Kasetsart J. (Nat. Sci.) 48 : 294 - 300 (2014)

Some Further Study Based on Cohen's Kappa Statistic: Theory and Applications

Pornpis Yimprayoon

ABSTRACT

The statistical inference of the problem of measuring agreement between two raters who employ measurements on a two-point nominal scale was studied. The objective was to propose some further characteristics on Cohen's kappa statistic. Another approach was proposed to test whether or not the observed estimate of kappa is significantly greater than a high predetermined value for a given value of an observed agreement. Moreover, the proposed test procedure was illustrated using a real data example.

Keywords: measuring agreement, Cohen's kappa, nominal scale, hypothesis testing, tuberculosis

INTRODUCTION

Researchers in many fields have become increasingly aware of the problem of errors in measurements with investigations into the scientific bases of measurement errors commencing over one and a half centuries ago (Shoukri, 2002). A number of statistical problems in several fields of application in the social and biomedical sciences require the measurement of agreement between two or more raters. One of the most popular indices of agreement was introduced by Cohen (1960), namely Cohen's kappa statistic (κ_C), as a reliability index for measuring chance-corrected agreement between two raters employing nominal scales. The current research contributes some further study on Cohen's kappa statistic when the proportion of rater pairs exhibiting agreement (θ_o) is given.

Brief description of Cohen's Kappa

Let us consider a reliability research where two raters, referred to as rater A and rater B, are required to classify subjects into one of two possible response categories. The two response categories, labeled as 1 and 2, are assumed to be disjoint. An interpretation of the probabilities π_{ij} for i, j = 1, 2, in a cross-classification of the reliability research can be depicted as given in Table 1.

Table 1 Joint distribution of classification probabilities for two raters and two response categories.

Rater A	Rater B		Total
	1	2	
1	π_{11}	π_{12}	$\pi_{1\cdot}$
2	π_{21}	π_{22}	$\pi_{2\cdot}$
Total	$\pi_{\cdot 1}$	$\pi_{\cdot 2}$	1

Department of Mathematics, Faculty of Liberal Arts and Science, Kasetsart University, Nakhon Pathom 73140, Thailand.
E-mail: faasppy@ku.ac.th

Received date : 17/09/13 Accepted date : 19/11/13

It follows from Table 1 that π_{ij} represents the chance that rater A classifies a subject into category i, while rater B classifies the same subject into category j, $1 \le i, j \le 2$. Let $\pi_{1\cdot} = \sum_{j=1}^{2} \pi_{1j}$ and $\pi_{2\cdot} = \sum_{j=1}^{2} \pi_{2j} = 1 - \pi_{1\cdot}$ be the probabilities of being classified by rater A into categories 1 and 2, respectively. Also define $\pi_{\cdot 1} = \sum_{i=1}^{2} \pi_{i1}$ and $\pi_{\cdot 2} = \sum_{i=1}^{2} \pi_{i2} = 1 - \pi_{\cdot 1}$ in the same manner. The joint rating probability matrix has π_{ij} as its elements; $1 \le i, j \le 2$.

In this arrangement, Cohen's kappa for assessing agreement between the two raters is defined by Equation 1:

$$\kappa_C = \frac{\theta_o - \theta_e}{1 - \theta_e} \tag{1}$$

where

$$\theta_o = \pi_{11} + \pi_{22},$$
$$\theta_e = \pi_{1\cdot}\pi_{\cdot 1} + \pi_{2\cdot}\pi_{\cdot 2}.$$

