การศึกษาบางอย่างเพิ่มเติมสำหรับตัวสถิติโคเฮนแคปปา

Some Further Study Based on the Cohen's Kappa Statistic

พรพิศ ยิ้มประยูร[1] และสิทธิพงศ์ รักตะเมธากูล[1]

Pornpis Yimprayoon[1] and Sittipong Ruktamatakul[1]

บทคัดย่อ

ในงานวิจัยนี้ ผู้วิจัยจะมุ่งเน้นพิจารณาปัญหาทางสถิติของการวัดความเห็นพ้องต้องกันระหว่างสองผู้สังเกตการณ์ซึ่งใช้วิธีการวัดแบบมาตรานามบัญญัติที่มี 2 ระดับ วัตถุประสงค์ของงานวิจัยนี้คือ การศึกษาคุณลักษณะบางอย่างเพิ่มเติมของตัวสถิติโคเฮนแคปปา นอกจากนี้ วิธีดำเนินการทดสอบที่ผู้วิจัยนำเสนอได้ถูกอธิบายโดยการประยุกต์กับตัวอย่างข้อมูล

คำสำคัญ : การวัดความเห็นพ้องต้องกัน ตัวสถิติโคเฮนแคปปา มาตรานามบัญญัติ การทดสอบสมมติฐาน

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 show some further characteristics on Cohen's kappa statistic. Moreover, our proposed test procedure is also illustrated using an application to a data example.

Key Words : measuring agreement, Cohen's kappa, nominal scale, hypothesis testing

E-mail : faasppy@ku.ac.th

INTRODUCTION

Researchers in many fields have become increasingly aware of the problem of errors in measurements. The investigations into the scientific bases of measurement errors began over one and a half centuries ago. 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 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. In this research, we will show some further study on Cohen's kappa statistic when the proportion of rater pairs exhibiting agreement (θ_o) is given.

[1] สายวิชาคณิตศาสตร์ คณะศิลปศาสตร์และวิทยาศาสตร์ มหาวิทยาลัยเกษตรศาสตร์ วิทยาเขตกำแพงแสน นครปฐม 73140
Department of Mathematics, Faculty of Liberal Arts and Science, Kasetsart University, Kamphaeng Saen Campus, Nakhonpathom 73140, Thailand

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 that most of this interest has focused on high measures of agreement. It would indicate consensus in the judgments by the two raters. In this section, we will propose another approach to test whether of 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, we wish to test the hypothesis

$$H_0 : \kappa_C \geq \kappa_0 \text{ versus } H_1 : \kappa_C < \kappa_0, \quad (1)$$

where κ_0 is a predetermined high value.

Since $\kappa_C \geq \kappa_0$,

or

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

or equivalently,

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

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

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

or

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

or equivalently,

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

Therefore, we obtain $\frac{\kappa_0(1-\theta_o)}{1-\kappa_0} \leq 2\left[\pi_{11}(\theta_o - \pi_{11}) + \pi_{12}(\theta_o + \pi_{12} - 1)\right]$.

We then have the graph of

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

or equivalently,

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

(2)

We note that (1) will be meaningful if $\kappa_0 \leq \frac{\theta_o^2}{1+(1-\theta_o)^2}$. If $\kappa_0 = \frac{\theta_o^2}{1+(1-\theta_o)^2}$, then we have that $\pi_{11} = \pi_{22} = \frac{\theta_o}{2}, \pi_{12} = 1-\theta_o, \pi_{21} = 0$ or $\pi_{11} = \pi_{22} = \frac{\theta_o}{2}, \pi_{12} = 0, \pi_{21} = 1-\theta_o$.

Given θ_o, we generate data on raters' judgement as follows:

We 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 we ask the raters to confine to the agreement cells $[(1,1)$ and $(2,2)]$ for the without loss of generality, we 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, we will consider an illustrative example and examine such situation below.

NUMERICAL RESULT

Let us consider a hypothetical reliability research scenario in Table 1. Raters A and B have classified 200 subjects into one of 2 possible response categories each.

