บางรูปแบบเชิงสถิติของความเห็นพ้องต้องกันของผู้ประเมิน

Some Statistical Aspects of Raters' Agreements

พรพิศ ยิ้มประยูร[1] และ สิทธิพงศ์ รักตะเมธากูล[1]

Pornpis Yimprayoon[1] and Sittipong Ruktamatakul[1]

บทคัดย่อ

ในงานวิจัยนี้ ผู้วิจัยจะมุ่งเน้นปัญหาทางสถิติของการวัดความเห็นพ้องต้องกันระหว่างสองผู้ประเมินซึ่งใช้วิธีการวัดแบบมาตรานามบัญญัติ 2 กลุ่ม วัตถุประสงค์ของงานวิจัยนี้คือ การอธิบายลักษณะบางอย่างของตัวสถิติโคเฮนแคปปาและตัวสถิติแคปปาที่ถูกปรับปรุง ผลที่ได้แสดงให้เห็นว่า ถ้าความน่าจะเป็นที่ผู้ประเมินไม่เห็นพ้องต้องกัน $\pi_{12}+\pi_{21} >> 0.50$ แล้ว $\kappa_C, \kappa_M < 0$ ในทางตรงกันข้าม ถ้า $\pi_{12}+\pi_{21} << 0.50$ แล้ว $\kappa_C, \kappa_M > 0$

คำสำคัญ: การวัดความเห็นพ้องต้องกัน ผู้ประเมิน มาตรานามบัญญัติ ตัวสถิติโคเฮนแคปปา ตัวสถิติแคปปาที่ถูกปรับปรุง

ABSTRACT

In this research, the statistical inference of the problem of measuring agreement between two raters who employ measurements on a 2-point nominal scale is focused. The purpose of this study is to illustrate some characteristics of the Cohen's kappa statistic and the modified kappa statistic. The results show the case that all kappa statistics κ_C and κ_M are less than zero, if the proportion of units in which the two raters disagree $\pi_{12}+\pi_{21} >> 0.50$. On the other hand, if the proportion of units in which the two raters disagree $\pi_{12}+\pi_{21} << 0.50$, then all kappa statistics κ_C and κ_M are greater than zero.

Keywords: measuring agreement, rater, nominal scale, Cohen's kappa statistic, modified kappa statistic

e-mail address: **faasppy@ku.ac.th**

INTRODUCTION

Researchers have become increasingly aware of the problem of errors in measurements more than one and a half centuries ago. The scientific bases of measurement errors have been investigated by statisticians, clinicians, epidemiologists, psychologists and many other scientists.

[1] สายวิชาคณิตศาสตร์ คณะศิลปศาสตร์และวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน นครปฐม 73140

[1] Department of Mathematics, Faculty of Liberal Arts and Science, Kasetsart University, Kamphaeng Saen Campus, Nakhonpathom 73140, Thailand.

There are numerous examples that illustrate these situations: the agreement of laboratory measurements collected in various laboratories, the agreement of a newly developed method with gold standard method, the agreement of manufacturing process measurements with specifications, the agreement of observed values with predicted values, and the agreement in bioavailability of a new or generic formulation with a commonly used formulation. By the way, measuring agreement has been used very often to designate the level of agreement between different data-generating sources referred to as raters. So 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 originally presented by Cohen (1960) as a reliability index for measuring agreement between two raters employing nominal scales.

BRIEF DESCRIPTION OF COHEN'S KAPPA STATISTIC

Let us consider a reliability research where 2 raters, referred to as rater A and rater B, operate independently and are required to classify subjects into one of 2 possible response categories. The subjects are independent. The 2 response categories, labeled as 1 and 2, are independent, mutually exclusive, and exhaustive. We denote by π_{ij} the chance that rater A classifies a subject into category $i,$ while rater B classifies the same subject into category j for $i, j = 1,2$. 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. We 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.

In this set-up, Cohen's kappa statistic for measuring agreement between the two raters is defined as

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

where $\theta_o = \pi_{11} + \pi_{22}$, $\theta_e = \pi_{1\cdot}\pi_{\cdot 1} + \pi_{2\cdot}\pi_{\cdot 2}$. (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 is given by

$$\hat{\kappa}_C = \frac{\hat{\theta}_o - \hat{\theta}_e}{1 - \hat{\theta}_e} \tag{3}$$

where $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$, $\hat{\pi}_{i\cdot} = \frac{n_{i\cdot}}{n}$, $\hat{\pi}_{\cdot j} = \frac{n_{\cdot j}}{n}$, $\hat{\theta}_o = \frac{n_{11} + n_{22}}{n}$, $\hat{\theta}_e = \frac{n_{1\cdot}n_{\cdot 1} + n_{2\cdot}n_{\cdot 2}}{n^2}$. (4)

The difference $\hat{\theta}_o - \hat{\theta}_e$ is the proportion of agreement beyond what is expected by chance. If $\hat{\theta}_o - \hat{\theta}_e$ is positive, two raters agree more often than expected based on chance. But negative values

of $\hat{\theta}_o - \hat{\theta}_e$ indicate they agree less than expected based on chance. The maximum possible discrepancy between $\hat{\theta}_o$ and $\hat{\theta}_e$ is $1-\hat{\theta}_e$. This discrepancy results when all decision appear in the agreement cells of the cross-classification. In this case, agreement is perfect.