In applications, if there are n subjects and n_{ij} represents the number of subjects classified in category i by rater A and in category j by rater B, the sample estimate of κ_C is given by Equation 2, 3, 4, 5:

$$\hat{\kappa}_C = \frac{\hat{\theta}_o - \hat{\theta}_e}{1 - \hat{\theta}_e} \tag{2}$$

where $\hat{\theta}_o = \frac{n_{11} + n_{22}}{n}$,

$$\hat{\theta}_e = \frac{n_{1\cdot}n_{\cdot 1} + n_{2\cdot}n_{\cdot 2}}{n^2}.$$

$$\hat{\pi}_{ij} = \frac{n_{ij}}{n}, \tag{3}$$

$$\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n}, \tag{4}$$

$$\hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n}, \tag{5}$$

Some features of Cohen's Kappa statistic

The value of Cohen's kappa statistic κ_C ranges from -1 to 1 dependent on the strength of agreement with most of this interest having been focused on high measures of agreement. It would indicate consensus in the judgments by the two raters. Another approach is proposed to test whether or not the observed estimate of kappa is significantly greater than a high predetermined value κ_0 for given value of θ_o. That is, when θ_o is a given value, the hypothesis tested is Equation 6:

$$H_0 : \kappa_C \ge \kappa_0 \text{ versus } H_1 : \kappa_C < \kappa_0 , \tag{6}$$

where κ_0 is a predetermined high value.

Since $\kappa_C \ge \kappa_0$,

or

$$\frac{\theta_o - \theta_e}{1 - \theta_e} \ge \kappa_0 ,$$

or equivalently,

$$\theta_e \le \frac{\theta_o - \kappa_0}{1 - \kappa_0} .$$

We then substitute $\theta_e = (\pi_{11} + \pi_{12})(\pi_{11} + \pi_{21}) + (\pi_{21} + \pi_{22})(\pi_{12} + \pi_{22})$ in the above expression, that takes the form

$$(\pi_{11} + \pi_{12})(\pi_{11} + \pi_{21}) + (\pi_{21} + \pi_{22})(\pi_{12} + \pi_{22}) \le \frac{\theta_o - \kappa_0}{1 - \kappa_0} ,$$

or

$$(\pi_{11} + \pi_{12})(\pi_{11} + 1 - \theta_o - \pi_{12}) + (1 - \theta_o - \pi_{12} + \theta_o - \pi_{11}) (\pi_{12} + \theta_o - \pi_{11}) \le \frac{\theta_o - \kappa_0}{1 - \kappa_0}$$

or equivalently,

$$\theta_o + 2\left[\pi_{12} + (\pi_{11}^2 - \pi_{12}^2) - \theta_o(\pi_{11} + \pi_{12})\right] \le \frac{\theta_o - \kappa_0}{1 - \kappa_0}.$$

Therefore, we obtain Equation 7:

$$\frac{\kappa_0(1 - \theta_o)}{1 - \kappa_0} \le 2\left[\pi_{11}(\theta_o - \pi_{11}) + \pi_{12}(\theta_o + \pi_{12} - 1)\right]. \tag{7}$$

We then have the graph of Equation 8:

$$\pi_{11}(\theta_o - \pi_{11}) + \pi_{12}(\theta_o + \pi_{12} - 1) \ge \frac{\kappa_0(1 - \theta_o)}{2(1 - \kappa_0)},$$

or equivalently,

$$(\pi_{12}^2-\pi_{11}^2)+\theta_o(\pi_{11}+\pi_{12})-\pi_{12}\ge\frac{\kappa_0(1-\theta_o)}{2(1-\kappa_0)}. \quad (8)$$

We note that Equation 6 will be meaningful if

$\kappa_0\le\frac{\theta_o^2}{1+(1-\theta_o)^2}$. If $\kappa_0=\frac{\theta_o^2}{1+(1-\theta_o)^2}$,

then $\pi_{11}=\pi_{22}=\frac{\theta_o}{2},\pi_{12}=1-\theta_o,\pi_{21}=0$ or

$\pi_{11}=\pi_{22}=\frac{\theta_o}{2},\pi_{12}=0,\pi_{21}=1-\theta_o.$

Given θ_o, data on raters' judgment is generated as follows:

