Korean J Fam Med. 2014;35:325-326 http://dx.doi.org/10.4082/kjfm.2014.35.6.325

Comments on Statistical Issues in November 2014

Commentary

Yong Gyu Park

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

In this section, we continue to explain Cohen's kappa coefficient, which was described in the commentary titled, "Comments on statistical issues in May 2014", by Park[1] published in May 2014, and its related topics.

KAPPA, DIAGNOSTIC ACCURACY, AND PARADOX OF KAPPA

Cohen's kappa coefficient is a measure of inter-observer or inter-device agreement for qualitative (categorical) items used in many scientific fields; it accounts for the possibility that the agreement occurred by chance. However, as we mentioned previously,[1] kappa should not be used 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 than the other), kappa may no longer be appropriate.

The term diagnostic accuracy, which is usually expressed by sensitivity, specificity, and positive and negative predictive values, presupposes that there is an underlying gold-standard (true values). Therefore, these two terms cannot be used interchangeably to explain the same phenomenon. Additionally, 'over-estimate' and 'under-estimate' cannot be used together with 'kappa,' but 'concordance' and 'correlation' can be shared with 'kappa.'

The interpretation of kappa has been under severe criticism. A well-known problem has been referred to as 'the first paradox of kappa' by Feinstein and Cicchetti.[2] We usually expect the value of kappa to be high when both observers assess one of several categories with a high probability, and low when both observers assess all categories evenly. However, the greater the imbalance in the marginal distributions of each category, the higher the probability of chance agreement; in this case, the magnitude of kappa is reduced considerably, even when the observed agreement is quite high. Such imbalanced marginal distributions often occur when the sample data are obtained from a population with a very low prevalence of the disease under consideration.

For this reason, 'kappa' is considered an overly conservative measure of agreement. It seems to be more appropriate to use the interclass kappa[3] and the first-order agreement coefficient[4] as measures of agreement for analyses of data with severely imbalanced marginal distributions.

CONFLICT OF INTEREST

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

REFERENCES

1. Park YG. Comments on statistical issues in May 2014. Korean J Fam Med 2014;35:167-8.
2. Feinstein AR, Cicchetti DV. High agreement but low kappa: I. the problems of two paradoxes. J Clin Epidemiol

1990;43:543-9.
3. Bloch DA, Kraemer HC. 2 × 2 kappa coefficients: measures of agreement or association. Biometrics 1989;45:269-87.
4. Gwet KL. Handbook of inter-rater reliability. Gaithersburg: STATAXIS Publishing Company; 2001.