{R}_{ij}^{(Y)} = \sum_{j1=1}^{j-1} x_{\bullet j_1} + \sum_{i1=1}^{i-1} x_{i_1 j} + 0.5(1 + x_{ij})$$

An empirical measure of random differences between two ordered categorical judgements on the same individual, calleded the Relative Rank Variance (RV) is defined by:

$$RV = \frac{6}{n^3} \sum\sum x_{ij} [\overline{R}_{ij}^{(X)} - \overline{R}_{ij}^{(Y)}]^2$$

RV ($0 \le RV < 1$) expresses the level of disagreement from a total agreement in rank ordering, given the marginals.

1.3 Standard error of RV, RP and RC (From Reference[10])

According to the Jackknife technique,the variance of the empirical Relative Rank Variance, Var(RV),is estimated by

$$\hat{\sigma}^2_{jack}(RV) = \frac{n-1}{n} \sum_{\kappa=1}^{n} (RV_{(\kappa)} - RV_{(\bullet)})^2 = \frac{(n-1)^2}{n} \hat{V}arRV_{(\kappa)}$$

Where $RV_{(\kappa)}$ denotes the Relative Rank Variance of the disagreement pattern with one observation, κ, deleted.

$RV_{(\bullet)}$ is the average of all possible Relative Rank Variances with one observation deleted, $\kappa = 1,...,n'$.

Similarly,the variance of the empirical measure of Relative Position, Var(RP) is estimated by

$$\hat{\sigma}^2_{jack}(RP) = \frac{n-1}{n}\sum_{\kappa=1}^{n}(RP_{(\kappa)} - RP_{(\bullet)})^2 \quad = \frac{(n-1)^2}{n}\hat{V}arRP_{(\kappa)}$$

Where $RP_{(\kappa)}$ denotes the Relative Position of the disagreement pattern with one observation, κ, deleted.

$RP_{(\bullet)}$ is the average of all possible Relative Position with one observation deleted, $\kappa = 1,...,n'$.

The variance of the empirical measure of Relative Concentration, Var(RC) is estimated by

$$\hat{\sigma}^2_{jack}(RC) = \frac{n-1}{n}\sum_{\kappa=1}^{n}(RC_{(\kappa)} - RC_{(\bullet)})^2 \quad = \frac{(n-1)^2}{n}\hat{V}arRC_{(\kappa)}$$

Where $RC_{(\kappa)}$ denotes the Relative Concentration of the disagreement pattern with one observation, κ, deleted.

$RC_{(\bullet)}$ is the average of all possible Relative Concentration with one observation deleted, $\kappa = 1,...,n'$.

1.4 Empirical results

For the data in Table 6.1, $RV_{(\kappa)}, RP_{(\kappa)}, RC_{(\kappa)}$ are listed in the following table (Using the software R)

Winnpeg	RP	RC	RV
(1, 1)	0.292549	0.1199232	0.03893772
(1, 2)	0.296658	0.1277411	0.0336277
(1, 4)	0.30305	0.1106093	0.03857496
(2, 1)	0.287071	0.1111164	0.03802897
(2, 2)	0.29118	0.1188997	0.03794938
(2, 3)	0.294878	0.1128389	0.03849723
(3, 1)	0.28488	0.1139419	0.03718684
(3, 2)	0.289034	0.1217005	0.03857496
(3, 3)	0.292732	0.1156242	0.03878781
(3, 4)	0.295334	0.1047564	0.03893772
(4, 1)	0.283601	0.1191546	0.03718684
(4, 2)	0.287756	0.1269259	0.03841949
(4, 3)	0.291499	0.1207654	0.03893772
(4, 4)	0.294102	0.1099144	0.03893772
New Orlean	RP	RC	RV
(1, 1)	0.165225	0.0958711	0.04430847
(1, 2)	0.170632	0.1090262	0.04199954
(2, 1)	0.156791	0.07103207	0.04350702
(2, 2)	0.162197	0.0841856	0.04280099
(2, 3)	0.170632	0.09088234	0.04602585
(3, 1)	0.14814	0.06441178	0.03377519