Select n individuals and divide them into two groups sizes $n_0=n\theta_o=n_{11}+n_{22}$ and $n_1=n(1-\theta_o)=n_{12}+n_{21}$. Then, ask the raters to confine to the agreement cells [(1,1) and (2,2)] the without loss of generality, assume

$$\pi_{11}=\min(\pi_{11},\pi_{22})\le\frac{\theta_o}{2}$$

and

$$\pi_{12}=\min(\pi_{12},\pi_{21})\le\frac{1-\theta_o}{2}.$$

Moreover, in sampling, n_{11} and n_{12} are independent random variables with binomial distributions $b\left(n\theta_o,\frac{\pi_{11}}{\theta_o}\right)$ and $b\left(n(1-\theta_o),\frac{\pi_{12}}{1-\theta_o}\right)$, respectively. Next, consider an illustrative example with such a situation examined below.

NUMERICAL RESULT

Consider a hypothetical reliability research scenario as shown in Table 2. Raters A and B have classified 200 subjects into one of two possible response categories each.

Table 3 displays an interpretation of the estimated probabilities $\hat{\pi}_{ij}=\frac{n_{ij}}{n}$ for i, j = 1, 2, in a cross-classification of Table 2.

The hypotheses of interest for testing are:

$H_0:\kappa_C\ge 0.80$

$H_1:\kappa_C<0.80$

where this predetermined value corresponds to strong agreement proposed by Landis and Koch (1977), Kraemer (1979), Fleiss (1981), Altman (1991), and Lantz and Nebenzahl (1996).

Therefore, when θ_o = 0.94 is given, in this case

n = 200,

$n_0=n\theta_o=200\times0.94=188$,

$\kappa_0=0.80$,

and $\kappa_{C,\max}=\frac{\theta_o^2}{1+(1-\theta_o)^2}=\frac{0.94^2}{1+(1-0.94)^2}=0.8804$.

From Equation 8, can be derived the graph of Equation 9:

$$(\pi_{12}^2-\pi_{11}^2)+0.94(\pi_{11}+\pi_{12})-\pi_{12}\ge\frac{0.80(1-0.94)}{2(1-0.80)}, \quad (9)$$

Table 2 Distribution of 200 subjects by rater and response.

Rater A	Rater B		Total
	1	2	
1	80	4	84
2	8	108	116
Total	88	112	200

Table 3 Joint distribution of classification probabilities.

Rater A	Rater B		Total
	1	2	
1	0.40	0.02	0.42
2	0.04	0.54	0.58
Total	0.44	0.56	1.00

or the graph of Equation 10:

$$(\pi_{12}^2 - \pi_{11}^2) + 0.94(\pi_{11} + \pi_{12}) - \pi_{12} \geq 0.12. \qquad (10)$$

Thence can be plotted the (π_{11}, π_{12}) -plane that satisfies the inequality in Equation 10 as shown in Figure 1. For given $\theta_o = 0.94$, this figure shows that $\pi_{11} \geq 0.1538$ whereas $\kappa_C \geq 0.80$.

To satisfy Figure 1, consider the test of

$H_0 : \pi_{11} \geq 0.1538$

against

$H_1 : \pi_{11} \geq 0.1538$

instead of the testing problem $H_0 : \kappa_C \geq 0.80$.

The test statistic is given by Equation 11:

$$z = \frac{\hat{\pi}_{11} - 0.1538}{\sqrt{\hat{V}ar(\hat{\pi}_{11})}} \qquad (11)$$

where $\hat{V}ar(\hat{\pi}_{11})$ is the estimated variance. A test with the approximate significance level α for doing this is to reject H_0 if $z \leq -z\alpha$ at $(1 - \alpha)100\%$ level of significance.

Applying the test procedures to analyze this example results in $\hat{\pi}_{11} = 0.40, \hat{\pi}_{22} = 0.54, \hat{V}ar(\hat{\pi}_{11})$ = 0.0011 and z = 7.4232.