In addition, a number of authors has proposed guidelines for the interpretation of κ_C. For example, Landis and Koch (1977) suggest the categories that the largest value of kappa is 1.00, indicating perfect agreement. A value of 0.00 indicates that the observed agreement is the same as that expected by chance, and the minimum value of kappa falls between -1.00 and 0.00.

We then can consider a general interpretation of κ_C focuses on the characteristic of κ_C in the following:

1. $\kappa_C = 1.00$ if and only if $\pi_o = 1.00$, this means that no controversial judgment by the raters, that is, the probability in the disagreement cells (off-diagonal cells) is zero.

2. $\kappa_C = 0.00$ only if $\pi_o = \pi_e$ (that is $\pi_{11} = \pi_{1.} \times \pi_{.1}$ or $\pi_{22} = \pi_{2.} \times \pi_{.2}$ or $\pi_{12} = \pi_{1.} \times \pi_{.2}$ or $\pi_{21} = \pi_{2.} \times \pi_{.1}$) or only says raters A and B perform independent.

3. $\kappa_C = -1$ if the probability in the agreement cells (diagonal cells) is zero and the probability of cell (2,1) is equal to the probability of cell (1,2) or it is only occurred $\pi_{12} = \pi_{21} = 0.50$ and $\pi_{11} = \pi_{22} = 0.00$.

Sinha *et al.* (2006) pointed out some undesirable features of κ_C and also said that the case of "$\kappa_C = -1$" seems to impose too restrictive behavior on the part of the raters. When $\pi_{11} = \pi_{22} = 0,$ there is already an indication of total disagreement between the two raters. Therefore, in such situations, irrespective of the values assumed by π_{12} and π_{21} $(0 < \pi_{12}, \pi_{21} < 1,$ $\pi_{12} + \pi_{21} = 1),$ the kappa coefficient is desired to assume the value -1. With this in mind, set $\pi_{12} = \alpha$ and $\pi_{21} = 1-\alpha,$ $0 < \alpha < 1$ and analyzed the situation with the purpose of modifying the definition of κ_C to deal with the full strength of disagreement between the two raters while the ratings are given independently in 2-point nominal scale.

Their modification is aimed at the value $\kappa_C = -1.$ They modified κ_C as

$$\kappa_M = \frac{\theta_o - \theta_e}{A - \theta_e} \tag{5}$$

and suggested a value of A to take care of the situations:

$$\pi_{11} = \pi_{22} = 0, \tag{6}$$

$$\pi_{12} = \alpha, \tag{7}$$

$$\pi_{21} = 1-\alpha, \tag{8}$$

when $0 < \alpha < 1$ along with $\kappa_M = -1$. Under (6)-(8), κ_M reduces to

$$\kappa_M = \frac{-2\alpha(1-\alpha)}{A - 2\alpha(1-\alpha)}. \tag{9}$$

and $\kappa_M = -1$ yields

$$A = 4\alpha(1-\alpha).$$

(10)

Then, replacing α by $\frac{\pi_{1.} + \pi_{.2}}{2}$ in (10). That is

$$A = 4 \cdot \frac{\pi_{1.} + \pi_{.2}}{2} \cdot \frac{\pi_{.1} + \pi_{2.}}{2}$$
$$= (\pi_{1.} + \pi_{.2})(\pi_{.1} + \pi_{2.}). \quad (11)$$

Next, substituting (11) in (7), we obtain

$$\kappa_M = \frac{\theta_o - \theta_e}{(\pi_{1.} + \pi_{.2})(\pi_{.1} + \pi_{2.}) - (\pi_{1.}\pi_{.1} + \pi_{.2}\pi_{2.})}.$$

(12)

Hence, the modified kappa statistic κ_M is defined as

$$\kappa_M = \frac{\theta_o - \theta_e}{\pi_{1.}\pi_{2.} + \pi_{.1}\pi_{.2}}.$$

(13)

This modification is based on the analysis of situations leading to total disagreement between the two raters. In addition Sinha *et al.* [3] verified below that all the three essential features of the kappa statistic are retained by κ_M. Clearly $\kappa_M = 0$ if and only if $\theta_o = \theta_e$. The other two values (± 1) are examined below in Theorem 1.