（3，2）	0.153763	0.0777992	0.04146525
（3，3）	0.162197	0.08444577	0.04413673
（3，4）	0.17128	0.06360774	0.04430847
（4，1）	0.141869	0.07892075	0.041408
（4，2）	0.147491	0.09225036	0.04135075
（4，3）	0.156142	0.09882186	0.04430847
（4，4）	0.165225	0.07808862	0.04430847

For Winnipeg group,

$$\hat{\sigma}^2_{jack}(RV) = \frac{n-1}{n}\sum_{\kappa=1}^{n}(RV_{(\kappa)} - RV_{(\bullet)})^2 \quad = 2.54686\text{E-}05$$

$$\hat{\sigma}^2_{jack}(RP) = \frac{n-1}{n}\sum_{\kappa=1}^{n}(RP_{(\kappa)} - RP_{(\bullet)})^2 \quad = 0.018397399$$

$$\hat{\sigma}^2_{jack}(RC) = \frac{n-1}{n}\sum_{\kappa=1}^{n}(RC_{(\kappa)} - RC_{(\bullet)})^2 \quad = 0.000566074$$

The estimated standard errors of RV, RP RC are 0.005, 0.0184 and 0.0238 respectively.

For New Orleans group, the estimated standard errors of RV, RP RC are 0.010, 0.033and 0.046 respectively.

References

[1] Altman, D.G. *Practical statistics for medical research.* Chapman and Hall. 1991 pp.403-415

[2] Fleiss JL. *et al. Statistical methods for rates and proportions.* Wiley series in probability and statistics. 2002 pp.598-618

[3] Agresti A. *Modelling patterns of agreement and disagreement.* Statistical methods in medical research 1992:1

[4] Cohen J. *A coefficient of agreement for nominal scales.* Educ Psychol Measure 1960; 20: 37-46

[5] Feinstein AR, Cicchetti DV. *High agreement but low kappa:* I. *The problems of two paradoxes.* J Clin Epidemiol 1989 Vol.43, No. 6, pp. 543-549

[6] Feinstein AR, Cicchetti DV. *High agreement but low kappa:* II. *Resolving the paradoxes.* J Clin Epidemiol 1989 Vol.43, No. 6, pp. 543-549

[7] Byrt et al *Bias, prevalence and Kappa*. J Clin Epidemiol 1993 Vol. 46, No. 5, pp 423-429

[8] Fleiss JL. *Measuring nominal scale agreement among many raters.* Psychological Bulletin 1971, Vol.76, No.5, pp.378-382

[9] Busch C. *Personal communication*. 1994 Uppsala

[10] Svensson E. *Analysis of systematic and random differences between paired ordinal categorical data*. Göteborg 1993

[11] Svensson E. *Application of a rank-invariant method to evaluate reliability of ordered categorical assessments*. Journal of Epidemiology and Biostatistics 1998 Vol. 3, No. 4, pp. 403-409

[12] Svensson E, Starmark J. *Evaluation of individual and group changes in social outcome after aneurismal subarachnoid haemorrhage: A long-term follow-up study*. J Rehabil Med 2002; 34: pp.251-259

[13] Svensson E et al. *Analysis of interobserver disagreement in the assessment of subarachnoid blood and acute hydrocephalus on CT scans.* Neurological research, 1996, Vol. 18 pp.487-193

[14] Landis JR, Koch GG, *The measurement of observer agreement for categorical data. Biometrics*, 1977, Vol. 33, No. 1, pp. 159-174

[15] Spearman C , *The proof and measurement of association between two things* Amer. J. Psychol. , 15 (1904) pp. 72–101

[16] Agresti A. *Modelling patterns of agreement and disagreement*. Stat Methods Med Res, 1992, pp 201-218

[17] Hoehler FK. *Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity*. Journal of Clinical Epidemilogy 53, 2000, pp. 499-503.

[18] Glass M. T*he kappa statistic: A second look. Computational Liguistics*. 2004, Vol 30: 1, pp. 95-101

[19] Adam T. *To judge the judges-Kappa, ROC or what?* 1994, Uppsala University