Using a significance level α = 0.05, the critical region is $z \leq -1.645$. Thus, it can be concluded that π_{11} is not significantly less than 0.1538. That is, $\kappa_C \geq 0.80$. Further, a 95% lower confidence limit based on $\hat{\pi}_{11}$ for π_{11} is $\hat{\pi}_{11} - z_\alpha \sqrt{\hat{V}ar(\hat{\pi}_{11})} = 0.40 - 1.645\sqrt{0.0011}$ = 0.3454.

Application to tuberculosis patients

Tuberculosis is a major cause of morbidity and remains one of the deadliest diseases in the world, with the World Health Organization (WHO) estimating that each year more than 8 million new cases of tuberculosis occur and approximately 3 million people die from the disease (World Health Organization, 1996). Ninety-five percent of tuberculosis cases occur in developing countries, where few resources are available to ensure proper treatment and where human immunodeficiency virus (HIV) infection may be common (World Health Organization, 1996). It is estimated that between 19 and 43% of the world's population is infected with *Mycobacterium tuberculosis*, the bacterium that causes tuberculosis infection and disease (Sudre and Kochi, 1992; World Health Organization, 1996).

The tuberculin skin test is an important aid in the detection of tuberculosis. Numerous studies have confirmed the specificity and accuracy of the Mantoux test and it has become the standard (Capobres *et al*., 1962; Sudre and Kochi, 1992; Campos-Neto *et al*., 2001). However, there are many drawbacks and inconveniences associated with the use of the Mantoux test, such as its requirements of freshly prepared solution, syringes and a needle which presents an emotional problem, and makes it necessary for a physician and nurse to be involved in the testing and reading (Sudre and Kochi, 1992).

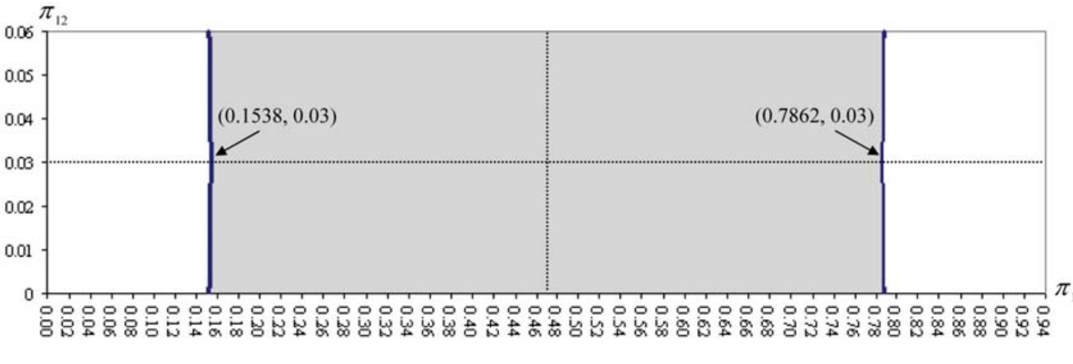

Figure 1 Graph of (π_{11}, π_{12}) -plane in testing the agreement.

The need for a more convenient test has been recognized for some time (Sudre and Kochi, 1992). One of the recommendations was to find a new tuberculin skin test which would be simple and accurate, and could be applied and read by non-medical personnel (Sudre and Kochi, 1992). The new test—namely, the tine test, here reported—is offered to meet the requirements of a new skin test as recommended by the Arden House Conference (A Report of the Arden House Conference on Tuberculosis, 1952).

Part of the real data reported by Capobres *et al*. (1962) was used to illustrate the procedures discussed. This study was conducted at the Missouri State Sanatorium, one of three tuberculosis hospitals in Missouri (the others are in Kansas City and St. Louis). The majority of the persons tested were inpatients and those newly admitted to the hospital, although many employees of the hospital were included.

In this example, the tine test was evaluated against the standard Mantoux test for the detection of tuberculosis. The tuberculin used in the Mantoux tests was intermediate strength purified protein derivative in the dosage of 5 tuberculin units.