Theorem 1: *Let* κ_M *be the modified kappa, then*

$$\kappa_M = \begin{cases} +1.00 \text{ iff } \begin{pmatrix} \alpha & 0 \\ 0 & 1\text{-}\alpha \end{pmatrix}, \text{for } 0 < \alpha < 1 \\ -1.00 \text{ iff } \begin{pmatrix} 0 & \alpha \\ 1\text{-}\alpha & 0 \end{pmatrix}, \text{for } 0 < \alpha < 1. \end{cases}$$

Theorem 2: *Suppose* $\pi_{12}, \pi_{21} \neq 0$ *and* $\pi_{12}\pi_{21} > \pi_{11}\pi_{22}$. *Then* $\kappa_C, \kappa_M < 0$.

Corollary 1: *Suppose* $\pi_{11}, \pi_{22} \neq 0$ *and* $\pi_{12}\pi_{21} < \pi_{11}\pi_{22}$. *Then* $\kappa_C, \kappa_M > 0$.

Theorem 3: $\kappa_C = \kappa_M$ *if and only if* $\pi_{1.} = \pi_{.1}$ *or equivalently,* $\pi_{2.} = \pi_{.2}$.

So, the purpose of this study is to illustrate some characteristics of the Cohen's kappa statistic and the modified kappa statistic in the following claims.

SOME CHARACTERISTICS OF THE COHEN'S KAPPA AND MODIFIED KAPPA

The next claim we show the case that all kappa statistics κ_C and κ_M are less than zero, if the proportion of units in which the two raters disagree $\pi_{12}+\pi_{21} >> 0.50$.

Claim 1: *Suppose* p_0 *be an arbitrary probability that satisfy* $p_0 >> 0.50$, $\pi_{12}+\pi_{21}=p_0$ *and* $\pi_{12} \neq 0$ *(or* $\pi_{21} \neq 0$*), then* $\kappa_C, \kappa_M < 0$.

Proof: Assume that for all probability, denoted by p_0, such that $p_0 >> 0.50$.

We first will show that $\kappa_C < 0$. The Cohen's kappa statistic is given by

$$\begin{aligned}\kappa_C &= \frac{\theta_o-\theta_e}{1-\theta_e} \\ &= \frac{(\pi_{11}+\pi_{22})-\left[(\pi_{11}+\pi_{12})(\pi_{11}+\pi_{21})+(\pi_{21}+\pi_{22})(\pi_{12}+\pi_{22})\right]}{1-\left[(\pi_{11}+\pi_{12})(\pi_{11}+\pi_{21})+(\pi_{21}+\pi_{22})(\pi_{12}+\pi_{22})\right]} \\ &= \frac{\left[1-(\pi_{12}+\pi_{21})\right]-\left[(\pi_{11}+\pi_{12})(\pi_{11}+\pi_{21})+(\pi_{21}+\pi_{22})(\pi_{12}+\pi_{22})\right]}{1-\left[(\pi_{11}+\pi_{12})(\pi_{11}+\pi_{21})+(\pi_{21}+\pi_{22})(\pi_{12}+\pi_{22})\right]}.\end{aligned}$$

(14)

Replacing 1 by $(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})$ in (14), we have

$$\begin{aligned}\kappa_C &= \frac{\left[\left(\pi_{11}(\pi_{12}+\pi_{21})+\pi_{22}(\pi_{12}+\pi_{21})\right)+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2\right]-(\pi_{12}+\pi_{21})}{\left[\left(\pi_{11}(\pi_{12}+\pi_{21})+\pi_{22}(\pi_{12}+\pi_{21})\right)+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2\right]} \\ &= \frac{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21})+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2-(\pi_{12}+\pi_{21})}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21})+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2} \\ &= \frac{(\pi_{11}+\pi_{22}-1)(\pi_{12}+\pi_{21})+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21})+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2}.\end{aligned}$$

(15)

Again, we substitute $1=(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})$ in (15) then the result is written as

$$\kappa_C = \frac{2(\pi_{11}\pi_{22}-\pi_{12}\pi_{21})}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21})+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2}.$$

(16)

We now consider the right hand side of (16), the denominator is always positive, that is $(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21})+2\pi_{11}\pi_{22}+\pi_{12}^2+\pi_{21}^2 > 0$. For nominator, recall that $\pi_{12}+\pi_{21}=p_0$ where $p_0 >> 0.50$ and $\pi_{12} \neq 0$ (or $\pi_{21} \neq 0$), so we can obviously see that $\pi_{12}\pi_{21} > \pi_{11}\pi_{22}$. That is, $\kappa_C < 0$.

