Comments on Statistical Issues in May 2014

Commentary

Yong Gyu Park

Department of Biostatistics, The Catholic University of Korea College of Medicine, Seoul, Korea

In this section, we explain Cohen's kappa coefficient, a measure of agreement, which appeared in the article titled, 'Impact of Clinical Performance Examination on Incoming Interns' Clinical Competency in Differential Diagnosis of Headache', by Park et al.¹⁾ published in March 2014.

KAPPA, A MEASURE OF AGREEMENT

Cohen's kappa coefficient is a statistical measure of inter-rater agreement or inter-instrument agreement for qualitative (categorical) items. It is generally thought to be a more robust measure than a simple percent agreement calculation since kappa takes into account agreement occurring by chance. A chance-corrected measure originally introduced by Scott,²⁾ was extended by Cohen³⁾ and has come to be known as Cohen's kappa. It comes from the notion that the observed cases of agreement include some cases for which the agreement was by chance alone.

Let us assume that there are two raters, who independently rate n subjects into one of two mutually exclusive and exhaustive nominal categories. Let p_{ij} be the proportion of subjects that are placed in the i, j th cell, i.e., assigned to the i th category by the first rater and to the j th category by the second rater ($i, j = 1, 2$). Also, let $p_{i+} = p_{i1} + p_{i2}$ denote the proportion of subjects placed in the i th row (i.e., the i th category by the first rater), and let $p_{+j} = p_{1j} + p_{2j}$ denote the proportion of subjects placed in the j th column (i.e., the j th category by the second rater). Then the kappa coefficient is

$$\kappa = \frac{(p_o - p_e)}{(1 - p_e)},$$

where $p_o = p_{11} + p_{22}$ is the observed proportion of agreement

and $p_e = p_{1+}p_{+1} + p_{2+}p_{+2}$ is the proportion of agreement expected by chance.

If there is complete agreement, $\kappa = 1$. If observed agreement is greater than or equal to chance agreement, $\kappa \geq 0$, and if observed agreement is less than chance agreement, $\kappa < 0$. The minimum value of κ depends on the marginal proportions. If they are such that $p_e = 0.5$, then the minimum equals -1 . Otherwise, the minimum is between -1 and 0 .

Example. Two doctors independently classified 100 people into one of two diagnostic categories, abnormal/normal as follows.

Table 1. Diagnoses on $n = 100$ people by two doctors

		Doctor B	
		Abnormal	Normal
Doctor A	Abnormal	40	10
	Normal	20	30

$$\text{Observed agreement } (p_o) = (40 + 30) / 100 = 0.7$$

$$\begin{aligned} \text{Chance agreement } (p_e) &= (40 + 10) / 100 \times (40 + 20) / \\ &\quad 100 + (20 + 30) / 100 \times (10 + 30) / 100 \\ &= 0.5 \times 0.6 + 0.5 \times 0.4 = 0.3 + 0.2 = 0.5 \\ \kappa &= (0.7 - 0.5) / (1 - 0.5) = 0.4. \end{aligned}$$

However, we should not use kappa as a measure of agreement when all raters or devices cannot be treated symmetrically. When one of the sources of ratings may be viewed as superior or a standard, e.g., one rater is senior to the other or one medical device is more precise measuring instrument than the other, kappa may no longer be appropriate.

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article was reported.

REFERENCES

1. Park SM, Song YM, Kim BK, Kim H. Impact of clinical performance examination on incoming interns' clinical competency in differential diagnosis of headache. Korean J Fam Med 2014;35:56-64.
 2. Scott WA. Reliability of content analysis: the case of nominal scale coding. Public Opin Q 1955;19:321-5.
 3. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960;20:37-46.
1. Park SM, Song YM, Kim BK, Kim H. Impact of clinical