Table 1: 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 2 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 1.

Table 2: 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

We are interested in testing the hypotheses that

$$H_0 : \kappa_C \geq 0.80$$

$$H_1 : \kappa_C < 0.80$$

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 $\theta_o = 0.94$ is given, in this case we obtain that

$$n = 200,$$

$$n_0 = n\theta_o = 200 \times 0.94 = 188,$$

$$\kappa_0 = 0.80,$$

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 (2), we have the graph of

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

or the graph of

$$(\pi_{12}^2 - \pi_{11}^2) + 0.94(\pi_{11} + \pi_{12}) - \pi_{12} \geq 0.12. \quad (3)$$

We can plot the (π_{11}, π_{12})-plane that satisfies the inequality (3) 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, we can consider the test of

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

against

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

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

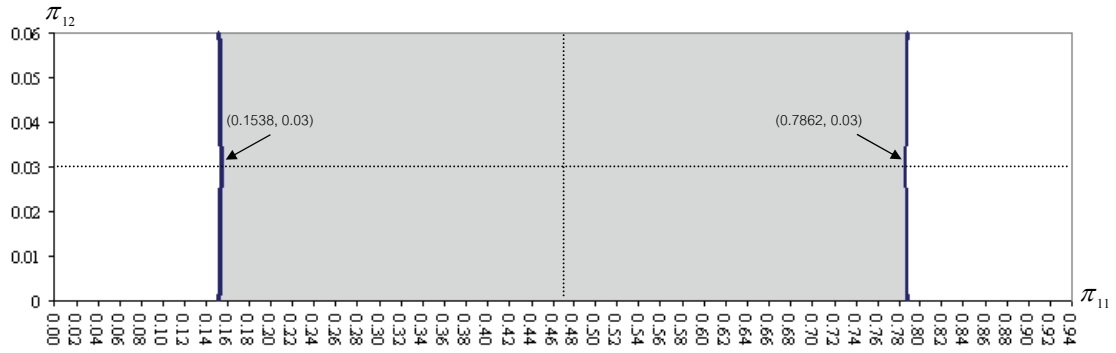

Figure 1. (π_{11}, π_{12})-plane for the test of agreement

The test statistic is given by

$$z = \frac{\hat{\pi}_{11} - 0.1538}{\sqrt{\hat{V}ar(\hat{\pi}_{11})}} \quad (4)$$

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 $z \le -z_\alpha$ at $100\alpha\ \%$ level of significance.

After we our test procedures to analyze this example, we then get that $\hat{\pi}_{11} = 0.40, \hat{\pi}_{22} = 0.54, \hat{V}ar(\hat{\pi}_{11}) = 0.0011$ and $z = 7.4232$.

If we use a significance level $\alpha = 0.05$, the critical region is $z \le -1.645$. So we conclude that π_{11} is not significantly less than 0.1538. That is, $\kappa_C \ge 0.80$. Further, a 95% lower confidence limit based on $\hat{\pi}_{11}$ for π_{11} is $\hat{\pi}_{11} - z_\alpha \sqrt{\hat{V}ar(\hat{\pi}_{11})} = 0.40 - 1.645\sqrt{0.0011} = 0.3454$.

CONCLUSION AND DISCUSSION

In this research, we discussed the problem of measuring agreement or disagreement between two raters while the ratings are given separately in 2-point nominal scale. We focused on some further study on Cohen's kappa statistic when θ_o is given. In addition, our proposed test procedure is also explained using data example.

ACKNOWLEDGMENT

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. 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.

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

Fleiss, J.L. 1981. Statistical methods for raters and proportions. John Wiley and Sons Inc, New York.

Altman, D.G. 1991. Practical Statistics for Medical Research. Chapman and Hall, London.

Lantz, C.A. and E. Nebenzahl. 1996. Behavior and interpretation of the kappa statistic: Resolution of the two paradoxes. Journal of Clinical Epidemiology. 49(4): 431-434.