Next, we prove a case that $\kappa_M < 0$. Since

$$\kappa_M = \frac{\theta_o-\theta_e}{\pi_{1\cdot}\pi_{2\cdot}+\pi_{\cdot 1}\pi_{\cdot 2}}$$

$$= \frac{(\pi_{11}+\pi_{22}) - \left[(\pi_{11}+\pi_{12})(\pi_{11}+\pi_{21}) + (\pi_{21}+\pi_{22})(\pi_{12}+\pi_{22})\right]}{\left[(\pi_{11}+\pi_{12})(\pi_{21}+\pi_{22}) + (\pi_{11}+\pi_{21})(\pi_{12}+\pi_{22})\right]}.$$

(17)

Then $(\pi_{11}+\pi_{22})$ is replaced by $(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22}) - (\pi_{12}-\pi_{21})$, equation (17) can be written in the form

$$\kappa_M = \frac{\left[\left(\pi_{11}(\pi_{12}+\pi_{21}) + \pi_{22}(\pi_{12}+\pi_{21})\right) + 2\pi_{11}\pi_{22} + \pi_{12}^2 + \pi_{21}^2\right] - (\pi_{12}+\pi_{21})}{\left[(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + 2\pi_{12}\pi_{21}\right]}$$

$$= \frac{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + \pi_{12}^2 + \pi_{21}^2 - (\pi_{12}+\pi_{21})}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + 2\pi_{12}\pi_{21}}$$

$$= \frac{(\pi_{11}+\pi_{22}-1)(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + \pi_{12}^2 + \pi_{21}^2}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + \pi_{12}^2 + \pi_{21}^2}.$$

(18)

When we make this substitution using the equation $1 = (\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})(\pi_{11}+\pi_{12}+\pi_{21}+\pi_{22})$, equation (18) thus becomes

$$\kappa_M = \frac{2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + 2\pi_{12}\pi_{21}}.$$

As the above reason, that is, since $\pi_{12}+\pi_{21} = p_0$ where $p_0 >> 0.50$ and $\pi_{12} \neq 0$ (or $\pi_{21} \neq 0$), then $\pi_{12}\pi_{21} > \pi_{11}\pi_{22}$. Therefore,

$$\frac{2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21})}{(\pi_{11}+\pi_{22})(\pi_{12}+\pi_{21}) + 2\pi_{11}\pi_{22} + 2\pi_{12}\pi_{21}} < 0. \quad \blacksquare$$

On the other hand, if the proportion of units in which the two raters disagree $\pi_{12}+\pi_{21} << 0.50$, or say that $\pi_{11}+\pi_{22} >> 0.50$, then all kappa statistics κ_C and κ_M are greater than zero as given in the following claim.

Claim 2: *Suppose* p_0 *be an arbitrary probability that satisfy* $p_0 >> 0.50$, $\pi_{11}+\pi_{22} = p_0$ *and* $\pi_{11} \neq 0$ *(or* $\pi_{22} \neq 0$*), then* $\kappa_C, \kappa_M > 0$.

Proof: The proofs of Claim 1 and Claim 2 are similar. If $\pi_{11}+\pi_{22} = p_0$ where $p_0 >> 0.50$ and $\pi_{11} \neq 0$ (or $\pi_{22} \neq 0$), then $\pi_{12}\pi_{21} < \pi_{11}\pi_{22}$. Clearly, $\kappa_C, \kappa_M > 0$. So the proof of this claim is completed. $\blacksquare$

CONCLUSION AND DISCUSSION

This research presents the statistical inference of assessing agreements with methods utilized to evaluate consistencies/inconsistencies in findings from different data-generating sources that collect the same or similar information. Measurements of agreement between raters are of great

importance in many fields, both scientific and non-scientific. When comparing a new method of measurement with an old standard method, one of the things we want to know is whether they agree sufficiently for the new to replace the old. If the new method agrees sufficiently well with the old, the old may be replaced. This makes sure that the new method of measurement is cheap, quick, simple, accurate and optimal. In addition, measuring agreement has been used very often to designate the level of agreement between two raters. In addition, in this research we focused on some characteristics of the Cohen's kappa statistic and the modified kappa statistic.

ACKNOWLEDGMENTS

Our sincere thanks are due to Assoc. Prof. Dr. Montip Tiensuwan at the Department of Mathematics, Faculty of Science, Mahidol University, for suggesting this problem. A special gratitude is expressed to Prof. Dr. Bikas K. Sinha from the Indian Statistical Institute, Kolkata, India, for his expert and excellent guidance and his useful comments. Furthermore, we also are particularly indebted to the Kasetsart University Research and Development Institute (KURDI), Kasetsart University, Thailand for the financial support which has enabled us to undertake this research.

REFERENCES

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement. 20(1): 37-46.

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

Sinha, B.K., P. Yimprayoon and M. Tiensuwan. 2006. Cohen's Kappa Statistic: A Critical Appraisal and Some Modifications. Calcutta Statistical Association Bulletin. 58: 151-169.