The tuberculin tine test unit consists of four 2 mm stainless steel points attached to a plastic handle. The points have been dipped in four-times-concentrated International Standard Old tuberculin, dried and gas sterilized. Thus, the apparatus for application and the testing material are supplied in one unit, that is easily stored and available for use at any time without wastage. The dried tuberculin maintains its potency for more than 2 yr at room temperature. The use of multiple punctures tends to eliminate test failures or repetitions. If only a few patients are to be tested, the unused units remain sterile in their individual compartments for use at a moment's notice. The technique of application is so easy that no special training is required to be able to apply the test speedily and uniformly. Moreover, the test unit minimizes the emotional reaction of the recipient compared to a needle and syringe. There is little pain and it is of short duration, comparable to that of a mosquito bite.

Both tests were skin tests applied to the arms. Readings were made at 48 hr, that is, readings of the Mantoux test were classified as negative (0–4 mm) and positive (greater than or equal to 5 mm); those of the tine test were classified as negative (0–1 mm) and positive (greater than or equal to 2 mm). The data and proportions of subjects are presented in Tables 4 and 5, respectively.

Table 4 Combined reading of Mantoux and tine tests at 48 hr.

Tine test	Mantoux test		Total
	Negative (0–4 mm)	Positive (≥ 5 mm)	
Negative (0–1 mm)	367	31	398
Positive (≥ 2 mm)	37	887	924
Total	404	918	1322

Table 5 Joint distribution of classification probabilities.

Tine test	Mantoux test		Total
	Negative (0–4 mm)	Positive (≥ 5 mm)	
Negative (0–1 mm)	0.2776	0.0234	0.3010
Positive (≥ 2 mm)	0.0280	0.6710	0.6990
Total	0.3056	0.6944	1.0000

The study tested the agreement between the Mantoux and tine tests regarding information on tuberculin skin testing. That is, interest was directed to the problem of when the tests illustrated that there was a high value of the observed estimate of kappa for a given value of θ_o. Thus to investigate how the estimate kappa is high in this problem, the problem was framed to plot a graph for considering the behavior of π_{11}. That is, when θ_o = 0.9486 is given, the hypothesis $H_0 : \kappa_C \geq 0.80$ was tested versus $H_1 : \kappa_C < 0.80$.

Equation 8 produces the graph of $(\pi_{12}^2 - \pi_{11}^2) + 0.9486(\pi_{11} + \pi_{12}) - \pi_{12} \geq 0.1028$ as shown in Figure 2. For given $\theta_o = 0.94$, this figure shows that $\pi_{11} \geq 0.1257$ whereas $\kappa_C \geq 0.80$. To satisfy Figure 2, consider the test of $H_0 : \pi_{11} \geq 0.1257$ versus $H_1 : \pi_{11} < 0.1257$ instead of the testing problem $H_0 : \kappa_C \geq 0.80$.

The values of the statistics for these data are given by $\hat{\pi}_{11} = 0.2776$, $\hat{\pi}_{22} = 0.6710$, and $\hat{Var}(\hat{\pi}_{11}) = 0.0001$, then z = 15.19. Using a significance level α = 0.05, the critical region is $z \leq -1.645$. Thus it can be concluded that π_{11} is not significantly less than 0.1257. Thus, in this example, there is no evidence for suggesting that the value of κ_C is less than 0.80. That is, $\kappa_C \geq 0.80$; this result also suggests a strong level of agreement between the Mantoux and tine tests. This result indicated there was sufficient agreement for the new tine test to replace the Mantoux test.

CONCLUSION

This research discussed the problem of measuring agreement or disagreement between two raters where the ratings were given separately in a two-point nominal scale. Some further study on Cohen's kappa statistic proposed another approach to test whether or not the observed estimate of kappa was significantly greater than κ_0 for a given value of θ_o. In addition, the proposed test procedure was explained using sample data.

ACKNOWLEDGEMENTS

The authors sincerely thank Assoc. Prof. Dr. Montip Tiensuwan at the Department of Mathematics, Faculty of Science, Mahidol University, for suggesting this problem. Special gratitude is expressed to Prof. Dr. Bikas K. Sinha from the Indian Statistical Institute, Kolkata, India, for his expertise, excellent guidance and useful comments. The Kasetsart University Research and Development Institute (KURDI), Kasetsart University, Thailand provided financial support which enabled this research to be undertaken. The insightful comments from the referees were extremely useful for revision and for this, the authors are also grateful.

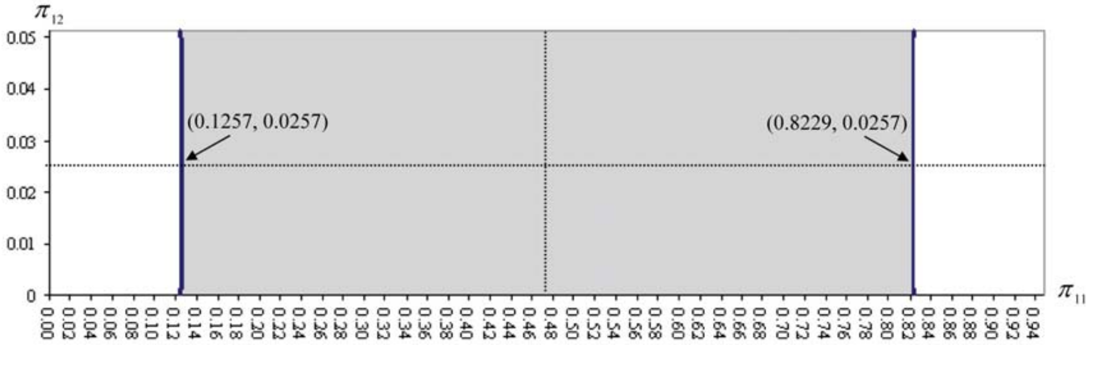

Figure 2 Graph of (π_{11}, π_{12}) -plane in testing the agreement for tuberculosis patients.

LITERATURE CITED

A Report of the Arden House Conference on Tuberculosis. 1952. **TB Control: The Big Push Ahead (Treatment Is Tool).** New York, NY, USA. November 29–December 2, 1952.

Altman, D.G. 1991. **Practical Statistics for Medical Research.** Chapman and Hall. London, UK. 611 pp.

Campos-Neto, A., V. Rodrigues-Junior, D.B. Pedral-Sampaio, E.M. Netto, P.J. Ovendale, R.N. Coler, Y.A. Skeiky, R. Badaro and S.G. Reed. 2001. Evaluation of DPPD, a single recombinant *Mycobacterium tuberculosis* protein as an alternative antigen for the Mantoux test. **Tuberculosis (Edinburgh).** 81: 353–358.

Capobres, D.B, F.E. Tosh, J.L. Yate and H.V. Langeluttig. 1962. Experience with the tuberculin tine test in a sanatorium. **J. Am. Med. Assoc.** 180: 1130–1136.

Cohen, J. 1960. A coefficient of agreement for nominal scales. **Educ. Psyc. M.** 20(1): 37–46.

Fleiss, J.L. 1981. **Statistical Methods for Raters and Proportions.** 3th ed. New York, NY, USA. 352 pp.

Kraemer, H.C. 1979. Ramifications of a population model for kappa as a coefficient of reliability. **Psychometrika** 44(4): 461–472.

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

Lantz, C.A. and E. Nebenzahl. 1996. Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. **J. Clin. Epid.** 49(4): 431–434.

Shoukri, M.M. 2000. **Measures of Interobserver Agreement.** CRC Press LLC. Boca Raton, FL, USA. 168 pp.

Sudre, P., G.T. Dam and A. Kochi. 1992. Tuberculosis: A global view of the situation today. **Bull. WHO** 70: 149–159.

World Health Organization. 1996. **Groups at Risk: WHO Report on the Tuberculosis Epidemic.** WHO. Geneva, Switzerland.